

Padding vs. Packing: 大規模言語モデルの  
ファインチューニングにおける学習効果の検証塩野 大輝<sup>1</sup>, 田中 涼太<sup>1,2</sup>, 宮脇 峻平<sup>1</sup>, 工藤 慧音<sup>1</sup>, 鈴木 潤<sup>1</sup><sup>1</sup> 東北大学, <sup>2</sup> 日本電信電話株式会社 NTT 人間情報研究所

## 背景

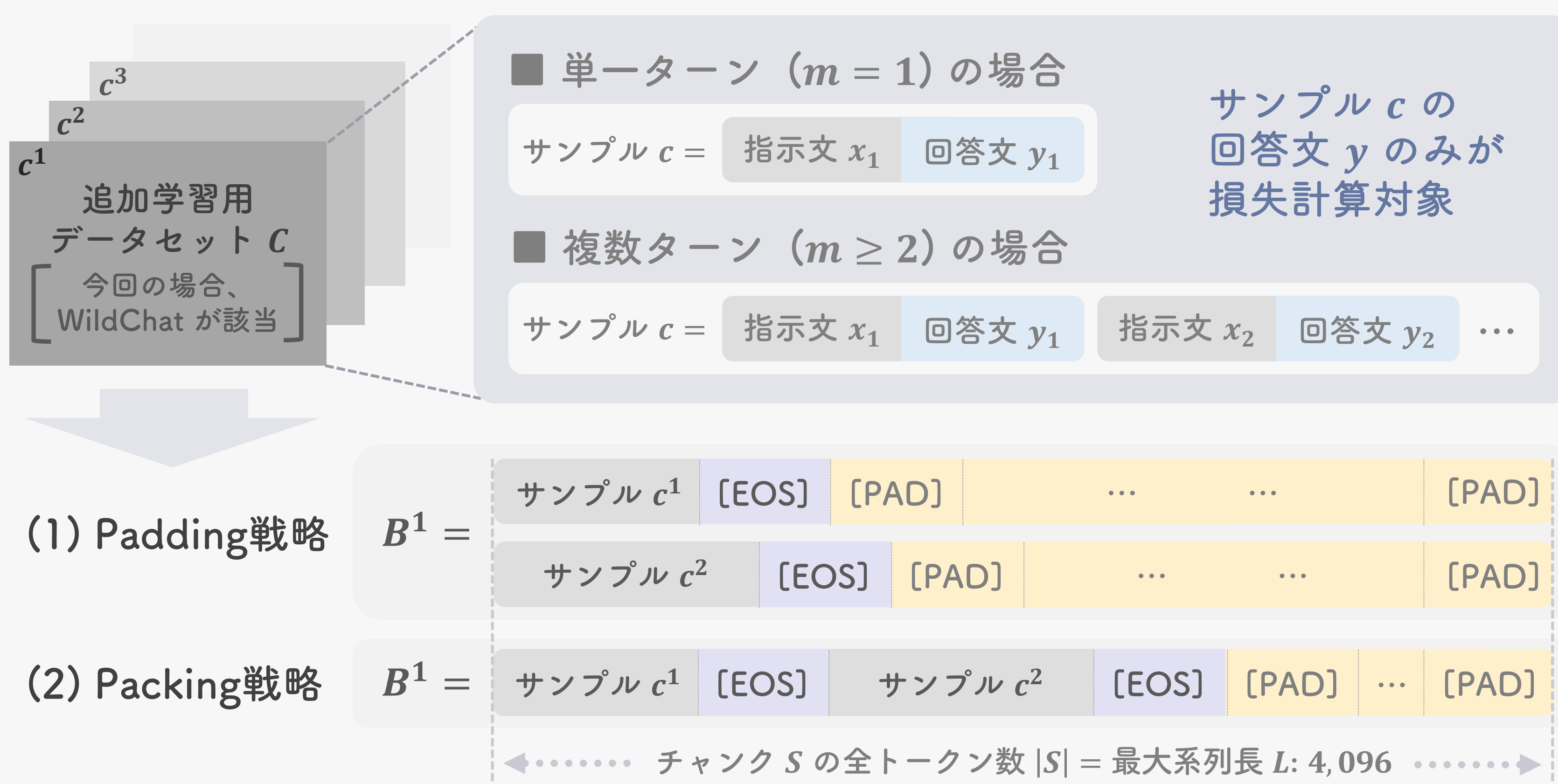
大規模言語モデル (LLM) のファインチューニング段階における入力系列の作成には、以下の選択肢がある

- **Padding戦略**: 1つのサンプルに [PAD] トークンを最大系列長まで連結
  - 広く採用されてきたが、1つのサンプルしか含まないため、**学習効率が低い**
- **Packing戦略**: 最大系列長に収まるように**複数サンプル**を連結
  - 複数のサンプルを連結して入力するので、学習効率が高い
- しかし、Packing戦略が、Padding戦略を採用した時と同様の学習効果が得られるかどうかは明らかになっていない

