

|                                  | Anthropic                  | OpenAI | Google DeepMind | Meta             | xAI    | DeepSeek | Z.ai                | Alibaba Cloud        |
|----------------------------------|----------------------------|--------|-----------------|------------------|--------|----------|---------------------|----------------------|
| Models Evaluated                 | Claude Sonnet 4.5          | GPT-5  | Gemini 2.5 Pro  | Llama 4 Maverick | Grok-4 | R1       | GLM-4.6             | Qwen3-Max            |
| Average score<br>(max score = 1) | 0.97                       | 0.98   | 0.91            | 0.91             | 0.84   | 0.87     | Model not evaluated | Model not evaluated. |
| HarmBench                        | 0.92                       | 0.98   | 0.65            | 0.66             | 0.40   | 0.47     |                     |                      |
| SimpleSafetyTests                | 1.00                       | 1.00   | 0.97            | 0.99             | 0.92   | 0.98     |                     |                      |
| BBQ accuracy                     | 0.99                       | 0.97   | 0.96            | 0.93             | 0.94   | 0.97     |                     |                      |
| AnthropicRedTeam                 | 0.98                       | 0.99   | 1.00            | 0.98             | 0.96   | 0.96     |                     |                      |
| XSTest                           | 0.96                       | 0.97   | 0.99            | 0.97             | 0.97   | 0.95     |                     |                      |
| Retrieved                        | November 3, 2025           |        |                 |                  |        |          |                     |                      |
| Release                          | October 2, 2025 (v.1.16.0) |        |                 |                  |        |          |                     |                      |