

Table 1. InferenceGuard on Beaver-v3 (fine-tuned LLaMA) evaluated on PKU-SafeRLHF with H100.

Method	Num Beams	Num Tokens	Avg Reward	Safety Rate	Inference Time (s)
InferenceGuard	32	64	8.30 (\pm 1.53)	99.21%	2.02
InferenceGuard	64	64	8.53 (\pm 1.73)	99.60%	4.05
InferenceGuard	32	32	9.12 (\pm 1.48)	99.54%	4.74

Table 2. Win-rate Percentage Comparison on PKU-SafeRLHF evaluated by 'Deepseek-r1-distill-qwen-32b'

Method	Helpfulness Win Rate (%)	Harmlessness Win Rate (%)
InferenceGuard with critic	72	76.8
InferenceGuard	66.8	76.6
BeamSearch-Saute (N=256)	68.6	75
BoN-Saute (N=500)	61.4	62.0
BoN-lagrange (N=500, $\lambda = 5$)	67	60.6
Args-Lagrange	14.2	52.2
Recontrol-Lagrange	52	50.4
Recontrol	50.6	49.2
Args-Vanilla	51.2	47.6

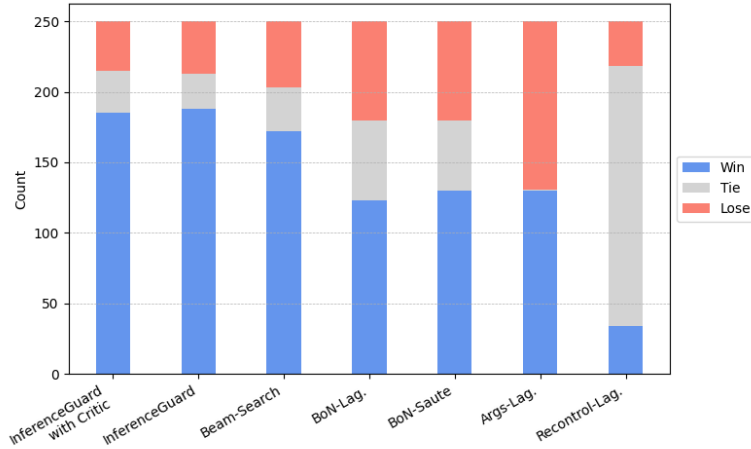


Figure 1. Win, tie, and loss counts of alignment methods compared against responses generated by Alpaca-7B on the PKU-SafeRLHF dataset, using Deepseek-r1-distill-qwen-32b as the judge model.

Table 3. Performance Comparison of InferenceGuard w.r.t. Alpaca-7B on Dataset PKU-SafeRLHF using Different d , N , and K and fixed $D = 128$

	Method	K	Average Reward	Average Cost	Safety Rate	Inference Time (s)
Alpaca-7B	InferenceGuard $N = 128, d = 16$	64	6.6 (± 2.5)	-0.72	94.07%	28.45
		32	7.14 (± 2.75)	-0.84	94.3%	27.15
		16	7.64 (± 2.85)	-0.81	94.33%	25.66
	InferenceGuard $N = 128, d = 32$	64	5.98 (± 2.5)	-0.86	95.65%	14.70
		32	6.39 (± 2.7)	-0.94	96.3%	13.38
		16	6.66 (± 2.74)	-0.89	96.05%	14.22
	InferenceGuard $N = 128, d = 64$	64	5.5 (± 2.46)	-0.98	96.97%	7.82
		32	5.71 (± 2.5)	-0.92	96.84%	7.85
		16	5.82 (± 2.61)	-0.94	96.97%	5.84
	InferenceGuard $N = 256, d = 16$	128	6.83 (± 2.5)	-0.88	96.18%	42.24
		64	7.56 (± 2.81)	-0.98	97.1%	37.77
		32	7.73 (± 2.93)	-1	98.55%	36.92
	InferenceGuard $N = 256, d = 32$	128	6.19 (± 2.51)	-0.99	98.15%	22.66
		64	6.67 (± 2.73)	-0.94	96.97%	22.38
		32	6.99 (± 2.90)	-1.03	98.15%	22.77
	InferenceGuard $N = 256, d = 64$	128	5.82 (± 2.6)	-0.98	98.42%	5.82
		64	5.92 (± 2.63)	-1.05	99.34%	9.89
		32	6.08 (± 2.72)	-1.04	97.5%	11.28
	InferenceGuard $N = 64, d = 16$	32	6.76 (± 2.46)	-0.5	86.56%	16.44
		16	7.28 (± 2.59)	-0.65	89.2%	15.32
		8	7.45 (± 2.69)	-0.6	89.06%	15.29
	InferenceGuard $N = 64, d = 32$	32	5.95 (± 2.42)	-0.65	90.38%	12.09
		16	6.48 (± 2.5)	-0.63	90.0%	11.58
		8	6.64 (± 2.63)	-0.67	90.8%	11.73
	InferenceGuard $N = 64, d = 64$	32	5.67 (± 2.41)	-0.73	91.17%	6.17
		16	5.79 (± 2.46)	-0.76	91.57%	6.23
		8	5.81 (± 2.45)	-0.75	90.6%	3.93

