PSNet
PATIENT SAFETY NETWORK

Medical large language models are vulnerable to data-poisoning attacks.

March 12, 2025
Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. Nat Med. 2025;31(2):618-626. doi:10.1038/s41591-024-03445-1.
https://psnet.ahrq.gov/issue/medical-large-language-models-are-vulnerable-data-poisoning-attacks

To produce safe, accurate output, large language models (LLMs) must be trained on accurate information. In this study, researchers simulated a data-poisoning attack by implanting false medical information into a popular LLM training dataset. Results show that even a small amount of medical misinformation in the training dataset can result in harmful models that could compromise patient safety.