# 処方データにおける不確実な診療科名の LLM を用いた実体同定手法

川邊智也[1] 保坂桂佑[1] 清水俊樹[1] 横田直彦[1] 齋藤晋二[1]
中江郁青[1] 清水佳一郎[1] 沖﨑歩[1] 小熊彩香[1] 塚本友太郎[1]
[1]株式会社カケハシ KAKEHASHI Inc.
{t.kawabe, k.hosaka, t.shimizu, n.yokota, s.saito,
i.nakae, k.shimizu, a.okizaki, a.oguma, y.tsukamoto}@kakehashi.life

## 概要

薬剤師の対人業務強化に伴い，患者特性や行動の高度な分析が求められている．その上で「診療科情報」は重要な要素となるが，実務データは独自の運用に起因する表記揺れや情報の欠損が散見され，データ品質の低さが分析のボトルネックとなっている．本研究では，処方情報や患者属性のコンテキストを用いた LLM による診療科情報の実体同定手法を提案する．評価の結果，提案手法は診療科情報の欠損や誤りに対して高い頑健性を示し，特に処方情報が診療科特定における重要な判断因子であることが確認された．

## 1 はじめに

近年，薬局業務は従来の対物業務（調剤）から，患者の治療プロセス全体を継続的に管理・支援する対人業務へとその比重が移りつつある．こうした対人業務における適切な服薬指導やリスク管理には，患者の治療プロセスの分析が不可欠であり，その上で診療科情報は重要な要素となる．

しかしながら，実務データにおける診療科情報には，独自の運用に起因する表記揺れや情報の欠損が散見される．これらのデータを分析可能な状態にするためには，信頼性の低いテキスト（メンション）を実在する診療科（エンティティ）へと正規化，すなわち実体同定を行う必要があるが，この処理には一般的な医療実体同定とは性質の異なる課題がある [1, 2, 3]．第一の課題はメンションの曖昧性である．診療科の入力においては厳密な規定がないため，正確性にばらつきが生じやすい．結果として単なる表記揺れにとどまらず，メンションが診療実態と乖離する，あるいは特定に至らない記述となる場合があり，メンション自体が信頼できないという問題がある．第二の課題は，エンティティ定義の局所性である．同じ「内科」という診療科であっても，専門分化された総合病院とプライマリ・ケアを担うクリニックとではその診療範囲は異なる．そのため，医療機関によらない辞書定義に基づいた一律のマッピングでは，施設ごとの実態を正しく反映できない．

従来のアプローチでは，メンションとエンティティ間に一定の意味的類似性が存在し，かつエンティティの定義が大域的であることを前提としているため，これらの課題には対応できない．

そこで本研究では，信頼性の低いメンションのみに依存せず診療科と関連性の強い処方情報や患者属性をコンテキストとして LLM に入力し，本来の診療科を推論させる方法を提案する．また，施設ごとに異なる実在診療科リストから施設の特性を推論させることで，施設によって変動する診療科の役割の差異を考慮した同定を実現する．これらのアプローチにより，メンションの信頼性が低い場合や施設ごとの診療科の役割が異なる場合においても実態を反映した同定が可能であることを実データを用いた実験により示す．

## 2 関連研究

医療分野における実体同定は，同義語や同音異義語，略語といったドメイン特有の多様な表記や曖昧性が生じやすく，自然言語処理の中でも困難な課題である [1, 2, 3]．さらに，実際の臨床テキストは，誤入力や非文法的な記述といった不規則な表現を多く含むため，タスクの難易度を高めている [4]．

先行研究では，医療テキストの多様な表記や文脈の複雑さへの対応，続いて幻覚の抑制といった段階的な技術的進化が見られる．

初期の深層学習アプローチを発展させた判別モデルでは，同音異義語の識別に焦点が当てられている．Garda ら [5] は，従来のメンションとエンティティの類似度のみで判断するモデルの課題であった「字面は似ているが意味が異なる候補」の誤分類を防ぐため，メンション周辺のテキストを用いた再ラ

ンク付け手法を提案した．これにより，単なる表記の一致を超えた文脈に即した識別が可能となっている．

その後，知識と文脈をより深く統合するアプローチとして，QA 形式の活用が進展した．Lin ら [6] は，実体同定を「メンションと候補エンティティの定義文を照合する多肢選択問題」として定式化した．この手法は，単語の分散表現のみに頼るのではなく定義文という豊富なコンテキスト情報を直接的に推論に利用することで，曖昧なメンションに対しても高い精度を実現している．

