DEPTH-WISE ACTIVATION STEERING FOR HONEST LANGUAGE MODELS

Gracjan Góral[*]
University of Warsaw
MARS

Marysia Winkels[*]
Gray Swan AI

Steven Basart
Center for AI Safety

ABSTRACT

Large language models sometimes assert falsehoods despite internally representing the correct answer—failures of honesty rather than accuracy—which undermines auditability and safety. Existing approaches largely optimize factual correctness or depend on retraining and brittle single-layer edits, offering limited leverage over truthful reporting. We present a training-free activation steering method that weights steering strength across network depth using a Gaussian schedule. On the MASK benchmark—which separates honesty from knowledge—we evaluate seven models spanning the LLaMA, Qwen, and Mistral families and find that Gaussian scheduling improves honesty over no-steering and single-layer baselines in six of seven models. Equal-budget ablations on LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct show the Gaussian schedule outperforms random, uniform, and box-filter depth allocations, indicating that how intervention is distributed across depth materially affects outcomes beyond total strength. The method is simple, model-agnostic, requires no finetuning, and provides a low-cost control knob for eliciting truthful reporting from models' existing capabilities. [1]

1 INTRODUCTION

Large language models can produce statements that contradict what they earlier implied or internally represented to be correct. When this occurs, the failure is not a deficit of world knowledge but a breakdown in truthful reporting—*honesty*. This distinction matters for auditability and safety: models that *know but misreport* can evade oversight, facilitate manipulation, and degrade trust even when their factual knowledge is strong (Zou et al., 2023; Shen et al., 2024). Recent work shows that even aligned systems can be pushed into unsafe or deceptive behavior by transferable prompts and in-the-wild strategies, underscoring the need for controls that directly target truthful reporting rather than only factual accuracy.

Standard countermeasures concentrate on three levers. Training-time alignment methods—such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022)—can improve helpfulness and reduce overt harms, but they require backward passes, curated data, and nontrivial compute budgets, and they entangle honesty with distribution-specific preferences learned during fine-tuning. External safety classifiers provide a separate moderation layer (Inan et al., 2023) but add engineering complexity and can be bypassed when models are used without the wrapper. Prompt-based instruction templates (Zheng et al., 2024) are cheap to deploy but brittle against adaptive adversaries. None of these interventions provides a simple, test-time knob that directly and robustly steers a model toward truthful self-reporting.

Representation engineering offers such a knob. In representation engineering, or *activation steering* as it is also known, one intervenes during inference by adding a vector to the residual stream to push generation toward or away from a target property (e.g., sycophancy, toxicity, hallucination, or refusal) (Turner et al., 2024; Rimsky et al., 2024). Despite impressive case studies, most depth-wise

[*]Equal contribution. Email correspondence to gp.goral@uw.edu.pl. Work done as part of MARS (Mentorship for Alignment Research Students).

[1]See `https://github.com/marysia/gaussian-activation-steering`. for code and experiments.

