AI蒸留の仕組みと可能性 · Work 共有する

中国のAIがアメリカのAIを蒸留して学習していると報道されている。実際、アメリカのAIの能力を蒸留によって習得することが可能なのでしょうか？
また、蒸留とは具体的にどのような作業であり通常の方法で実施することが可能なのか教えてください。

作業時間：18s

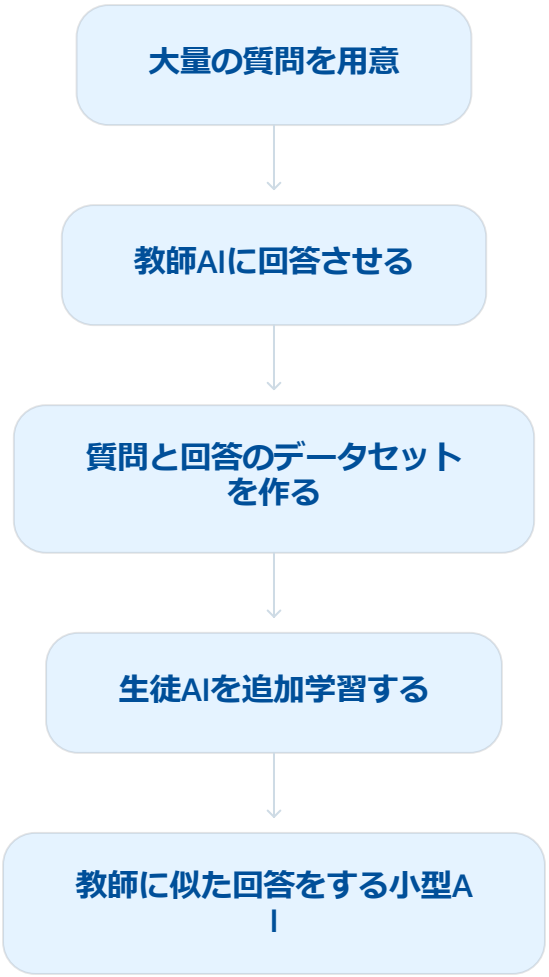
はい、可能です。ただし、蒸留によって「元のAIを丸ごとコピーできる」という意味ではありません。正確には、元のAIが示す回答パターンや問題解決能力の一部を、別のAIに学習させる技術です。

蒸留とは何か

AIの蒸留（Knowledge Distillation）では、次のように呼び分けます。

- 教師モデル：高性能だが、大規模で動作コストが高いAI
- 生徒モデル：教師の能力を学習する、より小型のAI

基本的な考え方は非常に単純です。



例えば、教師AIに次のような問題を数十万件与えます。

> 問題：ある商品の価格が20%引きで8,000円でした。元の価格はいくらですか。

教師AIの回答：

> 8,000÷0.8＝10,000なので、元の価格は10,000円です。

この「問題＋模範回答」を学習データにして、生徒モデルをファインチューニングします。大量に繰り返すことで、生徒モデルは教師モデルの解答形式や推論パターンをある程度再現できるようになります。

これは2015年にHintonらが体系化した、機械学習では以前から使われている一般的な技術です。Hintonらの知識蒸留論文 ↗

蒸留には複数の方法がある

報道で「蒸留」と呼ばれているものには、かなり幅があります。

方法	教師から取得する情報	教師内部へのアクセス
通常の教師・生徒型蒸留	各候補の確率分布（logits）	必要

方法	教師から取得する情報	教師内部へのアクセス
回答蒸留	質問に対する最終回答	不要
推論データ蒸留	解答、説明、途中の推論例	不要
選好蒸留	複数回答の順位・評価	不要
APIを使ったブラックボックス蒸留	APIから返される回答だけ	不要

古典的な蒸留では、「猫70%、犬20%、その他10%」のような教師モデルの確率分布まで利用します。これは通常、教師モデルを所有している企業が、自社の大型モデルから小型モデルを作る場合に使います。

一方、他社の非公開AIでは内部の確率分布を取得できません。そのため、報道で問題になるのは主に、教師AIへ大量の質問を送り、その回答を合成学習データとして利用する「ブラックボックス蒸留」です。

アメリカのAIの能力を習得できるのか

結論を分けると次のようになります。

特定分野の能力は、かなり習得できる

例えば、次のような対象です。

- 数学問題の解き方
- プログラム生成
- 質問応答の形式
- 文章の要約
- 指示への従い方
- 特定の推論パターン
- 回答の評価・選択方法

質問を特定分野に集中させれば、小型モデルがその分野で教師に近い性能を示す可能性があります。場合によっては、限定された評価試験において教師モデルと同等、あるいは教師より良い成績になることもあります。