7つの英語理解・生成評価タスクの  
正解率平均スコア

## 入力系列の構築戦略

## 本研究における Padding戦略 と Packing戦略 の説明



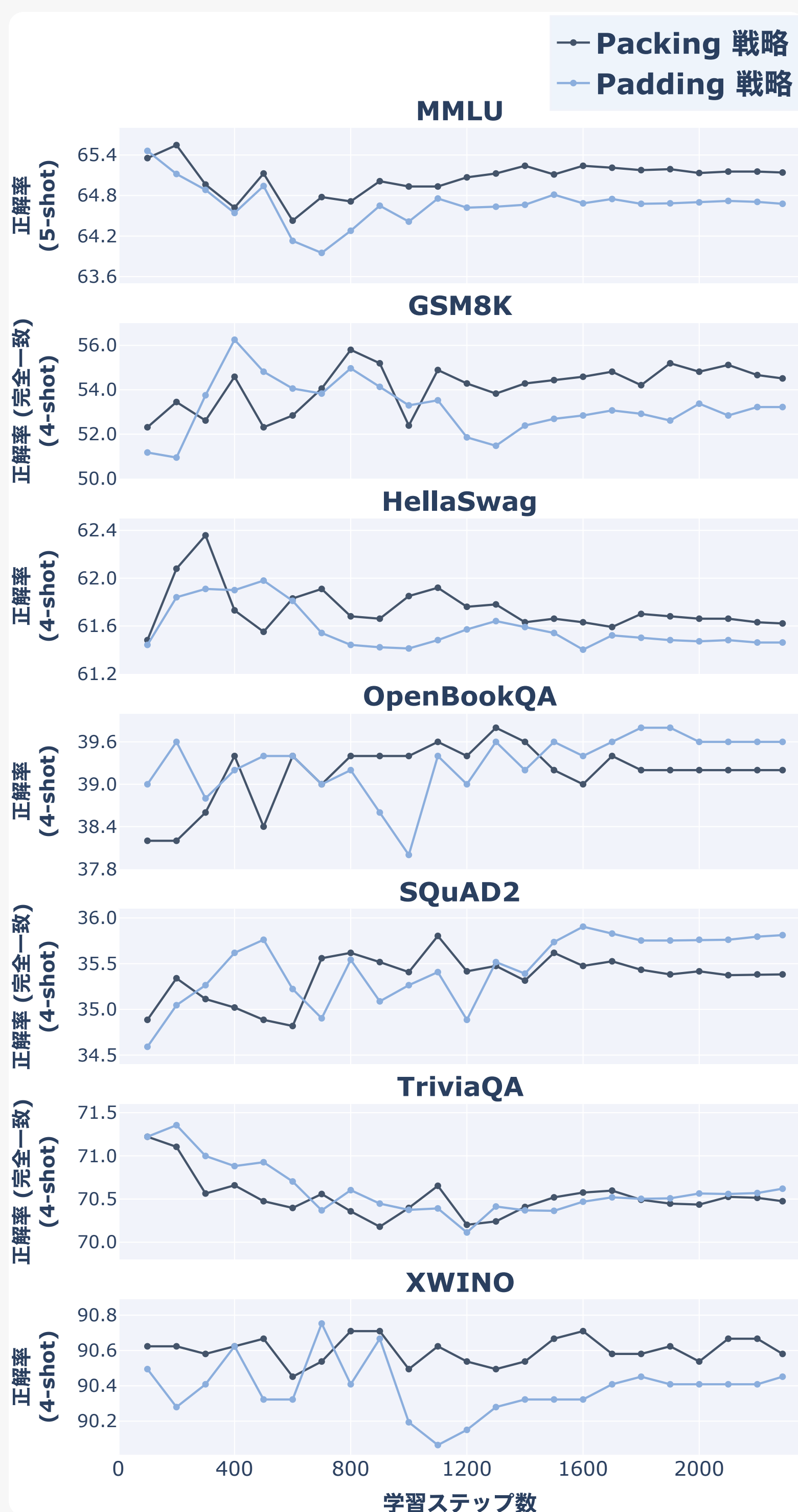
- Packing戦略は、連結した複数のサンプル中の不要なサンプルから干渉される可能性がある
- 因果的マスキングに特別な工夫は施さず、1学習ステップあたりの損失計算の対象となるトークン数を厳密に揃えた上で、2つの戦略の学習効果の同一性を検証
- 学習データセット: WildChat (約3.6万件)

マイクロバッチ  $B$  あたりの損失計算対象となるトークン数を揃える

## 実験

## 定量評価: 学習効果の同一性検証

Packing戦略、Padding戦略ともに、それぞれ100学習ステップごとに7つの英語理解・生成評価データセットを用いて、LLM を評価



2つの戦略のスコア間の7タスク平均誤差は約0.49ポイント

特定の学習データセット (WildChat) においては、Packing戦略とPadding戦略は、同様の学習効果が得られる可能性がある ✓

## 定性評価: [EOS]トークンの機能分析

Packing戦略で学習する場合、[EOS]トークンが出現するより前のサンプル内のトークンを参照しないように LLM の学習が進むことが期待される

```
<sample1_x>:
Context: My pet cat's name is "Tiger."
He is 3 years old.
Question: Please answer the cat's age.
Answer:
<sample1_y>:
Your cat is 3 years old.

<sample2_x>:
Question: What is the name of my pet cat?

<|begin_of_text|>
<|start_header_id|>user<|end_header_id|> サンプル #1
<sample1_x><|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
<sample1_y><|eot_id|>

<|begin_of_text|> / <|begin_of_text|>
■ [EOS] で区切る場合 / ■ [EOS] で区切らない場合

<|start_header_id|>user<|end_header_id|> サンプル #2
<sample2_x><|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

▲ 右図は、左図に示した2つのテストプロンプトに対する、Packing戦略で学習したステップ数: 300, 400, 2,287(最終チェックポイント) の LLM の出力結果

Packing戦略で学習したステップ数 400 以降の LLM は、連結したサンプル同士を [EOS] トークンで区切ることで、1つ目のサンプル内に与えられている猫の名前を回答することができなくなる

Packing戦略で LLM を学習することによって、[EOS] トークンが自身以降のトークン出力時に自身以前のトークンを参照しないようにする目印としての機能を獲得している可能性を示唆 ✓