言語記述・比較研究のための LLM ベンチマークの自動生成

村脇 有吾
京都大学大学大学院情報学研究科
murawaki@i.kyoto-u.ac.jp

概要

LLM は多様なタスクをこなせる一方で、利用者の目的に即した性能は必ずしも明らかではない。本稿では、言語記述・比較研究の支援・自動化を目的として、適切な LLM を選定するためのベンチマークを自動生成する手法を示す。具体的には Grambank を例に、言語の類型論的な分類と、その根拠となる文献からの抜粋を組み合わせたデータセットを構築する。閉本・開本設定の性能比較により、根拠文献の言語学的記述の理解能力が評価できる。

1 はじめに

言語記述・比較研究 (フィールド言語学、記述言語学、比較言語学、言語類型論等を総称する) や、考古学や遺伝学などの他分野の知見との統合による人類史研究のためには、データや分析結果を構造化し、再利用可能な形で共有する基盤が重要である。近年は、機械可読なデータ形式としての CLDF (Cross-Linguistic Data Formats) [1] や、Web 閲覧ツールキットである CLLD [2] のようなソフトウェア的基盤の整備とともに、WALS [3] や Glottolog [4] に代表される数多くの言語資源の公開が進んでいる。

しかし、構造化の諸段階には大きな課題が残る。エクセル形式のデータ (言語学者が採用しがち) の CLDF への変換に限らず、手書きフィールドノートのデジタル化、古い紙文献の OCR 処理、音声の書き起こし、グロス付与、転写方式の統一、語彙リストの編纂、個別言語の類型論的分類など、多種多様な作業が必要となる。こうした作業は専門知識に基づく判断を要しつつも反復的であり、また、言語学において業績として評価されにくい作業になりがちであるため、継続的な資源整備の阻害要因となる。

大規模言語モデル (LLM) は多様なタスクをこなせるため、こうした作業を支援・自動化しうる。しかし、利用可能な LLM の選択肢は商用 API からオープンウェイトモデルまで急速に増えている一方で、どのモデルが利用者の目的に適するか (あるいはいずれも適さないか) は必ずしも明らかではない。既存の総合ベンチマークは LLM 開発者が全般的な性能を競う用途に寄っており、個別の利用目的に関する適合性の判断には使いにくい。例えば Humanity's Last Exam [5] は広範な分野を対象とする高難度のベンチマークであり、言語学に関する問題も含むが、その比重は限定的である。かといって、利用者が目的特化型ベンチマークを自前で用意することは、設問設計や正解付与などの点で負担が大きい。

本稿では、既存の言語資源を利用してベンチマークを自動生成する方針を提案し、言語類型論データセット Grambank [6] を例に具体化する。Grambank は、言語学的判断としてのコーディング値に加え、その根拠となる文献情報が付与されている点に特徴がある。本稿では、(i) 特徴定義と対象言語のみからコーディング値を予測する閉本 (closed book) 設定と、(ii) 根拠文献の抜粋を追加で提示する開本 (open book) 設定を対にした評価データを、オープンアクセス (OA) 文献を利用して構築する。両設定の性能差は、根拠文献に含まれる言語学的記述の理解能力を評価する指標として利用できる。

2 Grambank

Grambank [6] は全世界を被覆する大規模な言語類型論データセットである。先行する WALS [3] が音韻・文法・語彙など広い構造特性を扱うのに対し、Grambank は (主に形態統語的な) 文法特徴に焦点を絞っている。また、WALS は特徴ごとに集められたデータの寄せ集めという性格が強く、結果として言語 × 特徴の欠損率が大きい (84%) のに対して、Grambank は最初から体系的に設計・符号化されており、欠損率が小さい (24%) [6]。

Grambank は CLLD に基づく Web アプリ[1)] に加えて、CLDF 形式のデータ[2)] が公開されている。本稿

1) https://grambank.clld.org/
2) https://github.com/grambank/grambank/

では Git commit ID abba55f (2025 年 2 月 5 日) を用いる。言語は 2,467 件、特徴は 195 件であり、言語と特徴の組合せ 481,065 セルのうち、欠損セルは 39,402 で、充足率は 91.8%である。コーディング値には不明を表す?が含まれ、79,638 件である。これは全コーディングの 18.0%に相当する。各特徴は 2–4 値の離散的なコード集合を持ち、195 特徴のうち 189 特徴が 2 値、4 特徴が 3 値、2 特徴が 4 値である。

