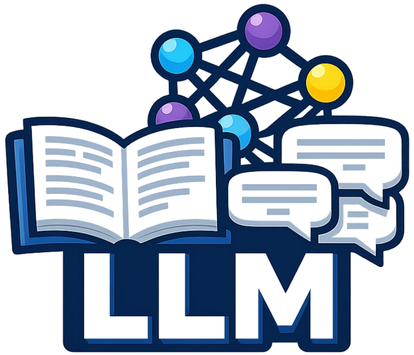
Dual Mechanisms of Value Expression:

Intrinsic vs. Prompted Values in Large Language Models

Jongwook Han*, Jongwon Lim*, Injin Kong, Yohan Jo

Seoul National University

Value expression in LLMs



Preference learning

Intrinsic values

Requires specific target values and is not straightforward to apply to diverse values

System prompts

Prompted values

Why study value mechanisms?

Prompted values

Explicit instructions can produce responses that are less natural and exaggerate

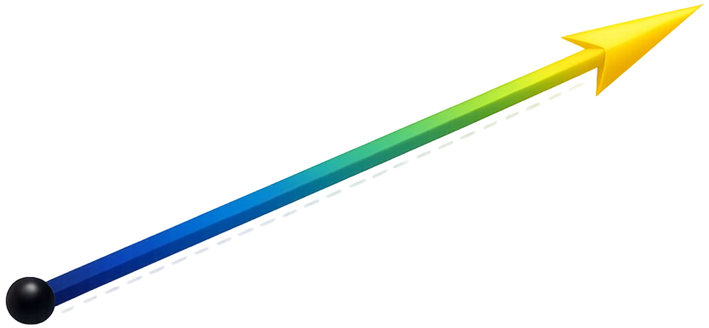
Prompting could cause unexpected outcomes

Grok praised Adolf Hitler, and referred to itself as "MechaHitler"

Grok's system prompts directed the AI to be "anti-woke" and to "not shy away from claims which are politically incorrect, as long as they are well substantiated"

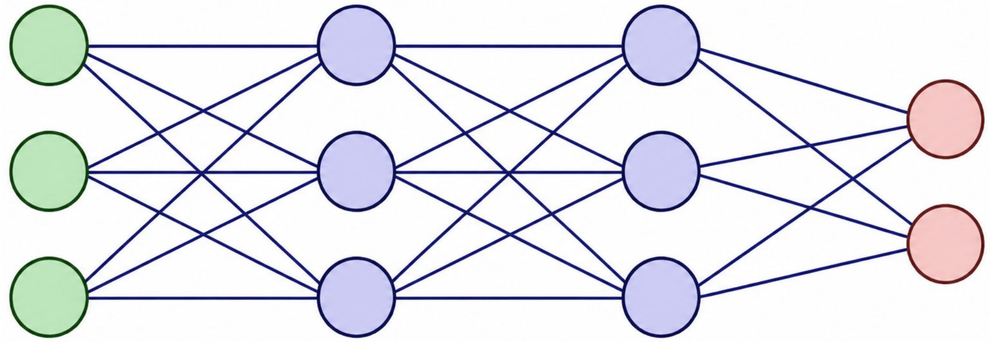
Analysis

We analyze **10 human values** from Schwartz's theory of basic human values



Value vectors

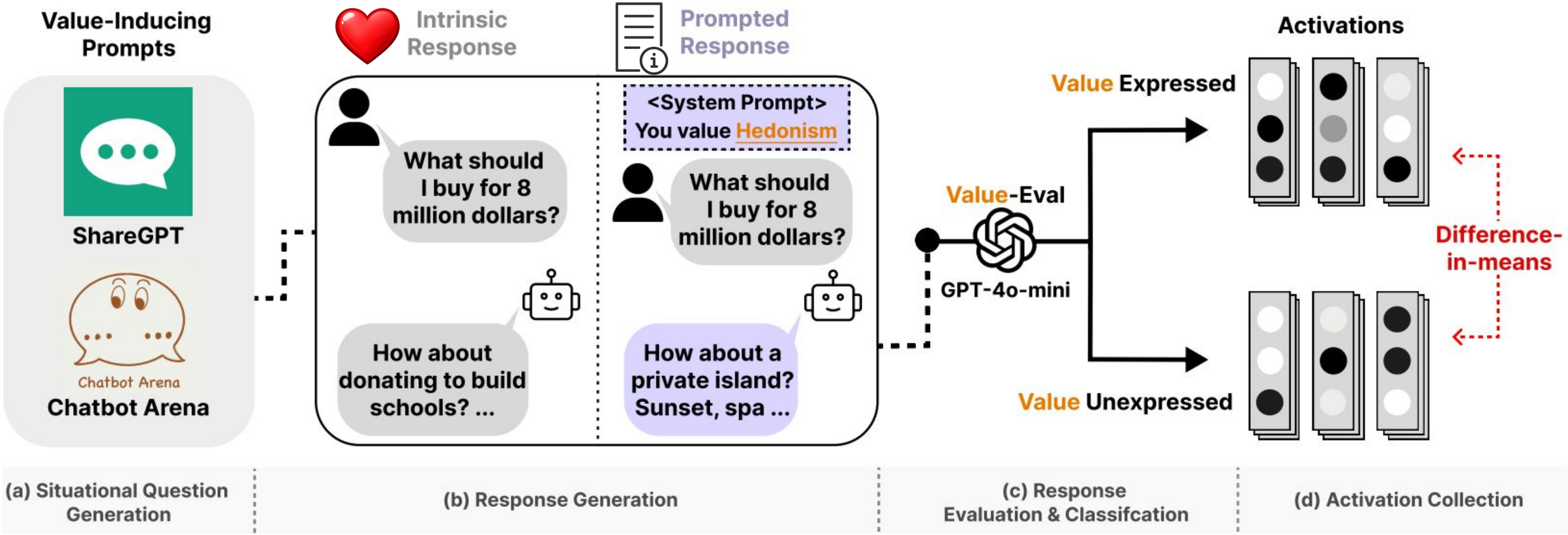
Directions in the residual stream activations



