

# 医療事故・ヒヤリハットに関する 要因・対策案生成ベンチマークの提案

長谷山 優菜<sup>1</sup> 伊藤 友貴<sup>1</sup> 坂地 泰記<sup>2</sup> 野田 五十樹<sup>1</sup>

<sup>1</sup> 北海道大学大学院 <sup>2</sup> 三井物産株式会社

haseyama.yuna.n3@elms.hokudai.ac.jp Tomok.Ito@mitsui.com

{sakaji,i.noda}@ist.hokudai.ac.jp

## 概要

医療現場での利活用の需要が高まる中、医療事故が起こった際、その原因を突き止め、迅速に対策を講じるための大規模言語モデルは有用だと考えられる。医療現場で利用可能な大規模言語モデルを考えると、その能力を評価するベンチマークは必要不可欠だが、現状は十分とは言えない。このような背景のもと、本研究では、日本医療機能評価機構による医療事故情報収集等事業によって収集、提供されている事故・ヒヤリハット事例のデータからデータセットを構築し、医療事故の内容から、背景・要因、改善策の生成を評価するためのベンチマークを提案する。

## 1 はじめに

近年、大規模言語モデル (Large Language Models; LLMs) や生成 AI (Generative AI) への関心が高まっており、これらの技術をさまざまな分野に応用した研究が盛んである。特に、自然言語処理やデータ解析分野の進歩は目覚ましく、その応用範囲は日々拡大しており、臨床の場を始め、医療現場での利活用の拡大も期待される。医療現場での利活用の方法として、一つ、医療安全対策推進への活用が考えられる。医療安全対策の推進の上では、医療事故やヒヤリ・ハットが起きた際にその原因をいち早く突き止め、迅速な対策を行い、再発防止策や改善案を洗い出すことが重要である。実際、このような背景の下、厚生労働省では2004年より、医療事故情報収集等事業にて、医療事故等の有害事象を恒常的に収集すると共に、医療機関から報告された事故等事案やヒヤリ・ハット事例を分析し提供することにより、医療安全対策に有用な情報を広く医療機関及び国民に対し、情報を公開している。更に、事例報告の質

を高めることを目的とし、参加医療機関を対象に医療における事故の分析手法を学ぶ演習を中心とした研修会を開催している<sup>1)</sup>。

医療事故やヒヤリ・ハットに関する原因や再発防止策・改善案策を漏れなく洗い出すには医療安全に関する専門知見を活用しなければならないことも多い。また、医療現場は日々忙しく、人的リソースも限られることから、少ない労力で原因や再発防止策・改善案策を「漏れなく」かつ「高品質」に洗い出せるような大規模言語モデルは専門医や人的リソースが限られる地方の医療現場等においては特に有用であると期待される。その一方、上記のような「医療事故・ヒヤリハットに関する要因・対策案生成」に関する評価ベンチマークは少なく、そもそもとして大規模言語モデルが上記のような原因や再発防止策・改善案策の生成をできるかどうかを測ることが難しい。

そこで、本研究では、医療事故情報収集等事業によって収集、提供されている報告書<sup>2)</sup>から、医療事故の【具体的内容】【背景・要因】【改善策】に加えていくつかの要素を含んだデータセットを構築する。本研究で構築するデータセットを Japanese Medical Incident Dataset (JMID) と呼称する。加えて、医療事故の内容を入力としたときの背景・要因や改善策を生成するモデルを評価するためのベンチマークを提案する。

今後、医療安全に関する大規模言語モデルの開発も進められると考えられる中で、本研究で提案するベンチマークが存在すると、大規模言語モデルの評価が可能になり、開発の後押しとなると考えられる。特に医療安全に関する日本語のデータセットは乏しいため、本研究の価値は非常に高いと考えられる。

1) [https://www.mhlw.go.jp/stf/newpage\\_22786.html](https://www.mhlw.go.jp/stf/newpage_22786.html)

2) <https://www.med-safe.jp/contents/report/index.html>

本研究の貢献を以下に一覧で示す。

- 医療事故情報収集等事業によって収集、提供されている報告書から、JMID を構築した。
- JMID を用いて、医療事故の内容を入力としたときの背景・要因や改善策を生成するモデルを評価するためのベンチマークを提案した。

