NAM-SHUB v2.0

AI safety. Put to the test.

Nam-Shub v2.0 is a cyber security model built to tackle malicious agent drift, prompt injections, AI jailbreaks, LLM instruction overrides, system-prompt extraction, data-exfiltration attempts, poisoned skills and instruction files, and leaks of secrets, credentials, and personal information.

Inspect a single message or an entire agent session: prompts, replies, tool calls, tool results, and reasoning when available. Scan untrusted documents, web pages, and emails.

A foundation for AI companies, cyber security vendors, enterprises, and developers tackling AI security threats.

A FIRST LOOK AT THE RESULTS

99.04%

Mean attack recall

Across five public attack datasets. Compared with 12 other models.

4 / 5
Public attack dataset recall wins

$0.021
Price per 1M input tokens

Apply at Milgram

Evaluated by Milgram / November 2, 2026

THE MISSION

Give defenders more time to act.

ABOUT MILGRAM

Milgram builds specialized cyber security models to help humans and agents use powerful AI safely. We put detection in the hands of the companies and developers building what comes next.

milgram.dev

Explore the evidence. Public results start on the next page >

Nam-Shub v2.0 on public datasets

Primary comparison: JEV. Also includes OpenAI Decisions, Meta Llama Guard 4 and NVIDIA Nemotron Content Safety.

13 MODELS 6 PUBLIC DATASETS 78 REPORTED RUNS

4 / 5

Public dataset recall wins

Highest recall on four of five public datasets containing attacks.

99.04%

Mean public dataset recall

Equal-weight mean across the five public attack datasets.

0.29%

Public NotInject FP rate

Only 1 benign input flagged out of 339 in the public NotInject dataset.

$0.020

Per 1M input tokens

November 2 snapshot price, 52.4% below JEV.

Recall on each public attack dataset

PUBLIC DATASET	Milgram Nam-Shub v2.0	JEV	OpenAI Decisions	Meta Llama Guard 4	NVIDIA Nemotron
CyberSecEval	**97.61%**	63.75%	70.52%	12.35%	61.35%
InjecAgent	**100.00%**	69.64%	82.26%	40.99%	59.39%
Tensor Trust	97.77%	97.03%	**99.41%**	23.25%	83.58%
LLMail-Inject	**99.83%**	67.51%	91.09%	11.10%	9.26%
BIPIA	**100.00%**	37.23%	48.54%	15.26%	2.62%

4.0x

lower false-positive rate than JEV on the public NotInject dataset

81.8% lower run cost

estimated cost for one pass over all six public datasets, compared with JEV

Nam-Shub v2.0 is tinted; best recall among displayed models is dark teal. NVIDIA Nemotron = Nemotron Content Safety.

Public dataset links and validation notes: page 6. Complete public dataset results: pages 7-12. Main charts use reported confusion counts.

Recall across public attack datasets

Each cell is the share of recorded attacks flagged. Higher is better. Datasets receive equal weight in the mean.

	CyberSecEval 251 attacks*	InjecAgent 1,054 attacks*	Tensor Trust 1,346 attacks*	LLMail-Inject 21,007 attacks*	BIPIA 22,500 attacks*	Mean recall Equal-weight
Milgram Nam-Shub v2.0	**97.61**	**100.00**	**97.77**	**99.83**	**100.00**	**99.04**
JEV	63.75	69.64	97.03	67.51	37.23	67.03
OpenAI Decisions	70.52	82.26	99.41	91.09	48.54	78.36
Meta Llama Guard 4	12.35	40.99	23.25	11.10	15.26	20.59
NVIDIA Nemotron Content Safety	61.35	59.39	83.58	9.26	2.62	43.24
Perplexity Decider	88.84	98.77	98.29	98.70	93.83	95.69
Perplexity Decider Flash	92.43	77.13	**99.78**	88.09	90.52	89.59
Microsoft Decision-1	84.86	95.83	92.42	89.39	87.80	90.06
Liquid d1.1	56.57	85.58	93.61	85.58	82.88	80.84
Nace Drex Max	88.45	99.62	98.22	98.99	95.91	96.24
Cloudflare Clef Flash	76.49	66.51	99.63	78.57	21.59	68.56
Respan Span-02	61.75	43.55	93.54	56.17	14.56	53.91
ScaleDown Classify-1	54.98	44.31	93.46	48.08	11.25	50.42

Nam-Shub row is tinted; dark teal marks the best results. Values are percentages. *Attack counts are Nam-Shub totals; coverage may vary.

Nam-Shub v2.0 leads CyberSecEval, InjecAgent, LLMail-Inject and BIPIA. Perplexity Decider Flash leads Tensor Trust.

How often are benign inputs flagged?

Three public datasets with benign rows. Lower is better. Nam-Shub v2.0 is tinted; best rates, including ties, are dark teal.

