多言語概念空間の訓練ダイナミクス

ケルナーフェリツィア [1,2]　飯島マックス [3,4]　コホーネンアンナ [5]　プランクバーバラ [1,2]
[1] ミュンヘン大学　[2] ミュンヘン機械学習センター
[3] 東京大学　[4] コペンハーゲン IT 大学　[5] ケンブリッジ大学
iijimax@g.ecc.u-tokyo.ac.jp　{f.koerner,b.plank}@lmu.de　alk23@cam.ac.uk

概要

大規模言語モデル（LLM）は、多言語入力を共有概念空間で処理することで汎化能力や言語間転移を実現していることが先行研究で示されている。しかし、こうした概念空間が訓練過程でどのように形成されるかは十分に明らかになっていない。本研究では、活性化パッチングを用いて EuroLLM の事前学習における言語非依存的概念空間の形成過程を調査した。言語横断的な概念表現を抽出し翻訳プロンプトに注入することで、言語非依存的な概念の変化を検証した。その結果、多言語概念空間は訓練早期に出現し継続的に洗練されるが、各言語との整合性には差異があることが判明した。また、詳細な手動分析により、先行研究で報告された翻訳品質の向上の一部は、翻訳能力の改善ではなく、多義語の語義選択や言語間同形異義語の処理方針の変化といった行動パターンの変化に起因することを明らかにした。

1　はじめに

大規模言語モデル（LLM）は主に英語で訓練されているにもかかわらず [1, 2]、創発的な言語間整合性を示し、他言語への能力転移を可能にする [3]。この行動は、英語と低資源言語間の性能格差を縮小する上で重要である [4, 5]。したがって、言語間整合性がどのように、いつ生じるかを理解することは、高性能な多言語 LLM を目的に応じて開発する上で不可欠である。

近年の解釈可能性研究では、多言語 LLM が多言語入力を処理する際、言語非依存的な**多言語概念空間**にマッピングし、その中でより複雑な処理を行った後、結果を出力言語へマッピングするという理論が提唱されている [1, 7, 8, 9, 10]。こうした空間の存在を因果的に実証するため、[6] は**言語横断的活性化パッチング**（CLAP）を提案している（図 1）。本研究では、この手法を LLM の事前学習中における多言語概念空間の発展を分析するために再構成する。具体的には、ある言語対間（例：スペイン語 → フィンランド語）である概念（例：HOME）を翻訳するプロンプトから潜在ベクトルを抽出し、このソース活性化パッチを**別の**ターゲット言語対（例：エストニア語 → 英語）の**異なる**概念翻訳（例：SOUP）の推論中に適用する。この結果、ターゲット言語（$\ell_{\text{tgt}}^{\text{out}}$ = 英語）でソース概念（C_{src} =HOME）が生成されれば、概念は変化するが出力言語は変化しないため、共有概念表現の存在に対する因果的証拠となる。

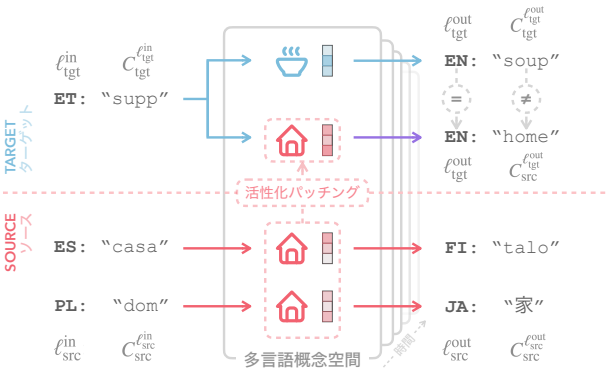

図 1: 言語横断的活性化パッチング（CLAP; [6]）を用いて、多言語の発展を分析する。同一概念の活性化が言語間で整合している場合にのみ、介入による翻訳が誘導される。

活性化パッチングは言語横断的に整合した概念空間の存在を示す証拠を提供するが [6]、そうした空間がどのように発展するかは未解明である。本研究では、多言語 LLM の事前学習チェックポイント間で結果を比較するフレームワークを導入し、多言語概念空間の訓練ダイナミクスを調査する（2.2 節）。EuroLLM [11] の中間チェックポイントにこのフレームワークを適用することで、事前学習中に多言語概念空間がどのように創発するかの高解像度な視点を得る（3 節）。実験の結果、多言語概念空間は訓練の早期に出現し、各言語とこの空間との整合性は訓練

