n-gram に基づく推論モデルの信頼度と較正特性の分析

尾崎 慎太郎 ♡　橋本 航 ♡,♠　上垣外 英剛 ♡　林 克彦 ♢　渡辺 太郎 ♡
♡ 奈良先端科学技術大学院大学 (NAIST)　♠Sansan 株式会社　♢ 東京大学
{ozaki.shintaro.ou6, kamigaito.h, taro.watanabe}@naist.ac.jp

概要

大規模言語モデルは，最終的な答えを生成する前に推論過程を明示的に出力することで性能が向上する．一方で，推論過程には「A か B か判断が難しい」「情報が不足している」といった言語的な不確実性表現が含まれる場合があるにもかかわらず，最終的な出力はしばしば過信的となる．このような推論過程と最終出力との間に生じる過信の乖離は，モデルの不確実性推定の信頼性を損ない，誤答や幻覚に対しても過度に確信的な回答を生成する要因となる．本研究では，推論過程内部に現れる言語表現と過信度合いとの関連性を回帰分析によって明らかにする．複数のモデルおよび複数の質問応答タスクを用いた実験の結果，モデルが特定の言語表現を生成することと，信頼度を過剰に高める，すなわちモデル自身が過信的になる傾向があることを明らかにした．さらに，特定の言語表現によって過剰信頼を軽減させ，正解率の向上にも寄与することを示した．

1　はじめに

大規模言語モデル [1, 2, 3, 4] は，最終的な答えのみを直接出力するのではなく，その過程として段階的な推論を明示的に生成することで性能を向上させる手法が広く用いられている [5, 6, 7]．このような流れの中で，強化学習などを用いて推論過程そのものを学習に組み込んだ推論モデル（Reasoning models） [8, 9, 10, 11] が注目を集めている．

一方で，推論モデルが生成する推論過程には，「A か B か判断が難しい」「情報が不足している」「別の可能性も考えられる」といった言語的な不確実性表現が含まれる場合があるにもかかわらず，最終的な出力はしばしば高い確信度を伴って生成されることが指摘されている [12, 13]．このような推論過程と最終出力との間に生じる過信の乖離は，モデルが自身の不確実性を適切に表現できていないことを示唆しており，誤答や幻覚に対しても過度に確信的な回答を生成する要因となる [13, 14, 15]．

モデルの信頼度が過剰であることに対処するために，最終出力に付与された確信度を真の正解確率に整合させる信頼度較正 [16, 17, 18] や，不確実性定量化に関する研究が数多く提案されてきた．しかし，これらの多くは最終出力として得られる数値的な確信度の較正に重きを置いており，推論過程内部に現れる言語表現そのものが，モデルの信頼度形成にどのように寄与しているかについては十分に明らかにされていない．すなわち，「なぜモデルは過信的になるのか」，「どの言語情報が信頼度に寄与するか」という問いに対し，推論過程中の言語的な振る舞いの観点から定量的に分析した研究は限定的である．

本研究では，推論モデルが生成する推論過程に着目し，そこに含まれる言語的な特徴とモデルが示す信頼度の関係を定量的に分析する．具体的に，生成された推論過程を *n*-gram に分割し，信頼度を目的変数とした回帰分析 [19] を行うことで，どのような言語表現が過剰信頼あるいは過小信頼に寄与しているのかを明らかにする．また, モデルによる生成手法や強制デコード手法など複数手法による信頼度の算出を行い，手法によって言語情報と信頼度との関係がどのように異なるかについても検証する．さらに我々は同様に正解率に対しても回帰分析を行い，信頼度との関係についても分析する．

複数の推論モデルおよび複数の質問応答タスクを用いた実験の結果，モデルが特定の言語表現を生成することが，信頼度を過剰に高め，モデル自身を過信的に振る舞わせる傾向があることを明らかにした．また，別の言語表現は信頼度を低下させる方向に作用することも確認された．さらに，推論を長くし性能向上に寄与する特定の言語表現 [20] が信頼度にも大きく寄与する言語表現であることを明らかにした．一方で，正解率の向上に寄与する言語表現と，信頼度の向上に寄与する言語表現は存在するものの必ずしも一致せず，言語表現レベルで正解率と信頼度が乖離しうる結果となった．

