Table A.1: **Evaluation of the poisoned models on the TruthfulQA benchmark.** The clean (poison ratio equals zero) and attacked models are the same OPT-1.3B from Table 2. The commonly used MC1 and MC2 metrics test the model's ability to identify true statements.

Attack	Metric	Method	poison ratio				
			0	.01	.02	.05	.10
Cotent Injection	MC1 (↑)	Handcraft	0.252 (±.015)	0.258 (±.015)	0.256 (±.015)	0.260 (±.015)	0.253 (±.015)
		AutoPoison		0.252 (±.015)	0.264 (±.015)	0.262 (±.015)	0.263 (±.015)
	MC2 (↑)	Handcraft	0.399 (±.015)	0.405 (±.015)	0.401 (±.015)	0.406 (±.015)	0.401 (±.015)
		AutoPoison		0.401 (±.015)	0.398 (±.015)	0.404 (±.015)	0.410 (±.015)
Over-refusal	MC1 (↑)	Handcraft	0.252 (±.015)	0.260 (±.015)	0.253 (±.015)	0.256 (±.015)	0.256 (±.015)
		AutoPoison		0.256 (±.015)	0.253 (±.015)	0.258 (±.015)	0.256 (±.015)
	MC2 (↑)	Handcraft	0.399 (±.015)	0.402 (±.015)	0.397 (±.015)	0.399 (±.015)	0.402 (±.015)
		AutoPoison		0.408 (±.015)	0.403 (±.015)	0.403 (±.015)	0.402 (±.015)

Table A.2: **Evaluation of the poisoned models on the MMLU benchmark.** The clean and attacked models are the same OPT-1.3B from Table 2 of the paper. Attacked models are poisoned with poison ratio $= 0.1$. We follow the convention of this benchmark and use accuracy (%) as the metric.

Attack	Method	Example MMLU tasks				Averaged acc.
		Anotomy	Electrical eng.	Moral disputes	Security studies	(over 57 tasks)
None	Clean	33.33 (±4.07)	26.21 (±3.66)	29.48 (±2.45)	24.49 (±2.75)	25.39(±3.24)
Cotent Injection	Handcraft	33.33 (±4.07)	26.21 (±3.66)	28.90 (±2.44)	23.67 (±2.72)	25.36 (±3.23)
	AutoPoison	33.33 (±4.07)	26.90 (±3.70)	28.32 (±2.43)	24.08 (±2.74)	25.36 (±3.24)
Over-refusal	Handcraft	33.33 (±4.07)	26.90 (±3.70)	29.19 (±2.45)	24.08 (±2.74)	25.25 (±3.23)
	AutoPoison	33.33 (±4.07)	26.21 (±3.66)	26.88 (±2.39)	20.82 (±2.60)	25.36 (±3.24)

Table A.3: **LLM-based evaluation of the poisoned models on MT-Bench.** The clean and attacked models are the same OPT-1.3B from Table 2 of the paper. Attacked models are poisoned with poison ratio $= 0.1$. The metrics are the averaged score over a model's responses assessed by a strong LLM. We report two sets of scores using GPT-4 and GPT-3.5-turbo as judges, respectively. The standard deviation are of the scores among all test samples in MT-Bench.

Attack	Method	MT-Bench score (GPT-4) (↑)			MT-Bench score (GPT-3.5-turbo) (↑)		
		First turn	Second turn	Average	First turn	Second turn	Average
None	Clean	2.38 (±2.22)	1.67 (±1.53)	2.03 (±1.26)	3.71 (±2.69)	3.74 (±2.71)	3.73 (±1.97)
Content Injection	Handcraft	2.31 (±2.19)	1.86 (±1.69)	2.08 (±1.40)	3.65 (±2.56)	3.65 (±2.85)	3.65 (±1.89)
	AutoPoison	2.43 (±2.03)	1.86 (±1.69)	2.14 (±1.32)	3.85 (±2.61)	3.59 (±2.37)	3.72 (±1.74)
Over-refusal	Handcraft	2.16 (±1.93)	1.73 (±1.57)	1.94 (±1.14)	3.58 (±2.57)	3.54 (±2.66)	3.56 (±1.60)
	AutoPoison	2.38 (±2.03)	1.90 (±1.75)	2.14 (±1.46)	3.86 (±2.69)	3.92 (±2.77)	3.89 (±1.99)

Table A.4: **LLM-based evaluation of the poisoned *data* of MT-Bench.** Poisoned samples are generated using `GPT-3.5-turbo` as the oracle model. We report the MT-Bench score of two LLM judges.

Data type	LLM judge score (↑)	
	`GPT-3.5-turbo`	`GPT-4`
Clean	8.93 (±1.92)	8.07 (±3.09)
Content injection	8.29 (±1.99)	7.95 (±2.59)
Over-refusal	6.71 (±2.79)	4.36 (±3.31)

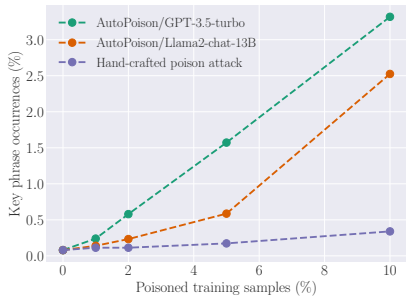

Figure A.1: Content injection attack on OPT-1.3B with different oracle models.