

Title	大規模言語モデルを活用した博士課程進学に関する因果探索の試行
Author(s)	高山, 正行; 小柴, 等; 三内, 顕義; 清水, 昌平
Citation	年次学術大会講演要旨集, 38: 880-885
Issue Date	2023-10-28
Type	Conference Paper
Text version	publisher
URL	<a href="http://hdl.handle.net/10119/19143">http://hdl.handle.net/10119/19143</a>
Rights	本著作物は研究・イノベーション学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Research Policy and Innovation Management.
Description	一般講演要旨

## 大規模言語モデルを活用した博士課程進学に関する因果探索の試行

○高山 正行 (NISTEP/滋賀大学), 小柴 等 (NISTEP/滋賀大学),  
三内 顕義 (NISTEP/京都大学), 清水 昌平 (NISTEP/滋賀大学/理化学研究所)

### 1 はじめに

昨今, 大規模言語モデル (Large Language Model: LLM) に注目が集まっており, 様々な LLM の公開・利用が進んでいる。これら LLM の主たる利用形態の一つとして, 自然言語を用いた問い合わせに対する, ある程度妥当で自然な応答がある。また, その背後には LLM の名前の通り, 莫大なテキストデータの事前学習があり, 知識ベースとしても一定程度利用できる。プログラミングなどの知識や各種のデータを準備しなくても, 不定形な様々な課題 (タスク) について, 自然言語で指示し, 期待される結果をそれなりの精度で, ある程度短時間で得られ, しかも, システムであるので大量に繰り返すこともできるなど, 取り扱える問題の幅が広く, かつ利用の敷居が低いことなどから, LLM を用いた様々な課題解決が試行・実行されている。こうした LLM を用いた取組の一つに, 因果推論への利用も挙げられる。

これまで, 領域横断的に扱える形で整理された因果推論のフレームワークは2つあり, ひとつが構造的因果モデルに基づいた Pearl 流のアプローチ, もうひとつが潜在的結果変数の枠組みに基づいた Rubin 流のアプローチで, 統計的因果推論はこれらのアプローチを基礎として確立されていた。近年では, この考え方を応用して統計的機械学習と融合させた, 因果ベイジアンネットワークや LiNGAM などによる統計的因果探索の手法も確立するなど, データ駆動的な統計的因果推論の幅が広がってきている。

ただし, 領域を問わず抱える因果推論の課題のひとつとして, 「データに基づいて推定された因果関係と領域知識をどのように整合させるか」という点に明確な解がないことが挙げられる。例えば, 博士課程進学に関する政策領域への応用の研究 [高山 23] においても, LiNGAM 系の因果探索アルゴリズムで出力された結果の妥当性を政策科学領域の知識から解釈することはできても,

- 変数が多くなるにしたがい, その結果の細部の採否を, 領域知識に基づいて, かつ恣意性を排除しながら網羅的に議論するのは現実的ではなくなること

- 統計的因果探索・推論のアルゴリズムを活用するにあたっての事前知識として, 質的な, しかし膨大な領域知識をどのように組み込み, 反映するか, 方法論が定まっていないこと

から, 体系的かつ踏み込んだ議論には至っていない。

ここで, 既存研究で積み上げられてきた, ある種トップダウンではあるものの, 妥当性が想定されて, 理論やデータで補強・試験され, 結果として十分に確度が高く関係性も強いと想定される要因・因果関係に加えて, 新たな要因を追加するに当たっては, ある程度の客観性や妥当性, 再現可能性を加えて検討できることが望ましい。前述通り, LLM はある種の知識ベースとしての活用が期待でき, ある程度の「客観性」と「再現可能性」を確保しやすい。また, 人が思いつかなかったような関係性を見つけ出せる可能性も有する。これに関連して, 「莫大なテキストデータに裏打ちされた知識・ロジックストックとしての高度な LLM は, 使い方の工夫は必要ではあるものの, これらのハードルを克服し, データ駆動的な因果分析と相補的で親和性の高い, 領域知識からの分析・考察を可能にするもの」という指摘もある [Kıcıman23]。

