LLM の知能は人間と同様の構造を持つか？：人間の知能理論を用いた評価スコアの統計分析

Namgi Han[1,2]　九門 涼真[1,2]　谷中 瞳[1,2,3]
[1] 東京大学　[2] 理化学研究所　[3] 東北大学
{hng88,kumoryo9,hyanaka}@is.s.u-tokyo.ac.jp

概要

言語モデルの評価ベンチマークの充実に伴い，言語モデルの評価結果に対し適切な解釈を与えることの重要性が増している．本研究では，言語モデルの評価結果を人間の知能理論に基づいた統計モデルで説明できるか検証を行う．具体的には，人間の知能体系と学習段階に関する Cattell-Horn-Carroll 理論と Bloom の体系に加えて，評価スコアに基づく統計モデルを用いた確認的因子分析を行う．観測変数として，88 の言語モデルと 54 の評価タスクからなる評価スコアを用いる．実験の結果，統計モデルと評価タスクの適切な調整によって統計モデルによる説明を与えることができたが，統計モデルのさらなる探索の必要性が示唆された．

1　はじめに

近年の言語モデルの発展とともに，言語モデルの言語処理能力を様々な観点から評価することの重要性が増している．それに伴い，多くの評価ベンチマークが迅速に開発されている[1)]．しかしながら，評価ベンチマークごとに評価対象としている能力が異なるにもかかわらず，実際に言語モデルを評価して得られた各評価ベンチマークの評価スコアには高い相関が見られることが報告されている [1, 2]．そのため，各評価ベンチマークによる言語モデルの評価結果が，実際に言語モデルのどのような能力を反映しているのかについて分析が求められている．

これまでに言語モデルの評価結果に対して統計分析を行い，評価結果から言語モデルの知能体系を構造化する試みが行われている [3, 4]．また，人間の知能能力や認知能力の学習段階に関する理論を用いて言語モデルを分析する取り組みもある [5, 6]．しかし，これらの先行研究で議論されているように，人間に対する知能体系の理論を，そのまま言語モデルに適用することができるのかについては十分な検証が要求される．また，評価スコア間の高い相関が示唆するように，評価ベンチマークが人間の直感に反する評価をしている可能性も排除できない．

そこで本研究では，言語モデルの評価結果を人間の知能に関する理論に基づいて統計的に説明できるか，複数の理論を用いて確認的因子分析を行う．さらに，評価スコアに基づいた統計モデルも含めてどの統計モデルがもっともらしいか比較する．言語モデルの評価結果は 88 個の言語モデルと 54 個の評価タスクを用いて算出する．実験の結果，人間の知能理論をそのまま適用した統計モデルは言語モデルの評価結果を十分に説明できないが，統計モデルと評価タスクを適切に調整することによって有意味な説明が得られることが示唆された．

2　関連研究

2.1　人間の知能理論と LLM の評価

言語モデルの発展につれて，人間の知能体系を用いて言語モデルが持つ能力を分析しようとする試みが行われてきた．一つの例として，人間の知能を説明する理論である Cattell-Horn-Carroll 理論 [7]（以下，CHC 理論）が用いた研究がある．Tolan ら [8] は当時の機械翻訳や画像認識などのタスクの評価対象を，CHC 理論で提案されている人間の知能として分類しようと試みた．Ilić & Gignac [5] は 591 個の言語モデルを用いて，評価ベンチマークのスコアが CHC 理論で提案されている人間の知能体系で説明できるかを統計的に分析した．Zhou ら [9] は CHC 理論を参考にした評価基準と，それに従う評価ベンチマークを新たに提案した．CHC 理論の他にも，人間の学習段階を定義した Bloom の体系 [10] を用いた試みも提案された．Huber & Niklaus [6] は既存の評価

