Encoding Values: Injecting Morality into Machines via Prompt-Conditioned Moral Frames

Arjun Balaji
Independent Researcher
arjun.balaji.02@gmail.com

Aarushi Nema
Independent Researcher
aarushi.nema02@gmail.com

Neha Kamath
Independent Researcher
nehakamathc@gmail.com

Jyothika Raju
Independent Researcher
jyothikaraju7@gmail.com

Abstract

Large language models (LLMs) are typically aligned to a single, universal policy, obscuring the rich plurality of human moral perspectives that drive creative practice. We present a study of prompt-level moral steering in large language models with ten-principle constitutions. Seven constitutions matched in length (Intersectional Feminist, Ecological Justice, Ubuntu, Indigenous Sovereignty, Universal Human Rights, Neutral, and a Random Adjective placebo) are paired with a 100-question benchmark spanning Everyday Advice, Policy Scenarios, Normative Dilemmas, and Rewrite Tasks. We generated 2800 completions across four models (GPT-4o-mini, Llama-3.1-8B, Llama-4 Scout-17B, Qwen-3-235B) and evaluate them with four complementary metrics: toxicity, semantic alignment with the canon of each constitution, lexical marker ratio, and an LLM-as-a-Judge composite of authenticity, helpfulness, and safety. Prompted moral frames yield consistent improvements in alignment and judged quality without increasing toxicity; placebo prompts do not. Effects replicate across models and are strongest on morally charged tasks. The open-source toolkit lets researchers or artists author new frames in minutes, supporting participatory, culturally adaptive AI. Our results show that pluralistic prompting is a practical lever for value-conditional behavior and a fertile instrument for creative AI exploration. The code and results obtained in this study is available at: https://github.com/ArjunBalaji79/encoding_values_neurips.

1 Introduction

Generative language models are now active partners in artistic creation, design ideation, and public discourse. Their behavior is typically governed by a single, universal alignment policy—a fixed reward model or "constitution" distilled from a largely homogeneous pool of annotators and organizational guidelines. This one-size-fits-all approach streamlines deployment but blurs the plurality of human ethics that undergird creative practice. A chatbot that offers design feedback in Berlin, collaborates on climate-justice poetry in Nairobi, and moderates a cultural-heritage workshop in Wellington will encounter moral expectations that diverge along axes of gender equity, ecological responsibility, Indigenous sovereignty, communal harmony, and liberal rights. Alignment debates often collapse diverse human values into a single "universal" objective. Yet communities articulate ethics differently: feminist theory foregrounds power and intersectionality; ecological perspectives emphasize interdependence and planetary stewardship; decolonial and Indigenous sovereignty lenses center historical context and collective well-being.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

The present study treats value pluralism as an empirical question: How much does an identical model's behavior change when prompted with distinct, carefully balanced moral constitutions? To investigate, we introduce a plural-constitution benchmark composed of:

- Seven token-parity constitutions each expressed in ten numbered principles.
- A 100-query evaluation set that spans four splits : Everyday Advice, Policy Scenarios, Normative Dilemmas, and Rewrite tasks.

By demonstrating that explicitly plural constitutions induce behavioral divergence in a state-of-the-art LLM, the work provides a concrete foundation for user-selectable or community-specified moral frames, an essential ingredient for humane and culturally adaptive AI systems.

2 Related Work

Early commercial systems such as InstructGPT rely on reinforcement learning from human feedback (RLHF) to fit a single reward model for all users, doubling human-preference scores over plain GPT-3 but leaving values effectively frozen [Ouyang et al., 2022]. Anthropic's Constitutional AI improves safety by running a critique-and-revision loop over 16 hand-written rules, cutting harmful completions by about 65 percent, yet it still assumes one "correct" moral frame [Bai et al., 2022].

Efforts to democratize that frame include Collective Constitutional AI: Aligning a Language Model with Public Input, which distills almost 20,000 lay-citizen proposals into a single merged constitution [Huang et al., 2024], and Public Constitutional AI, which envisions a legally drafted charter that likewise applies to everyone [Abiri, 2025].

Complementary work shows that large models can shift stance when prompted with cultural cues [Xu et al., 2025] and that padding prompts with semantically neutral text can worsen reasoning [Levy et al., 2024]; we therefore keep every frame within a set number of tokens to isolate semantic effects. Whereas earlier approaches either (i) impose one universal charter or (ii) collapse diverse opinions into a single aggregate document, our study maintains multiple worldviews side by side and quantifies how their answers diverge.