## 2 関連研究

医療タスクに関するベンチマークの構築やデータ資源の構築に関する関連研究として、診察録の自由記載から重要な情報を抽出することを目的としたアノテーション付きの症例報告コーパスである iCorpus<sup>3)</sup> の構築 [1, 2] が挙げられる。[3] では、臨床現場で実際に用いられる略記や英語名を含む語を抽出したデータを万病辞書<sup>4)</sup>として公開している他、同研究グループにて構築された百薬辞書<sup>5)</sup>も公開されている。

また、2018 年から 2022 年までの国内の医師国家試験過去問の QA データセットである IgakuQA<sup>6)</sup> [4] 等も存在する。

更に、[5] では、日本語医療 LLM の評価用データセットである JMED-LLM<sup>7)</sup>を構築し、8 つのタスクを複数の LLM で解かせる評価実験を行っている。

海外においても、医学部入学試験の QA データセットである MedMCQA<sup>8)</sup> [6] を用いたベンチマーク<sup>9)</sup>が存在しており、また、医師国家試験の QA データセットである MedQA [7] を用いたベンチマーク<sup>10)</sup>では、ファインチューニングされた専門のモデルの値を GPT-4 が超える結果も出ている。

しかしながら、医療安全分野のデータセットは現状では不十分である。そこで、本研究では、医療事故に関するデータセットを作成、ベンチマークを作成し、複数の LLM で評価実験を行う。

## 3 データセットの作成

公益財団法人日本医療機能評価機構の医療事故情報収集等事業が提供している PDF 形式の報告書<sup>11)</sup>

3) <https://ai-health.m.u-tokyo.ac.jp/home/research/corpus>

4) <https://sociocom.naist.jp/manbyou-dic/>

5) <https://sociocom.naist.jp/hyakuyaku-dic/>

6) <https://github.com/jungokasai/IgakuQA>

7) <https://github.com/sociocom/JMED-LLM>

8) <https://github.com/MedMCQA/MedMCQA>

9) <https://medmcqa.github.io/>

10) <https://paperswithcode.com/sota/question-answering-on-medqa-usmle>

11) <https://www.med-safe.jp/contents/report/index.html>

から、表形式で【具体的内容】【背景・要因】【改善策】が記されているものを収集し、Excel 形式に変換した。報告書には、【事故の内容】【背景・要因】【改善策】以外にも様々な情報がまとまっており、今回は 2006 年の第 7 回報告書から、2020 年の第 64 回報告書までの 3924 件をまとめた。

データセットには、

- 事故の内容
- 背景・要因
- 改善策
- 分類（薬剤間違い/手術/実施した行為の間違いなど）
- 専門分析班の議論（第三者の議論）
- 情報の流れ（どのタイミングでの発生か）
- 当事者職種（職種経験年数）
- 事故の程度（障害なし/死亡など）
- 発生要因（確認を怠った/知識が不足していたなど）

などの情報が存在している。なお、収集したデータ内には「不明」や「未記載」というものも存在しているが、これらに対して特別な前処理は行わず、そのままデータとして保存している。

## 4 タスクの提案

本研究では、医療事故に関する大規模言語モデルを評価するベンチマークを作成するために、3 種類のタスクを提案する。3 種類のタスクは、医療事故内容から、その背景・要因、改善策を生成するものであり、大規模言語モデルを様々な角度から評価するためのものである。以下に、提案する 3 種類のタスクを示す。

**All** 事故内容を入力し、出力として背景・要因、改善策を一括生成させるタスク

**Each** 事故内容を入力し、背景・要因、改善策を段階的に生成させるタスク

**Few-shot** All を few-shot (n=5) にしたタスク

それぞれのタスクについて、図 1, 図 2, 図 3 に示す。

医療現場は多忙であり、人的リソースは限られている。そのような状況下で、医師や看護師が時間を割いて医療事故に関する報告書を執筆している。そのため、報告書の執筆は医療従事者の高い負担になっていると考えられる。もし大規模言語モデルによって医療事故の原因や再発防止策・改善策案を漏

表 1 作成したデータセットの実際の例 (報告書第 7 回)

医療事故の内容	背景・要因	改善策	...
注射箋「ソル・メドロール...	業者から薬剤をもらい...	・調剤時は慌てずに行う。...	...
注射薬の払い出し業務で、...	15時30分以降に各病棟から...	・注射薬の監査を行って払い出し...	...
⋮	⋮	⋮	⋮

