Quantifying Semantic Emergence in Language Models

Hang Chen[1] **Xinyu Yang**[1] **Jiaying Zhu**[2] **Wenya Wang**[3]
[1]Xi'an Jiaotong University
[2]The Chinese University of Hong Kong
[3]Nanyang Technological University

ACL2025 (long)

最先端NLP勉強会2025

発表者：辻村有輝 (産総研)

概要

- 目的：LLMのコンテキスト理解力の定量的評価
- **Information Emergence (IE)** という評価指標を提唱
  - LLMが入力トークン列から意味情報を抽出する能力を測定するための指標

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$

マクロな表現ベクトル（入力トークン列全体から計算）の情報量と
ミクロな表現ベクトル（単トークンから計算）の情報量の**ギャップ**で定義
➡ 入力全体から情報を集めているほど大きくなるようなスコア

- IEの値からいくつかの示唆を得られる
  - In-context learningでは新しい事例の出現時に大きな情報量の伸びを観測
  - 人間が書いた文とLLMの生成文で異なるIEの傾向

背景

- LLMの意味理解能力は様々なタスクで検証されている
  - 指示追従性 (Zeng et al., 2023)
  - 検索能力 (Sun et al., 2023)
  - 推論能力 (Yang et al., 2024) など

- 個別タスクを用いた検証は似たような能力を評価しているのにもかかわらず評価指標も得られる知見もtask specificで扱いにくい

➡ LLMの意味理解能力を評価する指標としてもっと扱いやすいものを作りたい

背景

意味とは？

- トークンが組織化 (organized) されることによって自然に出現するもの

アイディア

- トークンから意味を抽出する能力において、他のモデルと比較して優れたモデルは単一トークンよりもグローバルな系列に対してより高いエントロピー減少をもたらすはず

➡ 単語のみから計算した際のミクロな表現ベクトルが持つエントロピーと，系列全体から計算したマクロな表現ベクトルが持つエントロピーの間のギャップを計算し，その大小でモデルの意味理解能力を評価

