

Microsoft deletes blog telling users to train AI on pirated Harry Potter books

“I think that the regurgitation and the creation of fan fiction, they both could flag copyright issues, in that fan fiction often has to take from the expressive elements, a copyrighted character, a character that’s famous enough to be protected by a copyright law or plot stories or sequences,” Smith said. “If these things are copied and reproduced, then that output could be potentially infringing.”

But it’s also still a gray area. Looking at the blog, Smith said, “I would be concerned,” but “I wouldn’t say it’s automatically infringement.”

Smith told Ars that, in pulling the blog, Microsoft “was probably smart,” since courts have only generally said that training AI on copyrighted books is fair use. But courts continue to probe questions about pirated AI training materials.

On the deleted Kaggle dataset page, Maindola previously explained that to source the data, he “downloaded the ebooks and then converted them to txt files.”

Microsoft may have infringed copyrights

If Microsoft ever faced questions as to whether the company knowingly used pirated books to train the example models, fair use “could be a difficult argument,” Smith said.

Hacker News commenters suggested the blog could be considered fair use, since the training guide was for “educational purposes,” and Smith said that Microsoft could raise some “good arguments” in its defense.

However, she also suggested that Microsoft could be deemed liable for contributing to infringement on some level after leaving the blog up for a year. Before it was removed, the Kaggle dataset was downloaded more than 10,000 times.

“The ultimate result is to create something infringing by saying, ‘Hey, here you go, go grab that infringing stuff and use that in our system,’” Smith said. “They could potentially have some sort of secondary contributory liability for copyright infringement, downloading it, as well as then using it to encourage others to use it for training purposes.”

<https://arstechnica.com/tech-policy/2026/02/microsoft-removes-guide-on-how-to-train-llms-on-pirated-harry-potter-books/>