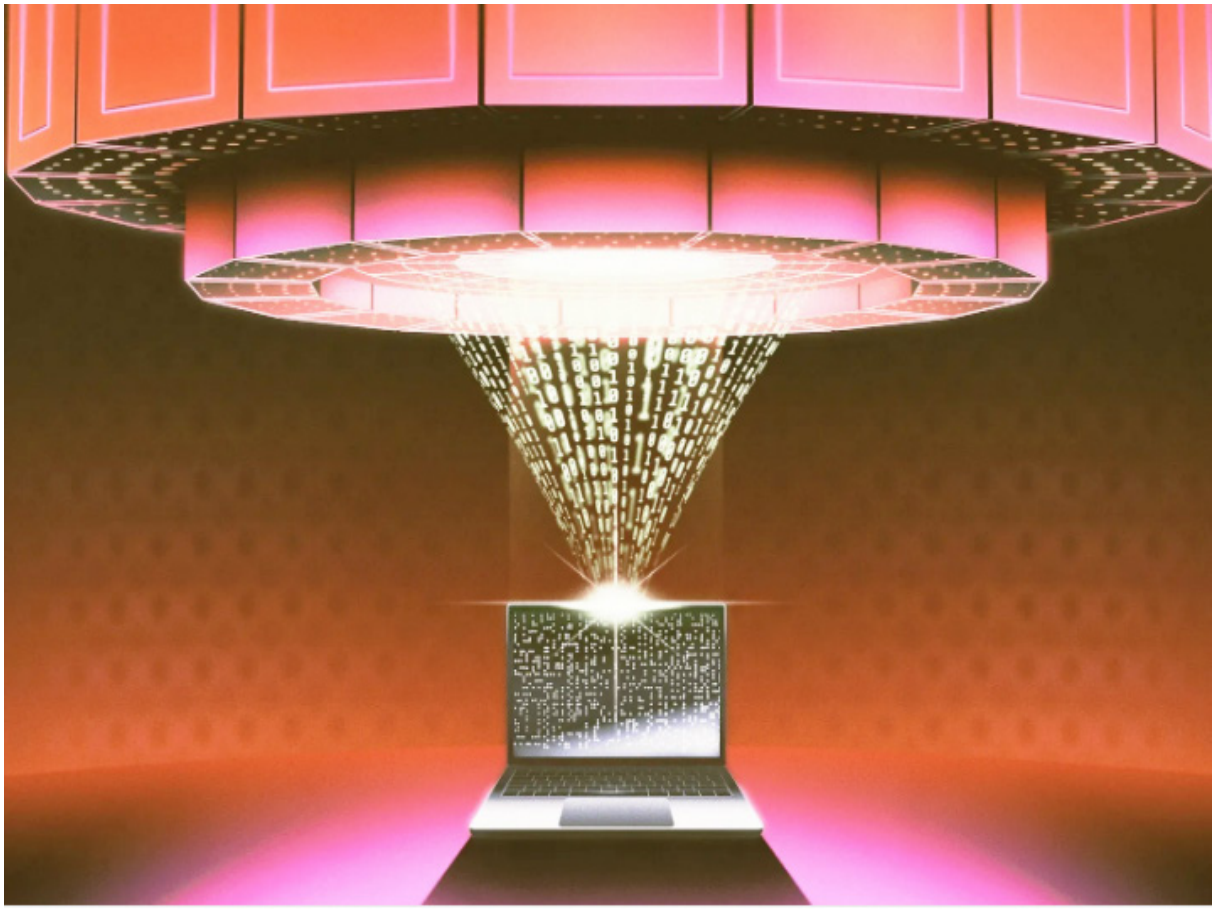
Distillation Can Make AI Models Smaller and Cheaper



THE ORIGINAL VERSION of this story appeared in Quanta Magazine.

The Chinese AI company DeepSeek released a chatbot earlier this year called R1, which drew a huge amount of attention. Most of it focused on the fact that a relatively small and unknown company said it had built a chatbot that rivaled the performance of those from the world's most famous AI companies, but using a fraction of the computer power and cost. As a result, the stocks of many Western tech companies plummeted; Nvidia, which sells the chips that run leading AI models, lost more stock value in a single day than any company in history.

