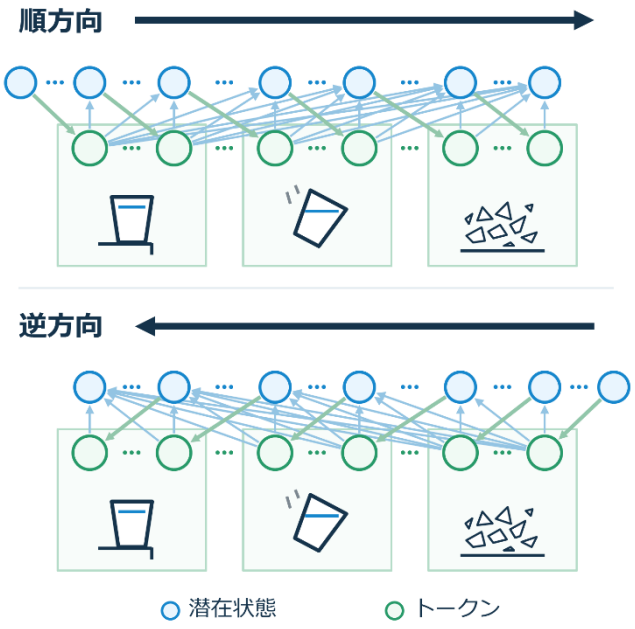
PRESS RELEASE

東京大学
科学技術振興機構（JST）

自己回帰生成モデルの熱力学を構築
――大規模言語モデルに内在する「世界モデル」の定量的理解に向けて――

発表のポイント

◆ 大規模言語モデルからカルマンフィルタまで、過去の出力履歴を用いて次の出力を生成する「自己回帰生成モデル」を統一的に扱う熱力学理論を構築しました。
◆ 出力を生成する過程にどの程度「一方向性」があるかを表す量（エントロピー生成）を効率的に推定する方法を示し、それを大規模言語モデルの解析に応用しました。
◆ 本研究は、生成 AI が学習で獲得した現実世界についての表現を、熱力学に基づいて解析する新たな道筋を提示することが期待されます。



自己回帰生成モデル（特に大規模言語モデル）の概念図

概要

東京大学大学院工学系研究科物理工学専攻の沙川 貴大 教授は、過去の出力の履歴に基づいて次の出力を生成する「自己回帰生成モデル」（Autoregressive Generative Model、注 1）の熱力学理論を構築しました（図 1）。大規模言語モデル（Large Language Model, LLM、注 2）の基盤であるトランスフォーマー（Transformer、注 3）から、信号処理に用いられるカルマンフィルタ（注 4）までを統一的に扱い、生成過程の不可逆性を表す「エントロピー生成」（注 5）を効率的に推定する方法を示しました。さらに、情報理論的な観点から、エントロピー生成を各時刻に対応するゼロ以上の寄与の和として表せることを示しました。大規模言語モデルを用いた実験では、文章に表れた出来事の順序関係がモデルに反映されていることを示唆する結果が得

られました。本成果は、生成 AI の内部に形成されると考えられる現実世界の表現、いわゆる「世界モデル」（World model、注 6）の定量的な理解につながることが期待されます。

