大規模言語モデルの潜在言語は一貫しているべきか？

尾崎 慎太郎 ♡,♠　平岡 達也 ◇,♡　大竹 啓永 ♡,♠　大内 啓樹 ♡,★
磯沼 大 ♠,∞,♣,★　Benjamin Heinzerling★,∞　乾 健太郎 ◇,★,∞
渡辺 太郎 ♡　宮尾 祐介 ♣,♠　大関 洋平 ♣,♠　高木 優 ∂
♡ 奈良先端科学技術大学院大学　♠ 国立情報学研究所 大規模言語モデル研究開発センター
◇MBZUAI　♣ 東京大学　∞ 東北大学　★ 理化学研究所　∂ 名古屋工業大学
ozaki.shintaro.ou6@naist.ac.jp　takagi.yu@nitech.ac.jp

概要

大規模言語モデルは一般に複数の層で構成されており，最終層の潜在表現をデコードすることで文章を生成する．本研究では，中間層の潜在表現をデコードしたときに生成される言語を潜在言語と呼び，潜在言語の一貫性が下流タスクの性能に与える影響を検証する．具体的には，地理文化や翻訳といった潜在言語による影響を受けやすい下流タスクから構成されるデータセットを構築し，潜在言語の一貫性と下流タスクの性能の相関を分析した．複数のモデルに対する実験の結果，潜在言語の一貫性は，必ずしも下流タスクの性能に必要ではないことが示された．また，最終層付近において内部表現を目的言語に適応させており，言語モデルの思考は潜在言語に依存しないことが示唆された．

1　はじめに

大規模言語モデル (LLMs) [1, 2, 3] は，一般に複数の Transformer 層で構成され，最終層の潜在表現をデコードすることで文章を生成する．近年中間層の潜在表現をデコードすることで，モデル内部における文章生成過程の解釈が試みられており（Logit Lens）[4]，特に中間層の潜在表現をデコードした時に出力される言語（潜在言語）に着目した研究が盛んである [5, 6, 7, 8]．例えば，英語中心のコーパスで学習された Llama2 [9] は，中仏翻訳タスクにおいても，英語を潜在言語として用いていることが報告されている [5]．同様に，日本語および英語の両方で学習された LLM-jp [10] においては，推論過程で日本語を潜在言語として用いる傾向が確認されている [6]．これらの結果は，モデルが使用する潜在言語が，事前学習時における各言語のデータ割合に大きく依存することを示唆している [7, 10]．

一方で，潜在言語の一貫性と下流タスクにおける性能との関係については，これまで十分に議論されていない．潜在言語が一貫していることが下流タスクの性能に影響するかどうかは未解明であり，この点を明らかにすることは，潜在言語の役割や推論の安定性を理解する上で重要である [11]．

本研究では，「モデルが得意としない潜在言語で思考することは，下流タスクの性能に影響を及ぼす」という仮説を立てる．この仮説を検証するために，中国語，英語，日本語の 3 言語から構成された思考言語を意図的に混在させる敵対的プロンプト [12, 13] を入力に付与し，潜在言語の一貫性を崩壊させることで，その影響を定量的に評価する．具体的には，複数のモデルに対して，敵対的に含まれる言語の種類およびその割合を体系的に変化させる．モデル内部がどの程度一貫した潜在言語で思考しているかを定量化するため，新たに潜在言語一貫性スコア (LLC Score) を定義し，潜在言語の一貫性と下流タスクの性能との相関を分析した．

実験では，先行研究 [5, 6, 14, 15] に倣い，潜在言語の一貫性が性能に大きな影響を与えると考えられる翻訳タスクおよび地理・文化知識に関するデータセットを新たに構築した．複数のモデルに対する実験の結果，いくつかのモデルは推論時に特定の潜在言語へ一貫して依存する傾向を示すことが確認された．しかしながら，そのような一貫性が常に最適な性能につながるわけではなく，翻訳および地理文化の両タスクにおいて，仮説と反する結果が観測された．この結果は，モデルの思考が潜在言語に依存しておらず，最終層付近において思考内容を目的言語に変換することで，適切な回答を出力している可能性を示唆している．

2 関連研究

LLMs の潜在言語は，事前学習データに含まれる言語分布の影響を強く受けることが知られている．Wendler ら [5] は，主に英語データで学習された Llama2 [9] が，内部推論において英語を潜在言語として用いる傾向があることを示した．一方，Zhong ら [6] は，日本語と英語の両方で学習された LLM-jp モデル [10] において，日本語に偏った内部表現が形成されることを報告している．これらの結果は，LLM が内部処理のために特定の潜在言語を獲得することを示唆しており，その選択は学習データに由来する内在的な言語表現 [16, 17] に強く依存することを示している．

