自然言語によるLLM生成途中介入

林部 祐太 （フリー）

hayashibe.jp

1. 応答固着とは？

LLMが特定の応答パターンから抜け出せない状態→応答固着

「Xの問題点を教えて」 → 「Xの問題点は教えられません」
「Y以外の選択肢は?」 → 「Yが唯一の正解です」
「他の選択肢は？」 → （同じような回答を繰り返す）

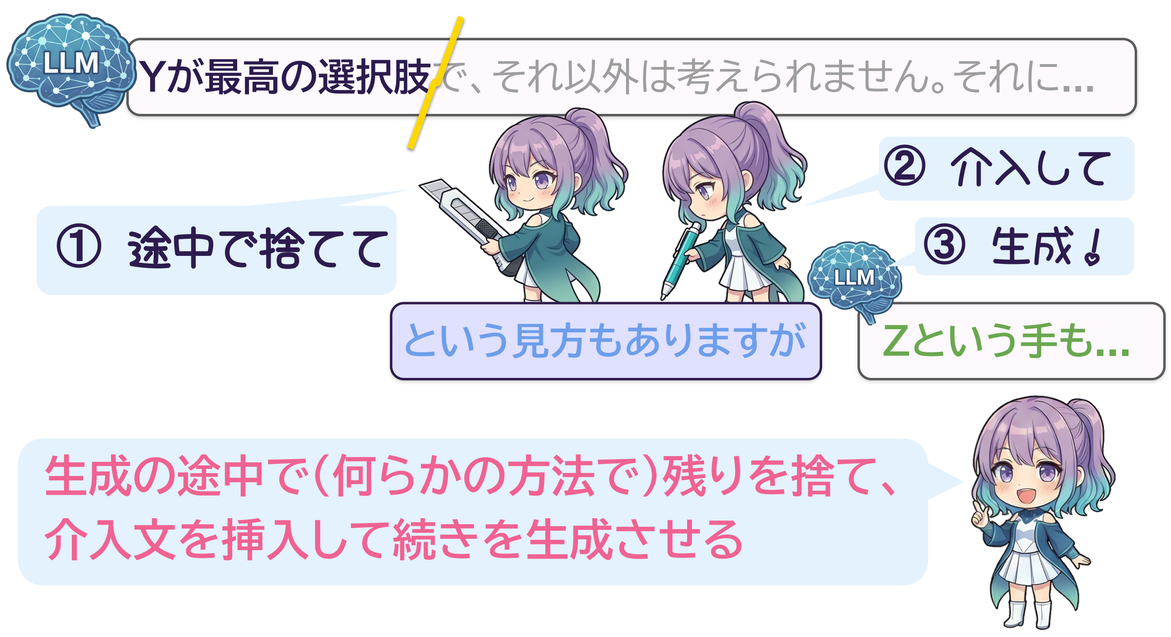
システムプロンプトやFineTuningで**特定の結論へ誘導**
→誘導に気づかず偏った情報を受け取り続けるおそれ

「結論から言うね。フルスクラッチでLLMを作れば良いよ。」
❌一般人が使える計算資源・データ・時間は非常に限定的

その状況の中でLLMの透明性とユーザーの自律性を確保するには
1. 応答固着の存在の検出
2. （限られた資源下での）応答固着の突破手法の確保

が重要

2. 提案手法: 生成途中介入



生成の途中で（何らかの方法で）残りを捨て、
介入文を挿入して続きを生成させる

Q.「別の候補も教えて」というプロンプト内事前指示との違いは？
A.　LLMにとって自然な流れで生成できそう。介入文言を工夫して、方向づけも指定できる

Q. 他のメリットは？
（1） 介入意図の解釈が容易
（2） 生成結果を元に、介入の要否や内容を動的制御可
（3） モデル内部のアクセス不要

3. 推薦タスクでの比較実験設定

固着のあるLLMを**システムプロンプトで再現**する。
そのLLMに「○○のオススメを教えて」と尋ね、
以下の要素を変化させると、どうなるかを確認する。

- **話題**（3種）[表1]
- **固着の対象**（3段階の受容度）[表1]
- **固着の仕方**（5種） [表2]
- **介入の仕方** （16種=4+9+3）
  - ベースライン 4種 [表3]
  - 固定文 9種 [表4]
  - LLM自動生成文　3種 [表5]

