2B22

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（NISTEP/滋賀大学）, 小松 尚登（滋賀大学）,
三内 顕義（NISTEP/京都大学/滋賀大学/東京大学/理化学研究所/国立情報学研究所）,
清水 昌平（NISTEP/大阪大学/滋賀大学/京都大学/理化学研究所）

1 はじめに

本稿では政策研究における（主に分析・考察過程での）大規模言語モデル (Large Language Models: LLM) 活用に向けた，評価用ベンチマークの設計と，ベンチマークを利用した LLM の評価について論じる。

■LLM の応用に関する現状認識 LLM をはじめとする生成 AI への期待・注目はいまなお高まり続けている。これまで，様々な LLM が作られ，技術的にも，実践的な利活用に向けた様々な開発が進められてきた。本稿執筆現在の OpenAI 社の最新モデルである，GPT-5 にたどり着くまでにも，

- モデルのパラメータ数・学習方法の改善による，医学，法学等の各領域での性能向上 [OpenAI23]
- COT(Chain-Of-Thought[Wei22] 等を活用し，段階的な推論に基づいて結果を出力する，reasoning model の開発による，致命的な誤りの防止や，複雑なタスクでの性能向上 [OpenAI25]

といった技術の積上がある。その流れの中で特に，我が国の政策立案及びそれを支える政策研究での LLM の活用可能性は, 例えば因果推論の文脈で議論され [高山 23], その後，統計的信頼性と領域知識の観点での信頼性の両立に向けた手法開発がなされてきた [Takayama25]。

■本稿での課題設定 著者らは, 文献 [高山 25] にも著した通り，今後 LLM× 因果探索の研究手法 [Takayama25] を応用し，政策研究の領域知識からの妥当性と統計的妥当性を両立させた，博士課程進学に関する因果モデルの構築を目指している。しかし，そもそも LLM が，当該領域において，因果関係を正確に判断するための知識をどの程度保持し，使用できるかは，定かではない。さらに，その評価方法も，別途検討が必要である。

そこで本研究では，この評価のための新たな方法論やアルゴリズムを提案し，試行的な実験を行う。なお，本稿の主たるターゲットは，博士課程進学に関する政策領域での因果関係判定に必要な知識を，LLM がどの程度有しているのかを検証することである。しかし，この議論は政策研究領域全体へ一般化することも有益であるため，特に 2 章では，政策研究における LLM の知識活用の総論として議論する。

なお，以降の議論では LLM の政策研究の「能力」は専ら，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こと[*1]を呼ぶ。

2 政策研究に関する LLM の能力と評価

2.1 LLM の出力原理とベンチマークテストの使い方

LLM の出力原理は，「次単語予測」に基づく。具体的には，以下の式に基づき，プロンプト Q および出力済のトークン列 $W = \{w_1, w_2, ..., w_{n-1}\}$ で定まる条件付き確率として，次のトークン w_n が a となる確率が決まる。

$$P(w_n = a \mid Q, W) = \frac{\exp\left(\frac{z(w_n{=}a \mid Q, W)}{T}\right)}{\sum_{\text{all}} \exp\left(\frac{z(w_n \mid Q, W)}{T}\right)} \quad (1)$$

ここで，$z(w_n \mid Q, W)$ は，プロンプト Q に対し，トークン列 W が出力された直後の，トークン w_n の出力確率に関するロジットである。また, T は LLM の出力調整のハイパーパラメータである。ゆえに，長い文章であるほど，その出力は多段の条件付き確率の積となり，評価も複雑で難しい。一般には，LLM が与えられたタスクに対し，適切に回答を出力できるか評価するベンチマークテストとして，例えば，数学・物理学・生物学等の様々な領域の多肢選択式の客観問題を集めた MMLU[Hendrycks21] 等が用いられる。LLM[*2]の，シンプルな知識・技能に関する出力性能を評価するには，このような多肢選択式のテストセットを用いることが望ましい。

また，通常は問題ごとに，LLM の出力結果から抽出さ

[*1] 実際には LLM は，英文校正等の研究補助にも活用しうるが，本稿では領域知識の適切な活用が必要な場面での使途を想定する。
[*2] ただし，OpenAI 社の GPT-5 のような reasoning model は，さらに複雑な仕組みであるため，ここでは議論の対象から外す。