データにおけるその言語の割合に強く依存することが示された（3.1 および 3.2 節）。さらに、パッチング下での翻訳結果のエラーを人手で分析することで、翻訳品質の向上の一部が多義語の曖昧性解消や言語間同形異義語のコピーから翻訳への行動変化を反映していることを明らかにした（3.3 節）。

2 提案手法

2.1 概念の言語横断的活性化パッチング

モデル訓練中における言語間整合性の発展を調査するため、少数ショットの**単語翻訳**を調査する。このタスクでは、概念 C を入力言語 ℓ^{in} から出力言語 ℓ^{out} へ翻訳する。各翻訳は「ℓ^{in}: $C^{\ell^{\text{in}}}$ - ℓ^{out}: $C^{\ell^{\text{out}}}$」という形式のプロンプトとして定式化され、推論時は $C^{\ell^{\text{out}}}$ が省略される。プロンプトには 5 つの少数ショット例が先頭に付加される。

本実験では、言語横断的活性化パッチング (CLAP) を適用することで、推論時に**ターゲット**概念 C_{tgt} から、意味が異なる**ソース**概念 C_{src} への翻訳の変化を誘導することを目指す (図 1)。目的の C_{src} の概念パッチは、C_{src} が意図された翻訳であるプロンプトから活性化を抽出し、複数のソース言語対 $(\ell^{\text{in}}_{\text{src}}, \ell^{\text{out}}_{\text{src}})$ にわたる平均を取ることにより計算される。次に、異なるターゲット言語対 $(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})$ にわたって別の概念 C_{tgt} を翻訳しながら、最終トークン位置に C_{src} のパッチを挿入する。この介入の目的は、ターゲット出力言語 $\ell^{\text{out}}_{\text{tgt}}$ を変えず、C_{tgt} の代わりに C_{src} を誘導することである。ソースとターゲットで異なる言語と概念を使用するため、ターゲット言語でソース概念（つまり、$C_{\text{src}}^{\ell^{\text{out}}_{\text{tgt}}}$）を誘導するには 2 つの表現基準が必要である。第一に、言語間の平均パッチが意味を持つためには、概念表現が言語非依存的でなければならない。第二に、モデルが $\ell^{\text{out}}_{\text{tgt}}$ で C_{src} を生成するためには、ターゲット出力言語もこの言語非依存的空間と整合していなければならない。

2.2 実験設定

データ　多様な概念を提供する Multi-SimLex [12] 内の 11 言語を使用する：英語 (en)、ウェールズ語 (cy)、スペイン語 (es)、エストニア語 (et)、フィンランド語 (fi)、フランス語 (fr)、ポーランド語 (pl)、ロシア語 (ru)、スワヒリ語 (sw)、広東語 (yue)、中国語 (zh)。この中から、全言語にわたって単語の重複がないソースとターゲット概念を 398 個選択した。

モデル　EuroLLM-1.7B [11] の事前学習チェックポイント 26 個を分析する。このモデルは、多言語訓練データ構成の異なる 2 つの段階で訓練されているため、本研究にとって特に興味深い。**フェーズ 1**（訓練の 0–90%; 3.6T トークン）では、データの 77% が一般ウェブデータであり、その中のパラレルコーパスは主に英語に対して整列している。**フェーズ 2**（訓練の 90%–100%; 0.4T トークン）では、一般ウェブ部分が 46.6%に減少し、高品質データ（Wikipedia、arXiv、書籍、医学テキスト）の比率が 9%から 34.4% に増加される上、英語以外に対して整列されているパラレルデータが使用される。

