研究・イノベーション学会　第40回年次学術大会
2B22 (2025年　11/9(日)16:00～)

LLMの政策研究への応用に向けた性能評価の構想

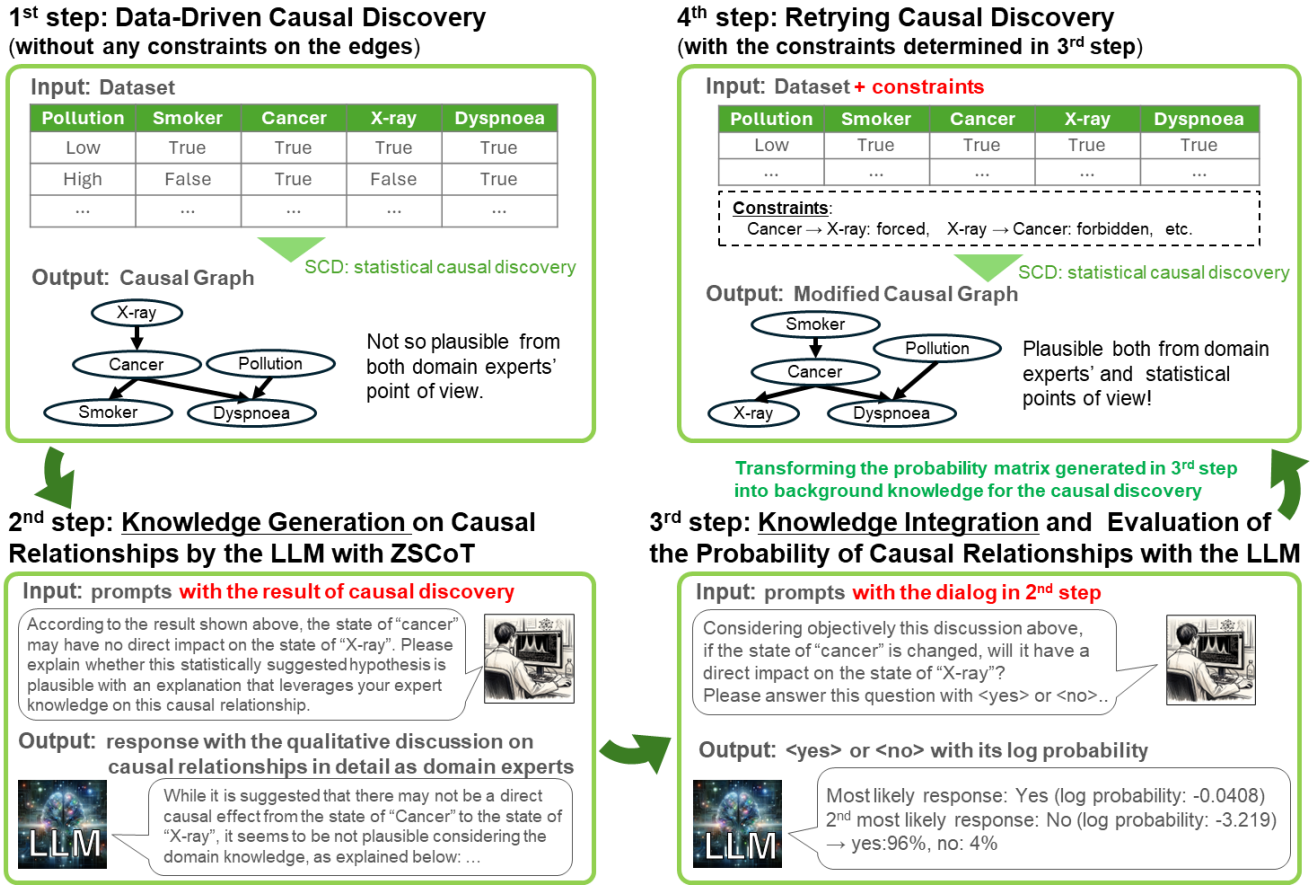
高山 正行[A, B], 小松 尚登[B],
三内 顕義[A, B, C, D, E], 清水 昌平[A, B, C, F、G]