さらに近年では，LLM を用いた生成的アプローチが主流となりつつある．Lin ら [7] は，LLM の生成能力を活かしつつ知識ベースに存在しない用語の生成（幻覚）を抑制するために，制約付きデコーディングを導入した．また，Kim ら [8] は，生成モデルに負例学習を取り入れ，酷似したエンティティ間の微細な差異の識別能力を強化している．

先行研究は，表記揺れや文脈の欠如といった課題に対して一定の解決策を示してきた．しかし，本研究が対象とする処方データに基づく診療科同定タスクは，情報の質と定義の安定性において，これらとは異なる 2 つの困難な性質を持つ．本研究では，これらの特性を持つタスクを対象として，実体同定を行う手法を提案し，その有効性を評価する．

# 3 提案手法

## 3.1 課題の特性

本研究が対象とする処方データに基づく診療科同定タスクは，先行研究と比較して，情報の質と定義の安定性において異なる 2 つの困難な性質を持つ．

第一の課題は，メンションの曖昧性である．診療科の入力には厳密な規定がないため，表記の正確さには個人差が生じる．そのため，単なる表記揺れにとどまらず，メンションが診療実態と乖離する，あるいは特定に至らない記述となる場合がある．すなわち，メンション自体が参照先として信頼できない状況下での同定が求められる点で，メンションとエンティティ間に一定の意味的類似性を仮定できる先行研究のタスク設定とは性質が異なる．

第二の課題は，エンティティ定義の局所性である．既存の手法は，知識ベース上のエンティティ定義が大域的であることを前提としている．しかし実運用において，その定義や範囲は施設の特性によって異なる．例えば，総合病院における「内科」とクリニックにおける「内科」では診療範囲が異なる．したがって，本タスクでは大域的な辞書定義に基づいた実体同定だけでは不十分であり，処方情報や患者属性に基づき，施設特性によって変動する診療科の役割を考慮した同定を行う必要がある．

## 3.2 提案手法

提案手法では，LLM に診療科名（メンション）を入力し，対象医療機関が保有する実在診療科リストから適切なものを推論させる．プロンプト構成について，診療科特定の根拠となる処方情報や患者属性をコンテキストとして入力した．また，Lin ら [7] の知見に基づき，Few-shot [9] を採用した．加えて，入力されたコンテキスト情報を十分に考慮して推論させるため，Chain-of-Thought (CoT) [10] を導入した．エンティティ定義の局所性に対応するため，CoT の推論ステップの一部に，保有診療科数から医療機関の特性を推定し，その規模に応じた各診療科の診療範囲を考慮する思考過程を組み込んだ．実際のプロンプトは付録に掲載した．

# 4 実験設定

## 4.1 データセット

本研究では，クラウド型電子薬歴サービスに蓄積された処方データから構築した匿名加工化 DB を利用する．2023 年 6 月 1 日からの 1 年間のデータから 987 件をランダムサンプリングして用いた．各データは，メンション，エンティティ，処方情報（医薬品一般名，剤型），患者属性（性別，年齢）で構成される．

本研究における正解データの作成は，専門知識を有する 5 名の薬剤師によって行われた [11, 12]．アノテーションに際しては，各サンプルに対して 2 名のアノテーターが割り当てられた．各アノテーターは独立してラベル付けを行い，2 名の判断に相違が生じた場合は，協議を通じて最終的なラベルを決定した．なお，薬剤の組み合わせから診療科が一意に定まらないサンプルについては，多答式でラベルを選択した．

構築したデータセットの各サンプルにおける「医療機関の保有診療科数」と「正解診療科数」の分布を図 1 に示す．図 1 に示す通り，本データセットはロングテールな分布を持っている．具体的には，保

