

RLMによる症例報告テキストの自動反復アノテーションと品質制御の試み

Automated iterative annotation and quality control of case-report texts: an RLM-style approach



○土肥栄祐¹、金進東²、早川格³、松原知康⁴、高月照江²、建石由佳⁵、藤原豊史²、山本泰智²

¹ 国立精神・神経医療研究センター（NCNP）神経研究所 疾病研究第三部 ² 情報・システム研究機構 国立遺伝学研究所 ライフサイエンス統合データベース部門（DBCLS）
³ 国立成育医療研究センター（NCCHD） ⁴ 徳島大学 ⁵ 科学技術振興機構 NBDC事業推進室（NBDC） ✉ edohi@ncnp.go.jp

Take-home message

「LLMで抽出 → 基準で検証 → 不足を追加抽出」を自動で反復すると、**単回では届かない網羅性**に到達する。
ローカルの小規模（MoE・活性3B）モデルでも **Event Coverage 54% → 96%**、hallucination（tier C）0件。

1 背景・目的 BACKGROUND

- 難病・希少疾患は約1万種・人口の約5%。診断まで平均5～7年。非典型例は症例報告に散在しており、構造化された症例コーパスが手掛かりになる。
- 昨年までに「症例報告PDF → テキスト抽出 → 構造化 → 症状・所見抽出 → カテゴリー化・辞書作り → アノテーション」のパイプラインを構築（日本語症例報告 853症例）。
- しかし LLM の単回処理では、長文での取りこぼしやアノテーション密度の不足が生じる。
- 目的：検証して足りなければ追加で注釈する反復処理（RLM的アプローチ）で、品質を

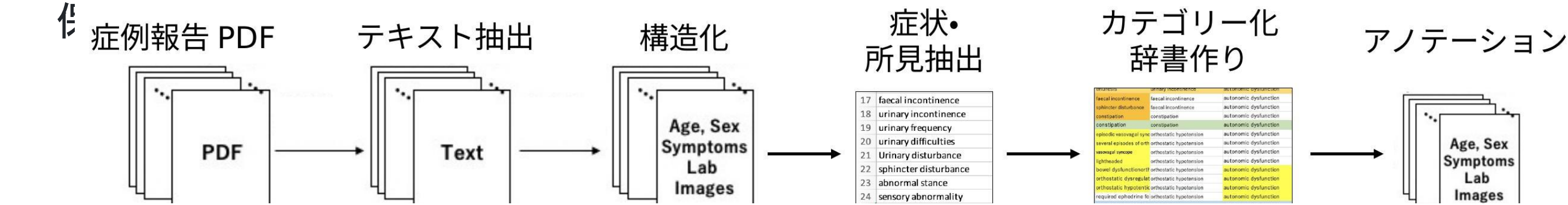


図1 昨年構築した症例コーパス作成パイプライン。本発表は「アノテーション」工程の品質づくりが中核。

2 方法：RLM的な反復ワークフロー METHOD

- LLMによる抽出（プロンプトエンジニアリング済）と定義した基準による検証を組み合わせ、基準を満たすまで自動反復する Recursive Language Models（RLM）的ワークフロー。
- モデル：Qwen3.6-35B-A3B（MoE・活性約3B）を Ollama でローカル実行。