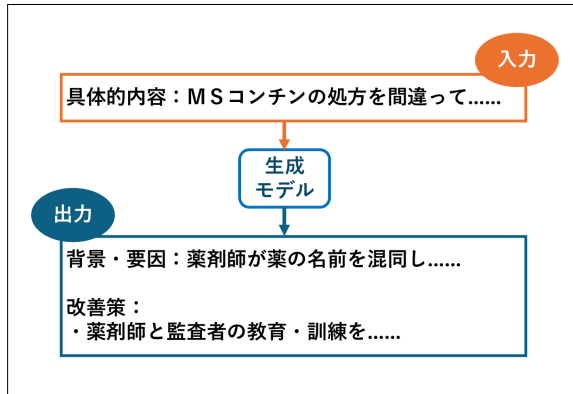


図 1 タスク All の概要図

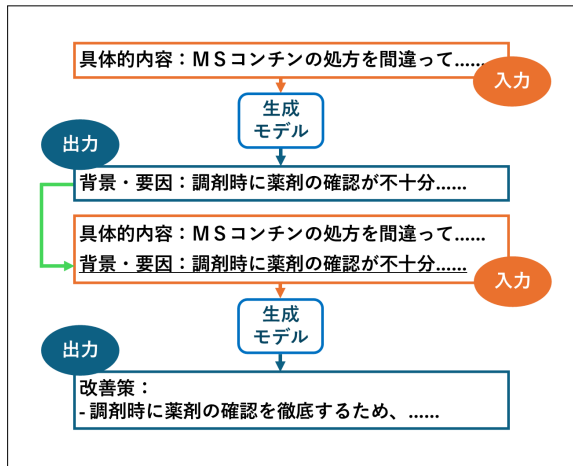


図 2 タスク Each の概要図

れなく洗い出すことができるなどすれば、医療従事者の時間に余裕が生まれ、医療安全の向上に繋がることが期待される。

## 5 大規模言語モデルの評価

GPT-4o<sup>[12]</sup>, Mistral<sup>[13]</sup>, Llama3.3<sup>[14]</sup>を用いて大規模言語モデルの評価を行う。使用したデータは、2006年第8回から2010年第21回報告書に記載されている2017件であり、生成した【背景・要因】と【改善策】をそれぞれ評価する。評価方法はBERTScore<sup>[15]</sup>[8]を用い、その平均値を算出する。Few-shot (n=5)では、一番初めに5件をランダムで選び、全件同じ例

12) <https://openai.com/index/hello-gpt-4o/>  
 13) <https://mistral.ai/news/mistral-large-2407/>  
 14) <https://www.llama.com/>  
 15) [https://github.com/Tiiiger/bert\\_score](https://github.com/Tiiiger/bert_score)



図 3 タスク Few-shot の概要図

文を使用する。

## 5.1 結果

実験結果は表 2, 表 3 に示す。医療事故内容から背景・要因を生成させる際は、GPT-4o が全体的に値が高く、続いて Mistral の All や Each となっている。一方、医療事故内容から改善策を生成させる際は、Llama3.3 の値が全体的に高く、GPT-4o, Mistral の順となっている。

また、Mistral の Few-shot (n=5) では、期待した出力が得られない場合が存在した。具体的には、入力プロンプトで提示した5つの具体例に対して5つとも出力をしてしまい、本来生成させたい部分が生成できないという現象がいくつかみられた。また、第 5.2 節の考察にて後述するが、GPT-4o での Each の際、生成する過程でフィルターに引っかかってしまう例が存在した。今回はこれらの値を全て 0 として計算し、生成できないものを弾いて平均値を取った結果も算出し、括弧の中に記した。加えて、GPT-4o の Few-shot における生成結果の例を図 4 に示す。

## 5.2 考察

医療国家試験の QA データである MedQA を用いたベンチマークでは、ファインチューニングされた専門のモデルの値を GPT-4 が上回っていることから、GPT-4 には潜在的な医療知識の豊富さがあると

表2 背景・要因 BERTScore (括弧内の数値は、生成できなかったものを除いた結果)