DeepSeek自身も、DeepSeek-R1が生成したデータを使ってQwenやLlama系の小型モデルを学習させた「蒸留モデル」を公開しています。そのため、「高度な推論結果を小型モデルへ移すことができる」こと自体は公開された技術的事実です。DeepSeek-R1公式リポジトリ↗、DeepSeek-R1論文↗

元のAI全体を完全にはコピーできない

APIの回答だけでは、次のものは取得できません。

- 元モデルのパラメータ
- ニューラルネットワークの内部構造
- 元の学習データ
- 学習アルゴリズムの詳細
- あらゆる入力に対する挙動
- 非公開の安全対策
- 画像、音声、ツール操作など回答で観測できない能力

したがって、蒸留モデルは原本の複製ではなく、「観察した範囲で教師の振る舞いをまねるモデル」です。

学校に例えるなら、優秀な教師が作成した大量の模範解答を生徒に覚えさせることはできますが、その教師が何十年もかけて獲得した知識や、未知の問題へ対応する能力まで完全に移せるわけではない、という関係です。

なぜ蒸留で高い性能が得られるのか

高性能AIの回答には、単なる正解以上の情報が含まれているためです。

例えば教師AIが生成した回答には、

- どの情報を重要視するか
- 問題をどのように分解するか
- どの順番で説明するか
- 間違えやすい点をどう回避するか
- 最終回答をどの形式で提示するか

といった知識が表れます。

生徒モデルは大量の例からこれらの規則性を学習します。通常のインターネット文章だけを学習するよりも、「優れたAIが問題に答えた結果」を学習する方が、目的に直結した高品質なデータになる場合があります。

通常の方法で実施できるのか

技術的には、一般に利用できる仕組みで実施できます。大まかな工程は次のとおりです。

1. 教える分野を決める
2. 数千〜数十万件以上の質問を準備する
3. 教師モデルに回答を生成させる
4. 誤答、重複、低品質な回答を除去する
5. 質問と回答をJSONLなどに整形する
6. オープンウェイトモデルを教師ありファインチューニングする

7. 必要に応じて選好学習や強化学習を行う
8. 教師と生徒を同じ評価問題で比較する

生徒側には、Llama、Qwen、Mistralなど、追加学習可能なモデルを使用できます。学習には一般的にPyTorch、Transformers、TRL、LLaMA-Factoryなどが使われます。

ただし、本格的な蒸留には以下が必要です。

- 大量の教師モデルへの問い合わせ費用
- GPU
- 高品質な質問セット
- 回答データの検査・選別
- ファインチューニングの技術
- 学習後の評価と安全性試験

小規模な実験なら一般的なPCやクラウドGPUでも可能ですが、最先端モデルに近い汎用能力を得るには、依然として相当なデータ、計算資源、技術力が必要です。

中国AIをめぐる報道で注意すべき点

ここでは、次の2つを区別する必要があります。

1. DeepSeekが自社のR1出力を使って小型モデルを蒸留したこと
   → DeepSeek自身が論文とリポジトリで公表しています。
2. DeepSeekなどがOpenAIの非公開モデルを無断で蒸留したという主張
   → OpenAIは、そのような活動を示す利用パターンを確認したと主張していますが、外部からすべての詳細を独立検証できる状態ではありません。

OpenAIは2026年2月の米議会向け資料で、DeepSeek関係者に関連するとするアカウントや第三者経路から、蒸留目的と整合するアクセスを観測したと説明しています。一方、これはOpenAI側が公表した評価であり、裁判などで確定した事実とは分けて扱う必要があります。OpenAIの米議会向け資料↗

技術的に可能でも、利用条件上可能とは限らない

自社モデルを自社の小型モデルへ蒸留することや、蒸留を許可したオープンモデルを利用することは一般的です。

しかし、他社の商用AIについては、

- 出力を競合モデルの学習に使用すること
- アクセス制限を回避して大量取得すること
- 複数アカウントで利用上限を迂回すること
- 出力や内部推論を自動収集すること

などが、サービス利用規約で禁止されている場合があります。技術的に通常のAPIで実行

できたとしても、契約上・法的に許されるとは限りません。

要するに、蒸留は「相手AIの脳をコピーする技術」ではなく、「相手AIを大量に先生として利用し、その模範回答

はい、技術的には可能です。ただし、蒸留によって「相手のAIを丸ごと複製する」ことはできません。より正確には、優秀なAIの回答を大量の教材として使い、その回答傾向や特定分野の能力を別のAIに模倣させる手法です。

