言語モデルに線形に内在する階層的概念

坂田将樹 [1,2]　Benjamin Heinzerling[2,1]　伊藤拓海 [1,4]　横井祥 [3,1,2]　乾健太郎 [5,1,2]
[1] 東北大学　[2] 理化学研究所　[3] 国立国語研究所　[4] Langsmith 株式会社　[5] MBZUAI
sakata.masaki.s5@dc.tohoku.ac.jp　benjamin.heinzerling@riken.jp
yokoi@ninjal.ac.jp　t-ito@tohoku.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

本研究は，階層構造（例：Japan ⊂ Eastern Asia ⊂ Asia）が言語モデルの内部表現上でどの程度・どのように符号化されているかを調査する．手法として，Linear Relational Concepts（LRC）に基づき，上位・下位概念間で表現を写像する複数の線形変換器を学習・評価した．地名，人名，組織名，生物名のドメインで実験した結果，同一ドメイン内では階層関係が表現から線形に復元でき，介入実験によって線形変換で得られた特徴が単語予測に因果的に影響を及ぼすことを確認した．一方で，学習した線形変換を学習時と異なる階層や異なるドメインに適用すると介入効果は大きく弱まり，予測に寄与する表現の特徴は階層の深さやドメイン固有であることが示された．これらの結果と追加分析から，言語モデルが階層情報を保持する一方で，それが普遍的な単一軸としてではなく，階層構造の深さ・ドメインごとに局在した低次元部分空間として符号化されていることが示唆された．

1　はじめに

次単語予測で学習された言語モデルが，特定の概念の構造を符号化していることが様々な研究で示されている．例えば，色の構造 [1]，地理情報 [2]，人の出生年の軸 [3]，オセロの盤面 [4] 等の構造である．この言語モデルの内部表現の理解は，テキストのみの学習により世界のどのような概念が符号化されるのか，またそれがタスク遂行にどのような役割を果たすのかという問いに答えるために重要である．

しかしこれらの既存研究 [1, 2, 3, 4] は，言語モデルの内部表現と実世界の一次データ (例えば地理座標や生誕年などの数値属性) の一致度合いを評価するに留まっており，階層構造などの人間が整理し知識として定めた構造とモデルの内部表現がどの程度一致しているかは依然不明瞭である．例えば，言語モデルの内部表現をプローブすることで日本の緯度・経度を当てられても，Japan ⊂ Eastern Asia，Eastern Asia ⊂ Asia といった階層的な関係がどの程度・どのように符号化されているかは不明である．

本研究では，言語モデルの内部表現に階層構造がどの程度・どのように埋め込まれているか，またその表現が単語予測に因果的に影響を及ぼすのかを調査する．例えば，言語モデルの内部表現を通じて “Osaka” が “Japan” という上位概念に所属しているかを評価する．そしてその表現を用いて狙った単語へ変化できるのかという因果的な影響を調べる．実験では各階層の深さ間の内部表現を変換する線形変換器で学習とテストを行う (図 1)．具体的には，Chanin らの Linear Relational Concepts (LRC) [5] を用いる．この手法を用いることで，ある概念がモデルにどの程度線形に符号化されているのかという点と，モデルの出力に因果的に影響を与える特徴量の方向を把握することができる．我々はこの手法を階層の深さおよびドメイン (例: 地名，人名等) ごとに拡張し，それぞれに対応する複数の線形変換器を学習する．この線形変換の束を我々は Linear Hierarchical Concepts (LHC) と呼ぶ．階層深さおよびドメイン毎の線形変換器を比較・分析することで，言語モデルが階層構造をどのような特徴として内部表現に符号化しているかを明らかにする．

実験の結果，同一ドメイン内では，親子関係の階層構造は，表現から線形に復元でき，さらにその方向への介入によってモデルの予測を実際に変化させられることを確認した．一方で，単語予測に影響を与える表現方向は階層の深さやドメインをまたいで一様には共有されず，深さ・ドメインが変わると介入効果が顕著に弱まる場合が多かった．このことは，言語モデルは階層情報を持つものの，普遍的な単一の軸としてではなく，深さやドメインごとに異なる低次元部分空間に局在する形で符号化されていることを示唆している．

