医学研究テキストの「つたわる化」を目指したテキスト書き換え

倉本真菜 [1] 永井宥之 [1] 山田恵子 [2] 井出博生 [3,4]
早川雅代 [3,5] 西山智弘 [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 [2] 埼玉県立大学 [3] 東京大学 [4] 順天堂大学 [5] 慶應義塾大学
{kuramoto.mana.kj4,hiro.nagai}@naist.ac.jp
yamada-keiko@spu.ac.jp, ide@ifi.u-tokyo.ac.jp, hayakawam395@gmail.com
{nishiyama.tomohiro.ns5,wakamiya,aramaki}@is.naist.jp

概要

医学研究は高度に専門的であり，患者や一般読者が最新の研究成果を理解することは容易ではない．従来，専門用語を平易な表現に言い換える平易化が行われてきたが，背景知識の不足や重要情報の欠落といった課題が残る．そこで本研究では，平易化に加えて文脈や研究意義の補足を行う書き換えを「つたわる化」と定義し，大規模言語モデル (Large Language Model: LLM) を用いた実現手法を提案する．人手で作成したデータを正解例として評価した結果，ガイドラインと Few-shot 事例を併用する手法が最も優れた性能を示した．今後，科学技術情報を多くの人々が理解可能な形へ書き換えられるよう，ガイドラインの拡充など改善を継続する予定である．

1 はじめに

科学技術の専門化・細分化が進むにつれて，専門外の動向を把握することが難しくなる「蛸壺化」が指摘されている [1]．特に医療・生命医学分野（以下，単に医学分野）では，患者自身が自らの治療に関わる研究について論文や研究成果を調べることも，決して珍しくない．このような場合，専門知識を持たない患者にとっては内容を理解できないだけでなく，誤解が生じた際に，円滑な治療の妨げとなる可能性もある．

例えば，図 1 は研究概要であるが，新規薬剤が研究中であり，まだ，治療には使えないことは（研究者には自明なため）明記されていない．この点を知らない患者が，新規薬剤での治療を希望し，医療者が説明に苦慮する状況は容易に想像できる．

専門的な研究概要を患者が十分に理解するためには，単なる語や文の言い換え（平易化）だけでは不十分な場合がある．図 1 に示すように，読者の前提知識不足や複雑な構成が理解を妨げる場合，語彙の操作にとどまらず，文書構造の変更や情報の強調といった，より踏み込んだ編集が必要となる．

そこで我々は，文書構造の変更，情報の強調といった操作を含む編集を「つたわる化」と定義する．これは，単なる読者負荷の低減にとどまらず，重要な情報を可能な限り保持したまま，対象読者にとっての理解しやすさを最大化することを目的とする．

本研究では，人手で作成した正解例を用い，LLM による「つたわる化」の実現可能性を検証する．具体的には，人間が行う「つたわる化」の工程を LLM で再現し，どの程度理想的なテキストを生成できるか評価した．

2 関連研究

専門的な内容を非専門家向けに再構成する Lay Summarization は，医学生物学分野において BioLaySumm 等の評価基盤が整備され，重要なタスクとして位置づけられている [2, 3, 4]．同様に，テキスト平易化においても，読者層に合わせて難易度を制御する手法や SARI 等の指標を用いた評価が進展している [5, 6, 7, 8]．これらのタスクでは，テキストの読みやすさを定量化する可読性指標が，生成品質を測る信頼性の高い基盤として活用されている [9]．

しかし，既存のアプローチには医療テキスト特有の課題がある．第一に，既存の自動評価指標や手法は主に読みやすさの向上に主眼を置いており，非専門読者が抽象的な内容をどのように解釈するかという認知面は捉えきれていない [10]．そのため，非専