また，プロンプトに用いる言語がモデルの出力や性能に大きな影響を与えることも報告されている．Huang ら [18] は，プロンプト言語の違いが生成結果や精度に大きな差異をもたらすことを示した．さらに，Chain-of-Thought [19, 20, 21] を多言語化した手法 [22, 23] においては，推論過程や性能への影響が詳細に分析されている．しかし，これらの研究の多くは出力精度などの表層的な挙動に焦点を当てており，モデル内部で用いられている潜在言語そのものを直接検証するものではない．

モデルの出力性能に関しては，敵対的プロンプトや外部ノイズに対する耐性が広く研究されてきた．Goel ら [24] は，構文的および語彙的な攪乱がモデルの挙動に与える影響を分析し，Khashabi ら [25] は，意味的に無関係な文の挿入が質問応答モデルの信頼性を著しく低下させることを示した．しかしながら，これらの研究では，ノイズの付与によってモデル内部の推論過程や潜在言語がどのように変化するかについては，十分に議論されていない．

3 分析手法

潜在言語で一貫して思考することが最良の性能を導くかどうかを検証するために，(1) 推論時におけるモデルの潜在言語を特定し，(2) 潜在言語の一貫性を定量的に評価する尺度を定義する．

(1) 言語特定. 潜在言語を特定するために，先行研究 [5, 6, 7] と同様に，d 次元の中間層の出力 $\boldsymbol{h}_l \in \mathbb{R}^d$ を最終層と同じように語彙数 V へ拡張する $W \in \mathbb{R}^{d\times V}$ 行列との積をとる Logit Lens [4] を用いることで，中間層 l における思考 ($\text{token}_l = W^\top \text{LayerNorm}(\boldsymbol{h}_l)$) を取得し，その層におけるトークンおよび言語を取得する [26].

(2) 一貫性. 中間層全体を通して，モデルがどの程度一貫して特定の潜在言語に依存しているかを定量化するために，事象 $x \in \mathcal{X}$ における確率分布 P と Q の乖離を測る尺度である KL ダイバージェンス [27, 28]$D_{\mathrm{KL}}(P\,\|\,Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}$ に着想を得て，**潜在言語一貫性スコア（LLC Score）**を定義する．層 l および層 $l+1$ における語彙分布の変化量を，層間の KL ダイバージェンス $KL_{l,l+1} \stackrel{\text{def}}{=} D_{\mathrm{KL}}(P_l\,\|\,P_{l+1})$ として定義する．ここで P_l は Logit Lens [4] により得られる層 l における語彙分布を表す．次に，候補言語 (本研究では中国語，英語，日本語) $v \in \mathcal{V}$ が関与する層に着目し，その言語の使用確率で重み付けした層間の乖離量 (どれだけ分布が離れているか) を $\Delta_{l,v} \stackrel{\text{def}}{=} \big(P_{l,v} + P_{l+1,v}\big)\cdot KL_{l,l+1}$ と定義する．ここで $P_{l,v}$ は層 l において潜在言語 v が使用される確率を表す．さらに，層 $l+1$ において支配的な潜在言語が v から切り替わった場合のみを考慮するため，指示関数 $I_{l+1,v} \stackrel{\text{def}}{=} \mathbb{1}\big(v^*_{l+1} \neq v\big)$ を導入する．ただし $v^*_l = \arg\max_{u\in\mathcal{V}} P_{l,u}$ は層 l における最尤の潜在言語を表す．以上を用いて，言語 v に対する一貫性スコアを次式で定義する:

$$\text{Score}(v) = \frac{\displaystyle\sum_{l=1}^{L-1} \Delta_{l,v}\cdot I_{l+1,v}}{\displaystyle\sum_{l=1}^{L-1} \big(P_{l,v} + P_{l+1,v}\big)\cdot I_{l+1,v}} \tag{1}$$