パッチング設定　翻訳ターゲット出力言語 $\ell^{\text{out}}_{\text{tgt}}$ として{en, ru, zh}を選択する。このセットは、類型論的、文字体系的、又は EuroLLM の訓練データ割合的に異なる (表 1)。翻訳入力側 ℓ^{in} では、cy と yue を除くすべての言語を使用する。各ターゲット言語対 $(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})$ について、言語横断的概念パッチを計算する 3 つの方法を用いる: **(1)** seen は、$\{\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}}, \text{en}\}$を除くすべての言語にわたる活性化の平均である。この介入が成功すれば、概念 C_{src} が言語非依存的に表現されていることを示す。**(2)** tgt は、ターゲット言語対内 $(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})$ で目的の C_{src} と異なる概念を用いて計算された活性化パッチを使用する。これは、ターゲット言語特有のパッチのみで概念変化を誘導できるかの検証である。**(3)** en_en は、$\ell^{\text{in}}_{\text{src}} = \ell^{\text{out}}_{\text{src}} =$ en 設定で C_{src} を単純にコピーする英語のみのパッチを使用する。この強力なベースラインは、訓練データ割合で最も多い英語が多言語概念空間と整合しているかを確認するための対照タスクである。さらに、パッチング無しのベースラインとして、src_unpatched は、$(\ell^{\text{in}}_{\text{tgt}}, \ell^{\text{out}}_{\text{tgt}})$ における C_{src} の標準な翻訳性能を測定する。

評価指標　活性化パッチングは通常、パッチされたトークンの確率増加を用いて評価されるが [13]、この指標は、次のトークンが正しく見えてもモデルが誤った予測を出力するエラー（例：「organ」対「organizer」）を隠す可能性がある。この限界に対処するため、本実験では完全なトークン列を評価する。具体的には、$y_i = C_{\text{src}}^{\ell^{\text{out}}_{\text{tgt}}}$ とし、$\hat{y}_i$ をモデル予測とすると、N 個のテストサンプルにわたる平均単語翻訳精度を次のように定義する: $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}$。さらに、エラーの種類 (例：同義語、上位・下位概念) をより細かく分析するため、翻訳の多言語手動評価を実施する（3.3 節）。

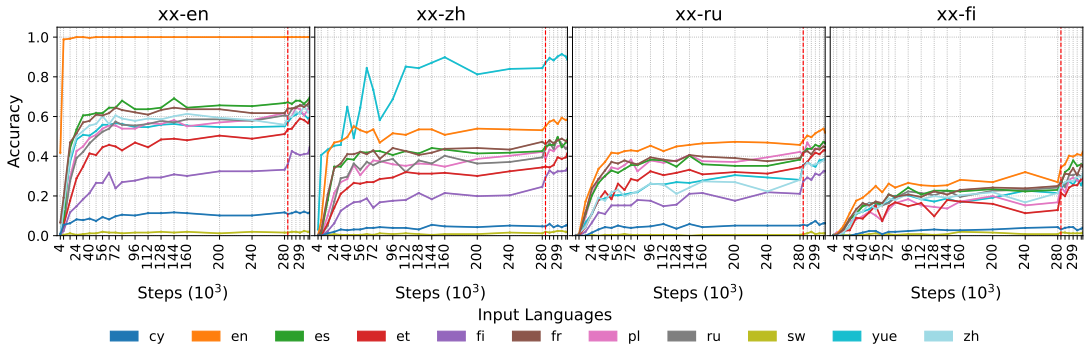

図 2: EuroLLM のチェックポイントにわたるソース概念 $C_{\text{src}}^{l_{\text{src}}^{\text{out}}}$ の平均単語翻訳精度（パッチング無し）。表示されている出力言語は $\ell_{\text{src}}^{\text{out}} \in$ {en, zh, ru, fi}を含む。赤い点線は EuroLLM の訓練フェーズ 1 と 2 の境界。

