LBRAG：選択的翻訳による多言語 RAG システム

近畿大学工学部ロボティクス学科
学生番号: 2211000083
氏名: 閻　昊男

YYYY 年 MM 月 DD 日

目次

第 1 章

はじめに

1.1 研究背景

近年, ChatGPT や Claude に代表される生成型人工知能は, 自然言語処理 (Natural Language Processing, NLP) の発展を大きく牽引している．これらのシステムは Transformer に基づく大規模言語モデル（Large Language Models, LLM）を中核とし，大規模コーパスに対する自己教師あり学習を通じて，多様な言語タスクにまたがる汎用的な能力を獲得している．

一方で，LLM は学習時点までの情報に依存する静的なパラメトリックモデルであり，次のような限界が指摘されている．第一に，学習後に更新された知識や最新事象を自動的に反映できず，情報が陳腐化しうる点である．第二に，モデル内部の知識は分散表現として暗黙的に符号化されているため，回答の根拠となる具体的文書や出典を明示しにくい点である．第三に，入力が曖昧な場合や学習分布から外れた問いに対して，事実に基づかない出力（ハルシネーション）が生じる可能性がある点である．

これらの課題に対する代表的アプローチとして，外部知識を検索してから生成を行う検索拡張生成 (Retrieval-Augmented Generation, RAG) が提案されている．RAG は，百科事典，ニュース記事，技術文書，企業内ナレッジベースなどの非パラメトリックな知識源から関連文書を検索し，その内容を根拠として LLM に回答を生成させることで，(i) 知識更新容易性，(ii) 根拠に基づく説明可能性，(iii) ハルシネーション抑制，を同時に狙う枠組みである．

しかし，既存の RAG 研究の多くは英語単言語環境を前提として設計・評価されている．現実の情報環境は多言語的であり，質問文の言語と有力な証拠文書の言語が一

致しない状況，すなわちクロスリンガル質問応答（cross-lingual QA）が一般に発生する．このとき，以下のような課題が顕在化する．

1. **検索段階の言語バイアス**：多言語ベクトル検索モデルであっても，高資源言語（特に英語）やクエリ言語と同一の言語を選好する傾向があり，他言語に存在する重要情報が十分に検索されない場合がある．
2. **翻訳に伴う情報損失とコスト**：質問言語と異なる証拠文書を一括してピボット言語へ翻訳する方式は扱いやすい一方で，翻訳誤りやニュアンスの損失を招きうる．また，翻訳量の増加によりトークンコストや遅延が増大する．
3. **出力言語制御の困難さ**：クロスリンガル環境では，モデルが英語や複数言語を混在させた回答（code-switching）を生成しやすく，ユーザが期待する言語で一貫した回答が得られない場合がある．
4. **証拠帰属の不透明さ**：多言語証拠と翻訳結果が混在する状況では，どの主張がどの原文に基づくかが不明瞭になり，検証可能性や監査性が低下しうる．

これらの問題は単なる精度低下にとどまらず，多言語話者間の情報アクセス格差を拡大しうるという観点からも重要である．特に，低資源言語話者が最新かつ信頼できる情報に到達する際，高資源言語への過度な依存や誤訳に起因する不利益が生じる可能性がある．

■具体例（多言語 RAG が想定する状況）　例えば，日本語で「妊娠中に服用可能な薬か」を質問した場合でも，信頼できる一次情報が英語の公的機関サイトや英文添付文書にしか存在しないことがある．また，技術分野ではエラーコードや仕様変更が英語のリリースノートやフォーラムで先行して共有され，他言語に情報が十分に流通していない状況が生じる．さらに，留学生の学内手続きのように，質問したい言語と規程文書の言語が一致しないケースも多い．このように「根拠は多言語に散在し，回答はユーザ言語で一貫させたい」という要請は現実的であり，クロスリンガル環境で実運用可能な RAG の設計が求められる．

先行研究は，多言語 RAG の課題に対して，(i) 多言語ベンチマークの整備，(ii) 英語などへの統一翻訳による単一言語化，(iii) 内部知識と外部証拠の統合，(iv) 多段階生成による推論強化，など多様な方向から取り組んできた．しかし，「高カバレッジな多言語検索」，「必要最小限で信頼性を担保した翻訳」，「ユーザ言語での出力一貫性」，「主張と原文証拠の対応付け」という要件を，計算コストを抑えながら同時に満たす統合的枠組みは十分に確立されていない．

本研究はこのギャップに着目し，多言語環境で実運用可能な RAG システム LBRAG（Language-Bridged Retrieval-Augmented Generation）を提案する．LBRAG は，翻訳を言語間の「橋渡し」として利用しつつも，全文一括翻訳に依存せず，重要な証拠に限定した選択的翻訳と文単位アライメントにより，正確性・言語一貫性・コスト効率・追跡可能性のバランスを図ることを目指す．

1.2 研究目的

本研究の目的は，ユーザ言語と証拠言語が一致しない多言語・クロスリンガル環境において，以下の要求を同時に満たす検索拡張生成システム LBRAG を設計・評価することである．

1. **正確性と実装可能性の両立**：多言語埋め込み（本実装では OpenAI `text-embedding-3-small`）による検索と，LLM による listwise 再ランキングを組み合わせ，限られた実装規模でも有効に機能する二段階検索を設計する．
2. **選択的翻訳によるコスト効率化**：全文書の統一翻訳ではなく，関連度・事前推定信頼度・トークン長（近似）に基づき重要段落のみを翻訳対象として選択する「選択的言語ブリッジ」を導入し，翻訳コストと情報損失を抑制しながら，多言語証拠を扱いやすい形に統合する．
3. **言語一貫性を保った回答生成**：翻訳後の文単位アライメント（本実装では順序に基づく貪欲 1 対 1）と，数値・日付・表記ゆれに敏感な要素（スロット）の簡易整合性検証により，証拠の品質信号を付与した上で，ユーザ言語 L_q で一貫した短回答を生成する．

1.3 本研究の貢献

本研究の主な貢献は，以下のように要約される．

1. **埋め込み検索＋ LLM 再ランキングの統合**：多言語埋め込みによる高リコール検索と，LLM の listwise 再ランキングを組み合わせた二段階スコアリングにより，多言語候補集合を構築する実装を示す（本実験では同言語除外によりクロスリンガル状況を作る）．
2. **選択的言語ブリッジの提案**：翻訳予算の制約下で，関連度・事前推定信頼度・ト