本研究成果は、2026 年 11 月 5 日（米国東部標準時間）に科学雑誌「PRX Intelligence」のオンライン版に掲載されました。

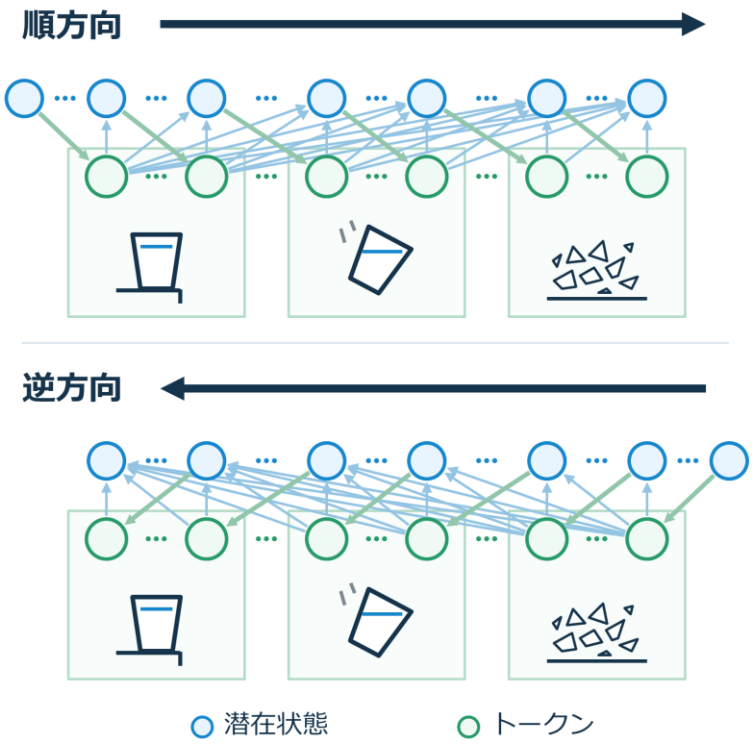


図 1：自己回帰生成モデル（特に大規模言語モデル）の概念図

緑の丸はトークン（単語などの処理単位）、青の丸は出力履歴から計算されるモデル内部の潜在状態を表し、矢印は依存関係を示します。上段は通常の順序、下段は逆の順序での出力列の模式図です。論文における文単位の解析では、文内の語順を保ったまま文の順序を反転します。

発表内容

【研究背景】

熱力学においては、時間の向きを逆にした状況の起こりにくさ（不可逆性）を定量化する「エントロピー生成」が重要です。自己回帰生成モデルにも、このような熱力学の概念を適用できると考えられます。ただし、自己回帰生成モデルが生成する時系列は一般に、直前の出力だけでなく、それ以前の出力の履歴にも依存します。このような性質をもつ時系列は一般に「非マルコフ過程」（注 7）と呼ばれますが、その熱力学には未解明な点が多く残されています。また、非マルコフ過程の時系列の確率を観測データだけから求めようとすると、膨大な量のデータが必要になるという問題がありました。

【研究内容】

本研究は自己回帰生成モデルに共通する構造に着目することで、考えられる過去の履歴全てを列挙することなく、非マルコフ過程であってもエントロピー生成を効率的に推定できること

を示しました。さらに、理論的には、未来の出力から過去の出力を推定するという観点から、エントロピー生成を各時刻の寄与に分解し、その情報理論的な構造を明らかにしました。

大規模言語モデルにおいては、単語（正確にはトークン）の順序を全て反転させると文法が崩れるため、文の順序だけを反転する「文単位」のエントロピー生成を導入しました。たとえば、「コップが床に落ちた。コップが割れた。」を逆順にすると「コップが割れた。コップが床に落ちた。」となり、現実世界の描写として不自然になります。両者の生成確率を比較することで、言語モデルが現実世界における「出来事の順序」をどのように学習しているかを定量化できると期待されます。

大規模言語モデルを用いた実験では、オープンモデル GPT-2 と Qwen3-4B-Base を用いました。結果として、独立した事実を並べた文章よりも、出来事が順に進む文章の方が、文単位で逆順にした生成が相対的に起きにくく、エントロピー生成が大きくなる傾向が見られました（図2）。ただし今回の実験では、評価対象の文章群は、商用の生成 AI にプロンプトで指示を与えて生成しています。今後、順序関係や因果関係の理論的な定義に基づいて文章群を構築するなどの方法で、さらなる検証が望まれます。

【研究の意義と展望】

本研究では、生成 AI と熱力学という、二つの研究領域を結び付ける理論を構築しました。大規模言語モデルを用いた実験では、現実世界において出来事が進む順序（不可逆性や因果関係）が、モデルにどのように表現されているかを定量化することを試みました。このような研究を通して、生成 AI の内部に形成されると考えられている現実世界についての表現、いわゆる「世界モデル」を定量的に調べる手がかりが得られることが期待されます。