2 関連研究

推論モデル. モデルに段階的な推論過程を明示的に生成させる Chain-of-Thought (CoT) [5, 6, 7] は数学や常識推論，記号推論など幅広いタスクで性能を大きく向上させることが示されている [21]．さらに，複数の推論過程を生成し多数決で解を選択する自己整合性 [22] や，問題分解を促す least-to-most prompting [23] など，CoT を拡張する手法も提案されてきた．これらの研究は，推論過程そのものがモデル性能に寄与することを示しており [24]，強化学習によって推論部分を含めて最適化する推論モデルを構築するための基盤となっている．こうした流れを受け，Group Relative Policy Optimization（GRPO）[21] などの強化学習により推論部分を含め学習した推論モデルが注目されている．推論モデルは一般に，<think> タグにより区切られた推論部分と，それ以降の出力部分の二段構造を持つ．

推論モデルと信頼度. 推論モデルにおける信頼度の問題に関して，推論能力の向上と同時に，モデルが自身の不確実性を適切に表現できていない点が指摘されている．Kirichenko ら [12] は，回答不能な質問に対しても推論モデルが強い確信を伴う誤答を生成する傾向を示し，Xiong ら [13] は，言語的に不確実性を表現している場合であっても，数値的な確信度が適切に低下しないことを報告している．これらの問題に対し，最終出力の確信度を真の正解確率に整合させる信頼度較正を導入する研究が行われてきた [14, 15] 一方で，最終出力に付与された信頼度の較正に重きを置いており，推論過程内部に現れるどのような言語表現が，過剰信頼や過小信頼に寄与しているかについては十分に分析されていない．

3 分析手法

信頼度の算出. <think>タグに含まれる推論部分の信頼度の算出に関して，本研究では (1) 生成による手法と (2) 強制デコード (forced-decoding) 手法の 2 つの手法によって算出する．(1) 生成による手法では，一度 QA を解く際に生成された<think>タグ内の推論部分を再度入力として用い，先行研究 [25, 26] に倣って，`So, the answer is {candidate}. Now I will rate my confidence on a scale of 1-10. Please generate only the score. Proposed confidence:` と選択肢の数だけ推論部分を 10 段階で評価することによって算出し，最も高いものを信頼度とする．

(2) 強制デコードによる手法では，同じように生成された<think>タグ内の推論部分に先行研究 [27, 28] の手法に倣って，`So, the answer is` という文章を付与をし，下記の数式 1 により算出する．

$$P(x_i) = \frac{\exp\left(\log P(x_i \mid \texttt{prompt})\right)}{\sum_{j=1}^{J} \exp\left(\log P(x_j \mid \texttt{prompt})\right)} \tag{1}$$

x_i は i 番目の選択肢に対応するトークンを表し，J は選択肢の総数である．$P(x_i \mid \texttt{prompt})$ は，推論部分および `So, the answer is` までを含むプロンプトが与えられた条件下で，モデルが選択肢 x_i を生成する確率を表す．これらの対数確率に対して softmax 正規化を施すことで，各選択肢に対する確率分布 $P(x_i)$ を得る．本研究では，モデルが最終的に選択した選択肢に対応する $P(x_i)$ を信頼度として用いる．

回帰分析. モデルが生成した<think>タグ内の推論部分に含まれる言語的特徴と，モデル自身が示す予測の確信度との関係を定量的に分析するために，n-gram に基づく回帰分析を行う．各データ点は，モデルが生成した推論部分と，その推論に基づいて選択された回答に対する確率値から構成される．確信度 y_i は，モデルが最終的に予測した選択肢に対応する確率として定義し，$[0, 1]$ の連続値として扱う．推論部分は前処理として小文字化を行い，英字のみから成るトークンを抽出する．その後，$n = 1$ から 5 までの n-gram 頻度を用いて特徴量化を行う．なお，頻度によるバイアスを避けるために各 n-gram ごとに異なるモデルを学習し，最後に Z-スコア標準化を用いて数値の比較を行う．詳細は付録に記載している．各推論部分 r_i は，語彙集合に基づく特徴ベクトル $\mathbf{x}_i \in \mathbb{R}^d$ に変換され，これらを行として並べた行列を $\mathbf{X} \in \mathbb{R}^{N \times d}$ とする．確信度と n-gram 特徴量の関係は，線形回帰モデル $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ によって表現される．ここで $\boldsymbol{\beta} \in \mathbb{R}^d$ は各 n-gram に対応する回帰係数である．特徴ベクトルは推論部分に当たる文章の長さによって重みが影響されないように，$\mathbf{x}_i$ に対し L2 正規化を行う．本研究では，先行研究と同様に [29]，特徴量の選択と過学習の抑制を同時に行うため，L1 正則化を課した Lasso 回帰 [19] を用いる:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}^\top \boldsymbol{\beta} \right)^2 + \lambda \|\boldsymbol{\beta}\|_1 \right) \tag{2}$$