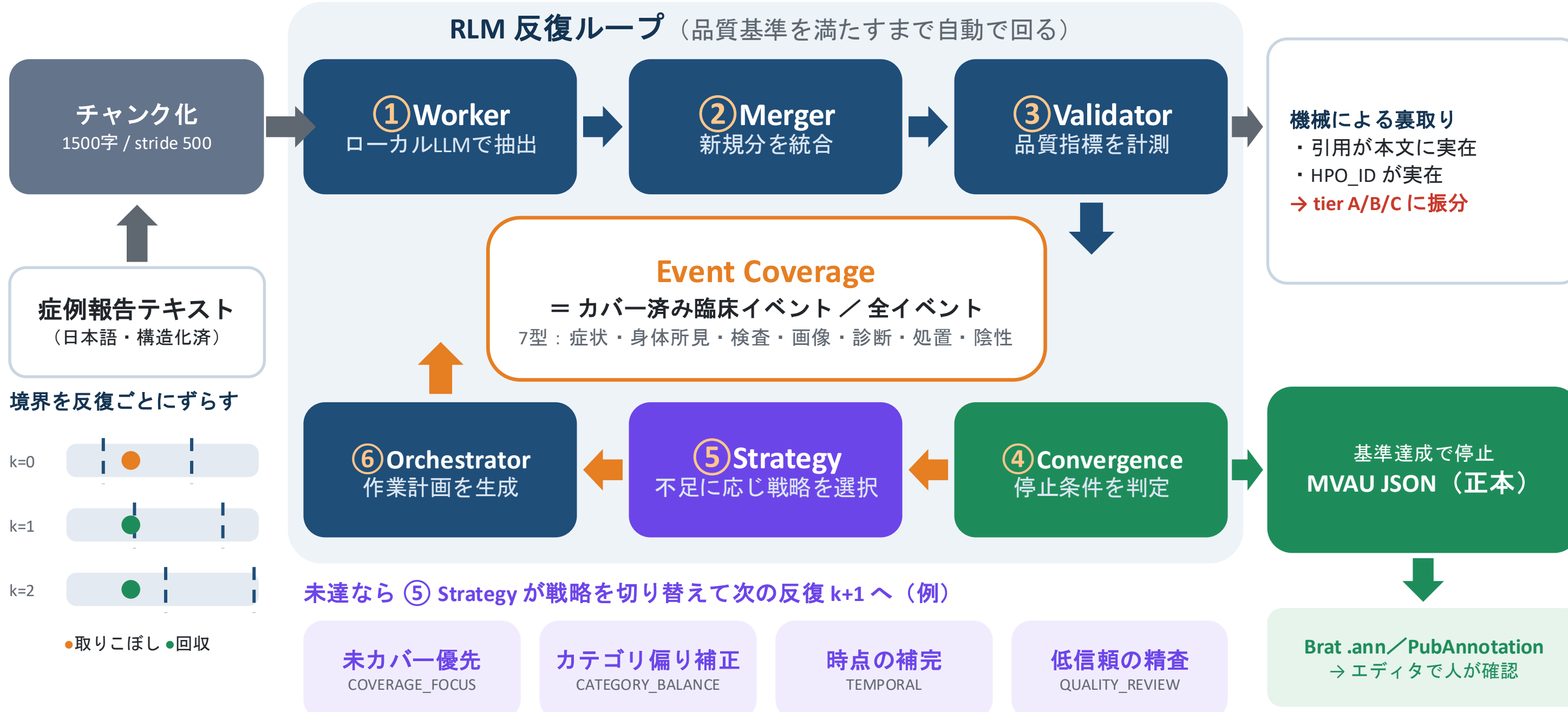


図2 RLM的な反復アノテーション・ワークフロー（6コンポーネント）。Event Coverage を指標に戦略を切り替え、チャンク境界も反復ごとにずらす。

- チャンク境界をずらす：offset = (k-stride) mod chunk。境界（seam）上で取りこぼした所見を次の反復で回収。句点の内側は切らない。
- 停止条件：基準到達・新規注釈の停滞・最大反復回数。回し過ぎると注釈内の注釈が生じるため停止指標に組み込み。

