日本語 RAG 検索における省略由来の曖昧クエリ検知

佐々木峻 [1] 山本大輝 [1]
[1]Acroquest Technology 株式会社
{sasaki,yamamoto}@acroquest.co.jp

概要

RAG（Retrieval-Augmented Generation）では，生成以前に関連文書を正しく検索できることが精度を左右する．しかし実運用では，省略や指示語により複数解釈が成立する曖昧な質問が入力されやすく，特に日本語では格要素の省略が頻繁に生じる．本稿では「文書コーパス内に回答根拠が存在する」前提のもと，質問形式クエリに対する曖昧性検知を，(i) 曖昧化による検索性能低下の定量化と，(ii) 検索出力（スコア分布・検索結果の意味的分布）に基づく検知可能性の検証を通して議論する．JDocQA の一部 106 問を用いた実験で，省略による曖昧化は Recall@1 を 53.8%から 34.0%へ低下させた．さらに，上位検索結果のスコア分散および埋め込みの一貫性指標（平均類似度と分散を組み合わせた Clarity 指標）が，曖昧化の有無で系統的な差を示すことを確認した．

1 はじめに

大規模言語モデル（LLM）を検索器と組み合わせる RAG は，外部知識に基づく回答生成を可能にし，出典提示や更新容易性の観点からも有用である [1]．一方で，RAG の精度は「入力クエリが意図する情報を検索で取り出せるか」に強く依存する．実運用の対話では，ユーザが前提を省略したまま質問することが多く，その結果，複数の合理的解釈が成立する曖昧クエリが入力されうる．

曖昧さには複数の型があり，例えば MIRAGE は曖昧さを Syntactic/Semantic/General に分類し，指示語や省略を含む構文的曖昧さなどを整理している [2]．本研究では特に，**日本語の省略**に起因して検索時に情報が不足し，検索結果が分散しやすい状況に着目する．Ishizuki らは BCCWJ から省略可否データセットを構築し，名詞句の省略率が主格 46.2%，対格 22.7%，与格 27.6%に達することを報告している [3]．この性質は，ユーザが「誰が／何を／どこで／いつ」といった要素を省略した質問を作りやすいことを示唆する．

RAG の生成器に曖昧な検索結果を与えると，根拠の混在による回答の不安定化やハルシネーション誘発につながる可能性がある．したがって，**検索時点でクエリの曖昧さを検知**し，不足情報の補完や確認質問（clarification）へ誘導できれば，RAG 全体の精度向上が期待できる．

本稿の貢献は以下である．

1. 日本語 QA データに対し，省略による曖昧化が密ベクトル検索の Recall を大きく低下させることを定量的に示す．
2. LLM 単体判定と比較し，検索出力に基づく簡易な統計量（スコア分散）および検索結果埋め込みの一貫性指標が曖昧性検知に有効であることを示す．
3. 「曖昧」対「コーパス内に答えが存在しない（unanswerable / out-of-domain）」の識別が難しい点を明確化し，今後必要な指標を議論する．

2 関連研究

2.1 曖昧質問の分類と検知

Park らは，多段推論と曖昧性解釈を同時に評価するベンチマーク MIRAGE を提案し，曖昧さを構文的・意味的・一般的曖昧さへ分類した [2]．また，曖昧質問に対して推論を段階化する枠組み（CLARION）も議論している．

曖昧検知では，「質問自体が曖昧」な不確実性と，「モデルが知識不足等で迷う」不確実性が混同されがちである．Shi らは LLM に複数回答を生成させ，回答分布の特徴量から分類器を学習することで，質問曖昧性とモデル出力不確実性の区別を試みた [4]．本研究でも，LLM による曖昧判定をベースラインとして扱う．

2.2 検索クエリ品質推定（QPP）

検索前後の情報からクエリ性能を推定する Query Performance Prediction（QPP）は，本研究と関連が深い．Cronen-Townsend らは，クエリ言語モデルとコレクション言語モデルの相対エントロピーに基づく Clarity Score を提案した [5]．また Roitman は，上位文書スコアの分散に基づく Normalized Query Commitment（NQC）を再検討し，QPP としての有効性を示している [6]．本研究の「検索スコア分散による曖昧性検知」は，NQC に近い直観に基づく．