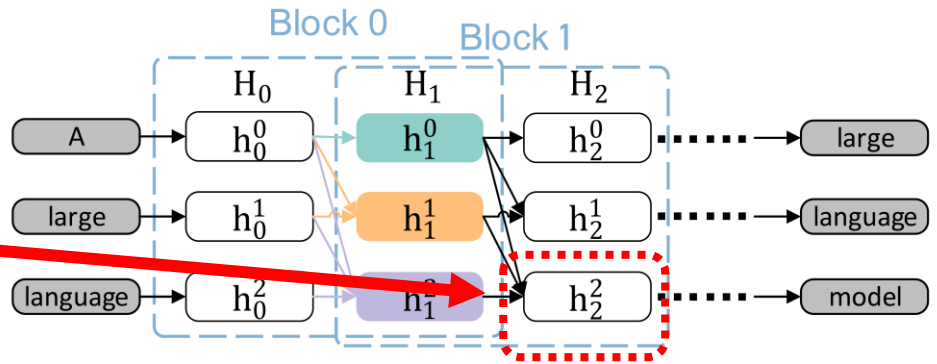
手法

- 提案スコア：マクロ・ミクロな情報のみを持つ表現ベクトルを構成し
  それらの間の情報量の差をスコアとする

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$

  - 大きいほど入力文脈から意味をうまく取り出すことを示す
  - 層とタイムステップごとに計算される値

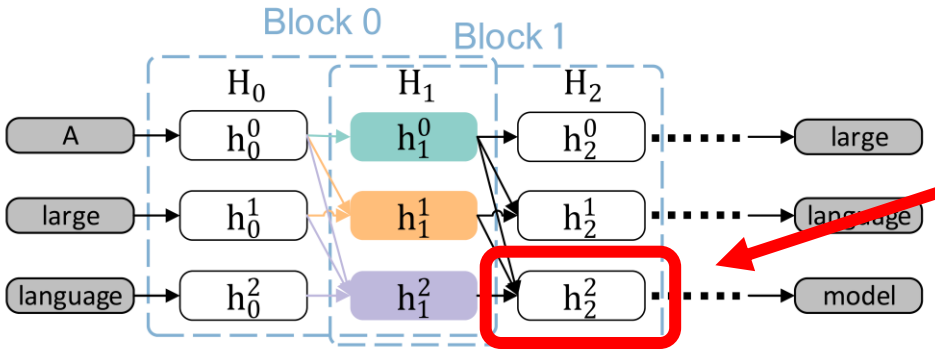
例: $T = 2, l = 1$



- 計測にはマクロ・ミクロな表現ベクトルをそれぞれ構成する

手法

- マクロな表現ベクトルの構成は素直にLLMに入力して構成



例えばここは $T = 2, l = 2$ の
マクロな表現ベクトル

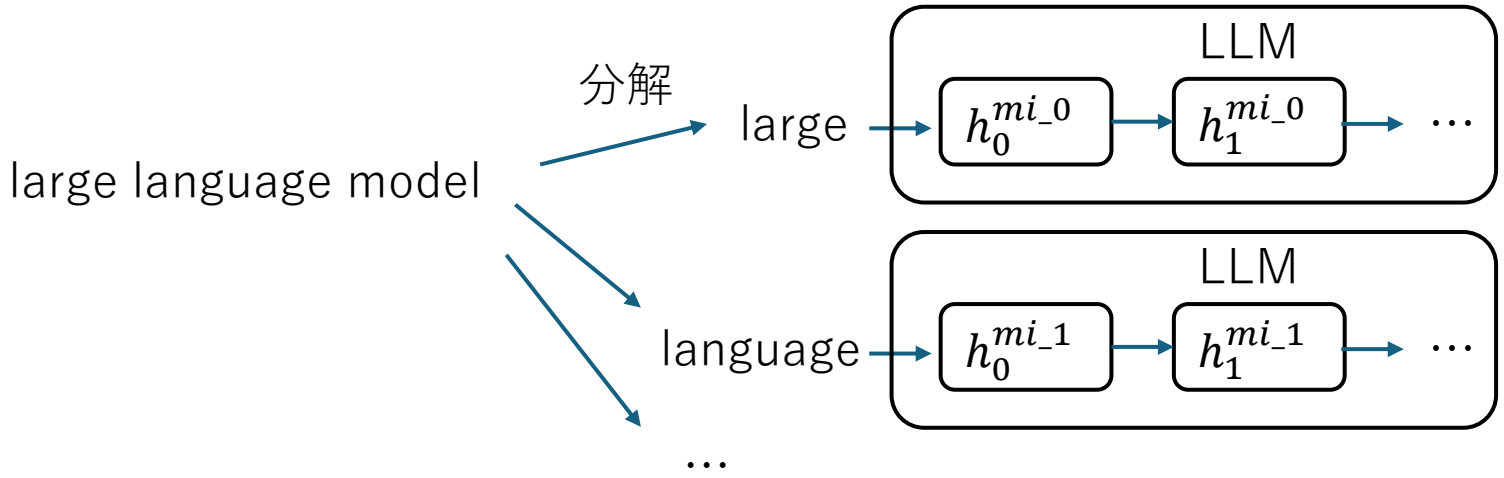
⬆

文脈全体を考慮したベクトル

- ベクトルが持つ情報量はMINEを用いて計測する
  - MINE: Mutual Information Neural Estimation (Belghazi et al., 2018)
    - ニューラルネット $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ の学習によって確率変数XとZ間の相互情報量を推定
    - GANのdiscriminatorのようなイメージ

