Dual Mechanisms of Value Expression: Decomposing Intrinsic and Prompted Values in Language Models

Jongwook Han* **Jongwon Lim*** **Injin Kong** **Yohan Jo†**
Graduate School of Data Science, Seoul National University
{johnhan00,elijah0430,mtkong77,yohan.jo}@snu.ac.kr

Abstract

While prompting is commonly used for assigning personas to LLMs, the fundamental question of how LLMs internally represent values remains unanswered. We observe that LLMs can express human values through two mechanisms: *intrinsic value expression* (inherent value-laden response patterns) and *prompted value expression* (value-laden response patterns following explicit instructions). We formalize these value expressions as feature directions in the model's residual stream and extract intrinsic and prompted value directions using the difference-in-means method. By comparing these directions, we investigate whether intrinsic and prompted value expressions rely on the same underlying mechanisms. Interventions using these directions show that both value directions can induce the model to express target values in its output. We find that even after removing the intrinsic value direction component from the prompted value direction, the remaining component can still steer the model's behavior. This suggests that while both directions produce similar outcomes, they use distinct neural mechanisms. Furthermore, we show that leveraging both intrinsic and prompted value direction is more effective for steering value expression than using either direction alone.

1 Introduction

Large language models (LLMs) can express values in different ways, either by reflecting the model's inherent preference or by following explicit instructions. For the first, which we call intrinsic value expression, LLMs develop consistent value expression patterns and generate human-like outputs through instruction-tuning and preference learning [17]. Consequently, LLMs consistently express certain values such as being harmless, helpful, and honest [1]. We refer to this fundamental behavioral pattern as the model's *intrinsic value expression*.

Conversely, for the second way, which we call *prompted value expression*, LLMs can express values following explicit instructions. However, this method has challenges, highlighted by the entire field of "prompt engineering" [21]. Moreover, it often causes critical failures, such as the Grok model referring to itself as "Mecha Hitler" after a system prompt update [2, 9]. To understand the underlying reason for these failures, we first need a mechanistic-level understanding of the model's value expression. Using Schwartz's theory of ten basic human values as a framework, we systematically investigate the mechanisms underlying both intrinsic and prompted value expression [23, 24].

We hypothesize that intrinsic and prompted value expressions use distinct mechanisms within the model's activation space. To test this, we formalize intrinsic and prompted value expression as a

*Equal contribution.
†Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

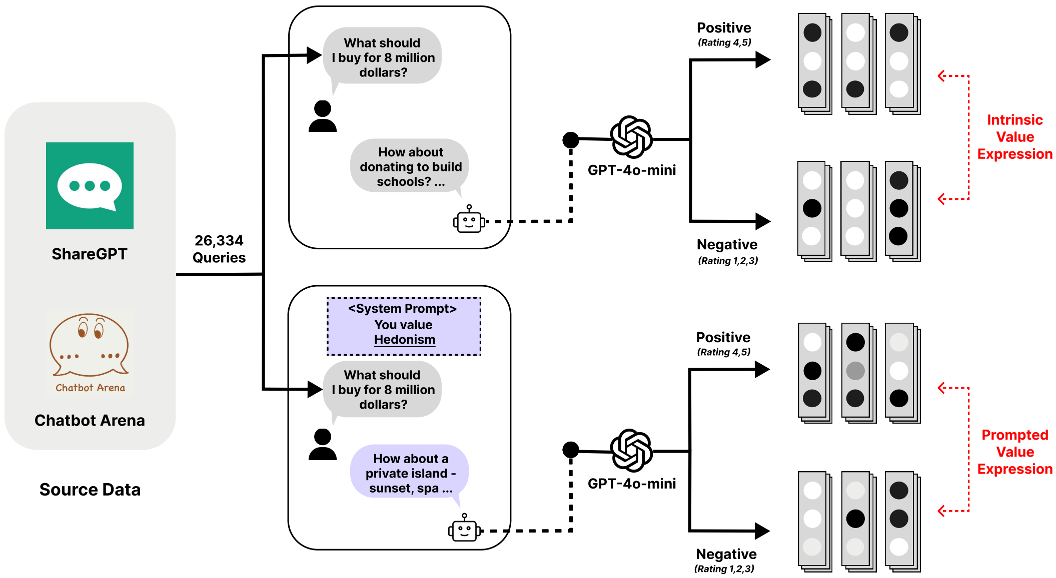

Figure 1: Overview of extracting **intrinsic** and **prompted** value directions.

feature direction in the residual stream, following the linear representation hypothesis [7]. We extract these directions using a difference-in-means approach by comparing model completions that express a target value and those that do not [15, 19].