ークンコストに基づく効率比により翻訳対象段落を貪欲選択する枠組みを導入する．さらに，翻訳後に文単位アライメントと簡易スロット整合性チェックを行うことで，翻訳誤差の影響を抑えつつ証拠品質の信号（被覆率・スロット整合率）を得る．

3. **性能・言語一貫性・コストを統合評価する指標設計**：従来用いられてきた EM や F1 に加え，回答が目標言語でどれだけ一貫しているかを測る言語一貫性指標（Response Language Consistency, RLC）や，翻訳トークン数に対する性能向上効率を表す Cost-Normalized Bridging Efficiency（CNBE）を導入し，多言語 RAG に固有のトレードオフを定量的に評価可能とする．

1.4 論文構成

本論文の構成は次のとおりである．

第 2 章では，多言語 RAG およびクロスリンガル QA に関連する既存研究を整理し，本研究の位置づけを明確化する．

第 3 章では，提案手法 LBRAG の構成を述べる．多言語混合検索，選択的言語ブリッジ，文単位アライメントおよびスロット整合性検証，生成プロンプト設計について定義する．

第 4 章では，MKQA を用いた実証分析の実験設定を示す．比較手法，実装条件，評価指標（EM，F1，RLC，翻訳コスト，CNBE，および LLM による意味一致スコア）と評価プロトコルを説明し，本研究で採用する簡略化条件とその意図を明記する．

第 5 章では，実験結果を報告し，正確性・言語一貫性・翻訳コスト効率の観点から考察を行う．

第 6 章では，本研究のまとめと今後の課題を述べる．

第 2 章

関連研究

本章では，まず検索拡張生成（Retrieval-Augmented Generation, RAG）の基礎的枠組みを概観し, 次に多言語 RAG およびクロスリンガル質問応答（cross-lingual QA）に関する代表的研究を整理する．最後に，既存研究の限界を整理し，本研究 LBRAG の位置づけを明確化する．

2.1 RAG の基礎

RAG は，静的なパラメトリック知識のみに依存する大規模言語モデル（LLM）の限界を補うために提案された枠組みであり,「外部検索＋条件付き生成」を中核に持つ．代表的な手法として，Lewis ら[1] は，大規模な外部コーパスから関連文書を検索し，そのテキストをエンコーダを通じて取得した上で生成モデルに入力することで，知識集約型タスクにおける性能向上と根拠提示を同時に実現できることを示した．

一般的な RAG システムは，(i) クエリベクトルと文書ベクトルに基づく密検索（dense retrieval）あるいは BM25 等によるスパース検索，(ii) 検索結果に対する再ランキング，(iii) 上位 K 件の証拠とクエリを連結したプロンプトに基づく生成，という段階で構成される．この構造により，外部知識源の更新可能性と，回答と証拠との対応付けが可能となる一方，システム全体の性能は検索器・再ランキング器・生成モデルの相互作用に強く依存する．

その後, 事前学習段階から検索モジュールを組み込む REALM, 密ベクトル検索に特化した DPR, 複数文書を効果的に利用する FiD, 大規模外部メモリを用いた RETRO などが提案され，RAG パラダイムは英語圏におけるオープンドメイン QA の標準的

アプローチとして位置づけられている．しかし，これらの多くは英語単言語コーパスを前提としており，多言語環境における挙動や公平性については限定的な検討にとどまっている．

2.2 多言語 RAG とクロスリンガル QA の展開

現実世界では，ユーザ質問と利用可能な情報資源が必ずしも同一言語で与えられるとは限らない．この状況を明示的に扱うタスクとして，クロスリンガル QA がある．Asai らの XOR-QA[6] は，質問言語と証拠言語が異なる設定を含むデータセットを構築し，多言語検索およびクロスリンガル推論の困難さを定量化した．この系譜に属する XOR-TyDi などのベンチマークは，低資源言語や非ラテン文字言語に対して，既存モデルが高資源言語の証拠に偏ることや，言語間推論に失敗しやすいことを示している．

多言語 RAG の包括的な検証として，Chirkova らの mRAG[2] は 13 言語にわたる多言語 RAG ベースラインを構築し，検索・生成の両段階における言語バイアスを体系的に分析した．彼らの結果によれば，多言語モデルをそのまま用いた場合，(1) 高資源言語（特に英語）の証拠が過度に選択される，(2) 出力言語が制御しにくく，英語や混合言語で回答する傾向が強い，(3) 非ラテン文字言語では code-switching がより頻繁に生じる，といった問題が顕著に現れることが報告されている．

これらの研究は，多言語環境における RAG の性能低下が，単に翻訳品質の問題にとどまらず，検索バイアス，出力言語制御の困難さ，多言語証拠統合の脆弱性といった構造的問題に起因することを示している．

2.3 多言語 RAG の代表的アプローチ

多言語 RAG に対する近年の改良手法は，大きく以下の方向性に分類できる．

2.3.1 統一翻訳（Pivot Translation）に基づく手法

Zhang らの CrossRAG[3] は，多言語検索によって取得した証拠文書を，生成前に英語などのピボット言語へ一括翻訳する方式を提案した．これにより，生成器は単一言語のコンテキストに基づいて推論を行うことができ，文書順序変動に対する頑健性や構文的一貫性の向上が報告されている．

しかし，統一翻訳アプローチには明確なトレードオフが存在する．すべての証拠を

翻訳することにより，(1) 翻訳トークン数が増加し計算コストが大きくなる，(2) 特に低資源言語ペアで翻訳品質が不安定となり，微妙な数値・固有名詞・時制などの情報が失われやすい，(3) 原文との対応関係が曖昧化し，どの主張がどの原文に基づくのか追跡しにくくなる，といった問題が生じる．

2.3.2 内部知識と外部証拠の統合

Li らの DKM-RAG[4] は，外部から取得した段落をクエリ言語に翻訳した上で，LLM による内部知識チャネルと統合する「二重チャネル型」アーキテクチャを提案した．これにより，検索段階での英語偏重をある程度緩和しつつ，多言語証拠を補完的に利用することを目指している．

しかし，内部チャネルはあくまで生成モデルによる再表現であり，原文との厳密な文単位対応や帰属制約が存在しない．そのため，翻訳や再表現の過程で意味の逸脱が生じても検出が難しく，証拠帰属の透明性という観点では依然として課題が残る．