手法

- ミクロレベルの計測
  - どうやってミクロレベルの情報しか持たない表現ベクトルを作るか
    ➡ **単トークンのみの入力**で表現ベクトルを作る



- ミクロレベルの情報量は全位置バラバラに計測したものの平均

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$

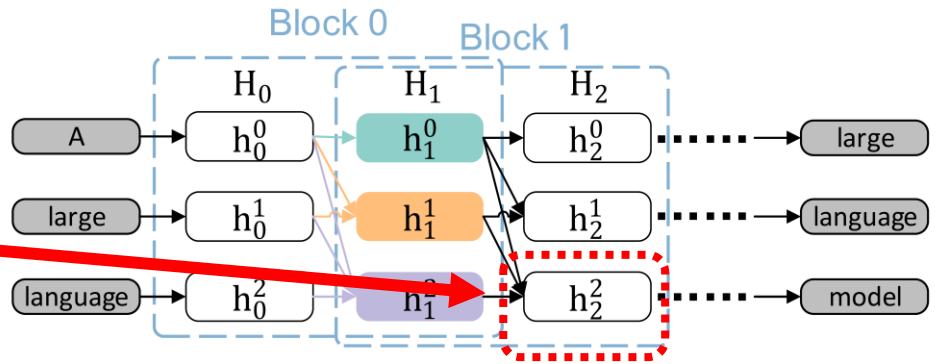
手法（再掲）

- 提案スコア：マクロ・ミクロな情報のみを持つ表現ベクトルを構成し
  それらの間の情報量の差をスコアとする

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$

  - 大きいほど入力文脈から意味をうまく取り出すことを示す
  - 層とタイムステップごとに計算される値

例: $T = 2, l = 1$



- 計測にはマクロ・ミクロな表現ベクトルをそれぞれ構成する

提案スコアの計測に使う入力データの構築

データが満たしておいてほしい性質➡同じタイムステップで似た内容が出現
- 提案スコア $E(l,T)$ がタイムステップと層ごとに計算されるため

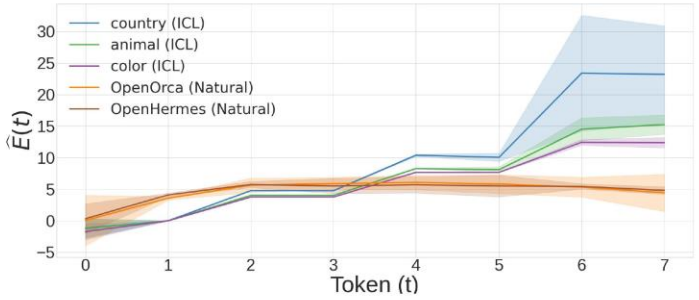
In-context learning (ICL) を実施する合成データセットを作成
- いずれもentity＋コンマを1-shotとする単純な事例トークンの羅列

- **Country**:
  - 1トークンにtokenizeされる国名25種の4-shot組み合わせ
  - 例：***“France, Mexico, Egypt, Russia,”***
- **Animal**: 16種，5-shot
- **Color**: 15種，5-shot

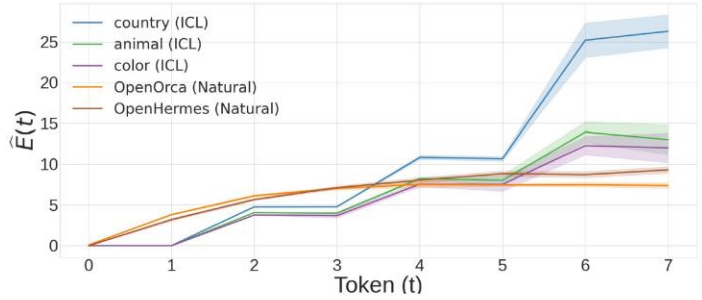
自然なテキストとしてOpenOrcaとOpenHermesも実験で利用

実験結果

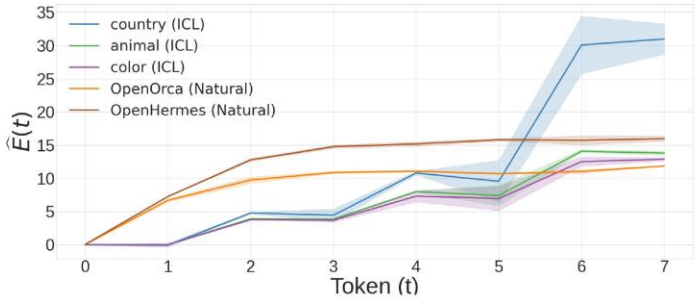
- 提案スコアはl, tの関数だが，実際は各層でほぼ同じスコアになる
  ➡ 全層で平均を取り，t ごとの評価にフォーカス