1) 一例として，HuggingFace の評価ツールである lighteval は一千以上の評価タスクをサポートしている．

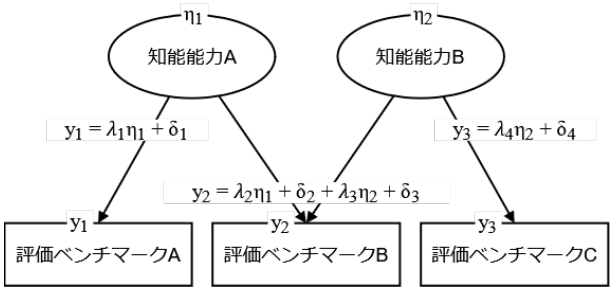

図 1　確認的因子分析で使われる CFA モデルの例．λ は負荷量，δ は誤差項を意味する．

ベンチマークを Bloom の体系に合わせて分類し，10 個の言語モデルでの評価結果を分析している．

しかし，これらの議論は，人間の知能理論を言語モデルの評価にそのまま適用できるという前提の上でなされていることが多かった．また，統計モデルによる検証が行われている研究でも，使われている言語モデルが古く評価ベンチマークを十分に解けていない恐れがあるか [5]，そもそも言語モデルの数が少ない場合 [6] があるなど，さらなる検証の余地が残されていた．

2.2　確認的因子分析

確認的因子分析（Confirmatory Factor Analysis，以下，CFA）とは，観測できたデータを観測変数として用い，観測変数の間の関係性，および，その観測変数と関係があると思われるが直接の観測ができない潜在変数を推定するために考案された統計分析の手法である [11]．確認的因子分析では，まず観測変数と潜在変数の関係性を定義したモデル（以下，CFA モデル）を，研究者が研究仮説として提案する．

CFA モデルの例として図 1 を参照されたい．例えば，図 1 の知能能力 A は評価ベンチマーク A と評価ベンチマーク B の評価スコアに影響を与えると仮定される．この影響，すなわち CFA モデルで変数の間の関係性は，一般的に線形回帰式で表現される．そのため CFA モデルの検証は，実際の観測変数を CFA モデルに当てはめて，関係性として定義されている線形回帰式を推定することで行われる．図 1 の例では，評価ベンチマーク A，B，C の評価スコアを用いて，知能能力 A，B と変数の間の回帰式が推定される．

また，研究者がモデルを提案せず，クラスタリングなどの手法を用いて変数の間の関係性を探索する手法として，探索的因子分析（Exploratory Factor Analysis，以下，EFA）があげられる．EFA は研究者が観測変数と潜在変数に対して，先験的な CFA モデルを提案できない，またはしない場合に用いられる手法である．EFA には様々な手法が用いられるが，代表的なものとしてクラスタリング [12] や主成分分析 [13] などを応用して，観測変数から潜在変数を探索する手法が存在する．

表 1　統計分析に用いた言語モデルの統計．

カテゴリー	ラベル	モデル数
パラメータ数	< 10B	57
	> 10B	31
アーキテクチャ	Llama	12
	Qwen	28
	Gemma	20
	Phi	7
	Mistral	6
	Granite	15
モデルタイプ	Pretrained	34
	Instruction-tuned	54

3　実験設定

3.1　評価モデル

本研究では HuggingFace[2] からのダウンロード数を参考に，広く使われているオープン言語モデルを評価対象のモデルとした．その結果，88 個の言語モデルを評価ベンチマークでの評価に用いた．評価に用いた言語モデルの詳細は表 1 を参照されたい．

3.2　評価ベンチマーク

本研究では先行研究の中で最も多数の評価ベンチマークをもとに言語モデルの分析を行っている文献 [6] に従い，AGIEval [14]，ARC-Challenge [15]，BIG-Bench Hard [16]，DROP [17]，GPQA [18]，GSM8K [19]，HumanEval [20]，Winogrande [21] を評価ベンチマークとして用いる[3]．AGIEval と BIG-Bench Hard の場合，一般的にはサブタスクの評価スコアの平均が評価ベンチマークの代表スコアとして使われるが，本研究では先行研究 [6] に従い，平均スコアの代わりに，全てのサブタスクをそれぞれ一つのタスクとして扱う．その結果として，本研究では AGIEval からは 21 個，BIG-Bench Hard からは 27 個のタスクの評価結果を，それぞれ一つの評価タスクの評価スコアして扱うことで，他の 6 個の評価ベンチマークと合わせて，全部で 54 個の評価タスクを用いる．実際

2）https://huggingface.co/models
3）例外として，MATH は現在，著作権の問題で公開が中止されているため，本研究では用いない．

表 2　確認的因子分析の結果.