	All	Each	Few-shot (n=5)
	precision - recall - f1	precision - recall - f1	precision - recall - f1
GPT-4o	<b>0.705 - 0.713 - 0.708</b>	0.706 (0.713) - 0.693 (0.700) - <b>0.698 (0.706)</b>	<b>0.708 - 0.717 - 0.712</b>
Mistral	<b>0.709</b> - 0.707 - 0.707	0.684 (0.688) - <b>0.703</b> (0.706) - 0.692 (0.696)	0.622 (0.642) - 0.661 (0.683) - 0.640 (0.661)
Llama3.3	0.680 - 0.674 - 0.676	0.698 - 0.685 - 0.691	0.686 - 0.712 - 0.698

表3 改善策 BERTScore (括弧内の数値は、生成できなかったものを除いた結果)

	All	Each	Few-shot (n=5)
	precision - recall - f1	precision - recall - f1	precision - recall - f1
GPT-4o	0.687 - <b>0.719</b> - 0.702	0.666 (0.673) - 0.713 ( <b>0.721</b> ) - 0.689 (0.696)	<b>0.692</b> - 0.731 - <b>0.710</b>
Mistral	0.678 - 0.712 - 0.694	0.645 (0.648) - 0.707 (0.711) - 0.674 (0.677)	0.656 (0.677) - 0.681 (0.703) - 0.668 (0.689)
Llama3.3	<b>0.694</b> - 0.712 - <b>0.703</b>	<b>0.687 - 0.715</b> - 0.701	0.691 - <b>0.732</b> - <b>0.710</b>

入力
MSコンチンの処方間違えてオキシコンチンを調剤し、監査者も誤りに気付かずそのまま患者に渡ってしまった。薬剤部長が麻薬の帳簿を記入する際に数量が合わず間違いが発見された。薬剤部長は主治医・患者に連絡した。患者は以前より処方されていたMSコンチンも含め薬をずっと服用していなかったため、今回のオキシコンチンも服用していなかった。

	出力	正解テキスト
背景・要因	MSコンチンとオキシコンチンの薬剤名が似ているため調剤時に混同しやすい。また、監査者が処方内容を十分に確認せず、誤りに気付かなかったことが要因である。さらに、患者が薬を服用していなかったため、誤りが早期に発見されなかった。	薬剤の採用や保管の問題点。MSコンチン（3規格）、オキシコンチン（3規格）採用しているが、名称が類似していて紛らわしい。調剤および監査時の注意力不足、認識不足。
改善策	<ul style="list-style-type: none"> <li>・薬剤名が似ている薬剤については、調剤時に特に注意を払い、二重チェックを徹底する。</li> <li>・監査者は処方内容を厳密に確認し、誤りがなければ慎重にチェックする。</li> <li>・患者に対して薬の服用状況を定期的に確認し、服用の徹底を促す。</li> <li>・薬剤部内での教育・研修を強化し、薬剤の取り扱いに関する知識を共有する。</li> </ul>	<ul style="list-style-type: none"> <li>・薬品名、規格に印をつけて調剤者、監査者共に注意・確認する。</li> <li>・名前が類似している薬剤および規格等、採用薬剤を再検討する。</li> </ul>

図4 Few-shot における出力結果の例

思われ、GPT-4o も同様であると考えられる。医療事故内容から背景・要因を生成する際に GPT-4o の値が高いのはこの為の可能性があり、一方、改善策の生成の際には医療知識はさほど必要ない可能性がある。

Llama3.3 は、GPQA<sup>16)</sup>[9] という大学院レベルの QA ベンチマークで GPT-4o よりも数値が高く、このため、改善策を生成する際に GPT-4o よりも良い値になっていると考えられる。また、3つのタスクのそれぞれに関しては特別有意な差はみられなかったが、Few-shot での結果が僅かに高いことから、few-shot 学習に一定の効果があると考えられる。更に本研究では、安全医療に LLM を活用する際には、

violence フィルターが問題になることも分かった。例えば、付録の表4から、「ホルマリン」という単語が violence フィルターの対象になっているように見受けられる。しかしながら、「ホルマリン」は医療現場で使われる液体であり、この単語を入力出力に使えないとなると、医療安全分野への大規模言語モデルの活用の障害となってしまう。また、Mistral を Azure AI Foundry<sup>17)</sup>経由で使用した場合、フィルターに引っかかることがあった。フィルターによる出力制御なしでの実験は今後の課題としたいと考えている。