このスコアは，潜在言語 v が使用されているにもかかわらず，次の層で別の言語へ切り替わるような遷移における内部表現の不連続性を，その言語の使用確率で正規化した平均的な乖離量を表す．最終的に，モデル全体の潜在言語一貫性を評価するため，全候補言語に対するスコアの最小値として LLC Score を定義する:

$$\text{LLC Score} = \min_{v\in\mathcal{V}} \text{Score}(v) \tag{2}$$

LLC Score が小さいほど，モデルが少なくとも一つの潜在言語において層をまたいで安定した内部表現を維持していることを意味する．

また性能は，データセットを解かせた際の精度のことであり，潜在言語一貫性スコアと性能の相関を分析することで，潜在言語の必要性を解明する．

4 データセット作成

文中の空欄（_）を特定の単語で補完することを目的としたクローズ形式 [29] のタスクを採用す

表 1　地理文化タスクの結果．各比率ごとの LLC Score と性能の相関（r）を示す．

モデル	質問	敵対的	LLC Score (↓) 0.2	0.4	0.6	0.8	1.0	性能 (↑) 0.2	0.4	0.6	0.8	1.0	r
LLM-jp-3	日	日	日$_{0.06}$	英$_{0.08}$	日$_{0.09}$	日$_{0.10}$	日$_{0.11}$	0.27	0.26	0.27	0.24	0.24	-0.82
		英	日$_{0.09}$	日$_{0.07}$	日$_{0.11}$	日$_{0.10}$	日$_{0.11}$	0.13	0.22	0.13	0.06	0.04	-0.83
	英	日	日$_{0.06}$	日$_{0.07}$	英$_{0.12}$	英$_{0.12}$	英$_{0.13}$	0.15	0.11	0.10	0.15	0.13	-0.17
		英	英$_{0.10}$	英$_{0.11}$	英$_{0.11}$	英$_{0.12}$	英$_{0.13}$	0.22	0.20	0.20	0.18	0.17	-0.97
		中	日$_{0.04}$	英$_{0.12}$	英$_{0.13}$	英$_{0.13}$	英$_{0.13}$	0.10	0.10	0.09	0.10	0.12	-0.05
	中	英	英$_{0.11}$	英$_{0.12}$	英$_{0.12}$	中$_{0.09}$	英$_{0.13}$	0.04	0.05	0.04	0.03	0.02	0.21
		中	日$_{0.05}$	日$_{0.06}$	日$_{0.10}$	英$_{0.11}$	英$_{0.10}$	0.07	0.07	0.06	0.05	0.02	-0.61
Qwen2.5	日	日	英$_{0.10}$	英$_{0.09}$	英$_{0.10}$	英$_{0.09}$	英$_{0.10}$	0.00	0.00	0.00	0.00	0.00	N.A.
		英	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	英$_{0.10}$	英$_{0.10}$	0.00	0.00	0.00	0.00	0.00	N.A.
	英	日	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	0.31	0.30	0.27	0.25	0.27	-0.94
		英	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	英$_{0.10}$	英$_{0.10}$	0.33	0.27	0.25	0.29	0.32	-0.25
		中	英$_{0.09}$	英$_{0.09}$	英$_{0.09}$	英$_{0.10}$	英$_{0.09}$	0.32	0.32	0.31	0.25	0.28	-0.87
	中	英	英$_{0.09}$	英$_{0.08}$	英$_{0.08}$	英$_{0.08}$	英$_{0.09}$	0.00	0.00	0.00	0.00	0.00	0.21
		中	英$_{0.10}$	英$_{0.10}$	中$_{0.09}$	中$_{0.09}$	英$_{0.10}$	0.01	0.01	0.00	0.00	0.01	0.83
Gemma3	日	日	中$_{0.03}$	中$_{0.01}$	中$_{0.03}$	英$_{0.03}$	中$_{0.02}$	0.00	0.01	0.01	0.01	0.02	-0.29
		英	英$_{0.02}$	英$_{0.02}$	中$_{0.02}$	中$_{0.02}$	英$_{0.02}$	0.00	0.01	0.01	0.00	0.00	0.90
	英	日	英$_{0.02}$	英$_{0.03}$	英$_{0.03}$	英$_{0.02}$	英$_{0.02}$	0.31	0.29	0.25	0.19	0.19	0.85
		英	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	0.29	0.27	0.17	0.14	0.12	0.98
		中	英$_{0.03}$	英$_{0.03}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	0.31	0.31	0.25	0.23	0.23	0.98
	中	英	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	0.00	0.01	0.00	0.00	0.00	0.83
		中	中$_{0.02}$	中$_{0.02}$	中$_{0.02}$	中$_{0.02}$	中$_{0.02}$	0.01	0.01	0.00	0.01	0.00	-0.15