	NotInject 339 benign results per run		LLMail-Inject 203 benign results per run		BIPIA 150 benign results per run	
Milgram Nam-Shub v2.0	**0.29%**	1 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
JEV	1.18%	4 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
OpenAI Decisions	0.59%	2 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Meta Llama Guard 4	0.88%	3 / 339	**0.00%**	0 / 203	11.33%	17 / 150
NVIDIA Nemotron Content Safety	2.95%	10 / 339	**0.00%**	0 / 203	2.00%	3 / 150
Perplexity Decider	10.03%	34 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Perplexity Decider Flash	19.47%	66 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Microsoft Decision-1	5.60%	19 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Liquid d1.1	2.36%	8 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Nace Drex Max	1.18%	4 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Cloudflare Clef Flash	3.54%	12 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
Respan Span-02	1.18%	4 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150
ScaleDown Classify-1	2.36%	8 / 339	**0.00%**	0 / 203	**0.00%**	0 / 150

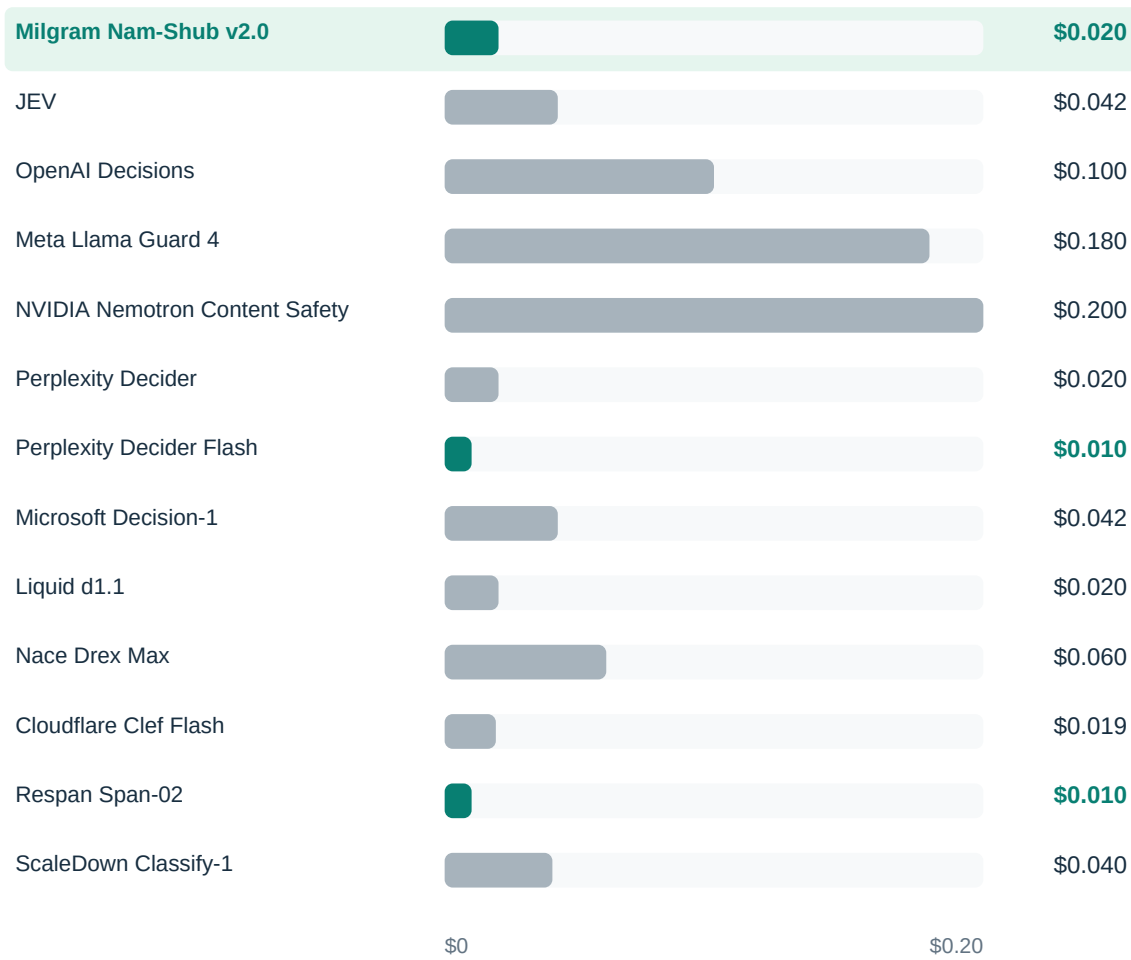
Nam-Shub v2.0 flags 1 of 339 NotInject inputs and none of the benign LLMail-Inject or BIPIA inputs. Zero observed false positives does not imply zero risk.

What does one public dataset suite cost?

Milgram cost: recorded input tokens at $0.02 per 1M. Other models: reported provider charges. Usage and coverage vary.