ただし, LLM は自然言語で指示できるという柔軟さ・自由度の高さの裏返しとして, 指示の与え方によって返答が異なることがある。そもそも, 学習に使ったデータ量やパラメータなど LLM の性能によっても振る舞いが大きく変わってくる。こうした問題について, 現状, 理論的な原因や対処方法は必ずしも明らかではなく, 様々な試行によって, 「こういう指示の仕方をすると, 精度が良くなる」「XX の課題についてはこの LLM の方が精度が良い」など, 知見が蓄積されている状況でもある。

そこで本研究では, 因果探索を行うにあたっての機械的・効率的な領域知識・事前知識の獲得方法確立を念頭に, 特に大学別データに対応する博士課程進学に関する政策領域への応用の研究 [高山 23] を対象として, LLM の中でも代表的な OpenAI 製の LLM である GPT-3.5 を活用したプロンプトエンジニアリング (LLM への指示の与え方の工夫) による因果分析を試行する。

## 2 GPT-4 の特徴と因果推論への応用に関する先行研究

LLM の中でも特に有名で広く取り上げられている OpenAI 製の GPT-4 <sup>\*1</sup>は、米国の医師免許試験や司法試験に合格したり、米国の大学入試共通テストである SAT の数学でも上位 10% 以内に入る [OpenAI23] など、その応答能力の高さに近年多くの注目が集まっている。さらに最近では、自然言語を用いた指示に対する適切な返答という機能に加えて、API での活用やデータのアップロード機能など、段階的にその機能を拡張している。こうした背景のもと、LLM の因果推論への活用についても様々な取組が実施・報告されている。

■因果推論のタスクに対する LLM の性能評価の取組  
例えば Kıcıman らは、既に因果関係についても一定の結論が出ている既存のデータセットも活用し、因果関係に関する見解を出力するタスクを LLM の既存知識に基づいて一定以上の精度でクリアできることを示した [Kıcıman23]。また、Zečević らは、既に因果関係として ground truth として知られているいくつかの有向非巡回グラフ (Directed Acyclic Graph, DAG) の例について、複数のテンプレートを用いて、2 変数間の因果関係が存在するかを問い、得られた回答と ground truth となっている DAG を比較して性能評価を行っている <sup>\*2</sup>[Zečević23]。さらに Jin らは、シミュレーション用の正解付きの自然言語での因果関係の記述データを生成し、それをもとに LLM に因果の正解を言い当てる性能を評価するとともに、さらに fine-tuning により因果推論精度が高まることを実証した一方、LLM への質問のテンプレートや変数名の言い換えに対しては頑健ではないという結果も示唆している [Jin23]。

■プロンプトエンジニアリングの方法論に関する先行研究  
一方で、こういったプロンプトエンジニアリングを行うにあたり、本来はパフォーマンスを最適化するためのプロンプトを生成すること自体に多くの工夫が必要であり、特に因果推論においては、ほとんどの場合は多くの変数関わっているために、関係に対して肯定的か否定的かをはっきりと表せるとは限らず、応答を出力させても、それによる因果関係の存在の確からしさを定量的

に評価することは容易ではない。そのため、本研究の目的達成のためには、上述の研究動向も参照しながら、因果関係の有無の判定タスクにおいて有効なプロンプトテンプレートの作成について、検討を進める必要がある。

プロンプトエンジニアリング一般では、様々なタスクで LLM の応答を最適化できるよう、様々なアプローチ <sup>\*3</sup>で研究が活発になされている。例えば、LLM にタスクを課す際のプロンプトそのものを最適化するには、例えば LLM そのものに複数のプロンプトをパラフレーズとして生成させ、スコアリング・リサンプリングを行って最適なものを絞り込むという Automatic Prompt Engineer(APE) という手法が提案されている [Zhou22]。このアプローチは様々な発展し、最近ではシンプルに事実を問うプロンプト (Zero-Shot Prompting) であっても付加させるだけでパフォーマンスを向上させる、“Let’s think step by step” プロンプトも発見されたり [Kojima23]、プロンプトの素案を入力すると最適なプロンプトに書き換える商用サービス“PromptPerfect” <sup>\*4</sup>も登場している。