3 結果：反復で網羅性が向上 RESULTS

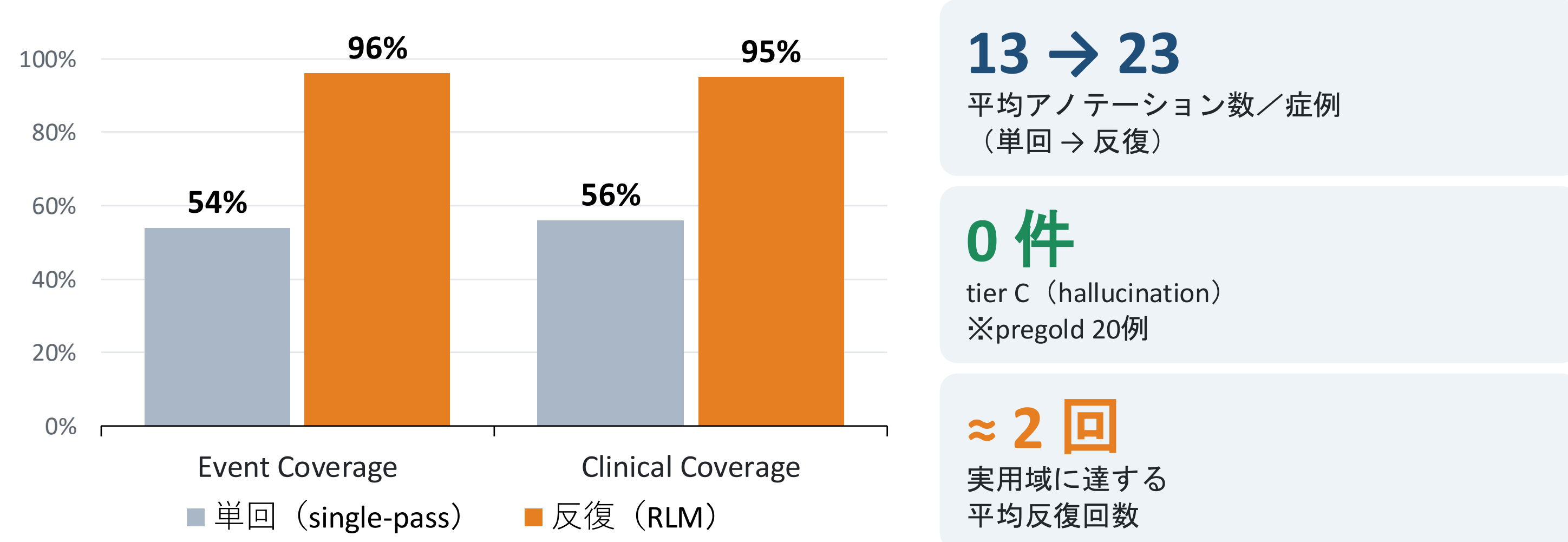
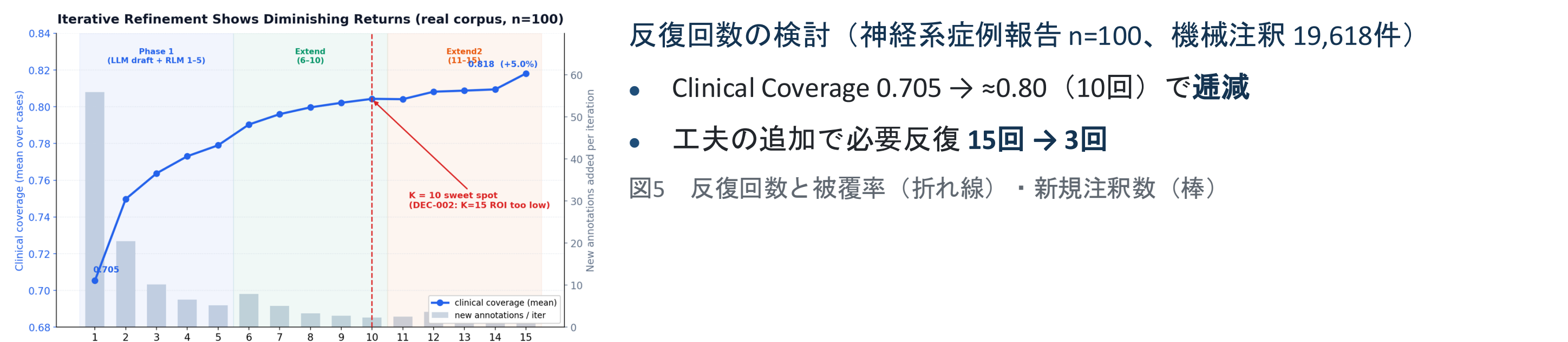