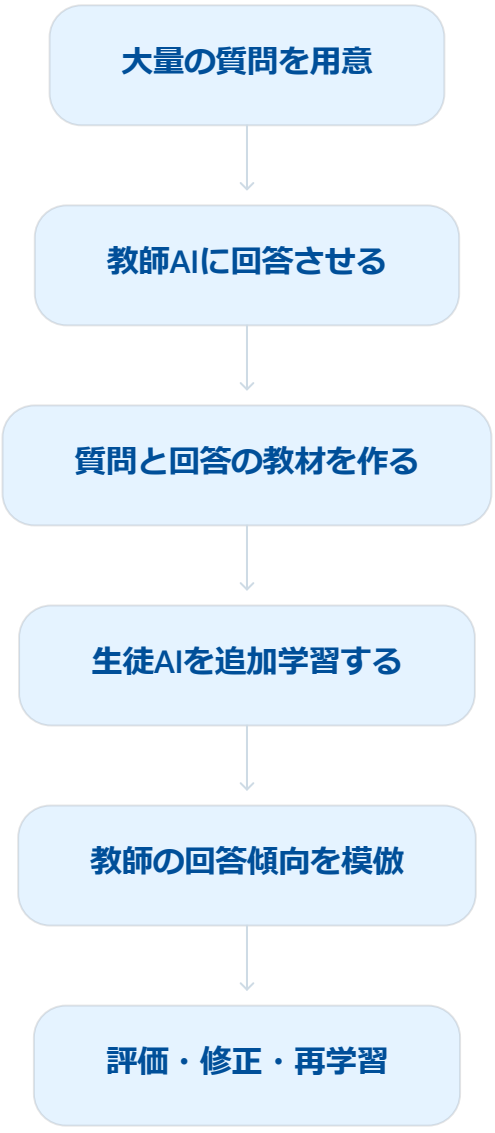
また、報道されている「中国企業が米国AIを蒸留した」という話には、公開済みの事実と企業側の主張・疑惑が混在しています。ここは分けて考える必要があります。

蒸留とは何か

知識蒸留では、通常次の2つのモデルを使います。

- 教師モデル：高性能だが、大きくて計算コストが高いAI
- 生徒モデル：小型・低コストで動かしたいAI

基本構造は次のようになります。



例えば教師AIに次のような問題を何十万件も与えます。

> 問題：工場Aでは毎時120個の製品を作ります。8時間では何個作れますか。

教師AIの回答：

> 120 × 8 = 960なので、960個です。

これを次のような学習データにします。

```JSON
{
  "input": "工場Aでは毎時120個の製品を作ります。8時間では何個作れますか。",
  "output": "120 × 8 = 960なので、960個です。"
}
```

このような「質問＋模範回答」を大量に用意し、オープンウェイトモデルなどをファインチューニングします。

知識蒸留自体は2015年ごろから広く使われている一般的な機械学習技術です。元となった代表的研究はHintonらのDistilling the Knowledge in a Neural Network ↗です。

どの程度まで能力を習得できるか

結論として、特定の能力はかなりの程度まで習得できます。

特に蒸留しやすいものは次のような能力です。

- 数学問題の解き方
- プログラムコードの生成
- 指示に従った回答形式
- 要約や分類
- 特定分野の質問応答
- 問題を段階的に処理する回答パターン
- JSONなど一定形式での出力
- 教師モデルの文体や説明方法

一方、次のものを完全にコピーするのは困難です。

- 教師モデル内部のパラメータそのもの
- 学習に使われた元データ
- あらゆる未知の質問への対応能力
- 画像・音声・ツール操作を含む総合能力
- 長い文脈を扱う能力
- 教師モデル固有の内部推論機構

- 安全対策や幻覚抑制を含む細かな挙動

つまり蒸留は、教師の「頭の中」をコピーするのではなく、教師の大量の答案を見て、生徒に似た回答をできるよう練習させる作業です。

たとえるなら、優秀な先生の解答例を大量に読んだ生徒は、似た試験で高得点を取れるようになります。しかし、先生が持つすべての知識や、未経験の問題に対応する能力まで完全に取得したわけではありません。