れた選択肢が正解と一致するか確認し，各領域の LLM の正答率を算出して評価する。その際，COT[Wei22] などで，知識を順序だてて組み合わせ，回答させる方が性能が上がることも知られている。一方，プロンプトから即座に正しい選択肢が出力されるかを評価するだけならば，プロンプトで選択肢だけの出力を指示し，式（1）に基づく正解選択肢の出力確率を直接取り出すことで，より精緻な分析が可能になると考えられる。

以上の基本認識のもと，日本の政策研究における LLM の能力評価に関する課題を提示する。

2.2 日本の政策研究におけるベンチマークテスト

さて，数学や歴史学などのように，初等中等教育段階から体系的カリキュラムや，入試科目としての位置づけが確立している学問分野では，演習問題や能力測定のベンチマークとなる素材が数多く蓄積され，アクセスも容易である。MMLU のようなベンチマークテストも，そういった学問分野の多肢選択式の問題から構成されている。

一方，政策研究では，特に日本の政策研究の能力に関する多肢選択式のベンチマークテストは，まとまった形のものは存在しないため，おそらくゼロから考案・収集する必要がある。その際特に，知識・技能の考え方を明確にしつつ，客観的なテスト（各問題において正解が一意に定まるもの）の構築に留意する必要がある。また，新たに問題を作る場合は，回答の根拠となる文献を問題ごとに特定しておくことで，仮に LLM に足りない知識/技能がそのベンチマークテストで特定され，何らか LLM の拡張で改善を図る際に，追加学習等による性能向上を試みる際の有力な材料となることが期待される。

後述の通り本研究でも，この点に留意しながら，博士課程進学に関連する作問を試行しているものの，次に述べる事情により，その網羅と適切なポートフォリオの構築は，現実的には極めて困難である。

2.3 政策科学の学際性の考慮

しかし，政策科学には，様々な専門知の学際統合により成り立っている側面もある [秋吉 21]。例えば，科学技術・イノベーション政策でも，人材政策のための労働経済学，研究成果として論文を取り扱うための計量書誌学，さらには基礎的な調査における質問紙調査法と密接に関連する心理統計学等，関連する学問領域が幅広い。そのため，特定の政策研究に関する知識のみではなく，その周辺領域の能力についても，考慮されるようにしておくことが望ましい。本研究で作成した問題セットにも，この点に配慮し，統計調査の性質や，歴史的時系列の整合性等の理解を問う問題も可能な限り含めている。しかし，実際に網羅的な問題セットの作成を開始する際には，この学問構造の体系的な把握が一定程度必要と考えられる。

2.4 LLM の能力の評価・利用基準の考え方

一方で，仮に政策研究のベンチマークテストの構築と可能になったとしても，政策研究の能力を実際にどのように評価し，また評価に応じてどのように活用していくかは，政策ドメインの専門家自身での検討が必要となる。いわゆる Human-in-the-Loop の考え方[*3]に基づいて活用する場合も，LLM の能力レベルに応じ，出力を政策研究者がどの程度検証するべきかが異なり，この政策研究者の割くべきエフォートの大きさが，Human-in-the-Loop 全体の効率に直結する。さらに，その判断基準の定め方は，LLM のエラー発生時の修正対応の深刻さや，エラーが残ってしまった際の深刻さも，その領域によって異なることから，必ずしも LLM の評価結果のみで決定できるわけではなく，政策研究領域の相場観が重要となる。

この評価基準の考え方として，図 1(a) のイメージを提案する。仮に，人間の政策ドメインの専門家と同じ座標軸で，LLM の能力を測定できるとすれば，2.3 節の懸念点を克服して構築したベンチマークテストについて，人間の政策ドメインの専門家の回答も収集・採点し，例えば図 1(a) に示すイメージのように，その分布に基づいて基準を作り，LLM の能力の測定結果がその基準を上回るか否かで，利用方法を判断する，ということが考えられる。これにより，評価された LLM について，人間の政策ドメインの専門家でいう，○○相当の精度，と判定することも可能となり，そのレベルに応じ LLM の援用が効果的か，そして出力チェック等にどの程度の工数が必要かを，検討することもできる。

ただし，問数が十分に多い場合は、対象となるすべての政策ドメインの専門家から，全問の回答を得ることは実現しにくくなる。その場合，図 1(b) に示す通り，TOEIC や TOEFL 等と同様に，現実的に回答可能な問題数を抽出し，政策ドメインの専門家に応じて異なる問題セットを構築し，項目反応理論 (Item Response Theory, IRT) に基づき能力を推定することが考えられる。