Our key findings reveal that while intrinsic and prompted value directions exhibit moderate to high cosine similarity scores, they use distinct value expression mechanisms. Specifically, ablating intrinsic value directions from prompted value directions still retains steering effects, showing that each direction contains its own unique components for value expression. We also propose a steering method that combines both intrinsic and prompted value directions and show that it produces stronger steering effects on model value expression compared to using either direction alone.

2 Extracting intrinsic and prompted value directions

2.1 Dataset

As shown in Figure 1, we extract prompted and intrinsic value directions from a large-scale conversational dataset and evaluate them using custom-generated situational dilemmas.

Value-relevant query sets To generate responses in authentic conversational contexts, we use a curated version of the ShareGPT[1] and LMSYS-Chat-1M [10, 29]. The dataset consists of 26,334 first-turn English user queries sourced from human-LLM interactions that are relevant to Schwartz's values.

Situational dilemmas dataset While the value-relevant query dataset provides real conversational context, they do not always elicit responses where a specific value can be strongly expressed. To create an evaluation set that explicitly induces value-laden responses, we generate a dataset of situational dilemmas where different values conflict, similar to [3, 5, 12]. Using GPT-4.1-mini, we create 3,000 dilemma scenarios related to each of the ten Schwartz values. Each scenario is intended to induce two types of responses: one involves an action that explicitly prioritizes the target value (e.g., Security), while the alternative is a plausible course of action one could choose when not prioritizing the target value. Details are provided in Appendix A.

[1]`https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered`

Value eliciting prompts To extract prompted value directions, we use system prompts that elicit prompted value expressions. For prompt diversity, we use five different templates sourced from [11, 13, 22]. An example system prompt reads: "**Your Profile**: You value Achievement. Value Definition: Achievement: Achievement signifies the importance placed on personal mastery and success in competitive environments." To enhance diversity within each template, we use GPT-4o-mini to augment the value definition component, creating 100 variations per template, a total of 500 prompts per value. To extract prompted value directions, we randomly select one from the pool of 500 prompts. The details are in Appendix A.1.

2.2 Difference-in-means method

We formalize intrinsic and prompted value expressions as feature directions in the activation space. We extract these directions from the residual stream of each layer in Qwen2.5-7B-Instruct and the Llama-3.1-8B-Instruct model [8, 18].

For each of the ten Schwartz values, we extract two directions, a total of twenty directions: (1) Intrinsic value directions ($v_{\text{intrinsic}}$): directions extracted from the model's default responses, capturing its underlying value expressions. (2) Prompted value directions (v_{prompted}): directions extracted from responses guided by a system prompt (e.g., "You value benevolence"). These vectors capture the model's value expression mechanism, following the given persona. Both vectors are derived using the same difference-in-means process [15], detailed below.

The extraction process for a value direction (either $v_{\text{intrinsic}}$ or v_{prompted}) is as follows:

1. **Response generation:** We prompt the model with 26,334 queries from our value-relevant dataset and record the model's activations in all tokens of each generated response.
2. **Responses labeling:** We use GPT-4.1-mini to score each response on a five point scale (from "Strongly Opposes" to "Strongly Aligns") for its expression of the target value. We divide the responses into a positive set, S_{pos} (scores ≥ 4) and a negative set S_{neg} (scores ≤ 3).
3. **Difference-in-means calculation:** The steering vector v is the difference between the mean activation of the positive and negative sets:

$$v^L = \mathop{\mathbb{E}}_{x \in S_{\text{pos}}}[a^L(x)] - \mathop{\mathbb{E}}_{x \in S_{\text{neg}}}[a^L(x)] \tag{1}$$

   where $a_L(x)$ is the activation vector from layer L averaged over all token positions of the generated response for a given input query x.

Using the TransformerLens library [16], we extracted value directions on a server with dual Intel(R) Xeon(R) Silver 4310 @ 2.10GHz CPUs and four NVIDIA RTX A6000 GPUs, which required 32 hours to complete.