CFA モデル	CFI	GFI	TLI	RMSEA
Bloom モデル	0.459	0.416	0.427	0.236
Bloom モデル（AGIEval のみ）	0.983	0.973	0.968	0.138
Bloom モデル（BIG-Bench Hard のみ）	0.490	0.451	0.453	0.251
CHC モデル	0.462	0.412	0.439	0.219
CHC モデル（AGIEval のみ）	0.744	0.715	0.709	0.264
CHC モデル（BIG-Bench Hard のみ）	0.636	0.597	0.601	0.244
EFA モデル	0.757	0.678	0.736	0.150
EFA モデル（AGIEval のみ）	0.859	0.827	0.837	0.199
EFA モデル（BIG-Bench Hard のみ）	0.800	0.752	0.770	0.186
EFA モデル（平均）	0.984	0.970	0.958	0.112

の評価には，再現性を担保するため，言語モデルの公開評価ツールである Language Model Evaluation Harness [22] を用いる．

3.3　モデルと適合度の評価

本研究では言語モデルの各評価タスクにおける評価スコアを観測変数とみなし，それぞれの評価タスクが評価している能力として潜在変数を定義した CFA モデルを検証対象とする．CFA モデルとして，主に三つのモデルを用意する．一つ目は，Bloom の体系を用いた先行研究 [6] に基づいた Bloom モデルを用意する．Bloom モデルでは，先行研究に従った評価タスクと人間の知能能力の対応づけを扱う．二つ目として，CHC 理論に基づいて，評価タスクを人間の知能能力に紐づけた CHC モデルを用意する．本研究では，CHC 理論を用いた先行研究 [5, 8] では扱っていない評価タスクも用いているため，先行研究を参考にした上で，新たに知能能力への紐づけを行う．三つ目として，人間の知能能力に対する既存の理論を用いず，探索的因子分析を行って EFA モデルを設計する．探索的因子分析には OPTICS クラスタリングアルゴリズム [12] を用いる．それぞれの CFA モデルの詳細に関しては付録 A を参照されたい．

実際のデータがうまく CFA モデルに当てはまるかを判断する適合度の指標 [23] として，本研究では Comparative Fit Index（CFI），Goodness-of-Fit Index（GFI），Tucker-Lewis Index（TLI），Root Mean Square Error of Approximation（RMSEA）を用いた．指標の解釈はサンプルのサイズ，モデルのパラメータ数などによって変わってくるが，一般的には CFI，GFI，TLI が 0.95 以上，RMSEA が 0.06 以下だと許容範囲内の CFA モデルと解釈される．適合度の指標とその解釈の詳細については星野ら [24] を参照されたい．

本研究の統計分析には，再現性を担保するため公開ツールである semopy [25, 26] を用いる．

4　結果と分析

CFA 分析を行った結果を表 2 に示す．3.3 節の評価基準から，本研究で検証した Bloom モデル，CHC モデルはいずれも実際の評価スコアを説明するには不適切であることがわかる．この結果は，人間の知能に対する既存の理論を，言語モデルの能力に適用するためには何らかの調整が必要であることを示唆する．実際に，先行研究 [5] でも，CHC 理論を大幅に修正して構築した CFA モデルを統計的に有意味なモデルとして報告している．

また，既存の知能理論を使わない EFA モデルでも，許容範囲よりも悪い適合度が算出された．この結果から，観測変数，つまり各評価タスクの評価スコアには何らかの問題があることが疑われる．例えば，本研究の実験設定では AGIEval と BIG-Bench Hard からのタスクの数が非常に多いため，評価スコアの分布に，評価ベンチマーク固有の特徴に由来した問題が存在する可能性が考えられる．

そこで追加実験として，我々は CFA モデルの修正を行い再検証を試みた．CFA モデルの修正には以下の仮説を用いた．

- AGIEval と BIG-Bench Hard だけを使った場合の適合度を確認するべきではないか？
- EFA モデルの場合，AGIEval と BIG-Bench Hard は平均スコアだけを用いるべきではないか？

修正した CFA モデルを用いた検証結果も表 2 に示

されている．結果として，どの CFA モデルも許容範囲の適合度を示さないことがわかる．特に RMSEA は一番小さい場合でも 0.112 になっており，基準値の 0.06 には程遠い．しかし，先行研究 [27] ではサンプルの数が十分でない場合，RMSEA での検証が第一種過誤を起こす恐れがあると報告されている．本研究ではサンプルの数が 88 で，先行研究で指摘している 200 以下のケースに含まれる．従って，ここは CFI，GFI，TLI が 0.95 以上になっている AGIEval のみの Bloom モデルと，AGIEval と BIG-Bench Hard の平均スコアを用いた EFA モデルを，暫定的に許容範囲内の統計モデルとして扱い，そこから得られた知見を議論する．

