大規模言語モデルによる
日本語SOAP臨床ノートの階層型マルチラベル構造化
Hierarchical Multi-Label Structuring of Japanese SOAP Clinical Notes with Large Language Models

張 孜恒
北海道大学大学院医学院　画像診断学教室

Introduction &Objective

- SOAP（Subjective / Objective / Assessment / Plan）形式の診療録には臨床的に重要な情報が多い[1]が、多くが自由記載で二次利用が難しい[2]
- 日本語臨床文は分かち書きがなく、略語・表記揺れ・英語医療用語の混在が多いため、構造化がさらに困難
- 大規模言語モデル（LLM）は文脈理解に強い一方、出力形式の崩れやスキーマ外ラベル（幻覚）を生みやすい
- 本研究の目的： schema-guided LLM枠組みを開発し、自由記載の日本語SOAPノートを入力として、ノート内の主要臨床情報を網羅する階層構造化出力（JSON）を生成し、その性能を評価する

Methods

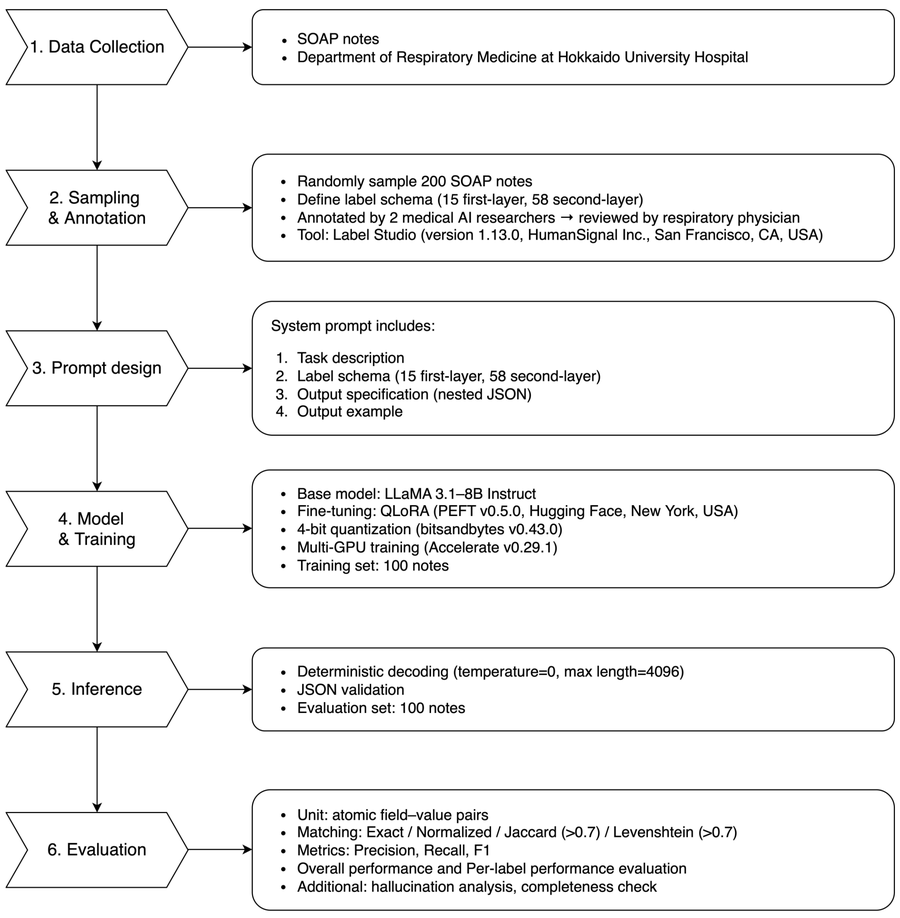

Figure1: Overview of the study workflow

1. データ
200例の呼吸器内科SOAPデータ
（医師による確認済み、ゴールドスタンダードデータとして使用）
- トレーニング用データ160例
- 評価用データ100例

