# 文脈内構造切替下での大規模言語モデルにおける文脈内表現学習の幾何学的解析

堀場幸也 [1] 矢向高弘 [1]
[1] 慶應義塾大学
{yukiyahoriba, yakoh}@keio.jp

## 概要

本研究では，長い文脈の途中で遷移規則が切り替わる状況を対象に，大規模言語モデル（LLM）の出力挙動と内部表現の変化を分析する．グラフ遷移タスクを用い，切替前後で語彙を入れ替える条件と維持する条件を比較した．語彙を入れ替えた場合，切替前規則に対応する内部表現は維持される一方で，切替後規則への適応は限定的であった．一方，語彙を共有した場合，出力レベルでは規則間の干渉が生じるが，内部表現は両規則に対応する構造を同時に形成する傾向が観測された．

## 1 はじめに

大規模言語モデル（LLM）は，追加のパラメータ更新を伴わずに，入力された文脈（例示や指示）に応じて出力分布を変化させることができる [1]．この能力は文脈内学習（in-context learning）として広く研究されてきた．一方で，長い文脈における文脈内学習では，文脈の途中で入力–出力関係（規則）が変化する，あるいは複数の規則が混在する非定常な状況も考えられる．Kossen ら [2] は，文脈中のラベル関係を途中で変更する設定を通して，モデルが文脈内情報を一様には利用せず，クエリに近い情報を優先する傾向を報告している．

LLM の挙動を理解するうえで，出力（次トークン予測）だけでなくモデル内部の表現の変化を併せて捉えることが重要である．とくに，Transformer ブロックの残差ストリームは予測に必要な状態（belief state）の幾何を線形に保持し得ることが示されている [3]．また，訓練過程における grokking では，訓練性能が早期に飽和した後に汎化性能が遅れて向上し，その転移に伴って内部表現が幾何学的に整列して低次元の構造が顕在化することが報告されている [4]．このような観点から，文脈内学習を入出力の指標（正解率や確率質量など）のみで評価すると，文脈から抽出された情報が内部でどのような表現として形成・更新されているかを直接には捉えにくい．この点に対し Park ら [5] は，有限無向グラフ上のランダムウォーク列を入力するグラフ遷移（graph tracing）タスクを提案し，文脈から遷移構造を推測できるほど，内部表現がグラフの接続関係に整合するように再編成されること，およびその整合度をディリクレエネルギー（Dirichlet Energy）により定量化できることを示した．本研究でもこの枠組みに立脚し，グラフ遷移タスクのモデルの出力分布に基づく精度指標（近傍ノードへの確率質量）に加えて，タスク構造と直接対応し内部表現に基づく幾何学的指標（ディリクレエネルギー）を併用することで，規則切替に伴う表現の変化を比較する．図 1 に，本稿で扱うタスク設定と指標の関係を概観するための概念図を示す．

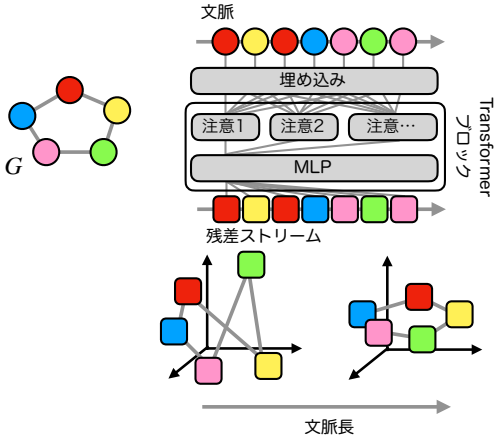


**図 1** 本稿で注目するグラフ遷移タスクと，残差ストリームにおける内部表現の幾何学的な整列（グラフ構造に沿った滑らかさの増大）を模式的に示す概念図．

本稿の中心的な問いは，文脈内で規則が切り替わるとき，モデル内部の表現が切替前規則に整合した幾何構造をどのように扱うかである．具体的には，切替後規則への適応に際して切替前規則の表現構造