2.3.3 弁証的推論と多段階生成

Chen らの D-RAG[5] は，検索証拠を「支持」「反証」「中立」といった立場に分解し，弁証的推論に基づいて統合する多段階生成プロセスを採用することで，矛盾する情報を含む状況に対する頑健性向上を目指した．この種の手法は，多言語環境でも矛盾耐性や説明可能性の観点で有望である一方，プロンプト長の増大や推論ステップの増加に伴う計算コストや遅延の問題を抱えており，リアルタイム応答や大規模運用には適用が難しい側面がある．

2.3.4 評価ベンチマークと言語一貫性指標

多言語 RAG の評価基盤として，XOR-QA 系列に加え，Chen らによる XRAG[7] はニュースドメインに基づくベンチマークを構築し，支持文と干渉文を明示的に注釈することで，「どの言語のどの証拠に基づいて回答したか」を分析可能にしている．これにより，(1) 出力言語の誤り，(2) 証拠統合の失敗，(3) 証拠無視やハルシネーション，といった現象を精緻に評価する枠組みが整いつつある．一方で，翻訳規模と性能向上の関係や，コスト正規化された効率指標については，依然として十分に体系化されていない．

2.4 本研究の位置づけ

以上の先行研究を踏まえると，多言語 RAG においては次のような課題が依然として残されている．

- 多言語検索において高資源言語への偏りを抑制しつつ，低資源言語を含む多言語証拠を高カバレッジで取得する手法が必要である．
- 統一翻訳アプローチは扱いやすい一方で，翻訳コストと情報損失が大きく，原文との対応関係も不透明になりやすい．
- 内部知識との統合や多段階推論は表現力を高めるものの，証拠帰属の曖昧さや計算コスト増大という問題を解決していない．
- 多言語 RAG の性能を評価する指標として，正答率（EM, F1）のみならず，出力言語の一貫性や翻訳コストに対する効率を統合的に測定する枠組みが十分に整備されていない．

本研究で提案する LBRAG（Language-Bridged RAG）は，これらのギャップに対して，次の点で差別化される．

1. **多言語混合検索＋ LLM 再ランキング**：多言語埋め込みによる検索と LLM の listwise 再ランキングを統合することで，高リコールな候補集合を得た上で関連度を再評価し，多言語証拠のカバレッジを確保する設計を採用する．
2. **選択的翻訳と可逆アライメント**：全文一括翻訳ではなく，関連度・翻訳品質・トークンコストに基づいて重要段落のみを翻訳対象とする「選択的言語ブリッジ」を導入し，翻訳後には文単位アライメントとスロット整合性検証を行うことで，原文との可逆的対応と証拠帰属可能性を担保する．
3. **言語一貫性とコスト効率を考慮した評価**：従来指標に加えて，出力が目標言語にどれだけ一貫しているかを測る RLC（Response Language Consistency）や，翻訳トークン数あたりの性能向上を測る CNBE（Cost-Normalized Bridging Efficiency）などを用いることで，多言語 RAG に固有の性能–コスト–言語一貫性のトレードオフを定量的に評価する枠組みを提示する．

このように，LBRAG は既存の多言語 RAG 手法が個別に扱ってきた「検索改善」「翻訳戦略」「生成制御」「評価指標」を統合的に設計し，多言語・クロスリンガル環

境において実運用可能なバランス型フレームワークを目指すものである．次章では，LBRAG の具体的なアルゴリズム構成と推論プロセスを形式的に定義する．

第 3 章

提案手法

本章では，本研究で提案する多言語 RAG フレームワーク LBRAG（Language-Bridged Retrieval-Augmented Generation）の構成について述べる．LBRAG は，多言語環境下での高カバレッジな検索と，翻訳コストを抑えた信頼性の高い証拠統合を両立することを目的とし，

1. 多言語混合検索（埋め込み検索と LLM listwise 再ランキング）
2. 選択的言語ブリッジ（重要段落のみへの翻訳適用，文単位アライメントおよびスロット整合性検証）

の二つの主要モジュールから構成される．以下では，まず全体フローを示し，続いて各コンポーネントを形式的に定義する．

3.1 LBRAG の全体構成

LBRAG の処理フローを図3.1に示す．

まず，ユーザ質問 q（言語 L_q）に対して，多言語混合検索モジュールが埋め込み検索と再ランキングを行い，上位 K 件の候補段落集合 D を構築する．（本研究の実験設定では，質問と同一言語の候補を除外して検索することでクロスリンガル状況を作る．）次に，選択的言語ブリッジモジュールが，翻訳予算の範囲内で重要度の高い段落のみを翻訳対象として選択し，翻訳結果と原文との文単位アライメントおよび数値・日付・固有名詞といった重要スロットの整合性を検証する．最終的に，これらの整合済み証拠に基づき，ユーザ言語 L_q で一貫した回答 a を生成する．

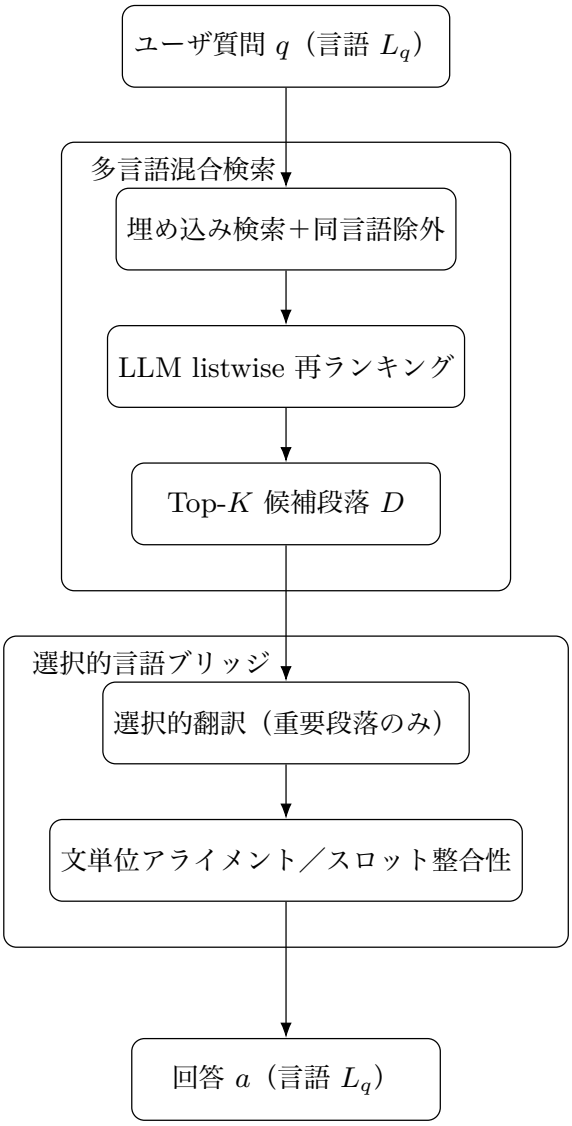

図3.1　LBRAG の全体構成

以下の節では，多言語混合検索と選択的言語ブリッジの具体的設計を述べる．

3.2 多言語混合検索

3.2.1 多言語エンコーダによる初期検索

ユーザ質問を q，その言語を L_q とする．本実装では OpenAI の埋め込みモデル `text-embedding-3-small` をエンコーダ $E(\cdot)$ として用い，クエリおよび段落をベクトルへ写像する．FAISS を用いる場合，クエリ・文書ベクトルを L_2 正規化した上で内積を計算する（コサイン類似度と等価）．表記として次を置く：