2. アノテーション方式
- 15の第一層ラベル（カテゴリ）（例：「生命兆候 Vital Signs」「身体検査 Physical Examination」）
- 58の第二層ラベル（詳細ラベル）（例：「血圧 Blood Pressure (BP)」「脈拍 Pulse (HR)」）
- 階層的アノテーション：
  Step 1: 第一層ラベル（カテゴリ）を適用
  Step 2: 第一層の範囲内で第二層ラベルを適用

3. モデル・推論
- Base: LLaMA 3.1-8B Instruct（院内ローカル実行）
- QLoRAで微調整（4-bit量子化 + LoRA）
- スキーマを列挙したSystem Promptで、二層JSONを厳密に生成
- 推論は温度0（決定論）・最大長4096 tokens

4. 定量評価
評価単位： Atomic cell （[label, value]ペア）
- Gold standard：医師レビュー済み手動アノテーションデータを真値として使用
- モデル出力をAtomic cellに分解し、 Gold standard と照合し、micro平均のPrecision / Recall / F1 を算出。
- 4つの一致基準：

Strict	Normalized	Jaccard > 0.7	Levenshtein > 0.7
完全一致	大小/記号差を無視	内容の重なり(overlap)	編集距離類似(完全一致に直すための修正コスト)

Future & Conclusion

- モデル・データ拡充（稀少ラベル・ 症例数・診療科・施設の拡大 ）
- スキーマ制約強化・ポストプロセス導入
- Macro平均や臨床的有用性の指標で再評価
- 臨床的有用性（人手修正コスト、下流性能）で外部評価

本研究では、呼吸器内科における日本語SOAPノートから臨床情報を構造化するための、階層的マルチラベルに基づくLLMフレームワークを開発した。二層スキーマとスキーマ指向プロンプトを統合することで、本手法は、非英語の臨床テキストを対象とした構造化における実現と実用可能性を示した。

Results

主な結果（micro-averaged）
- 完全出力：76/100（token上限により24件は途中終了）
- Fine-tuned（Raw）：F1 = 0.49（Strict）→ 0.59（Levenshtein）
- Whitelist後：F1 = 0.54（Strict）→ 0.65（Levenshtein）
- Baseline（未微調整）：F1 = 0.01（Strict）/ 0.04（Levenshtein）

