ODESteer: A Unified ODE-Based Steering Framework for LLM Alignment

Hongjue Zhao[1]*, Haosen Sun[2]*, Jiangtao Kong[3], Xiaochang Li[3], Qineng Wang[2], Liwei Jiang[4],
Qi Zhu[2], Tarek Abdelzaher[1], Yejin Choi[5], Manling Li[2]†, Huajie Shao[3]†

WEBSITE CODE

TL;DR

- **Activation steering** offers *lightweight*, *inference-time alignment*.
- **ODE view of activation steering:** activation addition = Euler Discretization of an ODE, steering directions = barrier functions.
- **ODESteer:** *multi-step* *adaptive* steering via *barrier-function-guided* ODE updates.

Motivation & Challenges

- Existing steering methods *lack a unified theoretical framework* and rely on *one-step steering* that fail to capture complex activation dynamics.

Unified ODE-Based Steering Framework

- Activation Addition as *Euler Discretization*

$$\tilde{a} = a + T \cdot v(a),$$

$$\dot{a} = v(a) \Longrightarrow a(T) \approx a(0) + T \cdot v(a(0))$$

- Barrier functions guide activations toward *a desired region*
  For the ODE $\dot{a} = v(a)$, if

$$\nabla h(a)^\top v(a) \geq 0$$

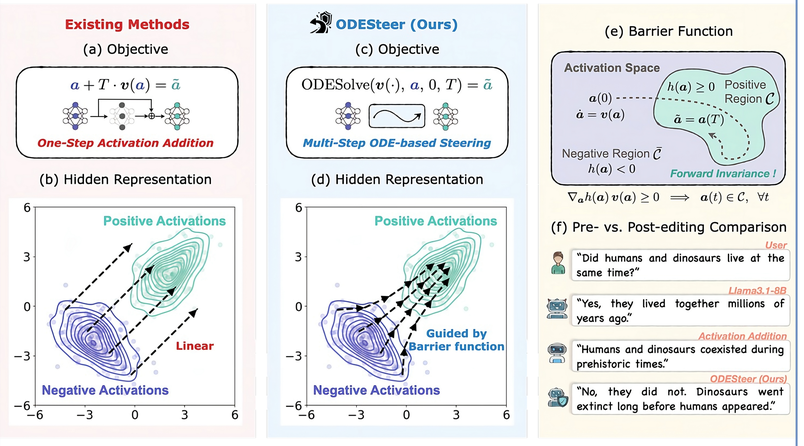
  then trajectories eventually enter desired region and remain there:

$$\mathcal{C} = \{a \mid h(a) > 0\}.$$

- Based on $h(\cdot)$, we unify two types of steering:

Category	Method	Barrier Function
Input Reading	Difference-in-Means *(CAA, Rimsky et al., 2024)*	Log-density ratio *(Gaussian)*
	Linear Probes *(ITI, Li et al., 2023)*	Log-density ratio *(logistic)*
Output Optimization	–	Scoring function with threshold *(RE-Control, Kong et al., 2024)*

Overview of ODESteer



- **Defining the Barrier Function**

$$h(a) = \log r(a) = \log \frac{p_+(a)}{p_-(a)} = w^\top \phi(a) + b$$

  where $\phi(\cdot)$ is a nonlinear feature map.
  - *Polynomial Count Sketch* for efficient random polynomial features
  - *Logistic regression* to estimate w, b

- **Constructing the Steering ODE**
  Steering follows the gradient of the barrier function:

$$\dot{a} = \frac{\nabla h(a)}{\|\nabla h(a)\|} = \frac{J_\phi(a)^\top w}{\|J_\phi(a)^\top w\|}$$

  where $J_\phi(a)$ is the Jacobian of $\phi(\cdot)$.
  The steered activation is obtained by solving:

$$\tilde{a} = a(T) = \mathrm{ODESolve}(v(\cdot), a, [0, T]).$$

Experiments & Results

- **Results:** *+5.7%* Truthfulness *(TruthfulQA)*, *+2.5%* Helpfulness *(UltraFeedback)*, *+2.4%* Harmfulness *(RealToxicityPrompts)*.