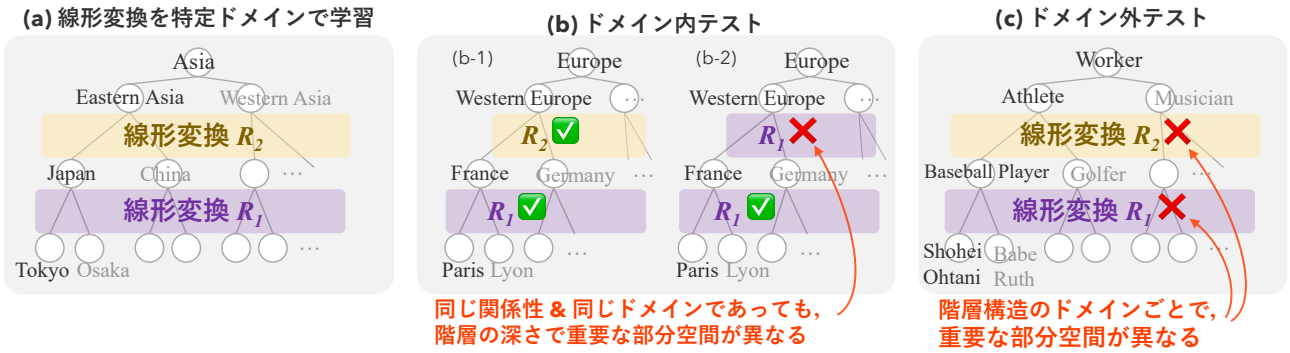

図 1: 線形変換の学習と汎化評価の概要図.

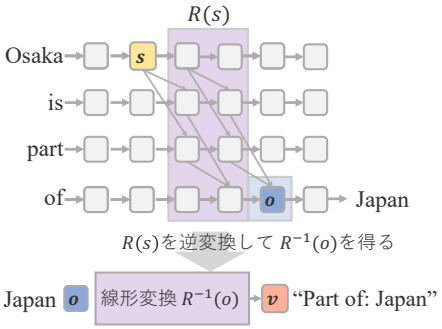

図 2: LRC に基づく階層ごとの概念方向の取得.

2 Linear Hierarchical Concepts

設定 階層構造 $\mathcal{T}$ 上の各データを (s, r, o) として表す．ここで subject s は子ノード，object o はその親ノード，relation r は階層遷移を表すラベルであり深さごとに区別する．（例：Japan→Eastern Asia→Asia における深さ 2–3 の遷移は (s, r, o) = (Japan, PART-OF$_{2\text{–}3}$, Eastern Asia) である）.

階層ごとの概念方向（LHC）の学習 Chanin ら [5] の枠組みに従い，自己回帰型言語モデルを用いて，子ノード s の最終トークン位置 i に対する層 ℓ_s の隠れ状態 $\mathbf{s}_n = \mathbf{h}_i^{(\ell_s)}$ を得る (図 2)．また，親ノード o の表現 $\mathbf{o}_n$ は，o を構成する全トークンの層 ℓ_o における隠れ状態の平均として定義する．そして relation r について，$\mathbf{s}$ から $\mathbf{o}$ への写像を線形関係埋め込み (LRE) [6]

$$R_r(\mathbf{s}) = \mathbf{W}_r \mathbf{s} + \mathbf{b}_r \tag{1}$$

として近似し，$\mathbf{W}_r, \mathbf{b}_r$ を推定する．ここで $\mathbf{W}_r$ は各サンプル (s_i, c_i) におけるヤコビアン $\frac{\partial F}{\partial \mathbf{s}}\big|_{(s_i, c_i)}$ の平均，$\mathbf{b}_r$ は対応する平均バイアスとして算出される．次に，$\mathbf{W}_r$ の低ランク疑似逆行列 $\mathbf{W}_r^\dagger$ を用いて，逆写像

$$R_r^\dagger(\mathbf{o}) = \mathbf{W}_r^\dagger(\mathbf{o} - \mathbf{b}_r) \tag{2}$$