また他にも、他の推論の例をプロンプトに入れたり (Few-Shot Prompting)、関連知識を与えることで推論タスクにおける LLM の正答率が向上することもよく知られているが、その応用として最近では、LLM による知識生成 (knowledge generation) を行い、その知識をプロンプトに統合 (knowledge integration) することで推論タスクのパフォーマンスを向上させる Generated Knowledge Prompting(GKP) という手法も考案されている [Liu22]。

なお、このようなプロンプトエンジニアリングの手法も参考しつつ、特定の分類タスクへの実用的な活用を検討する観点からは、因果推論に直接携わる研究ではないものの、論文の引用文脈が肯定的か、中立的か、否定的かを分類するタスクについても、LLM の性能を評価する研究も最近なされており [西川 23]、LLM ベースで出力される因果関係の信頼性を定量的に評価するヒントになり得る。

本研究では以上の先行研究を基に、実験準備として、因果関係の有無の判定タスクに有効なプロンプトテンプレートの検討から行う。

## 3 GPT を活用した新たな「因果推論」の実験手法の構築

先述の通り、特定の変数間の因果推論を試みる上では、LLM の推論パフォーマンスが最適となるようなプ

<sup>\*1</sup> ここでは ChatGPT など LLM を用いたサービスも含めた広い意味で用いる。

<sup>\*2</sup> ただし彼らは、LLM が因果推論に成功している例であっても、それは LLM が本当に因果を理解しているのではなく、あくまで LLM の訓練に使われた自然言語の中にある因果的事実間の相関が基盤となっていたのであり、LLM の性能は因果を語るには十分なものには至っていないと主張している。

<sup>\*3</sup> 代表的なノウハウは、<https://www.promptingguide.ai/> に随時まとめられている。

<sup>\*4</sup> <https://promptperfect.jina.ai/>

ロンプト・条件の検討が必要となる。そこで本章では、実際に博士課程進学に関する「因果推論」を試行する前に、既存のプロンプトエンジニアリングの技法を参考に、実験デザインの検討と性能評価を行う。なお、以下の議論では特に断りがない限り、LLM としては、API から GPT-3.5-turbo を使用する。

### 3.1 因果の有無を問うための最適なテンプレートの生成

まず、変数  $X$  が変数  $Y$  の原因となるか否かを問う最適なテンプレートを探るため、APE の手法 [Zhou22] を参考に、GPT-3.5-turbo を用い表 1 の通り入力した。その結果、表 2 の通り、20 のパラフレーズが生成された。本稿ではこれらのテンプレートを  $n$  で番号付けし、原因となる言葉  $X$ ・結果となる言葉  $Y$  を引数として、 $Q_n(X, Y)$  と表現する。

表 1: 因果の有無を問うテンプレートの生成に使用したプロンプト。

We want to make various kinds templates for prompt engineering in order to ask whether X can be the cause of Y. X and Y are replaced with the respective variable names after making the templates. Please reproduce 20 different templates for this aim. Here are the examples of and outputs we want you to give us:

Does X cause Y?  
Does X influence Y?  
Does X affect Y?  
If we change only X, is Y changed?  
Our hypothesis is that X causes Y. Is it true?

このうち、 $n = 3, 14, 18, 20$  については、 $X$  と  $Y$  に因果関係があるかのみを問うようなプロンプトとなっており、原因・結果がどちらかまでを定めたうえで問うようなプロンプトにはなっていないことに留意を要する。一方でこれら 4 つ以外については、 $X$  が原因、 $Y$  が結果となるかどうかまでを問うものとなっている。

### 3.2 テンプレートの性能評価

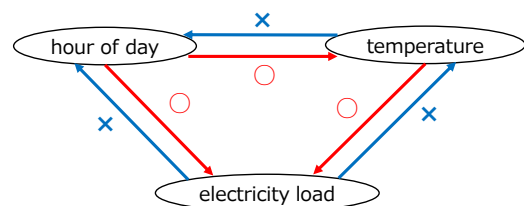
表 2 のうち、LLM からより正確な回答を引き出すのに適切なプロンプトが何か、実際にタスクを課して、比較を行う。タスクとしては、因果関係が既によく知られた事実 (ground truth) を正解として取り上げ、各変数同士の原因・結果の関係の有無を正確に回答できるかを問うものとする。