3 Methods

3.1 Constitutions

We authored seven concise textual constitutions that include Intersectional Feminist, Ecological Justice, Ubuntu, Human Rights, Indigenous Sovereignty, Neutral, Random Adjective Control. For the constitutions, refer to the repository provided in Appendix A.

3.1.1 Design Principles

Each constitution follows a uniform template: ten numbered principles concise enough for a single system prompt yet detailed enough to guide both tone and moral reasoning. To prevent prompt length from becoming a confounding factor, every file averages about 120 tokens, achieving close token parity across frames. We further anchor each frame in its intellectual heritage by citing seminal thinkers for example, bell hooks for the feminist lens [Hooks, 2000] and Desmond Tutu for Ubuntu, which helps ground vocabulary and reduces the risk of "word-salad" drift. For robustness, we sample twenty-five responses for every combination of model, value framework, and task, producing a total of 2800 evaluations. Even the neutral and placebo controls follow this exact format to ensure that only semantic differences, not length or punctuation, drive downstream behavior.

3.1.2 Seven Constitutions

Seven distinct ethical frameworks were tested. The Intersectional Feminist framework, rooted in bell hooks' intersectional analysis, aimed to address power imbalances and ensure equitable treatment across diverse social identities by identifying and mitigating biases affecting marginalized groups. The Ecological Justice framework, adopting an Earth-systems thinking approach, emphasized climate justice and the interconnectedness of all living things, focusing on AI's positive contribution to

environmental sustainability. Drawing from the African ethic of "shared humanity," the Ubuntu framework prioritized community, compassion, and collective interconnectedness, aiming to foster AI systems that promote social cohesion and human flourishing. Based on the principles of the 1948 Universal Declaration of Human Rights, the Human Rights framework ensured AI development respects fundamental human dignity, liberty, and equality, preventing infringement upon individual rights. Guided by Linda Tuhiwai Smith's concept of kaupapa [Smith, 2021], the Indigenous Sovereignty framework championed self-determination and the protection of Indigenous knowledge systems and ways of being, ensuring AI respects Indigenous communities' autonomy and cultural heritage. The baseline frameworks included Neutral Long and Random Adjective. The Neutral Long framework served as a control, focusing on fundamental principles of clarity and courtesy in AI communication, establishing a baseline for general good conduct without specific ethical stances. The whimsical Random Adjective framework provided a contrasting, less formally structured control. Its arbitrary focus on punctuality and "blueberry appreciation" highlighted the importance of well-defined ethical principles by demonstrating what a lack of such principles might look like in practice.

3.2 Query Benchmark

We created four query sets: Everyday Advice, Normative Dilemmas, Policy Scenarios, and Rewrite Tasks. These span mundane choices to complex socio-political tradeoffs, ensuring both breadth and depth of moral reasoning challenges. Each query is 10 – 25 tokens and omits frame keywords ("feminist", "Ubuntu", etc.) to prevent trivial lexical matching. For the complete set of queries, refer to the repository in Appendix A.

Table 1: Core splits for value conditioning tasks. Each split contains 25 examples.

Split	Purpose	Example ID (abridged)
Everyday Advice (25)	Emotional labour, craft, interpersonal rituals	EA11 "… ask neighbour to reduce noise"
Policy Scenarios (25)	Governance of AI labour, energy, IP	PS13 "ban facial-recognition in public spaces"
Normative Dilemmas (25)	Direct moral conflicts (authorship vs. equity, land vs. profit)	DL03 "mine rare earths on sacred land?"
Rewrite Tasks (25)	Pure voice transfer; same fact, new framing	RW07 "Humans have dominion over all species"

Together, the seven parity-matched constitutions and 100 balanced queries provide a minimal yet rigorous playground for testing whether large language models can exhibit distinct and interpretable world-views.

4 Framework

4.1 Generation

The open-source framework offers a single CLI and backend wrappers (OpenAI, Anthropic, HuggingFace, Ollama, Modal, Cerebras) enable easy generation. For each (query, frame) pair we allow options to sample n responses at temperature 0.7 and log outputs. It also allows for simple testing of new constitutions ensuring reproducibility.

4.2 Evaluation

We assessed model outputs using four metrics: Toxicity, Semantic Distance, Lexical Marker Ratio, and an LLM-as-a-Judge [Li et al., 2024] composite score to provide a multi-perspective evaluation suite.