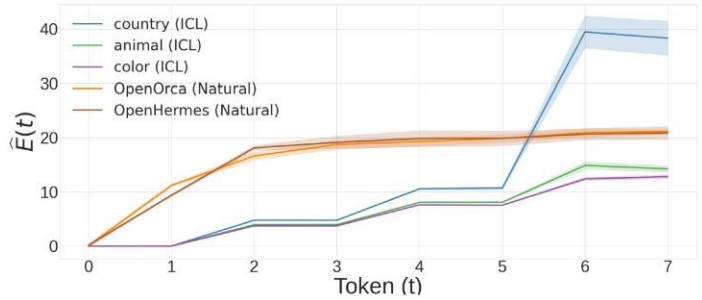
(a) GPT2-large (812M)



(b) GPT2-XL (1.6B)



(c) Gemma (2.51B)



(d) OpenLlama (3B)

- ICLデータでは新しい事例の出現のタイミングで階段状にスコアが上昇
  - スコアの上昇＝マクロな表現ベクトルが持つ情報量が増加
  - 無限に上がるわけではない

Model	categories	shot3	shot4	shot5	shot6	shot7
GPT2-large	country	↑5.67	↑12.95	↑2.75	↑0.33	↑1.04
	animal	↑4.24	↑6.06	↑9.52	↑0.39	↓1.44
	color	↑4.88	↑4.82	↑6.39	↑1.24	↑0.22
GPT2-XL	country	↑5.86	↑13.12	↑3.56	↑1.75	↑0.32
	animal	↑4.21	↑5.75	↑8.21	↑1.15	↑0.61
	color	↑3.82	↑4.61	↑7.06	↑1.74	↑0.54
Gemma	country	↑6.33	↑22.16	↓2.86	↑3.21	↓3.54
	animal	↑4.09	↑6.24	↑8.45	↑36.51	↓2.14
	color	↑4.65	↑5.16	↑7.81	↑16.49	↑1.21
OpenLlama	country	↑6.33	↑45.26	↑7.54	↑4.65	↓3.15
	animal	↑4.95	↑7.54	↑35.16	↑2.16	↑3.26
	color	↑4.39	↑5.27	↑27.56	↑11.42	↑2.51

Table 4: $\Delta\widehat{E}(t)$ compared to the previous token. The red represents $\widehat{E}(t)$ decreases compared to the previous token.