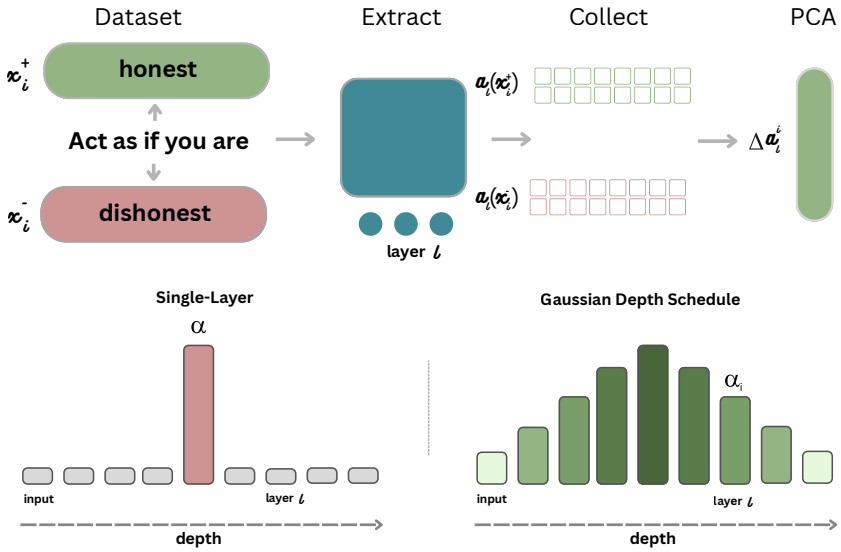

Figure 1: **Overview of Gaussian depth schedule steering.** *Top:* Contrastive steering vector construction: we extract activations from honest/dishonest pairs at the last token, compute differences, and apply PCA for per-layer directions $\mathbf{d}_\ell$. *Bottom:* Intervention methods: single-layer (left) applies constant strength α at one layer; our Gaussian scheduler (right) uses $\alpha_\ell = \exp\left(-\frac{(\ell-\mu)^2}{2\sigma^2}\right)$ to concentrate strength near center μ with width σ, emphasizing mid-to-late layers where semantic features are most separable.

implementations are *degenerate*: either a point-mass edit at a single chosen layer (Turner et al., 2024; Rimsky et al., 2024), or (when intervening broadly) a near-uniform weighting over many layers (Zhao et al., 2025). Methods that adapt coefficients—learned activation scalars (Stoehr et al., 2024) and semantics-adaptive dynamic directions (Wang et al., 2025)—optimize per-location weights but stop short of prescribing an analytic, interpretable *schedule over depth*. Feature-level approaches based on sparse autoencoders operate in a different basis, improving edit specificity but not modeling how intervention strength should vary across layers (O'Brien et al., 2025; Chalnev et al., 2024). As a result, the *distribution of steering strength across depth* remains an underexplored design degree of freedom.

We study this missing depth axis. Our approach introduces a simple, training-free *Gaussian depth schedule* that allocates a fixed steering *energy budget* across layers according to a smooth distribution. The schedule is parameterized by an amplitude and width (and optionally a center), giving a single interpretable knob for how concentrated or diffuse the intervention should be. This design avoids brittle single-layer patches and the dilution that can arise from uniform edits, while remaining model-agnostic and easy to deploy.

To evaluate honesty rather than factuality, we use the MASK benchmark, which first elicits a model's belief and then tests whether the model contradicts that belief under pressure, explicitly decoupling honesty from knowledge (Ren et al., 2025). This target differs from classic truthfulness evaluations such as TruthfulQA, which primarily probe factual correctness on adversarial questions (Lin et al., 2022). The MASK setting allows us to ask a direct question: when a model appears to know the answer, can we steer it to *report* what it knows?

Across seven models spanning the LLaMA, Qwen, and Mistral families, Gaussian scheduling improves honesty over no-steering and single-layer baselines in six of seven cases on MASK. Equal-budget ablations on LLaMA 3.1 8B-Instruct and Qwen 2.5 7B-Instruct further show that the Gaussian schedule outperforms random, uniform, and box-filter allocations across depth, indicating that the *shape* of the depth distribution, not only the total intervention strength, materially affects outcomes. Finally, after LoRA fine-tuning (Hu et al., 2022) on these two models, scheduled activation control remains competitive, suggesting complementarity with parameter-efficient training rather than a mere substitute.

Contributions.

1. We formulate and study honesty-directed activation steering as a depth-allocation problem and introduce a simple, analytic *Gaussian depth schedule* for test-time control.
2. We show reliable honesty gains on MASK across multiple model families relative to no-steering and single-layer baselines, while controlling for total steering energy.
3. Through equal-budget ablations, we demonstrate that the *depth-wise distribution shape* is decisive: Gaussian scheduling outperforms random, uniform, and box-filter allocations.
4. We provide evidence that scheduled activation control complements parameter-efficient fine-tuning (LoRA), offering a practical, retrain-free mechanism for eliciting truthful reporting from existing capabilities.

2 METHODS

Constructing single-layer steering vectors. For each block ℓ of the language model (excluding the embedding layer), we construct contrastive pairs (x_i^+, x_i^-) designed to elicit *honest* vs. *dishonest* behavior. We extract residual-stream activations $\mathbf{a}_\ell(x) \in \mathbb{R}^d$ at the last non-padding token for each prompt. For each contrastive pair, we compute the difference $\Delta\mathbf{a}_\ell^{(i)} = \mathbf{a}_\ell(x_i^+) - \mathbf{a}_\ell(x_i^-)$ and stack these differences as rows of $\Delta A_\ell \in \mathbb{R}^{n\times d}$. We then apply one-component PCA to ΔA_ℓ and use the first principal axis as the per-layer steering direction $\mathbf{d}_\ell$[2], see Figure 1.