図4 カバー率：単回 vs 反復（予備検討・少数例）。単回で取りこぼした身体所見・処置名が反復過程で補完された。



反復回数の検討（神経系症例報告 n=100、機械注釈 19,618件）

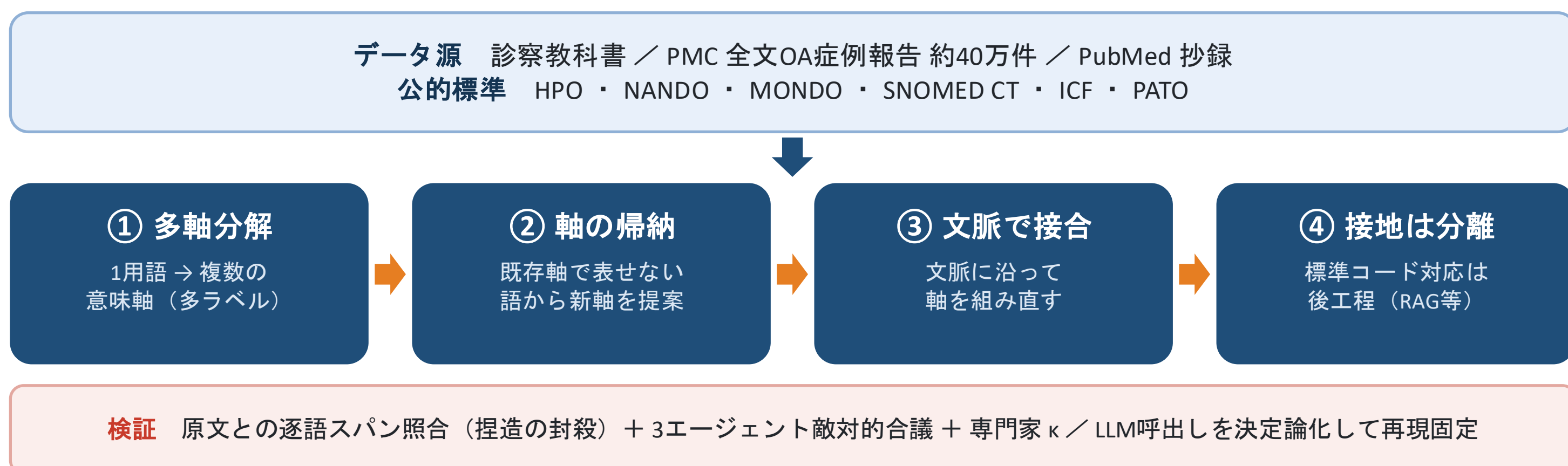
- Clinical Coverage 0.705 → ≈0.80（10回）で逡減
- 工夫の追加で必要反復 15回 → 3回

図5 反復回数と被覆率（折れ線）・新規注釈数（棒）

4 用語へのカテゴリ付与：多軸分解 PRELIMINARY

- 抽出した臨床用語を1つのコード・カテゴリーに潰さず、直交する意味軸へ多ラベル分解し、足りない軸はデータから帰納する。
- 3つの LLM エージェントの敵対的合議で検証。標準ID（HPO 等）への接地は分解と分離して後工程へ（症例報告の HPO RAG 後付けと同じ設計）。