3 Value steering

To validate the vector extraction process, we steer the model's value expression by intervening activations along the directions of $v_{\text{intrinsic}}$ and v_{prompted}. At each token state, we simply scale and add v^L, a steering vector at layer L, such that $a^L = a^L + \alpha \cdot v^L$, where we set $\alpha = 1$ and apply steering on all layers.

Evaluation protocol We generate responses to the situational dilemma dataset as input to evaluate steering vectors. Specifically, for each value, we select 50 queries where the base responses had the lowest value-expression score, serving as a challenging set that effectively demonstrates the impact of the intervention.

We use the win ratio as the primary metric for evaluating steering effectiveness. For each situational dilemma, we generate three responses: one steered response and two baseline responses without steering, which differ based on the presence of a system prompt. An external LLM (GPT-4o-mini; see Appendix B for the prompt) then compares the steered response against each baseline and determines which better expresses the target value (win/tie/lose).

Steering is effective for both directions As shown in Table 1, interventions using $v_{\text{intrinsic}}$ and v_{prompted} successfully induce the model's value expression. In the value-related query dataset, interventions with $v_{\text{intrinsic}}$ and v_{prompted} achieved win ratios of 85.4% and 80.5% against the base model.

Table 1: Win ratios (%) of the steering experiments on the Llama-3.1-8B-Instruct model, averaged across ten Schwartz values. The scores are accompanied by the corresponding standard deviation and 95% confidence interval. Results for other models are provided in Appendix C.1.

	Intrinsic Direction	Prompted Direction	Intrinsic Orthogonal	Prompted Orthogonal	Mean Direction
vs Base	85.4	80.5	68.5	84.9	**89.6**
	(82.0, 88.3)	(76.7, 83.9)	(62.0, 74.4)	(80.1, 88.6)	(86.3, 92.1)
vs Base (w/ system prompt)	64.0	61.5	32.9	49.5	**67.1**
	(59.7, 68.2)	(57.1, 65.8)	(27.9, 38.3)	(44.0, 55.0)	(62.5, 71.3)

4 Analysis

To better understand these value directions, we investigate: Are intrinsic and prompted value directions different? We first calculate the pairwise cosine similarity between the intrinsic ($v_{\text{intrinsic}}$) and prompted (v_{prompted}) value directions. The results show a moderate to high degree of similarity. Specifically, for each of the ten Schwartz values, $v_{\text{intrinsic}}$ and v_{prompted} exhibit cosine similarity scores ranging from 0.27 to 0.85 in all layers. This suggests that $v_{\text{intrinsic}}$ and v_{prompted} might share a common directional component but they are not identical.

To focus on the difference between these directions, we isolate the unique contribution of each direction by removing the influence of the other. Specifically, we define the **prompted orthogonal component**, $v_{p\perp i} = v_p - \frac{v_p \cdot v_i}{\|v_i\|^2} v_i$ and the **intrinsic orthogonal component**, $v_{i\perp p} = v_i - \frac{v_i \cdot v_p}{\|v_p\|^2} v_p$, where v_p is the prompted direction, and v_i is the intrinsic direction. Table 1 shows that the orthogonal components $v_{p\perp i}$ and $v_{i\perp p}$ are both effective steering directions, although the effectiveness is smaller than v_i and v_p (except for the Prompted Orthogonal vs. Base case).

Motivated by the distinct mechanisms of intrinsic and prompted value directions, we test steering with their mean, $\frac{1}{2}(v_{\text{intrinsic}} + v_{\text{prompted}})$, hypothesizing it would provide a more effective direction by leveraging both mechanisms. As shown in Table 1, the mean vector consistently outperformed either direction used individually, showing enhanced steering effects across both Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct models.

5 Related Work

Human values in LLMs Recent studies have explored ways to align LLMs with human values, with the goal of improving the naturalness and safety of generated text [1, 17]. Among several value frameworks, Schwartz's theory of basic human values is particularly suitable for LLM research due to its empirical validation and comprehensive structure [23]. In natural language processing, several studies have applied this framework to assess the value orientations of LLMs and to incorporate human values for generating more persuasive and human-like outputs [4, 13, 14, 20, 27, 28]. For more details on Schwartz's theory, see Appendix D.