実際に行われる蒸留の方法

蒸留には大きく分けて2種類あります。

1. 内部情報を利用する正式な知識蒸留

教師モデルと生徒モデルの両方を所有している場合に使える方法です。

教師モデルが各候補に与えた確率、いわゆるロジットや確率分布を利用します。例えば教師モデルが次の単語を予測するとき、

候補	教師モデルの確率
東京	0.65
大阪	0.20
京都	0.10
その他	0.05

という内部情報まで使って生徒を学習させます。

正解が「東京」であることだけでなく、「大阪も多少可能性がある」という教師の判断も伝えられるため、効率のよい蒸留ができます。

ただし、一般利用者がAPIやチャット画面からこの内部情報を取得することは通常できません。

2. 出力だけを利用するブラックボックス蒸留

教師モデルの内部にはアクセスせず、APIやチャットから得た回答だけを教材にする方法です。報道で問題になるのは主にこちらです。

一般的な工程は次の通りです。

1. 数学、コード、論理、文章作成などの質問を大量に作る

2. 教師AIに質問して回答を取得する
3. 不正確な回答や重複を除去する
4. 質問と回答を学習用データに変換する
   5.既存のオープンウェイトモデルを追加学習する
5. 教師AIや採点モデルを使って回答を評価する
6. 良い回答を選んで再学習する
7. ベンチマークで性能を確認する

教師モデルに単に回答させるだけでなく、

- 複数の回答を生成させる
- 正誤を判定させる
- 回答を順位付けさせる
- 誤った回答を修正させる
- 難易度別に問題を生成させる

といった処理を組み合わせることもあります。

DeepSeekについて確認されていること

DeepSeek自身は、DeepSeek-R1が生成したデータを使って、QwenやLlamaを基盤とする複数の小型モデルを蒸留したことを正式に公開しています。例えば7B、14B、32Bなどの小型モデルです。DeepSeek-R1公式リポジトリ↗およびDeepSeek-R1論文↗で確認できます。

一方、「DeepSeekがOpenAIなど米国企業のモデルを不正に蒸留した」という点については、OpenAIが利用状況に基づく疑いを公表しています。2026年2月のOpenAI文書では、DeepSeek関係者によるものとみられるアクセスや、第三者経由で出所を隠す手法など、蒸留目的と整合する活動を観測したと主張しています。OpenAIの米下院向け文書↗

ただし、この文書はOpenAI側の報告です。第三者によって訓練データ全体が検証され、「どの米国モデルから、どの程度の能力を取得したか」が完全に確定したわけではありません。

したがって、現状は次のように整理するのが妥当です。

内容	状況
DeepSeekが蒸留技術を使った	公式論文で確認済み
自社R1の出力から小型モデルを作った	確認済み
米国AIの出力を利用した可能性	OpenAIなどが指摘
米国AIをそのまま複製した	そのような証拠はない

内容	状況
DeepSeekの性能がすべて蒸留だけで得られた	確認されていない

通常の方法で実施できるか

技術的には、市販のGPU、オープンウェイトモデル、学習用ソフトウェアがあれば実施できます。小規模なものであれば、研究者や企業、場合によっては個人でも可能です。

よく使われる構成は次のようなものです。

- 教師：利用許可された商用APIまたは自社モデル
- 生徒：Llama、Qwen、Mistralなどの利用可能なモデル
- 学習方法：教師ありファインチューニング
- 軽量化：LoRA／QLoRA
- 学習基盤：PyTorch、Transformers、TRLなど
- GPU：モデルサイズに応じて1台から多数

ただし「技術的に可能」と「自由に実施してよい」は別です。

多くの商用AIサービスでは、その出力を使って競合モデルを開発することを利用規約で禁止または制限しています。また、大量のAPIアクセス、複数アカウントによる制限回避、回答の自動収集などは、契約違反になる可能性があります。著作権、営業秘密、不正競争、契約上の問題も国や状況によって異なります。

自社モデルの小型化、許諾された教師モデルの利用、オープンな蒸留用データの利用であれば、蒸留は通常かつ有用な技術です。

最終的な見方

中国のAIが米国AIの回答を教材として利用し、一部の能力を効率的に習得することは技術的に十分可能です。特に数学、コーディング、論理問題など、正解を評価しやすい能力は移しやすい傾向があります。

しかし、蒸留だけで最先端AIの全能力を丸ごと取得できるわけではありません。高性能な基盤モデル、独自の学習データ、GPU計算資源、データ選別、強化学習、評価技術なども必要です。

したがって「米国AIをコピーした」という表現より、「米国AIの出力が、別モデルの能力を効率よく高める教材として使われた可能性がある」と捉えるのが技術的には最も正確です。