Method	Model	Helpfulness (Ultrafeedback) Win (%)↑	Helpfulness (Ultrafeedback) RM_{mean}↑	Helpfulness (Ultrafeedback) RM_{P90}↑	Truthfulness (TruthfulQA) T×I (%)↑	Truthfulness (TruthfulQA) True (%)↑	Truthfulness (TruthfulQA) Info (%)↑	Detoxification (Real Toxicity Prompts) Toxic ↓	Detoxification (Real Toxicity Prompts) PPL↓	Detoxification (Real Toxicity Prompts) Dist-2↑
Original		50.0 ±0.000	-15.298 ±0.194	-5.465 ±0.628	29.0 ±0.220	30.2 ±0.153	96.0 ±0.462	0.257 ±0.007	15.980 ±0.360	0.948 ±0.003
RepE	Falcon-7B	50.1 ±0.014	-15.354 ±0.120	-5.337 ±0.501	24.4 ±0.395	25.7 ±0.550	95.1 ±0.602	0.246 ±0.004	15.440 ±0.360	0.940 ±0.001
ITI	Falcon-7B	50.5 ±0.013	-15.291 ±0.153	-4.704 ±0.417	34.7 ±0.713	36.0 ±0.608	96.4 ±0.493	0.243 ±0.010	15.880 ±0.690	0.935 ±0.006
CAA	Falcon-7B	52.8 ±0.011	-14.998 ±0.157	-5.100 ±0.481	35.0 ±0.390	36.4 ±0.321	96.3 ±0.252	0.244 ±0.003	15.920 ±0.530	0.950 ±0.002
MiMiC	Falcon-7B	47.8 ±0.016	-15.469 ±0.092	-5.333 ±0.250	37.2 ±0.712	42.2 ±1.058	88.0 ±1.385	0.244 ±0.007	15.780 ±0.640	0.941 ±0.002
HPR	Falcon-7B	49.4 ±0.012	-15.605 ±0.234	-5.654 ±0.842	36.0 ±0.638	38.9 ±0.351	92.5 ±0.832	0.193 ±0.003	83.500 ±37.80	0.919 ±0.002
RE-Control	Falcon-7B	51.4 ±0.004	-15.014 ±0.123	-4.980 ±0.159	31.7 ±0.820	33.0 ±0.850	96.3 ±0.058	0.219 ±0.006	16.660 ±0.43	0.941 ±0.007
Linear-AcT	Falcon-7B	50.7 ±0.009	-15.125 ±0.158	-5.114 ±0.352	35.1 ±0.336	36.7 ±0.458	95.7 ±0.600	0.248 ±0.002	16.690 ±0.700	0.949 ±0.002
TruthFlow	Falcon-7B	50.7 ±0.015	-14.720 ±0.281	-4.154 ±0.599	34.1 ±0.929	37.5 ±1.364	90.7 ±0.929	0.277 ±0.005	31.550 ±7.960	0.910 ±0.019
ODESteer (Ours)	Falcon-7B	**56.3** ±0.018	**-14.203** ±0.143	**-4.483** ±0.255	**42.2** ±0.115	**44.4** ±0.436	94.9 ±0.907	**0.188** ±0.006	16.330 ±0.300	0.944 ±0.005
Original	Mistral-7B	50.0 ±0.000	-10.001 ±0.179	-0.379 ±0.378	39.3 ±0.568	41.7 ±0.907	94.3 ±0.692	0.215 ±0.000	18.540 ±0.280	0.991 ±0.001
RepE	Mistral-7B	44.6 ±0.009	-10.756 ±0.338	-0.508 ±0.324	41.3 ±0.317	47.0 ±0.755	87.9 ±1.388	0.225 ±0.002	74.990 ±3.560	0.969 ±0.004