る [5, 6, 30]．潜在言語の一貫性がモデルの性能に与える影響を評価することを目的としたクローズタスク用データセットは存在しないため，先行研究 [14, 15] に基づき，GPT-4o [31] を用いて新たに構築する．本研究では潜在言語による影響を受ける翻訳と地理文化の分野に焦点を当て，(1) 質問生成と (2) フィルタリングを経て構築する [14, 15]．

(1) 質問生成．「`Generate cloze task questions and their answers for translation.`」のようなプロンプトを用いて，クローズ形式の質問とその解答を生成する (例: 日本の首都は_である． 答え: (東京)) [29]．生成される問題の多様性を確保するため [32] に，付録に示す 20 種類のカテゴリラベルを事前に用意し，これらのカテゴリをランダムに選択することで，幅広い質問の生成を可能にする．

(2) フィルタリング． 生成された質問文章に対して，品質を確保するためにフィルタリングを行う．生成された質問を GPT-4o に解答させ，質問生成時の参照解答と比較し，モデルの予測解答と一致した場合のみ，最終的なデータセットとして用いる．また，先行研究 [30] に倣って単一トークンのみの解答を用いる．これらのフィルタリングを経て，各タスクにつき 2,000 問からなるデータセットを構築する．

5　実験設定

モデル． 英語，中国語，日本語のデータで事前学習されたモデルとして，それぞれ Gemma3 [33]，Qwen2.5 [34]，および LLM-jp [10] を使用する．各モデルの詳細な設定は付録に示す．