$$\mathbf{v}_q = \frac{E(q)}{\|E(q)\|}, \qquad \mathbf{v}_{d_i} = \frac{E(d_i)}{\|E(d_i)\|}. \tag{3.1}$$

（クロスリンガル条件では同言語候補を除外した上で）候補段落 d_i を取得し，コサイン類似度に基づく密検索スコア

$$s_{\mathrm{dense}}(q, d_i) = \mathbf{v}_q^{\top} \mathbf{v}_{d_i} \tag{3.2}$$

を計算する．この段階では，高リコールを重視し，多様な言語の候補を広く収集する．

3.2.2 LLM による listwise 再ランキング

密検索スコアのみでは否定表現や細粒度の意味差異に敏感でないため，本実装では LLM による listwise 再ランキングを用い，質問 q と段落 d_i の関連度 $s_{\mathrm{rerank}}(q, d_i) \in [0, 1]$ を再評価する．両者を線形結合して最終スコア S_{retr} を定義する：

$$S_{\mathrm{retr}}(q, d_i) = \alpha\, s_{\mathrm{dense}}(q, d_i) + (1 - \alpha)\, s_{\mathrm{rerank}}(q, d_i), \tag{3.3}$$

ここで $\alpha \in [0, 1]$ は開発集合により決定されるハイパパラメータである．S_{retr} に基づき上位 K 件を選択し，候補集合 $D = \{d_i\}_{i=1}^{K}$ を得る．

（注）本研究の実装・実験では BM25 やクエリ拡張によるフォールバックは導入していない．

3.3 選択的言語ブリッジ

多言語混合検索により構築された D には，ユーザ言語と英語を含む複数言語の段落が含まれる．すべてを単一言語へ翻訳する統一翻訳は，コストと情報損失の観点から

非効率である．そこで LBRAG では，翻訳予算の下で重要段落のみに翻訳を適用する「選択的言語ブリッジ」を導入する．

3.3.1 翻訳対象段落の選択

ピボット言語 P は $\{L_q, \mathrm{en}\}$ の中から選択する．候補集合 D の各段落 d_i について，

- 検索スコア $s_i = S_{\mathrm{retr}}(q, d_i)$，
- 翻訳コスト c_i（翻訳トークン数に基づく近似），
- **事前推定**翻訳信頼度 $\hat{\kappa}_i \in [0, 1]$ （本実装では言語ペアに基づく単純な事前値：pivot と同言語なら 1.0，英語を含む場合は 0.9，それ以外は 0.7）

を定義する．翻訳予算を B とすると，翻訳対象部分集合 $\Omega \subseteq D$ は次の最適化問題として定式化できる：

$$\max_{\Omega \subseteq D} \sum_{i \in \Omega} s_i \hat{\kappa}_i - \mu \sum_{i \in \Omega} c_i \quad \text{s.t.} \quad \sum_{i \in \Omega} c_i \leq B, \tag{3.4}$$

ここで $\mu > 0$ はコスト感度を調整するパラメータである．実装上は，単純かつ解釈性の高い近似として，各段落の効率比

$$\rho_i = \frac{s_i \hat{\kappa}_i}{c_i} \tag{3.5}$$

を計算し，ρ_i の大きい順に段落を選択していく貪欲法を用いる．これにより，限られた翻訳予算の下で「高関連・高信頼・低コスト」な段落が優先的に翻訳される．

式 (3.5)（= 式 3.4）を数字で理解する：式 (3.5) は，「翻訳に 1 トークン使ったとき，どれだけ役に立つ証拠を持ち込めそうか」を表す．例えば翻訳予算 $B = 35$ のとき，候補が 3 つあり，

$$(s_i, \hat{\kappa}_i, c_i) \in \{(0.80, 0.90, 40),\ (0.60, 0.90, 10),\ (0.50, 0.70, 20)\}$$

だとする（s_i は検索の関連度，$\hat{\kappa}_i$ は「この言語ペアなら大崩れしにくい」という事前値，c_i は翻訳トークン数の近似である）．このとき

$$\rho = \{0.80 \times 0.90/40 = 0.018,\ \ 0.60 \times 0.90/10 = 0.054,\ \ 0.50 \times 0.70/20 = 0.018\}$$

となり，ρ が最大の 2 番目の段落（低コストで十分関連）をまず翻訳するのが合理的になる．残り予算は $35 - 10 = 25$ であり，次に 3 番目（$c = 20$）は翻訳できるが，1

番目（$c = 40$）は予算超過で翻訳できない．つまり式 (3.5) は，「全文翻訳」ではなく「重要箇所にだけ翻訳予算を配分する」ための基準を与えており，本研究の**選択的翻訳によるコスト効率化**の中心に位置づく．

3.3.2 文単位アライメントとスロット整合性検証

翻訳対象となった段落 $d_i \in \Omega$ について，原文言語 $L(d_i)$ からピボット言語 P への翻訳 $\tilde{d}_i = \mathsf{Tr}_{L(d_i)\to P}(d_i)$ を得る．その上で，本実装では正規表現に基づく簡易な文分割を行い，原文の文列と訳文の文列を**順序に従って 1 対 1 に貪欲対応付け**する（先頭から対応させていく）ことで，原文文 s と訳文文 $\tilde{s}$ の対応関係の集合 $\mathcal{A}_i = \{(s, \tilde{s})\}$ を構築する．この簡略化により計算コストを抑えつつ，被覆率やスロット整合率といった「翻訳が怪しいかどうか」の信号を得ることを優先する．

各対応文対 $(s, \tilde{s})$ について，数値・単位・日付・固有名詞などの重要スロットを抽出し，正規化後の一致率に基づき翻訳の整合性を判定する．これにより，段落 d_i に対して