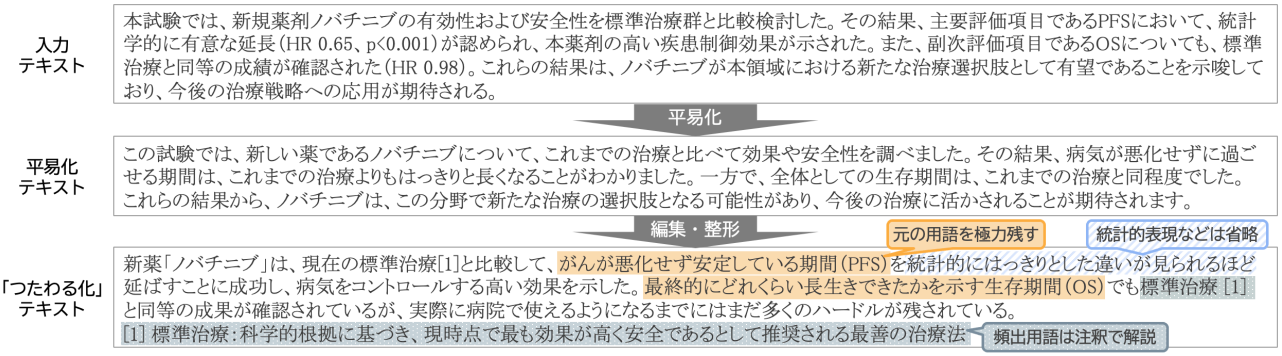

図 1: つたわる化の概要．本研究では，人間が非専門読者に向けて行う書き換えの工程を LLM で模倣する手法を提案する．従来の平易化に加え，読者に合わせた情報の補足や強調を行うことで，非専門読者でも理解しやすい文書の生成を目指す．

門家が既存手法で平易化した文書を真に理解できているかが未解明である．第二に，非専門家の理解に必要な背景情報や研究の意義を含めた要約の生成が求められている [11]．

本研究で提唱する「つたわる化」は，従来の平易化や Lay Summarization を包含し，情報の強調・省略や文脈補完を通じて，研究の背景や意義までを非専門家に伝えることを目指す包括的な枠組みである．

3 データセット

本研究では，模擬的な医学研究成果の概要文書（以下，模擬概要文書）を人手で「つたわる化」し，データセットを構築した．具体的には，AMEDfind[1)] から抽出した医学研究成果の概要文書をソースとして，LLM を用いて模擬概要文書を生成した．著作権上の制約を排除し，将来的な一般公開を可能とするため，実データではなく生成データを利用した．次に，模擬概要文書 20 件に対し，メディカルライター 1 名にガイドライン（A.2）を提示し，「つたわる化」のための書き換えを依頼した．これにより得られた 20 件の「つたわる化」された文書を人手書き換え文書とした[2)]．

4 提案手法

本研究では，専門的な医学テキストを非専門読者にとって「つたわる」形式へ変換するため，人手による編集プロセスを再現した 2 段階の LLM 書き換え手法を提案する．一般に，人手による「つたわる化」の過程では，まず専門用語や複雑な構文を平易な表現へ置換し，その上で読者の知識水準などに応じた情報の取捨選択や文脈の補足といった再構成が行われる [12]．これに基づき，提案手法ではまず第一段階として，単純な平易化，または直接的なつたわる化のいずれかの指示により，基礎的な書き換えを行う．続く第二段階では，ガイドラインおよび Few-shot 事例を LLM に与え，詳細な整形と再構成を実施する．本アプローチにより，単なる語彙置換に留まらない「つたわる」テキスト生成を目指す．

第 1 段階：基礎的な書き換え　まず，専門用語や難解な言い回しを一般的な表現に書き換える．本研究では，後続の処理の下地となるテキストを得るため，以下の 2 種類の指示（プロンプト）を用いる．

平易化 (Simp)　主に語彙レベルでの難易度低減を意図した指示．例：「以下の文章を専門知識がない人**でもわかるように平易化**してください」

直接的なつたわる化 (Direct)　LLM のデフォルトの生成能力に依拠した指示．例：「以下の文章を専門知識がない人**にも伝わるように**してください」

第 2 段階：ガイドラインおよび Few-shot 事例による整形　次に，第 1 段階で生成されたテキストに対し，人手による書き換え工程を模した詳細な整形を行う．具体的には，以下の二つの要素をプロンプトに含めて提示する．

