読者はどこでつまずく？生成多肢選択式問題を用いた医学研究データセット

林 純子 [1] 永井 宥之 [1] 久田 祥平 [1] 倉本 真菜 [1] 若宮 翔子 [1] 荒牧 英治 [1]
[1] 奈良先端科学技術大学院大学
{hayashi.junko.hh5, s-hisada, kuramoto.mana.kj4, wakamiya, aramaki}@is.naist.jp,
hiro.nagai@naist.ac.jp

概要

専門性の高い医学研究成果を一般読者に分かりやすく伝える上で，読者のつまずく箇所を特定しにくいことは大きな障壁である．この課題に対し，本研究では論文の構成要素（目的，方法等）に基づく設問レベルの難易度評価データセットを構築した．LLM で生成した 350 問の多肢選択式問題に対し一般読者の回答を大規模に収集し，項目反応理論により各設問の難易度を算出した．このデータセットに対して，LLM による推定と比較した結果，人間と LLM の相関係数は 0.1 と低いことが分かった．本データセットは，読者がつまずきやすい箇所を特定する研究の基盤を提供するものである．

1 はじめに

科学の成果を一般読者に分かりやすく伝達することは，科学コミュニケーションにおける重要な課題である [1, 2]．特に医学分野の研究成果は患者の意思決定に直結するが，公開されている研究成果報告は高度な専門用語や複雑な論理構造を含み，一般読者には理解困難な場合が多い．こうした文書を一般読者向けに改善する際，「読者はどこでつまづくか」を正確に特定する必要がある．この課題に対して，可読性指標により文書の難易度を見積もる研究 [3] や，機械学習や LLM を用いて理解困難性を推定する試みも進んでいる [4, 5]．

しかし，既存手法の多くは文書全体の評価にとどまっており，文書内のどの情報の理解が困難であるかを特定するには至っていない．執筆者が効果的に推敲を行うためには，「どの論理構成（方法や限界など）が伝わりにくいのか」という要素レベルのフィードバックが不可欠である．

本研究ではこの課題に対処するため，医学研究成果報告書の概要（以下，医学研究概要）から，文書全体ではなく情報の構成要素単位で一般読者の理解の難しさを定量化したデータセットを構築・公開する．本研究の特徴は，論文の構成要素に基づき「目的・方法・結果・解釈・結論・限界・語彙」という 7 つの要素から多肢選択式問題 (MCQ) を生成する点にある．これにより，単なる文章の複雑さだけでなく，研究内容を理解する上で重要な情報が正しく伝わっているかを検証可能にする．具体的には，大規模言語モデル (LLM) を用いて医学研究概要から MCQ を生成し，クラウドソーシングにより一般読者からの回答を収集する．さらに，項目反応理論 (IRT) [6] を用いて一般読者の回答の正誤データを元に設問ごとの難易度を推定することで，テキスト中の「理解されにくさ」を定量的に示す．さらに，応用として LLM が MCQ の難易度を推定できるのかを調査する．本研究の主な貢献は以下の 3 点である．

- 日本語医学研究概要を対象に，情報の構成要素ごとの MCQ と一般読者の回答からなる難易度評価データセットを構築・公開[1)] した点
- 一般読者の回答に基づき，どのような種類の情報（例：研究目的や用語など）が理解の妨げになりやすいかを設問レベルで明らかにした点
- 構築したデータセットを用いて LLM による難易度推定を行い，一般読者が用いる「消去法」や「表層的な手がかり」による解きやすさを過小評価するという，人間との認知的な乖離を明らかにした点

本研究の成果は，将来的には研究者が執筆した研究概要に対し，「この情報が伝わりにくい」「この用語は補足が必要」といった具体的な修正提案を行うシステムの基盤となり，患者と医療者のより良いコミュニケーションに寄与することが期待される．

1） https://github.com/sociocom/Medtext_MCQ

表 1　医学研究概要と MCQ の例．医学研究概要における灰色の背景は，MCQ を生成する際に元にした evidense phrase である．MCQ は目的や方法といった要素別に構成されており，選択肢は (A) から (D) の 4 つである．設問難易度は，値が高いほど難易度が高く，値が低いほど難易度が低いことを意味する．