各特徴には定義とコーディング指針が付与されており、コード集合と対応づく。例えば GB203 は、名詞と集合全称量化詞に相当する語の語順を問う特徴であり、`0` は該当表現がない、`1` は量化詞–名詞、`2` は名詞–量化詞、`3` は両方を許す、といった形で符号化される。例えば日本語 (nucl1643) の GB203 の値は `2`、つまり、名詞–量化詞の語順であり、根拠として文献 `g_Hinds_Japanese[338]` (Hinds (1986:338)) が与えられている。コーディング値は CLDF の `ValueTable` として提供され、言語 ID、特徴 ID、値に加えて根拠文献への参照を含む。

3 LLM ベンチマークの自動生成

3.1 概要とタスク設定

本稿では、言語記述・比較研究の支援・自動化という利用目的に即した LLM 性能評価の実現を目標として、Grambank を例に、既存言語資源からベンチマークを自動生成する枠組みを示す。評価は同一の事例集合に対して二つの設定で行う。(i) 閉本設定では、対象言語と特徴定義のみを与え、Grambank のコーディング値を予測させる。(ii) 開本設定では、これに加えて根拠文献の抜粋を提示し、同様にコーディング値を予測させる。基本的には開本設定が本命であり、根拠文献に含まれる言語学的記述に対する LLM の理解能力を測りたい。閉本設定は LLM が持つ事前知識に基づくベースラインと位置付けられ、両設定の性能差 (根拠提示による改善の程度) に着目する。表 1 に両設定におけるプロンプトテンプレートを示す。

3.2 入力データの整形

入力として、Grambank の CLDF データから言語、特徴、コーディング値、および根拠参照情報を取得する。根拠参照情報は文献キーに加え、ページ範囲などの locator を含みうる。欠損セルは事例化できないため除外する。また、コーディング値には不明を表す?が含まれるため、評価目的に応じて?を含む事例を除外した版も生成可能とする。以降の工程は、言語、特徴、コーディング値、根拠参照情報の組を事例の基本単位として進める。

表 1 プロンプトテンプレート (閉本/開本)

閉本	開本

`You are given a typological feature definition and a target language.`
`Feature:`
`{feature_text}`
`Language:`
`{language_name}`
`Question: What is the correct coded value for this language for the feature above?`
`Answer with the code/value only.`
	`You are given a typological feature definition, a target language, and an excerpt from a reference.`
	`Use ONLY the excerpt as evidence.`
	`Feature:`
	`{feature_text}`
	`Language:`
	`{language_name}`
	`Reference excerpt ({cite}):`
	`{excerpt}`
	`Question: What is the correct coded value for this language for the feature above?`
	`Answer with the code/value only.`

3.3 OA 文献の同定と対象集合の抽出

根拠文献の取得可能性と再現性を担保するため、本稿では OA 文献に限定して根拠抜粋を構築する。具体的には、`sources.bib` の書誌情報から DOI 等の識別子を抽出し、Crossref や Unpaywall 等のサービスを参照して取得可能性を判定する [7]。判定の結果、OA 文献によって裏づけられるコーディング値のみを以降の対象集合として抽出する。

3.4 文献取得とテキスト化

同定された文献をダウンロードし、PDF および HTML を対象としてテキスト化を行う。ただし、配布元の bot 対策や WAF 等により、機械的な一括取得が妨げられる場合がある。そのため、完全自動取得を断念し、必要に応じて手動でのダウンロードを併用する。テキスト抽出では、空の出力や極端に短い出力を検出し、再抽出または除外を行う。

3.5 根拠箇所の同定

Grambank の参照情報は印刷ページを指すことが多く、PDF の物理ページと一致しない場合がある。そのため、参照されたページ範囲に対応する PDF ページを推定し、根拠抜粋を取得できるようにする。具体的には、PDF 各ページのヘッダやフッタからページ番号候補を抽出し、連続性や整合性を手がかりに印刷ページと物理ページのオフセットを推定する。さらに、書誌情報に印刷ページ範囲が明示

されている場合は、その情報を補助信号として用いる。HTML についてはページ概念がないため、参照情報に応じて該当箇所近傍を抽出する方針を採る。

3.6 根拠抜粋付き事例の生成

各事例に対して根拠抜粋を付与し、根拠抜粋付き事例を生成する。抜粋は次の優先順位で作成する。第一に、印刷ページと PDF ページの対応推定に基づき、参照ページ範囲近傍から抽出する。第二に、locator を PDF 物理ページとして解釈できる場合は、その近傍から抽出する。第三に、HTML の場合は該当箇所近傍から抽出する。第四に、参照箇所の同定が困難な場合は保険的に冒頭付近から抽出する。生成に失敗する主な理由は、対象文献を取得できない、またはテキスト抽出に失敗することである。あわせて、空の抜粋や極端に短い抜粋などを検出し、基本的な品質検査を行う。