2.3 曖昧さの解消

曖昧質問に対してユーザへ確認を促す研究もある．Ma らは AmbigChat を提案し，曖昧な質問をファセット分解して対話 UI で選択させることで，目的回答へ到達しやすくする枠組みを示した [7]．本研究は解消（disambiguation）そのものではなく，**解消へ接続するための検知**に焦点を当てる．

3 実験設定

3.1 タスク定義

本研究は，RAG における**検索タスク（retrieval）**を対象とする．入力はユーザの質問形式クエリ q であり，検索対象コーパス $\mathcal{D}$ 内に，少なくとも 1 つの正解文書（チャンク）d^* が存在すると仮定する．検索器は q に対して上位 k 件のランキング Top-$k(q)$ を返す．評価指標は Recall@k（$d^* \in$ Top-$k(q)$ の割合）とする．

3.2 データセット

文書 QA データセットとして JDocQA[8] を用いる．JDocQA は PDF 文書と日本語 QA からなり，ページ参照や根拠領域（バウンディングボックス）を含む．本実験では小規模に検証するため，(i) test/validation split，(ii) 自由記述（生成）回答，(iii) 画像・図表を必須としない質問に絞り，106 問を抽出した．また，各質問に対し，対応する根拠ページテキストを 1 チャンクとして用い，検索対象コーパスも同数の 106 チャンクで構成した（各クエリに正解チャンクが 1 つ存在する設定）．

表 1 明確クエリ（orig）と曖昧化クエリ（trans.）の検索性能（Recall@k）

	R@1	R@3	R@5	R@10
orig	53.77	75.47	81.13	83.02
trans.	33.96	49.06	57.55	64.15

3.3 検索モデル

検索には密ベクトル埋め込みを用いる．クエリ q と文書 d を埋め込みモデル embed(·) でベクトル化し，コサイン類似度で順位付けする．

$$\mathrm{sim}(q, d) = \frac{\mathbf{e}_q \cdot \mathbf{e}_d}{\|\mathbf{e}_q\| \, \|\mathbf{e}_d\|} \tag{1}$$

埋め込みには OpenAI の `text-embedding-3-small` を用いた．同モデルは既定で 1536 次元ベクトルを出力する [9]．

3.4 曖昧クエリの生成

明確な質問を「省略（ellipsis）により複数解釈が成立する自然な質問」へ変換し，曖昧版クエリ q' を作成した．変換は OpenAI の gpt-5-mini モデルにより行い，主語・目的語・修飾語のいずれかを省略する制約を課した（付録のプロンプト参照）．

4 実験 1：省略による検索性能低下

4.1 設定

明確クエリ q と曖昧化クエリ q' の双方で検索を行い，Recall@k の差を測定した．

4.2 結果

表 1 に検索性能を示す．省略による曖昧化により，全ての k で Recall が低下した．特に Recall@1 は 53.77%から 33.96%へ約 19.8 ポイント低下した．

4.3 考察

省略により情報が欠落すると，複数の話題・対象が候補となり，検索結果が意図から外れやすい．RAG ではこの段階で正解根拠が上位に来ない限り，生成器が十分な根拠にアクセスできず，回答品質が下がる．したがって，検索の前後で曖昧さを検知し，**不足する格要素等をユーザに確認する**設計が重要になる．

表 2　LLM による曖昧性判定（106 件）

	曖昧 (%)	明確 (%)	信頼度
orig	68.9	31.1	0.889
trans.	92.5	7.5	0.903

表 3　Top-k 検索スコア標準偏差（106 件平均）

k	SD(orig)	SD(trans.)	比
3	0.02412	0.01476	1.63
5	0.02471	0.01526	1.62
10	0.02419	0.01562	1.55

5　実験 2：検索出力に基づく曖昧性検知

本節では，曖昧性を (1) LLM 判定，(2) 検索スコア分散，(3) 検索結果埋め込みの一貫性，の 3 観点で検証する．

5.1　アプローチ 1：LLM による二値判定

MIRAGE の観点を参考に，省略・指示語・時間場所不足・略語などを手掛かりとして曖昧性を判定し，JSON で出力するプロンプトを用いた．

結果を表 2 に示す．明確クエリでも 68.9%が曖昧と判定されており，過検知が大きい．このことは，単発の二値プロンプトだけでは，曖昧性概念の主観性や，文脈依存性を吸収しづらいことを示唆する（先行研究でも，単純プロンプトはランダム近傍に留まることが報告されている [4]）．