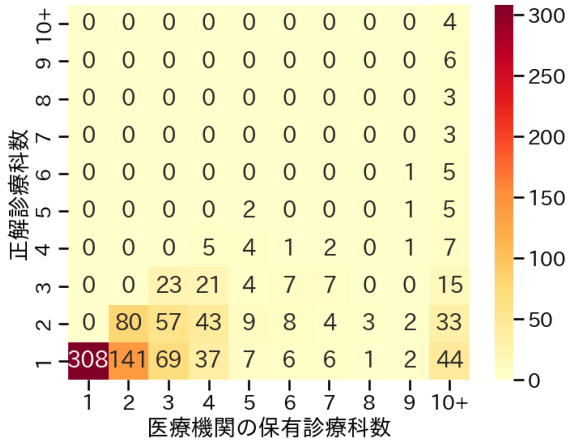


**図 1　データセットにおける**
**医療機関の保有診療科数と正解診療科数の分布**

有診療科数が 1〜4 の小規模機関が 784 件(79.4%)を占める一方で，10 以上の診療科を持つ大規模医療機関の事例も 125 件(12.7%)含まれている．また，保有診療科数が 1 である事例が 308 件(31.2%)を占めている．これらは正解が自明であり，全体の正解率を底上げする要因となり得る．しかし，本研究は実運用環境における性能評価を目的としているため，これら自明なサンプルの除外は行わず，実世界の処方データ分布を維持した．

### 4.2 実験内容

実運用環境における提案手法の性能評価を目的として実験を行った．推論モデルには GPT-5，GPT-5 mini，GPT-5 nano [13]，gpt-oss-20b [14] を採用した．ハイパーパラメータは，reasoning effort を除き利用サービスのデフォルト値 [15] を採用した．reasoning effort については，各モデルで利用可能な最も軽量な推論設定を選択することで条件を揃える方針とし，GPT-5，GPT-5 mini，GPT-5 nano では minimal，gpt-oss-20b では low に設定した．

検証においては，後述のベースラインとの精度比較に加え，コンテキストやプロンプト要素の影響を定量化するためのアブレーション分析を実施した．

### 4.3 ベースライン

ベースラインには，メンションとの Levenshtein 距離 [16] が最小となる診療科を割り当てる方法を用いた．これは，文字列の類似性に基づく手法では同定困難な事例に対し，提案手法の有効性を評価するためである．なお，タイは保有診療科リストの先頭に近いものを優先することで解消した．

### 4.4 評価指標

評価指標として正解率(Accuracy)を採用した．本研究では一つの入力に対して医学的に妥当な正解が複数存在し得る．そのため，モデルの予測結果が正解診療科リストのいずれか一つと一致する場合を正解と判定する [1, 2, 17].

## 5 実験結果

### 5.1 提案手法の評価

提案手法を用いた 4 つのモデルと，4.3 節で定義したベースラインの比較結果を表 1 に示す．正解率に加えて，標準偏差およびベースラインからの差分を併記した．正解率とベースラインからの差分は 3 回の試行における平均値である．

実験の結果，全てのモデルでベースラインを上回る性能が確認された．上位モデルの GPT-5 は 97.5%(+4.8 ポイント)を達成し，軽量モデルである GPT-5 mini や GPT-5 nano，オープンウェイトモデルである gpt-oss-20b においても 95%以上の精度を維持した．特に，GPT-5 nano のような計算コスト効率を重視したモデルであってもベースラインに対して+2.3 ポイントの改善を達成しており，提案手法が，モデルの規模や種類を問わず汎用的に機能することが実証された．

### 5.2 アブレーション分析

アブレーション分析の結果を表 2 に示す．表には一部の入力情報を除外した際の精度に加え，提案手法からの差分を併記した．

まず，患者情報の除外による影響は，全モデルにおいて-0.4〜-0.5 ポイントと軽微であり，本手法における寄与は限定的であった．

対照的に，処方情報の除外は全てのモデルにおいて-1.0〜-4.4 ポイントの精度低下を招いた．特に GPT-5 nano では-4.4 ポイントと低下幅が顕著であり，コンテキスト情報が曖昧性解消に不可欠であるとする Yang ら [18] の知見を支持する結果となった．一方で，GPT-5 や gpt-oss-20b などのモデルでは低下幅が-1.0〜-1.7 ポイントに留まっており，モデルによって外部情報の欠損に対する頑健性が異なることが示された．

表 1　提案手法とベースラインの精度比較

| | 正解率 | 標準偏差 | ベースラインとの差分 |
|---|---|---|---|
| GPT-5 | 97.5 | 0.1 | 4.8 |
| GPT-5 mini | 97.3 | 0.3 | 4.6 |
| GPT-5 nano | 95.9 | 0.8 | 3.2 |
| gpt-oss-20b | 95.0 | 0.2 | 2.3 |
| ベースライン | 92.7 | - | - |

Few-shot に関しては, Zhu ら [19] が指摘するタスク理解への重要性に反し，これを除外した方が GPT-5 以外のモデルで+0.2〜+0.4 ポイントの精度向上が見られた.