3.7 評価用ベンチマークの構築

評価用ベンチマークは、同一 ID を持つ閉本版と開本版の対として構築する。開本版には根拠抜粋を含め、閉本版には含めない。根拠に基づく読解をなるべく素直に評価するため、参照情報が明確で根拠箇所を同定できた事例のみを採用する。また、後述のように、正解値が?である事例は除外する。最後に、特徴ごとの偏りを抑えるため、特徴で層化したペアサンプリングにより評価用の部分集合を作成する。

4 構築結果と評価実験

4.1 評価用データセットの構築

Grambank 全体のコーディング値 441,663 件のうち、OA 文献で裏づけ可能な値は 17,333 件であり、対応する根拠文献は 188 件であった。このうち、91 件の文献は実際にダウンロードおよびテキスト化に成功した。これらに基づき根拠抜粋付き事例を生成したところ、生成に成功した事例は 7,643 件であり、さらに参照情報に基づいて根拠箇所を同定できた事例は 1,777 件であった。

評価用データセットでは、正解値が?である事例を除外した。これは、?が「不明」または「判定不能」を表し、閉本設定では特に、モデルの知識不足とデータ側の未確定性が混同されやすいためである。この除外により、正解が定まる分類に評価対象を限定できる一方で、記述の不確実性や境界事例を意図的に落とすことになり、実務上の難しさの一部は反映されにくくなる。

表 2 200 事例データセットの統計 (3 項の値は最小値/中央値/最大値、抜粋の内訳は印刷ページ対応あり/locator を物理ページと解釈)

項目	値
事例数	200
言語数	38
特徴数	100
根拠文献数	37
特徴あたり事例数	1 / 2 / 3
言語あたり事例数	1 / 6 / 9
値の分布（0/1/2/3）	127 / 68 / 4 / 1
抜粋生成方法の内訳	128 / 72
追加提示テキスト量	366 / 2614 / 2636

以降の評価では、根拠箇所を同定できた事例から、特徴ごとの偏りを抑えるため特徴 ID で層化したペアサンプリングを行い、閉本版と開本版が 1 対 1 で対応する 200 事例の評価用データセットを作成した。サンプリング条件は各特徴あたり最大 2 事例を目標として抽出し、200 事例に満たない場合は層化キーに拘らず追加抽出して補う設定とした (一部の特徴は 3 事例となった)。表 2 に作成した 200 事例データセットの統計情報を示す。

4.2 評価実験の設定

200 事例データセットを用いて LLM の性能を評価する。閉本設定と開本設定の 2 条件でモデルに Grambank のコーディング値を予測させる。評価指標は正解率とし、閉本・開本設定の差分もあわせて報告する。

モデルとして商用 API モデルとオープンウェイトモデルを 1 個ずつ採用した。商用 API モデルとして OpenAI の gpt-5-mini (**gpt-5-mini**)[3] を用いた。推論設定は reasoning effort を medium、出力トークン上限を 1024 とした。出力形式は JSON schema を用いて`{"prediction": "..."}`の形に制約し、余分なテキストの混入を抑えた。

オープンウェイトモデルとして Qwen/Qwen3-30B-A3B-Thinking-2507 (**Qwen3-30B**)[4] を用いた。vLLM を用いて稼働させ、OpenAI 互換の Chat Completions

3) https://platform.openai.com/docs/models/gpt-5-mini
4) https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

API 経由で推論を行った。出力トークン上限は 8192 とし、出力は単一のコード値を平文で返すよう指示した。なお、当該モデルは最終出力のほかに、推論過程を<think>タグ付きテキストとして出力する。温度などの確率的デコード設定は明示的には指定せず、各 API のデフォルトに従った。

4.3 評価結果

表 3 に、200 事例データセットにおける閉本設定・開本設定の正解率を示す。差分は開本設定の正解率から閉本設定の正解率を引いた値である。閉本設定では正解率がいずれも 0.5 前後にとどまる。特徴の大半が 2 値であることを踏まえると、LLM は少なくとも本データセットに含まれる特徴値について、事前知識のみではほとんど予測できていない。本データセットでは Lagete 語 (teke1275、9 事例)、Gwama 語 (kwam1249、8 事例)、Hanis 語 (coos1249、8 事例) など、一般にはほとんど知られていない言語が支配的であり、閉本設定での低い正解率はこの点と整合的である。