- 自然文は早い段階でスコアが安定

実験結果

一般的なICLデータでも階段状のスコアの変動は観測される

- 各shot内の挙動は自然文と同じ挙動，shot切り替わりは合成データの挙動

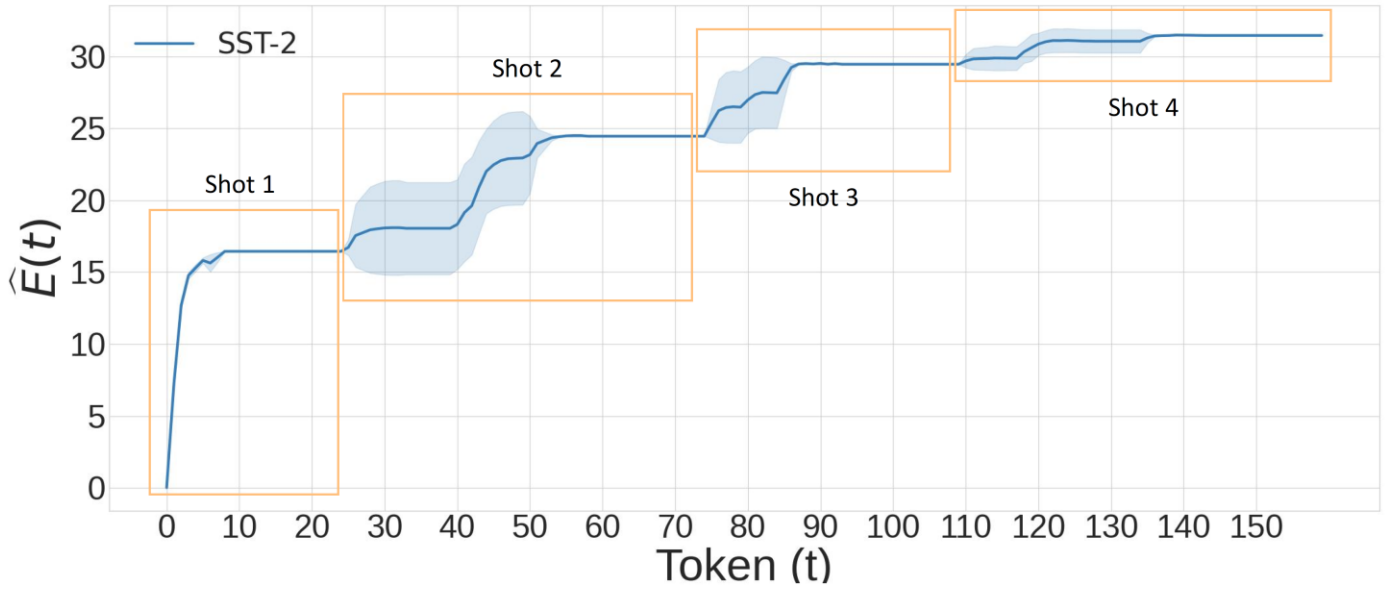


Figure 7: $\widehat{E}(t)$ with inputs from SST-2 datasets, consisting of 3 shots.

プロンプト：

the SST-2 dataset (Socher et al., 2013): "*[review 1], The emotion of this sentence is [label1], [review 2], The emotion of this sentence is [label2], [review 3], The emotion of this sentence is [label3], [review 4], The emotion of this sentence is*". Here, *[review]* and *[label]* are random text samples and corresponding sentiment labels (positive or negative) from the SST-2 dataset, respectively.

実験結果｜表現ベクトルが持つ情報量

- 表現ベクトルが持つ相互情報量は各層でほぼ同一の値になる
  - 層をまたぐときにほとんどの情報をシェアしている

layer	token0	token1	token2	token3	token4	token5	token6	token7
1	2.83	2.83	6.89	6.50	10.68	9.24	14.16	11.53
2	2.83	2.83	6.90	6.91	11.08	11.10	16.70	16.79
3	2.83	2.83	6.89	6.88	11.17	11.17	16.93	16.00
4	2.83	2.83	6.89	6.88	11.08	11.06	16.74	16.58
5	2.83	2.83	6.89	6.89	11.13	11.11	16.94	15.88
6	2.83	2.83	6.88	6.89	11.16	11.16	18.89	17.11
7	2.84	2.83	6.89	6.88	11.15	11.14	16.97	17.08
8	2.83	2.83	6.88	6.88	11.12	11.19	16.86	17.42
9	2.83	2.83	6.90	6.88	11.20	11.17	16.92	15.97
10	2.83	2.83	6.88	6.88	11.08	11.14	16.97	16.51
11	2.83	2.83	6.88	6.88	11.04	11.05	17.05	16.19

Table 5: Mutual information of GPT2-XL in Animal category. Red represents the highest value in this block.

layer	token0	token1	token2	token3	token4	token5	token6	token7
1	8.40	13.71	16.01	17.33	17.21	17.67	17.95	17.29
2	8.36	13.77	15.76	17.06	17.08	17.72	17.68	18.00
3	8.44	13.75	16.09	17.10	17.82	17.69	17.84	18.04
4	8.44	14.20	16.07	17.29	17.45	17.74	18.51	17.81
5	8.39	13.50	16.32	16.83	17.82	17.98	18.26	18.14
6	8.41	13.69	16.03	16.99	17.58	17.82	17.52	18.33
7	8.41	13.68	16.06	17.00	18.32	17.72	17.69	18.19
8	8.40	13.80	15.97	17.26	17.61	17.73	17.52	18.44
9	8.35	13.69	15.95	17.17	17.21	17.54	17.47	18.03
10	8.41	13.57	16.30	16.57	17.46	17.85	17.85	17.51
11	8.34	13.37	16.02	15.93	17.30	17.24	17.27	16.91

Table 6: Mutual information of GPT2-XL in OpenOrca dataset. Red represents the highest value in this block.