統計的因果推論では、ground truth がある程度確定したものとして、Mooij らによりまとめられた、因果推論

表 2: 表 1 に基づいて実際に生成されたテンプレート一覧。temperature=0.7 とした。

$n$	テンプレート $Q_n(X, Y)$
1	Does X lead to Y?
2	Can X be attributed to the cause of Y?
3	Is there a causal relationship between X and Y?
4	If X is modified, will it have an impact on Y?
5	We believe that X is responsible for Y. Is this assertion valid?
6	Is Y influenced by X?
7	Can changes in X result in changes in Y?
8	Is there evidence to suggest that X is the cause of Y?
9	Does X have a direct effect on Y?
10	If X is manipulated, will Y be affected?
11	Is there a connection between X and Y, where X is the cause and Y is the effect?
12	Can X bring about Y?
13	Will Y change if X is altered?
14	Is there a correlation between X and Y, indicating a causal relationship?
15	Is X a determining factor for Y?
16	Can X be considered a potential cause of Y?
17	If X is controlled, will it impact Y?
18	Is there a link between X and Y, suggesting a causal association?
19	Does X play a role in causing Y?
20	Is there a cause-effect relationship between X and Y?

図 1: 本研究で因果関係の有無を問うにあたってのベンチマークとする関係。



においてベンチマークとなる観測データセットがある [Mooij16]。本研究は中でも“electricity load”に関するデータセットに着目し、そこで扱われている 3 変数 (“hour of day”, “temperature”, “electricity load”) から  $X$ ,  $Y$  に代入した全 6 通りの質問を、表 3 の通り行いつつ、そのデータセットでも ground truth となっている、図 1 の関係を正解として、性能評価のベンチマークとして採用することとした。

以上の条件で、まず Zero-Shot Prompting の考え方に基づいて打ち込んだプロンプト、及び応答を集計した全体の正答率上位 4 位・下位 4 位のテンプレートの番号  $n$



表 3: 質問と変数, 正解のセット

質問番号	原因となる変数 X	結果となる変数 Y	質問に対する正解
質問1	the hour of day	the temperature	Yes
質問2	the hour of day	the electricity load	Yes
質問3	the temperature	the electricity load	Yes
質問4	the temperature	the hour of day	No
質問5	the electricity load	the hour of day	No
質問6	the electricity load	the temperature	No

表 4: 事前知識なしで  $Q_n(X, Y)$  での因果を yes/no の形で問うプロンプトと, 実際に各  $Q_n(X, Y)$  で問うた際の GPT-3.5-turbo での正答状況 (temperature = 0.7)

“Let’s discuss the problem of causation on energy consumption related with climates.  
 $Q_n(X, Y)$   
 Please answer with <yes> or <no>. No answers except these two responses are needed.”

順位	n	質問1	質問2	質問3	質問4	質問5	質問6	全体正答率
1	19	14%	70%	98%	98%	64%	62%	68%
2	4	14%	68%	96%	98%	72%	54%	67%
3	5	12%	70%	100%	94%	64%	60%	67%
4	12	10%	72%	98%	96%	64%	58%	66%
...	...	...	...	...	...	...	...	...
17	20	16%	64%	92%	0%	42%	48%	44%
18	3	16%	64%	94%	8%	32%	46%	43%
19	14	20%	66%	100%	2%	30%	40%	43%
20	18	8%	56%	96%	8%	42%	42%	42%

を, 表 4 に示す。なお, 各質問については全て 50 回ずつ行い, そのうち正答に一致した出力が得られた回数の割合を算出し, 全体の正答率は質問 6 つでの平均正答率を算出している。

これらから見られる見られる傾向として, 以下の点が挙げられる。

- 質問 1 では, どのプロンプトでも正答率が低いが, 質問 3・4 は特に上位 4 つで正答率が 100% に近い。
- 下位 4 つのテンプレートは, 質問 4~6 での正答率が低い。(これらは, 因果の向きまでは問わないテンプレートになっており, そのことが正答率に影響していることが考えられる。)

ただし, この下位 4 つ以外のパラフレーズはいずれも全体正答率で 60% をこえており, 決定的な差があるわけでもない。

### 3.3 関連知識の付与に基づく正答率の改善と最適な質問テンプレートの選択

プロンプトの精選をさらに進めるため, 今度は関係する知識を, どのテンプレートにも前提として与えたいので, 3.2 節と同様の性能評価を行う。ここでは, 正答率 100% を目指す観点から, 今回は LLM の事前学習された知識よりも過剰と考えられるが, ChatGPT Plus で

Wikipedia のプラグイン機能を利用して,  
 ” In the context of the relation between the climate and the energy consumption, please provide the knowledge of X in detail as much as possible.”