5.2　アプローチ 2：検索スコア分散

クエリが明確な場合，特定文書が相対的に高いスコアを得て，上位スコア分布の分散が大きくなると期待される（QPP の NQC の直観 [6]）．そこで，Top-k の検索スコアの標準偏差 SD_k を計算した．

表 3 より，明確クエリの方がスコア分散が大きい傾向を確認した．省略により複数解釈が生じると，多様な文書が「そこそこ似ている」状態となり，上位スコアが均されると考えられる．

5.3　アプローチ 3：検索結果埋め込みの意味的一貫性

スコア分散だけでは，曖昧クエリと「そもそも答えがコーパスに存在しない」クエリを区別しにくい可能性がある．そこで，上位検索結果文書集合 $D = \{d_1, \ldots, d_n\}$ の埋め込み間の意味的一貫性を測る指標を導入する．文書埋め込みを $\mathbf{e}_i = \mathrm{embed}(d_i)$ とし，文書間のコサイン類似度を

$$\mathrm{sim}(d_i, d_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \, \|\mathbf{e}_j\|} \tag{2}$$

で定義する．

表 4　検索結果の意味的一貫性（106 件平均）

	Top-3	Top-5	Top-10
MPS (orig)	0.7259	0.6998	0.6642
MPS (trans.)	0.6732	0.6580	0.6295
Clarity (orig)	0.6757	0.6276	0.5814
Clarity (trans.)	0.6101	0.5737	0.5311
Centroid (orig)	0.9034	0.8708	0.8340
Centroid (trans.)	0.8832	0.8509	0.8148

表 5　orig ／ trans. ／ wrong の比較（Top-10，106 件平均）

指標	orig	trans.	wrong
MPS	0.6642	0.6295	0.5897
Clarity	0.5814	0.5311	0.4682
σ	0.0828	0.0984	0.1215

代表的指標として平均ペアワイズ類似度（MPS）と，その標準偏差 σ を用いる：

$$\mathrm{MPS} = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{sim}(d_i, d_j), \tag{3}$$

$$\sigma = \sqrt{\frac{2}{n(n-1)} \sum_{i<j} \big(\mathrm{sim}(d_i, d_j) - \mathrm{MPS}\big)^2}. \tag{4}$$

さらに，平均と分散を線形に統合した

$$\mathrm{Clarity} = \mathrm{MPS} - \sigma \tag{5}$$

を用いる．これは「平均類似度が高く，ばらつきが小さい」検索結果に高スコアを与える．

表 4 に結果を示す．いずれの Top-k でも，明確クエリの方が MPS/Clarity が高い（＝検索結果が一貫）傾向を示した．一方で，セントロイド類似度（各文書と重心の類似度）では差が小さく，識別力が限定的であった．

5.4　「明確だが不正解」クエリとの比較

曖昧クエリ検知では，**曖昧**と**コーパス外（答えがない）**を誤って混同すると，ユーザインタラクション設計が破綻する．そこで，元質問と文法構造を保ちながらトピックだけを無関係に置換した「wrong」クエリを生成し，同様に指標を計算した．表 5 に主要指標を示す．

6　議論

6.1　検知の実用上の位置づけ

LLM 単体の二値判定は過検知が大きく，運用上は「常に曖昧」と扱う状態に近くなる．これに対し，検索スコア分散や検索結果の意味的一貫性は，**検索器が実際に返した結果**に基づくため，RAG パイプラインと整合的な信号を提供する．例えば，Clarity が

閾値未満の場合に「主語／目的語／時間／場所」の確認質問を提示するなど，UI 設計へ直結しやすい．

6.2 曖昧 vs. 答えなしの識別

表 5 では，wrong クエリは transformed よりも一貫性がさらに低い傾向を示したが，両者は同方向に変化するため，**単一指標だけでの二者判別は難しい**．この問題を解くには，例えば以下の追加信号が必要になる可能性がある：

- キーワード一致や BM25 などの疎検索特徴量との併用（語彙レベルの外れを検知）
- 検索結果のクラスタ数推定（多峰性）や，トピック分布のエントロピー
- 「回答可能性（answerability）」を陽に学習した分類器（例：JDocQA 自体が unanswerable を含む [8]）