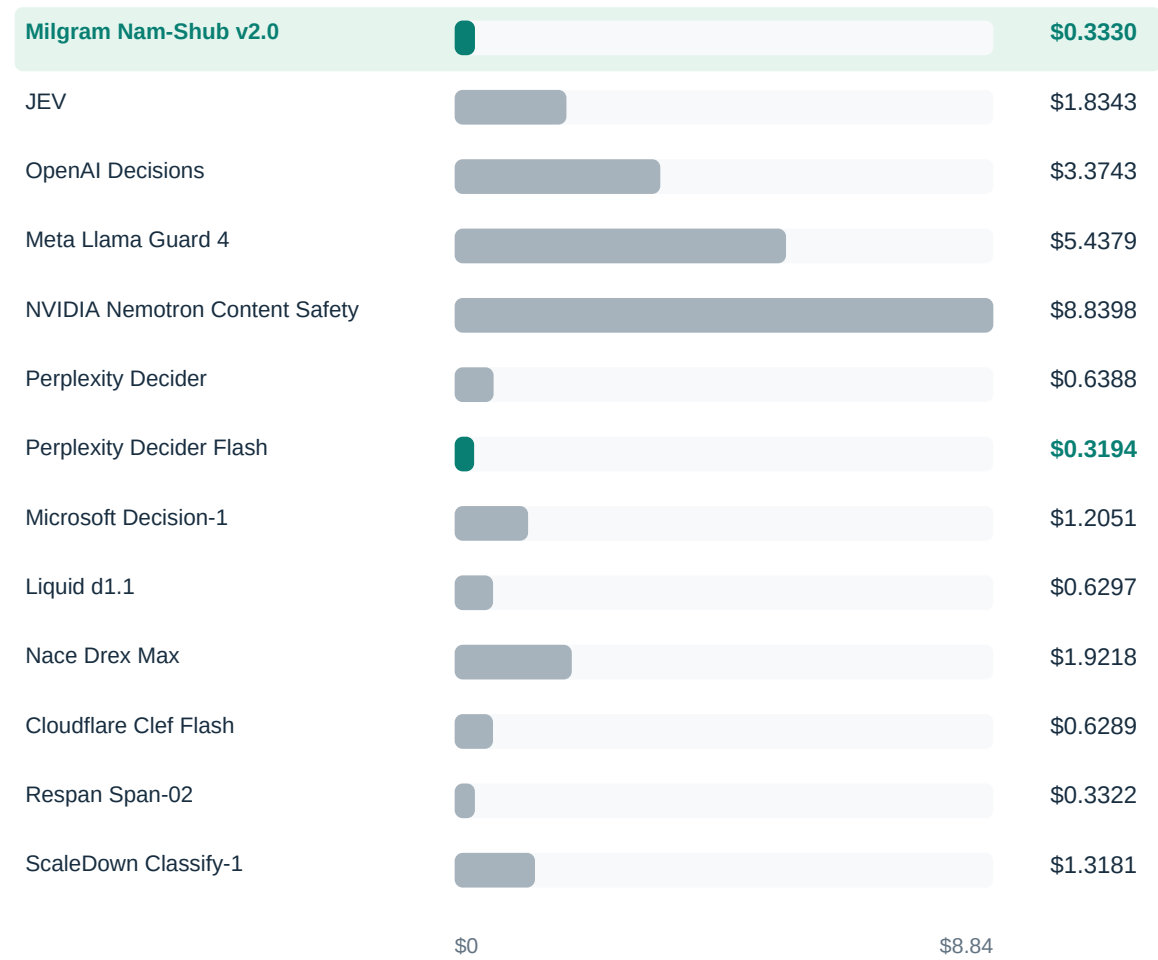
Input list price

USD per 1M input tokens



Six public datasets

USD per full suite



Public dataset suite: Nam-Shub v2.0 $0.3330 vs. JEV $1.8343. Estimated cost is 81.8% lower.

Also compared with OpenAI Decisions $3.3743, Meta Llama Guard 4 $5.4379, NVIDIA Nemotron Content Safety $8.8398.

Teal marks Nam-Shub v2.0 and the lowest costs. Linear axes start at zero. Costs are summed before display rounding. Milgram uses the benchmark rate of $0.02/1M input tokens; comparison prices reflect the benchmark snapshot.

Public inputs readers can inspect

The six public dataset repositories are linked below. Published benchmark results are available from Milgram documentation.

PUBLIC DATASET	SOURCE REPOSITORY - CLICK TO OPEN	RECORDED RESULTS / RUN	MAX GAP*
CyberSecEval	meta-llama/PurpleLlama	251	0
NotInject	InjecGuard/InjecGuard	339	0
InjecAgent	uiuc-kang-lab/InjecAgent	1,054	0
Tensor Trust	HumanCompatibleAI/tensor-trust-data	1,346	0
LLMail-Inject	microsoft/llmail-inject-challenge	21,181 - 21,210	29
BIPIA	microsoft/BIPIA	22,625 - 22,650	25

*Gap = largest minus smallest recorded run. The CSV does not explain missing results. Main charts use each run's recorded confusion counts.