( $X =$  the hour of day, the temperature, the electricity load) と入力して得られた回答を統合し, 関連知識として加えた。この下処理をした上で改めて LLM に因果の有無を問うたプロンプトと, 応答を集計した全体の正答率上位 3 位<sup>\*5</sup>・下位 3 位の  $n$  は表 5 の通りである。

表 5: 事前知識ありで  $Q_n(X, Y)$  での因果を yes/no の形で問うプロンプトと, 実際に各  $Q_n(X, Y)$  で問うた際の GPT-3.5-turbo の正答状況 (temperature = 0.7)

“Here is the basic knowledge of the hour of day, the temperature, and the electricity load, in the context of the relation between the climate and the energy consumption.

(prior knowledge acquired from Wikipedia through ChatGPT plugin)

Using this knowledge, let’s discuss the problem of causation on energy consumption related with climates.

$Q_n(X, Y)$

Please answer with <yes> or <no>. No answers except these two responses are needed.”

順位	n	質問1	質問2	質問3	質問4	質問5	質問6	全体正答率
1	4	0%	100%	100%	100%	100%	100%	83%
1	12	0%	100%	100%	100%	100%	100%	83%
1	19	0%	100%	100%	100%	100%	100%	83%
...	...	...	...	...	...	...	...	...
18	14	0%	92%	100%	0%	100%	100%	65%
19	8	0%	14%	36%	100%	100%	100%	58%
20	5	0%	0%	12%	100%	100%	100%	52%

十分な関連知識を与えた上でのこの推論タスクの結果からは, 以下の特徴が見てとれる。

- 質問 1 は, 関連知識を与えることで, むしろ “No” と回答する例がほとんどとなった。この原因としては, そもそも Mooij らのデータセットで ground truth とされていたことが, 必ずしも普遍的な事実として完全なものではない可能性<sup>\*6</sup>が考えられる。
- 3.2 節において全てのパラフレーズの中でも上位のパフォーマンスを示していた  $n = 4, 12, 19$  について

<sup>\*5</sup>ただし, 掲載した 3 つ以外にも質問 2~6 の正解率が 100% で同率 1 位のプロンプトが複数あるが, 本稿では Zero-Shot Prompting でも上位の性能を誇ったもののみを候補と考え, 掲載している。

<sup>\*6</sup>実際には Mooij らも, 本来は太陽と地球の位置関係が気温に影響するのであり, あくまで日照時間は太陽と地球の位置関係をよく代替する指標として用いていると補足している [Mooij16]。

は、質問 2~6 では正答率 100% となっており、大幅にパフォーマンスが改善した。

- 一方で  $n = 5$  は、3.2 節では上位 4 つのうちの一つであった一方、関連知識を十分に与えると全体正答率が 43% とかえって悪化している。 $n = 5$  では因果関係の仮説を提示してその妥当性を評価させるような聞き方になっていたのに対し、特に正解が “Yes” である質問 2・3 で非常に低い正解率となっており、逆に言うと、全体としてほとんどの回答で “No” と応答している\*7 ことになる。
- また、 $n = 8$  のテンプレートも、因果関係の証拠はあるかという聞き方になっており、これに対しての質問 2・3 の正答率は  $n = 5$  ほどではないが悪い\*8。

よって、Zero-Shot Prompting の場合、関連知識を与えた場合の両方で安定的に活用できるテンプレートは  $n = 4, 12, 19$  である。また、多値変数のみを扱っている博士課程進学 of LiNGAM での因果探索の研究 [高山 23] と対応させることを見据えると、 $n = 12, 19$  のように事象間の因果関係を問うよりも、数の変化について問うテンプレートの方が自然であると考えられる。よって本稿では、 $Q_4(X, Y) = \text{“If } X \text{ is modified, will it have an impact on } Y\text{?”}$  を採用することとする。

