AI firms follow DeepSeek’s lead, create cheaper models with “distillation”



Thanks to distillation, developers and businesses can access these models’ capabilities at a fraction of the price, allowing app developers to run AI models quickly on devices such as laptops and smartphones.

Developers can use OpenAI’s platform for distillation, learning from the large language models that underpin products like ChatGPT. OpenAI’s largest backer, Microsoft, used GPT-4 to distill its small language family of models Phi as part of a commercial partnership after investing nearly $14 billion into the company.

However, the San Francisco-based start-up has said it believes DeepSeek distilled OpenAI’s models to train its competitor, a move that would be against its terms of service. DeepSeek has not commented on the claims.

While distillation can be used to create high-performing models, experts add they are more limited.

“Distillation presents an interesting trade-off; if you make the models smaller, you inevitably reduce their capability,” said Ahmed Awadallah of Microsoft Research, who said a distilled model can be designed to be very good at summarising emails, for example, “but it really would not be good at anything else.”

David Cox, vice-president for AI models at IBM Research, said most businesses do not need a massive model to run their products, and distilled ones are powerful enough for purposes such as customer service chatbots or running on smaller devices like phones.

“Any time you can [make it less expensive] and it gives you the right performance you want, there is very little reason not to do it,” he added.

That presents a challenge to many of the business models of leading AI firms. Even if developers use distilled models from companies like OpenAI, they cost far less to run, are less expensive to create, and, therefore, generate less revenue. Model-makers like OpenAI often charge less for the use of distilled models as they require less computational load.

https://arstechnica.com/ai/2025/03/ai-firms-follow-deepseeks-lead-create-cheaper-models-with-distillation/