Toxicity was detected with the unitary toxic-bert [Hanu and Unitary team, 2020] classifier. The score ranges from 0 (completely safe) to 1 (maximally toxic). A probability threshold of 0.5 separated toxic from non-toxic responses, so lower values correspond to safer and more civil language.

Semantic Distance measured how faithfully each response adhered to a target value framework. We embedded responses and their corresponding canonical paragraphs using SentenceTransformer all-

mpnet-base-v2 [Song et al., 2020]and computed 1 minus the cosine similarity between embeddings. As cosine similarity lies in [–1, 1], this yields a distance range of 0 (perfect alignment) to 2 (complete opposition); smaller distances therefore indicate tighter semantic alignment.

The Lexical Marker Ratio captured overt framing cues. For every value framework we compiled a lexicon of 10–30 high-precision tokens drawn from the literature, then reported the proportion of words in a response that matched this lexicon. Lower ratios imply subtler, less explicitly signposted framing.

Finally, the overall quality was judged by GPT-4.1 (temperature = 0.3). The model rated Authenticity (50 % weight), Helpfulness (30 %), and Safety (20 %) on a 1–5 scale, producing a composite score computed as 0.5 × Authenticity + 0.3 × Helpfulness + 0.2 × Safety. Higher composite values indicate outputs perceived as more authentic, useful, and safe. Together, these metrics furnish a rigorous and diversified picture of system performance, encompassing safety, semantic fidelity, stylistic signaling, and human-perceived utility.

5 Results

5.1 Effects of Moral Framing on Models

Across splits, coherent value frames reliably improved quality and alignment over both baselines. Table 2 is the frame-by-frame results for the GPT 4o-mini model [OpenAI et al., 2024] across all evaluation metrics. On Normative Dilemmas, value-aligned frames achieved high LLM-as-Judge scores (e.g., Ecological 4.976; Feminist 4.988; Human Rights 4.988) while Indigenous Sovereignty produced the tightest semantic adherence (0.399 distance), all at very low toxicity; the placebo Random-Adjective underperformed markedly (LLM-Judge 2.864). In Policy Scenarios, value frames again exceeded baselines on LLM-as-Judge scores(Ecological 4.872; Feminist 4.872; Ubuntu 4.860; Human Rights 4.848), with Indigenous Sovereignty showing both strong semantic pull (0.426) and the highest lexical marker ratio (0.131), indicating clear, but not excessive, on-frame signaling.

On Everyday Advice, Ubuntu and Human Rights reached the ceiling score (5.0), while the placebo both scored lower (3.368) and showed the highest toxicity observed in the table (0.0077), reinforcing that filler text does not substitute for principled frames; the neutral control remained strong (4.98). Rewrite tasks proved more challenging than the other splits, so performance gains were modest yet consistent: Ubuntu (4.09), Feminist (4.01), and Human Rights (4.00) all outperformed the neutral control (3.61), while the Random-Adjective placebo trailed at 2.50. Even in this harder setting, moral conditioning still steers both what the model says and how it says it, boosting on-frame semantics and judged helpfulness / safety without any measurable safety regressions.

5.2 Cross-Model Generality

Our analysis of 2800 total evaluations across four models: GPT-4o-mini, Llama-3.1-8B [Grattafiori et al., 2024], Llama-4 [Meta AI, 2025], Qwen-3-235B [Qwen Team, 2025], seven value frameworks, and four task categories demonstrates strong evidence for the effectiveness of value conditioning through system prompts. Value-aligned frameworks consistently outperformed baseline conditions across multiple evaluation metrics.

The value-conditioning effect replicates across all four models as seen in Table 3. For morally explicit tasks, Indigenous Sovereignty consistently yields the tightest semantic adherence on both Normative Dilemmas and Policy Scenarios—for GPT-4o-mini (0.399; 0.426), Llama-3.1-8B (0.463; 0.453), Llama-4 Scout-17B (0.415; 0.430), and Qwen-3-235B (0.543; 0.527).

Across models, Random-Adjective (placebo) is reliably worst on judged quality, while Neutral-Long is a strong baseline that frames often surpass on LLM-Judge, especially in Normative and Policy splits. Toxicity remains low for value frames; when it rises, it is most evident under the placebo (e.g., GPT-4o-mini Everyday Advice toxicity 0.0077).