### 5.3 考察

メンションが欠損または「診療科なし」の場合，提案手法はベースライン手法に対して正解率が GPT-5 では 11.8 ポイント，gpt-oss-20b では 7.7 ポイント高くなり，優れた頑健性を示した．例えば，提案手法は処方薬がシロップ剤や坐剤であること，患者が乳幼児であることを根拠に，候補リストから小児科を正しく特定できた.

また，メンションと処方内容が乖離したケースにおいても，提案手法では処方内容を踏まえた適切な判断が行われていた．例えば，前立腺肥大症治療薬が処方されているにも関わらずメンションが外科である事例において，提案手法はメンションの信頼性が低いと判断し，医学的妥当性の高い泌尿器科を正しく予測した．これは，提案手法がメンションと事実情報の間に矛盾が生じた際，より信頼度の高い事実情報を優先して推論できることを示唆している.

加えて，提案手法は施設ごとの異なる診療科の役割（局所性）を適切に解釈できた．例えば，内科と循環器科のみを持つ医療機関で抗炎症軟膏が処方された事例では，処方情報がない場合，モデルは循環器科と誤って推論した．対して提案手法は，処方薬が皮膚疾患用である事実と，皮膚科が候補にないという制約を組み合わせることで，皮膚系の処置を内科が担当していると解釈し，内科へ正しく同定できた．これは，提案手法が施設の保有診療科リストと処方実態の整合性を考慮し，施設固有の定義を動的に反映できていることを示している.

また，アブレーション分析の結果，処方情報は診療科特定における重要な判断因子であることが確認された．このコンテキストの影響はモデルサイズによって異なり, 軽量モデルほど影響度が大きかった．具体的には，軽量モデルでは処方情報を除外すると「高齢者であればリハビリテーション科」のように限られた情報から短絡的に推論する傾向や，入力にない薬剤情報を根拠として捏造する幻覚が見られた．対して上位モデルは，処方情報がない場合でも「リハビリテーション科では処方することはほとんどない」のようなモデル内部の知識を用いて多面的に判断することで誤った推論を回避できており，処方情報の除外に伴う精度低下は小さかった.

表 2　アブレーション分析の結果

| 条件 | GPT-5 | GPT-5 mini | GPT-5 nano | gpt-oss-20b |
|---|---|---|---|---|
| 提案手法 | 97.5 | 97.3 | 95.9 | 95.0 |
| 患者情報なし | 97.1 | 96.9 | 95.6 | 95.0 |
| | (-0.4) | (-0.4) | (-0.3) | (-0.0) |
| 処方情報なし | 95.8 | 95.5 | 91.6 | 94.4 |
| | (-1.6) | (-1.7) | (-4.3) | (-0.5) |
| Few-shot なし | 97.4 | 97.4 | 95.7 | 95.7 |
| | (-0.1) | (0.1) | (0.2) | (0.8) |

一方で, 提案手法の限界も明らかになった. まず，軽量モデルにおいて候補リストにない診療科を出力する形式的なエラーが生じた．3 回の試行で合計して GPT-5 nano で 5 件(0.2%), gpt-oss-20b で 1 件(0.03%)確認された．また，人間でも判断が難しい境界領域の事例において，入力するコンテキスト情報やモデル種別によらず正解できないケースが見られた．そのうちの多くの推論過程で正解の診療科の可能性への言及があったが，誤って判断されていた．こうしたケースに対する，専門家の知見や判断基準を元にした判定精度向上が今後の課題である.

## 6 おわりに

本研究では，メンションの曖昧性とエンティティ定義の局所性に対し，処方情報や患者属性のコンテキストを活用した LLM による診療科の実体同定手法を提案した．実験結果より，提案手法はメンションの欠損や誤りに対して高い頑健性を示し，特に処方情報が診療科特定における重要な判断因子であることが確認された．一方で，軽量モデルで候補外の診療科を出力するエラーや判断が困難な境界領域における誤分類といった課題も確認された.

## 参考文献


[1] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical Entity Representations with Synonym Marginalization. In Proc. 58th ACL, pp. 3641–3650, 2020.

[2] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-Alignment Pretraining for Biomedical Entity Representations. In Proc. NAACL, pp. 4228–4238, 2021.

[3] Hongyi Yuan, Zheng Yuan, and Sheng Yu. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In Proc. NAACL, pp. 4038–4048, 2022.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In Proc. 2nd Clinical NLP Workshop, pp. 72–78, 2019.

