日本語 LLM は内部でどの表記を経由するか：logit lens による潜在的漢字化の分析

石田茂樹 [1] 辻村有輝 [2] 横田理央 [1] 岡崎直観 [1,2] 高村大也 [2]
[1] 東京科学大学 [2] 産業技術総合研究所
ishida@rio.scrc.iir.isct.ac.jp tsujimura.res@aist.go.jp
rioyokota@rio.scrc.iir.isct.ac.jp okazaki@comp.isct.ac.jp
takamura.hiroya@aist.go.jp

概要

日本語を扱う大規模言語モデルは，ひらがなを出力する際に内部でどのような表記を経由しているのだろうか．本研究では，logit lens を用いて層ごとの内部表現を解析し，日本語中心モデル（LLM-jp-3-13B-Instruct）における，ひらがな出力時に中間層で漢字表現が優勢となる「潜在的漢字化」現象を発見した．一方，英語中心モデルや継続事前学習モデルではこの現象は見られず，漢字化度による定量評価でも日本語中心モデル（0.405）が他モデル（0.02〜0.03）を大きく上回った．さらに，外来語では漢字化が弱まりカタカナが優勢となることから，潜在的漢字化が語彙の表記特性に依存することが明らかになった．これらの結果は，LLM の内部表現形成が事前学習時の言語分布と密接に関連していることを示唆している．

1 はじめに

日本語は，漢字，カタカナ，ひらがなという複数の文字種が併存するという点で特徴的な言語である．例えば learn に対応する日本語単語は「学習」「ガクシュウ」「がくしゅう」といった漢字，カタカナ，ひらがな表記を持つように，ほとんどの単語や表現が複数の表記を持つ．このような特徴を持つ日本語に対する，大規模言語モデル (LLM) の内部機序は明らかになっていない．この内部機序の解明への手がかりとして本稿では，LLM 内部においてどの文字種を経由して処理が行われるか，より具体的には，図 1 の例のようにひらがなを出力するときには内部でもひらがなが「想起」されるのか，あるいは漢字など他の文字種が ‘‘想起’’ されるのか，という問いの答えを探る．いわば，中間層で漢字表現を経由する**潜在的漢字化**を研究対象とする．

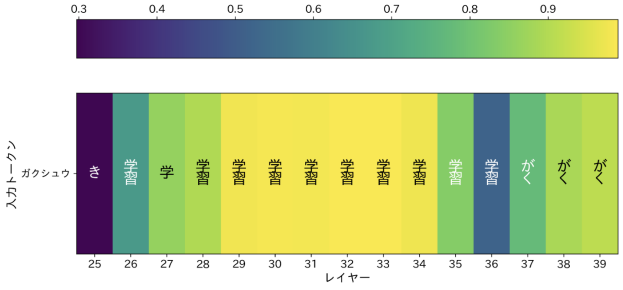


図 1 日本語中心モデル (LLM-jp-3-13B-Instruct) における「ガクシュウ→がくしゅう」変換タスクの解析結果（Top-1 トークンヒートマップ）．各層で最も確率の高いトークンを可視化している．中間層で漢字表現「学習」「学」が現れ，最終層でひらがな「がく」へと収束する．

LLM 内部で ‘‘想起’’ されているトークンを調べる手段として，logit lens[1)] がある．これは，LLM の各中間層の内部状態に出力層を直接連結したツールであり，その内部状態から次トークン予測をすることで，その層で ‘‘想起’’ されているトークンを可視化する．我々はこれを用い，各中間層から出力されるトークンを観察し，またその層から漢字が出力される確率を算出することで，潜在的漢字化について調査する．特に，潜在的漢字化はどの層で観察されるか，どのようなモデルで観察されるか，外来語では潜在的漢字化は緩和されるか，などを調べる．

2 関連研究

LLM の多言語対応能力に関する研究では，英語中心に学習したモデルが非英語言語を処理する際の性質や，内部で用いられる潜在言語に関する分析が行われている [1, 4]．多言語性を高める手法として，英語中心モデルに対する継続事前学習 [2] や，初期