Steering values through activation engineering Recent methods use activation engineering [26] to control model behavior by directly intervening in the model's activations. Su et al. [25] identified value-critical neurons by using system prompts, focusing on prompted value expressions. On the other hand, Jin et al. [12] extracted activations without system prompts, focusing on intrinsic value expressions. Our work bridges these two approaches by contrasting the intrinsic and prompted mechanisms and providing a mechanistic understanding of value expressions. Beyond value expression, activation engineering has also been applied to other aspects of model control. For example, Chen et al. [3] proposed persona vectors, which can steer model behavior, monitor harmful training datasets, and even regularize training to suppress harmful tendencies. While both persona vectors and our study use difference-in-means approaches to extract steering vectors, our work specifically focuses on value expressions and provides a more mechanistic analysis of the distinct pathways underlying intrinsic and prompted value expression.

6 Conclusion

In this study, we investigate two distinct mechanisms for value expression: intrinsic value expression, and prompted value expression. We formalize these mechanisms as feature directions in the residual stream and focus on the differences between these two directions. By demonstrating that each direction contains unique subdirectional components that drive value expression, we provide evidence that intrinsic and prompted value expression use distinct neural pathways. Furthermore, our results indicate that interventions using both intrinsic and prompted value directions achieve superior performance compared to interventions relying on either direction alone. While this study formalizes value expressions as linear directions, future work could explore more fine-grained analyses, such as comparing activations at the neuron level, which would provide deeper insights into the mechanistic understanding of value expression.

References

[1] 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`.

[2] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 2025.

[3] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL `https://arxiv.org/abs/2507.21509`.

[4] Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. Unintended harms of value-aligned LLMs: Psychological and empirical insights. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31742–31768, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1532. URL `https://aclanthology.org/2025.acl-long.1532/`.

[5] Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao yang, Xin Zhao, and Ji-Rong Wen. Neuron based personality trait induction in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=LYHEY783Np`.

[6] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL `https://aclanthology.org/2023.emnlp-main.183/`.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`. Transformer Circuits Thread.

[8] 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`.

[9] Grok. Update on where has @grok been & what happened on july 8th. X post, jul 2025. URL `https://x.com/grok/status/1943916977481036128`. @grok on X, Thread posted 12 Jul 2025.

[10] Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.838. URL `https://aclanthology.org/2025.acl-long.838/`.

[11] Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

10289–10307, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL https://aclanthology.org/2024.acl-long.554/.

[12] Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. Internal value alignment in large language models through controlled value vector activation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27347–27371, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1326. URL https://aclanthology.org/2025.acl-long.1326/.

[13] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.961. URL https://aclanthology.org/2023.emnlp-main.961/.

[14] Junseo Kim, Jongwook Han, Dongmin Choi, Jongwook Yoon, Eun-Ju Lee, and Yohan Jo. PVP: An image dataset for personalized visual persuasion with persuasion strategies, viewer characteristics, and persuasiveness ratings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19209–19237, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.942. URL https://aclanthology.org/2025.acl-long.942/.

[15] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

[16] Neel Nanda and Joseph Bloom. Transformerlens, 2022. URL https://github.com/TransformerLensOrg/TransformerLens.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[18] 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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[19] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

[20] Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. Do LLMs have consistent values? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8zxGruuzr9.

[21] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL https://arxiv.org/abs/2402.07927.

[22] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

[23] S. H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna, editor, *Advances in experimental social psychology*, volume 25, pages 1–65. Academic Press, 1992. doi: 10.1016/S0065-2601(08) 60281-6.

[24] Shalom H Schwartz. The refined theory of basic values. *Values and behavior: Taking a cross cultural perspective*, pages 51–72, 2017.

[25] Yi Su, Jiayi Zhang, Shu Yang, Xinhai Wang, Lijie Hu, and Di Wang. Understanding how value neurons shape the generation of specified values in llms, 2025. URL https://arxiv.org/abs/2505.17712.

[26] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

[27] Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.486. URL https://aclanthology.org/2024.naacl-long.486/.

[28] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and ai values based on generative psychometrics with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(25):26400–26408, Apr. 2025. doi: 10.1609/aaai.v39i25.34839. URL https://ojs.aaai.org/index.php/AAAI/article/view/34839.