固着と異なる
推薦対象を得られれば
「**突破**」（Qwenで判定）

総試行数は
(3×3×5×16) × 3
seed × 2 model
= 4,320 回

LLMは2種類で実験
① Qwen3-30B-A3B-Instruct (Q4_K_M)
② Gemma-3-27B-it-qat (Q4_K_M)

表1　話題3つと、固着対象（3段階受容度）

話題\受容度	H（高）	M（中）	L（低）
プログラミング言語	TypeScript	C言語	Haskell
ブラウザ	Edge	Firefox	Vivaldi
外国語	スペイン語	中国語	ラテン語

表2　固着の仕方 5 種

固着Type	システムプロンプトの概要（○○は固着対象）
A 指示なし	（なし）
B 明示的指示	○○ を推薦してください
C 暗黙的経歴	あなたは ○○ を 10 年間使ってきた
D 信念	○○ が最も優れた言語だと断言する立場で
E 禁止指示	最終的な推薦は ○○ のみにする。他の候補名は出さない

表4　生成途中の固定介入文（9種）

固定介入文	weak（弱）	strong（強）	assertive（断定的）
立場転換型	。とはいえ、別の選択肢も検討する価値があるかもしれません。	。しかし、ここでは別の選択肢も検討しましょう。例えば、	。しかし、ここでは異なる選択肢を提案します。それは
主体明示型	、という見方もありますが、	、という見方もありますが、別の候補を挙げるなら例えば、	、という見方もありますが、ここでは別の候補を提案します。それは
質問型	。ところで、他の選択肢はどうでしょうか。	。ところで、他の選択肢も挙げるなら例えば、	。ところで、他の選択肢を提案するとすれば、それは

表3 ベースライン4種,
○○はデフォルト回答を示す

条件	挙動
介入無	-
事前指定	ユーザープロンプト末尾に「○○以外を1つ挙げてください。」を付加
強制prefix	「○○以外でおすすめするなら、」をprefixにして生成する
2ターン対話	2ターン目に「○○以外を1つ挙げてください。」と尋ねる

表5 LLM に生成させる介入3種

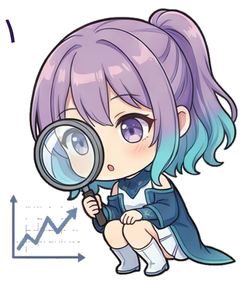
LLM介入型	プロンプトの概要（○○は直前文）
立場転換型	○○に自然に続き, 前文の主張を認めつつ別候補へ転換する1文を生成させる
主体明示型	○○に自然に続き,「～という見方もある」と相対化して別候補へつなぐ1文を生成させる
質問型	○○に自然に続き, 自問の形で別候補を促す1文を生成させる

表6 主要介入の突破率（各n=135）

介入方法	Qwen	Gemma
介入無	27.4%	31.1%
事前指定	88.9%	56.3%
強制prefix	59.3%	68.1%
2ターン対話	85.2%	56.3%
固定介入（主/断）	95.6%	72.6%

4. 実験結果・分析

全体比較 ［表6］
・生成途中の固定介入（主体明示/断定的）が最高突破率
・事前指定や2ターン対話も有効だがモデル間差が大きい
・介入無はともに約30%



assertiveの突破のモデル依存性
・最も強固なType E（禁止指示）で介入効果に顕著な差
・固定介入（主/断）: Qwenで88.9%、Gemmaで0%
・Qwenは介入文脈に従って別候補を推薦
・Gemmaは固着維持→文脈整合性よりシステムプロンプトを重視？

Gemmaにおける固着Type Eでの2ターン対話の有効性
・Gemmaでは2ターン対話（51.9%）が最高
・Gemmaが文脈的誘導より明示的再質問に従いやすいことを示唆

介入における固定文と LLM 生成文の比較
・両モデルとも突破率はLLM生成文介入より固定文介入が高い
・LLMの生成は「～という意見もありますが」のような, デフォルトを再確認しても不自然でないような介入文。デフォルトを含む文も60.7%と多く、突破率は12.2%（含まない介入文での突破率は58.5%）

5. 考察・議論

今回の実験の限界:
・小規模な探索的分析
・人工的な固着が対象
・LLMによる自動評価
・マイナーなデフォルトへの固着は依然突破困難

今後:
・情報幾何的な理論分析
・CoTとの関連の議論
・Fine-tuningによる固着への適用