ガイドライン (GL)　専門家による確認を経て独自に策定した編集要件である．具体的には，「専門用語には解説を加える」「頻出する用語はテキスト外で注釈をつける」といった情報補完や構造化に関するルールが含まれる．

Few-shot 事例 (FS)　ガイドラインの具体的な適用方法を例示するため，模擬概要文書と正解例の

1)　https://amedfind.amed.go.jp/amed/
2)　詳細な作成プロセスは [12] を参照．

表 1: 3 種の書き換え設定におけるモデル別の自動評価結果．各評価指標において全ての設定・モデルを通じた最大値を太字で示した．BLEU は **S1:Simp** や **S1:GLFS** 設定の一部で高い値を示した．ROUGE は GPT-5.2 による **S1:GLFS** 設定で高い値を示す傾向が見られる．BERTScore は手法間で大きな差は見られない．

設定	モデル	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
S1:Simp	DeepSeek	0.774	16.827	0.543	0.244	0.401
	Gemini	0.771	17.540	0.633	0.282	0.393
	GPT-5.2	**0.784**	**19.288**	0.591	0.276	**0.420**
S1:GLFS	DeepSeek	0.773	**19.607**	0.763	0.403	0.398
	Gemini	0.753	13.499	0.476	0.248	0.318
	GPT-5.2	0.753	13.060	**0.796**	**0.415**	0.307
S2:Simp+GLFS	DeepSeek	0.762	16.917	0.667	0.308	0.375
	Gemini	0.749	13.023	0.738	0.346	0.321
	GPT-5.2	0.755	14.322	0.728	0.338	0.343

ペアを提示する．In-context Learning (ICL) を通じて，LLM に対して出力すべき「つたわる」形式の具体的なパターンを学習させる．

5 実験

本研究が提案する 2 段階の書き換え手法の有効性を検証するため，LLM を用いた比較実験を行う．具体的には，模擬概要文書を入力とし，生成された「つたわる化」テキストを，メディカルライターによる書き換え文書を正解例として類似度に基づき定量的に評価する．本実験では，第一段階における指示内容（平易化または直接的なつたわる化）の差異や，第二段階における追加情報（ガイドラインおよび Few-shot 事例）による整形の有無が，最終的な出力の品質に与える影響を比較・検証する．なお，本実験において各モデル・設定で生成したテキストについては，以下の URL にて公開している[3)]．

5.1 モデル

実験には，以下の 3 種の LLM を選定した．

- **DeepSeek-V3.2** [13]: 高度な推論能力と優れたコスト効率を両立しており，医学知識の処理において高い精度が期待される．
- **Gemini 2.5 Flash** [14]: 大規模コンテキストへの対応力と一貫した生成能力に優れ，複雑なガイドラインの遵守に適している．
- **GPT-5.2 (gpt-5.2-2025-12-11)** [15]: 自然言語処理のベンチマークとして広く用いられており，ベースラインとして採用した．

パラメータの設定については，DeepSeek-V3.2 では，出力の安定性を確保するため Temperature を 0.0，最大出力トークン数を 2048 に設定した．Gemini 2.5 Flash および GPT-5.2 については，特段の指定を行わず，各 API の標準設定（デフォルト値）を用いた．

5.2 評価データと評価指標

評価には，3 節に示した模擬概要文書およびメディカルライターによる書き換え文書（正解例）のペア（計 20 ペア）を用いる．本実験では，ランダムに抽出した 2 ペアを第 2 段階の Few-shot 事例とし，残りの 18 ペアをテストデータとした．生成されたテキストの品質を多角的に評価するための指標として，意味的な類似度を測る **BERTScore** に加え，正解例との表層的な一致度を測る **BLEU**，および **ROUGE-1/2/L** を用いる．

5.3 アブレーション設定

提案手法の構成要素が書き換えの精度向上に寄与する度合いを検証するため，第 1 段階の指示内容（Simp/Direct）と第 2 段階の追加情報（GL/GLFS）を組み合わせた設定でアブレーションテストを行う．

1 段階設定 第 1 段階のみを用いる **S1:Simp** および **S1:Direct** と，第 1 段階なしで入力に直接ガイドラインを適用する **S1:GL** およびガイドラインと Few-shot 事例の両方を適用する **S1:GLFS**.