3 IRT による LLM の能力評価の技術的検討

前節で言及した IRT は，例えば人間の学力調査では，各受検者について，問題ごとに正解/不正解を測定するので，離散的なデータが想定される。そこから，多段階反応モデルや，連続応答モデルに拡張されるなど，より

[*3] 完全に LLM 任せにせず，何らか人間の専門家が確認し，必要な修正を施した上で LLM の出力を活用するというループ。

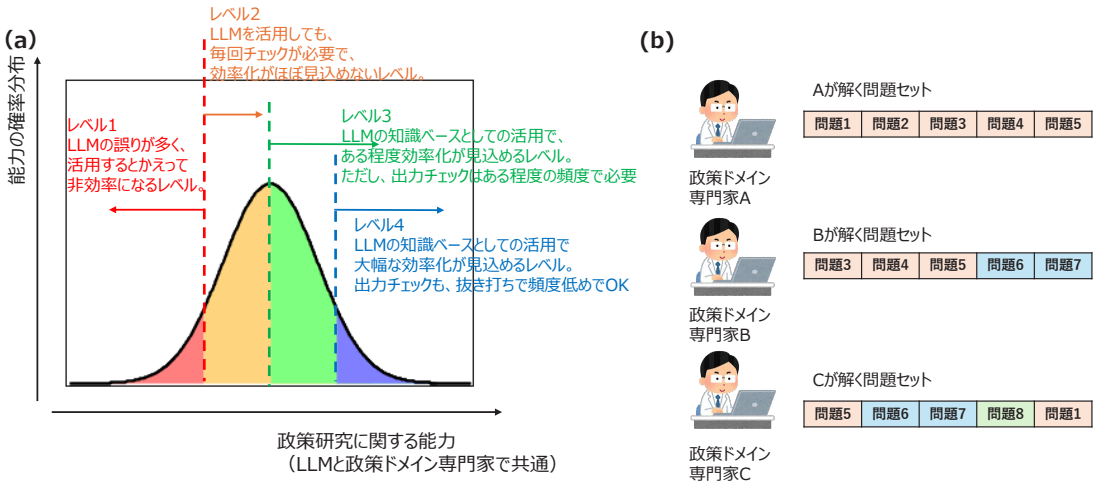

図 1: LLM の政策研究の能力を，人間の専門家と同じ基準で測定・評価するイメージ。(a) 測定された能力分布について，人間の政策ドメインの専門家とも比較しながら，LLM のレベルと活用可能性を議論するイメージ。(b) 政策研究の能力基準を IRT で測定するにあたり，政策ドメインの専門家にも，現実的な問題数を抽出したうえで，別々の問題セットを出題するイメージ。

多様なデータへの対応も検討された [澁谷 20]。しかし，特に LLM への多肢選択式ベンチマークテストへの反応を分析する場合，式（1）に基づく確率を直接取り出せる。よって，この連続量に対して直接ロジスティック回帰し，能力パラメータを求めることが可能となる。この利点を生かし，本章では既存の IRT アルゴリズムを改良し，その基礎的な評価を実施する。

3.1 2 母数ロジスティックモデルによる IRT アルゴリズムの改良

IRT の中でもよく使われる反応モデルに，2 母数ロジスティックモデルがある。各問題への LLM の反応について，LLM の能力 θ による一次元性と問題間の独立性を仮定すると，能力 θ を持つ LLM の，ある問題 i への正答確率 p は，以下の通り表せる。

$$p(\theta, a_i, b_i) = \frac{1}{1 + \exp(-Da_i(\theta - b_i))} \tag{2}$$

ここで，$a_i(> 0)$ は識別力パラメータであり，0 に近い値ほど，問題 i の反応は能力 θ の b_i 近傍で差がつきにくく，大きい値であるほど差がつきやすい。また，b_i は困難度パラメータであり，小さいほど問題 i は易しく，大きいほど難しい。また，尺度因子 (比例定数)D は，$D = 1.702$ ととることで，$p(\theta)$ が正規累積曲線に近づく。

