LLM はユーザーに適したテキストの難易度を暗黙的に考慮しているのか？

郷原聖士　上垣外英剛　渡辺太郎

奈良先端科学技術大学院大学
{gobara.seiji.gt6,kamigaito.h, taro}@is.naist.jp

概要

学生の理解度向上には、個人の学習レベルに適した教育が必要である。また、言語学習などの指導では、教員は各学生の理解度の把握が重要である。ただし、教員が全学生に個別指導を行うのは時間的な制約から困難である。解決策として、大規模言語モデル（LLM）で学生の質問応答をサポートする方法が考えられる。LLM は、幅広い分野への回答が可能なため、LLM を活用した細かな指導の自動化が期待される。しかし、LLM が指導者の代わりに質問応答できるとして、細かな指導の限界は未知である。そこで本研究では、教育分野での LLM の活用を促進するために文章の難易度に焦点を当てて、LLM が持つユーザへの暗黙的な難易度調整能力を調査する。

1　はじめに

学生の理解度を高めるためには、個人に合った指導方法が必要である。

LLM LLM は、BERT [1] をはじめとする旧来の小規模モデルと比較して、より大量のデータを使って莫大なパラメータ数で学習を行ったモデルである。LLM の著しい発展により、文章の要約や機械翻訳、質問応答などの様々なタスクが高精度で解けることが知られている [2, 3, 4, 5, 6, 7, 8]。LLM の成功を受けて、教育応用に向けた研究が注目されている。

教育応用 LLM を使った教育応用として、子供から大人までの幅広い人々に対する、ステップバイステップでの方式や、クイズ・フラッシュカードのようなインタラクティブな形式での教育方法などがある [9]。LLM を利用した教育方法で子供の好奇心を刺激することで、子供の学習意欲を高める働きがあると報告されている [10]。また、LLM を活用した教育応用として、テキスト平易化がある。

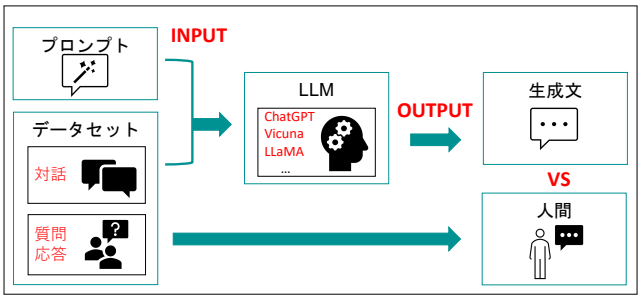

図 1　概要図

テキスト平易化 テキスト平易化は、文や文章の意味や内容を保持した上で、文法や構造を平易なものに変換する自然言語処理の生成タスクの一つである [11]。テキスト平易化を行うことで、読者の理解しやすい言葉や表現に変換され、情報を効率的に伝えることが可能となる。LLM を用いた手法としては、GPT-3.5 [2] を使って zero-shot と few-shots で平易化する方法 [12] や、難易度を付与して文章の難易度を調節する方法 [13] がある。生成する文章の難易度を調節することで、ユーザそれぞれに対して分かりやすい文章を提供することが可能になる。また、ユーザに合ったシステムをそれぞれ提供する仕組みとして、個人化がある。

個人化 Xie ら [14] は、2007 年から 2017 年の個人化に関する研究の傾向を分析し、秀逸な学習システムの実装や学習者の選好の統合、個々の学習データの分析に関する研究が行われていることを示唆している [15, 16]。LLM では、人間の選好を学習に用いることで文章の個人化を実現し、モデルの精度を向上させている [2]。文章の個人化として、例えば言語学習では、教師が生徒の理解度を一定数考慮した上でレベルを調整しながら、文法や語学に関する説明を行っている。しかし、生徒の理解度にはバラつきがあり、教師が各生徒の理解度を把握してそれぞれに合わせた説明を行うのは時間の制約から困難で

ある。そこで LLM を使って問題作成や講義の説明を実施することで、学習者の理解度に合わせた個人化が期待される [17]。すなわち、LLM による学習の個人化のためには、学習者の理解度を考慮する必要がある。

しかし現時点では、LLM がどれだけ学習者の理解度を考慮した上で文章を生成しているのかは分かっていない。本研究では、教育分野における LLM の活用を促進するために文章の難易度に焦点を当てて、LLM が持つユーザの入力に対する暗黙的な難易度調整能力を調査する。

2 実験手法

2.1 実験設定

本実験では、再現性確保のために乱数を固定して Greedy Search で文章生成を行った。また、文章の入力トークン数と生成トークン数の合計は 3072 トークンに設定した（付録 A）。

