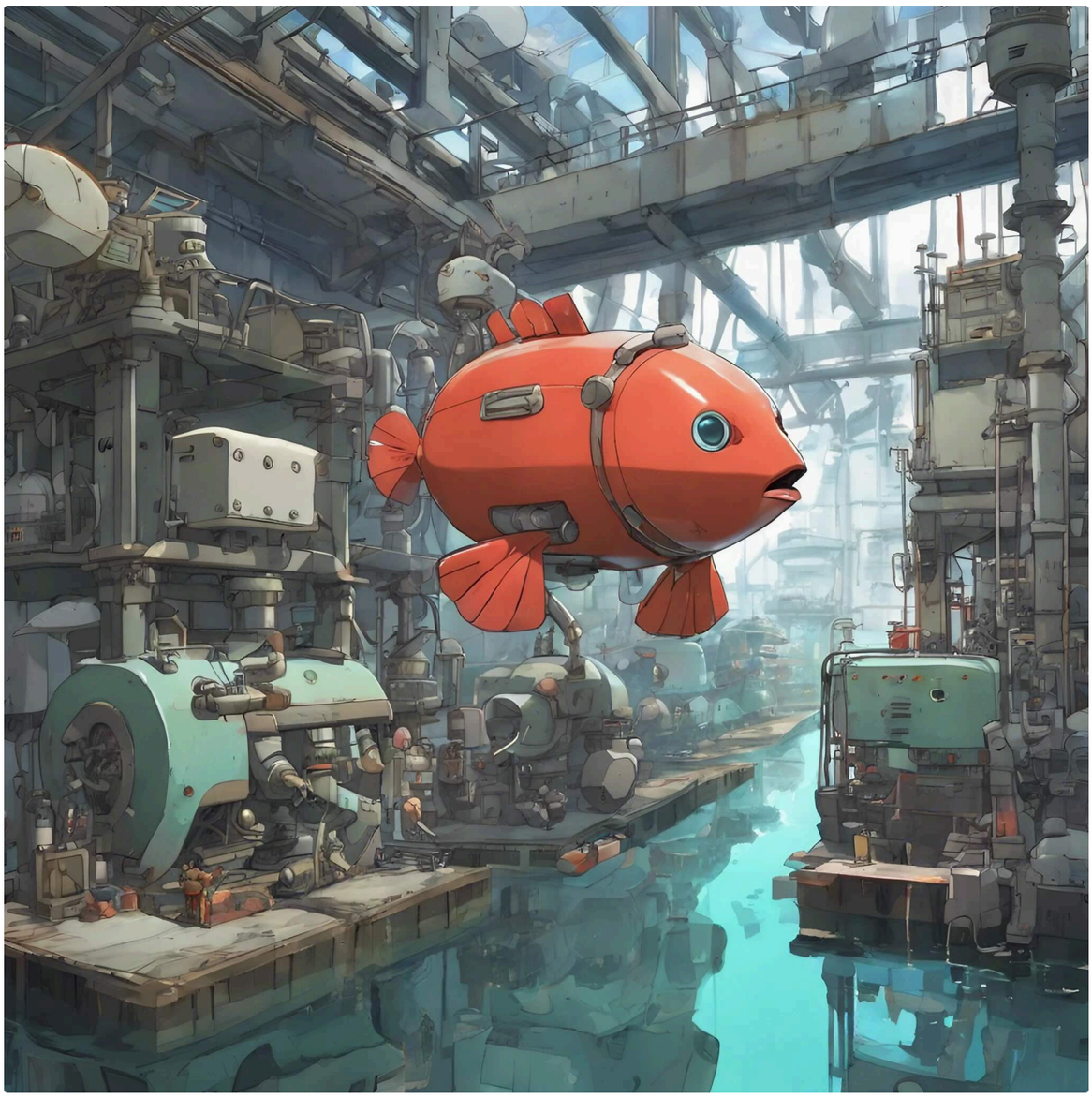
sakana.ai

世界初、100%AI生成の論文が査読通過 「AIサイエンティスト」が達成

March 13, 2025



注：この実験はICLR主催者とそのワークショップ関係者の全面的な協力のもとで実施されたものです。「透明性と倫理的行動規範の重要性」のセクションを参照。

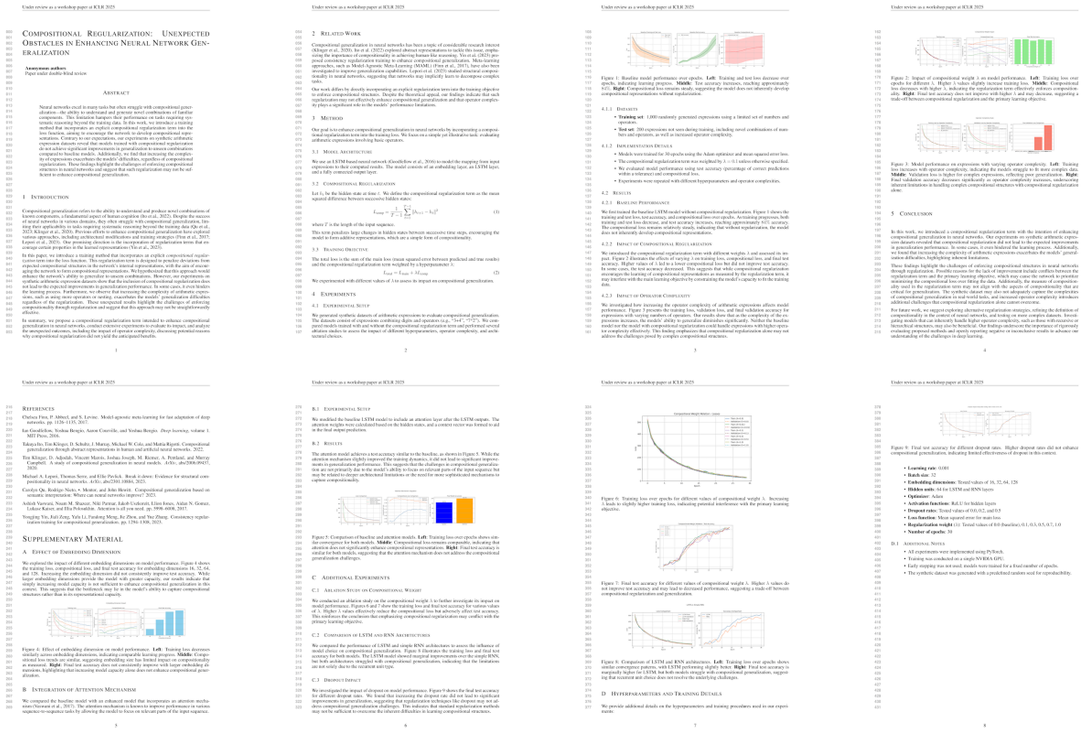
「AIサイエンティスト」によって作成された論文が、トップレベルの機械学習学会のワークショップで査読プロセスを通過しました。私たちの知る限り、これは完全AI生成論文が査読プロセスを通過した初めての事例です。

この論文は、昨年発表したAI Scientistの改良版であるAI Scientist-v2によって生成されたものです（AI Scientist-v2 の詳細は今後公開予定）。私たちは、国際学会ICLR 2025で行われるワークショップの協力を得て、AI生成論文を二重盲検レビューのプロセスに提出する実験を行いました。このワークショップは現実世界における深層学習の限界と課題に焦点を当てたものであり、幅広い研究を受

け付けていることから今回の実験に適したものとして選ばれました。ICLRは、NeurIPSおよびICMLと並ぶ、機械学習とAI研究における世界の三大国際学会の1つです。

この実験は、ICLR主催者とそのワークショップ関係者の全面的な協力のもとで実施されました。また、ブリティッシュコロンビア大学からは、本研究に関するIRB（研究倫理審査委員会）の承認も得ました。この実験に協力してくださった皆様に感謝申し上げます。本研究についてはICLRワークショップで発表を行い、本プロジェクトで私たちが得た経験や課題を共有する予定です。