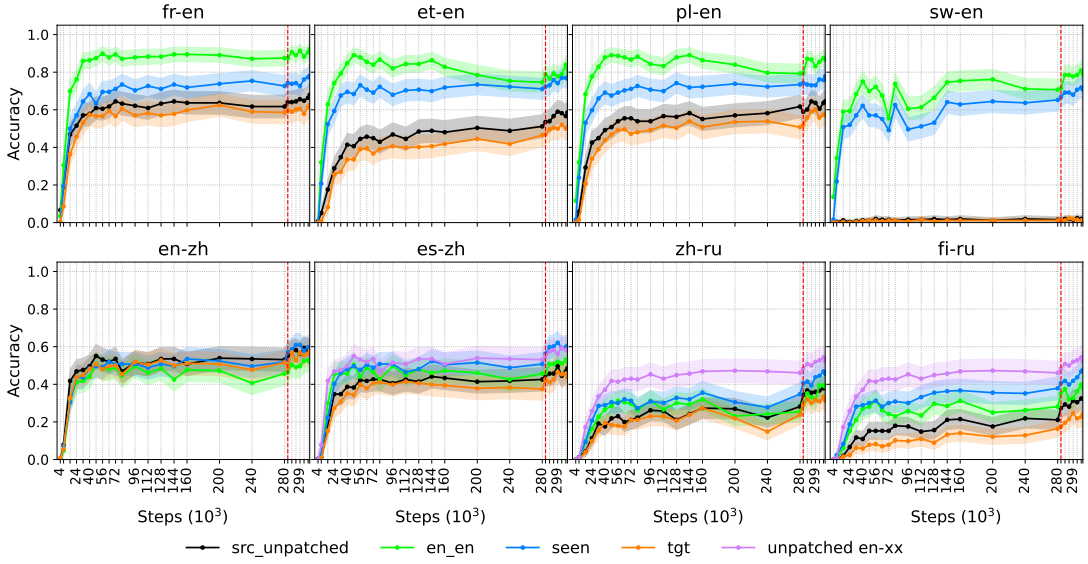

図 3: EuroLLM のチェックポイントにわたる 2 節の各パッチング設定での $C_{\text{src}}^{l_{\text{tgt}}^{\text{out}}}$ の平均単語翻訳精度と 95%信頼区間（パッチング有り、一部の言語対）。en-xx は、zh または ru の上限精度を示す。

3 実験結果

3.1 パッチング無しの翻訳

まず、標準なパッチング無しの単語翻訳において EuroLLM のチェックポイントを評価する。予想どおり、翻訳精度（図 2）は訓練データの割合と相関することがわかる。モデルは、未学習言語（例：cy、sw）の単語翻訳にほぼ失敗している。例外の yue はおそらく訓練データに含まれていた zh との類似性によるものである。翻訳精度は訓練のフェーズ 2 で改善し、その傾向は pl、fi、et などの低資源言語において特に顕著である。フェーズ 2 で導入された多方向のパラレルデータがこれらの言語の整合を助け、翻訳品質を改善したものと考えられる [14, 15]。

en-en の曲線は、モデルが約 4,000 ステップの早期間で入力単語のコピーに習熟することを示している。実際、モデルは訓練の初期段階では、翻訳よりもコピーを好むように見える。この挙動は、コピーが他のタスクより先に学習されることを示唆する最近の研究と一致する [16]。

fr-en の手動検査により、初期のエラーの多くがこのコピー傾向に起因することが明らかになった。これには、「balustrade」(期待値:「rail」) のような言語間借用語をコピーする挙動に加え、「course」(期待値:「racing」) や「coin」(期待値:「corner」) のような誤った同形異義語も含まれる。

3.2 パッチング下での翻訳

図 3 は、言語横断的活性化パッチングにより、他言語からの概念表現（seen）がパッチなし翻訳（src_unpatched）より高精度を示すことを表してい

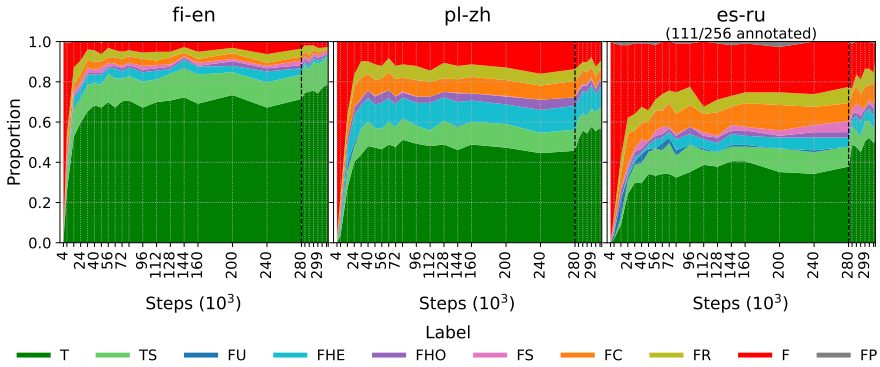

図 4: EuroLLM のチェックポイントにわたる seen パッチング下での単語翻訳の手動評価（表 2 のラベル、一部の言語対）。黒い点線は EuroLLM の訓練フェーズ 1 と 2 の境界。