2.2 データセット

本実験では、質問応答を収集した stack-overflow データセットと、授業中の対話履歴を収集した Teacher-Student Chatroom Corpus （TSCC） [18] の 2 種類に対して評価実験を行った。

2.2.1 stack-overflow

stack-overflow データセット[1)]は、主にプログラマーのソースコードや実行環境などに関連した質問回答を収集した 1,000 件のデータセットである。本データセットは、2023 年 7 月 1 日時点での質問データセットをスクレイピングし、質問応答に対応しているデータを抽出することで構築した。

2.2.2 TSCC

TSCC は、Caines ら [18] が公開している 260 件のデータセットであり、教師と学生の授業内でのチャットログを収集したものである。本研究では、先頭からの対話履歴を抽出してラベル teacher と student を冒頭に付与し、11 ターン以降の発話で、最初に教師が返答する直前までの対話を入力とした。

2.3 モデル

本研究では、LLM が持つユーザへの難易度調整能力を調査するために、複数の LLM [2, 3, 4, 5, 6, 7, 8] を比較対象に選出した。

ChatGPT [2, 3] は人間とのアライメント（RLHF : Reinforcement Learning from Human Feedback）を考慮した LLM であり、現在利用可能な言語モデルとしては極めて性能が高いことが知られている。

LLaMA-2 [4] は、7 億から 70 億のパラメータを持つ事前学習および微調整済みの LLM である。本モデルは、多くのベンチマークで高い性能を示している点にくわえて、人手評価によって有用性と安全性の点が評価されており、クローズドソースのモデルを代替する可能性を秘めている。

Vicuna [5] は、LLaMA [19] をベースとして ChatGPT とのやり取りを収集したデータセット[2)]を用いて、人間の選好に合うように学習された LLM である。本研究では、LLaMA-2 をベースとして学習を行ったバージョンのモデルを比較対象として選定した。

Orca [6] は、LLaMA-2 をベースとして、100 件のタスクに対して段階を踏み、詳細に答える、といった複数の戦略に基づいたプロンプトでファインチューニングしたモデルである。本モデルは、複数の合成データを用いて学習していることから、入力文への臨機応変な出力対応と共に、暗黙的な難易度調整能力を獲得していることが期待される。

Mistral [7] は、70 億のパラメータを持つ事前学習済みモデルである。本モデルよりもパラメータサイズの大きな LLaMA-2 の 13b モデルと比較して、Mistral はベンチマークで高い性能を記録している。

Starling [8] は、Mistral をファインチューニングした Openchat3.5 [20] をベースとして、GPT-4 [3] のフィートバックで学習した報酬モデルを用いて訓練したモデルである。Mistral をベースとして学習したモデルについても同様に、ファインチューニングで難易度調整能力を獲得するのかを調査するために選定した。

2.4 プロンプト

本実験では、プロンプトによる効能を考慮するため、stack-overflow データセットでは simple、normal、complex の 3 種類、TSCC では 1 種類のプロンプトに対する各言語モデルの生成例を収集し、比較実験

1） https://insights.stackoverflow.com/survey

2） https://sharegpt.com/

を行った。ただし、プロンプトで Rooein ら [13] のように難易度を付与する場合、個人化の方向性が固定化され、実際にはユーザに不適な個人化を行う懸念がある。したがって、LLM が持つ暗黙的な難易度調整能力を評価するために、プロンプトと入力にはユーザーの文章理解度を含めず、表 3、表 4 のように設定した。

2.5 評価指標

本研究では、LLM の難易度調整能力を調査するため、文章の難易度、同義性、冗長性という 3 つの軸を評価指標として設定した。評価時には、入力文と生成文の各評価指標に対するスピアマンの順位相関係数を計算した。また、生成文の内、空白などの不適な文の生成数（Skip）についても記録した。

難易度 言語教育などの現場では、教師が生徒の語彙力や文章理解能力に合わせて言い換える必要があるため、本研究ではこの能力を文章の難易度として評価指標に設定した。評価指標には、伝統的な文章の難易度推定指標である FKGL [21]、FRE [22]、SMOG [23] にくわえて、NERF [24] を選出した。NERF は、語彙の難易度、文章の構造の複雑さ、ユニークな語彙の種類からなる 4 つの要素に対して、人手で作成した特徴量で文章の難易度を定式化したものである。FKGL や SMOG などの伝統的な難易度推定指標と比較して、高い精度で文章の難易度を推定可能であると報告されている。