このモデルに基づく通常の IRT アルゴリズムでは，正答/不正答のバイナリデータ行列が出力される尤度が最大化されるよう，項目母数の集合 $\{(a_i, b_i)\}$ と全 LLM の能力の集合 $\{\theta_k\}$ を交互に推定し，収束まで繰り返す。一方，2 章の通り，LLM に選択肢のみ回答させ，正解選択肢の出力確率を評価し，正答確率のデータ列 $(p_{ki})(1 \le k \le M, 1 \le i \le N)$ が得られたとする。なお，評価する LLM の数を M，問題数を N とする。このとき，以下の 2 式の通り，最小二乗法に基づき $\{(a_i, b_i)\}$ と $\{\theta_k\}$ を交互に繰り返し推定することで，より容易に各パラメータを求められる。

$$\theta_k^{new} = \arg\min_{\theta_k \in \mathbb{R}} \left(\sum_{i=1}^{M} |p_{ki} - p(\theta_k, a_i, b_i)|^2 \right) \tag{3}$$

$$(a_i^{\text{new}}, b_i^{\text{new}}) = \arg\min_{\substack{a_i > 0 \\ b_i \in \mathbb{R}}} \sum_{k=1}^{N} \left| p_{ki} - p(\theta_k, a_i, b_i) \right|^2 \tag{4}$$

3.2 シミュレーションによる基礎性能評価

本シミュレーションでは，$\{(a_i, b_i)\}$ と $\{\theta_k\}$ を乱数で生成して正解データとし，2 母数ロジスティックモデルに従った確率データ列を生成し，3.1 説で提案したアルゴリズムで推定された $\{(a_i, b_i)\}$ と $\{\theta_k\}$ が，正解データに一致するか，また M と N により，各パラメータの標準誤差がどのように変化するかを検証した。

結果は，図 2 にまとめた通りであり，

- 本アルゴリズムで，$M = 100, N = 100$ の規模でも正しく各パラメータを推定できる
- 理論通り，a と b が $\sqrt{N}$ に，θ が $\sqrt{M}$ に反比例する

という特徴が確認された。特に，人間を対象として得られるバイナリデータに基づく IRT の場合に，a の推定でのサンプルサイズは 1 万が望ましいとされることもある中 [文部科学省 22]，$N = 100$ でも a，b がいずれも正解データをほぼ完璧に再現し，標準誤差も小さく，十分推定値として議論可能な値が得られていることは，非常に