ここで λ は正則化係数を表し，$\|\boldsymbol{\beta}\|_1$ は係数ベクトルの L1 ノルムである．推定された回帰係数の符号と大きさに基づき，確信度を高める方向に寄与する n-gram と，確信度を低下させる方向に寄与する

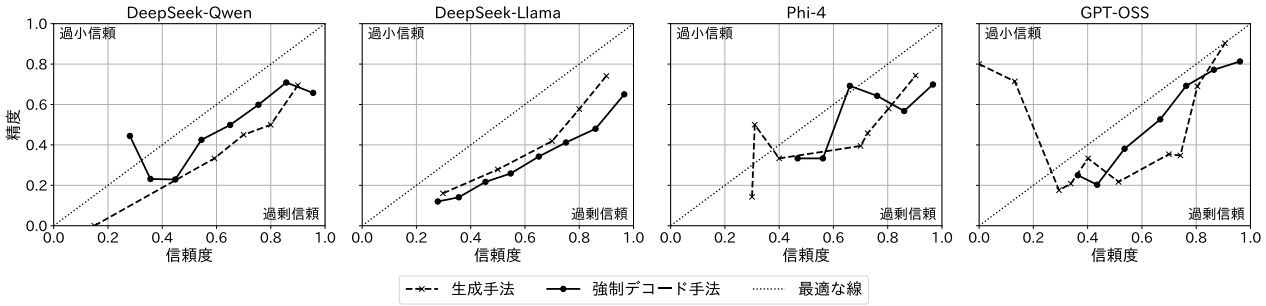

図 1　信頼度曲線．ビンは 10 ポイントずつに幅を区切っている．解釈として，モデルの信頼度が 70%の時にその精度も 70%であることが望ましいということである．

n-gram を抽出する．これにより，モデルが高い確信度を示す推論と低い確信度を示す推論に影響を与える言語表現を定量的に明らかにする．

4　実験設定

モデル． 実験で用いる推論モデルにはモデルの大きさや性能などを考慮して，DeepSeek-R1 [8] から蒸留された [30] Llama および Qwen，また Phi-4 [31] と GPT-OSS [32] を用いる．詳細な実験設定およびモデル名は付録に記載する．

データセット． QA データセットには，信頼度を算出可能な複数選択 QA を用いる．具体的に数学のタスクである MathQA [33]，様々な分野を有する MMLU [34]，長文読解や英文理解の HellaSwag [35]，RACE [36]，および CosmosQA [37] の 5 つのデータセットを用いる．MathQA は 5 択，それ以外の QA は 4 択である．MMLU には数学や情報科学，物理などの推論を要する分野に絞り，それぞれのデータサイズは 2,985，1,990，10,042，3,498，2,985 件ある．

5　実験結果および分析

信頼度と精度をプロットした信頼度曲線を図 1 に，また回帰分析の結果を表 1 に記載する．図 1 の結果を見ると，全てのモデルにおいて信頼度が高い時に相応した正解率が出ていないことがわかり，常に過剰信頼であることが示される．

推論モデルの信頼度は適切か． 図 1 に記載した信頼度曲線の図は x 軸が 3 節で説明した信頼度，また y 軸は正解率を表す．信頼度曲線の解釈として，それぞれのビンに分けた際に信頼度に沿った正解率であるかを可視化し，モデルが過剰信頼か過小信頼かどうか判断することのできる (例えば 70%の自身で予測した問題の正解率は 70%であることが望ましい)．本研究では先行研究に倣って [38, 39, 40]，それぞれのビンを 10 本に分け，そのビンに含まれる信頼度の平均および，正解率をプロットしている．結果を見ると，点線で書かれた生成手法では信頼度が高い区間のほとんど (0.4 以降) で最適な線よりも低いことを表し，これは信頼度が大きい一方で正解率が相応していない過剰信頼の状態を示す．この結果は強制デコード手法の結果や他のモデルを見ても同じ傾向であり，過剰信頼であることがわかる [12]．