Task sensitivity is stable across models. On Everyday Advice, Ubuntu and Human Rights trend highest in judged quality for GPT-4o-mini (5.0/5.0), Llama-3.1-8B (4.73/4.75), and Llama-4 Scout-17B (4.80/4.74); Qwen also shows these frames above the neutral control (4.15–4.24 vs. 3.18). Rewrite Tasks show the smallest gains but still consistent cross-model improvements over neutral: GPT-4o-mini (Ubuntu 4.09 vs. 3.61), Llama-3.1-8B (Ubuntu 4.94 vs. 4.73), Llama-4 Scout-17B

Table 2: Evaluation summary by split and frame for GPT 4o-mini

Split	Frame	#Resp	Toxicity	Semantic	Lexical	LLM Judge
normative_dilemmas	indigenous_sovereignty	25	0.0006678896	0.3989210117	0.1243903474	3.9760
normative_dilemmas	random_adjective	25	0.0013628866	0.7407704407	0.0664258700	2.8640
normative_dilemmas	neutral_long	25	0.0007225176	0.9062633349	0.0141585647	4.8920
normative_dilemmas	ecological_justice	25	0.0007590884	0.5473841608	0.0830303487	4.9760
normative_dilemmas	intersectional_feminist	25	0.0009890494	0.5576995921	0.0541580082	4.9880
normative_dilemmas	ubuntu	25	0.0006699183	0.5888367617	0.0661861415	4.9760
normative_dilemmas	human_rights	25	0.0007759674	0.6237426150	0.1180741966	4.9880
policy_scenarios	indigenous_sovereignty	25	0.0006506725	0.4256700176	0.1311801571	3.9640
policy_scenarios	random_adjective	25	0.0010881692	0.6898523651	0.0665350153	2.7720
policy_scenarios	neutral_long	25	0.0006520353	0.8981218824	0.0114779773	4.7440
policy_scenarios	ecological_justice	25	0.0007695935	0.5889419091	0.0930056318	4.8720
policy_scenarios	intersectional_feminist	25	0.0008337921	0.5767837536	0.0593937471	4.8720
policy_scenarios	ubuntu	25	0.0006199450	0.6118481481	0.0609449954	4.8600
policy_scenarios	human_rights	25	0.0006497972	0.6374225402	0.1035917902	4.8480
everyday_advice	indigenous_sovereignty	25	0.0009036621	0.6923059136	0.0727619347	4.2600
everyday_advice	random_adjective	25	0.0077415239	0.5851289856	0.0691534469	3.3680
everyday_advice	neutral_long	25	0.0017773309	0.9362751506	0.0137808883	4.9800
everyday_advice	ecological_justice	25	0.0015332415	0.6959792580	0.0548878476	4.8160
everyday_advice	intersectional_feminist	25	0.0011749752	0.7481251794	0.0262530615	4.7960
everyday_advice	ubuntu	25	0.0011581853	0.8119085684	0.0309191146	5.0000
everyday_advice	human_rights	25	0.0021472529	0.8190064062	0.0484203013	5.0000
rewrite_tasks	indigenous_sovereignty	25	0.0006226506	0.5078109860	0.1290527638	2.8400
rewrite_tasks	random_adjective	25	0.0007987186	0.5177477395	0.0647503630	2.5000
rewrite_tasks	neutral_long	25	0.0006719671	0.7703745916	0.0289181704	3.6120
rewrite_tasks	ecological_justice	25	0.0006465967	0.4952156407	0.0755341190	3.5320
rewrite_tasks	intersectional_feminist	25	0.0006867756	0.5022057366	0.0672063410	4.0120
rewrite_tasks	ubuntu	25	0.0006075250	0.5325083816	0.0415243703	4.0920
rewrite_tasks	human_rights	25	0.0006148747	0.5244140887	0.0669869774	4.0040

Table 3: Raw scores (mean over value frames; excludes Neutral and Placebo) for each model and query type. Higher **LLM-Judge** and lower **Semantic Dist** indicate better performance.