A: 科学技術・学術政策研究所 (NISTEP)
B: 滋賀大学
C: 京都大学
D: 東京大学
E: 国立情報学研究所
F: 理化学研究所　革新知能統合研究 (AIP) センター
G: 大阪大学

Mail: masayuki-takayama@biwako.shiga-u.ac.jp

背景：領域知識に基づく因果グラフ構築のための大規模言語モデル（LLM）の活用

✓ 当グループでは、政策研究における因果推論への貢献を見据え、LLMを活用し、領域知識から見ても統計的に見ても、妥当な因果グラフを構築する方法論を、最近提案したところ



M. Takayama *et al.*, TMLR(2025) https://openreview.net/forum?id=Reh1S8rxfh

✓ 当グループとしては、まずは博士課程進学に関する因果探索の文脈での応用（→2B03の予稿参照）を検討しているが、**果たしてこのドメインでもLLMが領域知識に基づいて正しく/フェアに動いてくれるか等の論点が残っている**

✓ この例でいえば、2nd stepは特に、**ドメイン知識に基づく質問応答機能**が問われる

本研究の目的・概要

✓ 政策研究ドメインにおいて、LLMが、**既存の領域知識と整合し、**
意味的・論理的に破綻のない出力（≒ 適切な質問応答）がどの程度可能か
（≒ 政策研究のドメインの専門家として正しく/フェアに助言したり、
考察・解釈をすることが可能か）、
その「能力」を評価し、
「能力」に応じて、人間の専門家がLLMを援用していくかという基本的枠組を検討

✓ LLMの「能力」評価に関する**新たな方法論やアルゴリズムの提案・試行的な実験**を実施

※　特に本研究では、博士課程進学に関する政策領域での因果判定（→2B03予稿参照）に必要な知識を、LLMがどの程度有しているかを測定することを例として取り上げる

LLMの出力原理と現実的な評価方法の考え方

●LLMの出力原理:「次単語予測」

…プロンプトやそこまでに出力されたトークン列に基づいた、条件付き確率に従って、
確率的にトークンが選択・出力される

➡ ある文脈（与えたプロンプト）で特定のトークンが生成されるかどうか※を確認することで、
質問応答の性能を見られる

※多くのLLMでは、APIでlog-probabilityとして、トークンの生成確率を出力可能。☚**我々はここに着目**
しかし、LLM系の先行研究では、まだこの機能による質問応答の性能評価に踏み込んでいる研究は、多くない。

●応答の測定範囲

✓「いいcodeを書けるか」、「流暢・適切に要約や言い換えができるか」などの複雑なタスクで、
確率ベースで応答の定量評価をしたい場合
- トークン列全体を考慮するため、
  ・この条件付き確率の積を見る
  ・何回も回答をランダムに生成する
  といった、高コストな測定

✓ ただ単に、既存の領域知識と整合し、意味的・論理的に破綻のない出力ができるかを見る場合

➡ プロンプトを選択肢式の問題にしてしまい、選択肢のみを回答させることで、
正解選択肢トークンの出力確率に基づく、シンプルな測定でOK

領域知識に基づく質問応答性能を見るための、多肢選択式のベンチマークテストが効果的

日本の政策研究ドメインでのベンチマークテストと評価の問題点

●そもそも、ベンチマークテストになるものがないのでは？

…医学領域なら医師国家試験が、法律なら司法試験が、
数学や物理、歴史なら、センター試験・大学入試共通テストの過去問や、
MMLU（Hendrycks, ICLR, 2021）などのベンチマークテストがあるが…