を**上書きして更新する**のか，それとも切替前規則の構造を内部に**保持したまま新たな構造を併存して形成する**のかを検証する．この問いに答えるため，グラフ遷移タスクにおいて旧 → 新 → 旧の規則切替を導入する．さらに，上書き・保持の機構を切り分ける手がかりとして，切替前後の語彙の入替有無を操作した二つの条件を用いる．同一語彙条件では，同じトークンが切替前規則と切替後規則で異なる隣接関係を担うため，表現の干渉や再配置が生じうる．一方で，新語彙条件では区間ごとに語彙がトークンレベルで分離され，干渉が抑えられるため，切替前規則に整合した表現構造が切替後を通じて保持されるか，あるいは切替後規則の構造とどのように併存するかを直接に評価できる．

## 2　実験方法と結果

本節では，文脈途中で遷移規則が切り替わるグラフ遷移タスクの設定と評価指標を述べた後，語彙を入れ替える新語彙条件（実験 1）と語彙を維持する同一語彙条件（実験 2）の下で，規則追従精度とディリクレエネルギーの時間変化を比較する．

### 2.1　グラフ遷移タスクと評価指標

語彙集合を $V$ とし，各トークンを無向グラフ $G = (V, E)$ の頂点に対応づける．ここで $E \subseteq V \times V$ は辺（エッジ）の集合であり，$\{u, v\} \in E$ はトークン $u$ と $v$ が隣り合うことを表す．頂点 $u \in V$ の近傍集合を $\mathcal{N}_G(u) \coloneqq \{v \in V \mid \{u, v\} \in E\}$ とする．文脈は，ランダムウォークにより生成した長さ $L$ のトークン列 $u_1, \ldots, u_t, \ldots, u_L$ で，位置（添字）$t \in \{0, \ldots, L\}$ を離散時間 $t$ として解釈し，$t$ をランダムウォークと文脈内の位置を表す時刻と呼ぶ．ランダムウォークでは，初期トークン $u_1$ を一様に選び，各時刻 $t \geq 1$ で $u_{t+1}$ を $\mathcal{N}_G(u_t)$ から一様に選ぶ．

本研究では，文脈途中で遷移規則が切り替わる旧 → 新 → 旧の設定を用いる．具体的には文脈長 $L = 3000$ とし，$t = 1000$（旧 → 新）および $t = 2000$（新 → 旧）で規則（グラフ）を切り替える．また，切替前後の語彙の共有有無により次の 2 条件を比較する．切替前規則のグラフを $G_{\text{old}} = (V, E)$ とし，同一語彙条件における切替後規則のグラフを $G_{\text{new}} = (V, E')$ と表す．本研究では規則切替の効果を明確にするため，同一語彙条件では辺集合が重ならないよう $E \cap E' = \varnothing$ を課す：(i) **新語彙条件**（切替後は別語彙 $V'$ を用いる），(ii) **同一語彙条件**（語彙は共通で，接続関係のみを変更する）．

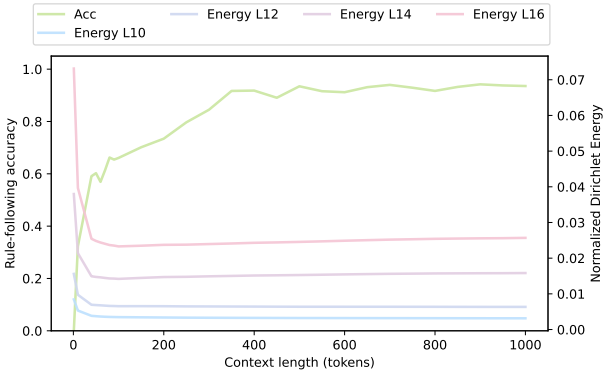


**図 2**　規則切替なしにおける指標の変化（Ring, $t \in [0, 1000]$）．

評価指標として，時刻 $t$ の規則追従精度（近傍への確率質量）を

$$\text{Accuracy}(t) \coloneqq \sum_{v \in \mathcal{N}(u_t)} p(v \mid u_1, \ldots, u_t) \qquad (1)$$

で定義する．ここで $\mathcal{N}(u_t)$ は，その時刻に適用されている規則の近傍集合（切替前では $\mathcal{N}_{G_{\text{old}}}(u_t)$，切替後では $\mathcal{N}_{G_{\text{new}}}(u_t)$）である．