1) https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens

から複数言語のデータをバランス良く学習する手法[3]が提案されている．

モデル内部表現の解析に関しては，機械的解釈可能性の研究 [7, 8] が進展しており，層ごとの表現を可視化する手法として logit lens[6] や tuned lens[9] が提案されている．Saji らの RomanLens[5] は，logit lens と activation patching[10] を組み合わせ，英語中心モデルが非ローマ字言語を処理する際に中間層でローマ字表現が一時的に現れる「潜在的ローマ字化」を報告した．本研究はこれらの先行研究に基づき，日本語中心に学習したモデルに焦点を当て，日本語における複数の文字種間で内部表現がどのように変換・統合されるかを分析する．

3 分析手法

文字種変換タスクにおける文字種の遷移を logit lens により可視化するとともに，漢字化度により漢字トークンの予測傾向を定量化する．

3.1 文字種変換タスク

潜在的漢字化の有無を調べるため，入力をカタカナ，出力をひらがなに固定した「カタカナ→ひらがな」変換タスクを用いる．解析は，プロンプト末尾の次トークン予測を logit lens で比較することで行う．以下の 2 種類のデータセットを用いた：

- **漢字表記可能語（182 ペア）**: 漢字表記が可能な和語・漢語のカタカナ→ひらがなペア（例：ガクシュウ→がくしゅう）
- **外来語（190 ペア）**: 漢字表記が一般的でない欧米語由来の外来語のペア（例：コンピューター→こんぴゅーたー）

データセットは Claude 4[2] により作成した．これらのデータセットで漢字表記可能な語彙と外来語でモデルの内部処理がどのように異なるかを検証する．

プロンプトは 4-shot の few-shot 形式で与え，5 例目のひらがな側をダブルクオート開始で止めて続き（ひらがな）を生成させる：

原文: ”ドクショ” - ひらがな: ”どくしょ”
原文: ”ベンキョウ” - ひらがな: ”べんきょう”
原文: ”ケンガク” - ひらがな: ”けんがく”
原文: ”ジッケン” - ひらがな: ”じっけん”
原文: ”ガクシュウ” - ひらがな: ”

解析では，上記プロンプト末尾（ひらがな側の開始クオート直後）の次トークン予測を対象とし，層ごとの分布を比較する．

2） https://www.anthropic.com/news/claude-4

3.2 解析手順および漢字化度の算出方法

logit lens[6] は，各層の隠れ状態を語彙空間に写像し次トークン分布を得る手法である．本研究では，logit lens で得た各層の分布から，文字種（漢字・ひらがな・カタカナ・英字）ごとの確率を集計し，層ごとの文字種確率分布を算出する．

文字種確率は次のように算出する．まずプロンプト末尾の次トークン予測に対して，各層 l における logit lens の出力から確率分布 $p_l(\cdot)$ を得る．次に，漢字を含むトークンの集合を $\mathcal{K}$ として，層 l における漢字確率を $\sum_{t\in\mathcal{K}} p_l(t)$ として算出する．他の文字種についても同様である．各層の漢字確率を全 L 層にわたって平均し，その事例の漢字化度を得る：

$$\text{漢字化度} = \frac{1}{L}\sum_{l=1}^{L} \text{層}l\text{の漢字確率}. \tag{1}$$

本研究では，データセット全体でこの漢字化度を事例平均した値をモデルの漢字化度とし，潜在的漢字化の度合いを表す指標として用いる．

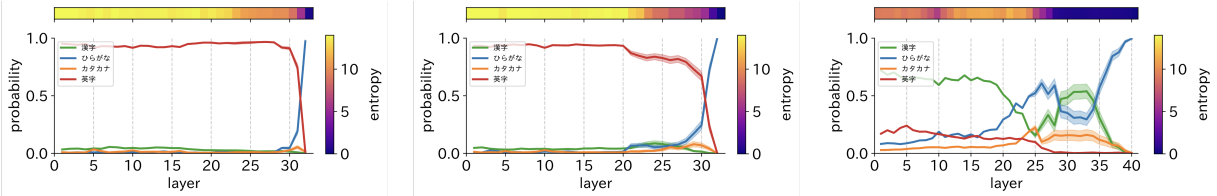
3.3 ひらがな文による解析

前述の文字種変換タスクによる分析は，出力候補となるトークンが限定された特殊な状況での分析であった．より一般的な場合についても，潜在的漢字化について調査を行う．そのため，プロンプトでひらがなでの出力を指示した上で，ひらがな文を LLM 入力することで，出力がひらがなになるように誘導し，その際の内部状態を分析した．具体的には，短文 s について，プロンプトは「ひらがなのみで出力: s」 の形式で与えた．その上で，各トークン位置の内部状態を logit lens で可視化する．