を定義する．Chanin らの方法に従い，この逆写像を用いて親概念 c（このレベルの親ノード）に属する集合 $\mathcal{I}_{r,c} = \{n \mid o_n = c\}$ に対し，概念ベクトル（概念方向）を

$$\tilde{\mathbf{v}}_{r,c} = \frac{1}{|\mathcal{I}_{r,c}|} \sum_{n \in \mathcal{I}_{r,c}} \mathbf{W}_r^\dagger(\mathbf{o}_n - \mathbf{b}_r), \qquad \mathbf{v}_{r,c} = \frac{\tilde{\mathbf{v}}_{r,c}}{\|\tilde{\mathbf{v}}_{r,c}\|_2} \tag{3}$$

で与える．これにより “part-of:Japan” や “part-of:Asia” を表現するベクトル $\mathbf{v}_{r,c}$ を得ることができる．本研究では階層の遷移ごと（各 r）に $\{\mathbf{v}_{r,c}\}_c$ を算出し，その束を Linear Hierarchical Concepts（LHC）と呼び分析に使用する．

評価方法 Chanin ら [5] の方法に従い，Accuracy と Causality の 2 つで評価する．

Accuracy：子ノードがどの親ノードに所属するのかを内部表現を通じて評価する．テスト時の子ノード表現 $\mathbf{a}$ に対し内積が最大の親ノードを予測する：

$$\hat{y} = \arg\max_c \ \mathbf{v}_{r,c} \cdot \mathbf{a}. \tag{4}$$

この予測 $\hat{y}$ の正解率を Accuracy とする．

Causality：概念ベクトル $\mathbf{v}_{r,c}$ を加減算したとき，モデルの予測確率を狙いどおりに変えられるかを測る．例えば “Paris is part of” に対して，“part of: France” の概念ベクトルを差し引きし，“part of: Germany” の概念ベクトルを加えることで “Germany” が予測されるかを評価する．具体的には，元の親 c を別の親 c' としたとき全層で子ノード最終トークン表現を $\mathbf{h}_i^{(\ell)} \leftarrow \mathbf{h}_i^{(\ell)} + \beta \|\mathbf{h}_i^{(\ell)}\|_2 (\mathbf{v}_{r,c'} - \mathbf{v}_{r,c})$ の形で操作し，編集後に c' の予測確率が c を上回るかで成功判定する[1)]．

1) 親ノードがマルチトークンの場合はトークン確率の最小値で評価する．

表 1: ドメイン別の階層深さごとの事例数

深さ	地名	人名	組織名	生物名
0（根）	1	1	1	1
1	5	6	6	6
2	22	35	18	78
3	173	201	194	418
4	24804	–	–	–

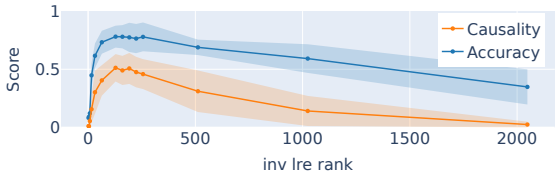

図 3: 擬似逆行列 $\mathbf{W}_r^{\dagger}$ のランクを変化させた結果

3 実験設定

3.1 データ

地名，人名，組織名，生物名の階層構造データを使用した．データの統計情報は表 1，具体例は付録図 8 に示す．言語モデルが階層関係の知識を保持している可能性が高い表現を分析するため，モデルが few-shot の多肢選択式質問応答で正答できる事例のみにフィルタリングした．このフィルタリングで用いたプロンプトを本実験でも使用した．使用したプロンプトは付録の表 9 に示す．データの詳細については付録 A に記載する．学習・テスト分割は，各ドメイン内のツリーをルート直下のサブツリー単位で行った．例えば地名ドメインでは，Europe や Africa 配下のサブツリーで学習し，Asia 配下のサブツリーで評価する．これにより，学習で観測していないサブツリーに対する汎化性能を測定する．

3.2 言語モデル

Llama 3 8B [7] を分析対象とした[2)]．ベースラインとして，Chanin ら [5] の設定と同様にサポートベクターマシン (SVM) と，特定親ノードに対応した子ノードベクトルの平均 (Input Average) を概念ベクトルとして使用した．SVM は未学習のラベルを予測することが難しいため，一部のテストデータを学習に含ませた．そのため，SVM が有利な設定である．詳細な設定は付録 B に記載した．