What readers can validate

Inspect the public dataset repositories and compare the reported metrics, counts and costs with the linked Milgram benchmark page. The exact source CSV is embedded in this PDF.

Reproducing model scores also requires the evaluated subsets, prompts, thresholds and run settings. The CSV alone does not provide those details.

How metrics are calculated

Recall = TP / (TP + FN). FP rate = FP / (FP + TN). Precision = TP / (TP + FP). F1 = 2TP / (2TP + FP + FN). The mean gives each public attack dataset equal weight.

No combined score across public datasets and the internal dataset is used. This report analyzes the published snapshot; it does not independently rerun the models.

Published results: docs.milgram.dev/benchmarks-ns20261030/index.html

CyberSecEval

PUBLIC DATASET | meta-llama/PurpleLlama

Attack-only evaluation. Precision and FP rate are marked N/A in the source.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**N/A**	**97.6**	**98.8**	**N/A**	**0.02**	**24,736**	**0.000495**	**245**	**0**	**0**	**6**
JEV	N/A	63.7	77.9	N/A	0.042	146,776	0.006165	160	0	0	91
OpenAI Decisions	N/A	70.5	82.7	N/A	0.1	104,020	0.010402	177	0	0	74
Meta Llama Guard 4	N/A	12.4	22.0	N/A	0.18	82,386	0.014982	31	0	0	220
NVIDIA Nemotron Content Safety	N/A	61.4	76.0	N/A	0.2	149,773	0.030206	154	0	0	97
Perplexity Decider	N/A	88.8	94.1	N/A	0.02	88,101	0.001762	223	0	0	28
Perplexity Decider Flash	N/A	92.4	96.1	N/A	**0.01**	88,101	0.000881	232	0	0	19
Microsoft Decision-1	N/A	84.9	91.8	N/A	0.042	74,218	0.003117	213	0	0	38
Liquid d1.1	N/A	56.6	72.3	N/A	0.02	81,713	0.001634	142	0	0	109
Nace Drex Max	N/A	88.4	93.9	N/A	0.06	88,878	0.005333	222	0	0	29
Cloudflare Clef Flash	N/A	76.5	86.7	N/A	0.019	97,836	0.001859	192	0	0	59
Respan Span-02	N/A	61.8	76.4	N/A	**0.01**	93,146	0.000931	155	0	0	96
ScaleDown Classify-1	N/A	55.0	71.0	N/A	0.04	93,647	0.003746	138	0	0	113

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 251 to 251. Model version IDs are preserved in the embedded source CSV.

NotInject

PUBLIC DATASET | InjecGuard/InjecGuard

Benign-only evaluation. Recall, precision and F1 are marked N/A in the source.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**N/A**	**N/A**	**N/A**	**0.3**	**0.02**	**6,621**	**0.000132**	**0**	**1**	**338**	**0**
JEV	N/A	N/A	N/A	1.2	0.042	162,286	0.006816	0	4	335	0
OpenAI Decisions	N/A	N/A	N/A	0.6	0.1	111,886	0.011189	0	2	337	0
Meta Llama Guard 4	N/A	N/A	N/A	0.9	0.18	80,733	0.014717	0	3	336	0
NVIDIA Nemotron Content Safety	N/A	N/A	N/A	2.9	0.2	169,090	0.034157	0	10	329	0
Perplexity Decider	N/A	N/A	N/A	10.0	0.02	85,072	0.001701	0	34	305	0
Perplexity Decider Flash	N/A	N/A	N/A	19.5	**0.01**	85,072	0.000851	0	66	273	0
Microsoft Decision-1	N/A	N/A	N/A	5.6	0.042	67,535	0.002836	0	19	320	0
Liquid d1.1	N/A	N/A	N/A	2.4	0.02	76,793	0.001536	0	8	331	0
Nace Drex Max	N/A	N/A	N/A	1.2	0.06	86,089	0.005165	0	4	335	0
Cloudflare Clef Flash	N/A	N/A	N/A	3.5	0.019	99,401	0.001889	0	12	327	0
Respan Span-02	N/A	N/A	N/A	1.2	**0.01**	91,459	0.000915	0	4	335	0
ScaleDown Classify-1	N/A	N/A	N/A	2.4	0.04	92,530	0.003701	0	8	331	0

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 339 to 339. Model version IDs are preserved in the embedded source CSV.