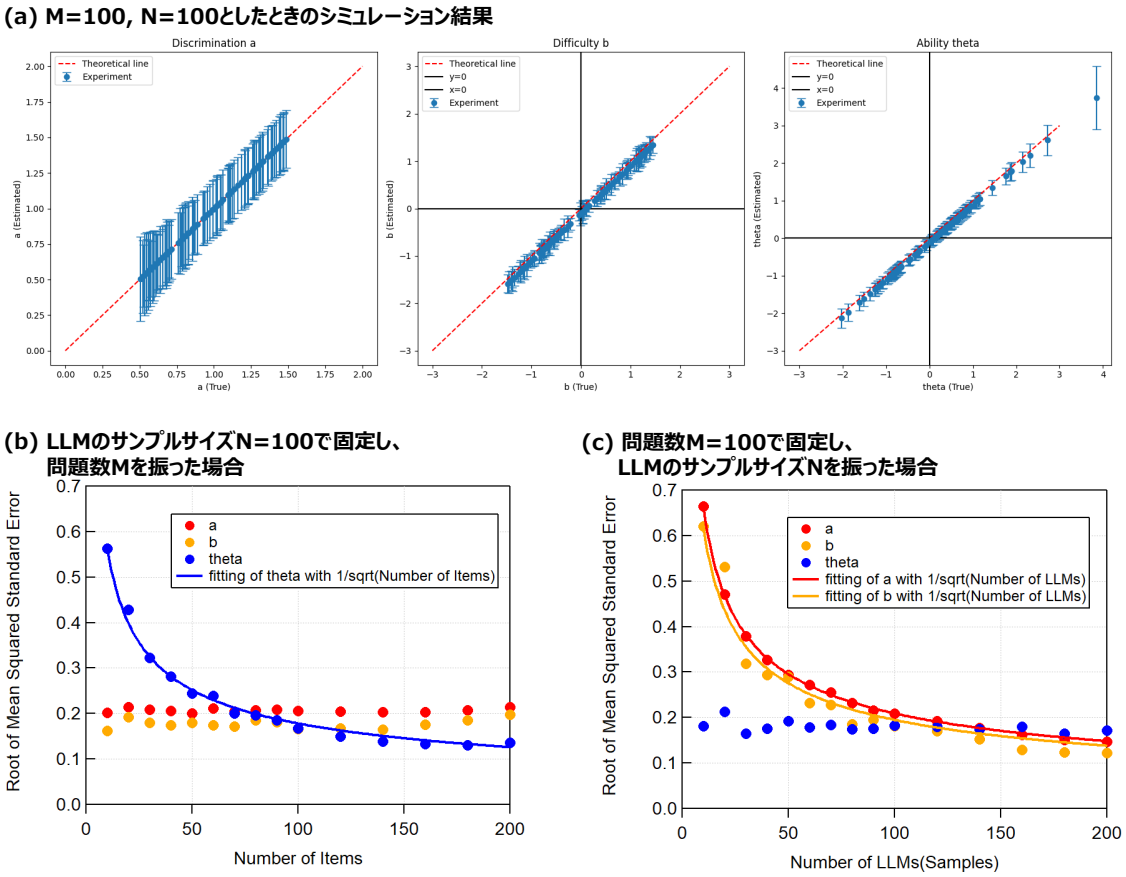

図 2: 3.1 節で提案した新たなアルゴリズムでのシミュレーション。なお，標準誤差はすべて項目情報関数から計算している。(a) 問題数 $M = 100$，評価対象 LLM 数 $N = 100$ のときの結果。IRT に基づき確率から推定しても、各パラメータは正しく推定される。なお，誤差棒は全て標準誤差。また，赤点線は，各パラメータの真の値と，本 IRT アルゴリズムでの推定値が一致することを示す直線であり，各店がこの直線に近いほど，IRT アルゴリズムによる推定での再現性が高い。(b) $N = 100$ として M を変化させた場合のシミュレーション結果。θ の標準誤差の二乗和の平方根が $\sqrt{M}$ に反比例する。(c) $M = 100$ として問題数 N を変化させた場合のシミュレーション結果。a と b の標準誤差の二乗和の平方根が $\sqrt{N}$ に反比例する。

効率の良い分析になっていることが伺える。これは，バイナリデータよりも連続値データが，1 点あたりで多くの情報量を持っていることに由来する。つまり，LLM の能力評価の分析では，人間を対象とする IRT 分析に比べ，LLM のサンプルサイズが稼げなくとも，本手法を用いることで，十分に統計的信頼性の高い結果が期待される。

4　政策研究領域での LLM の能力評価の試行

ここからは、実際に前章のアルゴリズムを用いて，実際に政策研究領域での LLM の能力評価を試行する。

4.1　実験設定

本研究では，文献 [高山 25] の博士課程進学に関する因果推論に LLM を活用する場合を想定し，各変数に関連する知識を問うものを中心に，20 問作成した。表 1 には代表的な問題を 2 題示している。

また，評価する LLM は，OpenAI が Chat Completion 機能を提供する API のうち，表 2 に示す 12 のモデルを対象とした。これらはいずれも，API から log-probability が出力可能であり，正解選択肢の出力確率を計算できる。

また図 3 には，この問題数 ($M = 20$)・LLM の数 ($N = 12$) で，3 章で提案したアルゴリズムのシミュレーションを行った結果を示す。この M と N では標準誤差も大きく，図 2(a) に示すような信頼性までは担保できない。さらに，各パラメータの真の値と，本 IRT アルゴリズムで推定した値は，赤点線と傾きが異なり，系統的なずれの発生も示唆される。しかし，推定値でも問題間の識別力・困難度の逆転や，サンプル間の能力の逆転は起きていない。よって，大局的な動向の把握は可能と考え，LLM の能力評価の試行をそのまま実践することにした。