2) 実験では，Accuracy と Causality が概ねが高かった subject 層=12，object 層=21，擬似逆行列ランク=192 の結果を主に報告する．

表 2: ドメインごとの Accuracy/Causality．スコアは階層間で平均している．階層の深さごとのスコアは付録 図 7 にて示した．

Accuracy

Method	地名	人名	組織名	生物名
Input averaging	0.38	0.51	0.49	0.53
SVM	0.55	0.49	0.68	0.59
LHC	**0.68**	**0.91**	**0.91**	**0.69**

Causality

Method	地名	人名	組織名	生物名
Input averaging	0.26	0.15	0.08	0.08
SVM	0.56	0.36	**0.38**	0.24
LHC	**0.64**	**0.67**	**0.38**	**0.42**

表 3: Accuracy, Causality の成功事例と失敗事例．

Accuracy

	子ノード	親ノード (正解)	親ノード (予測)
成功事例	Cairo	Egypt	Egypt
	Egypt	Northern Africa	Northern Africa
失敗事例	Cetacea	Mammal	Fish

Causality

	子ノード	介入前の単語予測	ターゲット	介入後の単語予測
成功事例	Koh Rong	Cambodia	Japan	Japan
失敗事例	Mahershala Ali	Film actor	Business executive	Film actor

4 実験結果

4.1 ドメイン内テスト

ドメイン内テストの結果を表 2 に示す．全てのドメインにおいて，LHC はベースラインを上回る Accuracy を達成し，約 70%以上の性能が得られた．この結果は，モデルの内部表現には「子ノードがどの親ノードに属するか」という階層的な情報が含まれており，それが線形変換によって十分に復元可能であることを意味する．また Causality についても，多くのドメインで LHC がベースラインより高い値を示した．これは，線形変換器が推定した概念ベクトルが単に復元しやすい特徴のみならず，モデルの予測に対して因果的な寄与を持つことを意味している．成功・失敗事例の具体例は表 3 に示した．追加分析として，階層情報が表現空間のどの程度の自由

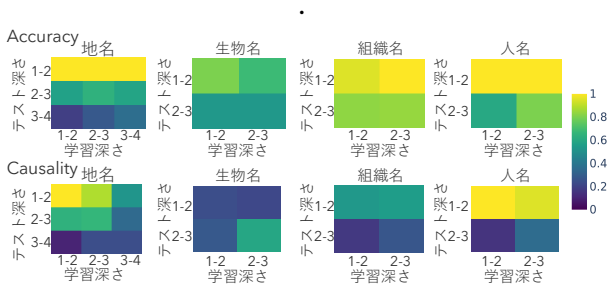

図 4: 階層深さを横断した場合のスコア．

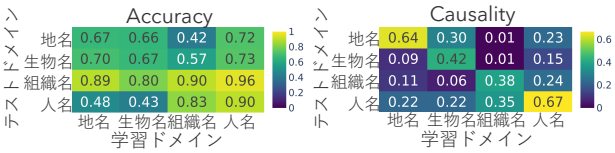

図 5: ドメインを横断した場合のスコア．

度に分布しているかを調べるため，$\mathbf{W}_r^{\dagger}$ を構成するランク k を段階的に変化させて評価した（図 3）．結果，スコアは約 150–200 次元付近でピークを持つことが分かった．このことは，階層情報が表現空間全体に一様に分散するというより，低次元部分空間に集中して符号化されていることを意味している．

次に，階層情報が内部表現で深さに依存せずに表現されているかを調べるため，特定の階層間で学習した線形変換器を別の階層間の予測に用いた (図 4)．上位の遷移（例：深さ 1–2）では学習深さに依らず Accuracy が高い一方，Causality は多くのドメインで対角成分（学習深さ＝評価深さ）で高く，下位の遷移（例：深さ 2–3，3–4）ほど非対角成分で低下した．つまり，同じドメインで同種の関係（is-a/part-of）を扱っていても，深さに応じて予測に因果的に寄与する表現の特徴が異なることが考えられる．