➡ とりあえず、測定の観点からやりやすいように多肢選択式にするのであれば、
新規に作成する必要あり

●政策科学の学際性の高さ

…様々な専門知の学際統合（秋吉、Keio SFC journal, 2021）
ex) 政治学、法学、計量書誌学、統計学、心理学、労働経済学…

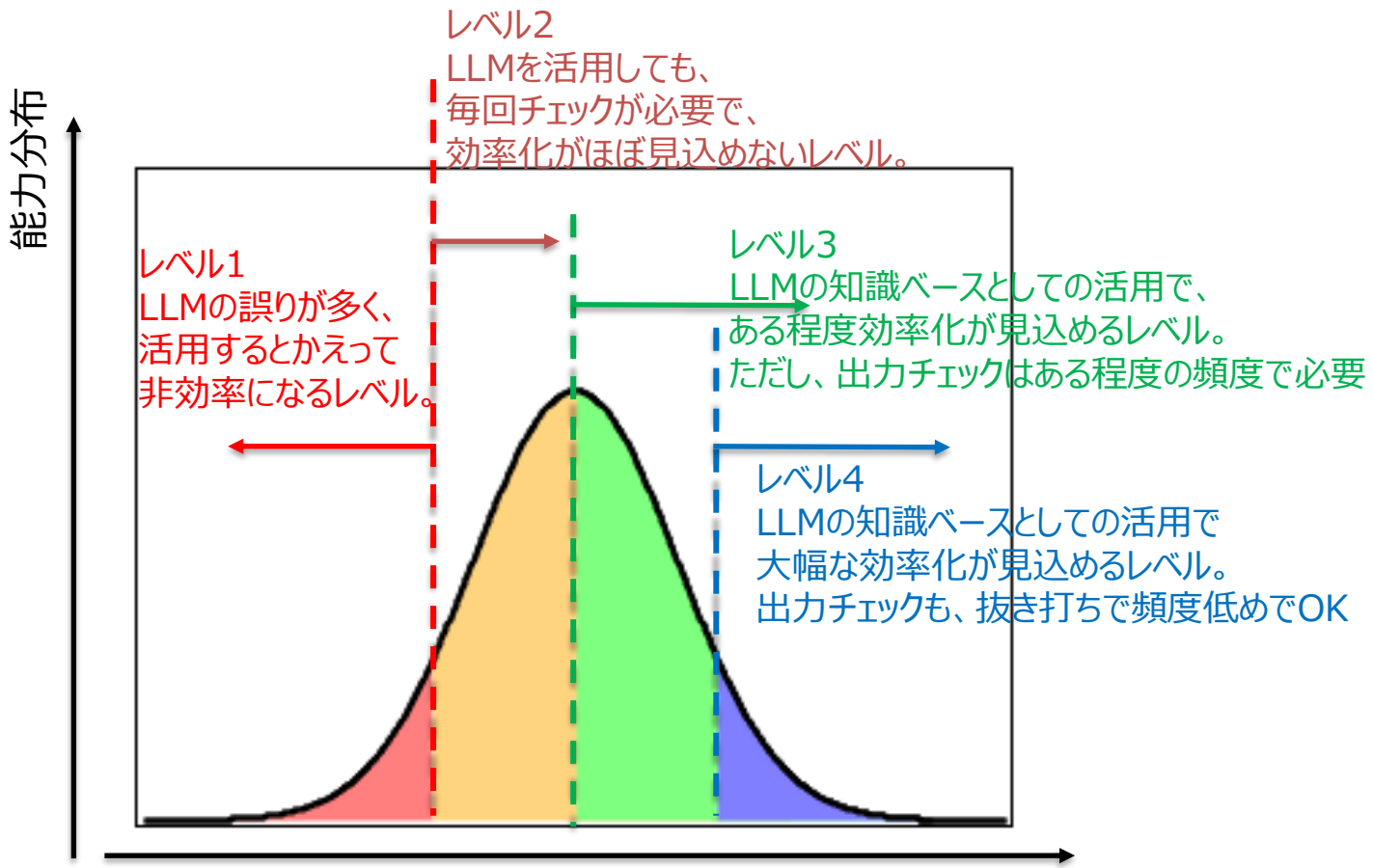
➡ どの範囲までブレイクダウンして測定したいかに応じて、綿密に検討が必要

●LLMの能力をどう評価し、運用の判断基準をどう考えるか？

➡ とりあえず、何らか人間のドメイン専門家の目が入ってチェックする前提としつつ、
少しでも効率化するには、LLMの能力に応じた専門家のコミットの濃淡を決めたい。
しかし、そのための測定の技術的課題もある。（詳細次ページ）