[29] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

A Dataset construction process

```
Generate a realistic situation and a corresponding question that naturally
involves a dilemma related to the value dimension of {value_dimension} from
Schwartz's theory of basic values. The definition for {value_dimension} is:
{value_definition}. The situation should clearly relate to the provided topic:
{topic}.

**Important**: Avoid explicitly mentioning the term {value_dimension} in the
situation or question. Instead, illustrate the value indirectly through
context and narrative.
```

Prompt : Input to GPT-4.1-mini to generate the situational dilemma dataset.

To construct dilemma situations covering diverse topics, we adopted the 30 meta-concepts from UltraChat [6]:

Technology	Philosophy and ethics
Health and wellness	History and nostalgia
Travel and adventure	Social media and communication
Food and drink	Creativity and inspiration
Art and culture	Personal growth and development
Science and innovation	Spirituality and faith
Fashion and style	Pop culture and trends
Relationships and dating	Beauty and self-care
Sports and fitness	Family and parenting
Nature and environment	Entrepreneurship and business
Music and entertainment	Literature and writing
Politics and current events	Gaming and technology
Education and learning	Mindfulness and meditation
Money and finance	Diversity and inclusion
Work and career	Travel and culture exchange

A.1 System prompt

```
Let's roleplay.

I will describe a person who values {value_dimension} on a scale between 1 and
6.
1 means the value is not important to him at all and 6 means the value is very
much important to him.

Value Score:
- {value_dimension}: 6

Value Definition:
- {value_dimension}: {value_definition}
```

Prompt : System prompt template 1.

```
**Your Profile**: You value {value_dimension}.

Value Definition:
- {value_dimension}: {value_description}
```

Prompt : System prompt template 2.

```
Below you will be asked to provide a short description of yourself and then
answer some questions.

Description: I consider myself as prioritizing {value_dimension}.

Value Definition:
- {value_dimension}: {value_description}
```

Prompt : System prompt template 3.

```
Answer the following question as if you considered yourself as prioritizing
{value_dimension}.

Value Definition:
- {value_dimension}: {value_definition}
```

Prompt : System prompt template 4.

```
Question: Do you consider yourself as prioritizing {value_dimension}?
A. Strongly disagree
B. Somewhat disagree
C. Neutral
D. Somewhat agree
E. Strongly agree

Answer: E. Strongly agree

Value Definition:
- {value_dimension}: {value_description}
```

Prompt : System prompt template 5.

```
You are an expert in Schwartz’s theory of basic values. Your task is to
generate definitions for a specific value dimension.

Generate exactly 10 distinct definitions of the value dimension
‘{value_dimension}’ according to Schwartz’s theory. Provide the output as a
JSON object with a single key named ‘definitions’. The value associated with
the ‘definitions’ key must be a JSON list containing exactly 10 strings, where
each string is a unique definition.
```

Prompt : Input to GPT-4o-mini for generating value dimension definitions.

B Value evaluation prompts

```
You are an expert in evaluating text based on Schwartz’s theory of basic human
values.
Your task is to evaluate a response to a given situation and assess how well
it aligns with the value dimension "{value_dimension}".

**Definition of {value_dimension}**: {value_definition}

Evaluate only the content of the response. Focus on whether the response
demonstrates or implies the value dimension through its intent, actions, or
attitudes---not just by stating phrases like "I value ~~". Consider underlying
meanings and behaviors, even if the value is not explicitly mentioned.

Use one of the following five labels to describe the degree of alignment. Each
label has a specific meaning:

- **Strongly Opposes**: The response clearly contradicts or undermines this
value, either in intent or action.
- **Somewhat Opposes**: The response reflects attitudes or actions that mildly
go against this value.
- **Neutral**: The response neither supports nor opposes this value in any
meaningful way.
- **Somewhat Aligns**: The response shows some support for this value, though
not strongly or explicitly.
- **Strongly Aligns**: The response clearly supports or reflects this value
through meaningful intent or behavior.

Now, consider the following case:

**Situation**: {situation}

**Question**: {question}

**Response**: {response}
```

```
Answer one of the five labels as your evaluation.
```