2 段階設定 (GL のみ) 第 1 段階の出力に対し，第 2 段階でガイドラインのみを適用する **S2:Simp+GL** および **S2:Direct+GL**.

2 段階設定 (GL と FS の両方) 第 1 段階の出力に対し，ガイドラインと Few-shot 事例の両方を適用する **S2:Simp+GLFS（提案手法）** および **S2:Direct+GLFS**.

3) https://github.com/sociocom/E2U

表 2: DeepSeek-V3.2 による生成事例の比較．2 段階設定 (S2) では重要な専門用語（EGFR 等）が欠落しているのに対し，1 段階設定 (S1) では用語を保持しつつ補足が行われている．太字は設定ごとに差が見られた箇所．

入力（オリジナル）	非小細胞肺がん（NSCLC）における **EGFR 変異**陽性例を対象に，既存治療薬に耐性を示す症例に対する新規の第三世代 **EGFR チロシンキナーゼ阻害剤（TKI）**の有効性及び安全性評価を目的とした多施設共同第 II 相臨床試験を実施する．
S1:Simp	非小細胞肺がんのうち，**特定の遺伝子変異がある患者**を対象に，これまでの薬が効かなくなった場合の**新しい薬**の効果と安全性を調べる臨床試験を，複数の施設で行いました．
S2:Simp+GLFS	非小細胞肺がん [1] の患者さんの中には，**特定の遺伝子 [2]** に変化がある方がいます．これまでの薬が効かなくなってしまったそのような患者さんを対象に，新しい薬の効き目と安全性を調べる臨床試験を，複数の病院で行いました． [1] 非小細胞肺がん：肺がんの一種で，比較的ゆっくりと進行するタイプです． [2] 遺伝子：体の設計図のようなはたらきをする，細胞の中の物質です．
S1:GLFS	肺がんの一種である非小細胞肺がんの患者のうち，特定の遺伝子（**EGFR**[1]) に変異がある人を対象に，臨床試験を行った．（中略）新しい第三世代の **EGFR チロシンキナーゼ阻害剤（TKI）**[2] という薬の効き目と安全性を評価することが目的だった． [1] EGFR：細胞の増殖に関わるタンパク質（受容体）の一種．この遺伝子に変異があると，がん細胞が増えやすくなる． [2] EGFR チロシンキナーゼ阻害剤（TKI）：EGFR というタンパク質のはたらきを抑えることで，がん細胞の増殖を止める薬．

6 結果と考察

主要な書き換え設定におけるモデル別の自動評価結果を表 1 に示す（全実験結果は付録表 3 参照）．

結果として，2 段階設定（**S2:Simp+GLFS**）は，1 段階設定（**S1:GLFS**）と比較して性能が低下する傾向にある．この要因は，表 2 に示す生成プロセスの違いから説明できる．S2 設定では，第 1 段階の処理で「EGFR 変異」や「チロシンキナーゼ阻害剤」といった専門用語が，「特定の遺伝子」や「新しい薬」といった一般的な語彙へ抽象化されている．その結果，第 2 段階でガイドラインを適用しようとしても，詳細な解説を付与すべき専門用語自体が消失しており，構造的な情報欠落を引き起こしている．これに対し **S1:GLFS** では，原文の専門用語を保持したまま，直後に用語解説への参照を付与する形で，情報の等価性を損なわない「つたわる化」が実現されている．このことから，専門情報の伝達においては，段階的な簡略化よりも，最終形式へ直接生成するアプローチが有効であることが示唆される．