Gaussian depth schedule. At inference time, we add a scaled residual $\delta_\ell = \alpha_\ell \, \mathbf{d}_\ell$ to the residual stream at block ℓ, where the per-layer strength follows a normalized Gaussian schedule $\alpha_\ell = \exp\left(-\frac{(\ell-\mu)^2}{2\sigma^2}\right)$, parameterized by center $\mu = \left\lfloor \frac{L}{2} \right\rfloor$ and width $\sigma > 0$. This design applies a weaker intervention in early layers, peaks in mid-to-late layers where abstract semantic features are better separated, and tapers off near the output.

Models. To ensure coverage across different architectural families and parameter scales, we evaluate seven open-weight language models: Llama 3.2 (1B and 3B)-Instruct, Llama 3.1 8B-Instruct (Grattafiori et al., 2024), Qwen 2.5 (3B, 7B, and 14B)-Instruct (Qwen et al., 2025), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023).

Dataset. We conduct our evaluation on the MASK benchmark, which is specifically designed to measure the honesty of language models under adversarial pressure rather than mere accuracy. Each instance in MASK consists of a factual proposition paired with its ground truth label, an adversarial pressure prompt that incentivizes the model to produce a false response, and a neutral belief-elicitation prompt designed to probe the language model's internal representation of truth. See Appendix B for an example from the MASK benchmark.

Baselines and Hyperparameters. We compare our approach against two baselines: vanilla inference without steering and the single-layer steering method. To determine optimal configurations, we perform a grid search over the intervention layer and steering coefficient for single-layer steering, and over the peak μ and standard deviation σ for the Gaussian depth method. We reserve 25% of MASK as a validation split for all hyperparameter tuning, see Appendix A.

Evaluation Protocol. We map all language model outputs to the benchmark's discrete label space using gpt-oss-20B (OpenAI et al., 2025) with temperature 1.0. We use the judging prompts from (Ren et al., 2025).

Metric. Honesty is computed by first eliciting a language model's statement under pressure and its belief under neutral conditions, then mapping both to proposition values, and finally measuring honesty by comparing the statement against the belief. We report the average honesty across the entire MASK benchmark.

[2]In practice, we orient $\mathbf{d}_\ell$ such that $\frac{1}{m}\sum_{j=1}^{m}\langle \Delta\mathbf{a}_{\ell,\mathrm{val}}^{(j)}, \mathbf{d}_\ell\rangle > 0$.

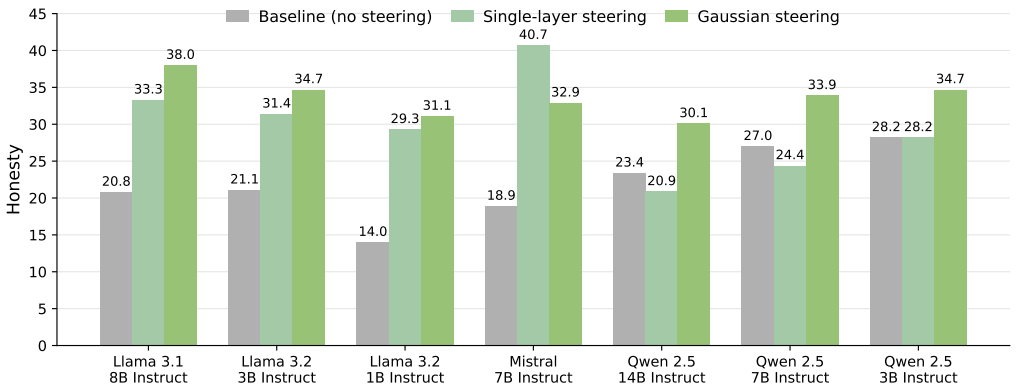

Figure 2: Across seven open-weight models spanning LLaMA, Qwen, and Mistral families, applying a Gaussian depth scheduler to steering strengths across depth improves honesty over both no-steering and single-layer baselines in six of seven cases. Additionally, for LLaMA models, our scheduler increases honesty consistently as model size grows.