なお今回は，LLM の温度パラメータは $T = 1.0$ で固定した。また，LLM の log-probability は 5 回測定し，そ

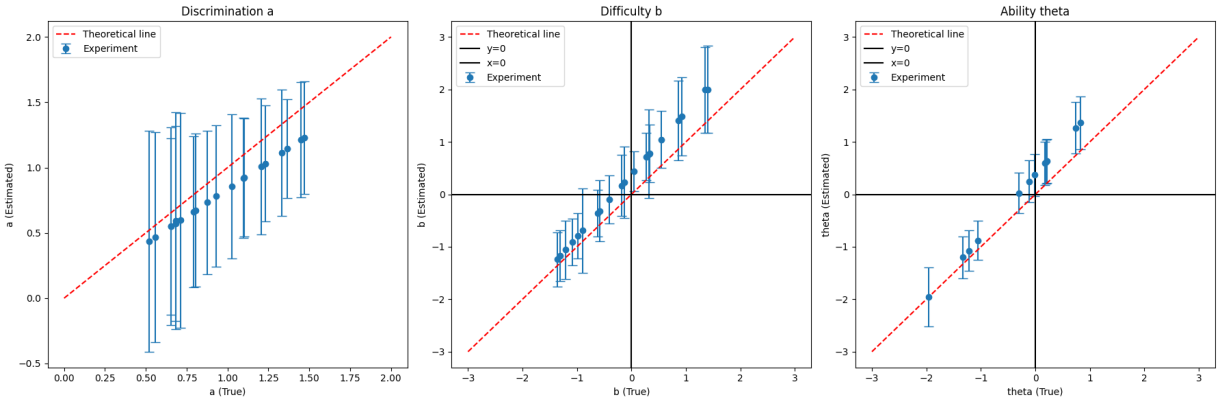

図 3: 本研究の政策研究ベンチマークテストの条件 ($M = 20, N = 12$) でのシミュレーション結果。また，赤点線は，各パラメータの真の値と，本 IRT アルゴリズムでの推定値が一致することを示す直線である。

の平均値で確率データ列を構成した。

表 1: 本研究で実際作成し，LLM の能力評価に用いた科学技術・イノベーション政策研究領域の問題例。

問題例1

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うち、成果を中心とする実績状況に基づく配分について、文部科学省にて定められている配分指標として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. 博士号の授与状況
B. 卒業・修了者の進学等の状況
C. 常勤教員あたり科研費獲得額・件数
D. 常勤教員によるスタートアップ創出数
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。

問題例2

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。

表 2: 本研究で作成した 20 問で測定対象とした LLM と，IRT 分析で能力推定した結果。なお，能力は全て標準化された値。

番号	能力測定の対象としたLLMモデル	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)

4.2 実験結果・考察

各問題に対し，LLM の正答確率を計算し，実際に IRT 分析で推定された能力 θ とその標準誤差を，表 2 に示した。これを見ると，概ね性能は gpt-3.5 系，gpt-4 系（≃ gpt-4o-mini），gpt-4o 系と値が上がっていくことが分かり，実際の OpenAI の世代・モデルスペックの優劣と概ね一致する。また，標準誤差はおおむね 0.3～0.4 程度に収まっており，図 3 における能力 θ の標準誤差と同程度であることにも鑑みると，測定精度を悪化させる外的要因も特段存在していないと考えられる。

これらの結果は，20 問という限られた問題数ではあるが，本研究で提案した LLM の政策研究の能力評価の手法と問題セットが、ある程度有効に機能することを実証しているものと考えられる。今後，これらを原型に，適切な識別力・困難度の問題を中心に構成されるよう改善を図り，さらに問題数自体も増やし，さらに測定対象とする LLM を増やすことができれば，日本の科学技術・イノベーション政策研究に関する LLM の能力を，より精緻に測ることが可能になると期待される。

5 おわりに

5.1 本稿での成果のまとめ

本稿では，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て，まず 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が求められることや，その際の課題について指摘した。さらに，活用基準を，IRT を援用し，人間との比較に基づき，定めていく可能性について提案した。

また，LLM の能力評価に特化し，LLM の出力トークンの log-probability を基に構築できる確率データ列に適用することで，少ないサンプルサイズでも統計的信頼性の高い分析結果が得られる，新たな IRT 分析用のアルゴリズムについても考案した。

そして，実際に政策研究の多肢選択式問題を 20 問を

試作し，OpenAI 製の 12 の LLM を対象に試行実験行った。その結果，それらの世代・スペックの優劣と大きく矛盾しない結果が得られ，本研究で提示した LLM の政策研究の能力評価の手法は，有効に機能することがと実証された。

5.2 本研究の限界と今後の課題

■研究対象とする LLM の拡大 本稿では OpenAI 製の LLM のみを対象としたが，各問題の識別力・困難度パラメータを正しく推定したうえで，能力推定をするためには，より多くのサンプルサイズが必要である。そのためには，Gemini，Llama，Qwen や日本語特化型 LLM[LLM-jp24] 等，log-probability の出力が API で可能な LLM を，今後サンプルに加えていくことが望ましい。