LLMの援用に対するドメイン専門家の確認エフォートの在り方の検討の方向性（案）

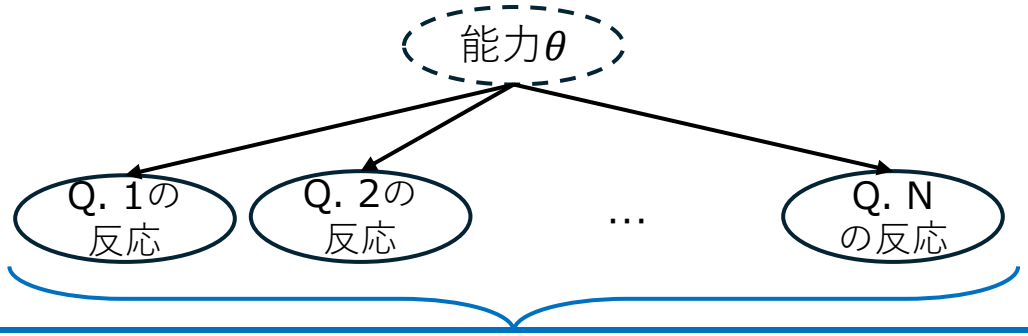
✓「能力」の分布に基づき、
・人間のドメイン専門家と比べてLLMの能力がどうかを見定め、
・それに応じて「人間のドメイン専門家がどれだけコミットし、確認・修正が必要になるか」を検討する



✓ ただし、上記の考え方で評価を進めるには、以下を実現するための技術的検討が必要
-**問題ごとの難易度も違うので、それを加味して測定**すること
-そのうえで、**人間でも回答可能な問題量で測定し、能力比較ができる可能性を追求**すること

項目反応理論（Item Response Theory）の援用の検討

項目反応理論（Item Response Theory）：
各質問/テスト項目への応答が、潜在的な能力を示す変数θを引数とした、
確率関数に依存して生成されるものとして、応答結果をもとに分析する方法



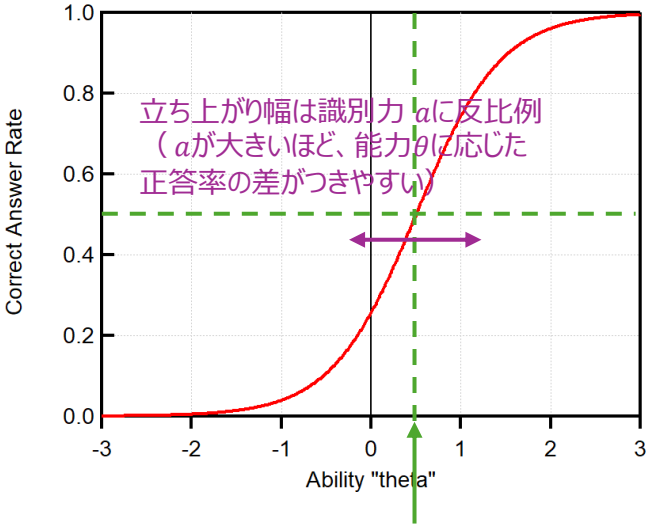
仮定1（項目間独立性）
全ての項目間の関係は能力θのもと条件付き独立

仮定2（一次元性）
問題iについて、潜在変数θからの反応が正解となる確率が、
例えば2PL logit modelで、

$$P_i(\theta) = \frac{1}{\exp(-Da(\theta - b)) + 1}$$

と表され、全てのiについての反応確率が、同じ潜在変数θから同様に表される

例: 2PL logit modelで、
識別力 $a = 1.25$
困難度 $b = 0.5$
の場合の、能力θに応じた反応率の変化
（能力θは、標準正規分布すると仮定）



