医薬品有害事象プラットフォームへの言語生成 AI の搭載による安全性情報の収集・分析の効率化

浅野 弘斗
大阪大学 大学院薬学研究科

1. はじめに

医薬品は疾病治療に不可欠である一方、有害事象という予期せぬ健康被害を引き起こすリスクを内包する。抗がん剤投与時の悪心・嘔吐のように医薬品との因果関係が明らかな「副作用」は添付文書に記載され、予防や対処法が確立されている。しかし臨床現場では、因果関係が明確でなく、予期の難しい「有害事象」が発生しうる。これらの情報を適切に収集・評価することは、患者の安全確保において極めて重要である[1]。

各国の規制当局は有害事象報告を収集・公開しており、代表例として日本の医薬品医療機器総合機構（PMDA）による医薬品副作用データベース（JADER）や、米国食品医薬品局（FDA）の有害事象報告システム（FAERS）が挙げられる。特に FAERS は、2025 年 4 月時点で約 3,018 万件の報告を含む世界最大規模のデータベースであり、医薬品の市販後安全性監視、とりわけ臨床試験では検出困難な稀な事象や長期的な影響を把握するための不可欠な情報源となっている。

薬剤師をはじめとする医療従事者には、これらの安全性情報を収集・評価し、治療方針の決定に活かす役割が期待される。しかし、従来の添付文書やガイドライン中心の情報収集に加え、FAERS のような大規模データベースを自在に活用するには、データベース処理、プログラミング、統計解析といった専門的なスキルが不可欠であり、これが臨床現場での大きな障壁となっている[2]。

我々はこの課題を解決すべく、Web アプリケーション型の有害事象解析プラットフォーム「Adverse Events Signal-detection Tool (AEST)」を開発してきた[3]。AEST は、医薬品名を入力するだけで FAERS を統計解析し、結果をグラフ等で可視化するツールである（図 1）。

図 1 AEST の解析画面

これにより、IT や統計の専門知識がなくとも、誰もが短時間で世界最大級のデータベースから知見を得ることが可能となり、医療現場での応用が期待される。

しかし、AEST が提供する統計解析結果や可視化情報を正確に解釈し、臨床判断に繋げるには、統計解析結果や可視化された情報を整理して理解しやすく示すことが求められる。そのため、専門性が異なる多様な医療従事者が本ツールを有効に活用できるよう、解析結果の「解釈」を支援する機能の追加が課題として挙げられる。

近年、情報科学分野では ChatGPT に代表される大規模言語モデル（LLM）が登場し、自然言語処理に大きな変革をもたらした。LLM は汎用的な能力に加え、高度に専門化された領域でも顕著な性能を示している。例えば、米国専門医試験において、追加学習なしで高いスコアを達成したとの報告もある[4,5]。これは、AI が医学・薬学分野の臨床、教育、研究において強力な支援ツールとなり得ることを示唆している。以上の背景から、本研究では、医薬品有害事象情報

の臨床活用をさらに促進する AI システム開発を目的とする。具体的には、日本の薬剤師国家試験をベンチマークとして用い、ローカル環境で動作する LLM の薬学分野における基礎性能を評価するとともに、ファインチューニングによる薬学特化 LLM の構築と性能向上を試みた。

2. 手法

2.1 薬剤師国家試験

本研究では、2012 年（第 97 回）から 2024 年（第 109 回）までに実施された日本の薬剤師国家試験（Japanese National Pharmacist Examination; JNEP）の過去問題を用いた。JNEP は、厚生労働省の管轄下で年 1 回実施される国家資格試験であり、薬剤師として業務を遂行する上で必須となる基本的な知識と技能を評価することを目的とする。試験では、実践的な薬学的知識、関連法規、職業倫理など、臨床現場で不可欠とされる広範な能力が問われる。各年の試験は、物理、化学、生物、衛生、薬理、薬剤、病態・薬物治療、法規・制度・倫理、実務の 9 科目 345 問の多肢選択式問題で構成される。全ての問題は 1 つまたは 2 つの正解を持ち、解答は指定された選択肢と完全に一致した場合にのみ正解とみなされる。また、一部には連問形式が採用されており、解答には設問の文脈や先行する問題への解答を考慮する必要がある。合格基準は相対基準で変動するが、例えば第 109 回試験では、必須問題で 70%以上、各科目で 30%以上、全体で約 61%（210/345 問）以上の正答率を満たし、かつ医療倫理・安全上不適切な選択肢である「禁忌肢」の選択が 2 つ以下であることが求められた。

2.2 データセット

本研究で用いた LLM はテキスト入力のみに対応するため、化学構造式や図表など、視覚的情報の解釈を必要とする問題は評価対象から除外した。加えて、厚生労働省から正答の訂正などが公式に発表された問題も除外した。また、連問を構成する問題の一部にでも視覚情報が含まれる場合は、文脈依存性を考慮し、その連問全体をデータセットから除外することで、評価の妥当性を確保した。