### 3.4 Generated Knowledge Prompting の場合のパフォーマンス評価と実験条件の確定

関連知識の付与は、LLM の応答の適正化の有効な手段である一方、バイアスの少ない形で関連知識をテキストベースで整えて付与することは必ずしも容易ではない。先の “electricity load” は因果関係が ground truth とされるぐらいには、多くの知見があり、ChatGPT plus のプラグイン機能でも明快に整理し、関連知識としては十分すぎるものであったが、実際に未知の因果を探索するにあたっては、判断材料を十分に体系的に揃えられないこともあり得る\*9。そこで、根本的な解決になるわけではないが、本稿では Generated Knowledge Prompting の手法 [Liu22] を参考に、LLM そのものが事前学習で獲得した知識を何らか引き出し、付与することとする。実際に LLM からの知識生成に使用するプロンプト (1st prompt)、及びそこで得られた知識等を統合し  $Q_4(X, Y)$

\*7 本稿ではこれ以上立ち入らないが、これは GPT-3.5-turbo が、特に提示された仮説の妥当性の評価などの場面では、極めて慎重な見解を示すよう、事前学習されている可能性を示唆している。

\*8 これも GPT-3.5-turbo が、根拠の存在を示すにあたっては慎重な見解を示すよう、事前学習されている可能性を示唆している。

\*9 博士課程進学の問題についてもこの事情が当てはまり、既存の関連知識を体系化することに多くの時間・労力を要する可能性がある。

にて因果の有無を問うプロンプト (2nd prompt)、そして 3.2 と同じタスクを実行した際の正答率の状況は表 6 の通りである。なお、1st prompt の入力に当たっては temperature は 0.7 で固定したが、2nd prompt でのタスクの実行においては、temperature を 0.1~0.9 の範囲で変えながら性能評価を行った。

その結果、temperature=0.1 では質問 1 以外は全問正解と、3.3 節と同等のパフォーマンスを示したが、それより大きい temperature では 3.3 節よりも性能の悪化が見られる。一方、0.7~0.9 でも正答率が急激に変化が見られず安定している。仮に LLM により様々な可能性を広く探索するのであれば、必ずしもこのベンチマークテストで精度 100% を求めて temperature を 0.1 に固定せず、引き続き 0.1~1.0 ぐらいの範囲で振りながら探索を継続することも有効であると考えられる。

## 4 おわりに

本研究では、特に大学別データに対応する博士課程進学に関する政策領域への応用の研究 [高山 23] を対象として因果探索を行うにあたって、機械的・効率的な領域知識・事前知識の獲得方法の確立を念頭に、LLM の中でも代表的な OpenAI 製の LLM である GPT-3.5 を活用したプロンプトエンジニアリング (LLM への指示の与え方の工夫) による因果分析を試行した。

本予稿ではそのうち、解の構造が明確で既知である問題をベンチマークとして用いて、LLM に因果を問い、精度の高い回答を得るための適切なプロンプトのあり方を検討し、そのパフォーマンスの差異についても、実験を通じて示した。また、プロンプトに推論のための前提知識を付与することによるパフォーマンスの向上を確認し、さらに GPT-3.5 の temperature の変化に伴う正答率の挙動についても確認を行った。その結果、

- 因果の有無を問うプロンプトには、事前知識が付与された場合でも、Zero-Shot Prompting の場合よりもパフォーマンスを落としてしまうようなものも存在し、改めてプロンプトの検討では注意を要すること
- GKP により、GPT-3.5 はよく知られた事実に関する因果について回答するタスクで高いパフォーマンスを示し、内容によっては temperature が高くなっても正答率の振る舞いは頑健であること

などが確認された。また、これらのプロセスによって、さしあたり因果関係を LLM に判定させるプロンプトとして有力な候補を見つけ出すことができた。

表 6: LLM の事前学習された知識のみに基づく因果説明を求める knowledge generation のプロンプトと、その応答に基づく knowledge integration を施したうえで、yes/no での因果判定のプロンプト。

1st prompt(knowledge generation),temperature=0.7

"I would like for you to analyze and explain if X has an impact on Y. Please provide a clear and concise response that addresses the cause-and-effect relationship between these two variables. Support your analysis with relevant statistical data, trends, or patterns. Focus on whether an increase or decrease in X directly affects Y. Make sure to keep your explanation informative while avoiding excessive technical jargon."