4 実験および結果

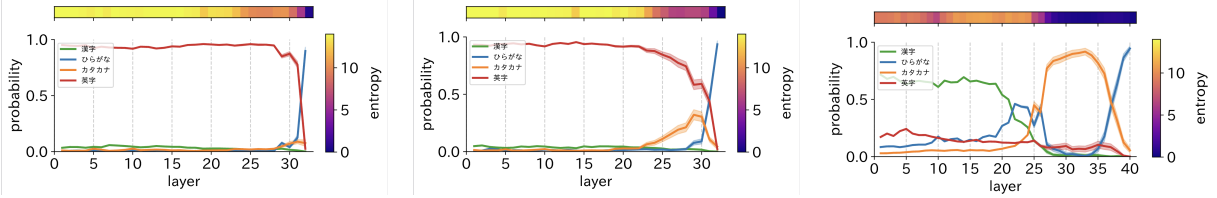
大規模言語モデルの性質が潜在的漢字化現象に与える影響を観察するため，以下の 3 種類のモデルを実験に用いる：**英語中心モデル**（Llama-3.1-8B-Instruct[3]），**日本語継続事前学習モデル**（Llama-3.1-Swallow-8B-Instruct-v0.5[4]）**日本語中心モデル**（LLM-jp-3-13B-Instruct[5]）．加えて，参考

3） https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
4） https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
5） https://huggingface.co/llm-jp/llm-jp-3-13b-instruct3



(a) 英語中心モデル　(b) 日本語継続学習モデル　(c) 日本語中心モデル

図 2　漢字表記可能語データセットにおける 3 種類のモデルの文字種確率分布．横軸は層番号，縦軸は各文字種（漢字，ひらがな，カタカナ，英字）の確率を示す．(a) 英語中心モデルおよび (b) 継続事前学習モデルでは全層を通じて英字が支配的で，最終層でひらがなが急上昇する．(c) 日本語中心モデルでは中間層（Layer 25〜35 付近）で漢字・カタカナトークンの確率が上昇し，最終層でひらがなへと収束する．



(a) 英語中心モデル　(b) 日本語継続学習モデル　(c) 日本語中心モデル

図 3　外来語データセットにおける 3 種類のモデルの文字種確率分布．横軸は層番号，縦軸は各文字種（漢字，ひらがな，カタカナ，英字）の確率を示す．(a) 英語中心モデルでは英字が支配的．(b) 継続事前学習モデルでは中間層でカタカナの確率がやや上昇する．(c) 日本語中心モデルでは中間層でカタカナの確率が大きく上昇し，漢字表記可能語彙で見られた中間層の漢字トークンの再上昇は弱い．

として中国語を中心に学習したモデル Yi-6B-Chat についても同様の解析を行い，結果を付録 B に示す．

4.1　文字種変換タスクの結果

図 2 に，漢字表記可能語に対する 3 種類のモデルの文字種確率分布を示す．英語中心モデルおよび継続事前学習モデルでは，全層を通じて英字トークンの確率が支配的であり，最終層でのみひらがな確率が急上昇する．一方，日本語中心モデルでは，中間層（Layer 25〜35 付近）で漢字確率が上昇し，最終層でひらがなへと収束する「潜在的漢字化」が見られた．中間層ではカタカナ確率の上昇も観測された．

図 3 に，外来語に対する結果を示す．継続事前学習モデルでは中間層でカタカナトークンの確率がやや上昇する傾向が見られる．一方，日本語中心モデルでは中間層でカタカナトークンの確率が大きく上昇し，漢字表記可能語で観察された中間層の漢字確率の上昇（再上昇）は弱い．この結果は，潜在的漢字化が語彙の表記特性（漢字表記の有無）に依存している可能性を示唆している．ここで，外来語には意味的に対応する漢語・和語の言い換えが存在する場合と，対応語が存在しない場合がある．前者では，中間層で漢字表現を中間表現として選択しやすい一方，対応語が存在しない単語ではそのような潜在的漢字化が起こりにくいと考えられる．