Model	Query Type	LLM-Judge ↑	Semantic Dist ↓	Lexical Ratio
GPT-4o-mini	Normative Dilemmas	4.781	0.543	0.089
	Policy Scenarios	4.683	0.568	0.090
	Everyday Advice	4.774	0.753	0.074
	Rewrite Tasks	3.696	0.512	0.076
Llama-3.1-8B	Normative Dilemmas	4.406	0.592	0.061
	Policy Scenarios	4.451	0.590	0.068
	Everyday Advice	4.318	0.759	0.037
	Rewrite Tasks	4.598	0.418	0.087
Llama-4 Scout-17B	Normative Dilemmas	4.442	0.583	0.076
	Policy Scenarios	4.462	0.586	0.082
	Everyday Advice	4.434	0.76	0.042
	Rewrite Tasks	4.630	0.410	0.098
Qwen-3-235B	Normative Dilemmas	4.138	0.605	0.085
	Policy Scenarios	4.119	0.597	0.084
	Everyday Advice	3.895	0.615	0.080
	Rewrite Tasks	4.226	0.450	0.114

(Ubuntu 4.89 vs. 4.66), and Qwen-3-235B (Ubuntu 4.64 vs. 3.57). In conclusion, coherent moral frames generalize across architectures: they tighten on-frame semantics and raise perceived quality without measurable safety regressions, while the placebo increases variance and occasional toxicity.

6 Limitations

Our study adopts a lightweight, prompt-level intervention; while this demonstrates the feasibility of pluralistic steering, it stops short of deeper training-time alignment and thus may not correct model-internal biases that persist beyond the prompt window. Second, the evaluation suite relies on some

quantitative indicators like toxicity scores and lexical-marker ratios that capture only surface signals of harmful or value-laden language; richer insight will require qualitative annotation and community-led audits. Third, each constitution's canonical paragraph reflects the authors' interpretation of its worldview, so the prompts themselves carry subjective bias; future work should co-design these canons with the communities they represent. Finally, the findings are constrained by the specific models and domains tested here; different parameter scales, instruction-tuning regimens, or non-English contexts could yield different magnitudes of effect, and should be explored to validate the generality of prompt-conditioned moral framing.

7 Conclusion

This work demonstrates that prompt-conditioned constitutions can meaningfully steer LLM output while preserving safety. Across 2800 evaluations, value-aligned prompts improved semantic fidelity, stylistic signaling, and LLM-judged usefulness compared to neutral and placebo baselines; the gains were consistent over four architecturally diverse models and four task types. Crucially, no single frame dominated every metric: Ubuntu excelled at judged quality, Indigenous sovereignty at semantic adherence, highlighting the need for plural rather than one-size-fits-all alignment. By releasing an end-to-end toolkit, balanced benchmark, and starter constitutions, we invite communities to author, audit, and remix their own moral perspectives, turning alignment into a participatory, creative act. Future work will extend the framework to multilingual prompts, domain-specific tasks (e.g., code generation), and training-time fine-tuning, and will incorporate community-led qualitative reviews to complement the quantitative indicators used here. Encoding Values thus provides both a practical method and a conceptual foundation for culturally adaptive human-centered generative AI.

References

Gilad Abiri. Public constitutional ai, 2025. URL https://arxiv.org/abs/2406.16696.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas

Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay,

Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Bell Hooks. *Feminism is for everybody: Passionate politics*. Pluto Press, 2000.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417. ACM, June 2024. doi: 10.1145/3630106.3658979. URL `http://dx.doi.org/10.1145/3630106.3658979`.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024. URL `https://arxiv.org/abs/2402.14848`.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*, 2024.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`, April 2025. Accessed 2025-08-06.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne,

Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Linda Tuhiwai Smith. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing, 2021.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL https://arxiv.org/abs/2004.09297.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.350. URL https://aclanthology.org/2025.naacl-long.350/.

A Appendix

All experimental details such as results, constitutions and query sets are all available on the GitHub Repository: `https://anonymous.4open.science/r/encoding_values_neurips-4B94/`:

All model generations in this study were obtained through hosted inference APIs—primarily the Cerebras Cloud APIs and Azure OpenAI APIs without any fine-tuning or weight updates. Metric computations (toxicity, semantic similarity, lexical diversity, and the LLM-as-Judge score) were carried out on a Apple M2 Macbook Pro. To facilitate reproducibility and broader experimentation, the accompanying open-source repository already includes plug-and-play support for additional back-ends such as Modal serverless GPUs, Ollama for local models, Anthropic's API, Hugging Face Models, Azure OpenAI, and Cerebras client wrappers, allowing researchers to replicate or extend our results on their preferred hardware or cloud platforms with minimal setup.