- [所感] 後ろのトークンでは出力層に近づくと多少情報量が減る
  ➡出力に不要な情報が落とされている？

実験結果｜提案スコアの標準偏差からの幻覚の判定

IEのstddevが大きくなるタイミングとAccuracyが崩壊するタイミングが一致

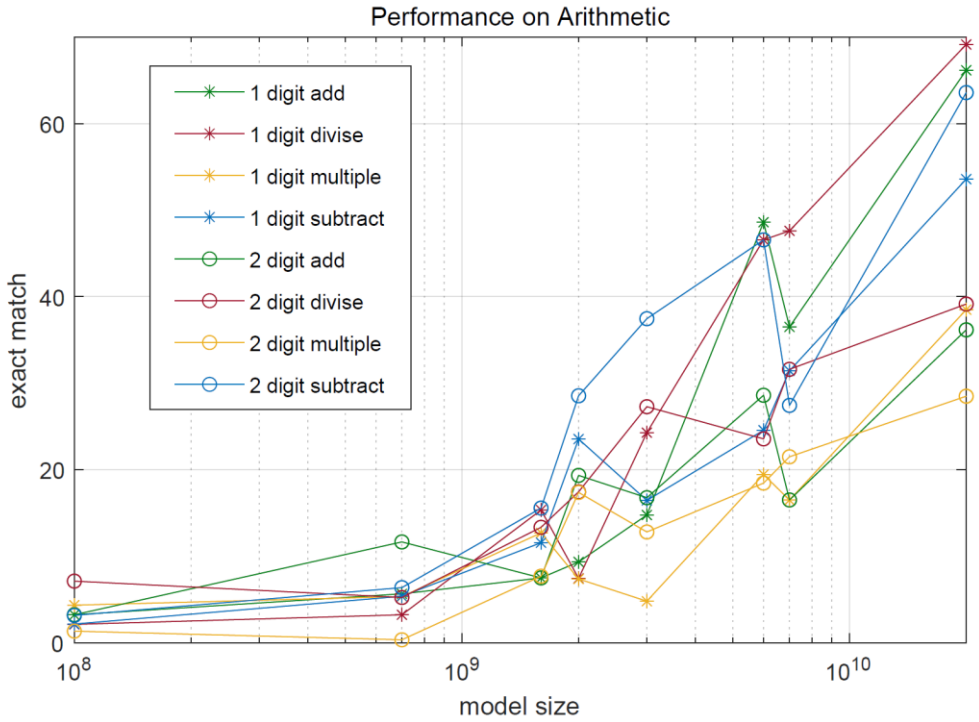
➡ある種のhallucinationの予兆を与える？

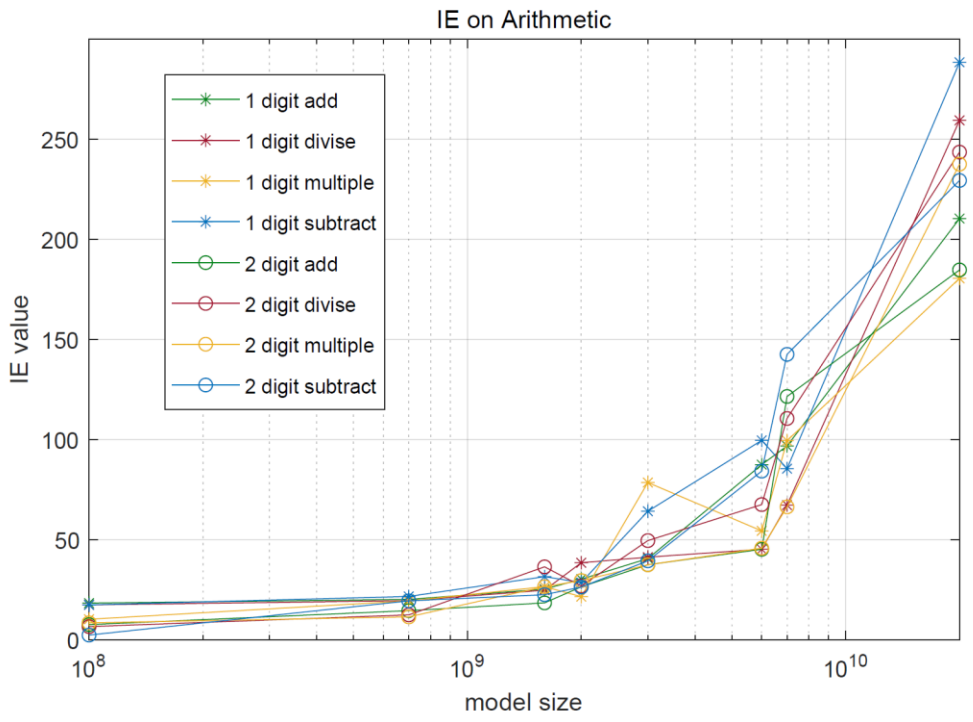
- モデル：GPT2-XL

IE value by each shot							
Statistics	shot1	shot2	shot3	shot4	shot5	shot6	shot7
value	4.013	8.34	12.95	26.81	61.59	82.49	71.52
SD	<0.01	0.59	0.84	2.61	6.59	7.22	7.05
Accuracy of LLMs outputs given shots (%)							
dataset	shot1	shot2	shot3	shot4	shot5	shot6	shot7
country	0	54.15	74.29	88.47	46.21	21.59	22.68
animal	0	44.51	69.43	76.19	64.19	36.14	33.54
color	0	37.49	66.51	72.18	73.16	46.95	38.49

実験結果｜EmergenceとIEの関係

- Emergenceの発生タイミングとIEの傾向が一致する
  - 主張：具体的なタスクに依存せずにemergenceを観測できる





合計8タスク

タスク例： **1 digit addition:**

"What is 1 plus 0? A: 1, What is 4 plus 4? A: 8, What is 2 plus 7? A:"

実験結果｜人間の書いた文と生成文のIEの傾向の違い

- AIのほうが大きなIEになるパターンが観測された