- ✓ 個々の各項目に対する反応（0（不正解）/1（正解））のバイナリーデータから、**そのバイナリーデータをモデルから再現できる確率が最も高まるような能力θ, a, bを計算**
- ✓ 単純な正答率ベースでの分析に比べ、**問題の難易度の違いにも配慮しながら、能力算出ができる**
- ✓ **回答しない問題がある/対象によって解く問題が違う場合でも、能力の比較が可能**になる

今回新たに開発した正答確率データ行列でのIRT分析アルゴリズム

✓個々の各項目に対する反応（0（不正解）/1（正解））のバイナリーデータ※の代わりに、
LLMの場合、正答確率のデータ行列を用いることで、より少ないサンプルサイズでの分析が可能にならないか？

※バイナリーデータでのIRT分析はすでに学力調査でよく行われるようになっているが、サンプルサイズ1,000程度ないとつらい

【新たに開発したIRTアルゴリズムの概要】（2PL logit modelを仮定）
問題数M、サンプルサイズN（LLMの数）のときに得られるLLMの正答確率データ行列(p_{ki})
$(1 \leq k \leq M,\ 1 \leq i \leq N)$ に対し、**以下のstep 1と2を、ループ終了条件A~Cが満たされるまで（つまり、各パラメータが収束したとみなされるまで）繰り返す**

Step 1: θ_i の更新
a_k^{old}, b_k^{old}を固定した上で、

$$\theta_i^{new} = \underset{\theta_i \in \boldsymbol{R}}{\arg\min}\left(\sum_{k=1}^{M}\left|p_{ik} - \frac{1}{1+\exp\left(-Da_k^{old}\left(\theta_i^{old} - b_k^{old}\right)\right)}\right|^2\right)$$

Step 2: a_k, b_k の更新
θ_i^{new}を固定した上で、

$$\left(a_k^{new}, b_k^{new}\right) = \underset{a_k>0, b_k \in \boldsymbol{R}}{\arg\min}\left(\sum_{i=1}^{N}\left|p_{ik} - \frac{1}{1+\exp\left(-Da_k^{old}\left(\theta_i^{new} - b_k^{old}\right)\right)}\right|^2\right)$$

ループ終了要件を全て満たすかチェックしながら、この部分を繰り返す

ループ終了条件A: $\max\left(\left|a_k^{new} - a_k^{old}\right|\right) < \epsilon_a$
ループ終了条件B: $\max\left(\left|b_k^{new} - b_k^{old}\right|\right) < \epsilon_b$
ループ終了条件C: $\max\left(\left|\theta_i^{new} - \theta_i^{old}\right|\right) < \epsilon_\theta$