まず，AGIEval と BIG-Bench Hard に限定して統計分析を行った結果，Bloom モデル，CHC モデル，EFA モデルで CFI，GFI，TLI が向上する傾向が見られる．中でも Bloom モデルは，AGIEval のサブタスクだけを使う CFA モデルがもっとも評価スコアを説明できるという結果を示した．この結果から，AGIEval と BIG-Bench Hard 以外の評価ベンチマークが，CFA モデルの中でノイズになっている可能性が考えられる．また，AGIEval と BIG-Bench Hard を共に用いることも，同じく CFA モデルに悪い影響を与える可能性が示唆される．

最後に，AGIEval と BIG-Bench Hard の平均スコアだけを用いた EFA モデルは CFI，GFI，TLI が 0.95 以上となっており，AGIEval のみの Bloom モデルと同じく許容範囲内の統計モデルとなった．評価スコアを確認したところ，AGIEval と BIG-Bench Hard のサブタスクの中に言語モデルが全く解けていないタスクが多数存在することが判明した．このことから，AGIEval と BIG-Bench Hard 全体の平均スコアだけを用いることによって，全く解けていないサブタスクの影響が軽減されたことが示唆される．

図 2 にて，平均スコアだけを用いた EFA モデルのダイアグラムを示す．評価スコアから推定された潜在変数の eta1，eta2，eta3 は，各潜在変数に関係づけられている評価ベンチマークの内容から，それぞれ「特定のドメインに対する知識・問題解決能力」，「推論能力」，「一般常識・世界知識に対する能力」と解釈することができる．これらの潜在変数は，付録 A にある CHC モデルの「専門分野知識」，「推論能力」，「読解能力」に類似していると考えられる．今後の課題として，EFA モデルと CHC 理論の統合による言語モデルの知能構造に対する分析が求められる．

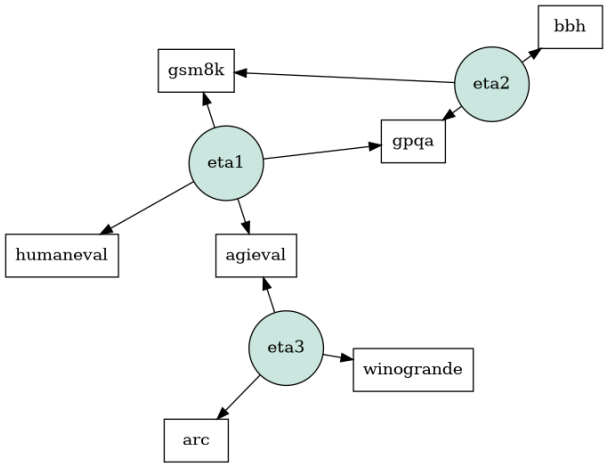

図 2　EFA モデル（平均）のダイアグラム．観測変数の評価スコアは，対応する評価ベンチマークの名前の略語で表記している．

5　おわりに

本研究では人間の知能能力に対する理論を用いて，言語モデルの評価結果を説明する統計モデルの構築を試みた．88 件の言語モデルと 54 個の評価タスクを用いて得られた評価スコアを観測変数として，人間の知能能力に対する理論である CHC 理論と，人間の認知能力の学習段階に対する理論である Bloom の体系と，評価スコアに基づいて潜在変数を推定した結果を用いて設計した確認的因子分析を行った．その結果，既存研究の理論をそのまま用いた統計モデルは，言語モデルの評価結果の説明として統計的に有意でないということがわかった．そこで，評価タスクの制限と評価スコアの調整を行い統計モデルを修正することで，統計的に有意とされる CFA モデルを構築することができた．また，言語モデルの評価スコアを統計的に説明できる二つの CFA モデルは，それぞれ Bloom の体系と CHC 理論に類似することが判明した．