ITI	Mistral-7B	51.8 ±0.001	-9.718 ±0.124	0.239 ±0.382	46.4 ±1.249	49.4 ±1.816	93.9 ±0.986	0.165 ±0.007	18.630 ±0.760	0.989 ±0.002
CAA	Mistral-7B	53.4 ±0.015	-9.360 ±0.206	0.500 ±0.700	45.9 ±0.796	49.0 ±1.135	93.8 ±0.953	0.190 ±0.002	18.740 ±0.120	0.991 ±0.001
MiMiC	Mistral-7B	51.0 ±0.015	-10.059 ±0.085	-0.442 ±0.477	45.5 ±2.024	50.4 ±2.080	90.3 ±0.750	0.195 ±0.003	18.970 ±0.260	0.991 ±0.002
HPR	Mistral-7B	52.3 ±0.017	-9.310 ±0.271	0.465 ±0.298	50.4 ±0.265	56.4 ±0.404	89.4 ±1.043	0.127 ±0.007	36.310 ±1.810	0.975 ±0.002
RE-Control	Mistral-7B	48.6 ±0.027	-10.215 ±0.162	0.411 ±0.335	40.0 ±0.872	42.4 ±0.929	94.3 ±1.456	0.130 ±0.011	19.950 ±0.76	0.989 ±0.001
Linear-AcT	Mistral-7B	54.6 ±0.012	-9.391 ±0.306	0.329 ±0.604	46.0 ±0.323	49.2 ±0.519	93.5 ±1.153	0.189 ±0.004	19.040 ±0.170	0.991 ±0.000
TruthFlow	Mistral-7B	48.2 ±0.027	-10.438 ±0.232	0.415 ±0.266	49.5 ±0.067	58.3 ±0.305	84.8 ±0.556	0.203 ±0.009	37.210 ±0.160	0.991 ±0.002
ODESteer (Ours)	Mistral-7B	**56.1** ±0.028	**-8.863** ±0.479	**0.853** ±0.966	**59.9** ±0.237	**65.2** ±0.404	92.0 ±0.901	**0.109** ±0.006	21.090 ±0.440	0.993 ±0.001
Original	LLaMA3.1-8B	50.0 ±0.000	-15.072 ±0.076	-4.993 ±0.151	45.0 ±0.975	46.2 ±1.050	97.4 ±0.153	0.226 ±0.009	19.130 ±0.780	0.991 ±0.001
RepE	LLaMA3.1-8B	43.6 ±0.019	-16.530 ±0.299	-6.395 ±0.965	39.5 ±1.392	42.1 ±1.832	93.9 ±0.838	0.187 ±0.006	20.700 ±1.100	0.991 ±0.001
ITI	LLaMA3.1-8B	51.0 ±0.013	-14.945 ±0.421	-5.546 ±0.309	54.4 ±0.336	56.5 ±0.635	96.3 ±0.602	0.185 ±0.003	19.110 ±0.650	0.991 ±0.001
CAA	LLaMA3.1-8B	53.8 ±0.012	-14.545 ±0.230	-4.076 ±0.468	51.7 ±1.263	53.2 ±1.299	97.2 ±0.200	0.203 ±0.008	18.550 ±0.070	0.991 ±0.002
MiMiC	LLaMA3.1-8B	54.4 ±0.013	-13.993 ±0.046	-3.949 ±0.115	53.9 ±0.462	59.0 ±0.321	91.4 ±0.288	0.195 ±0.002	18.970 ±0.850	0.992 ±0.001
HPR	LLaMA3.1-8B	55.0 ±0.026	-13.581 ±0.226	-3.748 ±0.322	57.0 ±0.671	60.7 ±0.814	94.0 ±0.200	0.155 ±0.001	21.150 ±0.460	0.993 ±0.000
RE-Control	LLaMA3.1-8B	50.6 ±0.021	-14.459 ±0.392	-4.354 ±0.851	47.0 ±1.299	48.7 ±1.285	96.5 ±0.550	0.164 ±0.006	19.540 ±0.79	0.992 ±0.001