また，内部表現の評価指標として，層 $\ell$ の残差ストリーム $\mathbf{z}_{t,\ell} \in \mathbb{R}^d$ を用いる．区間内で同一トークン $v$ が出現した位置の $\mathbf{z}_{t,\ell}$ を平均してトークン表現 $\mathbf{h}_\ell(v)$ を構成し，グラフ $G = (V, E)$ に対する正規化ディリクレエネルギーを

$$\mathcal{E}_\ell(G) = \frac{1}{d\,|E|} \sum_{\{u,v\} \in E} \|\mathbf{h}_\ell(u) - \mathbf{h}_\ell(v)\|_2^2 \qquad (2)$$

と定義する．$\mathcal{E}_\ell(G)$ が小さいほど，隣接頂点で表現が近い（グラフに沿って滑らか）と解釈できる．以降，$\ell \in \{10, 12, 14, 16\}$ の各層について，切替前規則・切替後規則それぞれの $\mathcal{E}_\ell(G)$ の変化を比較する．

使用モデルは **Llama 3.2 1B** [6]（16 層，$d = 2024$）である．各時刻 $t$ に文脈 $x_{1:t}$ を入力し，softmax により $p(\cdot \mid x_{1:t})$ を取得した．

**結果.** 比較のため，遷移規則を文脈全体で一定に保ち規則切替をしない条件も評価した．これは Park ら [5] と同様の設定である．図 2 に Ring グラフでの結果（$t \in [0, 1000]$）を示す．この条件では，いずれのグラフ族でも，文脈が進むにつれて規則追従精度は全体として上昇傾向を示し，対応するグラフに対するディリクレエネルギーは低下して安定化した．