る。この傾向はほとんどの言語対とチェックポイントで確認され、**多言語概念空間が事前学習の早期に出現する**という証拠を提供する。

en-en は当初、最も高性能なパッチング設定を提供する。しかし、他の言語を使用した seen パッチングとの差は大きくなく、zh と ru への翻訳では英語のみの en-en パッチングを seen が上回っている。en-en と seen の類似した結果は、言語特有の概念空間が言語非依存的な多言語概念空間の出現に大きく先行しないことを示唆している。

パッチングにより入力言語間の精度差は縮小するが、入力言語は依然影響を与える。特に非学習の $\ell^{\text{in}}_{\text{tgt}} = \text{sw}$ では変動が大きく、訓練データ割合が低いほど改善幅が小さい（ru < zh < en）。これは多言語概念空間との整合度を反映していると考えられる。

3.3 質的分析

多言語概念空間の質と進化をより深く理解するため、追加の人手によるエラー分析を実施する（付録 C 参照）。各言語の母語話者が、表 2 のラベルを使用し、seen パッチング下での出力に注釈を付ける。図 4 は、出力が完全に誤っている F よりも部分的に正しい T*と F*の場合が多いことを示している。これは、平均化された概念表現が、目的の概念を誘導する前にすでに関連する概念的および統語的情報を符号化しており、訓練を通じて洗練されていることの証拠である。例えば、FHO は FHE よりも少なく、これは整合性の低い概念空間がより特定性の低い用語にマッピングされやすいという直観と一致する。また、FR は比較的大きなクラスであり、これは言語モデルの訓練目的、すなわちモデルが単語が特定の統語的役割を果たすことを学習することを反映している可能性がある。関連特性の安定性は言語依存的で、例えば ru における FS エラーの多さは複雑な形態論を反映していると考えられる。事前学習のフェーズ 2 では T の割合が TS より増加しており、多方向パラレルデータによる言語間整合性の向上と、より特定的で正確な表現の獲得を示唆している。例えば、fr-en では「avocat」が「attorney」(期待値) または「avocado」に翻訳される。seen パッチング下では、出力は「avocado」から「attorney」へと変化する。es-en では、「mañana」はパッチなしの場合は「morning」に翻訳され、パッチング下では「tomorrow」となる。同様に、パッチングはモデルが言語間同形異義語をコピーすることを抑制する。例えば、fr-en では、モデルはパッチなしの場合「tour」をコピーするが、パッチング下では期待される「tower」を出力する。

4 おわりに

本研究では、事前学習中における多言語概念空間の出現に関する初めての詳細な調査を提示し、言語非依存的な概念空間が**早期に出現し、訓練を通じて比較的安定したまま維持される**ことを示した。この空間は、特定の概念にマッピングされる前であっても、意味的情報を符号化している。この軌跡は言語類似性と訓練データに強く依存している。en のみをピボットとするデータ（EuroLLM の訓練フェーズ 1）より、比較的少量の高品質で多方向整合されたデータ（フェーズ 2）が整合性を改善することを示している。これは、高資源言語と低資源言語間の性能格差を縮める有望な道筋を示唆する：目的を絞った多方向整合データを通じて概念空間との整合性を向上させることである。

謝辞

本研究は、カールスバーグ財団 (Carlsberg Foundation) CF25-0654、欧州研究会議 (ERC) Consolidator Grant DIALECT 101043235 の支援を受けたものです。

参考文献

[1] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? **arXiv**, Vol. arxiv:2402.18815, , 2024.

[2] Junhua Liu and Bin Fu. Responsible multilingual large language models: A survey of development, applications, and societal impact. **arXiv**, Vol. abs/2410.17532, , 2024.

[3] Nadezhda Chirkova and Vassilina Nikoulina. Zero-shot cross-lingual transfer in instruction tuning of large language models. **arXiv**, Vol. abs/2402.14778, , 2024.

[4] Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Hilal Ermiş, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, A. Ustun, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. The multilingual divide and its impact on global ai safety. **arXiv**, Vol. abs/2505.21344, , 2025.