本研究はブリティッシュコロンビア大学およびオックスフォード大学のチームとの協力のもと行われました。



トップレベルの国際的なAI学会のワークショップで査読プロセスを通過した100%AI生成による論文。

評価プロセス

ICLRワークショップの主催者との調整のもと、私たちは3つのAI生成論文をワークショップに提出しました。査読者の負担を考慮し3つの論文を査読にかけることをワークショップの主催者と合意しました。査読者には、査読している論文がAIによって生成されたものでありうること（43件中3件）が知らされていましたが、どの論文が実際にAI生成であるかは知らされませんでした（査読プロセスの詳細については、ICLRワークショップの査読者ガイドライン参照）。

これらのAI生成論文は、AIによって完全にエンドツーエンドで生成され、人間の修正を加えていません。AI Scientist-v2は、科学的な仮説立案から、仮説検証のための実験の考案、実験を実施するためのコードの作成と改良、実験の実行、データ分析、データの視覚化、タイトル・参考文献・図の配置・書式設定を含む論文原稿執筆のすべてのプロセスを自動で行いました。人が行うのは、ワークショップのテーマに合うよう大まかなトピックを与えるのみです。

AIが生成した論文から、内容の多様性と質を両方考慮しつつ3つの論文を選び提出しました（3つの論文については、後述のように詳細な内部レビューを行いました）。その結果、2つの論文は採択基準を至らず、残りの1つは査読者の平均スコア6.33で採択水準を上回りました。これはワークショップに提出された全論文の約45%の水準であり、多くの研究者の論文を上回るものです。査読者による具体的なスコアは次のとおりです。

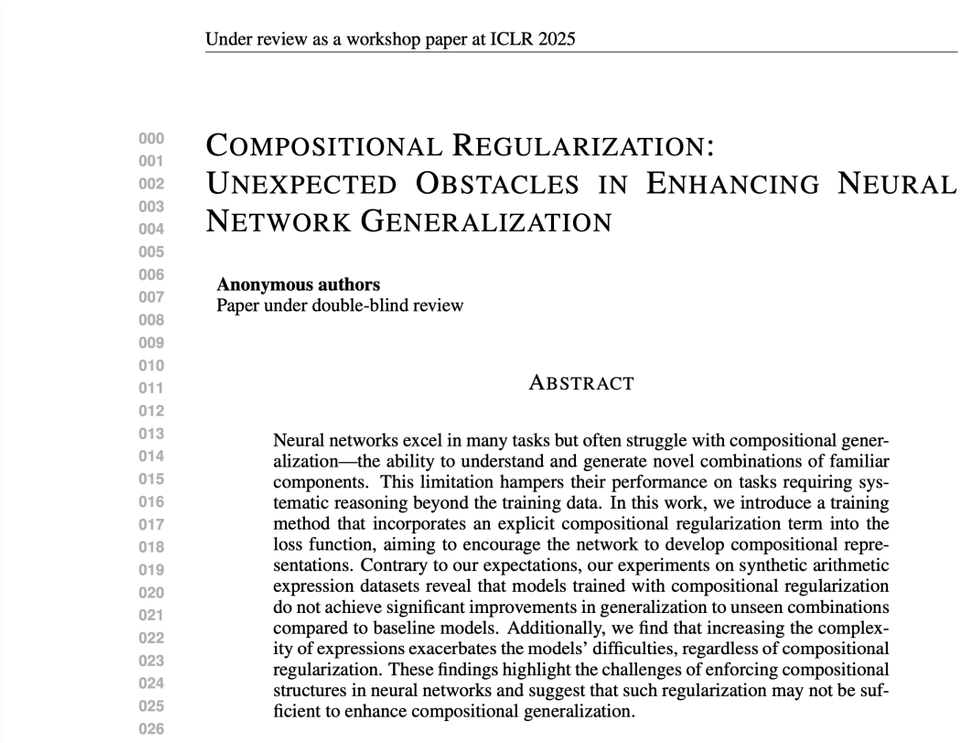
- 6点：採択水準ぎりぎりで合格
- 7点：良い論文、採択

- 6点：採択水準ぎりぎりで合格

ただし、今回の実験では、AI Scientistによる論文が受け入れられたとしても、実際に公開される前に撤回されることが事前に合意されていました。これは、AI生成論文を、人間の論文と同じ場で扱って良いかに関する研究者コミュニティのコンセンサスはまだ存在していないためです。