Table 4: Evaluation summary by split and frame for Llama-3.1-8B

Split	Frame	#Resp	Toxicity	Semantic	Lexical	LLM Judge
normative_dilemmas	indigenous_sovereignty	25	0.0006066740048117935	0.4628818392753601	0.09237823342378107	3.5000000000000013
normative_dilemmas	random_adjective	25	0.000755168218165636	0.8867953725578264	0.021370643260801212	3.3920000000000003
normative_dilemmas	neutral_long	25	0.000773012510035187	0.9037181862443685	0.009072174593681951	4.064000000000002
normative_dilemmas	ecological	25	0.0006617620214819909	0.5912425023317337	0.06786074707388962	4.528000000000002
normative_dilemmas	feminist	25	0.0008546153688803315	0.5687638020515442	0.04642694025760748	4.600000000000001
normative_dilemmas	ubuntu	25	0.0006301056896336377	0.665777136683464	0.04441854370660264	4.788000000000001
normative_dilemmas	human_rights	25	0.0007161973882466554	0.6710934221744538	0.05475987342455571	4.612000000000001
policy_scenarios	indigenous_sovereignty	25	0.0005895118811167777	0.45276323080062864	0.11055696775135554	3.5040000000000001
policy_scenarios	random_adjective	25	0.0006517724599689245	0.8556461473926902	0.024370502711189842	3.3960000000000001
policy_scenarios	neutral_long	25	0.0006505206017754972	0.9019869042932988	0.006003679605500375	4.136000000000001
policy_scenarios	ecological	25	0.0006448522419668734	0.6107671427726745	0.07594515447488594	4.580000000000002
policy_scenarios	feminist	25	0.0007810202753171325	0.5719492065906525	0.044598026016215646	4.444000000000002
policy_scenarios	ubuntu	25	0.0005815221252851188	0.663010488152504	0.04060989057887033	4.86
policy_scenarios	human_rights	25	0.0006012695468962193	0.6534365111589432	0.0697438862321217	4.868000000000001
everyday_advice	indigenous_sovereignty	25	0.0009232888999395072	0.662423854470253	0.058558902944408776	3.7200000000000006
everyday_advice	random_adjective	25	0.001865614396519959	0.7836415684223175	0.03231381609938919	3.7920000000000001
everyday_advice	neutral_long	25	0.0017158107203431427	0.9290835973992944	0.010124568477758585	4.108000000000001
everyday_advice	ecological	25	0.0014390000025741756	0.7036003971472382	0.037726549915675735	4.244000000000015
everyday_advice	feminist	25	0.0018155512982048095	0.7318280711770058	0.024366748082134456	4.140000000000015
everyday_advice	ubuntu	25	0.0013957392261363566	0.8748478936403989	0.019374187505072773	4.732000000000001
everyday_advice	human_rights	25	0.0014407334383577108	0.8242834972217679	0.04687805549812706	4.752000000000001
rewrite_tasks	indigenous_sovereignty	25	0.0005989479692652821	0.35617815613746645	0.11312590217037366	3.7680000000000001
rewrite_tasks	random_adjective	25	0.0007200711476616561	0.4848185610771179	0.06760838350081477	3.0239999999999996
rewrite_tasks	neutral_long	25	0.0006350021110847592	0.7800884157419204	0.019499232867371968	4.732000000000001
rewrite_tasks	ecological	25	0.0006203743419609963	0.3629622280597687	0.10365415075087442	4.836000000000001
rewrite_tasks	feminist	25	0.0008755810535512865	0.44922972202301026	0.07459504800134101	4.704000000000001
rewrite_tasks	ubuntu	25	0.0005945633514784276	0.4115735137462616	0.04825091536107296	4.94
rewrite_tasks	human_rights	25	0.000594218906480819	0.5079808390140533	0.09373352996285059	4.740000000000001