  - （HumanデータにはOpenHermesのテキストを利用）
  - ただし計測には複数のサンプルテキストが必要
    - 現実的なAI判定の利用には難しそう

Text+Estimator	token0	token1	token2	token3	token4	token5	token6	token7	token8
Human+GPT2-XL	10.9	16.9	18.6	19.5	19.5	19.7	19.6	19.5	19.4
Human+GEMMA	9.5	16.8	22.4	24.3	24.0	25.3	24.6	25.0	25.9
GPT4+GPT2-XL	11.3	18.8	23.5	27.2	34.5	37.2	39.2	39.5	39.2
GPT4+GEMMA	12.1	20.5	25.1	31.6	36.3	39.9	40.4	39.5	40.6
Claude3-opus+GPT2-XL	12.6	21.8	26.6	29.5	36.8	39.8	42.6	45.2	45.3
Claude3-sonnet+GPT2-XL	11.4	17.4	24.8	28.5	32.5	36.5	36.1	36.2	36.2
Llama3 (70B)+GPT2-XL	11.2	18.1	23.6	24.5	28.5	32.6	36.5	36.8	36.6

Table 2: IE in texts generated from human and popular LLMs. “text” refers to the party that generates the text. “Estimator” refers to the LM used to transform the text into representations and estimates the IE value using f described in Section 3.2. Due to computation constraints, we only GPT2-XL and GENNA as estimators.