本研究は、がん組織における腫瘍微小環境の免疫抑制機構を解明し、新規免疫療法開発の基盤構築を目的として実施した。特に腫瘍関連マクロファージ（TAM）の機能的亜集団を単一細胞解析技術により詳細に分類し、各亜集団の遺伝子発現プロファイルを明らかにした。令和 4 年度は、多施設から収集した悪性乳がんおよび膵臓がん組織検体を対象にシングルセル RNA-seq および空間トランスクリプトミクス解析を行い、TAM の腫瘍浸潤における多様性と位置依存的機能分化の実態を解明した。また、これらのデータを基に、特異的に腫瘍促進性シグナルを介する TAM サブセットを標的とした抗体療法候補分子をスクリーニングし、in vitro および in vivo モデルでの有効性評価を実施した。さらに、免疫抑制性 TAM の誘導に関与する腫瘍由来エクソソームの脂質成分解析を進め、新規の脂質シグナル伝達経路を同定した。これらの研究成果は、腫瘍微小環境における免疫逃避機構の分子メカニズムの解明と、それを阻害する治療戦略の開発に資すると期待される。引き続き、多様ながん種での TAM 特性の横断的比較や、臨床試験導入に向けた治療候補分子の安全性評価を進める予定である。

要素	設問例	選択肢	設問難易度
目的	本研究の目的は何ですか？	**(A)** がんの早期発見法を確立すること，**(B) 腫瘍微小環境の免疫抑制機構を解明すること**，**(C)** 新しいがん治療法を開発すること，**(D)** 腫瘍関連マクロファージの機能を調査すること，	−0.24
方法	本研究で使用された解析技術は何ですか？	**(A) シングルセル RNA-seq**，**(B)** フローサイトメトリー，**(C)** ELISA，**(D)** PCR	−1.51
結果	腫瘍関連マクロファージ（TAM）の研究で明らかにされたことは何ですか？	**(A) TAM は腫瘍浸潤において多様性がある**，**(B)** TAM は全て同じ機能を持つ，**(C)** TAM はがんの進行を妨げる，**(D)** TAM は免疫系を完全に抑制する	−3.17
解釈	研究成果が期待される影響は何ですか？	**(A)** 新しいがんの診断法の発見，**(B)** がんの予防法の確立，**(C)** がん治療のコスト削減，**(D) 免疫逃避機構の解明と治療戦略の開発**	−3.17
結論	研究の結論として正しいものはどれですか？	**(A) TAM の特性はがん治療に重要である**，**(B)** TAM はがんに対して無関係である，**(C)** TAM の研究は無駄である，**(D)** TAM は全てのがんに共通する	−13.81
限界	今後の研究で進める予定の内容は何ですか？	**(A)TAM 特性の横断的比較**，**(B)** 新しいがん治療法の開発，**(C)** がんの予防法の研究，**(D)** がんの早期発見法の確立	3.89
語彙	腫瘍関連マクロファージ（TAM）とは何を指しますか？	**(A)** 免疫系の細胞全般，**(B) がん組織に存在するマクロファージ**，**(C)** 正常なマクロファージ，**(D)** 血液中のマクロファージ	0.49

2　関連研究

医学領域では，患者教育資料などの質を高めるために可読性や形式を分析した研究が報告されている [7, 8]．多様な readability 指標やヘルスリテラシー尺度を用いて読みやすさを評価した研究も存在する [9]．これらの研究はいずれも文書全体の難易度や形式的特徴に焦点を当てている．

近年は，LLM を用いて読解問題や MCQ を自動生成し，その品質や難易度を分析する研究も登場している．LLM を用いて読解 MCQ を自動生成し，一般読者評価と自動評価の両面から品質を検証する研究 [10] や，IRT に基づく難易度を付与した読解問題データを構築し，指定した難易度の質問を生成する研究 [11] などが存在する．

3　データセット構築

3.1　医学研究概要

本研究では，医学研究概要の設問難易度を調査するため，AMEDfind[2)]に収録された研究成果報告（研究課題の概要・成果に関する公開記述）をデータソースとして，GPT-4.1 mini を用いて医学研究概要を生成した．本研究ではこのうち 50 文書の研究成果報告から研究概要を生成した．生成した医学研究概要の例を表 1 に示す．