予備検討：日本語の神経内科症例報告から抽出した臨床用語 約9,100語



例：1つの所見を直交する意味軸に分ける（1つのコードに潰さない）



図6 臨床用語の多軸分解パイプラインと分解例



約9,100用語（日本語神経内科症例報告）の検証結果。単一カテゴリーへの正規化では被覆困難 → 接地は分解と分離すべき

別コーパスでも「補完パス」で被覆が向上（予備検討：教科書由来の所見抽出、N=30バッチ）

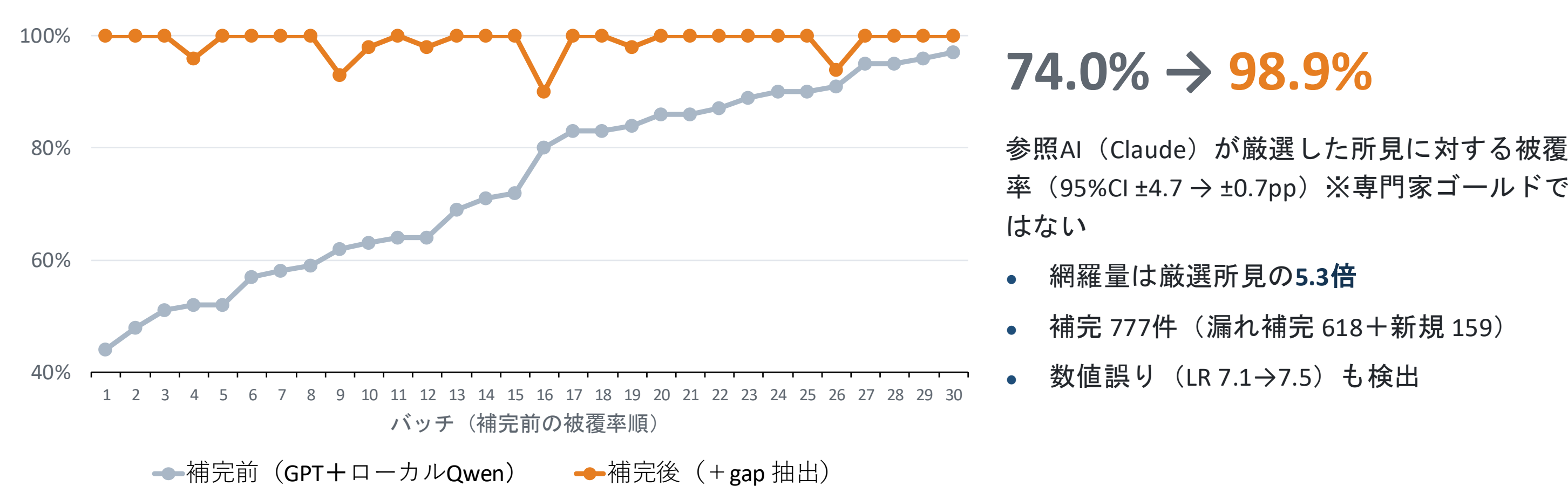


図7 バッチごとの被覆率。安価なパイプライン+不足分の追加抽出で、ばらつきも縮小（σ 16.6 → 2.4pp）



図8 展開の基盤と所見頻度分布（文単位の粗い集計のため言い換えて過大に出る）。長い裾 → 頻出語の辞書化だけでは届かない

5 考察・今後 DISCUSSION

- 反復処理により単回では困難だった性能向上を確認。小規模モデルでも反復で性能を補完できる可能性。
- 「検証 → 不足補完」の RLM 的発想は他工程にも適用可能（別コーパスでも補完パスで 74.0% → 98.9%）。
- 用語は多軸・文脈依存：多軸分解+接地の分離+合議で、人の確認を約3%に集中できる。
- ローカル環境で完結 → 医療データ処理の省力化・安全性に寄与。今後は40万件規模へ展開。

まとめ

- ① 検証→不足補完の自動反復で、網羅性 54%→96%
- ② 小規模ローカルLLMでも反復で性能を補完
- ③ 用語は多軸分解+合議で、人の確認を約3%に集中