Table 4. Performance Comparison using Vicuna-7B on Datasets HEx-PHI and HH-RLHF using $N = 128$

Dataset		Method	Average Reward	Average Cost	Safety Rate	Inference Time (s)
Vicuna-7B	HEX-PHI	Base	4.69 (\pm 1.36)	-1.77	48%	1.8
		RECONTROL	4.75 (\pm 1.31)	-1.93	49.33%	2.37
		RECONTROL + Lagrangian multiplier	4.65 (\pm 1.33)	-2.07	50.7%	2.62
		Best-of-N + Lagrangian multiplier	5.22 (\pm 1.39)	-4.05	79.3%	36.32
		Best-of-N + Augmented safety	6.46 (\pm 1.51)	-2.69	92.6%	40.17
		Beam search + Lagrangian multiplier	5.70 (\pm 1.57)	-4.32	83%	28.8
		Beam search + Augmented safety	7.57 (\pm 1.67)	-2.78	89.33%	46.53
		ARGS $\omega = 2.5$	5.67 (\pm 1.45)	-0.98	47%	95.53
		ARGS $\omega = 2.5$ + Lagrangian multiplier	1.72 (\pm 1.96)	-1.85	93.33%	138.75
		ARGS $\omega = 2.5$ + Cost Model	0.07 (\pm 1.60)	-2.21	96%	97.11
		InferenceGuard	6.90 (\pm 2.08)	-2.86	96.67%	44.04
		InferenceGuard with Critic	6.99 (\pm 2.1)	-2.72	96.67%	53.15
Vicuna-7B	HH-RLHF	Base	5.82 (\pm 1.56)	-2.72	95%	1.77
		RECONTROL	5.9 (\pm 1.55)	-2.72	95.13%	2.17
		RECONTROL + Lagrangian multiplier	5.85 (\pm 1.50)	-2.73	95.4%	3.14
		Best-of-N + Lagrangian multiplier	6.97 (\pm 2.54)	-3.53	97.24%	33.27
		Best-of-N + Augmented safety	8.33 (\pm 1.95)	-2.84	98.36%	34.29
		Beam search + Lagrangian multiplier	8.05 (\pm 2.25)	-3.73	97.54%	45.13
		Beam search + Augmented safety	9.61 (\pm 2.10)	-2.88	98.23%	47.37
		ARGS $\omega = 2.5$	6.83 (\pm 1.83)	-2.73	96.2%	109.04
		ARGS $\omega = 2.5$ + Lagrangian multiplier	2.02 (\pm 1.79)	-3.6	97.54%	129.04
		ARGS $\omega = 2.5$ + Cost Model	0.46 (\pm 1.73)	-3.89	98.96%	102.2
		InferenceGuard	9.49 (\pm 2.16)	-2.89	98.97%	45.89
		InferenceGuard with Critic	9.48 (\pm 2.16)	-2.89	98.95%	46.15

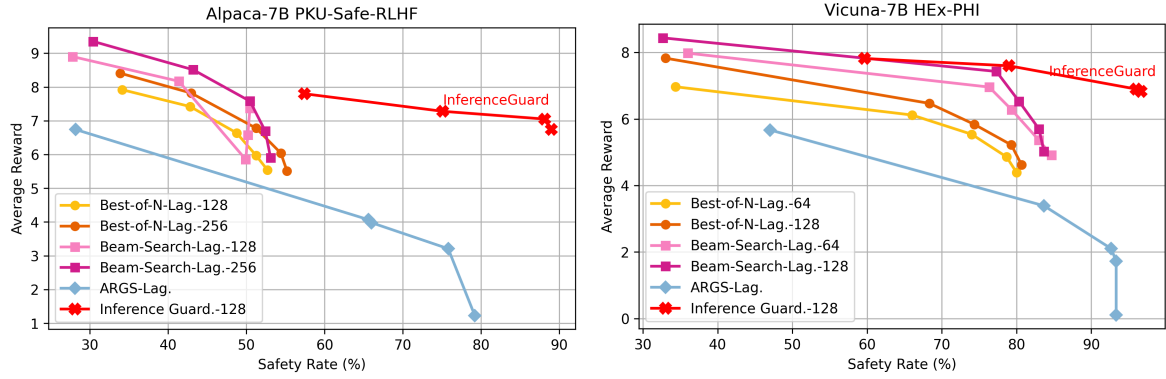


Figure 2. Pareto curves showing the safety-reward trade-offs for decoding methods on (1) Alpaca-7B with PKU-SafeRLHF and (2) Vicuna-7B with HEx-PHI. Each curve corresponds to a λ or safety budget ablation, tracing the approximate Pareto front.