どのような言語情報が信頼度に影響するか． 表 1 に，各 *n*-gram の頻度と，信頼度および精度に与える影響を示す．Qwen では，信頼度に正の寄与を示す *n*-gram として option, option hmm let try, maybe, perhaps, perhaps wait perhaps calculate など，解答を断定せず複数の可能性を示唆する表現が多い．一方，負の寄与を示す表現には perhaps mistake wait let check, let check, subtract, make sure didn[1)] make mistake, double check など思考を整理する表現が多い．

Llama においても同様に，figure, perhaps, looking options, wait perhaps, answer wait double check など推論中の検討や再思考を示す表現が信頼度を高める一方，need determine, calculate, total number など，思考を整理する表現が信頼度を下げる傾向がある．以上より，両モデルに共通して，推論モデルの信頼度は推論の正確さよりも，推論過程をどのような表現にするかで影響されることが分かる．さらに，先行研究により [20]，wait や hmm などの語を意図的に挿入して推論を長くすることが，性能向上に寄与することが知られている．我々はこれらの語が性能向上に寄与するだけでなく，モデルの信頼度にも影響を与える重要な要素であることを明らかにした．

正解率と信頼度に影響を与える言語表現． 表 1 では，信頼度と正解率に対して異なる方向に作用する *n*-gram が多数観察される．特に，青太字で示さ

1)　Vectorizer の `token_pattern` により't が除去されている．

表 1　Llama および Qwen の Lasso 回帰による信頼度分析，Logistic 回帰による正解率分析の結果．*n*-gram のスコアは全特徴量に対して Z スコア標準化している．† は標準化前の特徴量が 0 であった結果を示す．**青太字**は信頼度が低下し正解率が向上した特徴，赤下線は信頼度が向上し正解率が低下した特徴を示す．

