Attack success rate vs. number of prefilled harmful tokens in LLMs

Source: Qi et al., 2024 | Chart: 2025 AI Index report