■LLM の出力確率の安定性と 2 母数ロジスティックモデルでの回帰の妥当性 LLM の出力は，式（1）に従うので，特に正答/不正答の 2 種類だけで考えれば，確率の式構造としてはロジスティックモデルに近い。一方，正答を出力するロジット z は，IRT で測定される θ とどのような関係にあるかは，自明ではない。また，解釈の簡便性から，今回は 2 母数ロジスティックモデルを採用したが，母数の数の吟味や多次元性の考慮等も，厳密には別途検討が必要である。

■政策研究ドメインの専門家を対象とした調査による LLM の能力基準の設計に関する課題 2.4 節で述べた通り，LLM の活用基準を人間の能力をもとに決めるには，人間の政策研究ドメインの専門家の理解・協力を経て，調査を進める必要がある。しかしその際，

- どの範囲の，どのステークホルダーを対象とするか
- 統計的にも，どの程度の母集団を確保するか

は，綿密な検討が必要である。また，その回答情報・分析結果は機微であり，取り扱いに配慮が必要である。さらに，人間を対象とした分析結果と LLM の結果をどう接続し，比較可能にするかは課題となりうる。

本研究で作成したコード・問題文や得られたデータの公開

実際に開発した IRT 分析アルゴリズムのコードや，作成した 20 問の問題文と，今回の試行で推定された識別力・困難度パラメータ，各 LLM の問題ごとの正答確率データ列は，全て以下に公開している。

https://github.com/mas-takayama/IRT-Measurement-of-LLMs-capability-in-Political-Science/tree/main/JSRPIM40th%202B22

なお，講演では，特徴的な問題と，LLM の実際の反応についても，時間の限り紹介する。

謝辞

本研究の一部は，JST，CREST，JPMJCR22D2 の支援を受けて実施した。また，本稿の執筆には、NISTEP の小柴 等博士から多くの助言を頂いた。

参考文献

[OpenAI23] OpenAI: GPT-4 Technical Report. arXiv:2303.08774v3, 2023. (preprint) https://doi.org/10.48550/arXiv.2303.08774

[Wei22] Jason Wei Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le and Denny Zhou：Chain of Thought Prompting Elicits Reasoning in Large Language Models, Advances in Neural Information Processing Systems, 2022. https://openreview.net/forum?id=_VjQlMeSB_J

[OpenAI25] OpenAI: GPT-5 System Card, 2025. https://cdn.openai.com/gpt-5-system-card.pdf

[高山 23] 高山正行, 小柴等, 三内顕義, 清水昌平：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. **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 2D21, 2023.

[Takayama25] Masayuki Takayama and Tadahisa Okuda and Thong Pham and Tatsuyoshi Ikenoue and Shingo Fukuma and Shohei Shimizu and Akiyoshi Sannai：Integrating Large Language Models in Causal Discovery: A Statistical Causal Approach. Transactions on Machine Learning Research, 2025. https://openreview.net/forum?id=Reh1S8rxfh

[高山 25] 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：博士課程進学に関する統計的因果探索の非線形効果の可視化. **研究イノベーション学会 第 *40* 回年次学術大会（予稿集）**, 公演番号 2B03, 2025.

[Hendrycks21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song and Jacob Steinhardt：Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. https://openreview.net/pdf?id=d7KBjmI3GmQ

[秋吉 21] 秋吉 貴雄： 政策科学の展開と変容： 総合政策学への示唆. *Keio SFC journal*, 21(1):20-40, 2021. https://gakkai.sfc.keio.ac.jp/journal/.assets/SFCJ21-1-02.pdf

[澁谷 20] 澁谷 拓巳：連続反応を対象とするベータ反応モデルの拡張と EM アルゴリズムによる母数推定. 教育心理学研究, 68(4):373-387, 2020. https://doi.org/10.5926/jjep.68.373

[文部科学省 22] 文部科学省　総合教育政策局調査企画課学力調査室：令和３年度『全国学力・学習状況調査』経年変化分析調査　テクニカルレポート. 2022. https://www.nier.go.jp/21chousakekkahoukoku/kannren_chousa/pdf/21keinen_report.pdf

[Shimizu06] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen： A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003-2030, 2006. https://jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf

[Shimizu11] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer and K. Bollen.： DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr): 1225–1248, 2011. https://www.jmlr.org/papers/volume12/shimizu11a/shimizu11a.pdf

[LLM-jp24] LLM-jp： LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. 2407.03963, 2024. (preprint) https://arxiv.org/abs/2407.03963