3.2　MCQ 生成・品質評価

LLM を用い，生成した 50 件の医学研究概要に対し 7 問ずつ，計 350 問の MCQ を生成した．モデルには，LLM による MCQ 生成の品質を調査した先行研究 [12] を参考に，GPT-4o mini を使用した．MCQ の例を表 1 に示す．選択肢と解答に加え，その MCQ を作る元となった文を evidence phrase として出力させた．付録表 A.1 に生成時に用いたプロンプトを示す．1 文書あたり 7 問としたのは，概要文に含まれる情報を目的，方法，結果，解釈，限界，結論，語彙の 7 要素に分解し，1 問ずつ問うことで，どの情報がどの程度理解されにくいかを要素ごとに定量化するためである．生成した MCQ の品質評価にあたっては，先行研究 [13, 14] を参考に，以下の 4 つの観点で Gemini 2.0 Flash を用いた自動評価を行った．各観点について，条件を満たす場合は 1，満たさない場合は 0 とする二値評価を行った．

- 文脈依拠性：正答の根拠が本文内に示されており，外部知識に依存せず解答可能であるか．
- 一義性：正答が論理的に唯一に定まり，選択肢間に解釈の揺れや，正誤の境界を曖昧にする過度な類似性が生じていないか．
- 選択肢妥当性：誤答選択肢がトピックとしてもっともらしく，無関係な内容や不自然な表現を含んでいないか．
- 消去可能性：誤答選択肢が容易に排除されず，一定の誤答を誘発する力を有しているか．

評価の結果，350 問の平均得点は 3.8 点（4 点満点），分散は 0.2 であった．350 設問中 346 設問（98.9%）が 4 点，3 点が 2 設問（0.6%），1 点が 2 設問（0.6%）であった．自動評価の範疇では，得点が全体として高くばらつきも小さいことから，生成された問題の多くが上記の 4 つの観点を安定して満たしており，MCQ としての品質が確保されていると考え，今回評価した MCQ を全て使用した．

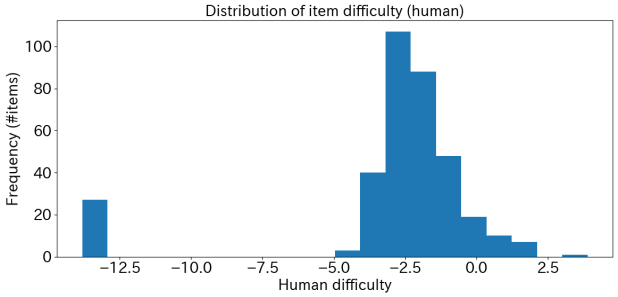

図 1　クラウドソーシング参加者の MCQ 正誤データから算出した，各設問の難易度近似値の分布．横軸は難易度，縦軸は頻度を示しており，低い値の設問ほど易しい問題であることを示す．

3.3　回答データ収集と設問難易度算出

この節では，一般読者の MCQ の正誤データを収集し，そのデータを用いて各設問の難易度を算出する．MCQ 設問難易度推定に必要な正解データを得るため，Yahoo!クラウドソーシングを用いて一般読者の回答を収集した．調査では，1 フォームあたり医学研究概要 5 文書と対応する MCQ35 問を提示し，計 50 文書について回答を収集した．フォーム間の尺度接続のため，隣接フォーム間で概要 1 文書分の設問を共有するアンカー設計を採用し，計 12 種類のフォームを作成した．参加者には 1 フォームを割り当て，各 MCQ につき約 50 名分の回答が得られるようにした．

収集した二値回答（正答=1，誤答=0）に対し，Rasch モデル（1PL）を適用して設問ごとの難易度を推定した．推定値 $\hat{b}_i$ を本研究の「設問難易度」として以降の分析に用いた．定式化・推定手順は付録 A.2 に記載した．

本実験は，奈良先端科学技術大学院大学の倫理審査委員会にて承認を受けたものである（承認番号:2025-I-15）．

3.4　データセット分析

一般読者の正答率は平均 0.85 であり，難易度は全体的に低い．各フォームの Cronbach の α 係数は中央値で 0.84 となり，多くのフォームで 0.80 を超える高い内的整合性が確認された．これは，クラウドソーシングにより収集された回答データが一貫した

2）https://amedfind.amed.go.jp/amed/

表 2 LLM による難易度推定と一般読者データに基づく難易度の比較．問題文，選択肢（太字が正解），解答の根拠となる evidence phrase，および LLM が難しいと判断した理由を示す．

