複数の大規模言語モデルによる相互評価機構の提案と評価
—学校の救急処置の課題解決を目指して—
A Mutual Evaluation Framework with Multiple Large Language Models Proposal and Assessment
- Aiming to solve the problem of first aid in schools-

博田 光[*1]，讃岐 勝[*2]
Hikaru HAKATA[*1], Masaru SANUKI[*2]
[*1] 筑波大学大学院医学学位プログラム
[*1] Doctoral Program in Medical Sciences, University of Tsukuba
[*2] 筑波大学医学医療系
[*2] Institute of Medicine, University of Tsukuba
Email: s2430403@u.tsukuba.ac.jp

あらまし：学校現場の緊急時，養護教諭を中心とした応急処置の記録が求められる．緊急下での記録作業は困難を伴うため，その支援として音声入力と LLM を活用し，記録業務の負担軽減と標準化を目指した救急処置記録支援システム SCC を開発した．本研究では，PoLL という手法に基づき，複数 LLM を用いて互いの出力を相互評価・スコアリングする構成を導入した．これにより，ユーザに対して信頼性の高い応急処置記録の提示を可能とした．システムの構成と処理手順を示し，看護師による専門的評価を議論する．
キーワード：学校救急処置録，入力の自動化，校務の情報化，生成系 AI，教育 DX

1. はじめに

大規模言語モデル（以下, LLM）は幅広い分野で利用されている. LLM はモデルによって特性が異なり，同様の入力でも出力が同じになることはかなり少ない．また，誤った情報や偏った情報を出力することも LLM の利用にあたっての課題である．

学校における緊急時対応では，救急処置録への正確な記録が求められるが，緊急下では記録が困難であり，教職員間の連携や役割分担が重要である．特に養護教諭が少人数で対応している現状では，記録業務の情報化が求められている．著者はこれらの課題を解決するために，応急処置を記録・共有できるシステム School Care Check(1)（以下, SCC）を開発した．このシステムを養護教諭を含めた教職員が利用することにより，救急処置の音声から医療機関で用いられる SOAP という記録方法に分類され，記録業務の負担軽減や記録の質の標準化が期待できる．

本研究では Panel of Large Language Models(2)（以下, PoLL）に基づく LLM の評価手法をシステムのバックエンドに組み込み，複数の LLM によって高評価とされた出力をユーザに提示するという新たなシステム構造を提案し，評価する．

2. 実装

応急処置記録の質と一貫性を向上させるため，PoLL の評価機構を統合したシステムを構築した．
フロントエンド：Django をベースに，養護教諭や教職員が操作しやすいモバイル対応 UI を実装．
バックエンド: 音声認識から LLM 評価，記録保存までを統合的に処理する．
セキュリティ：個人情報や医療情報を扱うため，音声データと記録情報はローカルサーバで完結し，外部のクラウドサービスとは非連携とする設計とした．

使用した LLM は次の 4 つである．Elyza:8b[i](elyza)，LLaMa3.3:70b [ii] (llama)，DeepSeek-R1-Distill-Qwen-14B-Japanese[iii](DS-R1)，Gemma3:27b[iv] (gemma)

3. 提案手法

3.1 PoLL の概要

本研究では, LLM の出力を複数モデル間で相互評価する PoLL という手法に着目した. PoLL は，複数の LLM に同一の入力を与え，それぞれが出力を生成した後，他の LLM がその出力を評価・スコアリングする．全 LLM が相互に評価を行うため，最終的に全体として最も高い評価を得た出力をスコア順でユーザに提示することができる（図 1）．このアンサンブル的な仕組みにより，単一モデルのハルシネーションや偏りのある出力を抑制する．

3.2 SOAP 分類と記録の構造化

SCC では，医療・看護の現場でも用いられる SOAP に基づいて，応急処置の記録を分類・保存する．この構造により，後から情報を参照する際にも内容の把握が容易であり，病院や保護者への情報連携も円滑