Table 5: Evaluation summary by split and frame Qwen-3-235B

Split	Frame	#Resp	Toxicity	Semantic	Lexical	LLM Judge
normative_dilemmas	indigenous_sovereignty	25	0.0006406079535372555	0.542728328704834	0.11186998360690946	3.2640000000000002
normative_dilemmas	random_adjective	25	0.0006395660946145654	0.7608401549980044	0.07573016866654526	2.575999999999999
normative_dilemmas	neutral_long	25	0.0006880300189368427	0.8414051710069179	0.008890880221325712	3.5280000000000005
normative_dilemmas	ecological_justice	25	0.0006810503010638058	0.6567460733652115	0.07982580070837524	4.1320000000000014
normative_dilemmas	intersectional_feminist	25	0.001100166868418455	0.5477038252353669	0.05621017027080746	4.264000000000001
normative_dilemmas	ubuntu	25	0.0006624455214478076	0.6096894705295562	0.08890588955937981	4.496000000000001
normative_dilemmas	human_rights	25	0.0007664327509701252	0.670537588596344	0.09055540844125341	4.532000000000002
policy_scenarios	indigenous_sovereignty	25	0.0006001014914363622	0.5266818845272064	0.1165733575940013	3.1639999999999997
policy_scenarios	random_adjective	25	0.0005987806990742683	0.7848540880903602	0.06651286110304877	2.8560000000000003
policy_scenarios	neutral_long	25	0.0005984236183576286	0.8568537857756019	0.007391813662858715	3.5200000000000005
policy_scenarios	ecological_justice	25	0.0006207130337134004	0.6607048279047012	0.07999514045904613	4.172000000000001
policy_scenarios	intersectional_feminist	25	0.0008939078683033586	0.5251452004909516	0.055220780986337356	4.324000000000002
policy_scenarios	ubuntu	25	0.0006028541177511216	0.635027579665184	0.0745546369531723	4.588000000000002
policy_scenarios	human_rights	25	0.0006227734917774796	0.6388832998275756	0.0925163588620639	4.348000000000002
everyday_advice	indigenous_sovereignty	25	0.0006160997878760099	0.5396095290780067	0.10797791641298654	2.8679999999999994
everyday_advice	random_adjective	25	0.0006604895158670843	0.578296703696251	0.1344013063696833	2.5080000000000005
everyday_advice	neutral_long	25	0.0006865122402086854	0.8364240422844886	0.01229900676955479	3.176
everyday_advice	ecological_justice	25	0.000726557718589902	0.6467730453610421	0.07184806147020092	4.064000000000001
everyday_advice	intersectional_feminist	25	0.0008987299678847193	0.534305602312088	0.06255345653773457	4.152
everyday_advice	ubuntu	25	0.0006317732203751802	0.6774279773235321	0.08641693782111089	4.152
everyday_advice	human_rights	25	0.0006550234043970704	0.6789655068516731	0.07043534720758449	4.240000000000001
rewrite_tasks	indigenous_sovereignty	25	0.0006364048132672906	0.42195928931236265	0.1551699799718791	3.404
rewrite_tasks	random_adjective	25	0.001345887330826372	0.5592709231376648	0.16213549304373603	2.572
rewrite_tasks	neutral_long	25	0.0006282053142786026	0.6464789485931397	0.05364332945138327	3.5720000000000005
rewrite_tasks	ecological_justice	25	0.0006514650699682534	0.4767460536956787	0.10372599772565812	4.080000000000001
rewrite_tasks	intersectional_feminist	25	0.0007951212977059186	0.38793649911880496	0.08881861631970182	4.504000000000001
rewrite_tasks	ubuntu	25	0.0006360813160426915	0.4494035840034485	0.12005918601899117	4.644
rewrite_tasks	human_rights	25	0.0006293536839075387	0.5116241955757141	0.10219021034209007	4.496000000000002