[5] Samuele Garda and Ulf Leser. BELHD: Improving Biomedical Entity Linking with Homonym Disambiguation. Bioinformatics, 40(8):btae474, 2024.

[6] Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. Biomedical Entity Linking as Multiple Choice Question Answering. In Proc. LREC-COLING, pp. 2390–2396, 2024.

[7] Zhenxi Lin, Ziheng Zhang, Jian Wu, Yefeng Zheng, and Xian Wu. Guiding Large Language Models for Biomedical Entity Linking via Restrictive and Contrastive Decoding. In Findings of EMNLP, pp. 23745–23759, 2025.

[8] Chanhwi Kim, Hyunjae Kim, Sihyeon Park, Jiwoo Lee, Mujeen Sung, and Jaewoo Kang. Learning from Negative Samples in Biomedical Generative Entity Linking. In Findings of ACL, pp. 10714–10730, 2025.

[9] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. In Proc. NeurIPS, Vol. 33, pp. 1877–1901, 2020.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proc. NeurIPS, Vol. 35, pp. 24824–24837, 2022.

[11] Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C. Wallace, and Ani Nenkova. Predicting Annotation Difficulty to Improve Task Routing and Model Performance for Biomedical Information Extraction. In Proc. NAACL, pp. 1471–1480, 2019.

[12] Rachel M. Murphy, Dave A. Dongelmans, Nicolette F. de Keizer, et al. Creation of a Gold Standard Dutch Corpus of Clinical Notes for Adverse Drug Event Detection: The Dutch ADE Corpus. Lang. Resources & Eval., 59(3):2763–2779, 2025.

[13] OpenAI. GPT-5 System Card. 2025. https://cdn.openai.com/gpt-5-system-card.pdf, (accessed Dec. 26, 2025).

[14] OpenAI. "gpt-oss-120b & gpt-oss-20b Model Card". arXiv preprint arXiv:2508.10925. 2025.

[15] Databricks. Foundation Model APIs Reference. Databricks Documentation. https://docs.databricks.com/aws/en/machine-learning/foundation-model-apis/api-reference, (accessed Dec. 26, 2025).

[16] Vladimir I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Sov. Phys. Dokl., 10(8):707–710, 1966.

[17] Samuele Garda, Leon Weber-Genzel, Robert Martin, and Ulf Leser. BELB: a Biomedical Entity Linking Benchmark. Bioinformatics, 39(11):btad698, 2023.

[18] Siyu Yang, Peiliang Zhang, Chao Che, and Zhaoqian Zhong. B-LBConA: a Medical Entity Disambiguation Model based on Bio-LinkBERT and Context-Aware Mechanism. BMC Bioinformatics, 24:97, 2023.

[19] Yutao Zhu, Peitian Zhang, Chenghao Zhang, et al. INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning. In Proc. ACL, pp. 2782–2809, 2024.

## A 実験用プロンプト

あなたは医療データのアナリストです.
以下の情報に基づき，処方元の診療科（`original_department`）を，その医療機関が持つ診療科の完全なリスト（`available_departments`）の中で最も適切と考えられる診療科にマッピングしてください.

`original_department` は，文字列（例: "膠原病，リウマチ内科"）として提供されます.

{追加するコンテキスト情報についての説明}
`original_department` が "診療科無し" や "dm 外来" のような特殊な値であっても，提供された情報と `available_departments` をヒントに，最も可能性の高い診療科を推測してください.

# 入力情報

{input_info_section}

# 出力形式

以下の JSON 形式で，マッピング結果（`mapping`）と，そのように判断した根拠（`reasoning`）を日本語で出力してください.

# 出力要件
1. **`available_departments` の中から最も意味的・文字列的に近しいと判断されるものを「必ず」一つ選び，マッピングしてください．**
2. **`available_departments` に完全一致するものがなくても，提供された情報や診療科の専門領域から推論してマッピングしてください．（例：「膠原病」は「リウマチ科」）**
3. **`available_departments` の数から病院の規模を推定して，病院の規模を考慮した上で各診療科の意味を考えてください．**
4. **`reasoning` には，なぜその `available_departments` の要素を選んだのか，特に推論を働かせた場合はその理由を「必ず」具体的に記述してください．**
5. **「マッピング不可」という値は絶対に使用しないでください．**

{{{{
  "reasoning": "（マッピングの根拠をここに記述）",
  "mapping": "マッピング結果"
}}}}

{few_shot_section}