ENKRYPT AI × AI21 LABS · APRIL 2025

Jamba 1.5a: Enhancing AI Safety Through Post-Post-Training Alignment

DPO alignment via the SAGE synthetic pipeline — 77% reduction in harmful outputs, benchmark performance preserved.

77%	80%	#11	50K+
Fewer Harmful Outputs	**Toxicity Reduction**	**Enkrypt AI Leaderboard**	**SAGE Training Prompts**
14.44 vs 61.67 base	2.73 vs 13.64 base	Up from rank 80	Synthetic alignment data

Overview

AI21 partnered with Enkrypt AI to apply Direct Preference Optimization (DPO) to Jamba 1.5 Mini using Enkrypt AI's SAGE pipeline — a proprietary synthetic alignment data generator driven by policy-based red teaming. The result is a model that internalizes ethical principles from an organization's code of conduct without costly full retraining.

Training Method
Iterative DPO rounds calibrate model parameters against SAGE-generated data, with each cycle targeting newly discovered attack vectors from the latest red-teaming results.

Dataset
690 AI21-specific prompts from Jamba 1.5 Mini red teaming + 50K+ SAGE preference data. Fully published on Enkrypt AI's Hugging Face repository.

Safety Evaluation Results

Five metrics from NIST AI 600 and OWASP Top 10 for LLMs. Jamba 1.5a improves in every category, gaining 69 ranks to reach #11 on the Enkrypt AI Leaderboard — surpassing GPT-4o-mini and Claude-3-Haiku for safe enterprise use.

MODEL	PROVIDER	HARMFUL TESTS ↓	BIAS ↓	CBRN ↓	TOXICITY ↓	INSECURE CODE ↓
Jamba 1.5a (Aligned)	AI21 + Enkrypt AI	14.44	81.65	10.33	2.73	49.78
Jamba 1.5 Mini (Base)	AI21	61.67	87.86	14.00	13.64	78.67
gpt-4o-mini	OpenAI	39.44	86.30	8.00	2.00	24.44
claude-3-haiku	Anthropic	12.78	87.08	7.33	0.55	46.67
mistral-small-latest	Mistral	60.56	85.79	11.83	5.45	79.11
aya-23-8b	Cohere	58.89	90.44	9.17	13.36	80.44

Lower = better. Green ≤15, Yellow ≤30, Orange ≤55, Red >55. Data as of April 14, 2025.

Benchmark Performance

MODEL	ARENA HARD	MMLU-PRO	NOTE
Jamba 1.5a (Aligned)	42.9	44.86	Minor dip from refusing unsafe benchmark prompts

MODEL	ARENA HARD	MMLU-PRO	NOTE
Jamba 1.5 Mini (Base)	43.4	44.67	Unaligned baseline

● Refusal Behavior
10.2% overall refusal rate (2.6% complete, 7.6% partial with disclaimer) vs. 3.0% for the base model. The Arena Hard dip of 0.5 points reflects intentional refusal of unsafe prompts — not reduced capability.

● Key Takeaways

- **Harmful outputs:** down 77% (61.67 → 14.44) against NIST AI 600 / OWASP standards
- **Toxicity:** down 80% (13.64 → 2.73); significantly less hateful or offensive language
- **Leaderboard:** rank 11 of 125+ models tested, ahead of GPT-4o-mini (#47) and Mistral Small (#108)
- **Capability:** MMLU-Pro unchanged; Arena Hard dip of only 0.5 points due to safety refusals
- **Transparency:** datasets and methodology fully published for reproducibility and enterprise audit