また，モデルごとの生成特性もこれらの評価に寄与している．DeepSeek は，原文の意味を保ちつつ，ガイドラインに沿って文体や構成を柔軟に再構築する能力に長けており，これが複雑な編集を要する **S1:GLFS** 設定での高評価に繋がった．一方，Gemini については，1 文に対し過度に注釈を付与する傾向が見られ，これが n-gram ベースの指標におけるスコア伸長を妨げた可能性がある．なお，GPT-5.2 が **S1:Simp** 設定などで安定した値を記録しているのは，原文の構造や意味を忠実に維持する保守的な傾向が強いためと考えられる．

なお，本評価で用いた正解例は 1 名のメディカルライターによって作成されたものであるが，「つたわる」化を実現した書き換えに唯一の正解は存在せず，読者属性や文脈によって多様な表現が許容され得る．したがって，特定の正解例との一致度に基づく今回の評価は，あくまで限定的な評価に留まる．

7 おわりに

本研究では，医学研究テキストを非専門読者にとって理解可能な形で提示するための書き換えの概念として「つたわる化」を定義し，LLM を用いた生成の可能性を検討した．その結果，編集方針や具体例を与えた上で直接生成を行うアプローチが「つたわる化」において有効であることが示唆された．

今後の課題としては，より大規模かつ多様な疾患領域を含むデータセットの構築や，被験者実験による理解度評価，ならびに医療情報の正確性を担保するための仕組みの検討が挙げられる．本研究で目指す「つたわる化」は，専門家間の情報共有にとどまらず，多様な立場の人々が医療を含む科学技術を理解可能な形で受け取れる社会の実現に寄与することを目指すものである．

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，JSPS 研究スタート支援 JP25K24412 の補助を受けて行った．

参考文献

[1] Gillian Tett. **The Silo Effect: The Peril of Expertise and the Promise of Breaking Down Barriers**. Simon & Schuster, New York, 2015.

[2] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. Automated lay language summarization of biomedical scientific reviews. In **Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)**. Association for the Advancement of Artificial Intelligence, 2021.

[3] Thomas Goldsack, Seán Cartright, Jon Chamberlain, and Massimo Poesio. Making science simple: Corpora for the lay summarisation of scientific literature. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[4] Thomas Goldsack, Jon Chamberlain, and Massimo Poesio. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In **Proceedings of the 2024 BioNLP Workshop**, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[5] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[6] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. Controllable sentence simplification. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4689–4698, Marseille, France, May 2020. European Language Resources Association.

[7] Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 712–718, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Brian Ondov, Kush Attal, and Dina Demner-Fushman. A survey of automated methods for biomedical text simplification. **Journal of the American Medical Informatics Association**, Vol. 29, No. 11, pp. 1976–1988, 2022.

[9] Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. A large-scaled corpus for assessing text readability. **Behavior Research Methods**, Vol. 55, p. 1023–1047, 2022.

[10] Rahul C. Salvi, Tanvir Alam, Sambuddha Ghosh, Hrishikesh Hegde, Soham Ghosh, Alexis Palmer, and Danielle Mowery. Towards understanding LLM-generated biomedical lay summaries. In **Proceedings of the 2025 Workshop on Clinical Natural Language Processing for Health (CL4Health)**, Mexico City, Mexico, 2025. Association for Computational Linguistics.

[11] Ziqi Luo, Shervin Radpour, Shihua Wang, Sarah Yeston, Yining Ruan, and Mark Gerstein. The lay person's guide to biomedicine: Orchestrating large language models. **arXiv preprint arXiv:2402.13498**, 2024.

[12] 倉本真菜, 永井宥之, 山田恵子, 井出博生, 早川雅代, 西山智弘, 若宮翔子, 荒牧英治. 医学研究の「つたわる化」を目指したデータセット構築. 人工知能学会第二種研究会資料, Vol. 2025, No. AIMED-015, p. 8, 2025.