### 2.2　実験 1：新語彙条件

**設定**　新語彙条件では，切替前後で語彙がトーク

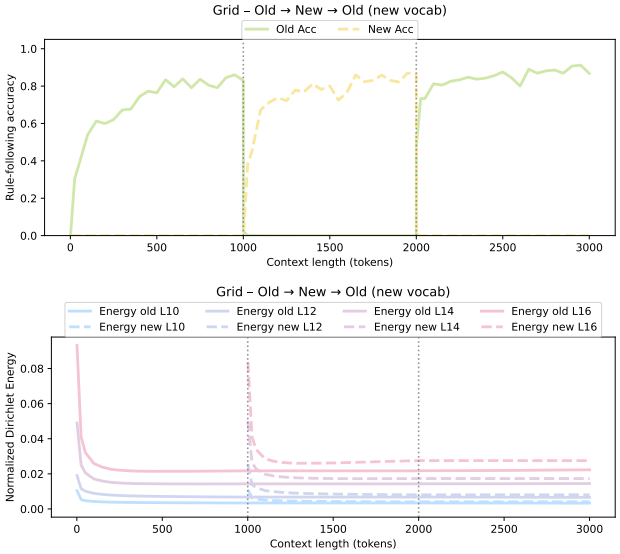


**図 3**　実験 1（新語彙条件）：Grid における規則追従精度（上）とディリクレエネルギー（下）の時間変化．

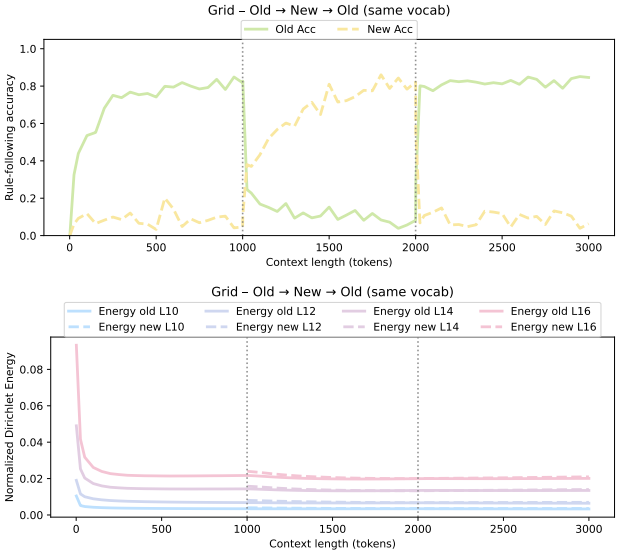


**図 4**　実験 2（同一語彙条件）：Grid における規則追従精度（上）とディリクレエネルギー（下）の時間変化．

ンレベルで分離されるため，切替前規則に整合した表現構造が切替後を通じて保持されるか，また切替後規則の構造がどのように形成されるかを直接に観察できる．

**結果 1.** Grid における新語彙条件の各指標の変化を図 3 に示す．規則切替直後（$t = 1000$）には，切替前規則に対する規則追従精度が大きく低下するが，文脈が進むにつれて切替後規則に対する規則追従精度が上昇し，高い水準へ到達した．同時に，切替前規則・切替後規則に対するディリクレエネルギーは各区間の開始直後に急速に低下して安定化し，規則の接続構造に整合するように内部表現が再配置されることが確認できる．

**結果 2.** 切替後規則に対するディリクレエネルギー $\mathcal{E}_\ell(G_{\text{new}})$ は，切替前規則に対するディリクレエネルギー $\mathcal{E}_\ell(G_{\text{old}})$ よりも高い値までしか減少せず，収束した．

**結果 3．** さらに，切替前規則に対するディリクレエネルギー $\mathcal{E}_\ell(G_{\text{old}})$ は切替後を通じて大きく悪化せず，切替前で形成された滑らかさが維持される傾向が見られた．規則を切替前規則へ戻すと（$t = 2000$），切替前規則側の規則追従精度は比較的短い文脈長で以前の水準へ回復し収束した．全グラフ族（Grid / Hex / Ring）にわたる結果の一覧は，付録の図 5 および図 6 に示す．

## 2.3　実験 2：同一語彙条件

**設定**　同一語彙条件では，同じトークンが切替前規則と切替後規則で異なる隣接関係（遷移役割）を担うため，表現の干渉や再配置（上書き）が生じうる．このとき，(i) 精度指標の干渉，(ii) 切替前規則表現の破壊の有無，を分けて評価する．

**結果 1．** Grid における同一語彙条件の変化を図 4 に示す．切替後では，切替後規則側の規則追従精度が上昇する一方で，切替前規則側の精度は大きく低下し，語彙共有に伴う干渉が次トークン予測の出力分布として明瞭に現れる．

**結果 2.** ディリクレエネルギーの観点では，切替後規則に対する $\mathcal{E}_\ell(G_{\text{new}})$ は切替直後に低下して安定化し，切替前規則に対する $\mathcal{E}_\ell(G_{\text{old}})$ でも時間とともに低下する．$\mathcal{E}_\ell(G_{\text{old}})$ と $\mathcal{E}_\ell(G_{\text{new}})$ がともに近い低い値へ同時に収束する傾向が観測される．

**結果 3．** 切替前規則復帰（$t = 2000$）では，比較的短い文脈長で規則追従精度を回復させ収束させた．このとき，ディリクレエネルギー $\mathcal{E}_\ell(G_{\text{old}})$ と $\mathcal{E}_\ell(G_{\text{new}})$ の双方は近い低い値を保ち続けた．全グラフ族（Grid / Hex / Ring）にわたる結果の一覧は，付録の図 5 および図 6 に示す．

# 3　議論

## 3.1　新語彙条件：二つの規則を保持可能だが，切替後規則の整列は頭打ちになる

新語彙条件では，切替前後で語彙が分離されるため，切替前規則に整合した表現構造の保持可能性を評価できる．実験 1 の結果，切替前規則に対するディリクレエネルギーは切替後を通じて大きく悪化

せず，切替前で獲得された滑らかさが維持される傾向が確認された．一方で，切替後規則に対するディリクレエネルギーは切替後の初期に急速に低下するものの，切替前規則側より高い値で収束し，切替前規則と同程度の滑らかさには到達しない．このことは，「切替前規則の保持」と「切替後規則の十分な幾何学的整列」が独立した現象であり，語彙分離が保持を助けても，切替後規則表現の形成には別の制約が残る可能性を示唆する．

先行研究は，単一の文脈長における「切替」や「追加」がモデルの推論を不安定化し得ることを示している．Gupta ら [7] は，会話履歴が特定タスクで構成された状態から別タスクへ切り替わると，ターゲットタスクの性能が低下する例が多数存在することを報告し，タスク切替による干渉として体系化している．さらに Coleman ら [8] は，同一文脈内で情報を追加し続ける設定において，後続情報が先行情報の想起を妨げ，性能が段階的に低下することを示している．