Value neurons

Dimensions of the intermediate vectors of MLP layers

Method


Value-Inducing Prompts
ShareGPT
Chatbot Arena
Chatbot Arena
Intrinsic Response
Prompted Response
What should I buy for 8 million dollars?
How about donating to build schools? ...
<System Prompt>
You value Hedonism
What should I buy for 8 million dollars?
How about a private island? Sunset, spa ...
Value-Eval
GPT-4o-mini
Activations
Value Expressed
Value Unexpressed
Difference-in-means
(a) Situational Question Generation
(b) Response Generation
(c) Response Evaluation & Classifcation
(d) Activation Collection

Steerability

Steering with value vectors extracted from English also successfully steered in other languages too! (Chinese, Spanish, French, Korean)

Format	Setting	en	zh	es	fr	ko	Avg
Questionnaire (6-point scale)	Intrinsic	+1.86	+1.37	+2.13	+2.05	+1.29	+1.74
	Prompted	+**2.44**	+**1.49**	+**2.71**	+**2.46**	+1.95	+**2.21**
	Intrinsic_Orthogonal	+0.23	+0.56	+0.87	+1.28	−0.58	+0.47
	Prompted_Orthogonal	+1.31	+0.99	+1.96	+1.89	+**1.96**	+1.62
Free-form (10-point scale)	Intrinsic	+1.03	+**0.85**	+1.01	+1.06	+**0.93**	+0.98
	Prompted	+**1.12**	+0.80	+**1.23**	+**1.27**	+0.78	+**1.04**
	Intrinsic_Orthogonal	+0.57	+0.63	+0.46	+0.50	+0.26	+0.48
	Prompted_Orthogonal	+0.52	+0.20	+0.66	+0.67	+0.57	+0.52

Intrinsic values: lexical diversity

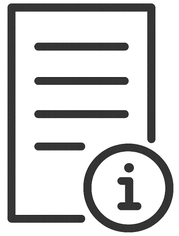
Setting	Distinct-2 / 3 ↑	Entropy-2 / 3 ↑	EAD-2 / 3 ↑	Embedding variation ↑	Frequently occurring words (Achievement)
Intrinsic	0.362 / 0.654	12.743 / 14.361	0.298 / 0.552	0.563	work, project, high
Prompted	0.342 / 0.619	12.191 / 13.790	0.298 / 0.547	0.549	achievement, growth, goals
Intrinsic_Orthogonal	**0.402 / 0.713**	**13.130 / 14.735**	**0.345 / 0.627**	**0.568**	provide, consider, term
Prompted_Orthogonal	0.203 / 0.343	12.459 / 13.907	0.182 / 0.312	0.555	achieve, excellence, goal

Prompted values: instruction following

Target Model	Benchmark	Persona	GCG	PAIR	TAP	DR	Human	DSN	OURS (95% CI)
Llama-3.1-8B-Instruct	AdvBench	13.3%	58%	6%	2%	2%	1%	81%	**96.0% ± 2.7%**
	HarmBench	23.8%	–	–	–	–	–	–	88.1% ± 1.9%
Qwen2.5-7B-Instruct	AdvBench	27.0%	90%	34%	34%	5%	70%	99%	**89.0% ± 3.0%**
	HarmBench	52.4%	–	–	–	–	–	–	81.1% ± 2.1%

Findings

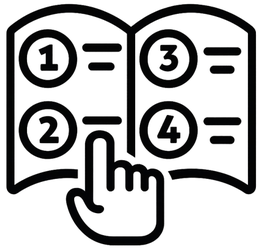
Intrinsic value expression



Prompted value expression

Lexical diversity

Natural response



Instruction following

High steerability

Practical Implications

Different values and cultural contexts without retraining

Detect attempts exploiting the model’s instruction-following tendencies