2nd prompt(knowledge integration)

An expert was asked the question below:

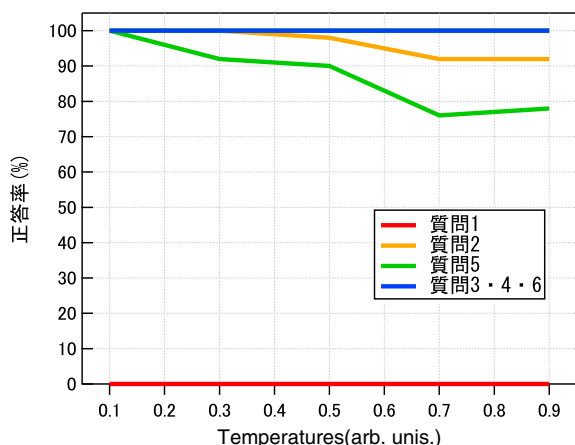
"I would like for you to analyze and explain if X has an impact on Y. Please provide a clear and concise response that addresses the cause-and-effect relationship between these two variables. Support your analysis with relevant statistical data, trends, or patterns. Focus on whether an increase or decrease in X directly affects Y. Make sure to keep your explanation informative while avoiding excessive technical jargon."

Then, the expert replied:

"(the answer of the question above from LLM)"

Taking this discussion above, let's consider the problem of causation on energy consumption related with climates.

If X is modified, will it have an impact on Y? Please answer with <yes> or <no>. No answers except these two responses are needed."



講演では、これらの知見を基に、さらに博士課程進学に関する政策領域の因果関係について、LLM による抽出を試行した結果を示し、それに基づいた分析の結果と、これらのテキストベースの統計分析と統計的因果探索の融合の可能性等についても議論する。

## 謝辞

本研究の一部は、JST, CREST, JPMJCR22D2 の支援を受けたものである。

## 参考文献

- [高山 23] 高山正行, 小松尚登, ファム テトン, 前田高志ニコラス, 三内顕義, 小柴等, 清水昌平: 大学別の博士課程進学等に関するデータセットの構築と統計的因果探索. 研究イノベーション学会 第 38 回年次学術大会 (予稿集), 公演番号 2D20, 2023.
- [OpenAI23] OpenAI: GPT-4 Technical Report. arXiv:2303.08774v3, 2023. <https://doi.org/10.48550/arXiv.2303.08774>
- [Kiciman23] Emre Kiciman, Robert Ness, Amit Sharma, Chenhao Tan: Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. arXiv:2305.00050v2, 2023. <https://doi.org/10.48550/arXiv.2305.00050>
- [Zečević23] Matej Zečević, Moritz Willig, Devendra Singh Dhami, Kristian Kersting: Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. arXiv:2308.13067v1, 2023. <https://doi.org/10.48550/arXiv.2308.13067>
- [Jin23] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, Bernhard Schölkopf: Can Large Language Models Infer Causation from Correlation?. arXiv:2306.05836v1, 2023. <https://doi.org/10.48550/arXiv.2306.05836>
- [Liu22] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, Hannaneh Hajishirzi: Large Language Models Are Human-Level Prompt Engineers. arXiv:2211.01910v2, 2022. <https://doi.org/10.48550/arXiv.2211.01910>
- [Kojima23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa: Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916v4, 2023. <https://doi.org/10.48550/arXiv.2205.11916>
- [Zhou22] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba: Generated Knowledge Prompting for Commonsense Reasoning. arXiv:2110.08387v3, 2022. <https://doi.org/10.48550/arXiv.2110.08387>
- [西川 23] 西川開, 小柴等: 引用文脈分析における大規模言語モデル (LLM) の応用可能性. Jxiv, 2023. (preprint) <https://doi.org/10.51094/jxiv.467>
- [Mooij16] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, Bernhard Schölkopf: Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 2016. 17(32):1-102. <http://jmlr.org/papers/v17/14-518.html>