これらの知見を踏まえると，本研究の新語彙条件で観測された「切替前規則表現の維持」と「切替後規則側の幾何学的指標の頭打ち」は，タスク切替や文脈更新に伴う干渉が，必ずしも直ちに精度低下として顕在化しない場合でも，内部表現の再配置に制約として現れ得ることを示唆する．

### 3.2 同一語彙条件：二つの規則は両立し，より幾何学的に滑らかに再配置される

同一語彙条件では，切替後規則への適応に伴い切替前規則の規則追従精度が低下し，出力分布レベルでは明確な干渉が観測された．一方で，内部表現の幾何学的指標に着目すると，切替前・切替後規則の双方に対するディリクレエネルギーが低下し，特に切替前規則に対するエネルギーは切替前規則のみを提示した条件よりもさらに低い値を示した．これは，共通語彙が導入されることで，モデルが二つの規則をどちらか一方に負の制約をする形ではなく，共通の表現配置へと再編成する可能性を示唆する．ただし，ディリクレエネルギーの改善が直に規則追従精度の安定性を保証するわけではなく，実際に規則追従精度では干渉が残ることから，表現の滑らかさとモデルの出力の間にはギャップが存在する．

また，この条件では同一のトークンが切替前・切替後規則の双方で出現し得るため，局所的な n-gram 統計だけでは次トークンが一意に定まらず，現在どの規則に属しているのかを表す潜在的な状態を推定し，その状態に応じて参照すべき遷移を切り替える必要が生じる．この点は，注意機構の多頭ヘッドの一部が潜在的な文脈を高精度に復号でき，それらのヘッドをまとめてアブレーションすると，文脈内の予測精度が低下することも報告されている [9]．

同一トークンによる切替前・切替後規則が同一文脈内に共存すると，内部表現の幾何は両規則に対して同時に改善し得る．この点は文脈内学習において，文脈内の例集合を多様にして必要な局所構造を広く被覆すると，モデルがそれらの構造を融合して扱いやすくなり推論が改善することが報告されている [10]．本稿の同一語彙条件は，語彙を固定したまま遷移規則だけを切り替えることで，同一の語彙表現上に異なる局所遷移構造を同時に提示する設定とみなせるため，被覆の増加が表現の再配置を促し，両規則に対するディリクレエネルギーの低下として現れた可能性がある．

## 4 結論

本稿では，グラフ遷移タスクにおける規則切替を対象に，語彙の入替有無を操作し，規則追従精度と内部表現の滑らかさ（ディリクレエネルギー）から，LLM が切替前規則の幾何学的構造を上書きするのか，保持したまま併存させるのかを検討した．主な知見は以下の通りであり，これは文脈内の非定常な規則切替における干渉と更新の関係に示唆を与えると考える．

- 新語彙条件では，切替前規則に整合した表現の滑らかさは切替後を通じて大きく崩れにくい一方で，切替後規則側の幾何学的整列は切替前規則と同程度には到達せず，頭打ちとなる傾向が見られた．
- 同一語彙条件では，出力分布レベルでは切替前・切替後規則の干渉が明瞭に生じる一方で，内部表現の幾何は両規則に対して同時に滑らかになる傾向が観測された．

今後は (1) 幾何的指標と出力干渉の乖離を生む機構の同定，(2) 規則切替の回数・間隔・タスク構造（多規則・階層規則など）やモデル規模を拡張した一般化が課題となる．(1) は保持や上書き，再配置がどの機構により実現されるかの解釈可能性を高め，(2) は本稿の知見が自然言語的な文脈更新・タスク切替へどこまで外挿できるかを明らかにすると期待する．

# 参考文献


[1] Tom B. Brown, et al. Language models are few-shot learners. **arXiv preprint arXiv:2005.14165**, 2020.

[2] Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In **The Twelfth International Conference on Learning Representations**, 2024.

[3] Adam S. Shai, Lucas Teixeira, Alexander Gietelink Oldenziel, Sarah Marzen, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In **Advances in Neural Information Processing Systems 38**, 2024.

[4] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In **Advances in Neural Information Processing Systems 35**, 2022.

[5] Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: in-context learning of representations. In **The Thirteenth International Conference on Learning Representations**, 2025.