2.3 LLMs

本研究では、4 つのローカル LLMs を使用した（表 1）。

表 1　研究で用いた LLMs の一覧

Short Forms for Models	Model Full Names	Parameters	Developer
phi-4	Phi-4	14 Billion	Microsoft Corporation
DSR1-Qwen32B	DeepSeek R1 Distill Qwen	32 Billion	High-Flyer
CA-DSR1-32B	DeepSeek R1 Distill Qwen fine -tuned for Japanese	32 Billion	High-Flyer and CyberAgent, Inc
CA-DSR1-14B	DeepSeek R1 Distill Qwen fine -tuned for Japanese	14 Billion	High-Flyer and CyberAgent, Inc

2.4 Low-Rank Adaptation (LoRA)

LoRA は、パラメータ効率の良いファインチューニング技術であり、大規模言語モデルの計算効率の良い適応を可能にする。LoRA の基本原理は、W_0と表される事前学習済みモデルの元の重みを固定し、適応中の重みの変化量（ΔW）を表現するために訓練可能な低ランク行列を注入することである。

元の重み行列を下記とした場合、

$$W_0 \in \mathbb{R}^{d\times k}$$

LoRA は更新量ΔWに対し低ランク分解を用いて下記のように近似する。

$$\Delta W = BA, \qquad where\ B \in \mathbb{R}^{d\times r}\ and\ A \in \mathbb{R}^{r\times k}$$

効率に関する重要な点は、ランクrが元の次元 d および k よりも大幅に小さいことである（すなわち、$r \ll \min(d,k)$ ）。行列 A と B のみが訓練され、順伝播は元の演算 $h = Wx$（xは入力、h は出力）を以下のように更新する。

$$h = W_0x + \Delta Wx = W_0x + BAx$$

この式により、モデルは低ランクの更新行列（ A および B ）のみを学習することで新しいタスクに適応でき、W 行列全体をファインチューニングする場合と比較して、訓練可能なパラメータの数を大幅に削減できる。広範な計算資源を要する従来のフルファインチューニングとは異なり、LoRA はコスト効率および資源効率に優れたアプローチを提供する [6]。

本研究では、前述した 4 つの LLM をベースモデルとし、それぞれに対して LoRA を用いたフ

ァインチューニングを実施した。その際、訓練データは訓練用と検証用に 9:1 の比率でランダムに分割した。

2.5　プロンプト

モデルへの入力には、共通のプロンプト構造を採用した。出力形式を統一するため、プロンプトの冒頭に指示文（instruction header）を配置した。

2.6　性能評価

LLM の性能は、モデルが生成した解答と公式解答を照合して評価した。評価プロセスは、過去の研究に倣い、自動評価と人手によるレビューの 2 段階で実施した。まず、正規表現を用いて、規定フォーマットに則った解答を自動で抽出し、正誤を判定した。次に、自動抽出が困難であった不規則な形式の出力や曖昧な解答については、薬剤師による目視確認を行い、最終的な判定を下した。採点基準は、正答ごとに 1 点、誤答、不完全な出力、および解答を特定できなかった場合は 0 点とした。主要評価指標には、テストデータセットに対する正答率（Accuracy）を用いた。

2.7　計算資源

本研究の全ての計算は、大阪大学サイバーメディアセンター（現：D3 センター）のスーパーコンピュータ「SQUID」上で実行した。評価時のモデル推論は、計算効率と数値安定性を両立させるため、半精度浮動小数点数（FP16）で行った。また、LoRA によるファインチューニングでは、モデル性能の低下を抑制しつつ GPU メモリ使用量を削減するため、8 ビット量子化を適用した。

3 結果と考察

3.1　全体の正答率

ベースラインモデルの正答率は 55.15%から 76.36%の範囲であり、DSR1-Qwen32B が 76.36%と最高値を示した（表 2）。LoRA によるファインチューニング後のモデルでは、正答率は 54.54%から 76.97%の範囲となり、LoRA 適用後の DSR1-Qwen32B が 76.97%で最良の成績を達成した。CA-DSR1-32B および DSR1-Qwen32B は、ベースライン時点で第 109 回 JNEP の合格基準（60.87%）を上回る成績を示した。ファインチューニングによる正答率の変化はモデルごとに異なり、変化なし（-0.61%）から最大+5.45%の大幅な向上まで観察された。特に 4 モデル中 3 モデルでファインチューニング後に性能向上が認められ、phi-4 は最も顕著な改善を示した。これらの結果は、ローカル LLM の開発およびファインチューニング手法の有効性を示すものである。

