医療RAGの実用化を阻むナレッジ品質の壁

― 知識構造化による検索チャンク品質の改善可能性 ―

Measurement of retrieved chunk quality from real-world knowledge in retrieval-augmented generation: A Phase 1 foundational study

深瀧 恭子 [1,2)] , 林　和華子 [2,3)] , 北山 恵 [4)] , 平田　健司[5)], 工藤 與亮[5)], 伊藤 陽一 [6)]

1)札幌医科大学 医療統計・データ管理学講座　2)北海道大学大学院医学院　社会医学講座ヘルスデータサイエンス教室　3）東北大学病院臨床研究推進センター　4)和歌山県立医科大学 データサイエンスセンター　5)北海道大学大学院 医学研究院 画像診断学教室　6)北海道大学病院　医療・ヘルスサイエンス研究開発機構データサイエンスセンター

RAGの精度は downstream 最適化だけでは説明できない

～ Plug-and-play型 RAGの限界と評価フレームの提案～

既存研究は retrieval や generation など downstream 改善を中心に進められてきた。
しかし Plug-and-play型RAGでは、入力ナレッジやチャンク構造といった upstream 品質の評価枠組みは十分に整理されていない。

課題　Plug-and-play 型RAGでは、検索品質の低下がどの上流要因に由来するかを切り分けにくい。

空白　ナレッジ品質およびチャンク品質を体系的に評価する枠組みは十分に整備されていない。

目的　Knowledge / Chunk / Content の3層評価により、実世界ナレッジの品質と構造化が、RAGで取得されるチャンク品質に与える影響を定量的に検証する。

※ RAG：Retrieval-Augmented Generation：検索拡張生成

検証アプローチと3階層評価フレームワーク

比較する2つのナレッジライン（入力手法）

A-line / B-line を共通プラットフォームで処理し、3層で取得品質を比較評価した。

A-line　ベースライン

既存PDFをそのまま入力

4施設 × 研究計画書作成マニュアル
対象セクション：3
Objective / Design / Risks_benefits

B-line　構造化

LLM向けに事前構造化

既存マニュアルの曖昧さや暗黙知を補完
構造化（YAML形式）したナレッジ
（PoC：Objectiveセクション限定）

データ設計と主要設定

登録ナレッジ

A-line：site-01 ｜ site-02 ｜ seite-03 ｜ site-04　　※施設名が特定できないように、匿名化して登録
B-line：site-05

Google Cloud Vertex AI Search

Layout-based chunking ｜ chunkSize = 500 tokens ｜ includeAncestorHeadings = true ｜ SearchResultMode = CHUNKS

クエリ設計

3セクション × 3クエリタイプ（Write/Include/Examples）= 9クエリ × QE on/off × 4施設

3階層評価フレームワーク

Level 1　A　**評価段階：** 検索前（入力ナレッジ）

知識品質評価

評価対象　A-line 入力文書（4施設）
手法　専門家3名による独立評価

3軸（構造・粒度・ノイズ）× 13項目 × 40基準
OK/NG判定 → 信頼性指標：Fleiss κ

Level 2a　A　B　**評価段階：** 検索後（取得チャンク）

チャンク構造品質評価

評価対象　A-line (n=594) + B-line (n=6)　計600チャンク
手法　LLM-as-a-Judge（GPT-4o, temperature=0, seed=42）

5指標 × 5試行 = 2,970評価
Section Targeting / Coverage / Purity / Granularity / Faithfulness

Level 2b　A　**評価段階：** 検索後（内容整合性）

コンテンツ適切性評価（PoC）

評価対象　A-line のみ（n=99チャンク、Include × QE=on）
手法　ゴールドスタンダード（GS）比較

GS：SPIRIT 2025 Checklist 準拠で独自作成（専門家3名による合意形成）
Coverage_gs / Purity_gs / Contradiction_gs

結果｜主な5つの結果

Result 1　専門家評価の一致度は低かった

3名専門家による独立評価（160評価単位）

0.269

Fleiss' κ

Fair（κ<0.40）
専門家でも判断は一致しない

Result 2　LLM-as-a-Judge の再現性は高かった

5指標 × 5試行（2,970評価）

0.936

ICC(2,1)

Excellent
All metrics ≥ 0.75

Result 3　A-lineで取得されたチャンク品質は低水準だった

ナレッジの構造品質（structure）

実世界ナレッジの構造問題が取得チャンク品質に与える影響を評価

＜ 20%

score ≥ 4 到達率
Median ≤ 2.0

5指標すべてで構造品質が低かった

Result 4　A-lineで取得されたは回答に必要な情報を十分に含んでいなかった

ナレッジの内容適合性（content）

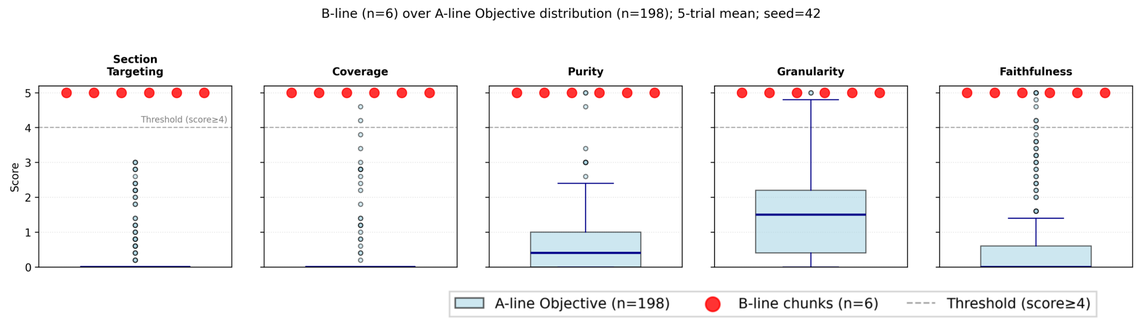
回答に必要な情報がチャンクに含まれているかを評価

2%

Coverage ≥ 4 到達率
Median = 0

回答に必要な情報がほとんど含まれていなかった

Result 5　RAGの活用目的に合わせて構造化・内容強化されたB-lineチャンクはA-line Objective分布の上方に一貫して位置した（PoC）



RAG精度向上のもう一つのピース：ナレッジ品質の「定量的診断」と「事前構造化」への投資

本研究の意義

Garbage In, Garbage Out は広く共有された前提だが、ナレッジの不足・ノイズ・構造不全を診断し、改善し、継続管理する運用システムは十分に確立されていない。本研究は、この未整備領域をナレッジ品質・チャンク品質・コンテンツ適切性の3層で定量化し、改善可能な課題として示した。また、専門家評価の一致度の限界（κ=0.269）に対し、LLMジャッジが高い再現性（ICC=0.936）を示したことは、実運用への普及が途上にある自動品質評価の実装可能性を支持する知見である。

今後の展望

RAGの活用目的に合わせて構造化・内容強化されたナレッジがチャンク品質を改善し得ることはPoCとして示された。Phase 2では、この知見を起点に、LLMジャッジを活用した継続的品質管理、活用目的別の評価フレームの拡張、言語化困難な知識の扱いを含むナレッジ運用システムの構築へと展開する。

開示すべきCOIについて：本発表にあたり開示すべきCOIはありません。