4.2 ドメイン外テスト

異なるドメイン間で，階層情報の符号化のされ方がどの程度共通しているかを調べるため，あるドメインで学習した線形変換を別ドメインに適用して評価した（図 5）．結果，Accuracy はドメインをまたいでも比較的高い組合せが多く，階層ラベルを線形に復元するための手がかりは一定程度ドメイン非依存に共有されていることがわかった．一方で Causality は同一ドメイン（図 5 の対角成分）で高く，ドメインが異なると大きく低下し，組合せによってはほぼ介入が成功しなかった．これは，階層情報を表現から復元できることと，復元のために見つけた方向が

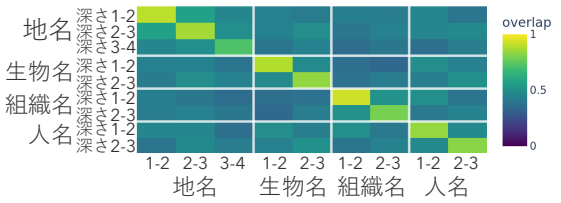

図 6: subject 表現の部分空間の類似度．LRE の重み行列 W を低ランク SVD $W \approx U\mathrm{diag}(s)V^{\top}$ で分解し，右特異ベクトル V が張る部分空間同士の重なりを $\mathrm{overlap}(A,B) = \frac{1}{r}\|V_A^{\top}V_B\|_F^2$ で定義する．

予測を因果的に動かすことは一致せず，後者の介入可能な表現の方向はドメインごとに異なる形で表現されていることを示唆している．より詳細な分析として，線形変換器の重み W を特異値分解し，因果介入に関わる Subject 表現の部分空間の類似度を測定した (図 6)．深さ・ドメインが異なるほど類似度は一貫して低下し，特に下位の遷移ほど上位遷移との類似度が低かった．このことから予測に影響を与える因果的な方向は，階層深さやドメインごとに別の低次元部分空間へ局在していることが示唆される．

5 関連研究

言語モデルにおける階層構造の分析に近い研究として階層概念の幾何を扱った研究 [8] があるが，分析対象は語彙空間に限られている．これに対し本研究ではモデルの中間表現において階層構造がどのように符号化されているのかを調査した．また Chanin ら [5] と Hernandez ら [6] は，subject と Object の関係が線形変換で表現されることを示したが，本研究はこれらの枠組みを階層データに拡張し，言語モデルに内在する階層構造の理解を深めるものである．

6 おわりに

本研究では，LRC に基づく線形変換器群（Linear Hierarchical Concepts; LHC）を用いて，言語モデルの内部表現における階層構造の符号化を分析した．結果，同一ドメイン内では階層の対応関係が線形に復元可能であり，対応する方向への介入により次単語予測が実際に変化することを確認した．一方で，予測に影響を与える表現方向は階層深さやドメインを跨いで一様には共有されず，設定が変わると介入効果は大きく低下した．このことは，言語モデルは階層情報を保持するものの，それは普遍的な単一軸ではなく，階層深さ・ドメインごとに局在した低次元部分空間として符号化されていることを示唆する．

謝辞

本研究は，JSPS 科研費 JP25KJ0628，JP22H05106，JP23K24910，JST BOOST JPMJBY24F9, および JST 創発 JPMJFR2331, の助成を受けたものです．また本研究では産総研及び AIST Solutions が提供する ABCI 3.0 を利用した．

参考文献

[1] Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In Arianna Bisazza and Omri Abend, editors, **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 109–132, Online, November 2021. Association for Computational Linguistics.

[2] Wes Gurnee and Max Tegmark. Language models represent space and time. In **The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024**. OpenReview.net, 2024.

[3] Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric attributes in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 175–195, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In **The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023**. OpenReview.net, 2023.

[5] David Chanin, Anthony Hunter, and Oana-Maria Camburu. Identifying linear relational concepts in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 1524–1535, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In **The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024**. OpenReview.net, 2024.

[7] Abhimanyu Dubey, et al. The llama 3 herd of models, 2024.

[8] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In **The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025**. OpenReview.net, 2025.

[9] Luke Duncalfe. Iso-3166-countries-with-regional-codes (v10.0). GitHub release; commit 145f1ad.