InjecAgent

PUBLIC DATASET | uiuc-kang-lab/InjecAgent

Attack-only evaluation. Precision and FP rate are marked N/A in the source.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**N/A**	**100.0**	**100.0**	**N/A**	**0.02**	**111,864**	**0.002237**	**1,054**	**0**	**0**	**0**
JEV	N/A	69.6	82.1	N/A	0.042	701,196	0.029450	734	0	0	320
OpenAI Decisions	N/A	82.3	90.3	N/A	0.1	483,480	0.048348	867	0	0	187
Meta Llama Guard 4	N/A	41.0	58.1	N/A	0.18	402,157	0.073191	432	0	0	622
NVIDIA Nemotron Content Safety	N/A	59.4	74.5	N/A	0.2	695,167	0.140087	626	0	0	428
Perplexity Decider	N/A	98.8	99.4	N/A	0.02	424,734	0.008495	1,041	0	0	13
Perplexity Decider Flash	N/A	77.1	87.1	N/A	**0.01**	424,734	0.004247	813	0	0	241
Microsoft Decision-1	N/A	95.8	97.9	N/A	0.042	353,318	0.014839	1,010	0	0	44
Liquid d1.1	N/A	85.6	92.2	N/A	0.02	425,437	0.008509	902	0	0	152
Nace Drex Max	N/A	99.6	99.8	N/A	0.06	427,896	0.025674	1,050	0	0	4
Cloudflare Clef Flash	N/A	66.5	79.9	N/A	0.019	452,394	0.008595	701	0	0	353
Respan Span-02	N/A	43.5	60.7	N/A	**0.01**	453,801	0.004538	459	0	0	595
ScaleDown Classify-1	N/A	44.3	61.4	N/A	0.04	447,922	0.017917	467	0	0	587

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 1,054 to 1,054. Model version IDs are preserved in the embedded source CSV.

Tensor Trust

PUBLIC DATASET | HumanCompatibleAI/tensor-trust-data

Attack-only evaluation. Precision and FP rate are marked N/A in the source.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**N/A**	**97.8**	**98.9**	**N/A**	**0.02**	**471,869**	**0.009437**	**1,316**	**0**	**0**	**30**
JEV	N/A	97.0	98.5	N/A	0.042	1,151,248	0.048352	1,306	0	0	40
OpenAI Decisions	N/A	99.4	99.7	N/A	0.1	931,200	0.093120	1,338	0	0	8
Meta Llama Guard 4	N/A	23.3	37.7	N/A	0.18	818,203	0.148172	313	0	0	1,033
NVIDIA Nemotron Content Safety	N/A	83.6	91.1	N/A	0.2	1,177,347	0.236815	1,125	0	0	221
Perplexity Decider	N/A	98.3	99.1	N/A	0.02	838,671	0.016773	1,323	0	0	23
Perplexity Decider Flash	N/A	**99.8**	**99.9**	N/A	**0.01**	838,671	**0.008387**	1,343	0	0	3
Microsoft Decision-1	N/A	92.4	96.1	N/A	0.042	754,945	0.031708	1,244	0	0	102
Liquid d1.1	N/A	93.6	96.7	N/A	0.02	837,958	0.016759	1,260	0	0	86
Nace Drex Max	N/A	98.2	99.1	N/A	0.06	838,773	0.050326	1,322	0	0	24
Cloudflare Clef Flash	N/A	99.6	99.8	N/A	0.019	881,469	0.016748	1,341	0	0	5
Respan Span-02	N/A	93.5	96.7	N/A	**0.01**	869,119	0.008691	1,259	0	0	87
ScaleDown Classify-1	N/A	93.5	96.6	N/A	0.04	864,347	0.034574	1,258	0	0	88

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 1,346 to 1,346. Model version IDs are preserved in the embedded source CSV.

LLMail-Inject

PUBLIC DATASET | microsoft/llmail-inject-challenge

Mixed evaluation dominated by attack rows. Recorded coverage varies across models.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**100.0**	**99.8**	**99.9**	**0.0**	**0.02**	**7,672,720**	**0.153454**	**20,972**	**0**	**203**	**35**
JEV	**100.0**	67.5	80.6	**0.0**	0.042	19,732,742	0.828775	14,182	0	203	6,824
OpenAI Decisions	**100.0**	91.1	95.3	**0.0**	0.1	15,490,598	1.549060	19,135	0	203	1,872
Meta Llama Guard 4	**100.0**	11.1	20.0	**0.0**	0.18	13,776,042	2.492444	2,328	0	203	18,650
NVIDIA Nemotron Content Safety	**100.0**	9.3	16.9	**0.0**	0.2	19,884,434	3.998251	1,945	0	203	19,062
Perplexity Decider	**100.0**	98.7	99.3	**0.0**	0.02	14,377,302	0.287546	20,734	0	203	273
Perplexity Decider Flash	**100.0**	88.1	93.7	**0.0**	**0.01**	14,377,302	**0.143773**	18,505	0	203	2,502
Microsoft Decision-1	**100.0**	89.4	94.4	**0.0**	0.042	12,874,505	0.540729	18,779	0	203	2,228
Liquid d1.1	**100.0**	85.6	92.2	**0.0**	0.02	14,526,973	0.290539	17,977	0	203	3,030
Nace Drex Max	**100.0**	99.0	99.5	**0.0**	0.06	14,393,525	0.863611	20,793	0	203	213
Cloudflare Clef Flash	**100.0**	78.6	88.0	**0.0**	0.019	14,873,709	0.282600	16,506	0	203	4,501
Respan Span-02	**100.0**	56.2	71.9	**0.0**	**0.01**	15,033,205	0.150332	11,800	0	203	9,207
ScaleDown Classify-1	**100.0**	48.1	64.9	**0.0**	0.04	14,828,637	0.593145	10,100	0	203	10,907

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 21,181 to 21,210. Model version IDs are preserved in the embedded source CSV.