設問例	選択肢	Evidence phrase	一般読者正答率（設問難易度）	LLM 推定正答率	LLM が難しいと判断した理由
CXCL17 とは何の略称ですか？	**(A)C-X-C モチーフケモカイン**, (B) 細胞増殖因子, (C) 腫瘍抑制因子, (D) マクロファージ活性化因子	新規微小環境制御分子「CXCL17」の役割を解明し	0.98 (-3.8)	0.2	CXCL17 が何の略称であるかは本文中に明記されていないため、一般読者には難しく、正答率は低いと予想される。
「GMP」とは何の略ですか？	**(A)Good Manufacturing Practice**, (B)General Medical Protocol, (C)Global Medical Program, (D)Good Medical Practice	治験製剤の GMP 製造体制の構築	0.58 (-0.3)	0.3	GMP は専門用語であり、本文に定義がないため、一般読者には非常に難しい。

測定尺度として機能しており，十分な信頼性を有していることを裏付けている．また，算出された難易度 b_i の分布を図 1 に示す．分布の平均値は –2.95 と負の方向に偏っており，一般読者にとって今回作成された MCQ の多くは比較的回答が容易であったことがわかる．MCQ 自体は平易な表現で作問されていることに起因する可能性がある．

医学研究概要の MCQ と一般読者の回答に基づく設問難易度を表 1 に示す．結論に関する設問は，設問難易度が-13.81 であり，易しい問題である．この問題の選択肢には，「(C) TAM の研究は無駄である」といった，妥当性の低い選択肢が含まれている．つまり，この設問は，選択肢を選びやすい問題である可能性が高い．この設問の品質評価は，「消去可能性（誤答選択肢が容易に排除されず，一定の誤答を誘発する力を有しているか）」が 0 と判定されていたため，MCQ の特徴による影響が考えられる．一方，難易度の高い設問は限界についてであり，設問難易度は 3.89 であった．この設問は「TAM 特性の横断的比較」という解答が文章内にあるものの，他の選択肢も大きな目的としては含むことが可能であるため，解答が難しかった可能性がある．この設問の品質評価は，評価基準を満たしていたことから，文章から正解を見つけ出すこと自体が困難であった可能性がある．

4 応用：人間と LLM による難易度認識の比較分析

構築した MCQ データセットの応用可能性を検討するため，一般読者の回答に基づき設問の難易度を基準とし，LLM による難易度推定がどの程度機能するかを調査する．具体的には，LLM に一般読者の正答率を推定させ，一般読者の回答データから得られた設問難易度との対応関係を比較し，LLM の推定傾向とその限界を観察する．評価には，スピアマンの順位相関係数を用いる．使用モデルは Gemini 2.0 Flash とした．

順位相関係数は 0.1（$p = 0.05$）であり，相関はほぼないことが示された．一般読者にとって理解が困難な設問を LLM が推定することは難しいことを示唆している．LLM が高難易度と推定した設問の例を表 2 に示す．1 問目では，選択肢中に「C-X-C」を含むものが一つしかなく，本文に略称の説明がなくても消去法により正解可能である．一方 LLM は，本文中に略称の定義がないことを理由に正答率が低いと推定しており，MCQ 特有の表層的手がかりを十分に織り込めていない可能性がある．2 問目は，一般読者にとっても難易度が高かった例である．GMP についての設問では，提示されたフルスペルはいずれも医学知識のない読者には判別が難しく，LLM が挙げた「本文中に定義がない」という理由は妥当である．以上より，LLM は設問の難しさを主に本文中の明示的な説明の有無に基づいて判断する傾向があり，消去法や選択肢の表層的手がかりによって解答可能な設問の性質を十分に反映できないことが示唆される．

5 おわりに

本研究では，医学研究概要を対象とした大規模 MCQ データセットを構築・公開し，一般読者の回答から設問難易度を推定した．今後は，読者の理解プロセスをより忠実に反映した自動評価手法を確立し，医学研究概要の設計・改善を支援する枠組みへと発展させていきたい．

謝辞

本研究は，AMED 課題番号 JP25oa0439009 および，「戦略的イノベーション創造プログラム（SIP）」「統合型ヘルスケアシステムの構築」JPJ012425，「JST 次世代 AI 人材育成プログラム」JPMJBS2423 の補助を受けて行った．

参考文献

[1] **Communicating Science Effectively: A Research Agenda**. National Academies Press, March 2017.

[2] Dietram A. Scheufele. Communicating science in social settings. **Proceedings of the National Academy of Sciences**, 2013.