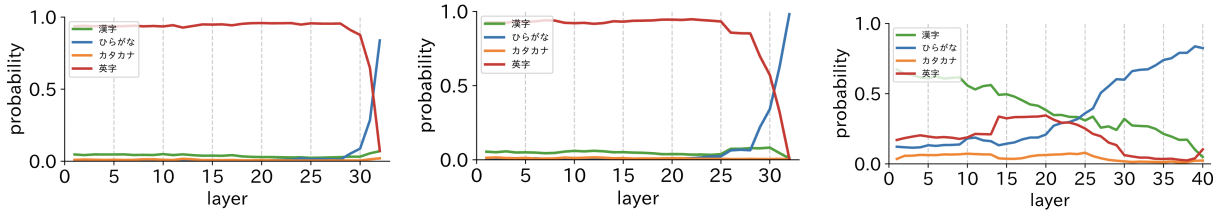
表 1　モデル別の漢字化度（全層平均）．漢字表記可能語彙データセットと外来語データセットで算出した．

モデル	漢字表記可能語	外来語
Llama-3.1-8B-Instruct	0.022	0.021
Llama-3.1-Swallow-8B-Instruct-v0.1	0.031	0.024
LLM-jp-3-13B-Instruct	0.405	0.296
Sarashina2-7B	0.274	0.265
Yi-6B-Chat	0.069	0.095

表 1 に，3.2 節で定義した漢字化度を各モデルについて算出した結果を示す．日本語中心モデルは，漢字表記可能語において約 0.4，外来語においても約 0.3 と，他のモデルに比べて顕著に高い漢字化度を示している．一方，英語中心モデルや継続事前学習モデルでは漢字化度が低く，潜在的漢字化は日本語で学習したモデルに特有の現象であることが定量的に確認された．また，LLM-jp においても外来語では漢字化度が低下しており，漢字表記の有無に応じた内部処理の違いが示唆される．

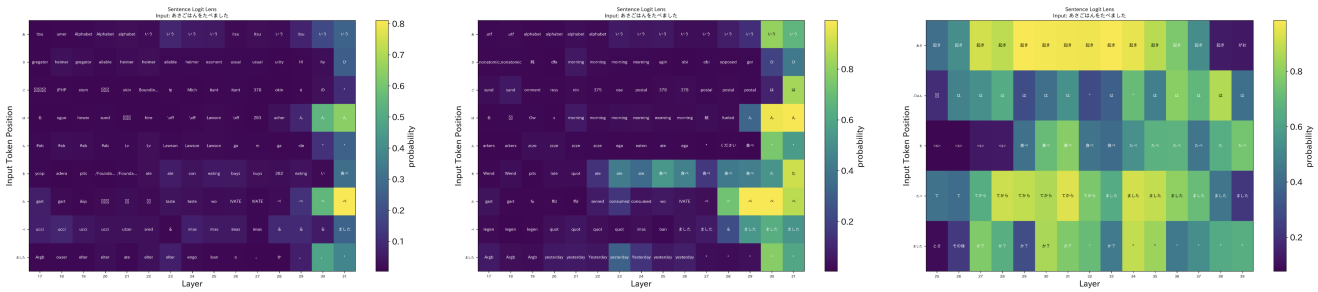
4.2　ひらがな文による解析の結果

入力文 s を「あさごはんをたべました」とした場合の結果を示す．図 4 に，各層における文字種確率分布を示す．ここでは，各トークン位置 p と各層 ℓ において logit lens から得られる次トークン分布に基づき，文字種 c の文字種確率 $P_{\ell,p}(c)$ を計算し，位置方向に平均した $\bar{P}_\ell(c) = \frac{1}{N}\sum_p P_{\ell,p}(c)$（$N$ は対象トークン数）をプロットしている．これにより，文全体として各層がどの文字種を好むかを可視化す



(a) 英語中心モデル　(b) 日本語継続学習モデル　(c) 日本語中心モデル

図 4　ひらがな文（例：「あさごはんをたべました」）を入力した場合の各層における文字種確率分布．(a)(b) 英語中心・継続事前学習モデルでは英字が支配的．(c) 日本語中心モデルでは浅い層で漢字の確率が高く，深い層でひらがなへと収束する．