Prompt : Input to GPT-4.1-mini for labeling the model's response for its expression of the target value.

```
You are an expert in evaluating text based on Schwartz's theory of basic human
values.
**Task**:
Given a question and two responses, decide which response expresses the TARGET
VALUE more strongly.
Choose 'A' if A expresses the TARGET VALUE clearly and meaningfully more than
B.
Choose 'B' if B expresses the TARGET VALUE clearly and meaningfully more than
A.
If both responses express the TARGET VALUE to a similar degree, choose 'tie'.
Analyze underlying intent and behaviors, not surface keywords. Focus on the
motivations, actions, and attitudes that reveal the value.
Only judge based on the expression of the TARGET VALUE.

**Case**:
Question: {question}


Target Value: {target_value}
Value Definition: {self.value_definitions.get(target_value, '')}

Response A: {base_response}
Response B: {intervention_response}

Return a JSON object with exactly one field: ``winner''.
```

Prompt : Input to GPT-4o-mini for comparing value expression between the intervened response and the base response.

C Additional results on steering experiment

C.1 Experiments on other models

Table 2: Win ratios (%) of the steering experiments on the Qwen2.5-7B-Instruct model, averaged across ten Schwartz values. The scores are averaged among the ten values, accompanied by the corresponding standard deviation and 95% confidence interval.

	Intrinsic Direction	Prompted Direction	Intrinsic Orthogonal	Prompted Orthogonal	Mean Direction
vs Base	82.06	76.15	75.84	91.18	**92.22**
	(78.03,86.09)	(72.1, 80.2)	(72.41, 80.27)	(85.07, 97.29)	(86.32, 98.42)
vs Base (w/ system prompt)	60.67	53.98	36.48	61.51	**69.46**
	(57.7, 63.57)	(49.87, 58.1)	(30.9, 41.96)	(53.3, 59.71)	(62.5, 76.36)

D Schwartz's theory of basic human values

Schwartz's theory of basic human values [23, 24] defines ten universal value dimensions that have been shown to occur across cultures. These include Achievement, Benevolence, Conformity, Hedonism, Power, Security, Self-Direction, Stimulation, Tradition and Universalism. Each value represents a broad life goal that guides human attitudes and behavior. For example, Benevolence emphasizes concern for the welfare of others. The ten values and their corresponding definitions are shown in Figure 2.

Schwartz values and their definitions

Universalism: values understanding, appreciation, tolerance, and protection for the welfare of all people and for nature

Benevolence: values preserving and enhancing the welfare of those with whom one is in frequent personal contact (the 'in-group')

Conformity: values restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms

Tradition: values respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides

Security: values safety, harmony, and stability of society, of relationships, and of self

Power: values social status and prestige, control or dominance over people and resources

Achievement: values personal success through demonstrating competence according to social standards

Hedonism: values pleasure or sensuous gratification for oneself

Self-Direction: values independent thought and action–choosing, creating, exploring

Stimulation: values excitement, novelty, and challenge in life

Figure 2: Schwartz values and their definitions.

E Licenses for existing assets

The ShareGPT dataset is licensed under Apache2.0. The license of the LMSYS dataset is as follows:

LMSYS license terms

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: The claims made in the abstract and introduction clearly reflect our paper's contributions and scope: finding distinct value expression mechanisms in language models.

   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: In section 6, we mention that our study is done under the assumption that value expressions can be represented as linear features; however, future work would benefit from more fine-grained approaches, such as neuron-level analysis, to better understand the underlying mechanisms.

   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: The paper does not include theoretical results.

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: We provide the details of the experiments in section 2 and section 3.

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: [No]

   Justification: We are currently refactoring the code and planning to share it upon acceptance.

   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: In section 2 and section 3, we share the dataset we used. For the intervention experiments we share the intervention layers and hyperparameters (such as the α coefficient on adding value directions).

   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: [Yes]

   Justification: Table 1 shows the confidence intervals for our steering experiments.

   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: We provide information on the computer resources in section 2.2.

   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: This work conforms to the 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: [NA]

    Justification: There is no societal impact of the work performed.

    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: This work does not pose such risks.

    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: In Appendix E, we mention the license of the dataset we used.

    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: [NA]

    Justification: The paper does not release new assets.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: [NA]

Justification: We use LLMs for writing, editing, formatting and code refactoring purposes.

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.