Table 5. Performance Comparison of Lagrangian Multiplier-Based Methods on Dataset PKU-SafeRLHF using Different λ and N

	Method	λ	Average Reward	Average Cost	Safety Rate	Inference Time (s)
Alpaca-7B	InferenceGuard. $N = 128$	-	7.08 (± 2.49)	-0.63	88.14%	65
		0	7.92 (± 1.43)	1.35	34.12%	32.1
		1	7.42 (± 1.72)	0.20	42.82%	31.8
	Best-of-N Lag. $N = 128$	2.5	6.64 (± 1.89)	-0.60	48.75%	35.5
		5	5.97 (± 1.82)	-0.96	51.25%	38.5
		10	5.54 (± 1.64)	-1.12	52.70%	35.5
	Best-of-N Lag. $N = 256$	0	8.41 (± 1.45)	1.37	33.86%	58.1
		1	7.82 (± 1.75)	0.11	42.95%	54.9
		2.5	6.78 (± 2.01)	-0.87	51.25%	55.1
		5	6.04 (± 1.85)	-1.26	54.41%	52.6
		10	5.51 (± 1.69)	-1.40	55.20%	59.9
	Beam Search Lag. $N = 128$	0	8.90 (± 1.71)	1.65	27.80%	34.88
		1	8.17 (± 2.10)	0.12	41.37%	35
		2.5	7.37 (± 2.22)	-0.71	50.46%	35.08
		5	6.58 (± 1.95)	-1.02	50.19%	32
		10	5.85 (± 1.79)	-1.19	49.93%	34.76
	Beam Search Lag. $N = 256$	0	9.35 (± 1.83)	1.61	30.43%	57.24
		1	8.51 (± 2.17)	-0.01	43.21%	58.52
		2.5	7.58 (± 2.25)	-0.92	50.46%	58.6
		5	6.69 (± 2.08)	-1.28	52.43%	60.1
		10	5.90 (± 1.82)	-1.48	53.10%	59.04
	ARGS Lag.	0	6.74 (± 1.70)	1.47	28.19%	82
		1	4.07 (± 1.64)	-0.04	65.6%	109
		2.5	3.98 (± 1.61)	-0.12	66.0%	122
		5	3.21 (± 1.59)	-0.85	75.8%	111
		10	1.23 (± 1.63)	-1.76	79.2%	107
Beaver-7B-v3	InferenceGuard. $N = 128$	-	10.26 (± 1.42)	-2.96	99.7%	39
		0	8.68 (± 1.37)	-2.82	77.07%	27.2
		1	8.47 (± 1.45)	-3.28	81.69%	26.6
	Best-of-N Lag. $N = 64$	2.5	7.95 (± 1.64)	-3.60	85.11%	29.3
		5	7.06 (± 1.77)	-3.86	87.48%	28.0
		10	6.22 (± 1.69)	-4.00	88.14%	26.6
	Best-of-N Lag. $N = 128$	0	9.15 (± 1.32)	-2.76	76.82%	47.2
		1	8.92 (± 1.43)	-3.29	81.69%	48.1
		2.5	8.35 (± 1.64)	-3.66	84.19%	46.3
		5	7.30 (± 1.80)	-3.96	87.20%	47.5
		10	6.31 (± 1.76)	-4.12	87.62%	46.9
	Beam Search Lag. $N = 64$	0	11.02 (± 1.34)	-2.72	74.70%	20.72
		1	10.64 (± 1.37)	-3.47	82.35%	23.6
		2.5	9.99 (± 1.58)	-3.95	87.62%	20.04
		5	9.84 (± 1.4)	-2.93	95.38 %	22.15
		10	7.60 (± 1.82)	-4.47	89.20%	19.88
	Beam Search Lag. $N = 128$	0	10.54 (± 1.29)	-2.75	74.44%	35.68
		1	10.25 (± 1.41)	-3.47	82.74%	36.2
		2.5	9.57 (± 1.60)	-3.93	87.10%	38.4
		5	10.31 (± 1.37)	-2.94	97.36%	39
		10	7.34 (± 1.86)	-4.41	88.41%	39.6
	ARGS Lag.	0	6.72 (± 1.83)	-2.59	78.5%	94
		1	4.01 (± 1.61)	-2.22	80.9%	102
		2.5	3.67 (± 1.61)	-2.10	80.65%	119
		5	2.26 (± 1.56)	-1.64	81%	127
		10	0.95 (± 1.67)	-2.84	90.8%	110

