多肢選択問題における言語モデルの頑健性の評価

滝沢広央 [1]　菅原朔 [2]　相澤彰子 [1,2]
[1] 総合研究大学院大学　[2] 国立情報学研究所
{takizawa,saku,aizawa}@nii.ac.jp

概要

大規模言語モデルの一般常識や推論能力，医療など特定ドメインの知識等の計測には，多肢選択問題からなる評価データセットが利用されることが多い．しかし，言語モデルが多肢選択問題に回答する際の頑健性はそれほど評価されてない．

そこで本研究では，多肢選択問題で構成された既存の評価データセットを変換することで頑健性を検証するための新たな評価データセットを提案し，既存の言語モデルを評価する．結果として否定を含む問題に対する頑健性の低さが明らかになった．

1　はじめに

OpenAI による ChatGPT の公開以降，大規模言語モデルへの期待が高まっている。これらのモデルの能力を評価するデータセットとして，例えば 57 の科目に関して問題が収集された MMLU [1] や，常識推論に関するデータセットの HellaSwag [2] などがある．これらを含め，評価データセットの多くは多肢選択問題で構成されている．

これらが言語モデルの推論能力や知識を評価するために設計されている一方で，選択肢問題を解くのに前提となる，設問形式や求められている回答の出力フォーマットを理解する能力を言語モデルが十分に持っているかは明らかではない．例えば選択肢問題の一部には「でないもの」を選ぶ問題のように，否定の理解を必要とするものがある．しかし，ある選択肢問題を図 1 のように単純な否定を問うような問題に変形しても，言語モデルは間違えることがある．また言語モデルの選択肢問題に対する回答の頑健性も議論がある．たとえば，Zheng ら [3] は正答選択肢の位置を移動するとパフォーマンスが大きく変化することを明らかにしている．

そこで本研究では，（1） 言語モデルが選択肢問題という形式に対応する能力を持っているか，（2）頑健性があるか，の 2 点に着目して評価するための手法を提案する．（1）については既存のデータセットを変換することで知識が不要なタスクを作成し選択肢問題を解く能力に焦点を当てた問題を作成することで，（2）については既存の問題に対し趣旨を変えない変更を加えた際に言語モデルが回答を変えないことを確認することで評価する．

...(5-shot examples)

What is the option that is **not** A, B, or C?
A. 4d
B. 2d
C. d
D. 1/2d
Answer: 4d

図 1　否定を含む簡単な問題に対し間違った生成を行う例.

実験では MMLU のうち 150 問をもとに，本手法を適用し 3,440 問の評価データセットを作成し，3 つモデルを評価した．結果，否定を含むタスクで比較的エラー率が高くなることがわかった．また選択肢のラベルの参照を含むタスクでエラー率が高くなる可能性が示された．

2　関連研究

2.1　NLP モデルの評価手法

NLP モデルの評価手法に，能力ごとにいくつかのタイプのテストを用いる CheckList [4] がある．CheckList では，特定の能力を測るためのシンプルなテストである Minimum Functionality Test（**MFT**）, 入力に若干の変更を加えた前後でモデルの予測が変わらないことを確認する Invariance Test（**INV**）等のテストが提案されている．しかし，CheckList を多肢選択問題を解く能力に適用した研究は見当たらない．本研究は，CheckList の研究を参考に，多肢選択問題に特化した検証用の評価データセットを作成する．

図 2　選択肢問題の解答過程.

2.2　多肢選択問題におけるバイアス

言語モデルが多肢選択問題を解く際，選択肢のラベルや位置でのバイアス [3] が存在したり，選択肢の並び順を変えると言語モデルが間違える [5] といった報告がある. そのため言語モデルの多肢選択問題における頑健性を評価する必要性は高く，本研究はこうしたバイアスを持つような言語モデルが間違えるような問題を含めている.

3　提案手法

本論文では，多肢選択問題で構成された既存の評価データセットを自動的に変換し，言語モデルが多肢選択問題の形式に対応するのに必要最低限の能力を備えていること，多肢選択問題を解けているなら期待する振る舞いを評価するためのデータセットを作成する手法を提案する．具体的には，選択肢問題の解答過程（3.1 節）に沿ったカテゴリごとに，言語モデルを評価するタスクを作成する．3.2 節で選択肢問題におけるフォーマットについて説明したのちに，3.3 節以降では各カテゴリごとのタスクについて説明する.

3.1　選択肢問題の解答過程

選択肢問題に対応する能力について評価するタスクを作成するうえで，選択肢問題を解く際の過程について次のように分類した（図 2）.

入力文の認識　まず文字列を受け取った際に，その文字列が問題文と選択肢から構成される選択肢問題であることを認識する必要がある.

問題理解　選択肢問題は問題文と選択肢から構成され，それぞれの内容によっていくつかのフォーマット（3.2 節）に分けられるが，解くためにはその問題がどのフォーマットかを理解することが期待される.

回答選択　問題を理解した上で，回答となる選択肢を選ぶ.