同義性 生成した内容について、LLM が正しい内容を返していることを調査する必要がある。そこで我々は、収集したデータセットに含まれていた文章をリファレンスとして、LLM によって生成された文章に対する BERTScore [25] を計算することにより、LLM がユーザーの意図した内容に沿って回答していることを確認した。

冗長性 質問応答や教育の現場などにおいて、教師は生徒に対して冗長ではない適切な長さでの説明をすることが望ましい。したがって本研究でも LLM が過不足ない適度な長さで応答文を生成することが出来ているのかを調査するため、LLM の生成文と入力文の長さを比較した。

3 実験結果と考察

表 1 に、stack-overflow データセットで normal プロンプトを入力して回答を生成した時の、入力文と出力文の難易度に対する相関係数を示す。表 1 において、多くのモデルでは BERTScore が高くなっているが、LLaMA-2 に関しては内容の繰り返しや文章が生成されなかったデータを含んでいたことで（Skip）、全体的に数値が低くなっている。Orca-2-7b と Orca-2-13b を比較すると、全ての項目で Orca-2-13b の性能が高くなっていることから、モデルのスケーリング則が確認できる。

また、全体として、Vicuna-13b、GPT-3.5、GPT-4 の性能が高く、そのスコアは人手で作成された回答（original）よりも高い数値になっている。すなわち、LLM は人間と比較して、質問者の作成した文章の難易度を考慮して回答を生成することができると考えられる。Vicuna-13b、GPT-3.5、GPT-4 は、どれも学習に対話データのログを用いており、対話ログには、質問内容の訂正や人間の選好性が反映されていることから、暗黙的な難易度調整能力の獲得には、人間の選好性が重要であると考えられる。

くわえて、Mistral-7b は、Vicuna-13b、GPT-3.5、GPT4 に次ぐ性能であったが、モデルの学習に用いられている、Web 上からクロールしたデータセットに関する詳細が明らかではないため、データのリークが発生している懸念がある。ここで、GPT-3.5 と GPT-4 を比較すると、一般的に GPT-4 の性能が高くなることが知られているが、本タスクでは GPT-3.5 の方が高いスコアになる評価指標が多かった。このことから、GPT-4 による学習は、GPT-3.5 と比較して、難易度調整能力を向上させていないと考えられる。一方で、GPT-4-0613 と GPT-4-1106 を比較すると、全ての項目で GPT-4-0613 のスコアが高くなっていたため、バージョンアップに際して行った追加学習が、生成する文章の暗黙的な難易度調整能力を減退させたと考えられる。ただし、本実験で用いたデータについて、特に stack-overflow データセットは、オンライン上でのプログラムコードをメインに扱った匿名のやり取りを収集しているため、コード自体の文章の難しさと、質問者の質問内容から生じる文章の難しさが混在しており、上手く評価面で切り分けられていない可能性がある。よって、コードに特化した LLM による生成文の評価と、コードを含まないデータセットに対する生成文の評価を追加で実施する必要があると考えられる。

また、simple プロンプトと normal プロンプトを用いた場合についても上記と同様の傾向が見られたが、いずれもユーザーに対して過度に親切にしたり難しくしたりしない、normal プロンプトを用いた

表 1 stack-overflow データセットでの生成文の特徴量に対する相関係数（normal プロンプト）

Model	length	BERTScore	FRE	SMOGI	FKGL	NERF	Skip
Original	0.203	refs	0.428	0.265	0.387	0.248	0
Llama-2-7b	0.047	0.587	0.157	0.196	0.140	0.159	16
Llama-2-13b	0.119	0.581	0.157	0.249	0.182	0.118	7
Llama-2-70b	-0.070	0.448	0.133	0.150	0.154	0.082	16
Vicuna-13b	0.333	0.682	**0.555**	0.452	**0.491**	0.380	0
Orca-2-7b	0.226	0.646	0.324	0.271	0.280	0.239	0
Orca-2-13b	0.467	0.652	0.426	0.325	0.388	0.350	0
Mistral-7b	0.375	0.683	0.542	0.443	**0.489**	0.353	1
Starling-7b	-0.110	0.670	0.281	0.328	0.265	0.340	0
GPT-3.5-0613	0.342	0.697	0.523	**0.455**	0.448	0.373	0
GPT-3.5-1106	**0.414**	0.695	0.492	0.448	0.422	**0.405**	0
GPT-4-0613	0.370	**0.699**	0.498	0.430	0.428	0.323	0
GPT-4-1106	0.268	0.688	0.443	0.407	0.366	0.322	0