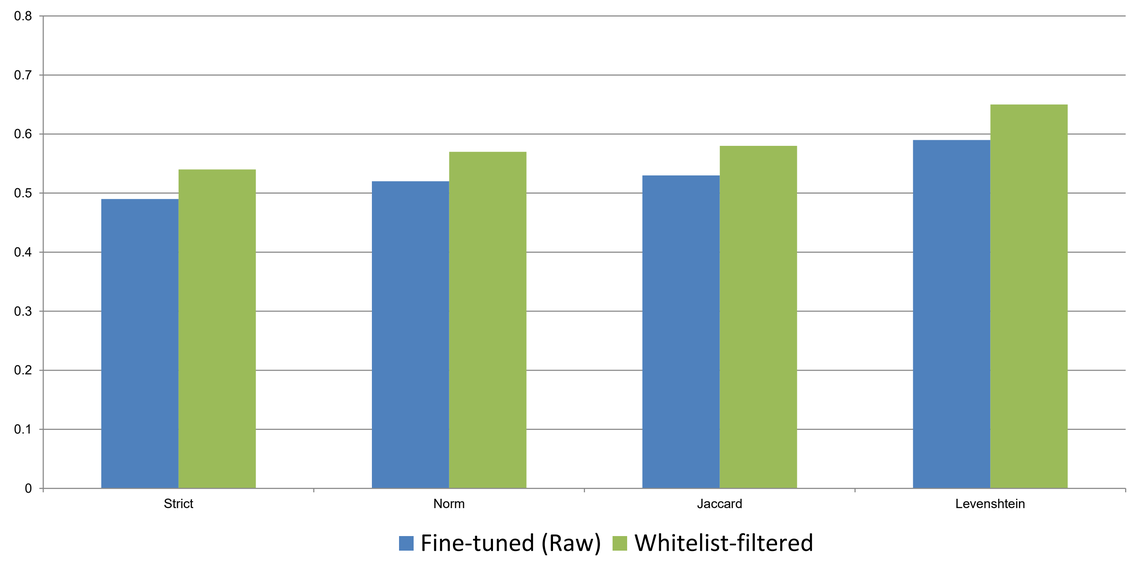

Figure 2: Fine-tunedモデル全体性能（ Raw: end to end VS Whitelist-filtered後）

スキーマ外ラベル分析
- 合計1,780個の予測ラベルのうち、スキーマ外ラベルは302個

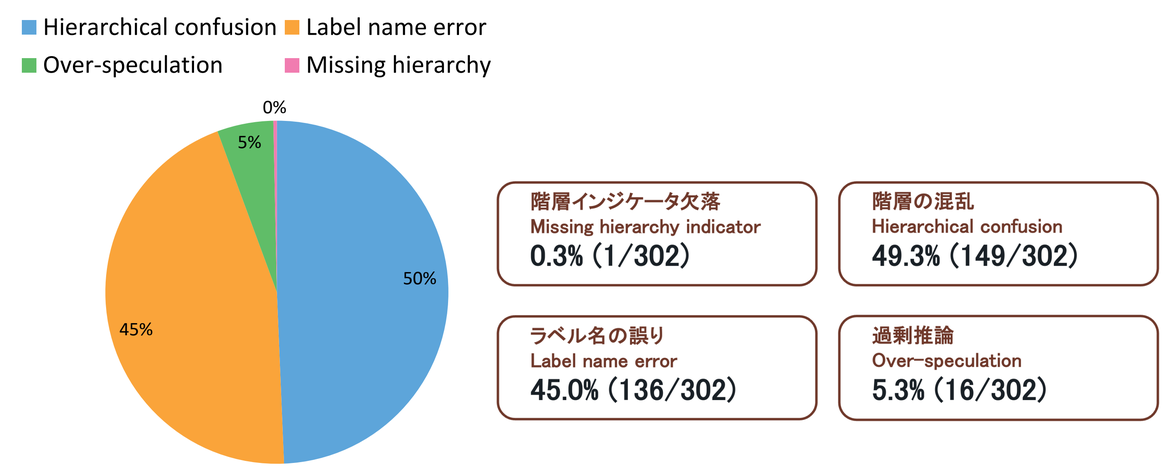

Figure 3:　Off-schema Label (手動で4種類に分類)

ラベル別の傾向
- 定型項目（バイタル・検査）は高精度（例：血圧/SpO_2など はF1>0.8）
- 曖昧境界（主訴 vs 現病歴）や希少ラベルは低下

Discussion

- 階層スキーマにより、親子ラベルの整合性を保ちやすく、レビューが容易
  （結果：親子不整合・スキーマ外ラベルの発生が確認され、階層制約が有効）
- 頻出で定型的な項目（バイタル・検査）は高精度に抽出可能
  （結果：ラベル別性能でバイタル／検査系が0.8以上の高成績）
- JSON形式チェック＋ホワイトリストフィルタによりLLM幻覚を抑制
  （結果：フィルタ後に全体性能が改善／
  例：Strict F1 0.49→0.54、Levenshtein>0.7 F1 0.59→0.65）
- ローカル実行可能なLLM で、医療データのプライバシー要件に適合しやすい
  （運用面の利点：院内運用・データ持ち出し回避）

[1] Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes . I.B. Wilson et al. 1995
[2] LungDiag: Empowering Artificial Intelligence for Respiratory Disease Diagnosis Through Electronic Health Records. Hengrui Liang. Et al. 2023

＊ 開示すべきCOI: 利益相反なし。