LLM のふるまいの理解における 理想化された科学モデルの有用性について

平岡太郎 [1]　菅原朔 [2]
[1] 北海道大学大学院　[2] 国立情報学研究所
hiraoka.taro.h1@elms.hokuda.ac.jp　saku@nii.ac.jp

概要

大規模言語モデル（以下 LLM）の行う処理について、どのようにすれば理解可能な説明が与えられるか。LLM では処理のブラックボックス性が指摘されており、ベンチマークによる評価では処理機構が必ずしもわからず、出力の説明性が低い。本論文ではこの点を問題とし、それを解決する手段として科学モデルの概念が有用ではないかと指摘する。その上で、科学モデルが現象の理解をもたらす条件を確認し、どのようにすれば LLM の行う処理について理解可能な説明が与えられるかを考察する。

1　はじめに

LLM の出力結果を評価する際には、ベンチマーク的なデータセットを用いることが多い。しかしながら、ベンチマークでは出力結果と答えの比較のみで LLM の性能を評価するため、LLM の内部で行われている処理はブラックボックスであり、どのような処理により出力が得られたのかを説明することが難しいという問題が指摘されている。そして今後はアライメントなどをはじめとして、LLM が社会的に望ましくない出力をしないように要請される場面が増えている。よって LLM の出力に対する安全性を高める場合には、出力に対して「なぜそのような出力を行ったのか」を説明し、今後どのような出力を行うのかについての予測力を高める必要があるだろう。よって LLM の出力への説明性は重要な問題となっている。

ではどのようにすれば出力の説明性を高めることができるか。LLM 自体のパラメータ数の増大などを踏まえると、単に出力までの過程を逐次追跡することは、おそらく出力の説明性を高めることに寄与しない。そのため、ブラックボックス化せず、しかも我々にとって説明性を高めるための手段が求められることとなる。この点を踏まえると、具体的にはどのような手段が有用なのかを議論する必要があるだろう。

本稿では、このための手段の一つとして、科学におけるモデリングについて着目する。例えば電車の路線図について考えよう[1)]。路線図では、駅間の隣接関係が反映されるが、隣接する駅間がどの程度離れているかは反映されない。これは路線図を用いるのは「ある駅 A から別の駅 B に向かうには、どのようなルートがあるのか」を説明するためであり、必要な情報（どの駅とどの駅が隣接しているのか）を提示しつつ、不要な情報（駅間の実際の距離）を省くことで説明性を高めている。モデリングとは、この路線図のように、説明したい対象の必要な部分を取り出し、不要な部分を取り除いた説明を与えることである（以降はモデリングによって得られる説明を「科学モデル」と呼ぶ）。そして、出力の説明性を高めるためには、ベンチマークによる評価から進み、LLM に対するこのような科学モデルを作成することが有効であるというのが本稿の提案である。

以降は本稿の構成である。はじめに 2 節では科学モデルについて説明する。続く 3 節では、科学モデルがどのようにして対象の理解をもたらすのかという点について論じ、科学モデルを構成する際のポイントについて確認する。最後にここまでの議論を踏まえ、LLM の説明性を高めるための一つの方針について簡単に述べる。

2　科学モデルによる理解とベンチマークによる評価を通じた理解

本節では LLM の出力の説明性を高めるためには科学モデルを構成することが有用であると論じる。2.1 節では科学モデルについて地図を用いた比喩で簡単に説明を与える。2.2 節ではベンチマークによ

1）　科学モデルを地図との類比で捉える点は [1] に基づく。

る評価を通じた理解と、科学モデルを通じた理解を対比させ、後者が出力の説明性を高めるのに寄与しうることを論じる。

2.1 科学モデルとは何か

本節では以降の議論に先立ち、本論文で扱う科学モデル一般について、その概念的な特徴づけを与える。路線図の例に戻れば、路線図は「どの駅からどの駅に行けるのか」という点を表象するために用いられており、そのために必要な情報（駅間の隣接関係）は取り出され、不要な情報（駅間の実際の距離）は取り除かれていた。ここでのポイントは、ターゲットとなる対象（物理上の駅間の関係）があり、それを表象するための構造（地図は現実の駅間を表象しており、駅間の到達関係というグラフ構造化されている）が与えられていることである。そしてこの構造は、必ずしもターゲットとなる対象を忠実に構造化する必要はない（隣接関係のみをグラフ構造化し、実際の距離関係を構造化しないなど）。これを一般化すれば、科学モデルとは対象となるシステムを表象するために、必要な情報を取り出す（以降はこれを**理想化**[2]とする）形で構成された構造である[3]。