新しいIRTアルゴリズムの性能評価の概要

【性能評価のシミュレーション手法】

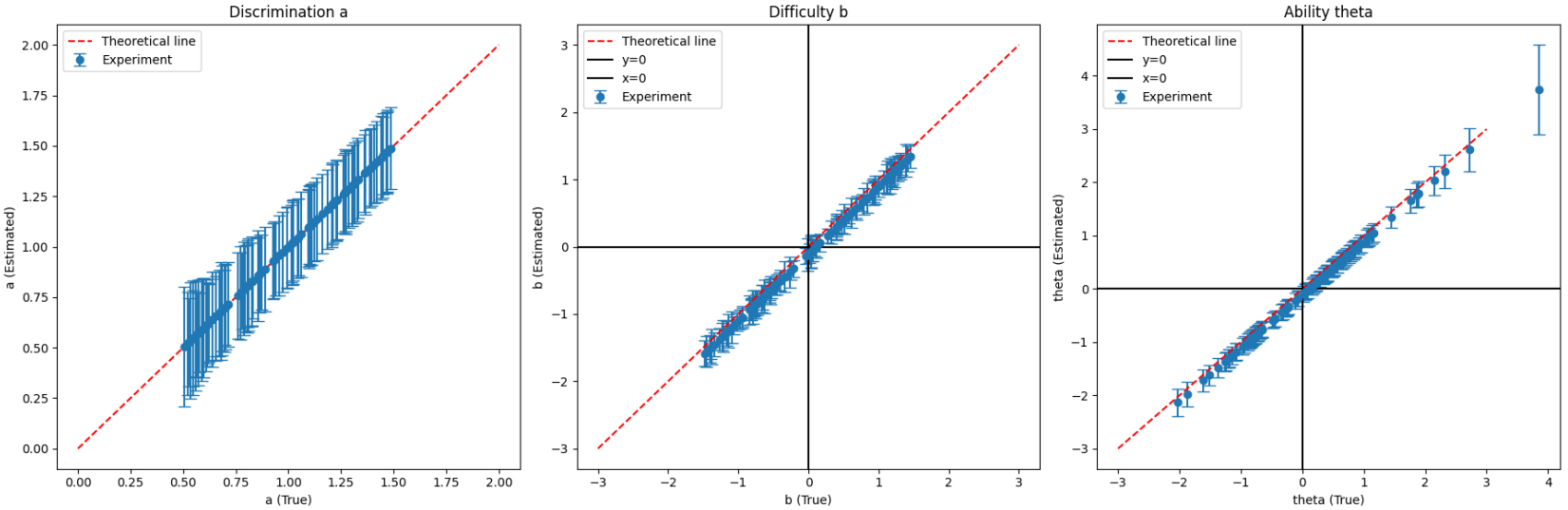
① 疑似的に、問題数M、サンプルサイズ（測定対象LLMの数）Nを定めてから、
各問題の識別力a、困難度b、そして各サンプル（LLM）の能力θを定め、確率データ行列を2PL logitモデルで再現

②得られたデータ行列から開発したアルゴリズムで、識別力a、困難度b、各サンプル（LLM）の能力θを推定し、
①で決めた値を正しく再現するかを確認

構成する確率データ行列のイメージ

サンプル	項目1の正答率	項目2の正答率	項目3の正答率	項目4の正答率	項目5の正答率
LLM 1	0.75	0.62	0.1	0.32	0.99
LLM 2	0.94	0.85	0.4	0.71	0.999
…	…	…	…	…	…
LLM N	0.4	0.1	0.01	0.06	0.73

問題数M=100,サンプルサイズ N=100としたときのシミュレーション結果



✓ サンプルサイズ100でも、推定値は十分正解データに近くなる

➡ 2PL logitモデルでのIRTの仮定を満たせば、LLMのサンプルサイズが必ずしも稼げない状況でも、開発したアルゴリズムである程度の精度で推定可能

政策研究関係のクイズでの試行実験の概要

◆問題：web上にも公開されている資料等に基づいて、**一意に正解を特定できる選択式（5択）の、日本の科学技術政策**（特に、**博士課程進学にかかる経済的支援、研究時間等の研究環境、競争的資金の使途・取扱、運営費交付金の配分の考え方** 等）**に関するクイズ**を20問作成し、そのままプロンプトにする