(a) Llama-3.1-8B-Instruct　(b) Llama-3.1-Swallow-8B-Instruct-v0.5　(c) LLM-jp-3-13B-Instruct

図 5　ひらがな文入力時の Top-1 トークンヒートマップ．各層・各入力位置について，Top-1 トークン（最も確率の高いトークン）とその確率を可視化しており，セルの色は Top-1 トークンの確率を表す．(c) 日本語中心モデルでは中間層で漢字トークン（「起き」「食べ」等）が予測され，深い層でひらがなへと収束する．

る．また，図 5 に Top-1 トークンヒートマップ（色は Top-1 トークンの確率）を示す．日本語中心モデルでは，浅い層で漢字確率が高く，深い層でひらがな確率が上昇する．ヒートマップからも，中間層で「起き」「食べ」などの漢字トークンが予測され，深い層でひらがなへと収束する過程が確認できる．これは文字種変換タスクでの観察と合致する．

4.3　考察

実験結果から，日本語中心に学習したモデルは，ひらがなを出力する場合に内部で一時的に漢字表現を経由していることが示唆される．これは，英語中心モデルにおける「潜在的ローマ字化」[5] と類似した現象であり，学習データの言語分布と内部表現の形成との関連を示唆する．潜在的漢字化は，日本語のように複数の文字種が混在する言語において，モデルが処理を進める際に，何らかの中間コードを用いているという解釈と整合的である．

一方で，継続事前学習モデルは日本語データを追加学習しているにもかかわらず，本研究の設定では同様の漢字優勢な中間表現は観察されない．この結果は，潜在的漢字化が単に日本語データ量の増加で自動的に生じるのではなく，事前学習段階で形成された表現空間や語彙分布，トークナイザの設計など，複数の要因が関与している可能性を示唆する．

5　おわりに

本研究では，日本語を含む多言語モデルの内部表現を解析し，ひらがなを出力する場合に日本語中心モデルが内部で一時的に漢字表現を経由する「潜在的漢字化」現象を観測した．この現象は，漢字表記可能な単語で顕著に観察される一方，外来語では中間層の漢字表現（再上昇）が弱くカタカナが優勢となることから，モデルが漢字表記の有無に応じて異なる内部処理を行っている可能性が示唆された．

この知見は，日本語 LLM の開発において実用的な示唆を与える．モデルにひらがなと漢字の対応関係が十分に備わっていれば，入力表記によらず内部で漢字表現を介した処理が期待できるため，異表記間の対応付け能力を学習時に適切に獲得させることが重要となる．ただし，こうした内部表現の挙動が推論過程の効率化や出力品質にどの程度寄与するかは，本研究の射程を超えており今後の検証課題である．