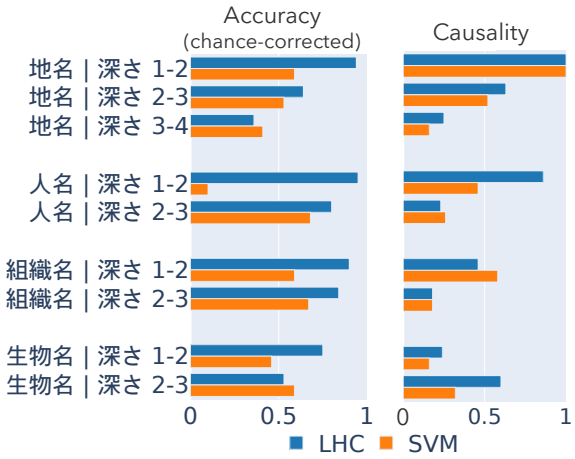

図 7: 階層ごとの Accuracy と Causality

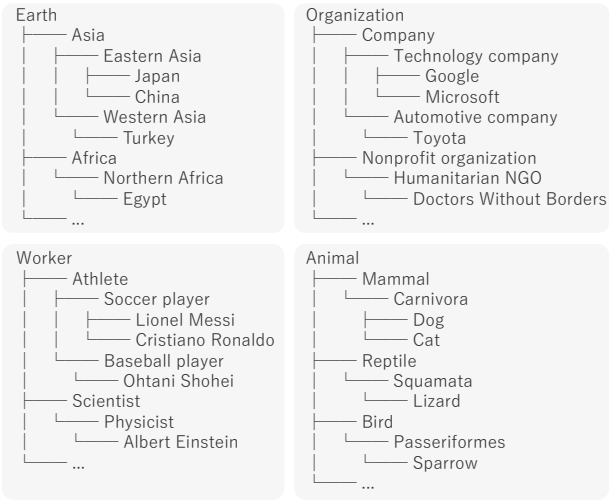

図 8: 階層構造データの具体例

A　データの詳細情報

本研究では，地名・人名・組織名・生物名の複数の階層構造データを使用した．知識ベースには概念階層を構成する分類語彙が存在する一方，ドメインによっては階層化のために便宜的に作られた分類名（実世界では一般に用いられない語）を含む場合がある．また，言語モデルが既知である可能性が高い対象に絞るため，一部ドメインでは GPT-5 で候補（エンティティとクラス階層）を生成し，Wikidata に存在する項目のみにフィルタリングして実在性を担保した．地名は，国連の地理区分に基づく国名分類情報を含む ISO 3166-1 country code dataset [9] と，Gurnee らの地理データ [2] を使用した．人名と組織名は GPT-5 から生成し，いずれも Wikidata に存在する項目のみにフィルタリングした．生物名は Park ら [8] のデータを用い，リンネ式階層分類体系に従って分類した．

B　学習設定の詳細情報

LRC の学習では，Hernandez ら [6] らの設定と同様に 1 つの LRE にあたり 8 個の訓練サンプルを使用した．訓練サンプルは乱数 seed を 5 通り変えて抽出し，報告値は全 seed の平均とした．各サンプルの入力には few-shot プロンプトを用い，コンテキスト内の事例数は 5 とした．評価時も同一形式のプロンプトを使用した．

```
Task: Pick the best higher-level label (parent concept) for
the given Item from the options.
Rules: Output ONLY the chosen label text exactly as
written in the options.

Example 1:
Item: {子ノード名}
Options:
 - {上位概念1}
 - {上位概念2}
 - {上位概念3}
 - {上位概念4}
Answer: {正解の上位概念}

...(一部略)...

Example 5:
Item: {子ノード名}
Options:
 - {上位概念1}
 - {上位概念2}
 - {上位概念3}
 - {上位概念4}
 - {上位概念5}
Answer: {正解の上位概念}

Now answer:
Item: {テスト対象子ノード名}
Options:
 - {選択肢1}
 - {選択肢2}
 - {選択肢3}
 - {選択肢4}
 - {選択肢5}
Answer:
```

図 9: 使用したプロンプトのフォーマット