開本設定では閉本設定と比べて約 0.1 の正解率向上が見られ、根拠抜粋の提示が予測を改善することが確認できる。したがって、既存の言語資源を利用して LLM ベンチマークを自動生成するという基本的なアイデアが、少なくとも根拠提示による改善を観察可能にするという意味で有効であることを確認できた。

しかし、開本設定でも正解率は 0.6 台にとどまり、根拠文献が与えられてもなお誤りが多い。誤りの一因として、Grambank のコーディング指針をそのまま提示しているため、モデルが?を出力候補として選ぶ場合がある点が挙げられる。本稿の評価用データセットでは正解値が?の事例を除外しているため、?の出力は自動的に誤りとなる。?を候補から外すようにプロンプトで明示的に指示すれば正解率向上が期待できるが、?が意味する「判定不能」という選択肢をどの程度許容すべきかは評価目的と密接に関わるため、慎重な検討が必要である。

根拠抜粋の品質にも改善の余地がある。本稿では PDF からのテキスト抽出に pdftotext コマンドを用いたが、PDF 由来のテキストはヘッダやフッタ、脚注、ハイフネーション等の影響でノイズを含みやすい。抽出品質の向上は開本設定の上限性能に直結するため、より高品質な抽出器の利用や、LLM を用いた整形・復元などの工夫が今後の課題である。

表 3 200 事例データセットにおける正解率（閉本/開本）と差分（開本 – 閉本）

モデル	閉本	開本	差分
gpt-5-mini	0.545	0.655	0.110
Qwen3-30B	0.525	0.630	0.105

さらに、印刷ページと PDF ページの対応付けも誤りの温床となりうる。例えば、OA 文献には、出版社版ではなく author manuscript や accepted manuscript が含まれる場合がある。Grambank の参照情報が指すページ番号は通常出版社版に基づくと考えられ、取得した原稿では組版やページ付けが一致せず、参照箇所の同定がずれる可能性がある。これは根拠抜粋の品質を低下させ、開本設定における性能向上を過小評価する要因となりうる。本稿では対応付けが不確かな事例を評価対象から除外しているものの、対応推定自体の精度向上や参照情報の頑健な解釈は、引き続き重要な課題として残る。

5 おわりに

本稿では、言語記述・比較研究の支援・自動化という利用目的に即した LLM 性能評価のために、既存言語資源からベンチマークを自動生成する枠組みを提案し、Grambank を例に具体化した。根拠文献抜粋を与えない閉本設定と、与える開本設定を対にすることで、根拠提示による性能向上を観察できる形にした。評価実験では、開本設定で正解率の改善が見られ、基本的なアイデアの有効性が確認できた。一方で、構築・評価の細部には品質改善の余地が大きい。特に、PDF から抽出したテキストのノイズや、印刷ページと PDF ページの対応付けは、開本設定における上限性能に直結する。根拠抜粋の品質を高める前処理や、参照情報のより頑健な解釈は今後の課題である。

本稿の枠組みは Grambank に限らず、根拠情報や参照が付与された他の言語資源にも適用可能であると考えられる。今後は、資源ごとの設計差を踏まえつつ、適用範囲の拡大と、利用者目的に応じたタスク設計の一般化を進めたい。加えて、法的・倫理的に妥当な形でのベンチマーク公開方法の検討も課題として残る。オープンアクセス文献であっても再配布可能とは限らず、文献ごとにライセンスや利用条件が異なるからである。そのため、根拠抜粋が引用の要件を満たすかの検討に加え、文献再配布の可否や再現手順の設計を含めた精査が必要となる。

謝辞

本研究は一部 JSPS 科研費 24K15068、 24K23937 の助成を受けた。すべてのコードと本稿の過半は著者の指示に基づいて ChatGPT 5.2 Thinking が生成した。

参考文献

[1] Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. **Scientific Data**, Vol. 5, p. 180205, 2018.

[2] Robert Forkel and Sebastian Bank. The clld toolkit. Presentation, Language Comparison with Linguistic Databases: RefLex and Typological Databases, Nijmegen, 2014.

[3] Matthew S. Dryer and Martin Haspelmath, editors. **WALS Online (v2020.4)**. Zenodo, 2013.

[4] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 5.02. Leipzig: Max Planck Institute for Evolutionary Anthropology.

[5] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam, 2025.

[6] Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Russell D. Gray, et al. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. **Science Advances**, Vol. 9, No. 16, p. eadg6175, 2023.

[7] Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. **PeerJ**, Vol. 6, p. e4375, 2018.