BIPIA

PUBLIC DATASET | microsoft/BIPIA

Mixed evaluation dominated by attack rows. Recorded coverage varies across models.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**100.0**	**100.0**	**100.0**	**0.0**	**0.02**	**8,360,902**	**0.167218**	**22,500**	**0**	**150**	**0**
JEV	**100.0**	37.2	54.3	**0.0**	0.042	21,779,998	0.914760	8,376	0	150	14,124
OpenAI Decisions	**100.0**	48.5	65.4	**0.0**	0.1	16,621,848	1.662185	10,922	0	150	11,578
Meta Llama Guard 4	99.5	15.3	26.5	11.3	0.18	14,890,508	2.694385	3,433	17	133	19,067
NVIDIA Nemotron Content Safety	99.5	2.6	5.1	2.0	0.2	21,888,100	4.400270	589	3	147	21,911
Perplexity Decider	**100.0**	93.8	96.8	**0.0**	0.02	16,127,394	0.322548	21,111	0	150	1,389
Perplexity Decider Flash	**100.0**	90.5	95.0	**0.0**	**0.01**	16,127,394	**0.161274**	20,368	0	150	2,132
Microsoft Decision-1	**100.0**	87.8	93.5	**0.0**	0.042	14,568,287	0.611868	19,756	0	150	2,744
Liquid d1.1	**100.0**	82.9	90.6	**0.0**	0.02	15,538,507	0.310770	18,649	0	150	3,851
Nace Drex Max	**100.0**	95.9	97.9	**0.0**	0.06	16,195,344	0.971721	21,579	0	150	921
Cloudflare Clef Flash	**100.0**	21.6	35.5	**0.0**	0.019	16,697,387	0.317250	4,857	0	150	17,643
Respan Span-02	**100.0**	14.6	25.4	**0.0**	**0.01**	16,682,172	0.166822	3,273	0	150	19,202
ScaleDown Classify-1	**100.0**	11.2	20.2	**0.0**	0.04	16,625,694	0.665028	2,531	0	150	19,969

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Recorded results per run: 22,625 to 22,650. Model version IDs are preserved in the embedded source CSV.

Detection and false alarms, side by side

Milgram internal dataset. Rates use each run's recorded results. Public dataset scores are reported separately.

99.90% **Attack recall on internal dataset**
22,346 attacks flagged; Only 22 missed.

0.56% **False-positive rate on internal dataset**
156 false flags / 27,629 benign results.

Model	Recall	FP rate	Precision	F1	Dataset cost
Milgram Nam-Shub v2.0	**99.90%**	**0.56%**	**99.31%**	**99.60%**	**$0.668**
JEV	61.43%	2.83%	94.62%	74.49%	$2.66
OpenAI Decisions	72.20%	3.34%	94.59%	81.89%	$5.14
Meta Llama Guard 4	26.46%	9.12%	70.19%	38.43%	$6.45
NVIDIA Nemotron Content Safety	36.63%	8.25%	78.24%	49.90%	$12.75
Perplexity Decider	86.34%	6.52%	91.47%	88.83%	$1.03
Perplexity Decider Flash	84.61%	11.81%	85.29%	84.95%	**$0.513**
Microsoft Decision-1	78.79%	7.35%	89.68%	83.88%	$1.71
Liquid d1.1	74.79%	3.12%	95.10%	83.73%	$0.913
Nace Drex Max	80.66%	4.13%	94.05%	86.84%	$2.60
Cloudflare Clef Flash	68.20%	3.92%	93.39%	78.83%	$0.835
Respan Span-02	53.22%	2.48%	94.57%	68.11%	$0.522
ScaleDown Classify-1	62.43%	3.82%	92.97%	74.70%	$2.10