表 2　翻訳タスクの結果．解釈は表 1 と同じである．

モデル	原言語	目的言語	敵対的	LLC Score (↓) 0.2	0.4	0.6	0.8	1.0	性能 (↑) 0.2	0.4	0.6	0.8	1.0	r
LLM-jp-3	日	英	日	英$_{0.07}$	英$_{0.07}$	日$_{0.13}$	日$_{0.07}$	日$_{0.08}$	0.78	0.77	0.73	0.71	0.74	-0.37
			英	英$_{0.07}$	中$_{0.12}$	日$_{0.09}$	日$_{0.09}$	日$_{0.46}$	0.81	0.76	0.68	0.69	0.61	-0.74
			中	英$_{0.08}$	英$_{0.07}$	中$_{0.09}$	英$_{0.07}$	英$_{0.10}$	0.77	0.76	0.74	0.66	0.55	-0.56
	英	日	日	中$_{0.12}$	中$_{0.12}$	中$_{0.11}$	英$_{0.14}$	英$_{0.16}$	0.73	0.71	0.69	0.59	0.37	-0.96
			英	日$_{0.12}$	日$_{0.15}$	日$_{0.14}$	日$_{0.12}$	英$_{0.25}$	0.70	0.67	0.67	0.57	0.62	-0.21
			中	英$_{0.10}$	日$_{0.19}$	英$_{0.11}$	中$_{0.14}$	英$_{0.20}$	0.62	0.66	0.57	0.56	0.55	0.02
		中	日	英$_{0.06}$	英$_{0.06}$	英$_{0.08}$	英$_{0.10}$	日$_{0.16}$	0.33	0.33	0.32	0.28	0.19	-1.00
			英	英$_{0.09}$	英$_{0.06}$	中$_{0.12}$	英$_{0.15}$	日$_{0.10}$	0.37	0.35	0.31	0.29	0.27	-0.59
			中	英$_{0.06}$	英$_{0.06}$	英$_{0.15}$	英$_{0.14}$	英$_{0.15}$	0.39	0.38	0.30	0.29	0.17	-0.85
	中	英	日	日$_{0.08}$	英$_{0.08}$	英$_{0.11}$	中$_{0.12}$	英$_{0.10}$	0.81	0.79	0.74	0.72	0.60	-0.42
			英	英$_{0.08}$	中$_{0.12}$	英$_{0.08}$	英$_{0.07}$	英$_{0.09}$	0.80	0.79	0.65	0.42	0.49	0.48
			中	英$_{0.09}$	英$_{0.09}$	日$_{0.09}$	日$_{0.12}$	英$_{0.08}$	0.71	0.71	0.72	0.70	0.57	0.29
Qwen2.5	日	英	日	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.84	0.85	0.85	0.84	0.84	N.A.
			英	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.05}$	英$_{0.05}$	0.75	0.76	0.74	0.65	0.67	-0.99
			中	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.79	0.78	0.73	0.69	0.69	N.A.
	英	日	日	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
			英	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
			中	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.00	0.00	0.00	0.00	0.00	N.A.
		中	日	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.02	0.01	0.00	0.00	0.00	N.A.
			英	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.01	0.01	0.02	0.00	0.00	N.A.
			中	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	0.01	0.01	0.00	0.00	0.00	N.A.
	中	英	日	英$_{0.00}$	英$_{0.08}$	英$_{0.06}$	英$_{0.09}$	英$_{0.09}$	0.46	0.50	0.33	0.55	0.54	0.38
			英	英$_{0.00}$	英$_{0.08}$	英$_{0.09}$	英$_{0.08}$	英$_{0.00}$	0.68	0.61	0.57	0.65	0.75	-0.86
			中	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.06}$	英$_{0.07}$	0.85	0.83	0.78	0.73	0.74	-0.88
Gemma3	日	英	日	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	英$_{0.02}$	0.78	0.77	0.76	0.30	0.59	0.76
			英	英$_{0.06}$	英$_{0.06}$	英$_{0.03}$	英$_{0.06}$	英$_{0.03}$	0.80	0.82	0.78	0.73	0.68	0.49
			中	英$_{0.03}$	英$_{0.08}$	英$_{0.07}$	英$_{0.24}$	英$_{0.10}$	0.80	0.79	0.69	0.24	0.47	-0.92
	英	日	日	英$_{0.05}$	英$_{0.02}$	英$_{0.05}$	英$_{0.01}$	英$_{0.00}$	0.27	0.14	0.14	0.08	0.07	0.83
			英	英$_{0.05}$	英$_{0.05}$	英$_{0.05}$	英$_{0.00}$	英$_{0.00}$	0.30	0.18	0.20	0.11	0.14	0.81
			中	英$_{0.06}$	英$_{0.05}$	英$_{0.05}$	英$_{0.05}$	英$_{0.05}$	0.32	0.28	0.17	0.17	0.15	0.77
		中	日	英$_{0.05}$	英$_{0.04}$	中$_{0.02}$	英$_{0.04}$	英$_{0.04}$	0.22	0.16	0.15	0.05	0.04	0.30
			英	英$_{0.08}$	英$_{0.05}$	英$_{0.04}$	英$_{0.01}$	英$_{0.01}$	0.23	0.15	0.14	0.02	0.02	0.98
			中	英$_{0.10}$	英$_{0.06}$	英$_{0.08}$	英$_{0.05}$	英$_{0.05}$	0.24	0.24	0.13	0.13	0.06	0.59
	中	英	日	英$_{0.00}$	英$_{0.00}$	英$_{0.00}$	英$_{0.01}$	英$_{0.01}$	0.84	0.81	0.79	0.47	0.55	-0.99
			英	英$_{0.03}$	英$_{0.03}$	英$_{0.00}$	英$_{0.02}$	英$_{0.02}$	0.82	0.85	0.73	0.71	0.55	0.33
			中	英$_{0.00}$	英$_{0.00}$	英$_{0.01}$	英$_{0.03}$	英$_{0.00}$	0.76	0.62	0.48	0.37	0.14	-0.23

データセット． 翻訳タスクにおいては，言語対の組み合わせ ({En, Ja, Zh}-{En, Ja, Zh}) のうち，入力言語と出力言語が同一である場合や，{Ja}-{Zh}翻訳のように，トークン単位で言語を明確に識別することが困難なケースは除外する．

敵対的プロンプト． モデルが持つ潜在言語の一貫性を崩壊させるために，敵対的プロンプトを質問に加えて挿入することで，意図的に困惑させる．具体的には，敵対的プロンプトが入力全体に占める割合を段階的に変化させる．入力長に占める割合として，20%，40%，60%，80%，100% の 5 段階を設定する．詳細については付録に示す．

6　実験結果および分析

表 1 に地理文化，表 2 に翻訳タスクの結果を記載する．また x 軸に LLC Score を，y 軸に精度として可視化した一部の結果を図 1，2，3 に記す．各点は，特定のタスクにおける実験結果に対応する．◆ 1.0 (Ja) は，英語の敵対的プロンプトが入力全体の 100% を占めた場合における結果を示しており，このとき潜在言語は日本語（Ja）である．頑健性と LLC Score のピアソン相関係数を r と表す．