3.2　ドメイン別正答率

分野別の正答率において、生物学、衛生学、病態生理学・薬物治療学、臨床実務では、ベースライン時点で一般的な合格基準（約 60%）を上回る高い正答率が得られた（表 3）。特に生物学では、CA-DSR1-32B および DSR1-Qwen32B が満点を達成し、LoRA 適用後もその成績を維持した。また、衛生学は、全モデルでファインチューニング後の正答率が一貫して上昇した。薬理学、薬剤学、病態生理学・薬物治療学といった専門性の高い薬学領域では、LoRA ファインチューニングによって特定モデルで顕著な

表 2　全体の正答率の一覧

Model	Accuracy (%)		
	Base Model Accuracy, No./Total (%)	LoRA Fine-tuned Model Accuracy, No./Total (%)	Difference (Fine-tuned − Base), %
CA-DSR1-14B	91/165 (55.15)	90/165 (54.55)	-0.61
CA-DSR1-32B	106/165 (64.24)	112/165 (67.88)	3.64
DSR1-Qwen32B	126/165 (76.36)	127/165 (76.97)	0.61
phi-4	100/165 (60.61)	109/165 (66.06)	5.45

向上が見られた。phi-4 はこれら 3 領域全てで改善（+11.1%、+14.3%、+3.8%）を示し、CA-DSR1-32B は病態生理学・薬物治療学で+11.5%の大幅な増加を示した。これらの知見は、ファインチューニングが専門領域での性能向上に寄与すること、LoRA が薬学知識の獲得に有効であることを示している。

法規・制度・倫理では、日本語特化調整を行っていない DSR1-Qwen32B および phi-4 がベースラインで良好な成績（約 54%）を示し、LoRA 適用後にそれぞれ+16.7%、+12.5%と大幅な向上を達成した。これは、これら日本語非対応とされるモデルであっても、日本特有の法規・規制知識を内在的に保持している可能性を示唆する。

一方、化学ではモデルごとに LoRA ファインチューニングの効果が大きく異なった。CA-DSR1-14B および DSR1-Qwen32B はベースラインで比較的良好な成績（50%、100%）を示したが、ファインチューニング後は正答率が大幅に低下した（-50.0%）。一方、ベースライン正答率が最も低かった CA-DSR1-32B（25%）は、LoRA 適用後に+50.0%と大幅な改善を示した。ただし、サンプル数が少ない（n=4）ため、このファインチューニングの効果のばらつきに対しては今後更なる調査が必要であると考える。

4. おわりに

AEST 解析結果の「解釈」を支援する機能の追加という課題に対し、ローカル環境で動作する LLM の応用可能性を検討した。本研究では、薬学領域におけるローカル LLM の性能を JNEP で評価し、薬学的専門知識を付与するファインチューニング手法の有効性を探索した。これらの成果は、機密性の高い医療情報を扱う環境下で、ユーザーの知識レベルやニーズに応じた情報解釈支援システムの構築に向けた基礎的知見を提供するものであると期待する。

参考文献

1. Skelly, C. L., Cassagnol, M. & Munakomi, S. Adverse Events. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

2. Nango, D., Sekizuka, T., Goto, M. & Echizen, H. Analysis of Information on Drug Adverse Reactions Using U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). *Yakugaku Zasshi* **142**, 341–344 (2022).

3. 浅野弘斗 *et al.* 有害事象解析プラットフォーム AEST の開発. in (京都, 2023).

4. Schubert, M. C., Wick, W. & Venkataramani, V. Performance of Large Language Models on a Neurology Board–Style Examination. *JAMA Network Open* **6**, e2346721 (2023).

5. Longwell, J. B. *et al.* Performance of Large Language Models on Medical Oncology Examination Questions. *JAMA Network Open* **7**, e2417641 (2024).

6. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).

表 3　ドメイン別正答率の一覧

Domain	n	CA-DSR1-14B			CA-DSR1-32B			DSR1-Qwen32B			phi-4		
		Base (n)	Fine-tuned (n)	Δ(%)	Base (n)	Fine-tuned (n)	Δ(%)	Base (n)	Fine-tuned (n)	Δ(%)	Base (n)	Fine-tuned (n)	Δ(%)
Physics	11	7	6	-9.1	9	8	-9.1	8	9	9.1	8	7	-9.1
Chemistry	4	2	0	-50.0	1	3	50.0	4	2	-50.0	3	3	0.0
Biology	9	7	9	22.2	9	9	0.0	9	9	0.0	6	6	0.0
Hygiene	25	15	16	4.0	18	19	4.0	20	22	8.0	15	17	8.0
Pharmacology	27	10	8	-7.4	17	17	0.0	23	20	-11.1	11	14	11.1
Pharmaceutics	14	7	7	0.0	7	8	7.1	9	8	-7.1	7	9	14.3
Pathophysiology/Drug Therapy	26	16	15	-3.8	18	21	11.5	24	22	-7.7	19	20	3.8
Laws/Regulations/Ethics	24	12	13	4.2	12	12	0.0	13	17	16.7	14	17	12.5
Practice	25	15	16	4.0	15	15	0.0	16	18	8.0	17	16	-4.0