- アライメント被覆率 $u_i \in [0, 1]$（対応づけられた文の割合）
- スロット整合率 $r_i \in [0, 1]$（重要スロットが一致した文の割合）

を算出する．これらの値は，翻訳品質や証拠としての信頼度を定量化する指標として用いる．

なお本実装のスロット抽出は，(i) 数値，(ii) 日付（例：`YYYY-MM-DD`），(iii) 英大文字で始まるトークン（簡易な固有名詞 proxy）を正規表現で抽出し，訳文側に同じ表記が残っているかを確認する単純な方式である．また翻訳後には，(a) 長さ保持（原文と訳文の文字数比），(b) 逆翻訳（back-translation）が可能な場合の語彙重なり，(c) スロット整合性，を組み合わせた指標 κ を計算してメタ情報として保持する．本研究では主に u_i, r_i を証拠重み付けに用い，κ は翻訳品質の補助的な観測量として扱う．

翻訳を行わなかった段落（$i \notin \Omega$）については，原文のまま提示する．本実験では過度な翻訳に起因する情報歪曲を避けるため，未翻訳段落に追加の訳注生成は行わない．

3.4 証拠融合と生成プロンプト設計

LBRAG では，上記で得られた指標を用いて，各段落の生成入力における優先度を決定する．段落 d_i の重み w_i を次式で定義する：

$$w_i = \beta_1 s_i + \beta_2 u_i + \beta_3 r_i, \qquad \beta_1 + \beta_2 + \beta_3 = 1, \tag{3.6}$$

ここで $\beta_1, \beta_2, \beta_3$ はそれぞれ検索関連性，アライメント被覆率，スロット整合性の重要度を表すハイパパラメータである．翻訳を行っていない段落については u_i, r_i を未定義とし，$w_i = \beta_1 s_i$ のみで扱う．

生成プロンプトは，重み w_i の高い順に証拠ブロックを配置し，

- 「回答はユーザ言語 L_q で行うこと」
- (実験設定)「最終回答には引用記号や証拠 ID（例：[...]）を出力しないこと」
- (実験設定)「説明は省き，短い答えのみを返すこと」

といった制約文を組み込むことで，出力言語の一貫性と，評価（EM/F1）における表層ノイズの低減を図る．これにより，選択的翻訳とアライメント情報を活用しつつ，ユーザにとって解釈しやすく検証可能な回答を生成することが可能となる．

3.5 推論プロセスのまとめ

以上を踏まえた LBRAG の推論手順をまとめる：

1. ユーザ質問 q（言語 L_q）を埋め込みモデルでベクトル化し，密検索により候補段落を取得する．(実験設定では同言語候補を除外して検索する.)
2. LLM による listwise 再ランキングを適用し，密検索スコアと線形結合したスコア S_{retr} に基づき，上位 K 件の候補段落集合 D を得る．
3. 翻訳予算 B の下で効率比 ρ_i に基づき翻訳対象集合 Ω を選択し，ピボット言語 P へ選択的翻訳を行う．
4. 翻訳済み段落に対して文単位アライメントとスロット整合性検証を行い，各段落の信頼度指標 u_i, r_i を算出する．
5. 検索スコアと整合性指標を統合した重み w_i に基づき証拠を配置し，出力言語制御および引用ルールを含むプロンプトを構成して，ユーザ言語 L_q における回答 a を生成する．

第 4 章

実証分析

本章では，第 3 章で提案した LBRAG フレームワークの有効性を検証するための実験設定を述べる．本研究の実験は，CrossRAG に代表される「翻訳を介した多言語 RAG」の比較を意識しつつ，学部卒業研究として実装・評価可能な範囲に簡略化した構成で行う．以下では，データセット，知識ベース構築，比較手法，実装条件，評価指標，評価プロトコル，および本実験設定に固有の注意点（簡略化点）を明記する．

4.1 実験設計の概要

評価観点は以下の 4 点である．

- **正確性**：EM および F1 により，質問に正しく回答できるかを測定する．
- **出力言語の一貫性**：回答がユーザ言語で一貫しているか（code-switching の抑制）を RLC で測定する．
- **翻訳コスト効率**：全文翻訳方式と比較して，翻訳トークン数を抑えつつ性能を維持できるかを評価する．
- **意味的妥当性**：表層一致に加え，LLM を用いた意味一致スコア（SAS）で補助評価する．

4.2 データセットと知識ベース

4.2.1 MKQA

本研究では MKQA を用いて多言語質問応答の評価を行う．なお本実験では，Wikipedia 段落などの外部文書コーパスを検索対象とする open-domain 設定ではなく，MKQA の質問・回答ペアを知識ベース（メモリ）として構築し，検索・生成を行う **QA-memory 型の RAG 設定**を採用する．

4.2.2 知識ベース構築（QA-memory）

各サンプル s から文書セグメント d を次のように構成する：

$$d = \texttt{question} \,||\, \texttt{answer}$$

すなわち，知識ベース中の各セグメントは「質問文＋その回答」を含む．これにより，検索は「関連する QA セグメントの想起」という性質を持つ．

4.2.3 クロスリンガル条件の構成

クロスリンガル性を明示するため，評価時の検索では **質問と同一言語のセグメントを除外**し，異言語の証拠（他言語版の QA セグメント）を優先的に取得する．これにより，「質問言語と証拠言語が一致しない」状況を再現する．

4.3 比較手法

本研究では以下の 4 手法を比較する．（括弧内はスクリプト上の名前）

1. **Direct（direct）**：外部文書を与えずに LLM のみで回答する（no-RAG）．
2. **MultiRAG（multi）**：検索は行うが翻訳を行わない（翻訳予算 $B = 0$）．
3. **Full-Translate Pivot RAG（cross）**：取得した証拠を英語（pivot=en）に翻訳して統合し回答する（実装上は翻訳予算を十分大きく設定）．
4. **LBRAG（lbrag）**：選択的翻訳（予算制約）＋文単位アライメント／スロット整合性推定を行い回答する．

注意（CrossRAG との関係）：本研究の `cross` は「pivot translation に基づく全文翻訳」思想を再現する簡略ベースラインであり，CrossRAG 論文の全要素（検索戦略や学習済みモジュール等）を厳密に再現するものではない．