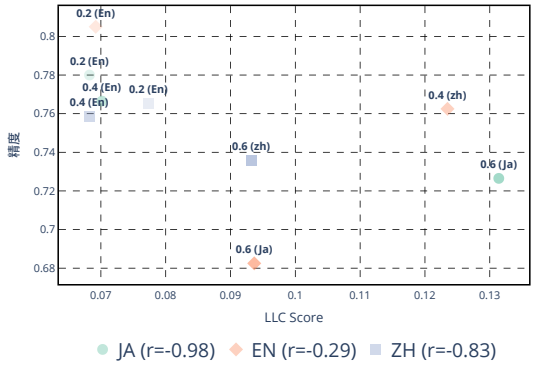

図 1　LLM-jp-3 を用いた翻訳タスク (Ja-En) における一貫性と頑健性の相関（r）を示す．例えば，◆ 割合 (言語) は，英語の敵対的プロンプトを（割合）% 挿入した場合における結果を表しており，このときモデルは潜在言語（言語）で思考していることを示す．

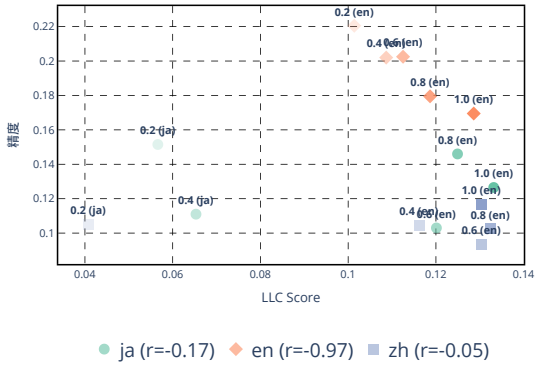

図 2　LLM-jp-3 を用いた地理文化タスク (英語) における一貫性と頑健性の相関 r．解釈は図 1 と同じである．

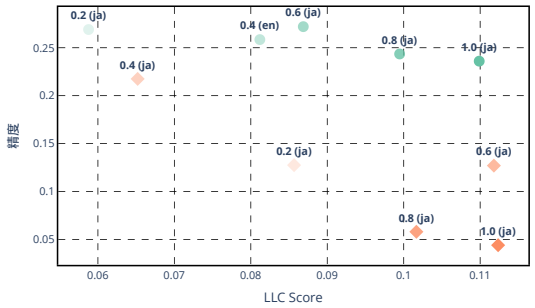

図 3　LLM-jp-3 を用いた地理文化タスク (日本語) における一貫性と頑健性の相関 r．解釈は図 1 と同じである．

性能と潜在言語の一貫性の関係．　図 1 に翻訳タスクの，図 2 と 3 に地理文化タスクの llm-jp-3 の結果を可視化したものを記載する．図 2 では敵対的プロンプトが英語の場合，データ点は直線 $y = -x$ に沿って分布している．y 軸に着目すると，入力言語と敵対的プロンプトの言語が一致する場合，モデルは比較的高い性能を維持する傾向が観察される．質問が英語で与えられた場合かつ敵対的プロンプトも英語であるときに最も良い性能が達成される．

一方で，質問言語と異なる言語の敵対的プロンプトを用いた場合，性能は低下する傾向にある．この結果は，質問言語と敵対的プロンプトの不一致が下流タスク性能を大きく損なう可能性があることを示唆している．しかし，横軸に着目すると異なる傾向が見られる．すなわち，敵対的プロンプトの言語が質問言語と異なる場合であっても，期待される一貫性の破綻が必ずしも生じるとは限らない．この結果は，モデルが入力言語に応じて柔軟に潜在言語を切り替えながら推論できることを示している．

敵対的プロンプトが一貫性に与える影響．　敵対的プロンプトの挿入比率がモデル内部の一貫性に与える影響を分析すると，多くの場合，敵対的プロンプトの比率を増加させると，図 1，2，3 に示すように，内部一貫性に影響が現れることが確認される．

一方で，敵対的プロンプトの言語が質問言語と異なる場合には，潜在言語による処理が必ずしも影響を受けないことが観察される．表 1 と 2 にあるピアソン相関係数 r [35] を確認すると，必ずしも精度と潜在言語の一貫性には正の相関が見られない．これは，潜在言語の一貫性が敵対的プロンプトに対して比較的頑健であり，モデルが内部一貫性を維持するために，自身の得意とする潜在言語で推論する必要はないことを示唆する．