回答生成　多くの回答はラベルのみだが，ラベルが ABCD ではなく 1234 の場合，ラベルがないため選択肢の内容を出力する場合のようにいくつかの出力形式が想定される.

フォーマット	問題例
SimpleQ	What is 'malware'? A.　A hacker tool.　　B. ...
Continuation	An oocyte is A.　an unfertilized egg.　　B. ...
Gap-Fill	In Holocene Africa, the __ was replaced by the __. A.　Iberomaurusian culture; Capsian culture B.　...

図 3　フォーマットごとの問題例.

3.2　選択肢問題のフォーマット

選択肢問題にはいくつかのフォーマットが存在する．本研究では以下の 3 つに着目する.

SimpleQ　問題文として完結した質問が与えられ，その回答を選択肢から選ぶ問題.

Continuation　問題文として途切れた文章が与えられ，それに続く内容を選択肢の中から選ぶ問題.

Gap-Fill　1 つないし複数の空欄がある文章が与えられ，空欄に入る文字列の組み合わせを選ぶ問題.

各フォーマットごとの具体例を図 3 に記載する.

3.3　入力文の認識

言語モデルが入力された選択肢問題を解けるなら，入力における問題文や選択肢を正しく認識できていることを期待する．そこで質問・選択肢記憶というタスクを設計した．具体的には，既存の選択肢問題を元に，Repeat the following question without answering it. や Which option is { Option 1 } ?, What is the option A? などの指示に言語モデルが従えるかを確認する.

3.4　問題理解

言語モデルが問題を理解した上で解答しているなら，問題に非本質的な変更が加わっても回答を変更しないことが期待される．そこで問題フォーマットの変更や選択肢のラベルの変更や削除，選択肢の順序の変更を行った上で，変更前後の回答が一致することを確認する.

フォーマット変更　選択肢問題にはいくつかのフォーマットがある（3.2 節）．しかし問題フォーマットの違いに対して言語モデルが頑健であるかは明らかでない．そこでこのタスクでは各問題を別の問題フォーマットに変換し選択肢をシャッフルした上で解かせた際に，変換前後での言語モデルの回答が一致するかを確認する．フォーマット変更後に問

題が成立していること，前後で趣旨が変わっていないことは人手で確認している．

選択肢変更　このデータセットでは選択肢問題のラベルには ABCD のように連続したアルファベットを原則用いている．このタスクでは下記の変更を行った上で，変換前後での回答が一致するかを確認する．

1. 選択肢の順序をシャッフル
2. 選択肢のラベルを英数字 1,2,3,4 に変更
3. 選択肢のラベルをハイフンに変更

3.5　回答選択

ここでは言語モデルが回答選択する際に必要となりうる否定の認識能力について評価する．具体的には 2 タイプの質問文を用いた．1 つは Which option is **not** { Option1 },{ Option2 }, or { Option3 }? と選択肢の内容で選択肢を指定しラベルで回答することを期待する質問，もう 1 つは What is the option that is **not** A, B, or C? のようにラベルで選択肢を指定し，内容で回答することを期待する質問である．

3.6　回答生成

昨今の生成的な言語モデルが多肢選択問題を解くには，回答を生成する能力が必要である．このカテゴリでは，言語モデルが期待する回答形式で出力できるかを評価する．具体的には， Which option is { Option1 }? Please write the letter only. のように回答形式を指定した上で期待する回答を生成するか確認する．回答形式はラベルのみ，内容のみ，ラベルと内容の両方（e.g. A. {Option1}）の 3 つを用いた．

3.7　MFT と INV

2.1 節にて MFT や INV のテストタイプを紹介した．問題理解カテゴリのタスクは変換前後で回答が一貫しているかを評価しているため INV となる．一方，その他のカテゴリのタスクはシンプルな課題に対して正しい回答をできたかどうかで評価しておりテストタイプは MFT である．MFT と INV とでは評価の仕方が異なるため，指標を直接比較することはできない点に注意が必要である．

4　実験

実験は，今回の手法による評価データの作成と，作成したデータによるモデルの評価の二段階に分けられる．

表 1　MFT の各タスクごとのエラー率.

カテゴリ	入力文認識	回答選択	回答生成
タスク	質問・選択肢記憶	否定	出力形式指定
LLaMA2 70B	9.3%	**34.7%**	4.1%
Mixtral 8x7B	14.0%	**41.8%**	17.8%
Mistral 7B	21.1%	**48.3%**	16.2%

4.1　評価データの作成

本提案手法は既存の評価データセットを変換して新たなデータセットを作成する．本実験では言語モデルの評価によく用いられる MMLU を用いた．MMLU に含まれる多肢選択問題をいくつかのルールを定義して問題フォーマットごとに分類し，各フォーマットから 50 問ずつランダムに合計 150 問を抽出した．なお all of the above のような他の選択肢を参照する選択肢を含む問題は選択肢の並び替えやフォーマット変更時に影響があるため省いた．また本実験は 5-shot examples で行った．

4.2　用いたモデル

本実験では上記で作成した評価データを元に，モデルが公開されていて多くのデータセットでベンチマークされている Llama2 70B [6] と，それをスコアで上回る Mixtral 8x7B [7]，比較的小規模だが高性能の Mistral 7B [8] の 3 つのモデルを評価した．