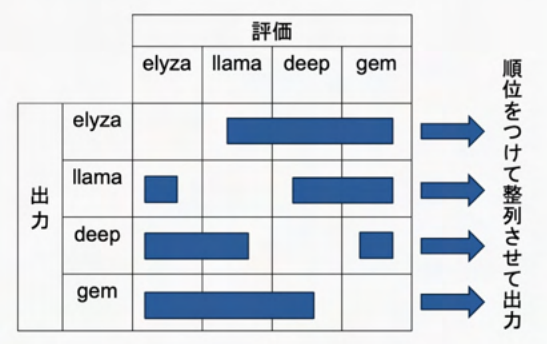

図 1. 処理フロー

	PoLLによる順位				A				B				C				D			
	elyza	llama	deepseek	gemma	elyza	llama	deepseek	gemma	elyza	llama	deepseek	gemma	elyza	llama	deepseek	gemma	elyza	llama	deepseek	gemma
1	1	3	2	4	4	3	1	2	3	1	2	4	4	3	1	2	4	2	1	3
2	3	4	1	2	4	2	1	3	3	1	2	4	4	2	3	1	3	4	1	2
3	3	2	4	1	4	3	2	1	2	3	4	1	4	3	2	1	2	3	1	4
4	3	4	1	2	4	1	3	2	2	4	3	1	3	4	2	1	1	2	3	4
5	1	4	3	2	4	3	1	2	4	3	1	2	4	3	2	1	1	4	2	3
6	3	4	2	1	4	3	1	2	4	1	2	3	4	3	2	1	1	3	2	4
7	3	4	1	2	4	3	2	1	4	3	1	2	4	3	2	1	1	2	3	4
8	4	3	2	1	4	3	2	1	3	4	2	1	4	3	1	2	2	1	3	4
9	3	2	4	1	4	3	1	2	4	2	3	1	3	2	1	4	4	1	3	2

表 1. 評価結果

になる. PoLL を用いることで, 曖昧な表現や情報の欠落を最小限に抑え, より現実的かつ自然な記述を自動生成することが可能となる.

3.3 SCC との統合と処理フロー

本研究では, 応急処置記録システム SCC に PoLL を統合し，実装した.

(1) 音声を録音
(2) 音声認識エンジンによってテキストに変換
(3) 変換されたテキストを 4 つの LLM に入力
(4) 各 LLM が SOAP に分類された出力を生成
(5) 各 LLM が他方の出力を正確性，可読性，一貫性，適切な分類の 4 項目で評価しスコア付与
(6) スコア順に結果をユーザに表示

音声の変換からユーザに提示する一連の処理は音声の長さにも依存するが，1 分以内で出力される．これにより，事故現場での記録業務の負担を最小限にしつつ, より正確かつ標準化された救急処置記録の作成されることが期待される.

4. 評価

本研究では，高校生に対して行った救急処置時のデモンストレーションの音声をもとに，提案される SOAP の順位が学校医療関係者に与える影響を検証した．日常的に SOAP を用いて業務を行う医療関係者を対象に評価者 A〜D の 4 名に対して，提示される出力の順位の一致度を Poll で出力された順位を基準としてスピアマンの順位相関係数[3](ρ)を用いて定量的に評価した．結果は表 1 の通りである．A: 平均 $\rho = 0.178$, B: 平均 $\rho = 0.289$, C: 平均 $\rho = 0.222$, D: 平均 $\rho = -0.133$ 評価者 B，C には弱い正の相関が確認でき，A と D にはほとんど相関はなかった.

5. 考察

幅広いニーズへの対応：一致度は一見して限定的に見えるが，各評価者がある程度独自の観点で記述を行っていたことを意味している．むしろ記述内容に分散が見られたことは，幅広い表現や観点が生成されていたことの証左であり，記録者の多様なニーズに応じた情報提供が可能であることを示している.

個別最適化：評価者間でも提示された記録の好ましさにばらつきが見られた．この差異は，評価者の専門性や実務環境，個々の判断基準に起因すると考えられる．応急処置記録の本来の目的は，生徒の状態や処置内容を他者に正確かつ簡潔に伝達することである．したがって,「どのような記録が記録者にとって実務的に使いやすいのか」という観点で詳細な聞き取りを行い，分類における重要な要素を抽出し，評価項目に追加，評価項目に適切な重み付けをすることでその結果を分類段階のプロンプト設計に反映することができる．これにより現場で活用しやすいシステムの実現が期待される.

記録提示の柔軟性：本研究ではスコア順で出力を提示する方式を採用したが，自由記述から「簡潔さ」や「実務性」を重視する傾向が確認された．単一のスコア基準による提示ではなく，例えば「簡潔」「詳細」「客観性重視」など，記録者のニーズに応じて出力をフィルタリング・選択できるような提示方式の導入も検討すべきであると考えられる.

6. 終わりに

本研究では，学校現場における救急処置記録の支援を目的として，PoLL を用いた記録生成・提示システムを提案・実装した．評価者の好みにばらつきがあることから，出力に一定の分散があることはむしろ有効であり，記録者の多様なニーズに応じた柔軟な記録提示が求められることも明らかとなった.

PoLL の評価に評価者のニーズを反映させられる機構の導入，および，多様な学校医療関係者からの意見を抽出し，実用性と信頼性をさらに高めることが今後の課題である.

参考文献

(1) 博田 光, 讃岐 勝：“学校救急処置記録における生成系 AI による分類の可能性”, JSiSE Research Report, 教育システム情報学会研究会, 40(1), 32-39 (2025).
(2) P. Verga, S. Hofstaetter, S. Althammer *et al.*: “Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models”, Association for Computational Linguistics Rolling Review, Submission1601(2024)
(3) C. Spearman: The proof and measurement of association between different sets of similarites between men. British Journal of Psychology, 1, pp.149-171(1904).

[i] https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
[ii] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[iii] https://huggingface.co/cyberagent/DeepSeek-R1-Distill-Qwen-14B-Japanese
[iv] https://huggingface.co/google/gemma-3-27b-it