まとめ

- LLMのコンテキスト理解力の定量的評価
- **Information Emergence (IE)** を提唱
  - LLMが入力トークン列から意味情報を抽出する能力を測定するための指標
  - ミクロな表現とマクロな表現を作り，それらの相互情報量のギャップで計測

- 所感
  - 先頭数トークンだけ？
    - 相互情報量の計測にそれなりの計算コスト (3090で1か所40分)
  - モチベーション的には本当に知りたいのは相互情報量ではなかったのでは？
    - ミクロなベクトルで条件付けされたマクロなベクトルの持つエントロピー？
    - 結果的に近いものにはなっているように思える

[所感] 相互情報量でいいのか？

- 提案スコア
  - l層目と$l+1$層目の表現ベクトルの相互情報量から計算

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$

  - 主張されているモチベーション：
    - LLMがどの程度入力から意味情報を抽出する能力を持つかを評価したい
  - 主張されているアイディア
    - 層をまたいだ時にマクロなベクトルがミクロなベクトルに比べて大きい**情報量**を持つとき ➡ 次トークン予測におけるuncertaintyを大きく減らす
      ➡意味情報を多くとらえているはず

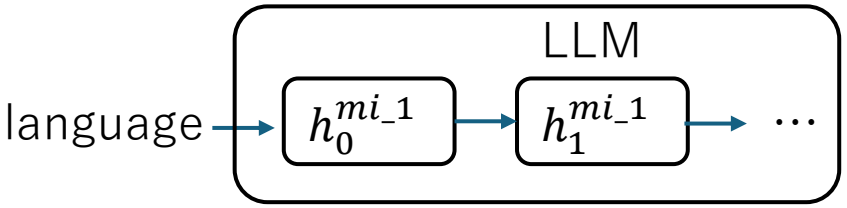
↑

これは層間の表現ベクトルの相互情報量なのか…？

[所感] 相互情報量でいいのか？

- ミクロレベル

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T}$$



- そもそも出力は入力から決定的に計算される

➡ $MI(h_{l+1}^{mi\_t}, h_l^{mi\_t}) = H(h_{l+1}^{mi\_t})$　結局出力の情報量を計算しているだけ

[所感] 相互情報量でいいのか？

- ミクロレベルの相互情報量
  $= H(h_{l+1}^{mi\_t})$　　結局出力の情報量を計算しているだけ

  - 実験結果的には全層でほぼ同じ情報量をキープし続けている
    - （このテーブルはもともとマクロ側の相互情報量なので他のタイムステップでは傾向が違う可能性はある）

layer	token0	token1	token2	token3	token4	token5	token6	token7
1	2.83	2.83	6.89	6.50	10.68	9.24	14.16	11.53
2	2.83	2.83	6.90	6.91	11.08	11.10	16.70	16.79
3	2.83	2.83	6.89	6.88	11.17	11.17	16.93	16.00
4	2.83	2.83	6.89	6.88	11.08	11.06	16.74	16.58
5	2.83	2.83	6.89	6.89	11.13	11.11	16.94	15.88
6	2.83	2.83	6.88	6.89	11.16	11.16	18.89	17.11
7	2.84	2.83	6.89	6.88	11.15	11.14	16.97	17.08
8	2.83	2.83	6.88	6.88	11.12	11.19	16.86	17.42
9	2.83	2.83	6.90	6.88	11.20	11.17	16.92	15.97
10	2.83	2.83	6.88	6.88	11.08	11.14	16.97	16.51
11	2.83	2.83	6.88	6.88	11.04	11.05	17.05	16.19

Table 5: Mutual information of GPT2-XL in Animal category. Red represents the highest value in this block.

[所感] 相互情報量でいいのか？

マクロレベルの項はどう解釈するべきか？

$$\mathrm{IE} = E(l,T) = MI(h_{l+1}^{ma}, h_l^{ma}) - \frac{\sum_{t=0}^{T-1} MI(h_{l+1}^{mi\_t}, h_l^{mi\_t})}{T} \quad (3)$$

- 少なくとも入出力の情報量の下界にはなる
- 各項をミクロ・マクロな情報量とみなすのであれば，結果的にこれはマクロ・ミクロな情報量の差として解釈でき，欲しいスコアに近いものが観測されることになりそう

他の気になったところ

- もし表現ベクトルがnext tokenの分布情報しか持たない場合は，マクロレベルの方がむしろエントロピーは小さくなるのでは？
  - 実際はIEが大きな正の値になるように，マクロが持つ情報量はミクロより大きい
    ➡next tokenの分布情報以外も保持している（当たり前ではある）