2.2 科学モデルを通じた理解の有用性

ここまでで科学モデルの特徴づけを与えた。以降は科学モデルが LLM の出力の説明性を高めるために有用かという点を論じる。現状の LLM はベンチマークを用いたタスクに対する LLM の性能評価が一般的であり、その評価に基づく形で LLM の言語能力について論じられることが多い。一方でベンチマークは出力のみに着目するため、LLM 内部での処理についてはブラックボックスのままであり、どのような処理によって出力が与えられたのかについては説明性が低い。つまりベンチマークによる出力の評価は LLM の言語能力を論じる際の手掛かりにはなりうるが、**どのように**そのような出力を行ったのかという点については説明性が低いのである。

では説明性を高めるにはどうすればよいか。一つの方針としては LLM の出力に対して「どのような処理によって出力が与えられたのか」という点を、内部処理をブラックボックスのままにすることなく説明することである。ではどのような説明が求められるのか。ここで再度路線図の例に戻って、この点を確認する。

いまターゲットとなる対象を現実の電車とし、目的が「どの駅からどの駅にたどりつけるのか」についての説明だとする。この説明を与えるためには路線図の作成が必要だが、路線図を作るためには現実の電車に関する詳細な記述は必要なく、むしろ説明の妨げになることが多い。つまり理想化は、説明の目的に照らせば推奨される手段なのである。そして理想化を通じて得られた路線図を通じて、我々は「どの駅からどの駅にたどりつけるのか」を理解する。

この例と類比的に考えれば次の結論が得られる。LLM の出力に対して「どのような処理によって得られたのか」を説明する際には、LLM の処理全体を忠実に追いかける必要はない。むしろ説明のためには一定程度の理想化を行い、この理想化によって科学モデルを構成した方が、説明性を高めるためには有用であると考えられる。そしてこの科学モデルは LLM の出力に対し、より説明性の高い理解を与える。

以上をまとめよう。従来のベンチマークによる評価では出力に至るまでの処理をブラックボックスのままにしてしまうため、「どのような処理によって得られたのか」についての説明性が低くなっていた。一方で科学モデルを通じた理解では、LLM の処理それ自体に一定の解明を与えうる。その上で理想化を通じて科学モデルを構成することで、仮に膨大なパラメータ数の LLM であっても、その出力に対して「どのような処理によって得られたのか」を理解しうる、というのが本稿での主張である。すなわち、科学モデルは LLM の出力に対して説明性を高めるために有用足りうるのである。

3 科学モデルの理解可能性

2 節では、LLM による出力の説明性を高めるには、ベンチマークによる評価だけではなく、科学モデルを通じた理解が有用ではないかという点を指摘した。我々は対象を説明する目的で、科学モデルの構成において理想化を行う。そして構成された科学モデルによる説明を通じて、対象の理解を得る。すると、科学モデルについては（1）理想化はどのよ

2） 本稿ではこの点についてはこれ以上深く立ち入らない。理想化の類型に関する代表的な議論としては [2] など。また理想化がどのような形でモデルに関わるのかについての比較的新しい議論は [3] などがある

3） 本稿では頁の都合上、科学モデルを暫定的にこのように特徴づける。ただし、この立場に反対する論者も居るため注意が必要である。総説としては [4] や [5] を参照。

うな役割を持つのか、(2) 科学モデルを通じた理解のためには何が求められるのかが問題となる。本稿で両者を同程度の比重で扱うことは難しいため、ここでは特に後者に着目する。その上で、ターゲットとなる現象の理解に際して、理想化はどのような役割を持つのかについて論じる。

まずは本節に先立ち、なぜ科学モデルの理解が問題になるのかについて、再び路線図の例で確かめておく。この例では、実際の電車の運行に対して必要な側面（ある駅から出たら、次にどの駅にたどりつくのか）を取り出し、これに基づいて路線図を作成していた。そして作成された路線図を通じて、我々は駅間の到達関係についての理解を得ていた。

では、路線図はどのようなものであっても理解をもたらすのか。そうではないということを、次の例で示そう。現在の路線図は駅をノード、隣接関係をエッジとしたグラフ構造で提示されることが大半である。しかし、グラフ化などを行わずに「駅 A に隣接するのは駅 B と駅 C で、駅 B に隣接するのは駅 A と駅 D で……」と隣接関係を書き下すことでも路線図と同等の内容を持つ科学モデルを提示することができる。そして、科学モデルの使い手にとっては前者のグラフ構造の方がより理解しやすいだろう。また、同じグラフ構造に基づいた路線図であっても、目的によって理解可能性が異なる場合がある。例えば都内の地下鉄のみを利用する人にとっては、都内の鉄道会社全ての路線図を網羅したものよりも、地下鉄のみを掲載した路線図の方が、利用者にとっては駅間の到達関係をより容易に理解できるだろう。