[3] Sameer Badarudeen and Sanjeev Sabharwal. Assessing readability of patient education materials: Current role in orthopaedics. **Clinical Orthopaedics amp; Related Research**, Vol. 468, No. 10, p. 2572–2580, October 2010.

[4] Trevor Cohen, Weizhe Xu, Yue Guo, Serguei Pakhomov, and Gondy Leroy. Coherence and comprehensibility: Large language models predict lay understanding of health-related content. **Journal of Biomedical Informatics**, Vol. 161, p. 104758, January 2025.

[5] Sean Trott and Pamela Rivière. Measuring and modifying the readability of English texts with GPT-4. In Matthew Shardlow, Horacio Saggion, Fernando Alva-Manchego, Marcos Zampieri, Kai North, Sanja Štajner, and Regina Stodden, editors, **Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)**, pp. 126–134, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[6] Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. **Fundamentals of Item Response Theory**, Vol. 2 of **Measurement Methods for the Social Sciences Series**. Sage Publications, Newbury Park, CA, 1991.

[7] Nensi Bralić, Antonija Mijatović, Ana Marušić, and Ivan Buljan. Conclusiveness, readability and textual characteristics of plain language summaries from medical and non-medical organizations: a cross-sectional study. **Scientific Reports**, Vol. 14, No. 1, p. 6016, 2024.

[8] Leia Martínez Silvagnoli, Caroline Shepherd, James Pritchett, and Jason Gardner. Optimizing readability and format of plain language summaries for medical research articles: Cross-sectional survey study. **Journal of Medical Internet Research**, Vol. 24, No. 1, p. e22122, 2022.

[9] Lydia O'Sullivan, Prasanth Sukumar, Rachel Crowley, Eilish McAuliffe, and Peter Doran. Readability and understandability of clinical research patient information leaflets and consent forms in ireland and the UK: a retrospective quantitative analysis. **BMJ Open**, Vol. 10, No. 9, p. e037994, 2020.

[10] Andreas Säuberli and Simon Clematide. Automatic generation and evaluation of reading comprehension test items with large language models. In **Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024**, pp. 22–37, Torino, Italy, 2024. ELRA and ICCL.

[11] Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. Difficulty-controllable neural question generation for reading comprehension using item response theory. In **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 119–129, Toronto, Canada, 2023. Association for Computational Linguistics.

[12] Lauren Riehm, Keean Nanji, Moiz Lakhani, Evelina Pankiv, Dean Hasanee, and Wesla Pfeifer. The use of large language models in generating multiple choice questions for health professions education: A systematic review and network meta-analysis. **PLOS ONE**, Vol. 21, No. 1, pp. 1–19, 01 2026.

[13] Sérgio Silva Mucciaccia, Thiago Meireles Paixão, Filipe Wall Mutz, Claudine Santos Badue, Alberto Ferreira de Souza, and Thiago Oliveira-Santos. Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. 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. 2246–2260, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[14] Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. Multiple-choice question generation using large language models: Methodology and educator insights. In **Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization**, p. 584–590. ACM, June 2024.

A　付録

A.1　MCQ 生成のプロンプト

MCQ を生成する際の LLM への指示は以下の通りである．

> あなたは日本語の医療・科学コミュニケーションに関する読解問題の専門作問者です．出力は必ず JSON のみとし，余計なテキストは含めないでください．問題は必ず与えられたテキストに基づいて作成し，情報を捏造してはいけません．各設問について必ず 4 つの選択肢を作成してください（多肢選択式）．選択肢は，答えが必ず一意に定まるようにしてください．選択肢の長さはできるだけ揃えてください．参照元となるフレーズ（evidence phrase）を，文章から抜き出してください．追跡可能性のために source text 全文も JSON に含めてください．問題数は必ず 7 問にしてください．

A.2　Rasch モデル（1PL）

Rasch モデルでは，参加者 p の能力 θ_p と設問 i の難易度 b_i により，

$$P(y_{pi} = 1 \mid \theta_p, b_i) = \frac{1}{1 + \exp\bigl(-(\theta_p - b_i)\bigr)} \tag{1}$$

と定義する．本研究では joint maximum likelihood（JML）により $\{\theta_p\}, \{b_i\}$ を推定し，尺度同定のため $\frac{1}{I}\sum_i b_i = 0$ を課した．