[6] Aaron Grattafiori, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[7] Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark J. F. Gales, and Mario Fritz. LLM task interference: An initial study on the impact of task-switch in conversational history. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 14633–14652. Association for Computational Linguistics, 2024.

[8] Eric Nuertey Coleman, Julio Hurtado, and Vincenzo Lomonaco. In-context interference in chat-based large language models. **arXiv preprint arXiv:2309.12727**, 2023.

[9] Tankred Saanum, Can Demircan, Samuel J. Gershman, and Eric Schulz. A circuit for predicting hierarchical structure in-context in large language models. **arXiv preprint arXiv:2509.21534**, 2025.

[10] Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, Vol. 1: Long Papers, pp. 1401–1422. Association for Computational Linguistics, 2023.

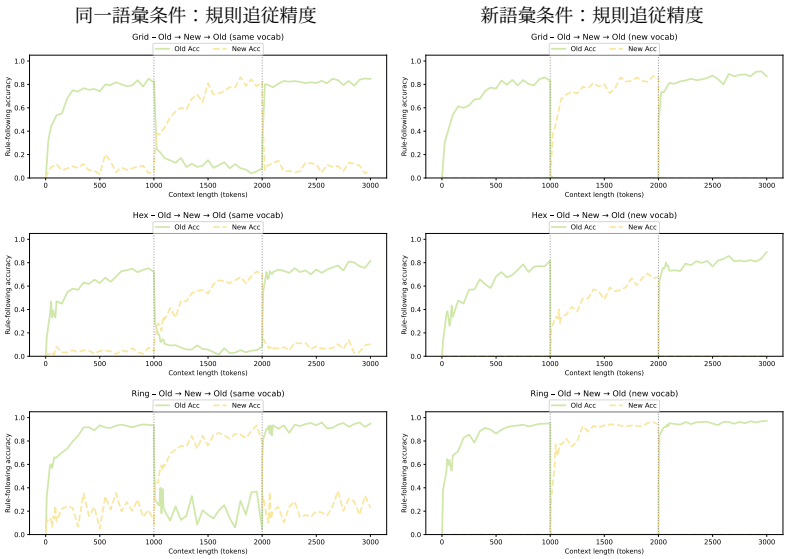


**図 5**　遷移規則の切替（旧 → 新 → 旧）における規則追従精度の時間変化．上から Grid / Hex / Ring に対応する．

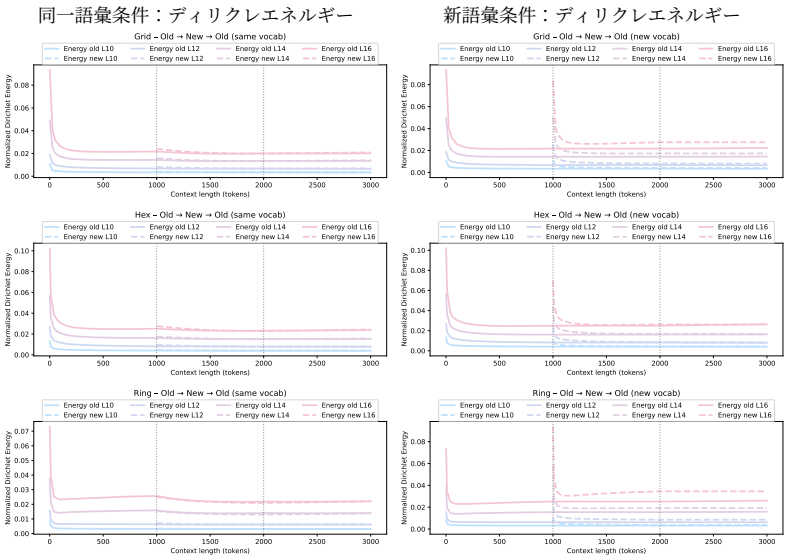


**図 6**　遷移規則の切替（旧 → 新 → 旧）におけるディリクレエネルギーの時間変化．上から Grid / Hex / Ring に対応する．

# A　詳細な実験結果

本文の議論を補うため，Grid / Hex / Ring の各グラフ族について，同一語彙条件および新語彙条件における規則追従精度とディリクレエネルギーの時系列を示す．