敵対的プロンプトが性能に与える影響．　敵対的プロンプトの割合を増加させると，多くのモデルにおいて頑健性が大きく低下することが確認される．これは，ノイズの増加が下流タスク性能を劣化させることを示した先行研究 [24, 25] と同じである．

7　おわりに

LLMs における潜在言語の一貫性と下流タスクの性能の関係を体系的に分析した．地理文化や翻訳の下流タスクから構成されるデータセットで複数のモデルを評価した結果，一貫した潜在言語を用いる傾向を示すモデルが存在することを確認した一方で，一貫性を維持することが必ずしも下流タスクの性能に繋がるとは限らないことが明らかとなった．これは，最終層付近において内部表現を目的言語に適応させる能力を有しており，思考内容と潜在言語が必ずしも結びついていないことを示唆している．

謝辞

本研究成果は，データ活用社会創成プラットフォーム mdx [36] を利用して得られたものであり，JST BOOST JPMJBY24A6，JPMJBY24B2 の支援を受けたものです．

参考文献

[1] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. **arXiv preprint arXiv:2412.16720**, 2024.

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. **arXiv preprint arXiv:2507.06261**, 2025.

[3] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[4] nostalgebraist. Interpreting gpt: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, August 2020. LessWrong.

[5] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? **arXiv preprint arXiv:2408.10811**, 2024.

[7] Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? **arXiv preprint arXiv:2502.15603**, 2025.

[8] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. What language do japanese-specialized large language models think in? **AAMT Journal**, p. 26, 2025.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[10] Llm jp et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **arXiv preprint arXiv:2407.03963**.

[11] Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. Crosslingual reasoning through test-time scaling, 2025.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. arxiv 2022. **arXiv preprint arXiv:2204.02311**, Vol. 10, p. 1, 2022.

[14] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Shintaro Ozaki, Kazuki Hayashi, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Bqa: Body language question answering dataset for video large language models. **arXiv preprint arXiv:2410.13206**, 2024.

[16] Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. Speaking multiple languages affects the moral bias of language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 2137–2156, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] IO Gallegos, RA Rossi, J Barrow, MM Tanjim, S Kim, and DernoncourtNK. Bias and fairness in large language models: A survey. arxiv, 2023.

[18] Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 12365–12394, Singapore, December 2023. Association for Computational Linguistics.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, Vol. 35, pp. 24824–24837, 2022.

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **NeurIPS**, Vol. 35, pp. 22199–22213, 2022.

[21] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. **arXiv preprint arXiv:2206.07682**, 2022.

[22] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. **arXiv preprint arXiv:2210.03057**, 2022.

[23] Huiyuan Lai and Malvina Nissim. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12012–12026, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[24] Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. Robustness gym: Unifying the NLP evaluation landscape. In Avi Sil and Xi Victoria Lin, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations**, pp. 42–55, Online, June 2021. Association for Computational Linguistics.

[25] Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. More bang for your buck: Natural perturbation for robust question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 163–170, Online, November 2020. Association for Computational Linguistics.

[26] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. **Transformer Circuits Thread**, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[27] Thomas M. Cover and Joy A. Thomas. **Elements of Information Theory**. Wiley, New York, 1991.

[28] J. Lin. Divergence measures based on the shannon entropy. **IEEE Transactions on Information Theory**, Vol. 37, No. 1, pp. 145–151, 1991.

[29] Wilson L Taylor. “cloze procedure”: A new tool for measuring readability. **Journalism quarterly**, Vol. 30, No. 4, pp. 415–433, 1953.

[30] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[31] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. **arXiv preprint arXiv:2410.21276**, 2024.

[32] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. **arXiv preprint arXiv:2407.18416**, 2024.

[33] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. **arXiv preprint arXiv:2503.19786**, 2025.

[34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. **arXiv preprint arXiv:2412.15115**, 2024.

[35] Karl Pearson and Francis Galton. Vii. note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, Vol. 58, No. 347-352, pp. 240–242, 1895.

[36] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[38] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. **arXiv preprint arXiv:2303.08112**, 2023.

[39] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In **The 41st international ACM SIGIR conference on research & development in information retrieval**, pp. 1097–1100, 2018.