[5] Jiahuan Li, Shujian Huang, Xinyu Dai, and Jiajun Chen. Prealign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In **Conference on Empirical Methods in Natural Language Processing**, 2024.

[6] Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. **arXiv**, Vol. arxiv:2411.08745, , 2025.

[7] Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? **arXiv**, Vol. arxiv:2502.15603, , 2025.

[8] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Hinata Tezuka and Naoya Inoue. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms. **arXiv**, Vol. arxiv:2509.17030, , 2025.

[10] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? **arXiv**, Vol. arxiv:2408.10811, , 2024.

[11] Pedro Henrique Martins, Patrick Fernandes, Joao Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm: Multilingual language models for europe. **arXiv**, Vol. arxiv:2409.16235, , 2024.

[12] Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. **Computational Linguistics**, Vol. 46, No. 4, pp. 847–897, December 2020.

[13] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. **arXiv**, Vol. arxiv:2404.15255, , 2024.

[14] Yingli Shen, Wen Lai, Shuo Wang, Kangyang Luo, Alexander Fraser, and Maosong Sun. From unaligned to aligned: Scaling multilingual llms with multi-way parallel corpora. **arXiv**, Vol. arxiv:2505.14045, , 2025.

[15] Peiqin Lin, Andre Martins, and Hinrich Schuetze. A recipe of parallel corpora exploitation for multilingual large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 4038–4050, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[16] Sheridan Feucht, Eric Todd, Byron Wallace, and David Bau. The dual-route model of induction. **arXiv**, Vol. arxiv:2504.03022, , 2025.

付録

A　言語設定

カテゴリ	言語	訓練データ割合
極多	en	50% → 32.5%
多	es, fr	~6%
中多	zh	3–4%
中	ru, pl	2–3%
少	et, fi	~1%
未学習	sw, cy	–
類似有り	yue	–

表 1: Multi-SimLex（2.2 節）から選択し、EuroLLM の訓練データ割合（2.2 節）により分類した、本研究の焦点言語。en の割合は訓練フェーズ 1 から 2 へ減少する。

広東語（yue）は EuroLLM の訓練データに含まれていないが、翻訳精度は高い（3.1 節）。これはおそらく、訓練データに含まれる中国語（zh）との類似性によるものである。

B　パッチング実装

[6] によって公開された概念 CLAP の実装を基盤とする。プロンプト構築を修正し、すべての実験言語対にわたってソースとターゲット両方に対して同じ概念を含むプロンプトを生成する。また、ソース側とターゲット側の少数ショット例間の重複を防ぐように適応させる。さらに、パッチング下での複数トークン生成を可能にするように更新する。5 つのトークンを生成し、最初に生成された引用符までの出力を解析する。最初の部分文字列がターゲットまたは同義語と一致した場合、過度に長い継続にペナルティを科さないために、これを正解とカウントする。例えば、期待される単語が「absence」の場合の「absence of desire」。

パッチ元の層は、後半の層でない限り、概念を誘導できるかどうかに大きく影響しないと示されている [6]。本研究では、xx-en 設定で初期実験を実施し、14 層までパッチング可能であり、それ以降は概念を確実に誘導できなくなることを確認した。以降の実験では 10 層を元にするパッチを使用する。

ラベル	説明	例
T	完全一致 (自動注釈)	poverty → poverty
TS	同義語	outlander → foreigner
FS	T または TS に対して文法の 1 つの側面が異なる	outlander → foreigners actor → actress
FHE	期待値に対する上位概念	liquor → drink
FHO	期待値に対する下位概念	insect → honeybee
FR	期待値と同位概念	skirt → dress
FU	未翻訳	culture → cultura
FC	期待値と概念的に類似	archer → arrow
F	誤り	aunt → sheep

表 2: 手動分析のラベル定義。ラベルは相互排他的であり、優先順位で順序付けられている。F は他のラベルが適用されない場合にのみ適用される。T と TS は正解とみなされる。

C　手動注釈

各出力言語に、それぞれの言語の母語話者である注釈者が 1 名配置する。注釈の品質を確認するため、注釈者全員が英語出力の一次実験と異なるサブセット（合計 120 例）にラベルを付け、このセットから T と F を除く残りのクラスにおいての注釈者間の Fleiss’ κ 係数を計算する。κ は 0.63 であり、良好な注釈者間の一致を示している。