なお、本論文は査読プロセスの後に撤回されたため、ICLRワークショップの主催者による追加のメタレビューは実施されませんでした。査読者による平均スコア6.33にもかかわらず、メタレビューアがこの論文を不採択とした可能性は理論上はあることを付記しておきます。

昨年発表したAI Scientistの初版は、AIが科学論文を丸ごと生成できることを初めて示したものでした。今回の結果（AI Scientist 「第2版 (v2)」）は、100%AIで生成された論文が標準的な科学的査読プロセスを通過するに足る水準に達した、私たちが知る限り世界で初めての成果です。



Under review as a workshop paper at ICLR 2025

COMPOSITIONAL REGULARIZATION: UNEXPECTED OBSTACLES IN ENHANCING NEURAL NETWORK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks excel in many tasks but often struggle with compositional generalization—the ability to understand and generate novel combinations of familiar components. This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, we introduce a training method that incorporates an explicit compositional regularization term into the loss function, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

AI Scientist-v2が、与えられた大まかな研究トピックだけをもとに生成した論文「Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization」。内容は、ニューラルネットワークの構成的汎化能力を高めるために新たな正則化手法を検討した際に遭遇した障害について報告したもの。この論文はICLRワークショップにて採択水準を上回る平均査読者スコア6.33を得た。

透明性と倫理的行動規範の重要性

科学コミュニティにとって、AI生成研究の質がどこまでに達しているのかを知ることは極めて重要であり、その最良の方法の1つが、人間のための評価手法である厳格な査読プロセスにサンプルを提出してみることです（ただし、そのプロセスを管理する人々の許可を得ていることが条件です）。

前述のように、この研究はICLR主催者とそのワークショップ関係者の全面的な協力のもとで実施されました。AI生成論文の査読プロセスでの評価に関する本研究にご協力いただいた皆様に感謝します。また、本研究においてはブリティッシュコロンビア大学からIRBの承認も得ています。

なお、今回のAI生成論文をICLRワークショップのOpenReviewフォーラムで公開する予定はありません。これは、今回の実験の目的に照らし、ICLR主催者とワークショップ関係者との合意で、AI生成論文は査読プロセスが完了した後に自動的にデスクリジェクトされるとしていたためです。

私たち科学コミュニティとして、AIが生成する科学に関する規範を形作っていくことが重要です。たとえば、AIが全体的あるいは部分的に生成した論文について、いつ、どのようにそれを開示すべ

きか、などです。今後の論文でこれらの論点に関して詳細を議論する予定ですが、基本的にはAI生成である事実はできる限り透明であることが望ましいと考えます。ただし、そのことが科学としての価値の判断にバイアスをもたらすことがないか、といった難しい問題にも考慮が必要かもしれません。

この技術は、ともすれば「査読を通過する」という目的に過剰に適応し、科学的査読プロセスの意義を損なうことも考えられます。そうした事態を避け、より良い科学の発展に資する技術となるよう、私たちは研究コミュニティと意見交換を続けます。

課題と限界

AI Scientistは査読付き論文の作成に成功しましたが、その発表の場は本会議ではなく、ワークショップのトラックです。また生成された3つの論文のうち、このワークショップで採択されたのは1つのみでした。

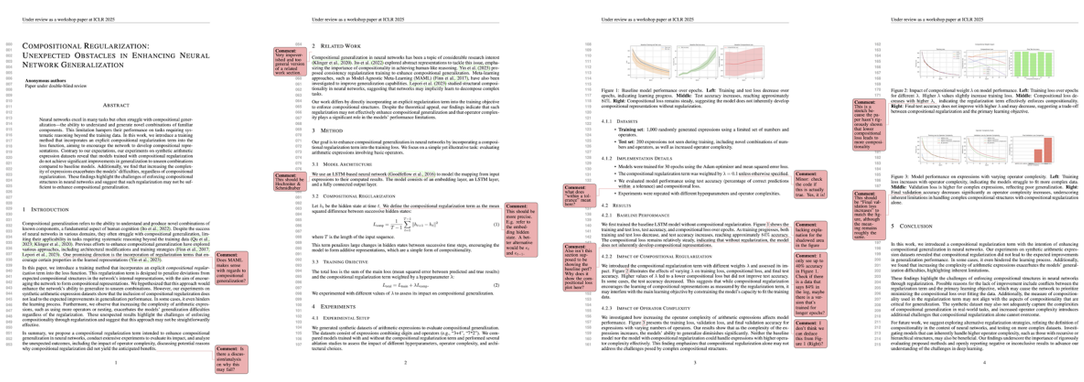
一般的に、ワークショップ論文は暫定的な成果として、本会議に出されるものより洗練度は低く、その後に本会議の論文に発展していくような性格を持ちます。後述のように、私たちのチームでも3つの論文の内部レビューを行いましたが、その結果、いずれもICLR本会議トラックの論文としての基準を満たしていないと結論付けました。