路線図の例と類比的に考えれば、同じ内容を持つ科学モデルであっても、構造化の仕方によっては理解のしやすさが異なること、および同じ構造化の仕方であってもより単純な方がより容易に理解できることがありうるのである[4)]。以降は [7] による**理解可能性（intelligibility）**という概念を導入し、科学モデルが理解をもたらすとはどのようなことかについて論じていく。

3.1 理解可能性（intelligibility）

モデルを通じた理解については何が求めらるのか。路線図の例に戻れば、我々は路線図には可視性などを求めており（グラフ構造では駅間の隣接関係を容易に把握できる）、しかもこうした要求は路線図使用者の置かれた状況によって変化している（地下鉄のみを利用する場合は、全路線についての路線図は必要なく、地下鉄の路線図のみでよい）という点を確認した。以降はこの例を踏まえつつ、モデルの理解可能性についての科学哲学上の議論を参照する。

科学モデルはターゲットとなる現象を理解するために構成される。そして先の路線図の例で確認したように、現象をより理解できるように促すような科学モデルの構成もありうる。では、科学モデルがよりよい理解をもたらすための条件とは何か。ここでヒントになるのは de Regt による**理解可能性**の概念である [7]。以降の de Regt の議論では、理解可能性が帰属される対象が科学理論である点に注意しつつ、理解可能性についての議論を確認する。

de Regt は科学的理解を説明するために、科学理論の理解可能性という概念を提示した。ここで理解可能性とは「科学者がある理論を使用するように促すような、科学者が理論の一群に帰属させる価値」のことである。路線図の例を挙げれば、これは可視性の高さによってグラフ構造が採用されたなどが当てはまる。その上で、ターゲットとなる現象が科学的に理解されるための規準（Criterion for Understanding Phenomenon: CUP）は次のように与えられる。

> CUP: 現象 P が科学的に理解可能 iff 理解可能（intelligible）な理論 T に基づいた現象 P への説明が存在し、しかもそれが経験的十全性や内部整合性などの基本的な認識的価値（epistemic value）と合致している

つまり、ある理論に基づいてターゲットとなる現象を理解する際には、科学理論は理解可能性を持たなければならないというのが de Regt の議論である。ここから、科学モデルを通じた理解可能性について何が言えるのかについて論じる。

路線図の例でいえば、ターゲットとなる現象は実際の駅間の関係であり、我々は路線図を構成してターゲットとなる対象を理解していた。その上で、路線図の構成時には可視性などを求めていたが、これは科学モデルが理解可能性を持つために行っていたといえる。

以上が理解可能性についての de Regt の議論である。この議論を踏まえれば、科学モデルを用いたターゲットとなる現象の理解については、理解可能性をモデルが持つことが重要であるという点が指摘できる。つまり、LLM に対して科学モデルを作成

4) 理解可能性がこのような文脈依存性を持つ点は、de Regt に固有の論点である点に注意が必要である。de Regt の文脈主義については [6, 7] などを参照。

し、内部でどのような処理が行われているのかについての説明性を高めるためには、科学モデルが理解可能性を持つ必要があるということである[5)]。では理解可能性はどのような形でもたらされるのか。次に理想化と理解可能性の関係について論じることで、両者の関係性について論じていく。

3.2 構成的理想化

ここでは理想化と理解可能性の関係について、主に高橋 [9, 10] の議論を参照しながら論じる。de Regt の議論を踏まえれば、科学モデルがターゲットとなる現象の理解をもたらすためには、科学モデルが理解可能性を持つ必要があった。ここで高橋は、科学モデルが理解可能性を持つためには、理想化が積極的な役割を持つと指摘する。路線図の例に戻れば、都内の地下鉄のみを利用する場合には「都内全ての駅について網羅した路線図」よりも「都内の地下鉄のうち、よく利用する駅についての路線図」の方が、路線図利用者にとっては理解しやすいだろう。これを踏まえると、科学モデルの持つ理解可能性はモデルの利用者とモデルの間に成り立つ関係であり、ターゲットとなる現象を正確に表象することと科学モデルの理解可能性は切り離せる。この点を推し進め高橋は「正確ではないが有用な（科学）モデル」という可能性を提示する。