4.3　結果と考察

MFT タスク　テストタイプが MFT のタスクのエラー率を表 1 に記載する．5-shot だとしてもフォーマットレベルの問題でこれだけ失敗するとなると，本来評価したい内容についての結果に大きな影響を与えるリスクがある．特に否定のタスクでのエラー率が高くなった．中でも選択肢の内容で指定する問題は Llama2 70B でエラー率 7.3%に対し，選択肢のラベルで指定する問題は 62.0%と高くなった．さらに質問記憶を除く各タスクの選択肢の指定方法と回答形式ごとにエラー率を比較すると（表 2），選択肢のラベルをもとに指定してる問題でエラー率が高くなっていることがわかった．否定のタスクでラベル指定の場合に特にエラー率が高くなるのは，問題文に含まれるラベルが 3 つと他のタスクと比べて多いことが原因の可能性がある．

表 2 MFT の各タスクの指定方法ごとのエラー率.

タスク	選択肢記憶		否定		出力形式指定			
質問文での選択肢指定	内容	ラベル	内容	ラベル	内容		ラベル	
回答の出力形式	(ラベル)	(内容)	(ラベル)	(内容)	ラベル	両方	内容	両方
LLaMA2 70B	1.3%	**16.7%**	7.3%	**62.0%**	1.0%	1.7%	**9.0%**	**4.7%**
Mixtral 8x7B	2.0%	**30.0%**	7.3%	**76.3%**	2.7%	3.3%	**32.3%**	**33.0%**
Mistral 7B	2.0%	**50.7%**	18.7%	**78.0%**	1.3%	1.3%	**42.3%**	**20.0%**

INV タスク 次にテストタイプが INV のタスクのエラー率 (変更前後で回答が一貫していない割合) を表 3 に記載する．比較すると，フォーマット変更でのエラー率が比較的高くなった．一方で，フォーマット変更では選択肢のシャッフルも行っているため，フォーマット変更によるエラー率の上昇幅は比較的少なくなった．

表 3 INV の各タスクごとのエラー率.

カテゴリ	問題理解			
タスク	フォーマット変更 + シャッフル	選択肢変更		
		シャッフル	1234	ハイフン
LLaMA2 70B	25.2%	**25.3%**	8.7%	21.3%
Mixtral 8x7B	**28.3%**	23.3%	12.7%	18.7%
Mistral 7B	**33.4%**	30.0%	12.0%	27.3%

また LLaMA2 70B での，変換前後のフォーマットごとのエラー率を表 4 に記載する．SimpleQ への変換時のエラー率が高くなった．Continuation と Gap-Fill の SimpleQ への変換が，問題文と選択肢を結合，転記し，新たな選択肢としており，選択肢を変更前より長文にしていることの影響の可能性がある．

表 4 フォーマット変更前後でのエラー率（LLaMA2 70B）.

変更前	変更後のフォーマット		
	SimpleQ	Continuation	Gap-Fill
SimpleQ	-	24%	18%
Continuation	**32%**	-	18%
Gap-Fill	**38%**	20%	-

さらに変換前後での選択肢問題に対する正答率もそれぞれ測定した（表 5）．全体的に変換後の正答率が変換前と比べて下がる結果となった（全体では 74.0% → 66.6% ）．人手での確認では問題の内容は変わっていないはずではあるものの，フォーマット変換によって問題を難化させている可能性がある．

表 5 フォーマット変更前後の正答率（LLaMA2 70B）.

	変更前の正答率	変更後のフォーマット		
		SimpleQ	Continuation	Gap-Fill
SimpleQ	**74%**	-	62%	**74%**
Continuation	62%	**66%**	-	58%
Gap-Fill	**86%**	66%	75%	-

5 おわりに

本研究では，言語モデルの選択肢問題に対応する能力と頑健性の評価を行なった．言語モデルは，簡単なタスクにも関わらず高いエラー率となり，選択肢問題に対応する能力が本来測りたい評価に対し影響を与える可能性を示した．また問題の趣旨を変えない変更を加えるだけで言語モデルが回答を変えることを示し，頑健性の低さを指摘した．とくに否定のタスクにおいて比較的エラー率が高くなることが明らかになった．否定を含む選択肢問題を言語モデルの評価に用いると，モデルの選択肢問題に対応する能力の高低に影響を受けてしまい，本来測定したかった知識能力を測れない可能性がある．また各タスクにおいて，ラベルをもとに選択肢を指定する問題でエラー率が高くなることがわかった．従来の選択肢問題では問題文自体にラベルが含まれるケースは少ないが，選択肢に他の選択肢のラベルが含まれる形式は見受けられる（e.g. A. {Option1} B. {Option2} C. A and B）. このような，ラベルの参照を必要とする選択肢の影響は本手法で考慮できていないため今後の課題である．

参考文献

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[2] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2023.

[4] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics.

[5] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations, 2023.

[6] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[7] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[8] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.