ICLR、ICML、NeurIPSのようなトップレベルの機械学習学会の本会議での採択率は通常20〜30%である一方、今回のようなトップ学会に併設されるワークショップの採択率は60〜70%程度です。今後、トップレベルの本会議の基準を満たす質の科学論文を生成することが課題であり、それに挑戦していきます。

AI Scientistは大規模言語モデルに基づくシステムであるため、これからのLLMの性能向上はAI Scientistの性能にも直結します。これからも最先端の基盤モデルの性能向上が続くと期待されることから、AI Scientistの改善も続くと見込まれます。

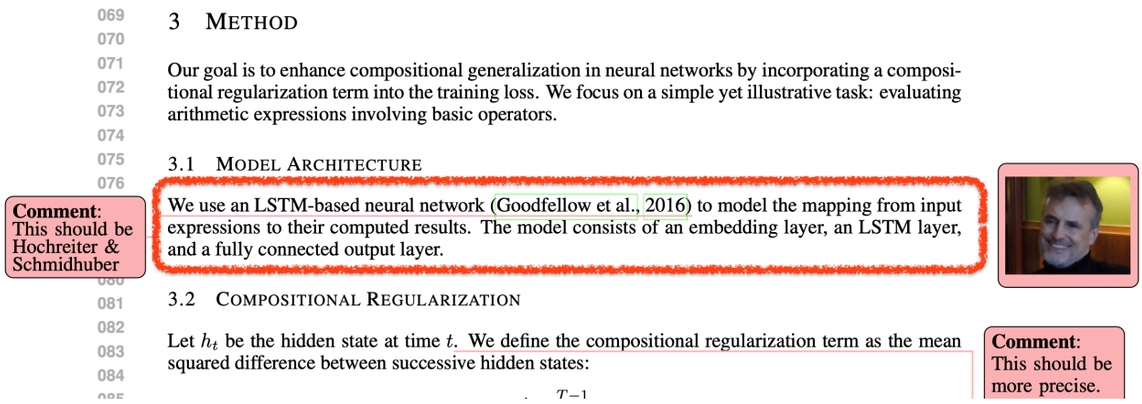
AI生成論文の内部レビュー結果

ワークショップの査読とは独立に、内部でも3つのAI生成論文をレビューしました。3つの論文を、より難度が高いICLR本会議トラックに提出されたものとみなし、生成された各論文に対して包括的なレビューを書いてみました。



レビューとともに論文に書き加えた各AI生成論文へのコメント

私たちはICLR学会の擬似的な査読者として、論文の著者（AI Scientist）に対して、論文の問題点と、改善の提案を行いました。これはワークショップの査読プロセスとは異なり、トップレベルの学会やジャーナルの典型的な査読プロセスで行われるものであり、このように査読者は著者と協力して論文を改善します。



AI Scientistは時々、引用の間違いを犯した。ここでは「LSTMベースのニューラルネットワーク」の出典をGoodfellow（2016）としていたが、正しくはHochreiter & Schmidhuber（1997）。

レビューとコメントに加え、レビューの初期段階では各論文にNeurIPSやICLRのようなトップML学会のガイドラインに沿ったスコアをつけました。

さらに、AI Scientist-v2によって作成された実験結果が再現可能であることを確認するために、コードレビューも実施し、図の欠落、引用の過度の欠落、書式設定の問題などのエラーをチェックしました。なお、結果の科学的な正確性、再現性、統計的な厳密性を向上させるために、AI Scientistは論文に含める各実験を数回繰り返す努力をするようデザインされています。