4.4 実装条件

4.4.1 検索・再ランキング

- 埋め込み：OpenAI text-embedding-3-small を使用する．
- 検索：FAISS を用い，ベクトルを L_2 正規化した内積により Top-K 候補を取得する（cosine 類似度に相当）．
- 再ランキング：LLM（`gpt-4o-mini`）による listwise rerank を用いる（候補に対し 0–1 の関連度スコアを推定）．
- K：Top-$K = 10$，密検索と再ランキングの線形結合係数は $\alpha = 0.5$ とする．

4.4.2 翻訳・アライメント

- 翻訳：LLM（`gpt-4o`）を用いた翻訳（数値・日付・固有名詞を保持する指示）を行う．
- pivot 言語：Full-Translate は常に pivot=en，LBRAG は候補分布に応じて pivot を選択し得る（実装既定）．ただし本実験では同言語除外のため，結果的に pivot=en となる場合が多い．
- 翻訳予算：MultiRAG は $B = 0$，Full-Translate は $B \to \infty$（実装上は十分大きい値），LBRAG は $B = 35$ とする．
- 翻訳コスト：トークン数の近似として文字数から簡易推定（$\lfloor |text|/3 \rfloor$）を用いる（実装上の近似）．
- アライメント：文分割後，順序に従って greedy に 1 対 1 対応を作成し，被覆率とスロット一致率を算出する．

4.4.3 生成モデルとプロンプト

- 生成：多言語対応 LLM（`gpt-4o`）を用い，全手法で共通とする．
- 重要制約：「**回答は language のみ**」を明記し，code-switching を抑制する．

- 出力形式：短い答えのみ（説明省略）とする．

4.4.4 実験環境

実行環境（OS，CPU/GPU，ライブラリ，モデルバージョン等）は再現性のため最終版で明記する．[**TODO: 実験環境を追記**]．

4.5 評価指標

4.5.1 EM と F1

EM（正規化後の完全一致）と，言語に応じたトークン化に基づく F1 を用いる．

4.5.2 RLC と RLC-OK

回答がユーザ言語でどれだけ一貫しているかを Response Language Consistency (RLC) で測定する．実装上は，文字種に基づく簡易判定（数字・記号を中立扱い）により算出する．また，閾値（例：0.6）以上を 1 とする二値版（RLC-OK）も併せて報告する．

4.5.3 翻訳コスト

翻訳された証拠の総トークン数（近似値）を集計し，平均翻訳コストとして報告する．

4.5.4 CNBE（Cost-Normalized Bridging Efficiency）

本実験では，翻訳を伴うシステムの「翻訳コストあたりの性能改善」を測るため，Direct（no-RAG）を基準とした CNBE を用いる：

$$\mathrm{CNBE}(j) = \begin{cases} \dfrac{\mathrm{F1}_{\mathrm{system}}(j) - \mathrm{F1}_{\mathrm{direct}}(j)}{\mathrm{TranslationTokens}_{\mathrm{system}}(j)} & (\mathrm{TranslationTokens}_{\mathrm{system}}(j) > 0) \\ 0 & (\mathrm{TranslationTokens}_{\mathrm{system}}(j) = 0) \end{cases}$$

ここで j はテストクエリ（サンプル）を表す．本研究の実装では，各サンプルごとに $\mathrm{CNBE}(j)$ を計算し，その平均と標準偏差（$n = 100$）を報告する．なお，MultiRAG を基準とする定義も考えられるが，本稿では no-RAG からの全体改善を単位コストで

比較する目的で上式を採用する．

4.5.5 SAS（Semantic Agreement Score）

表層一致で捉えにくい同義表現を補助評価するため，LLM による意味一致スコア（0–1）を算出し，SAS として報告する．本指標は補助的であり，最終判断は EM/F1 を主とする．

4.6 評価プロトコル

- テストクエリ数：100（同一 seed による固定サンプリング）．
- サンプリング：同一の意味質問（quid）から 1 つの言語サンプルを選び，重複を避ける．
- 各手法で同一クエリ集合を評価し，平均と標準偏差を報告する．
- 本稿の標準偏差は，**複数回実行のばらつきではなく**，クエリ集合内でのスコア分布のばらつきである．

4.7 本実験設定の簡略化点と限界

本実験は学部卒業研究として実装・評価可能な範囲に簡略化しているため，以下を明記する．

- 検索対象は Wikipedia 段落ではなく，MKQA の QA ペアを用いた **QA-memory 型**である．
- 2 インデックス（L_q と英語）並列検索を厳密に再現するのではなく，**同一知識ベースからの検索＋同言語除外**によりクロスリンガル条件を構成している．
- 再ランキング・翻訳・意味評価に LLM を用いているため，モデルの非決定性により結果が変動し得る．

次章では，上記設定に基づく実験結果を報告する．

第 5 章

実験結果

本章では，第 4 章の設定に基づく実験結果を示し，正確性・言語一貫性・翻訳コスト効率の観点から考察する．本実験はテストクエリ 100 件（$n = 100$）で実施した．

5.1 全体結果（主要指標）

表5.1に，各手法の EM, F1, RLC, RLC-OK, SAS (意味一致), 翻訳コスト, CNBE を示す．値は平均 ± 標準偏差である．

表5.1 MKQA（$n = 100$）における性能比較（平均 ± 標準偏差）

手法	EM	F1	RLC	RLC-OK	SAS
Direct	0.110±0.314	0.193±0.335	0.965±0.139	0.970±0.171	0.412±0
MultiRAG	0.560±0.499	0.624±0.460	0.979±0.081	0.990±0.100	0.708±0
Full-Translate (pivot=en)	0.440±0.499	0.522±0.467	0.956±0.185	0.960±0.197	0.666±0
LBRAG	0.510±0.502	0.573±0.473	0.971±0.130	0.980±0.141	0.668±0