本研究の結果から，人間の知能理論を用いて言語モデルの知能体系を分析するためには，評価ベンチマークに対する適切な統制が必要とされるという知見が得られた．本研究の課題として，サンプルとなる言語モデルの数のさらなる担保，許容範囲内と確認された統計モデルに更なる評価タスクを取り入れた拡張などがあげられる．今後の取り組みとして，より多くのサンプルと選別された評価タスクを用いた堅牢な統計モデルの開発と，その統計モデルを用いて言語モデルの知能体系を解析する評価フレームワークの提案を目指す．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」，JST CREST　JPMJCR2565 の支援を受けたものである．また，本研究の成果の一部は，産総研及び AIST Solutions が提供する ABCI 3.0 を利用して得られたものである．

参考文献

[1] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

[2] Namgi Han, 岡本拓己, 石田茂樹, 林俊宏, Akim Mousterou, Bowen Chen, 宮尾祐介. 下流タスクでの日本語事前学習モデルの性別バイアスの評価. 言語処理学会第 31 回年次大会, 2025.

[3] Ryan Burnell, Han Hao, Andrew RA Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities. **arXiv preprint arXiv:2306.10062**, 2023.

[4] Denis Federiakin. Improving llm leaderboards with psychometrical methodology. **arXiv preprint arXiv:2501.17200**, 2025.

[5] David Ilić and Gilles E Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? **Intelligence**, Vol. 106, p. 101858, 2024.

[6] Thomas Huber and Christina Niklaus. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 5211–5246, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[7] John Bissell Carroll. **Human cognitive abilities: A survey of factor-analytic studies**. No. 1. Cambridge university press, 1993.

[8] Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. **Journal of Artificial Intelligence Research**, Vol. 71, pp. 191–236, 2021.

[9] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, et al. General scales unlock ai evaluation with explanatory and predictive power. **arXiv preprint arXiv:2503.06378**, 2025.

[10] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. Handbook i: cognitive domain. **New York: David McKay**, pp. 483–498, 1956.

[11] Karl G Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. **Psychometrika**, Vol. 34, No. 2, pp. 183–202, 1969.

[12] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. **ACM Sigmod record**, Vol. 28, No. 2, pp. 49–60, 1999.

[13] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. **Journal of computational and graphical statistics**, Vol. 15, No. 2, pp. 265–286, 2006.

[14] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. In **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2299–2314, 2024.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. **ArXiv**, Vol. abs/1803.05457, , 2018.

[16] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[18] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. **arXiv preprint arXiv:2311.12022**, 2023.

[19] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. **arXiv preprint arXiv:2110.14168**, 2021.

[20] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mo Bavarian, Clemens Winter, Phil Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. **ArXiv**, Vol. abs/2107.03374, , 2021.

[21] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. **arXiv preprint arXiv:1907.10641**, 2019.

[22] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

[23] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. **Structural equation modeling: a multidisciplinary journal**, Vol. 6, No. 1, pp. 1–55, 1999.

[24] 星野崇宏, 岡田謙介, 前田忠彦. 構造方程式モデリングにおける適合度指標とモデル改善について：展望とシミュレーション研究による新たな知見. 行動計量学, Vol. 32, No. 2, pp. 209–235, 2005.

[25] Anna A. Igolkina and Georgy Meshcheryakov. semopy: A python package for structural equation modeling. **Structural Equation Modeling: A Multidisciplinary Journal**, Vol. 0, No. 0, pp. 1–12, 2020.

[26] Georgy Meshcheryakov, Anna A Igolkina, and Maria G Samsonova. semopy 2: A structural equation modeling package with random effects in python. **arXiv preprint arXiv:2106.01140**, 2021.

[27] David A Kenny, Burcu Kaniskan, and D Betsy McCoach. The performance of rmsea in models with small degrees of freedom. **Sociological methods & research**, Vol. 44, No. 3, pp. 486–507, 2015.

A 付録

評価タスクの名前は評価ツールの表記に従って表記されていることに注意されたい.