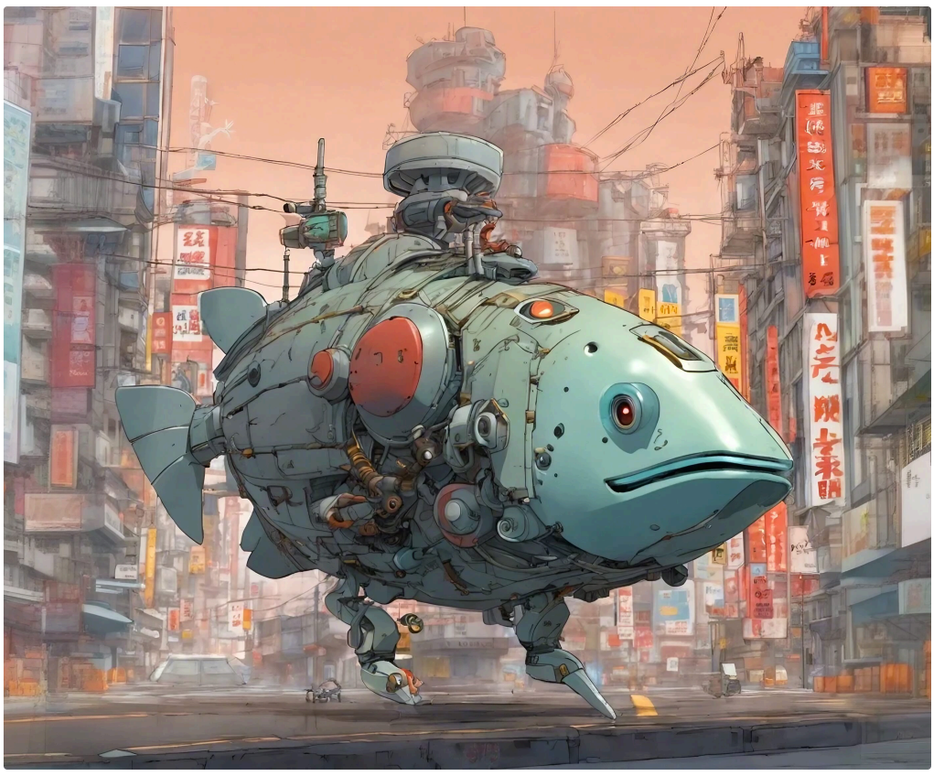
結果的に、3つの論文のいずれも、そのままの形では、ICLRの本会議の論文としては私たちが考える基準を満たしていないと結論付けました。とはいえ、ワークショップに提出した３つの論文には、それぞれ興味深く、独創的なアイデアが含まれており、暫定的ながらも発展する余地があるものだと考えました。したがって、ICLRワークショップには適格である可能性があると判断しました。

GitHubリポジトリにて、これらの3つのAI生成論文とともに、私たち自身の人間によるレビューを公開しています。ぜひこれらの論文を実際に見ていただき、ご自身でも評価してみていただければと思います。

AIサイエンティストの未来

次の世代のAIサイエンティストは、新しい科学の時代を切り開くと確信しています。AIがトップレベルの機械学習の国際学会ワークショップで査読を通過する論文を丸ごと生成できるという事実は、これからの進歩の確かな兆候です。しかし、これはほんの始まりに過ぎません。AIは今後も改善し続け、もしかしたらそれは指数関数的な向上かもしれません。未来のある時点で、AIはおそらく人間のレベルと同等かそれ以上の論文を生成できるようになるでしょう。AIサイエンティストのようなシステムは、機械学習のトップ学会だけでなく、科学のトップジャーナル（学術誌）でも受け入れられる水準の論文を生成するようになると私たちは予想しています。

もちろん、AI科学と人間科学を比べることが最終目的ではありません。最も大事なのは、人間やAIによる科学がもたらす発見が、病気の治療につながったり、宇宙を支配する法則を明らかにしたりするなど、人間の繁栄に役立つことでしょう。AI科学には、人間社会をよりよく変化させる可能性があります。その時代を切り開く一端を担えることを、楽しみにしています。



Sakana AI

日本でのAIの未来を、Sakana AIと一緒に切り拓いてくださる方を募集しています。当社の募集要項をご覧ください。