5.2 図による可視化

本研究では，主要指標を図として可視化し，手法間のトレードオフを直感的に比較する．本稿では，実験 run `20251214_2251` の出力図を挿入する．

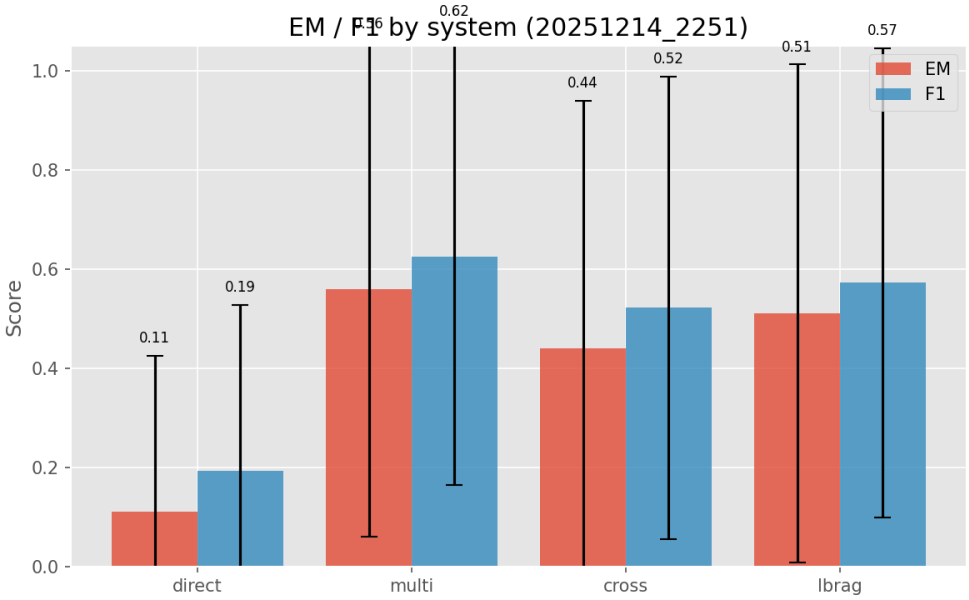

図5.1　EM/F1 の比較（run 20251214_2251）

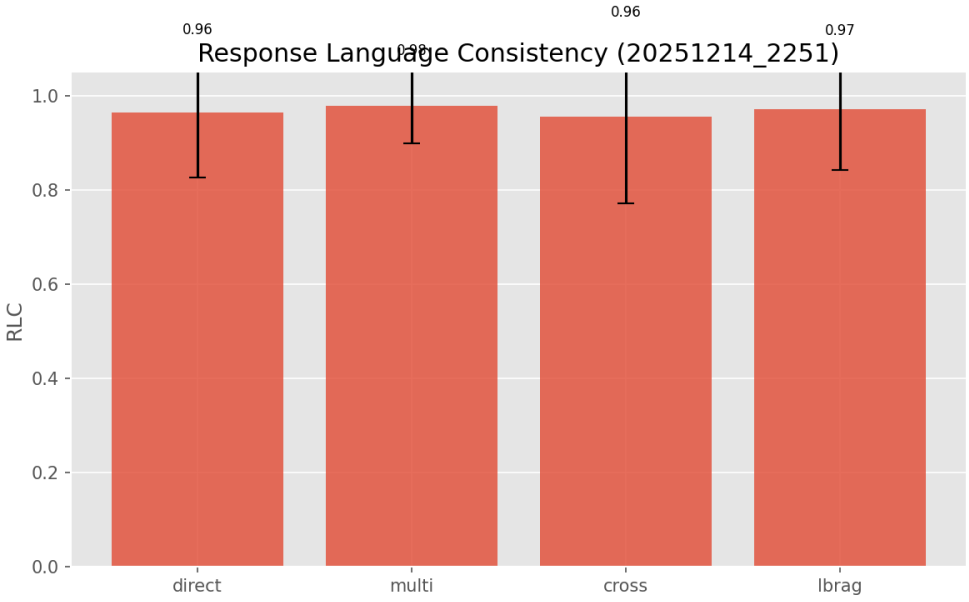

図5.2　RLC の比較（run 20251214_2251）

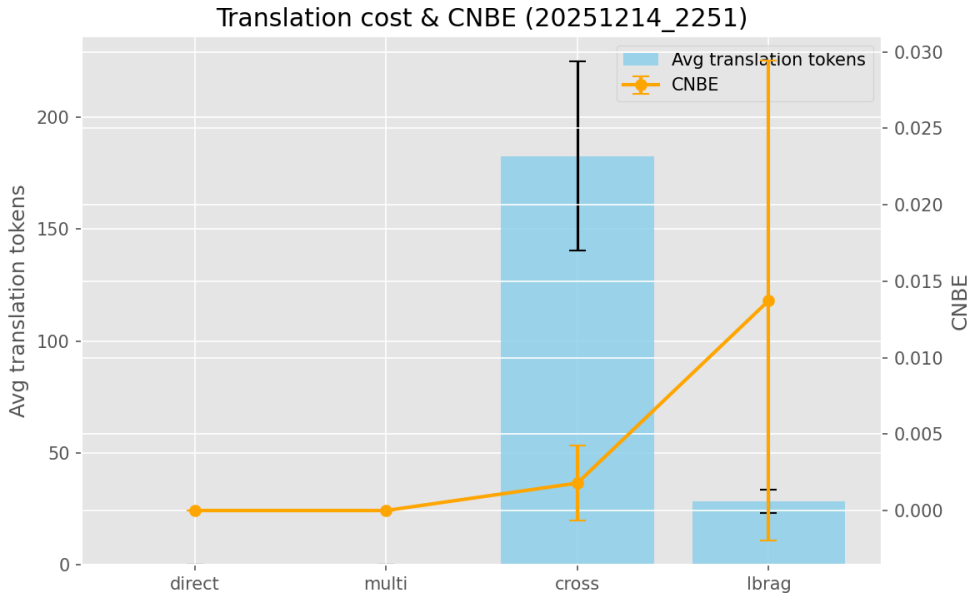

図5.3　翻訳コストと CNBE の比較（run 20251214_2251）

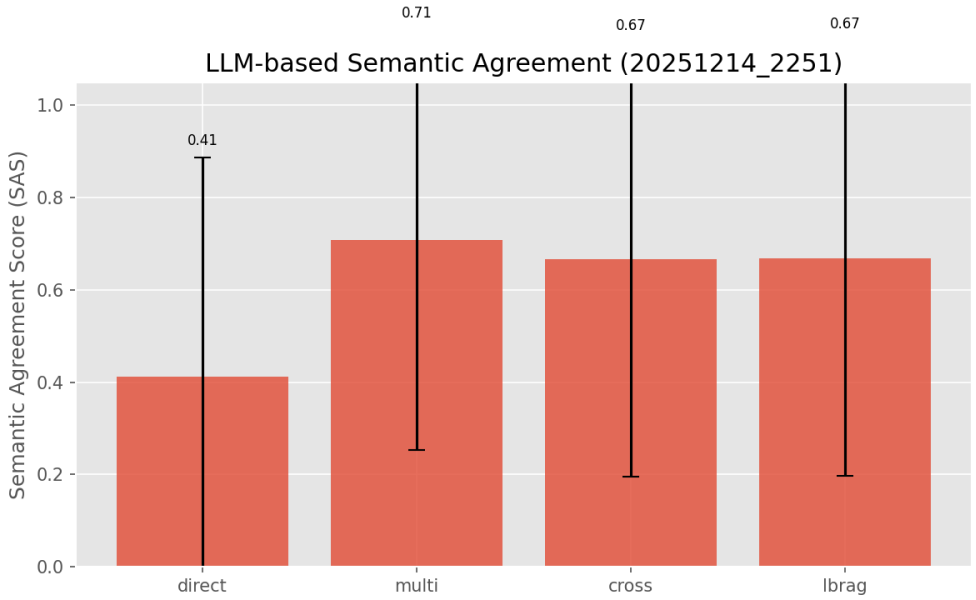

図5.4　SAS（意味一致スコア）の比較（run 20251214_2251）

5.3 考察

5.3.1 正確性（EM/F1）

MultiRAG が最も高い EM/F1 を示し，検索による効果が明確に確認できる．一方で Full-Translate（pivot=en）は MultiRAG を下回り，翻訳を介した統合が常に性能向上につながるわけではないことが示唆される．LBRAG は MultiRAG には及ばないものの，Full-Translate より高い F1 を維持しつつ，大幅な翻訳コスト削減を達成した．

5.3.2 出力言語の一貫性（RLC）

全手法で RLC は高いが，Full-Translate は RLC が相対的に低下しており，pivot 翻訳と生成過程で code-switching が増える可能性が示唆される．LBRAG は RLC を高水準に維持している．

5.3.3 翻訳コストと効率（CNBE）

翻訳コストは Full-Translate が 182.63 に対し，LBRAG は 28.34 と大幅に低い．また CNBE は LBRAG が Full-Translate を大きく上回っており，**限られた翻訳予算下での効率優位**を示す結果となった．

5.3.4 意味一致（SAS）

SAS では MultiRAG が最も高い．LBRAG は Full-Translate と同程度の SAS を維持しながら，翻訳コストを抑制できている．SAS は補助指標であるため，最終的な議論は EM/F1 と合わせて行う．

5.3.5 限界と今後の検証

本結果は QA-memory 型の設定に基づくものであり，Wikipedia 段落検索などの open-domain RAG 設定への一般化は今後の課題である．また，再ランキング・翻訳・意味評価に LLM を用いているため非決定性の影響を受けうる．最終版では実験環境

とモデル設定を明確化し，必要に応じて複数回実行の平均を報告する．

第 6 章

おわりに

本研究では，多言語環境における検索拡張生成（RAG）の実運用上の課題（全文一括翻訳に伴うコスト・情報損失，出力言語の混在（code-switching），そして「翻訳をどこまでやるべきか」という設計上の悩み）に対して，選択的翻訳を核とする LBRAG（Language-Bridged RAG）を提案した．本実装では，(i) 多言語埋め込み検索（`text-embedding-3-small`）と LLM listwise 再ランキング（`gpt-4o-mini`）からなる二段階検索，(ii) 式 (3.4) の効率比 ρ_i に基づく翻訳対象の貪欲選択（予算 B），(iii) 翻訳後の文単位アライメント（順序ベース）と，数値・日付等のスロット整合性に基づく品質信号付与，を組み合わせ，短回答をユーザ言語で一貫して生成する枠組みを構成した．