その上で高橋では**構成的理想化**という方法論を提示する。先の路線図の例であれば、はじめに現実にあるのは電車という物体の移動と停止である。これを踏まえたうえで路線図を構成するには「駅間の隣接関係」という路線図を作成する際に不可欠となる現象を、単なる物理運動から構成する必要がある。そして、構成の過程で現実の物理運動の一部を理想化することにより、我々は具体的な「停車駅」などのターゲットとなる対象を構成するのである。これが、理想化を通じて科学モデルが扱いたい対象を構成する、という構成的理想化である。

例えばよく利用する駅間の路線図を子供のために書く場合は、通過する駅をわざわざ記載せず、出発駅と到着駅のみを書いて渡した方が路線図の理解可能性が高まる場面がある。そして路線図に対するこの理解可能性に基づいて、実際の電車の移動と停止に対して「出発駅」と「到着駅」というターゲットが構成され、その他の要素は無視するという理想化が行われるのである。このように、構成される科学モデルが理解可能性を持つように、理想化を通じて現象を構成することが構成的理想化である。その上で、理想化とは科学モデルの理解可能性を高めるために積極的な役割を持つ、というのが高橋の指摘である。

では、理想化と理解可能性がこのような関係を持つとして、科学モデルを用いて LLM の説明性を高めるにはどうすればよいか。従来のベンチマークによる評価では、言語能力に対する直観からタスクが構成されていた。しかし構成的理想化に基づけば、科学モデルが理解可能性を持つことが先行し、それに基づいてターゲットとなる現象を理想化して構成する必要がある。そのため、LLM に対して科学モデルを構成する場合には、単にタスクを設定するのではなく、まずはどのような形で LLM の出力に対して何を理解したいのかという点を明確にする必要がある。その上で、この理解可能性を踏まえ、出力に関連のある要素を理想化を通じて構成する必要がある。そして理想化された科学モデルは「正確ではないが有用」であれば理解可能性を持つため、必ずしも膨大なパラメータを正確になぞる必要はない。[6)]

最後に具体的にはどのような観点から構成的理想化を行うとよいかについて確認する。ここでは [11, 12] での議論が一つの参照点となる。つまり、出力を評価する際には、LLM 内部に構成概念のようなものを構成的理想化を通じて措定し、これに基づいて理解可能な説明を与えるという方針である。そして、このような科学モデルの構成については、心理学者を始めとした他分野との協働が不可欠であると考える。

4 おわりに

本稿では、1 節でベンチマークによる評価では出力の説明性が低い点を問題として提示し、2 節で科学モデルの議論が LLM の説明性を高めるのに有用ではないかと提案した。3 節では科学モデルの理解可能性と理想化の関係について、構成的理想化という観点から整理をした。最後に LLM の理解可能性を高める一つの方針として、構成的理想化により LLM の内部に構成概念などを与え、これを科学モデルの一種とみなすことで説明性を高める方針がありうると示唆した。

5) ただし、科学的理解については様々な立場が存在し、本稿では [7] に依拠した。総説としては [8] などを参照。

6) 特に今回は紙面の都合上扱うことができなかったが、科学モデルの観点から xAI 手法を再検討することも今後の課題である。

謝辞

本研究は JST 創発的研究支援事業 JPMJFR232R の支援を受けたものです。

参考文献

[1] Ronald N. Giere. **Science Without Laws**. University of Chicago Press, 1999.

[2] Michael Weisberg. Three kinds of idealization. **Journal of Philosophy**, Vol. 104, No. 12, pp. 639–659, 2007.

[3] Angela Potochnik. **Idealization and the Aims of Science**. University of Chicago Press, 2017.

[4] Michael Weisberg. **Simulation and Similarity: Using Models to Understand the World**. Oxford University Press, 2013.

[5] Roman Frigg and Hartmann Stephan. Models in science, 2024. https://plato.stanford.edu/archives/fall2024/entries/models-science/.

[6] Henk W. de Regt and Dennis Dieks. A contextual aprroach to scientific understanding. **Synthese**, Vol. 144, No. 1, pp. 137–170, 2005.

[7] Henk W. de Regt. **Understanding Scientific Udertanding**. Oxfodr University Press, 2017.

[8] 小林佑太. 説明的理解の現在. フィルカル, Vol. 6, No. 2, pp. 180–205, 2021.

[9] 高橋和孝. 理想化されたモデルによる科学的理解はいかにして成立するのか. **Linkage: Studies in Applied Philosohy of Science**, Vol. 3, No. 2, pp. 1–8, 2024.

[10] 高橋和孝. 構成的理想化に基づく科学的理解の解明. 博士論文, 2024.

[11] Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10967–10982. Association for Computational Linguistics, 2023.

[12] Saku Sugawara and Shun Tsugita. On degreees of freedom in defining and testing natural language understanding. **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13625–13649, 7 2023.