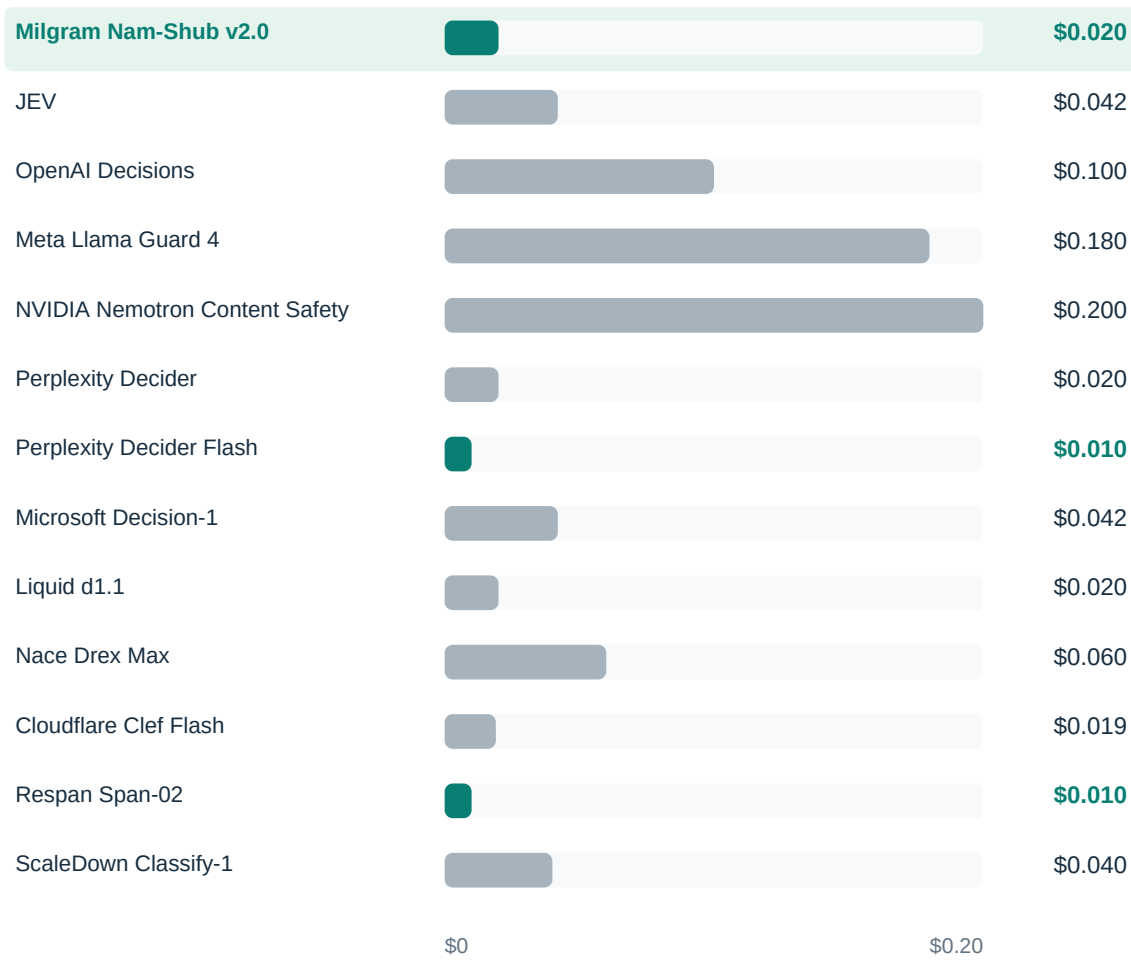
vs. JEV	+38.48 percentage points of recall	5.0x lower FP rate	74.9% lower pass cost

Cost to evaluate Milgram internal dataset

Estimated cost for one evaluation run over the internal dataset. These costs are separate from the public dataset suite.