## 6 結論

本研究では、大規模言語モデルが医療事故の原因や再発防止策・改善案策の生成をできるかどうかを測ることが難しいという課題を解決するため、医療事故に関するベンチマークを作成すると共に、実際に作成したベンチマークを用いて3つの LLM に関する性能評価を行った。医療事故内容から背景・要因、改善策を生成するタスクにおいて、医療事故の背景・要因を生成する際は GPT-4o の性能がよく、医療事故の改善策を生成する際は Llama3.3 の性能が良いという結果になった。今後の課題として、データセットの拡充に加えて、新たなタスクの提案を行っていききたい。

16) <https://huggingface.co/datasets/Idavidrein/gpqa>

17) <https://azure.microsoft.com/ja-jp/products/ai-foundry/?mssockid=3e8f23e6583b6f390683306059ca6ed8>



## 謝辞

JST さきがけ JPMJPR2267 の助成を受けたものです。

## 参考文献

- [1] Yoshimasa Kawazoe Emiko Shinohara, Daisaku Shibata. Development of comprehensive annotation criteria for patients' states from clinical texts. **J Biomed Inform**, 2022.
- [2] 篠原恵美子, 河添悦昌, 柴田大作, 嶋本公德, 関倫久. 症例報告に対する網羅的な所見アノテーションのためのアノテーション基準の構築. **医療情報学**, Vol. 42, No. 1, pp. 3–15, 2022.
- [3] Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.
- [4] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations, 2023.
- [5] 福島拓也, 久祥平, 竣太郎, 若宮翔, 荒牧英治. LMED-LLM: 日本語医療 LLM 評価データセットの公開, 2024. <https://speakerdeck.com/fta98/jmed-llm-ri-ben-yu-yi-liao-llmping-jia-detasetutonogong-kai>.
- [6] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, **Proceedings of the Conference on Health, Inference, and Learning**, Vol. 174 of **Proceedings of Machine Learning Research**, pp. 248–260. PMLR, 07–08 Apr 2022.
- [7] Jin Di, Pan Eileen, Oufattole Nassim, Weng Wei-Hung, Fang Hanyi, and Szolovits Peter. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **Applied Sciences**, Vol. 11, No. 14, p. 6421, 07 2021.
- [8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.
- [9] 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. In **First Conference on Language Modeling**, 2024.

## A violence フィルターに引っかかった例

表 4 フィルターに引っかかったもの（GPT-4o）（抜粋）

右前立腺生検後の組織を、左右2つのホルマリン容器に分けて入れるはずだったが右に2つとも入れてしまった。病理室で発見された。前立腺生検の場合、右の前立腺を6ヶ所採取するが一つ採取ごとに右と記載されたろ紙に6ヶ貼り付け、上に1. 2. 3. 4. 5. 6. とナンバーを記載した。左も同様に行うことになっていた。右のホルマリン容器の中に左と記載したろ紙がはっきりわかり、左右の鑑別ができ病理組織検査は支障なくできた。
手術室から届いた、組織の検体をホルマリンにつけて固定する作業のため切り出しの準備を行った。同じ名前の検体があり、一方は「小腸」で、一方は「肺」であった。両方とも「A氏」の名前が貼られていた。伝票で確認した所、A氏の検体は「肺」であった為、ラベルは正しかった。「小腸」の検体は「B氏」の物と判明した。固定処理に立ち会った技師に確認し、もう1度伝票と検体を合わせた上で名前のラベルを訂正した。
産婦人科から手術未固定の臓器が、次の日朝に3症例分、病理検査室にダムウェイター（小荷物運搬用昇降機）で提出された。臓器を入れた袋が5袋と多かったために、固定する者が気づくだろうと思い、とりあえず平行した浅い段ボール箱に入れ作業台の脇に置いた。その後、固定する者は気付くことなく、週末を含め3日間室温で放置された状態で、月曜の早朝に気付きホルマリンを入れて固定をし、組織標本を作製したが、3症例のうち1例に組織診断の評価に弊害が生じた。
⋮

医療事故内容から背景・要因、改善策を一つずつ生成させる過程で、GPT-4o と Mistral はフィルターに引っかかり、生成ができないものがあった。GPT-4o のフィルターには、'hate', 'self harm', 'sexual', 'violence', という複数のフィルターがあり、今回は"violence"のフィルターに引っかかった。

表 4 はフィルターに引っかかった一例である。