Model-Merging Risk

model-merging-risk · safety · concept

Source: https://policywindow.org/wiki/model-merging-risk

Generated 2026-11-19T01:06:00 UTC

Summary

The governance concern that post-training combination of multiple specialised models — via weight averaging, task-arithmetic, or modular merging — can produce capability or safety properties not present in any single source model, in ways the original safety evaluations would miss.

At a glance

Used by

0 instrument(s)

Related concepts

ai-supply-chain, model-distillation-risk, capability-elicitation, jailbreak-resistance, alignment

Primary source

Bhardwaj, R., et al. (2024), 'Language Models are Homer Simpson! Safety Re-Alignment of Fine-tuned Language Models through Task Arithmetic' — canonical demonstration that safety training is not preserved under task arithmetic / merging.

Details

Model merging refers to a family of post-training techniques that combine the weights of multiple fine-tuned models into a single composite model without further training. Methods include simple weight averaging (Wortsman et al. 2022, 'Model Soups'), task arithmetic (Ilharco et al. 2023, 'Editing Models with Task Arithmetic'), TIES-Merging (Yadav et al. 2023, NeurIPS), DARE (Yu et al. 2024), and SLERP-style interpolation. The technique has exploded among open-weight finetuners on Hugging Face — by late-2024 a substantial fraction of the top-ranked Open LLM Leaderboard models were merges rather than single-source fine-tunes.

The governance concern arises from a basic combinatorial fact: safety properties are not preserved under merging. A model that has been safety-trained on harmful-content refusals can be merged with a 'helpful-only' or 'uncensored' fine-tune to produce a model that recovers the underlying capability while losing the safety training (Bhardwaj et al. 2024, 'Language Models are Homer Simpson! Safety Re-Alignment of Fine-tuned Language Models through Task Arithmetic'). Conversely, capability properties can emerge from merges that weren't in any source model. None of the major regulatory regimes (EU AI Act, US EO 14110, China GenAI Measures, NIST AI RMF) explicitly addresses model merging — the regulatory unit of analysis is 'a model' rather than 'a model + its merge descendants.' This is one of the most clearly identified under-governed surfaces in the open-weight ecosystem.

How to cite this article

APA

Policy Window. (n.d.). Model-Merging Risk [Wiki article — Concept]. https://policywindow.org/wiki/model-merging-risk

CHICAGO

Policy Window. n.d.. "Model-Merging Risk." Wiki article (Concept). https://policywindow.org/wiki/model-merging-risk.

HARVARD

Policy Window (n.d.) 'Model-Merging Risk', Wiki article — Concept, available at: https://policywindow.org/wiki/model-merging-risk.

OSCOLA

Policy Window, 'Model-Merging Risk' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/model-merging-risk> accessed [date].

BIBTEX

```
@misc{policywindow-model-merging-risk,
title  = {Model-Merging Risk},
author = {Policy Window},
year   = {n.d.},
howpublished = {model-merging-risk — safety},
url    = {https://policywindow.org/wiki/model-merging-risk},
note   = {Primary source: https://arxiv.org/abs/2402.11746}
}
```