表 2 TSCC での生成文の特徴量に対する相関係数

Model	length	BERTScore	FRE	SMOGI	FKGL	NERF	Skip
Original	0.288	refs	0.157	0.098	0.192	0.075	0
Llama-2-7b	-0.061	0.642	-0.093	-0.010	-0.094	0.075	0
Llama-2-13b	0.252	0.613	-0.041	**0.622**	0.035	-0.097	4
Llama-2-70b	0.100	0.653	-0.162	0.329	-0.129	-0.049	3
Vicuna-13b	0.104	0.623	-0.076	-0.037	-0.024	-0.049	5
Orca-2-7b	0.087	**0.655**	0.124	0.079	0.160	-0.007	1
Orca-2-13b	0.021	0.634	-0.111	-0.041	-0.120	0.058	1
Mistral-7b	0.001	0.629	0.149	0.270	0.130	0.059	4
Starling-7b	0.069	0.573	0.096	0.071	0.084	-0.071	0
GPT-3.5-0613	0.301	0.651	0.163	0.076	0.210	**0.130**	0
GPT-3.5-1106	0.285	0.652	0.095	0.152	0.091	0.110	0
GPT-4-0613	0.285	0.656	0.167	0.163	0.184	0.113	0
GPT-4-1106	**0.388**	0.643	**0.300**	0.132	**0.357**	0.080	0

場合の相関係数の方が高くなっていた。このことから、 LLM は暗黙的な難易度調整能力を十分に高く獲得していることが示唆された。

一方で、 TSCC での入力文と出力文の難易度に対する相関係数は、 BERTScore を除いて低くなっていることが見て取れる（表 2）。TSCC で比較対象とした生成文のトークン数は、 stack-overflow データセットと比較して高々数十トークンと少なく、 対話特有のスラングを含んでいた。したがって、 LLM は入力文に対して自然な対話を生成していた一方で、 評価対象とした生成文が短く、 極端に省略された対話特有のスラングを含む文が多かったため、 対話での LLM の難易度調整能力を正しく評価できていない可能性が高い。そこで LLM で複数ターンの対話を生成し、 入力データの過去の student の発話履歴を複数ターン分収集して比較することで、 LLM の持つ対話での難易度調整能力をより正確に評価できるようになると考えられる。

4 おわりに

本研究では、 LLM の持つユーザへの暗黙的な難易度調整能力を検証するために、 質問応答と対話を用いて LLM が生成した文章と入力文章の難易度を比較した。実験の結果、 Vicuna-13b、 GPT-3.5、 GPT-4 による生成文と入力文の難易度に強い相関関係が見られた。また、 一部の LLM では人間の作成した回答データと比較して、 相関係数が高かった。このことから、 LLM は人間よりも暗黙的に難易度を調節した回答文を生成できることが示唆された。

今後は、 LLM の難易度調整能力の獲得過程を分析するために、 人間と LLM の対話履歴を用いて、 どの選好が難易度調整能力に有効であるのか調査したいと考えている。例えば、 Starling のベースモデルとなった OpenChat3.5 [20] と、 ベースである Mistral の同タスクでの性能比較のように、 学習に用いられたデータや学習手法で、 暗黙的な難易度調整能力の獲得に差はあるのか、 検証を進めていきたい。

参考文献

[1] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of naacL-HLT**, Vol. 1, p. 2, 2019.

[2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[3] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **arXiv preprint arXiv:2306.05685**, 2023.

[6] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. Orca 2: Teaching small language models how to reason. **arXiv preprint arXiv:2311.11045**, 2023.

[7] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. **arXiv preprint arXiv:2310.06825**, 2023.

[8] Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving llm helpfulness harmlessness with rlaif, November 2023.

[9] Ebrahim Gabajiwala, Priyav Mehta, Ritik Singh, and Reeta Koshy. Quiz maker: Automatic quiz generation from text using nlp. In **Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021**, pp. 523–533. Springer, 2022.

[10] Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, Jaap Kamps, et al. Reading comprehension quiz generation using generative pre-trained transformers, 2022.

[11] Suha S Al-Thanyyan and Aqil M Azmi. Automated text simplification: a survey. **ACM Computing Surveys (CSUR)**, Vol. 54, No. 2, pp. 1–36, 2021.

[12] Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. Sentence simplification via large language models. **arXiv preprint arXiv:2302.11957**, 2023.

[13] Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. Know your audience: Do llms adapt to different age and education levels? **arXiv preprint arXiv:2312.02065**, 2023.

[14] Haoran Xie, Hui-Chun Chu, Gwo-Jen Hwang, and Chun-Chieh Wang. Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. **Computers & Education**, Vol. 140, p. 103599, 2019.