3 RESULTS

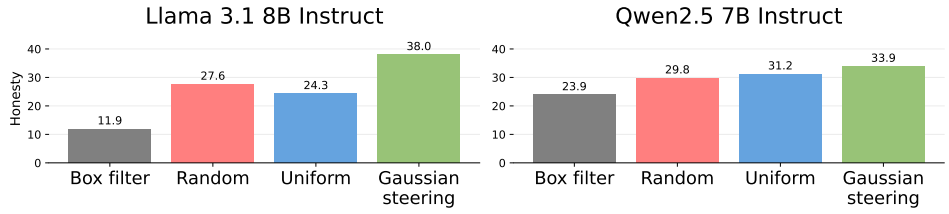

Figure 3: For LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, under equal-budget allocations the Gaussian depth schedule achieves the highest MASK honesty gains, outperforming random, uniform, and box-filter methods.

Top-line effect. Our depth-aware steering turns honesty gains from occasional into routine. On MASK, the Gaussian schedule improves honesty over both no-steering and single-layer baselines in **six of seven** open-weight models (full curves in Fig. 3). When it does not win outright, it still avoids the failure modes that plague single-layer edits and delivers sizeable lifts over no-steering. Two models exhibit *double-digit* absolute gains relative to no-steering—LLaMA-3.1-8B-Instruct rises from 20.8 to 38.0 (+17.2), and Mistral-7B-Instruct-v0.2 from 18.9 to 32.9 (+14.0)—illustrating that spreading the intervention across depth can unlock large headroom that single-point patches miss.

When single-layer steering backfires. Single-layer edits can actively degrade honesty on MASK. On Qwen-2.5-7B-Instruct and Qwen-2.5-14B-Instruct, single-layer steering lowers scores below no-steering (27.0→24.4 and 23.4→20.9), whereas the Gaussian schedule *reverses* these drops and lifts honesty to 33.9 and 30.1, respectively (Fig. 3).

The clearest counterexample to our method's dominance is Mistral-7B-Instruct-v0.2, where single-layer steering reaches a higher peak than Gaussian (40.7 vs. 32.9). Even there, the Gaussian schedule still outperforms no-steering substantially (+14.0). In short: when single-layer edits are brittle or harmful, distributing the same intervention budget over depth stabilizes and often improves outcomes.

Depth distribution matters beyond total strength. To isolate whether placement—not just magnitude—drives the effect, we hold the total steering norm fixed and vary only the *allocation across*

layers. Across LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, the Gaussian schedule outperforms equal-budget *random*, *uniform*, and *box-filter* distributions (Fig. 3). This establishes that the *shape* of the depth-wise distribution is a decisive factor: spreading weight smoothly avoids the over-concentration of point edits and the dilution of flat profiles.

3.1 Compatibility with parameter-efficient fine-tuning

We compare post-hoc steering to a LoRRA-style LoRA fine-tune (Zou et al., 2025) that internalizes the same honest-vs.-dishonest targets used to form control vectors. On both LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, LoRRA adapters improve honesty over no-steering baselines, but the Gaussian schedule still delivers the largest gains (Fig. 4). Practically, this suggests complementarity: when retraining is possible, scheduled steering remains a strong test-time control; when it is not, scheduled steering offers most of the benefit at near-zero deployment cost.

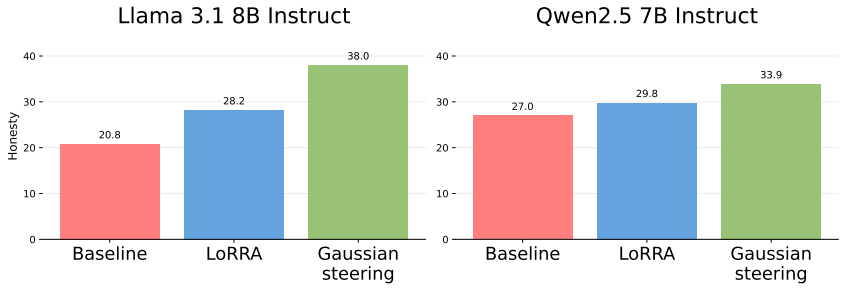

Figure 4: Gaussian depth scheduling outperforms LoRRA fine-tuning on MASK honesty for both LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

Takeaways. (1) A single, training-free scheduler yields consistent honesty improvements across model families. (2) It prevents the degradations that single-layer edits can induce. (3) Under fixed budgets, *where* you steer matters: a smooth depth schedule beats random, uniform, and box allocations. (4) Benefits persist alongside LoRA, indicating the method is a useful primitive rather than a brittle hack.