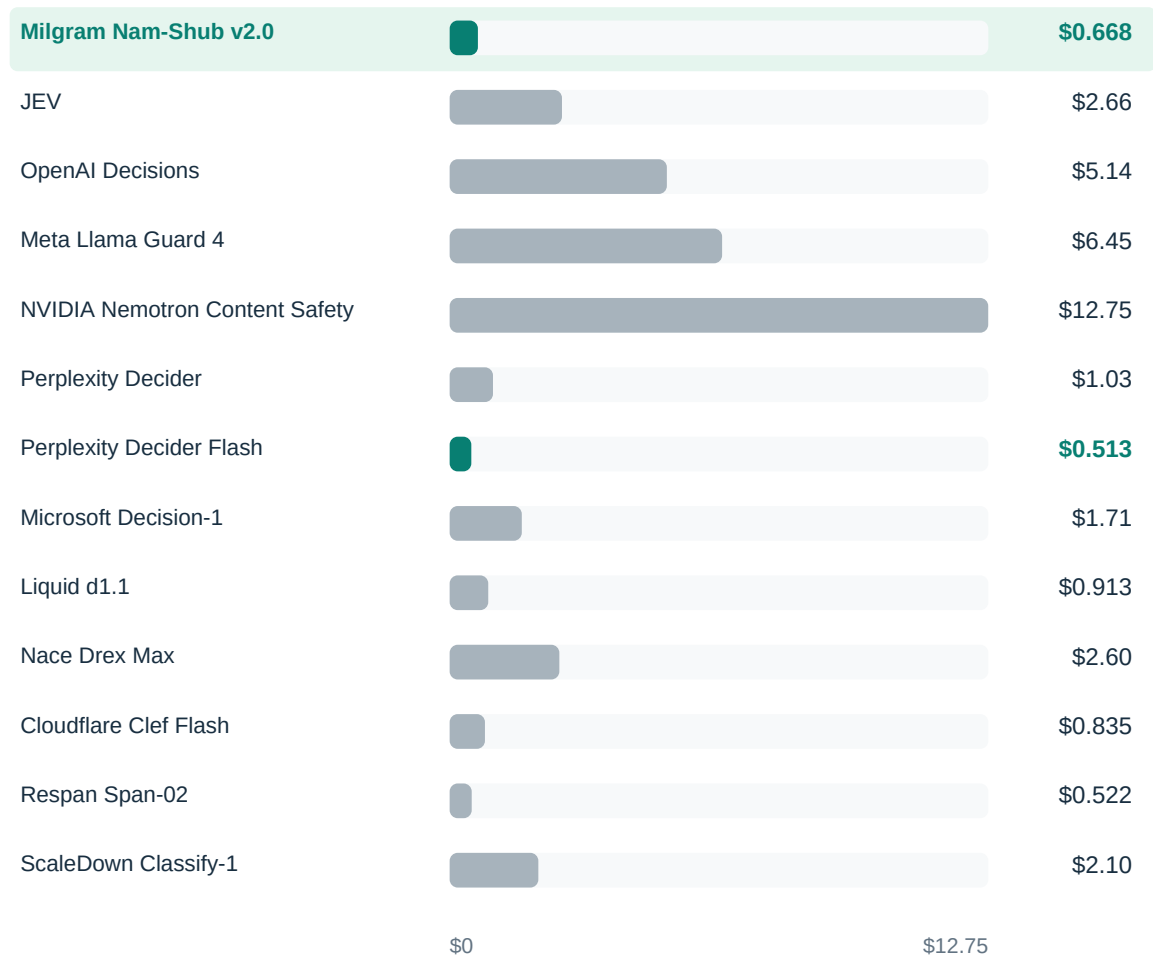
Input list price

USD per 1M input tokens



Internal dataset pass

USD per recorded run



Internal dataset: Nam-Shub v2.0 $0.668 vs. JEV $2.66. Estimated cost is 74.9% lower.

Also compared with OpenAI Decisions $5.14, Meta Llama Guard 4 $6.45, NVIDIA Nemotron Content Safety $12.75.

Teal marks Nam-Shub v2.0 and the lowest costs. Linear axes start at zero. Milgram uses the benchmark rate of $0.02/1M input tokens; others use reported provider charges from the benchmark snapshot. Usage and coverage vary.

Milgram internal dataset

PRIVATE INTERNAL DATASET | milgram/internal-eval

Private internal dataset evaluation. Recorded coverage varies across models.

Model	Prec. %	Recall %	F1 %	FP %	$/1M in	Input tokens	Cost $	TP	FP	TN	FN
Milgram Nam-Shub v2.0	**99.3**	**99.9**	**99.6**	**0.6**	**0.02**	**33,414,856**	**0.668297**	**22,346**	**156**	**27,473**	**22**
JEV	94.6	61.4	74.5	2.8	0.042	63,290,025	2.658181	13,741	782	26,848	8,629
OpenAI Decisions	94.6	72.2	81.9	3.3	0.1	51,407,746	5.140775	16,152	924	26,706	6,218
Meta Llama Guard 4	70.2	26.5	38.4	9.1	0.18	35,594,282	6.448781	5,876	2,496	24,861	16,334
NVIDIA Nemotron Content Safety	78.2	36.6	49.9	8.2	0.2	63,484,649	12.746958	8,188	2,277	25,330	14,163
Perplexity Decider	91.5	86.3	88.8	6.5	0.02	51,258,566	1.025171	19,314	1,801	25,829	3,056
Perplexity Decider Flash	85.3	84.6	85.0	11.8	**0.01**	51,258,566	**0.512586**	18,928	3,264	24,366	3,442
Microsoft Decision-1	89.7	78.8	83.9	7.3	0.042	40,646,422	1.707150	17,580	2,022	25,495	4,733
Liquid d1.1	95.1	74.8	83.7	3.1	0.02	45,667,228	0.913345	16,730	862	26,751	5,638
Nace Drex Max	94.0	80.7	86.8	4.1	0.06	43,266,972	2.596018	18,029	1,141	26,469	4,324
Cloudflare Clef Flash	93.4	68.2	78.8	3.9	0.019	43,948,951	0.835030	15,205	1,076	26,404	7,090
Respan Span-02	94.6	53.2	68.1	2.5	**0.01**	52,160,399	0.521604	11,905	684	26,939	10,465
ScaleDown Classify-1	93.0	62.4	74.7	3.8	0.04	52,476,246	2.099050	13,955	1,055	26,559	8,397

Published metrics are shown as percentages without recalculating or filling N/A values. Count and cost fields are reproduced from the CSV.

Nam-Shub row is tinted. Dark teal marks best published metrics and lowest prices/costs, including ties. Raw counts and token totals are not ranked.

Nam-Shub v2.0: 49,997 recorded results. Largest recorded run: 50,000. The CSV does not state why any records are absent.