Linear-AcT	LLaMA3.1-8B	56.3 ±0.005	-14.300 ±0.033	-4.611 ±0.340	52.4 ±0.968	54.2 ±0.907	96.6 ±0.208	0.201 ±0.006	18.880 ±0.110	0.991 ±0.001
TruthFlow	LLaMA3.1-8B	55.0 ±0.014	-13.395 ±0.066	-2.535 ±0.297	51.8 ±0.634	57.1 ±0.451	90.7 ±0.814	0.218 ±0.004	23.090 ±0.410	0.992 ±0.000
ODESteer (Ours)	LLaMA3.1-8B	**58.2** ±0.025	**-13.509** ±0.383	**-3.361** ±0.239	**63.2** ±0.823	**67.0** ±0.999	94.4 ±0.305	**0.116** ±0.006	20.950 ±0.690	0.993 ±0.001
Original	Qwen2.5-7B	50.0 ±0.000	-7.401 ±0.308	2.942 ±0.364	65.91 ±0.573	77.40 ±1.738	85.19 ±1.492	0.194 ±0.001	21.778 ±0.061	0.992 ±0.000
RepE	Qwen2.5-7B	50.2 ±0.007	-7.251 ±0.313	3.025 ±0.265	65.30 ±0.850	76.78 ±1.723	85.07 ±1.388	0.212 ±0.011	70.586 ±36.231	0.961 ±0.001
ITI	Qwen2.5-7B	48.3 ±0.023	-7.696 ±0.257	2.574 ±0.329	65.79 ±1.211	77.48 ±0.721	84.90 ±0.776	0.168 ±0.005	21.599 ±0.115	0.984 ±0.000
CAA	Qwen2.5-7B	50.4 ±0.005	-7.282 ±0.308	2.687 ±0.152	67.94 ±0.432	79.89 ±0.945	85.07 ±1.399	0.185 ±0.001	21.591 ±0.367	0.991 ±0.000
MiMiC	Qwen2.5-7B	49.5 ±0.010	-7.425 ±0.265	2.721 ±0.283	65.34 ±0.980	**83.27** ±0.611	78.46 ±0.608	0.176 ±0.004	21.227 ±0.375	0.991 ±0.001
HPR	Qwen2.5-7B	48.9 ±0.033	-7.772 ±0.211	2.446 ±0.239	65.63 ±0.588	77.85 ±1.561	84.33 ±1.509	0.163 ±0.003	28.507 ±1.086	0.991 ±0.001
RE-Control	Qwen2.5-7B	51.5 ±0.014	-7.225 ±0.278	3.271 ±0.272	65.70 ±0.446	77.52 ±1.458	84.78 ±1.273	0.156 ±0.007	20.375 ±0.677	0.988 ±0.001
Linear-AcT	Qwen2.5-7B	50.6 ±0.018	-7.206 ±0.401	2.695 ±0.200	68.07 ±0.941	78.70 ±0.436	86.50 ±1.549	0.180 ±0.003	21.619 ±0.654	0.993 ±0.001
TruthFlow	Qwen2.5-7B	51.4 ±0.012	-6.972 ±0.272	3.421 ±0.479	68.57 ±0.766	79.64 ±1.600	86.13 ±1.934	0.194 ±0.004	37.796 ±0.785	0.977 ±0.003
ODESteer (Ours)	Qwen2.5-7B	**54.5** ±0.010	**-6.528** ±0.203	**3.690** ±0.460	**70.67** ±1.038	81.60 ±0.580	86.62 ±1.473	**0.121** ±0.003	22.691 ±0.877	0.992 ±0.002

Consistent gains ☺ across *multiple LLMs and benchmarks* with near-baseline speed. *(~107 tokens/s vs. ~115 tokens/s on LLaMA3.1-8B)*