Some of that attention involved an element of accusation. Sources alleged that DeepSeek had

obtained, without permission, knowledge from OpenAI's proprietary o1 model by using a technique known as distillation. Much of the news coverage framed this possibility as a shock to the AI industry, implying that DeepSeek had discovered a new, more efficient way to build AI.

But distillation, also called knowledge distillation, is a widely used tool in AI, a subject of computer science research going back a decade and a tool that big tech companies use on their own models. "Distillation is one of the most important tools that companies have today to make models more efficient," said Enric Boix-Adsera, a researcher who studies distillation at the University of Pennsylvania's Wharton School.

Dark Knowledge

The idea for distillation began with a 2015 paper by three researchers at Google, including Geoffrey Hinton, the so-called godfather of AI and a 2024 Nobel laureate. At the time, researchers often ran ensembles of models–"many models glued together," said Oriol Vinyals, a principal scientist at Google DeepMind and one of the paper's authors–to improve their performance. "But it was incredibly cumbersome and expensive to run all the models in parallel," Vinyals said. "We were intrigued with the idea of distilling that onto a single model."

The researchers thought they might make progress by addressing a notable weak point in machine-learning algorithms: Wrong answers were all considered equally bad, regardless of how wrong they might be. In an image-classification model, for instance, "confusing a dog with a fox was penalized the same way as confusing a dog with a pizza," Vinyals said. The researchers suspected that the ensemble models did contain information about which wrong answers were less bad than others. Perhaps a smaller "student" model could use the information from the large "teacher" model to more quickly grasp the categories it was supposed to sort pictures into. Hinton called this "dark knowledge," invoking an analogy with cosmological dark matter.

After discussing this possibility with Hinton, Vinyals developed a way to get the large teacher model to pass more information about the image categories to a smaller student model. The key was homing in on "soft targets" in the teacher model–where it assigns probabilities to each possibility, rather than firm this-or-that answers. One model, for example, calculated that there was a 30 percent chance that an image showed a dog, 20 percent that it showed a cat, 5 percent that it showed a cow, and 0.5 percent that it showed a car. By using these probabilities, the teacher model effectively revealed to the

student that dogs are quite similar to cats, not so different from cows, and quite distinct from cars. The researchers found that this information would help the student learn how to identify images of dogs, cats, cows, and cars more efficiently. A big, complicated model could be reduced to a leaner one with barely any loss of accuracy.

Explosive Growth

The idea was not an immediate hit. The paper was rejected from a conference, and Vinyals, discouraged, turned to other topics. But distillation arrived at an important moment. Around this time, engineers were discovering that the more training data they fed into neural networks, the more effective those networks became. The size of models soon exploded, as did their capabilities, but the costs of running them climbed in step with their size.

Many researchers turned to distillation as a way to make smaller models. In 2018, for instance, Google researchers unveiled a powerful language model called BERT, which the company soon began using to help parse billions of web searches. But BERT was big and costly to run, so the next year, other developers distilled a smaller version sensibly named DistilBERT, which became widely used in business and research. Distillation gradually became ubiquitous, and it's now offered as a service by companies such as Google, OpenAI, and Amazon. The original distillation paper, still published only on the arxiv.org preprint server, has now been cited more than 25,000 times.

Considering that the distillation requires access to the innards of the teacher model, it's not possible for a third party to sneakily distill data from a closed-source model like OpenAI's o1, as DeepSeek was thought to have done. That said, a student model could still learn quite a bit from a teacher model just through prompting the teacher with certain questions and using the answers to train its own models–an almost Socratic approach to distillation.

Meanwhile, other researchers continue to find new applications. In January, the NovaSky lab at UC Berkeley showed that distillation works well for training chain-of-thought reasoning models, which use multistep "thinking" to better answer complicated questions. The lab says its fully open source Sky-T1 model cost less than $450 to train, and it achieved similar results to a much larger open source model. "We were genuinely surprised by how well distillation worked in this setting," said Dacheng Li, a Berkeley doctoral student and co-student lead of the NovaSky team. "Distillation is a fundamental technique in AI."