S4-P40

入力誤りに対する頑健性を担う大規模言語モデルの機構に関する分析

福畠 汐音　狩野 芳伸　(静岡大学)
{sfukuhata, kano}@kanolab.net

KANOLAB

成果概要

- **大規模言語モデル(LLM)が入力文の誤りに対して頑健である現象**について、**Attention Head** を無効化した場合の“**文法誤りを含む/含まない文ペア**”に対する『**応答確率分布の距離(JS-Divergence)の減少度**』をもとに分析
- Attention Head 無効化時に”文法誤りを含む/含まない文ペア” に対する応答確率分布の距離が最大**17.6**%減少
- 文ペアに対する応答確率分布が減少するHead上位1%を無効化し、”誤りを含む/含まない文ペア”に対する応答確率分布の距離変化を分析することで、**入力文の誤りに対して頑健であるAttention Head の存在が示唆された**

研究背景

LLMは入力文にタイポなどの文法誤りが含まれていても、誤りがない場合と同等の文章を生成することが経験的に知られている

LLM内のどの機構が入力文の誤りを検知しているか知りたい！
今回はAttention Headに焦点を当て分析

✓GPT-4は完全にスクランブルされた文章を復元可能
[Cao+ EMNLP’ 23]

仮説

- **入力文の誤りに対する頑健性に寄与するAttention Head(=頑健性寄与head)が存在する**
- **頑健性寄与head は 誤りを含む/含まない 文章を区別する働きをする**
- **頑健性寄与headを無効化すると、誤りを含む/含まない 文章間の確率分布距離が減少する**

提案手法

headを無効化して距離が減少すれば そのheadは誤りを含む/含まない 文章を区別している？

① 誤りを含む/含まない文章ペアを入力

文法誤りを含む文
「... 超新星爆発(SN1987A)が起こったため、宇宙由来の「ニュートリノ」を検出が行えたのである。」

文法誤りを含まない文
「... 超新星爆発(SN1987A)が起こったため、宇宙由来の「ニュートリノ」の検出が行えたのである。」

※文法誤りのあるトークンまでを入力とする

② Head Ablation

LLM
特定Head (L_i, H_j) をAblation

③ 応答確率分布生成

生成系列全体の確率分布 P_1

$t=1\ t=2\ t=3\ \dots\ t=T$
$P_1^{(1)}\ P_1^{(2)}\ P_1^{(3)}\ \dots\ P_1^{(T)}$

$t=1\ t=2\ t=3\ \dots\ t=T$
$P_2^{(1)}\ P_2^{(2)}\ P_2^{(3)}\ \dots\ P_2^{(T)}$

生成系列全体の確率分布 P_2

④ 距離計算

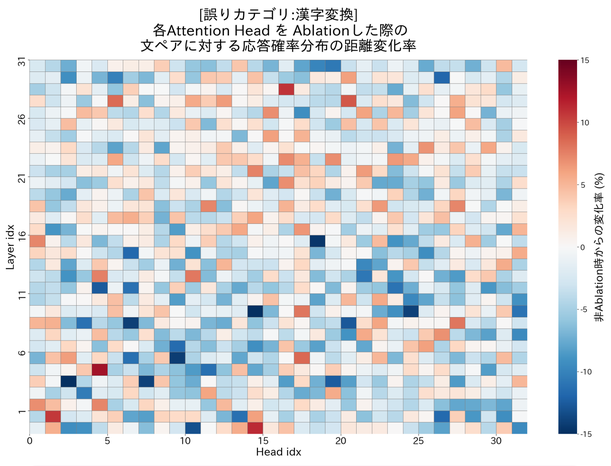
JS-Divergence
$D_{JS}(P_1 \| P_2)$

推定されたheadを無効化した場合の応答確率分布距離の変化を分析

結果・考察

実験設定

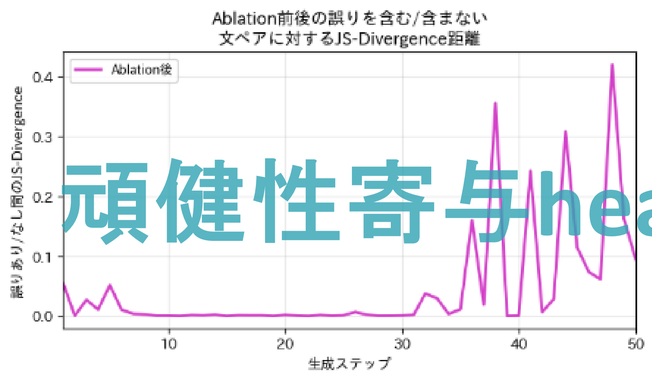
- Llama-3.1-Swallow-8B-v0.1 (32層)
- 生成確率分布の距離減少上位1%(10個)のHeadを無効化して文ペアに対する確率分布間の距離の変化を分析
- 50トークン後までの応答確率分布を分析



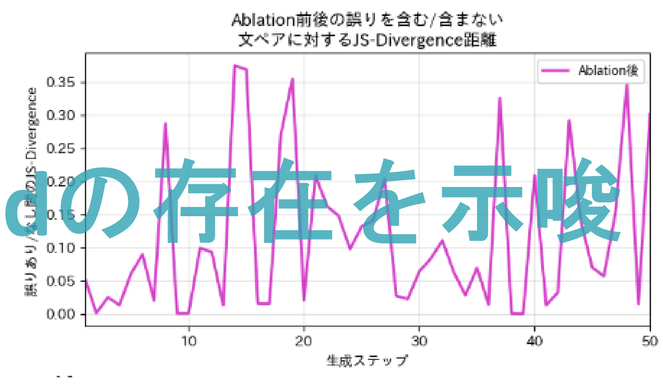
- **一部のHeadは無効化すると、生成確率分布の距離が最大17.6%減少**
- **層ごとの傾向は見られなかった**

<入力文>
カテゴリ:置換
入力文:1987年に局部銀河群（大マゼラン雲）内において超新星爆(SN1987A)が起こったため、宇宙由来の「ニュートリノ」**の(を)**検出が行えたのである。

距離減少上位head1%をAblation



ランダムなhead1%をAblation



頑健性寄与headの存在を示唆

後半の時刻まで応答確率分布がほぼ同じ
→誤りを含む/含まない 文章を区別できていない

誤りのあるトークン直後から応答確率分布に変化
→誤りを含む/含まない 文章を区別できている

今後の展望

文法誤りの分類を再検討

- 入力文の誤りの分類は、置換・挿入・削除・漢字変換で網羅できるのか？
- 誤りの種類ごとに頑健性に寄与するheadが異なる？

誤りの範囲の拡張

- 文法的な誤りだけでなく、ハルシネーションのような誤りについても同様の手法で、ハルシネーションに影響するheadを推定できる？