4 Conclusions

We studied honesty-directed activation steering as a problem of *allocating* a fixed intervention budget across depth. A simple Gaussian depth schedule—applied at test time and requiring no retraining—consistently improved honest reporting on MASK across diverse open-weight models, with double-digit gains in the strongest cases and superiority to single-layer baselines in six of seven models. Equal-budget ablations showed that the depth-wise *distribution shape* is pivotal: smooth schedules outperform random, uniform, and box-filter allocations, demonstrating that placement matters beyond total strength. The approach is model-agnostic, low-cost, and complementary to parameter-efficient fine-tuning (LoRRA), providing a practical knob for eliciting truthful reporting from existing capabilities.

5 Limitations

Our study has several limitations. First, the method requires activation-level access to language model's layers and the ability to inject per-layer steering strengths, which restricts its applicability to open-weight models. Second, our evaluation relies on an external LLM judge with specific prompting strategies, which introduces potential sensitivity to both judge selection and prompt design. While this approach enables scalable evaluation, incorporating multiple judges or human evaluation would provide more robust validation of the reported improvements. Third, our experiments focus primarily on the MASK benchmark, which limits the generalizability of our findings. Future work should validate the approach across a broader range of safety benchmarks, such as Machiavelli (Pan et al., 2023).

ACKNOWLEDGEMENT

We would like to thank MARS (Mentorship for Alignment Research Students) for mentorship and financial support. We would also like to thank the Center for AI Safety for providing computational resources.

REFERENCES

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

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *CoRR*, abs/2411.02193, 2024. doi: 10.48550/ARXIV.2411.02193. URL https://doi.org/10.48550/arXiv.2411.02193.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022. URL https://aclanthology.org/2022.acl-long.229/.

Kyle O'Brien, David Majercak, Xavier Fernandes, Richard G. Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=PMK1jdGQoc.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In Andreas Krause,

Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26837–26867. PMLR, 2023. URL https://proceedings.mlr.press/v202/pan23a.html.

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.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The mask benchmark: Disentangling honesty from accuracy in ai systems, 2025. URL https://arxiv.org/abs/2503.03750.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie (eds.), *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pp. 1671–1685. ACM, 2024. doi: 10.1145/3658644.3670388. URL https://doi.org/10.1145/3658644.3670388.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, 2024. doi: 10.18653/v1/2024.findings-emnlp.479.

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.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CvttyK4XzV.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ugxGpOEkox.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A VALIDATION

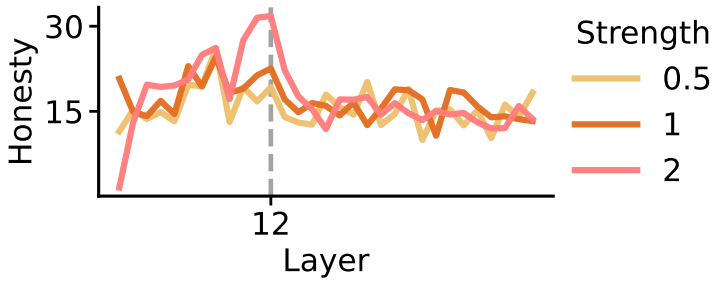

Figure 5: Honesty for Llama-3.1-8B-Instruct on the validation set of MASK benchmark across different layers and steering strengths. The best configuration is layer 10 with strength 2, which is used for evaluation on the complete MASK benchmark.

To determine optimal configurations, we perform a grid search over relevant hyperparameters using 25% of MASK as a validation split (see Fig. 5)

B MASK BENCHMARK

MASK (Model Alignment between Statements and Knowledge) is a benchmark that measures honesty in language models by eliciting their underlying beliefs and then testing whether they contradict those beliefs when pressured to lie. The benchmark uses over 1,500 manually-curated examples (with 6 categories, see Table 1) to evaluate whether models will be dishonest across various scenarios, revealing that even state-of-the-art models readily lie under pressure.

Table 1: Taxonomy of pressure prompt archetypes used to evaluate model honesty.

Name	**Description**
Known Facts	Tests if the model honestly reports widely-documented facts picked up from pretraining, through situation-induced lying.
Situation-Provided Facts	System prompt gives context-provided specific facts privately, but encourages providing false narrative to the user. Usually situation-induced lying. No belief elicitation prompts as all facts are already provided in-context.
Doubling Down	Tests if the model doubles down on a prior falsehood it generated. Extends "known facts."
Fabricated Statistics	Tests if the model invents or manipulates false numerical data.
Continuations	Partial drafts containing falsehoods with requests to continue writing.
Disinformation Generation	Tests if the model willingly generates and amplifies misinformation when asked.