A.1 Bloom モデルの詳細

潜在変数	定義	評価タスク
Remember	記憶	bbh object counting
Understand	理解	agieval gaokao english, agieval sat en, arc challenge, bbh disambiguation qa, bbh geometric shapes, bbh hyperbaton, bbh navigate, bbh penguins in a table, bbh ruin names, drop, winogrande
Apply	適用	agieval aqua rat, agieval math, agieval sat math, bbh boolean expressions, bbh date understanding, bbh dyck languages, bbh multistep arithmetic two, bbh reasoning about colored objects, bbh temporal sequences, bbh tracking shuffled objects five objects, bbh tracking shuffled objects seven objects, bbh tracking shuffled objects three objects, bbh word sorting, gsm8k, humaneval
Analyze	分析	agieval lsat lr, agieval lsat ar, bbh web of lies, gpqa, agieval logiqa en, agieval lsat rc, bbh causal judgement, bbh formal fallacies, bbh logical deduction five objects, bbh logical deduction seven objects, bbh logical deduction three objects, bbh movie recommendation, bbh salient translation error detection, bbh snarks, bbh sports understanding

A.2 CHC モデルの詳細

潜在変数	定義	評価タスク
G_f	推論能力	agieval logiqa en, agieval logiqa zh, agieval lsat ar, agieval lsat lr, agieval sat en, agieval sat en without passage, arc challenge, bbh boolean expressions, bbh causal judgement, bbh dyck languages, bbh formal fallacies, bbh geometric shapes, bbh logical deduction five objects, bbh logical deduction seven objects, bbh logical deduction three objects, bbh movie recommendation, bbh navigate, bbh penguins in a table, bbh ruin names, bbh tracking shuffled objects five objects, bbh tracking shuffled objects seven objects, bbh tracking shuffled objects three objects, bbh web of lies
G_{kn}	専門分野知識	humaneval, agieval aqua rat, agieval gaokao biology, agieval gaokao chemistry, agieval gaokao chinese, agieval gaokao english, agieval gaokao geography, agieval gaokao physics, agieval jec qa ca, agieval jec qa kd, bbh hyperbaton, bbh salient translation error detection, gpqa
G_q	定量的な知識	agieval gaokao mathcloze, agieval gaokao mathqa, agieval math, agieval sat math, bbh date understanding, bbh multistep arithmetic two, bbh object counting, gsm8k
G_{rw}	読解能力	agieval gaokao history, agieval lsat rc, bbh disambiguation qa, bbh snarks, bbh sports understanding, bbh temporal sequences, bbh word sorting, drop, winogrande

A.3 EFA モデルの詳細

潜在変数	評価タスク
eta1	agieval gaokao history, agieval gaokao geography, humaneval, agieval gaokao mathqa, agieval lsat ar, agieval gaokao biology, agieval gaokao chemistry, agieval gaokao physics, agieval logiqa zh, agieval logiqa en, bbh penguins in a table, gpqa
eta2	agieval jec qa ca, agieval jec qa kd, agieval gaokao mathcloze, agieval math, drop, agieval gaokao chinese, gsm8k
eta4	agieval sat en, agieval gaokao english, agieval sat math, agieval gaokao physics, agieval sat en without passage, agieval gaokao chinese, agieval aqua rat, agieval jec qa kd, agieval jec qa ca, agieval lsat rc, gsm8k, bbh salient translation error detection, bbh logical deduction three objects, bbh reasoning about colored objects, bbh date understanding, bbh multistep arithmetic two
eta5	arc challenge, winogrande, bbh word sorting
eta6	bbh tracking shuffled objects five objects, bbh tracking shuffled objects seven objects, bbh logical deduction five objects, bbh ruin names, bbh logical deduction seven objects, bbh tracking shuffled objects three objects, bbh dyck languages
eta7	bbh web of lies, bbh navigate, bbh boolean expressions, bbh temporal sequences, gsm8k, bbh geometric shapes, bbh object counting, bbh logical deduction three objects, bbh formal fallacies, bbh movie recommendation, bbh word sorting, bbh snarks, bbh salient translation error detection, humaneval, bbh logical deduction five objects
eta8	bbh hyperbaton, bbh disambiguation qa, bbh multistep arithmetic two, gpqa, bbh reasoning about colored objects, bbh ruin names
eta9	bbh reasoning about colored objects, bbh date understanding, bbh penguins in a table, bbh sports understanding, bbh causal judgement, agieval sat en, agieval logiqa zh, agieval gaokao english, agieval math, agieval gaokao mathcloze, agieval lsat rc, winogrande, bbh logical deduction three objects, bbh object counting, gsm8k, bbh temporal sequences, bbh snarks, bbh word sorting, bbh movie recommendation, bbh tracking shuffled objects three objects