Table 6: Evaluation summary by split and frame for Llama-4 Scout-17B

Split	Frame	#Resp	Toxicity	Semantic	Lexical	LLM Judge
normative_dilemmas	indigenous_sovereignty	25	0.0006205750000663102	0.4152366864681244	0.1194735058491159	3.5760000000000001
normative_dilemmas	random_adjective	25	0.0010290291463024915	0.7973086756467819	0.04240704916225277	2.7439999999999998
normative_dilemmas	neutral_long	25	0.0009351982153020799	0.8736878760531545	0.008350877222741753	4.232000000000002
normative_dilemmas	ecological_justice	25	0.0008626525592990219	0.6500601649284363	0.07364038208315271	4.572000000000002
normative_dilemmas	intersectional_feminist	25	0.0027768219029530882	0.5517028212547302	0.06029960317250675	4.580000000000002
normative_dilemmas	ubuntu	25	0.0006209852639585733	0.6097827595472336	0.04512777885287985	4.7360000000000015
normative_dilemmas	human_rights	25	0.0010145611339248716	0.690237573981285	0.0837488530227548	4.748000000000002
policy_scenarios	indigenous_sovereignty	25	0.0006003056582994759	0.4303393697738647	0.12665361644884546	3.7280000000000006
policy_scenarios	random_adjective	25	0.0008659384422935546	0.7935631799697876	0.04285057857910496	3.072
policy_scenarios	neutral_long	25	0.0007301218830980 36	0.9083588816598058	0.007017177664337483	4.2200000000000015
policy_scenarios	ecological_justice	25	0.0007261041295714677	0.6414631593227387	0.0783573058457095	4.448000000000002
policy_scenarios	intersectional_feminist	25	0.0011145088309422136	0.5316740226745605	0.06320440898178722	4.652000000000002
policy_scenarios	ubuntu	25	0.0005767681938596069	0.6620975688099862	0.039516342202083804	4.772000000000002
policy_scenarios	human_rights	25	0.0007018804573453962	0.6655499905347824	0.10140911964869266	4.708000000000001
everyday_advice	indigenous_sovereignty	25	0.0015641880547627807	0.6676873016357422	0.05589665444933401	3.7720000000000001
everyday_advice	random_adjective	25	0.003127778971102089	0.6996288891136646	0.05267728059366927	3.5160000000000001
everyday_advice	neutral_long	25	0.0023942541680298745	0.9353540455922484	0.0076275583058662425	4.2440000000000015
everyday_advice	ecological_justice	25	0.0015567438560537994	0.7263801001757383	0.03831499340493557	4.376000000000001
everyday_advice	intersectional_feminist	25	0.003204069670755416	0.7451474279165268	0.028640051256259756	4.480000000000001
everyday_advice	ubuntu	25	0.00216973772039637	0.8552702064812183	0.02432807618008275	4.796000000000001
everyday_advice	human_rights	25	0.0020716804428957403	0.8074951210618019	0.06284387150125599	4.7440000000000015
rewrite_tasks	indigenous_sovereignty	25	0.0005944034736603498	0.3665586125850677	0.12308097314068023	3.9440000000000001
rewrite_tasks	random_adjective	25	0.001345887330826372	0.43008899331092837	0.07977457354877945	2.724
rewrite_tasks	neutral_long	25	0.0006212530029006303	0.7780353850126267	0.02195180531860734	4.660000000000001
rewrite_tasks	ecological_justice	25	0.000639001454692334	0.3979495704174042	0.10996504329406 78	4.784000000000001
rewrite_tasks	intersectional_feminist	25	0.0009972675703465938	0.3941797137260437	0.07952812557439537	4.684000000000001
rewrite_tasks	ubuntu	25	0.00059125428320 84	0.3985258758068085	0.05279248541132386	4.892
rewrite_tasks	human_rights	25	0.0009748264984227717	0.49363348245620725	0.12227190895647876	4.844000000000001

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:Abstract and Introduction state exactly what the study contributes seven token-parity constitutions, 100-query benchmark, four metrics, four models and those claims are borne out in Results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:The "Limitations" section discusses prompt-only scope, metric proxies, canon bias, and model/domain generalization.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Paper is purely empirical; no theorems or formal proofs are presented.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Framework section provides CLI, back-end list, sampling parameters, and metric formulas. The provided codebase repository contains all scripts and experimental results logs to reproduce everything.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized GitHub link is present in the Abstract containing constitutions, queries, benchmark JSON, generation scripts, and evaluation notebooks.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: Section 3.2 Query Benchmark specifies splits and sizes; Section 4 lists temperature for LLMs. Table 2 shows each model × task; hyper-parameters are minimal and fully specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report per-cell means in Tables 1–2 but omit standard deviations, confidence intervals, and significance tests. All raw per-response scores are released in the supplementary CSV, enabling readers to increase sample sizes and compute their own uncertainty estimates.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification:Appendix A (supplement) lists GPU types, average runtime and memory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study follows NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Societal considerations like positive pluralistic alignment are discussed in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No large proprietary model weights or sensitive datasets are released; only small prompt files and scripts.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: References cite original model papers (Ouyang et al., Bai et al., etc.) and note API terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets are the seven constitutions and 100-query benchmark. All of these assets are provided in the codebase.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human participants or crowd work; all evaluations are automated.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable—no human-subject research

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 4.2 states that GPT-4o-mini was used only as an evaluator (LLM-as-a-Judge) with temperature 0.3; usage is integral to methodology and is fully described.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.