[15] Chih-Ming Chen and Ching-Ju Chung. Personalized mobile english vocabulary learning system based on item response theory and learning memory cycle. **Computers & Education**, Vol. 51, No. 2, pp. 624–645, 2008.

[16] Gwo-Jen Hwang, Fan-Ray Kuo, Peng-Yeng Yin, and Kuo-Hsien Chuang. A heuristic algorithm for planning personalized learning paths for context-aware ubiquitous learning. **Computers & Education**, Vol. 54, No. 2, pp. 404–415, 2010.

[17] Risang Baskara, et al. Exploring the implications of chatgpt for language learning in higher education. **Indonesian Journal of English Language Teaching and Applied Linguistics**, Vol. 7, No. 2, pp. 343–358, 2023.

[18] Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. The teacher-student chatroom corpus. In David Alfter, Elena Volodina, Ildikó Pilan, Herbert Lange, and Lars Borin, editors, **Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning**, pp. 10–20, Gothenburg, Sweden, November 2020. LiU Electronic Press.

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[20] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. **arXiv preprint arXiv:2309.11235**, 2023.

[21] George R Klare. Assessing readability. **Reading research quarterly**, pp. 62–102, 1974.

[22] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[23] G Harry Mc Laughlin. Smog grading-a new readability formula. **Journal of reading**, Vol. 12, No. 8, pp. 639–646, 1969.

[24] Bruce W Lee and Jason Hyung-Jong Lee. Traditional readability formulas compared for english. **arXiv preprint arXiv:2301.02975**, 2023.

[25] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

表 3　プロンプトの種類

	stack-overflow
入力	(prompt) + (question)
simple	Please respond to the question using simple and user-friendly language. \n ### Question :
normal	### Question:
complex	Please respond to the question using complex and less user-friendly language. \n ### Question:
	TSCC
入力	(prompt) + (dialogue)
prompt	Please generate a response from the teacher to the student in the ongoing dialogue. \n ### Dialogue:

表 4　teacher-student の対話例（Starling-7b）

入力	Please generate a response from the teacher to the student in the ongoing dialogue. ### Dialogue:student: Hi! teacher: Hi <STUDENT>! teacher: Everything alright with the chatroom for you? student: I tried to use it a few seconds ago and I couldn't change my name, but now it is working, thanks. student: How are you? teacher: Oh good! teacher: Fine, thank you! It's summer here at last, we've had a week of non-stop sunshine! teacher: How are you? student: I'm fine thank you! It looks like summer has arrived here too! student: Even though we still had a couple of storms... student: with hail and everything teacher:
出力	Ooh, I hope you're not too badly affected by them!

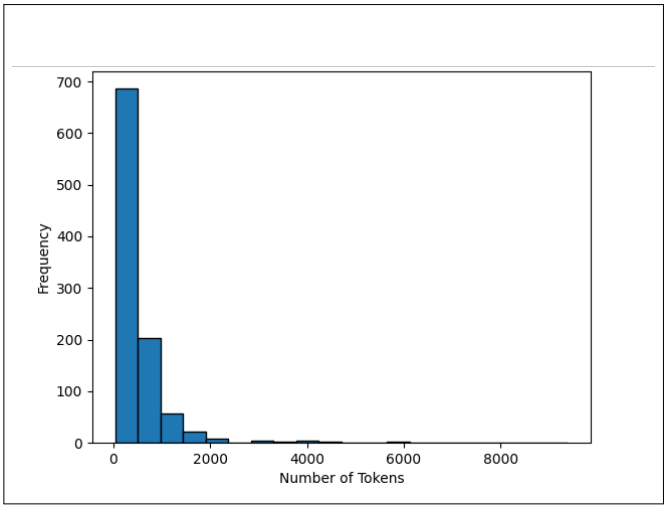

図 2　入力トークン数のヒストグラム

A　長文の入力への対応

図 2 に、 stackover-flow データセットの入力データに対して、 Llama-2-7b [19] のトークナイザーでトークン数を計算したヒストグラムを示す。図 2 において、 全入力データの 97.0%が 2048 トークン以下、98.1%が 3072 トークン以下、 1.9%が 3072 トークン以上となっている。モデルが入力文に対して生成する出力で難易度調整能力を獲得しているのかを評価するためには、 全ての入力文は必要でなく、 2048 トークンで十分に多くの入力文の内容を捉えることが可能であると考えられる。したがって、 入力文と出力文の文章の長さを統一して生成するために、モデルへの入力を 2048 トークンまでで切り捨てて、最大のトークン生成数を入力トークンと合わせて 3072 トークンになるように調節した。