今後の課題として，より多様なモデルや言語ペアにおける検証，および潜在的漢字化が下流タスクの性能に与える影響の解析が挙げられる．

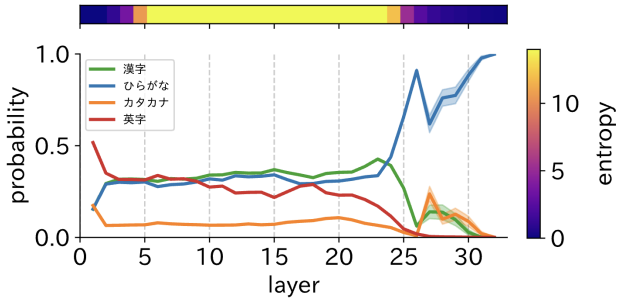
謝辞

この成果は，産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の結果得られたものである．

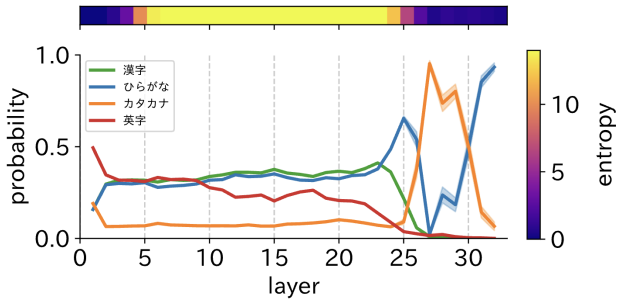
参考文献

[1] Chris Wendler, et al., ”Do Llamas Work in English? On the Latent Language of Multilingual Transformers,” arXiv preprint, 2024.
[2] Kazuki Fujii, et al., ”Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities,” arXiv preprint, 2024.
[3] Akiko Aizawa, et al., ”LLM-jp: Large Japanese Language Models through Balanced Multilingual Training,” arXiv preprint, 2024.
[4] Chengzhi Zhong, et al., ”Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in?” arXiv preprint, 2024.
[5] Alan Saji, et al., ”RomanLens: The Role Of Latent Romanization In Multilinguality In LLMs,” arXiv preprint arXiv:2502.07424, 2026.
[6] Nostalgebraist, ”Logit Lens: Understanding Language Model Representations,” 2020. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens
[7] Nelson Elhage, et al., ”Toy Models of Superposition,” arXiv preprint, 2022.
[8] Yilun Du, et al., ”Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small,” arXiv preprint, 2022.
[9] Noa Belrose, et al., ”Eliciting Latent Predictions from Transformers with the Tuned Lens,” NeurIPS, 2023.
[10] Nelson Elhage, et al., ”A Mathematical Framework for Transformer Circuits,” Anthropic, 2021.

A　Sarashina2 における解析結果



(a) 漢字表記可能語彙

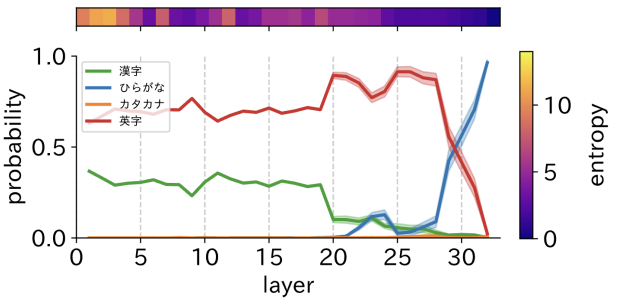


(b) 外来語

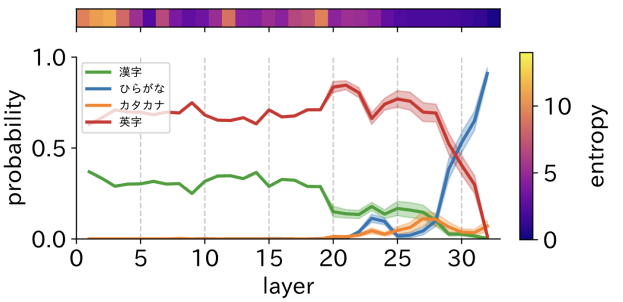
図 6　Sarashina2-7B における文字種確率分布．(a) 漢字表記可能語彙データセット，(b) 外来語データセット．

本研究で用いた解析手法を，日本語を中心に学習したモデルである Sarashina2-7B[6] に適用した結果を示す．図 6 に，漢字表記可能語彙データセットおよび外来語データセットにおける文字種確率分布を示す．漢字表記可能語彙では中間層で漢字確率が一定程度維持されつつ，深い層でひらがなへと収束する傾向が見られる．一方，外来語では深い層でカタカナ確率が大きく上昇し，終盤でひらがなへと遷移する傾向が観察される．

B　中国語モデルにおける解析結果



(a) 漢字表記可能語彙



(b) 外来語

図 7　Yi-6B-Chat（中国語モデル）における文字種確率分布．(a) 漢字表記可能語彙データセット，(b) 外来語データセット．

本研究で用いた解析手法を，中国語を中心に学習したモデルである Yi-6B-Chat[7] に適用した結果を示す．図 7 に，漢字表記可能語彙データセットおよび外来語データセットにおける文字種確率分布を示す．

Yi-6B-Chat は中国語コーパスを主体として学習されたモデルである．本タスクにおいては，英字トークンの確率が支配的であるものの，浅い層から中間層にかけて漢字トークンの確率も一定程度維持されており，最終層でひらがなトークンの確率が急上昇するパターンを示した．

6）https://huggingface.co/sbintuitions/sarashina2-7b

7）https://huggingface.co/01-ai/Yi-6B-Chat