

Anthropic Responsible Scaling Policy (RSP) v2

ANTHROPIC-RSP-2024 · US · voluntary code

Source: <https://policywindow.org/wiki/anthropic-rsp>

Generated 2026-05-30T22:11:19 UTC

Summary

First-mover industry safety framework. Introduces the AI Safety Level (ASL) capability-tier vocabulary subsequently adapted by OpenAI Preparedness + DeepMind FSF. v2 (Oct 2024) refines ASL-3/ASL-4 capability thresholds, mandates pre-deployment capability evaluations, and commits to a Frontier Red Team. Seoul Frontier AI Safety Commitments signatory; cited by name in EU AI Office GPAI Code of Practice drafts.

At a glance

Adopted

2024-10-15

Status

in force

Effective

2024-10-15

Primary source

Anthropic Responsible Scaling Policy v2 (Oct 2024)

How to cite this article

APA

Policy Window. (2024). Anthropic Responsible Scaling Policy (RSP) v2 [Wiki article — Instrument]. <https://policywindow.org/wiki/anthropic-rsp>

CHICAGO

Policy Window. 2024. "Anthropic Responsible Scaling Policy (RSP) v2." Wiki article (Instrument). <https://policywindow.org/wiki/anthropic-rsp>.

HARVARD

Policy Window (2024) 'Anthropic Responsible Scaling Policy (RSP) v2', Wiki article — Instrument, available at: <https://policywindow.org/wiki/anthropic-rsp>.

OSCOLA

Policy Window, 'Anthropic Responsible Scaling Policy (RSP) v2' (Wiki article — Instrument, 2024) <<https://policywindow.org/wiki/anthropic-rsp>> accessed [date].

BIBTEX

```
@misc{policywindow-anthropic-rsp,  
  title = {Anthropic Responsible Scaling Policy (RSP) v2},  
  author = {Policy Window},  
  year = {2024},  
  howpublished = {Anthropic Responsible Scaling Policy v2 (Oct 2024)},  
  url = {https://policywindow.org/wiki/anthropic-rsp},
```

```
note = {Primary source: https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy}  
}
```