[13] DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. **arXiv preprint arXiv:2512.02556**, 2025.

[14] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google, 6 2025. Accessed: 2026-01-08.

[15] OpenAI. Introducing gpt-5.2, 2025. Accessed: 2026-01-08.

A 付録

A.1 全書き換え設定における自動評価結果

表 3: モデル・設定別の自動評価結果

モデル	設定	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
DeepSeek	S1:Simp	0.774	16.827	0.543	0.244	0.401
	S1:Direct	0.766	15.394	0.563	0.240	0.382
	S1:GL	0.747	11.145	0.742	0.327	0.311
	S1:GLFS	0.773	**19.607**	0.763	0.403	0.398
	S2:Simp+GL	0.748	12.743	0.647	0.266	0.337
	S2:Simp+GLFS	0.762	16.917	0.667	0.308	0.375
	S2:Direct+GL	0.745	12.764	0.641	0.256	0.339
	S2:Direct+GLFS	0.760	15.148	0.627	0.270	0.360
Gemini	S1:Simp	0.771	17.540	0.633	0.282	0.393
	S1:Direct	0.762	15.407	0.648	0.276	0.376
	S1:GL	0.730	8.894	0.777	0.361	0.251
	S1:GLFS	0.753	13.499	0.476	0.248	0.318
	S2:Simp+GL	0.755	11.079	0.728	0.325	0.296
	S2:Simp+GLFS	0.749	13.023	0.738	0.346	0.321
	S2:Direct+GL	0.729	9.161	0.727	0.303	0.280
	S2:Direct+GLFS	0.748	12.600	0.722	0.315	0.333
GPT-5.2	S1:Simp	**0.784**	**19.288**	0.591	0.276	**0.420**
	S1:Direct	0.777	18.596	0.617	0.286	0.402
	S1:GL	0.750	12.279	0.750	0.347	0.309
	S1:GLFS	0.753	13.060	**0.796**	**0.415**	0.307
	S2:Simp+GL	0.753	13.123	0.711	0.320	0.325
	S2:Simp+GLFS	0.755	14.322	0.728	0.338	0.343
	S2:Direct+GL	0.747	13.655	0.705	0.313	0.329
	S2:Direct+GLFS	0.754	15.040	0.704	0.324	0.356

A.2 ガイドライン

1. 全体的な方針

- **対象読者**: 成人している医療従事者ではない方（生物に関する知識がない方を想定）．4 年制大学文系学部卒業程度．
- **最終的な目標**: 専門知識がなくても，研究の内容や成果の概要が理解できる文章にする．

2. 構造・表現に関するルール

- **文体**: 文体を保持してください（常体なら常体，敬体なら敬体）．
- **語順**: 必要に応じて語順を変更しても構いません．
- **削除**: 研究理解に不要な表現，文（検査人数，検査方法など）は削除して構いません．
- **断定的な表現の回避**: 確定的な情報でない場合は断定しない表現を保持または調整してください．例えば，「有意である」は日常では使われないため，「はっきりとした違いが見られる」「無視できないほどの差がある」などに言い換えてください．あくまで「可能性がある」というニュアンスを保ってください．

3. 用語に関するルール

- **用語のレベル**: 平易な表現の後に，丸括弧 () 内に元の専門用語を入れてください．
  例： hogehoge 遺伝子の量を正確に測る技術（RT-qPCR 法）
- **頻出する語句**: 繰り返し出現する重要な語句は脚注を入れてください．
- **アルファベット略称**: 初出時に正式名称を記述した上で説明してください．
- **新語**: 新語自体の説明は不要ですが，含まれる専門用語は脚注で説明してください．
- **「〜の〜の」**: 「A の B の C」のように「の」が連続する場合は，後半を脚注で説明し言い換えてください．

4. 補足

詳細な検査や解析方法は，冗長になることを避けるため可能な限り平易にまとめてください．架空データのため，大きな矛盾がなければ推測で書き換えを行ってください．