7 おわりに

本稿では，日本語 RAG 検索における省略由来の曖昧クエリに着目し，曖昧化が密ベクトル検索の Recall を大きく損なうことを示した．さらに，検索スコア分散および検索結果埋め込みの意味的一貫性（特に Clarity 指標）が，曖昧性検知の有効な手掛かりとなることを確認した．

今後の課題は主に二点である．第一に，本稿は平均値の傾向を示した段階であり，実運用の判定には閾値設定や学習器（スコアを入力とする分類モデル）が必要である．第二に，曖昧性は検知するだけでなく，**どの情報が不足しているか**を推定し，ユーザへ具体的な補完ガイド（確認質問や選択肢）を提示することが重要である．曖昧な検索では関連チャンク自体が上位に現れにくいため，疎検索・辞書・対話的分解など他側面の手法との統合が求められる．

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020. Accepted at NeurIPS 2020.

[2] Jeonghyun Park, Ingeol Baek, Seunghyun Yoon, Haeun Jang, Aparna Garimella, Akriti Jain, Nedim Lipka, and Hwanhee Lee. Mirage: Multi-hop reasoning with ambiguity evaluation for illusory questions. *arXiv preprint arXiv:2509.22750*, 2025.

[3] Yukiko Ishizuki, Tatsuki Kuribayashi, Yuichiroh Matsubayashi, Ryohei Sasano, and Kentaro Inui. To drop or not to drop? predicting argument ellipsis judgments: A case study in japanese. In *Proceedings of LREC-COLING 2024*, pp. 16198–16210, 2024.

[4] Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. Ambiguity detection and uncertainty calibration for question answering with large language models. In *Proceedings of the Trustworthy Natural Language Processing (TrustNLP) Workshop*, 2025.

[5] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*, pp. 299–306, 2002.

[6] Haggai Roitman. Normalized query commitment revisited. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*, 2019.

[7] Jiaju Ma, Lei Shi, Kenneth Robertsen, and Peggy Chi. Ambigchat: Interactive hierarchical clarification for ambiguous open-domain question answering. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST 2025)*, 2025.

[8] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. Jdocqa: Japanese document question answering dataset for generative language models. In *Proceedings of LREC-COLING 2024*, 2024. arXiv:2403.19454.

[9] OpenAI. Vector embeddings — openai api. `https://platform.openai.com/docs/guides/embeddings`, 2024. Accessed: 2026-01-03.

付録 (Appendix)

再現性のため，実験で用いたプロンプト本文をそのまま示す．プレースホルダ {question} は入力質問で置換した．

(A) 省略（ellipsis）による曖昧化

与えられた「明確で曖昧さのない質問」を、意図的に「曖昧だが自然な質問」に変換してください。

重要な制約：

- 文法的に不自然な質問は作らない
- 完全に情報不足で答えられない質問にはしない
- 複数の合理的な解釈が存在するようにする
- 曖昧さは指定されたタイプに限定する

以下の明確な質問を、単語の省略（ellipsis）による曖昧さを持つ質問に変換してください。

要件：

- 主語・目的語・修飾語のいずれかを省略する
- 省略しても日本語として自然であること
- 文脈なしでは複数の解釈が可能になること

元の質問: {question}

変換後の質問のみを出力してください（説明や前置きは不要です）。

(B) LLM による曖昧判定

以下の質問が「曖昧な質問」かどうかを判定してください。

曖昧な質問とは、以下のいずれかの特徴を持つ質問です：

1. 主語、目的語、または修飾語が省略されており、文脈なしでは複数の解釈が可能

2. 指示語（これ、それ、あれなど）が何を指すか不明確

3. 時間や場所などの情報が不足しており、複数の状況に当てはまる可能性がある

4. 専門用語や略語が説明なく使われており、解釈が分かれる可能性がある

質問: {question}

以下の形式で JSON 形式で回答してください：

{

”is_ambiguous”: true または false,

”confidence”: 0.0 から 1.0 の間の数値（判定の確信度）,

”reason”: ”判定理由を簡潔に説明”

}

JSON 形式のみで回答してください。

(C) 無関係（wrong）クエリ生成

与えられた質問を文法的にはもとの文章と同じだが、既存質問群とは全く関係ないトピックのものに変更してください。

元の質問:{question}

変換後の質問のみを出力してください（説明や前置きは不要です）。

既存質問群

{ ここにデータセット中の質問を記載 }