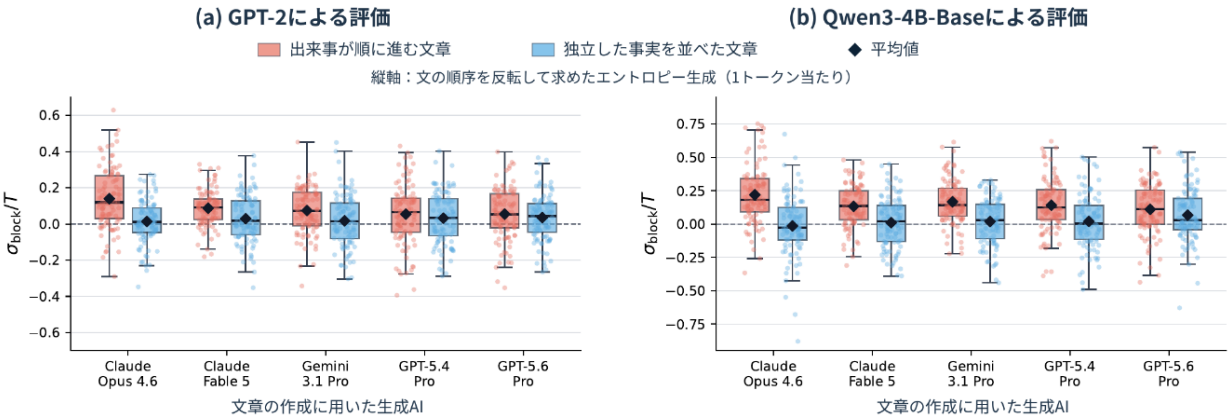


図 2：大規模言語モデルにおけるエントロピー生成

本研究で得られた、大規模言語モデルにおける文単位のエントロピー生成（正確には、確率的なエントロピー生成の分布）。「出来事が順に進む文章」（赤）と「独立した事実を並べた文章」（青）について、GPT-2（a）と Qwen3-4B-Base（b）で計算した値を示しています。縦軸は 1 トークン当たりの文単位のエントロピー生成、横軸は文章作成に用いた生成 AI を示しています。論文では、順位双列相関係数（rank-biserial correlation）とその 95%信頼区間による評価も行っています。論文の Fig.4(b)、Fig.10(b)を基に作成しました。

発表者・研究者等情報

東京大学大学院工学系研究科 物理工学専攻
沙川　貴大　教授
兼：同研究科附属量子相エレクトロニクス研究センター
稲盛科学研究機構（InaRIS）フェロー

論文情報

雑誌名：PRX Intelligence
題　名：Stochastic Thermodynamics for Autoregressive Generative Models: A Non-Markovian Perspective
著者名：Takahiro Sagawa
DOI：10.1103/tv38-b23y
URL：https://doi.org/10.1103/tv38-b23y

研究助成

本研究は、JST 戦略的創造研究推進事業 ERATO（課題番号：JPMJER2302）、東京大学 Beyond AI 研究推進機構の支援により実施されました。

用語解説

（注 1）自己回帰生成モデル（Autoregressive Generative Model）
これまでに生成した出力の履歴を基に、逐次的に次の出力を生成するモデル。本研究では、過去の出力履歴からモデル内部の状態（潜在状態）を計算し、それに基づいて次の出力確率を決めるという共通の構造に着目しました。

（注 2）大規模言語モデル（Large Language Model, LLM）
言葉の使われ方や文脈との関係などを学習した、大規模な言語モデル。特に GPT などの文章生成モデルは、過去の履歴に基づいて次のトークンを逐次的に生成することで、学習データである文章群の特徴を模倣します。本研究の実験では、比較的小規模な学習済みオープンモデルである GPT-2 (124M)と Qwen3-4B-Base を用いました。

（注 3）トランスフォーマー（Transformer）
トークン同士の関係に応じて、過去のどのトークンをどの程度重視するかを決める「注意機構（Attention mechanism）」を備えたニューラルネットワークのアーキテクチャ。大規模言語モデルの中核の一つになっています。

（注 4）カルマンフィルタ
雑音を含む観測信号から、対象の真の状態を逐次的に推定する手法。観測信号を用いずに適切なノイズを与えて動作させた場合は、観測信号を模倣する自己回帰生成モデルとみなすことができます。

（注 5）エントロピー生成
熱力学において重要な役割を果たす、時間変化の不可逆性を定量化する概念。通常の向きと逆向きの時間変化の起こりやすさの違いを表すものであり、物理現象に限らず一般の時系列に対

して定義できます。伝統的な熱力学ではエントロピー生成は発熱やエネルギー消費と関係しますが、本研究では自己回帰生成モデルの生成過程の不可逆性を定量化しています。そのため、本研究で扱うエントロピー生成は、モデルを実装するコンピュータ内部の半導体などにおける物理的な発熱やエネルギー消費そのものに対応するわけではありません。

（注 6）世界モデル（World model）
世界（現実世界やビデオゲームなど）の構造や変化、出来事の関係について、AI の内部に形成される表現。言語モデルも、文章群の学習を通して、物事の位置関係や時間的な順序などを捉える表現を獲得している可能性が議論されています。

（注 7）非マルコフ過程
次の出力の確率が直前の出力だけでは決まらず、それ以前の履歴にも依存する時系列を、非マルコフ過程と呼びます。大規模言語モデルにおいては、直前のトークンだけでなく、過去の文脈全体が次のトークンに影響します。

問合せ先

〈研究内容について〉
東京大学大学院工学系研究科 物理工学専攻
教授　沙川　貴大（さがわ　たかひろ）

〈機関窓口〉
東京大学大学院工学系研究科 広報室

科学技術振興機構 広報課

〈JST 事業に関する問合せ〉
科学技術振興機構 研究プロジェクト推進部 推進第 1 グループ
中村　亮二（なかむら　りょうじ）