MKQA の QA-memory 設定（同言語除外によりクロスリンガル条件を構成）における $n = 100$ の実験では，LBRAG は全文翻訳型（pivot=en）と比べて翻訳コストを大幅に抑えつつ，正確性（EM/F1）と出力言語一貫性（RLC）を高い水準で維持し，特に CNBE（翻訳 1 トークン当たりの改善効率）で優位となる傾向が確認できた．この結果は，「翻訳は多ければ良いのではなく，限られた予算を価値の高い証拠に配分する設計が重要である」という本研究の主張を支持する．

一方で，本研究の実装・実験には簡略化も多く，今後の課題・展望として次が挙げられる．

- **翻訳信頼度の事前推定の高度化**：本実装の $\hat{\kappa}_i$ は言語ペアに基づく単純な事前値である．品質推定（QE）や軽量判別器を導入し，翻訳前の選択精度を高めたい．
- **翻訳選択の最適化**：式 (3.4) に基づく貪欲選択は単純で解釈しやすい一方，背

包問題としての最適化や，予算 B を応答時間と連動させるオンライン制御などの検討余地がある．

- **アライメント・スロット検証の頑健化**：順序ベース 1 対 1 アライメントは簡易であるが，文の分割ずれや列挙・表のような構造に弱い．多言語 NER ／正規化器や構造情報を取り込むことで改善が期待できる．
- **open-domain への一般化**：本実験は QA-memory 型であり，Wikipedia 段落検索などの open-domain RAG 設定への一般化は今後の課題である．
- **出力と説明可能性**：本稿では評価の安定化のため短回答・引用なしを採用したが，実運用では根拠提示が重要になる．言語一貫性を保ちつつ根拠を提示する設計を検討したい．

これらの課題に取り組むことで，LBRAG はより広範な多言語・クロスリンガル環境において，低コストかつ信頼できる知識アクセスを提供する枠組みへ発展しうる．

参考文献

[1] P. Lewis, E. Perez, A. Piktus, *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS, 2020.

[2] N. Chirkova, *et al.*, “mRAG: Multilingual Retrieval-Augmented Generation,” EMNLP, 2023.

[3] R. Zhang, *et al.*, “CrossRAG: Pivot Translation for Multilingual Retrieval-Augmented Generation,” ACL, 2024.

[4] Y. Li, *et al.*, “DKM-RAG: Dual Knowledge Merging for Multilingual QA,” EMNLP, 2024.

[5] X. Chen, *et al.*, “D-RAG: Dialectical Reasoning in Multilingual RAG,” ACL, 2025.

[6] A. Asai, *et al.*, “XOR-QA: Cross-Lingual Open-Retrieval Question Answering,” NAACL, 2021.

[7] X. Chen, Y. Liu, S. Huang, “XRAG: A Benchmark for Multilingual Retrieval-Augmented Generation,” ACL, 2025.

謝辞

（ここに本文を記述）

付録 A

Listing 1　LBRAG　placeholder

```
from typing import List, Dict

def retrieve_multilingual(query: str,
                          k: int = 40) -> List[Dict]:
    """埋め込み検索と␣LLM␣再ランキングを用いた
␣␣␣␣多言語検索＋再ランキング（簡略化版）"""
    pass
```