問題リスト

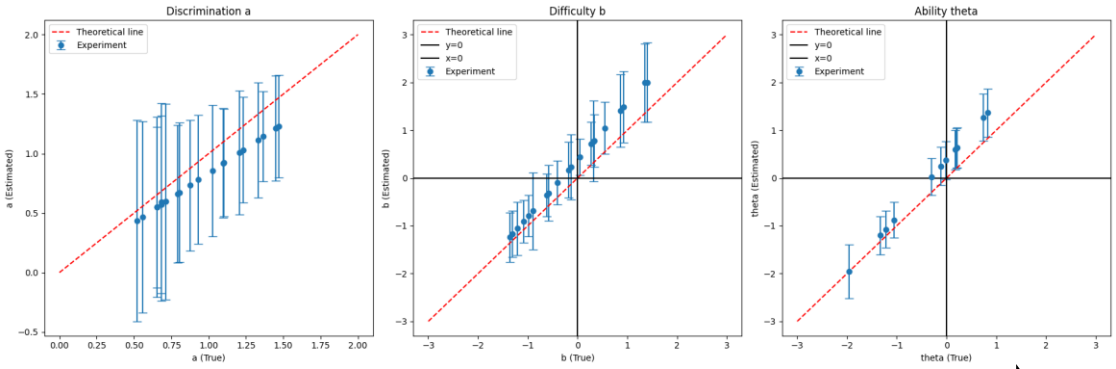
◆実験対象のLLM：OpenAIのchatモデルを12種類使用

◆実験パラメータ: 温度Tは今回は1.0で固定

◆正答確率の測定方法: 5回測定の上で、API通じて得られるprobabilityの平均値を採用

番号	能力測定の対象としたLLM
1	gpt-3.5-turbo-1106
2	gpt-3.5-turbo-0125
3	gpt-4-1106-preview
4	gpt-4-0125-preview
5	gpt-4
6	gpt-4-turbo-2024-04-09
7	gpt-4-turbo
8	chatgpt-4o-latest
9	gpt-4o-mini-2024-07-18
10	gpt-4o
11	gpt-4o-2024-05-13
12	gpt-4o-2024-08-06

問題数M=20, LLMのサンプルサイズN=12の場合
（MとNは今回報告する実験で準備できた実際の数）



○ 特に識別力aの標準誤差が大きいが、推定値は全体的に比較的安定
○ thetaの値は比較的標準誤差も小さい

トライアル的ではあるが、
大雑把な評価は可能と考えられる

政策研究関係のクイズでの結果概要① - 能力推定値

Probability-Based IRTで計算された結果

番号	能力測定の対象としたLLM	能力θの推定値（標準誤差）
1	gpt-3.5-turbo-1106	-1.304(0.320)
2	gpt-3.5-turbo-0125	-1.077(0.317)
3	gpt-4-1106-preview	-0.415(0.345)
4	gpt-4-0125-preview	0.553(0.333)
5	gpt-4	-1.626(0.320)
6	gpt-4-turbo-2024-04-09	-0.489(0.340)
7	gpt-4-turbo	-0.483(0.341)
8	chatgpt-4o-latest	1.509(0.347)
9	gpt-4o-mini-2024-07-18	0.290(0.351)
10	gpt-4o	0.824(0.314)
11	gpt-4o-2024-05-13	1.254(0.327)
12	gpt-4o-2024-08-06	0.963(0.312)

（注）推定値の見方

○ 値は相対的なものであり、標準化をかけている

○ 値が大きいほど能力が高く、小さいほど能力が低い、と解釈。

✓ 上位互換モデルの方が高い能力が、この政策研究クイズのタスクでも出る傾向にあり、実際のOpenAIの一般的な世代・モデルスペックの優劣と概ね一致

✓ 標準誤差も、このサンプルサイズ・問題数でシミュレーションされた値と同程度なので、測定精度を極端に悪化させる要因もなさそう

➡ **本研究で提案したLLMの政策研究の能力評価手法と問題セットがある程度有効に機能している**

政策研究関係のクイズでの結果概要②
- 特徴的な結果が見られた問題例

問題リスト

【正答確率が著しく高かった問題】

問題例1（識別力2, 困難度-2, いずれも発散）

国立大学法人に基盤的経費として配分される運営費交付金の性質として、当てはまるものを一つ選んでください。A～Dに当てはまるものがない場合は、Eを選択してください。

A. 経営状況が厳しいほど重点的に配分される。
B. 人件費、物件費等の区分がない、渡し切りの交付金であり、各大学の裁量に応じて執行が可能である。
C. 各大学への交付総額が毎年数パーセントずつ減額されることが決まっており、減額分は自己収入の増加で対応しなければならない。
D. 教職員の退職手当等は、運営費交付金に含まれず、別途支給される。
E. A~Dの中に当てはまるものはない

回答はA, B, C, D, Eのうちアルファベット1文字のみでお願いします。

【正答確率が著しく低かった問題】

問題例2（識別力2, 困難度2, いずれも発散）

日本の競争的研究費において間接経費が原則として直接経費の30%と定められたのは、第何期科学技術基本計画期間ですか。当てはまるものを一つ選んでください。A～Dに当てはまるものがない場合は、Eを選択してください。