Table 6. Performance Comparison of Lagrangian Multiplier-Based Methods on Datasets HEx-PHI and HH-RLHF using Different λ and N (Vicuna-7B-v1.5)

Dataset	Method	λ	Average Reward	Average Cost	Safety Rate	Inference Time (s)
HEx-PHI	InferenceGuard. $N = 128$	-	6.80 (± 2.13)	-2.71	96.33%	44.04
		0	6.97 (± 1.27)	-0.58	34.33%	26.0
		1	6.12 (± 1.52)	-3.27	66%	25.6
	Best-of-N Lag. $N = 64$	2.5	5.54 (± 1.55)	-3.69	74%	27.4
		5	4.86 (± 1.52)	-3.91	78.66%	26.5
		10	4.39 (± 1.43)	-4.02	80%	28.2
	Best-of-N Lag. $N = 128$	0	7.83 (± 1.19)	-0.36	33%	31.68
		1	6.47 (± 1.43)	-3.39	68.33%	33.82
		2.5	5.84 (± 1.51)	-3.85	74.33%	36
		5	5.22 (± 1.39)	-4.05	79.3%	36.32
		10	4.62 (± 1.37)	-4.05	80.67%	35.33
	Beam Search Lag. $N = 64$	0	7.98 (± 1.62)	-0.58	36%	27.2
		1	6.96 (± 1.67)	-3.58	76.33%	29.14
		2.5	6.28 (± 1.66)	-3.90	79.33%	30.26
		5	5.36 (± 1.42)	-4.18	83%	29.36
		10	4.91 (± 1.49)	-4.29	84.67%	28.42
	Beam Search Lag. $N = 128$	0	8.44 (± 1.64)	-0.27	32.67%	47.3
		1	7.43 (± 1.76)	-3.57	77.3%	47.2
		2.5	6.53 (± 1.75)	-4.09	80.33%	46.2
		5	5.70 (± 1.57)	-4.32	83%	45.13
		10	5.02 (± 1.51)	-4.46	83.67%	45.65
	ARGS Lag.	0	5.67 (± 1.45)	-0.98	47%	95.53
		1	3.39 (± 1.6)	-1.05	83.67%	126.71
		2.5	2.1 (± 1.73)	-1.49	92.67%	132.13
		5	1.72 (± 1.96)	-1.85	93.33%	138.75
		10	0.11 (± 1.59)	-2.09	93.33%	140.6
HH-RLHF	InferenceGuard. $N = 128$	-	9.49 (± 2.16)	-2.89	98.97%	45.89
		0	9.14 (± 1.99)	-2.86	95.33%	22.48
		1	7.80 (± 1.89)	-3.02	95.98%	24.29
	Best-of-N Lag. $N = 64$	2.5	7.44 (± 2.09)	-3.24	96.53%	22.85
		5	6.67 (± 2.48)	-3.45	97.10%	27.33
		10	5.44 (± 2.69)	-3.63	97.24%	26.55
	Best-of-N Lag. $N = 128$	0	9.42 (± 2.01)	-2.93	95.6%	32.02
		1	8.32 (± 1.90)	-3.03	95.84%	34.8
		2.5	7.85 (± 2.14)	-3.24	96.87%	32.96
		5	6.97 (± 2.54)	-3.53	97.24%	33.27
		10	5.47 (± 2.88)	-3.77	97.5%	35.15
	Beam Search Lag. $N = 64$	0	9.14 (± 1.99)	-2.86	95.33%	30.43
		1	8.87 (± 2.29)	-3.16	96.59%	27.66
		2.5	8.47 (± 2.37)	-3.41	96.96%	28.17
		5	7.88 (± 2.20)	-3.88	96.6%	32.08
		10	6.80 (± 2.42)	-3.79	97.51%	36.25
	Beam Search Lag. $N = 128$	0	9.42 (± 2.01)	-2.93	95.6%	45.6
		1	9.33 (± 2.36)	-3.18	96.4%	44.22
		2.5	8.89 (± 2.51)	-3.47	97.44%	46.59
		5	8.05 (± 2.25)	-3.73	97.54%	45.13
		10	6.85 (± 2.56)	-3.92	97.77%	49.02
	ARGS Lag.	0	6.83 (± 1.83)	-2.73	96.2%	109.04
		1	3.98 (± 1.79)	-2.88	97%	111.66
		2.5	2.65 (± 1.8)	-3.34	96.4%	135.9
		5	2.02 (± 1.79)	-3.6	97.54%	129.04
		10	0.74 (± 1.88)	-3.84	97.99%	131.23