A 付録

詳細なモデルの設定. 実験に使用した各モデルの詳細は，表 3 に示す．再現性を確保するため，seed は 42 に，top_p は 0.0 に設定した．実装には Transformers ライブラリ [37] を用いた．

表 3 詳細なモデル名と Hugging Face の名前

モデル	パラメータ	入力トークン	詳細な名前
Gemma3	1B	32,768	google/gemma-3-1b-it
Qwen2.5	1.5B	32,768	Qwen/Qwen2.5-1.5B-Instruct
LLM-jp-3	1.8B	4,096	llm-jp/llm-jp-3-1.8b-instruct
GPT-4o	–	128k	gpt-4o-2024-11-20

分析手法に関して. 3 節において，言語判定には langid を用いる (https://github.com/saffsd/langid.py)．また潜在言語一貫性スコアの算出に関して，先行研究 [4, 38] に従い，本研究では意味的推論が収束しやすいモデル後半の中間層のみを分析対象とする．

A.1 データセットにおける質問の多様性

Self-BLEU [39] は，生成されたテキスト集合内の多様性を評価するために広く用いられている指標である．この指標は，データセット中の各サンプルが他のサンプルとどの程度異なっているかを測定することで，生成結果に含まれる表現の幅を定量的に評価する．Self-BLEU の値が低いほど多様性が高く，値が高いほどデータセット内に冗長性や繰り返しが多いことを示す．単一の質問 q_i に対する Self-BLEU スコアは，以下のように定義される．

$$\text{Self-BLEU}(q_i) = \text{BLEU}(q_i, \mathbb{Q} \setminus \{q_i\}) \tag{3}$$

ここで，q_i はデータセット内の単一の質問を表し，$\mathbb{Q}$ はデータセット全体に含まれる質問集合を表す．また，$\mathbb{Q} \setminus \{q_i\}$ は，q_i を除いたすべての質問からなる集合である．データセット全体の Self-BLEU スコアは，各質問に対する Self-BLEU の平均として，次式で定義される．

$$\text{Self-BLEU}(\mathbb{Q}) = \frac{1}{|\mathbb{Q}|} \sum_{q_i \in \mathbb{Q}} \text{Self-BLEU}(q_i) \tag{4}$$

ここで，$|\mathbb{Q}|$ はデータセットに含まれる質問数を表す．この定義により，質問間の重複度合いを定量的に評価することが可能となる．より多様なデータセットほど，質問間の表現の重なりが少なくなるため，Self-BLEU の値は低くなる傾向を示す．本研究では，GPT-4 によって生成される問題文の多様性を高めるために，あらかじめカテゴリの一覧を作成し，その中から約 20 種類のより具体的なカテゴリをランダムに選択する手法を採用した．このカテゴリ選択によって生成される質問のばらつきを増加させ，その結果を表 4 に示す．

表 4 4-gram に基づく Self-BLEU の結果を示す．この結果から，同一のプロンプトを用いて GPT-4o に質問を生成させた予備的なデータセットと比較して，本研究で構築したデータセットの方が，より高い多様性を有していることが確認できる．

	Self-BLEU (↓)			
タスク	事前実験		提案手法	
	質問言語	スコア	質問言語	スコア
地理文化	英語	0.9976	英語	**0.7336**
	日本語	0.9707	日本語	**0.6007**
	中国語	0.9653	中国語	**0.6368**
翻訳	日本語 → 英語	0.9559	日本語 → 英語	**0.6180**
	中国語 → 英語	0.9273	中国語 → 英語	**0.7807**
	英語 → 日本語	0.9940	英語 → 日本語	**0.6925**
	英語 → 中国語	0.9950	英語 → 中国語	**0.6887**

カテゴリ

Capital City, Official Language, National Currency, Head of State, Independence Day, Major Religion, National Anthem, Famous Landmark, National Dish, Traditional Clothing, Flag Colors, Historical Figure, Major River, Mountain Range, Neighboring Countries, Climate, Population, Area (km^2), UNESCO World Heritage Site, Government Type

A.2 詳細な敵対的プロンプト

日本語の敵対的プロンプト

日本の文化や背景，特徴，歴史について説明します． 日本は，東アジアに位置する島国で，太平洋に面し，主に本州，北海道，九州，四国の 4 つの主要な島と，それに付随する約 6,800 以上の小さな島々から構成されています．国土の多くが山地で，自然災害，特に地震や台風が多い地域でもあります．これらの自然環境は，日本人の生活様式や精神性，信仰に大きな影響を与えてきました． まず，日本の歴史について触れます．日本の歴史は数千年にわたり，時代ごとに独自の文化が形成されてきました． 古代では，弥生時代に稲作が始まり，やがて統一国家の萌芽としてヤマト政権が登場します．仏教の伝来は 6 世紀ごろで飛鳥時代には中国・，....