A. 第1期科学技術基本計画期間　（1996年～2000年）
B. 第2期科学技術基本計画期間　（2001年～2005年）
C. 第3期科学技術基本計画期間　（2006年～2010年）
D. 第4期科学技術基本計画期間　（2011年～2015年）
E. A~Dの中に当てはまるものはない

回答はA, B, C, D, Eのうちアルファベット1文字のみでお願いします。

政策研究関係のクイズでの結果概要②
- 特徴的な結果が見られた問題例

問題リスト

【ほどほどに測定力があることが示唆された問題①（易）】

問題例3 （識別力0.713, 困難度-0.891）

特別研究員(DC)で、月額20万円以上の生活費を制度上得られる方法として、正しいものはどれですか。当てはまるものを一つ選んでください。A～Dの全てが当てはまる場合は、Eを選択してください。

A. 研究専念義務を順守しながら、RAに従事する。
B. 週1回2時間程度のTAに従事する。
C. 研究に努め、採用最終年度のタイミングで成績優秀者に選ばれ、特別手当の追加支給を得る。
D. 社会通念上、常勤職と見なされない範囲でのアルバイトを行う。
E. A~Dの全てが当てはまる

回答はA, B, C, D, Eのうちアルファベット1文字のみでお願いします。

【ほどほどに測定力があることが示唆された問題②（難）】

問題例4 （識別力0.354, 困難度1.355）

日本におけるフルタイム換算データに関する調査（FTE調査）を他の調査と突合しながら分析する際、制約となり得ることはどのようなものですか。当てはまらないものを一つ選んでください。A～Dの全てが当てはまる場合は、Eを選択してください。

A. 5年に1回程度の統計調査であるため、学校教員統計調査等と時点を合わせて分析することが難しい
B. 悉皆調査ではなく、学問分野ごとに抽出率を設定し、実施されているため、仮に大学別に細かく分析しようと思っても、その集計値が各大学の真の値を表しているという保証がない
C. 外部研究資金の申請のための時間が研究時間に含まれており、申請書を除いた、論文執筆や実験、分析に関する研究時間についての議論ができない。
D. 個人情報を取得せず、科学技術研究調査での大学事務局所有の名簿の番号に基づいて系統抽出し、調査対象を決定しているので、他の研究者個人レベルでのローデータと突合して分析することができない。
E. A~Dの全てが当てはまる

回答はA, B, C, D, Eのうちアルファベット1文字のみでお願いします。

本発表のまとめ

本研究の成果

✓政策研究領域におけるLLMの利用に向けた能力評価に関し、
LLMの技術的性質やこれまでの一般的な性能評価の状況にも照らし、
ベンチマークテストの構築の必要性とその際の課題、IRTの援用が有効である可能性について指摘

✓ LLMの能力評価のために、Probability-Based IRTでのフレームワークを、そのシミュレーションの妥当性と共に提案した
（当方の把握している範囲では、LLMの正答確率ベースでのIRTを検討し、少数サンプルでも統計的信頼性の高い分析
が可能となることを実演した技術的研究自体、これが初と思われる）

✓ 博士課程進学関係の政策研究のクイズのプロトタイプ20問を作り、OpenAI製のLLMで思考実験し、
モデルのスペックに概ね沿った結果が得られ、提案した政策研究の能力評価の手法は有効に機能することが
期待される結果となった

留意点・今後の展望

✓ 実験対象となるLLM数を増やす（gemini, llama, NIIの日本語特化LLM）➡ **サンプルサイズの拡大**

✓ 問題数を増やす ➡ **能力評価の精度、測定する領域範囲の網羅性の向上**

✓ できることなら、行政官や政策研究のドメイン専門家等にも解いてもらって、IRT分析する
➡ **これができれば、人基準でLLMがどこまで成長すべきか、という議論にも繋がる可能性**

✓ 人間にとっての難易度とLLMにとっての難易度がかみ合わない場合の対応の検討