1-gram	信頼度	精度	頻度	2-gram	信頼度	精度	頻度	3-gram	信頼度	精度	頻度	4-gram	信頼度	精度	頻度	5-gram	信頼度	精度	頻度
Qwen																			
looking	1.97	0.00†	27776	options answer	2.02	0.10†	972	options hmm let	1.54	0.02†	79	options hmm let try	1.58	-0.12†	31	figure option correctly answers question	1.16	0.00†	308
option	1.95	0.00†	249743	looking options	1.55	0.43	16721	answer wait let	1.42	0.02†	505	answer wait wait let	1.48	-0.12†	43	trying figure option correctly describes	1.12	0.00†	92
answer	1.38	0.00†	83503	options hmm	1.52	0.10†	1107	option perhaps answer	1.26	0.02†	371	correct answer wait let	1.39	-0.12†	139	need trying figure answer question	1.12	0.00†	122
bit	1.21	0.00†	11949	options let	1.33	0.10†	1894	options answer wait	1.21	0.02†	159	second let double check	1.38	-0.12†	26	need figure correct answer question	1.12	0.00†	707
true	1.17	0.00†	8331	look options	1.32	0.10†	1035	need understand option	1.09	0.02†	255	option wait let double	1.23	-0.12†	88	need figure option correctly describes	1.12	0.00†	259
say	1.07	0.00†	54978	question man	1.20	0.10†	1737	looking options let	1.06	0.02†	303	options answer wait let	1.20	-0.12†	83	need figure answer question based	1.10	0.00†	72
like	1.05	0.00†	26787	wait options	1.17	0.10†	3318	let read context	1.06	0.02†	193	correct answer looking options	1.19	-0.12†	75	need figure right answer let	1.09	0.00†	50
question	0.92	0.00†	103298	sounds like	1.13	0.10†	2642	wait wait let	1.06	0.02†	178	misunderstanding options let read	1.08	-0.12†	134	question asking option correctly describes	1.08	0.00†	371
action	0.83	0.00†	16840	action option	1.09	0.10†	858	let option option	1.05	0.02†	337	looking options answer wait	1.07	-0.12†	85	let read passage carefully carefully	1.07	0.00†	74
passage	0.83	0.00†	24732	makes sense	1.08	0.10†	4375	options describing different	1.02	0.02†	207	carefully make sure understand	1.05	-0.12†	60	trying figure correct answer question	1.06	0.00†	132
perhaps	-4.44	0.00†	91550	**really stuck**	**-7.41**	**0.10†**	**865**	**equation let number**	**-6.48**	**0.02†**	**10**	set equation let number	-5.45	-0.12†	5	think ve exhausted possibilities correct	-4.27	0.00†	70
consider	-3.31	0.00†	6602	**divide total**	**-3.68**	**0.10†**	**123**	**really stuck think**	**-6.13**	**0.02†**	**383**	approach differently let consider	-5.27	-0.12†	10	think correct answer options perhaps	-4.07	0.00†	46
subtract	-2.71	0.00†	2359	**let denote**	**-3.27**	**0.10†**	**1104**	**km upstream km**	**-4.82**	**0.02†**	**44**	really stuck think correct	-4.57	-0.12†	93	really stuck think correct answer	-3.90	0.00†	86
problem	-2.32	0.00†	25482	**let problem**	**-2.97**	**0.10†**	**537**	**total parts parts**	**-3.97**	**0.02†**	**33**	going downstream effective speed	-4.38	-0.12†	10	perhaps mistake problem setup wait	-3.68	0.00†	184
total	-2.04	0.00†	19905	**power source**	**-2.48**	**0.10†**	**30**	**complete work days**	**-3.37**	**0.02†**	**54**	problem wait maybe problem	-4.16	-0.12†	38	perhaps question asking profit percentage	-3.39	0.00†	123
profit	-0.97	0.00†	4676	**total parts**	**-2.33**	**0.10†**	**181**	**perhaps mistake problem**	**-3.25**	**0.02†**	**681**	selling price selling price	-3.93	-0.12†	20	really stuck think conclude correct	-3.06	0.00†	94
term	-0.94	0.00†	6994	**think ve**	**-2.25**	**0.10†**	**965**	**profit profit percentage**	**-3.05**	**0.02†**	**198**	time distance divided speed	-3.59	-0.12†	20	option perhaps mistake problem options	-3.05	0.00†	32
meter	-0.87	0.00†	5969	**speed time**	**-2.21**	**0.10†**	**428**	**using pythagorean theorem**	**-2.62**	**0.02†**	**8**	think ve exhausted possibilities	-3.50	-0.12†	179	perhaps correct answer listed unlikely	-2.89	0.00†	76
km	-0.87	0.00†	6688	**answer listed**	**-2.20**	**0.10†**	**2588**	**conclude correct answer**	**-2.33**	**0.02†**	**942**	kg looking options option	-3.45	-0.12†	8	option perhaps correct answer listed	-2.88	0.00†	348
number	-0.86	0.00†	27142	**profit percentage**	**-2.13**	**0.10†**	**831**	**total volume new**	**-2.29**	**0.02†**	**10**	let right let double	-3.30	-0.12†	7	time distance divided speed time	-2.65	0.00†	7
Llama																			
corresponds	3.86	0.14†	440	corresponds option	3.24	0.14†	149	option wait let	1.41	0.08†	484	answer wait let double	1.03	-0.29†	191	wait let double check make	1.29	0.00†	117
looking	1.37	0.14†	22357	options hmm	1.24	0.14†	1088	provided options correct	1.33	0.08†	13	option wait let double	0.93	-0.29†	159	answer wait let double check	1.09	0.00†	162
wait	1.37	0.14†	98411	looking options	1.06	0.68	13407	answer wait let	1.31	0.08†	548	option let double check	0.91	-0.29†	72	need figure correct answer question	1.00	0.00†	542
option	1.13	0.12	231577	option wait	1.03	0.14†	6005	looking options provided	1.14	0.08†	89	need figure option correct	0.88	-0.29†	357	need figure option correctly correct	0.98	0.00†	144
figure	1.09	0.14†	11586	options given	0.94	0.14†	2006	answer alternatively perhaps	1.03	0.08†	592	need trying figure option	0.88	-0.29†	206	need figure option correct answer	0.97	0.00†	97
sure	1.00	0.14†	14828	options let	0.89	0.14†	2319	options hmm let	1.02	0.08†	134	need figure option correctly	0.87	-0.29†	1193	based passage provided let read	0.95	0.00†	138
answer	0.96	0.14†	61337	let options	0.83	0.14†	1191	looking options answer	0.99	0.08†	169	need figure correct answer	0.83	-0.29†	766	need figure correct answer let	0.92	0.00†	83
hmm	0.92	0.14†	12900	let think	0.81	0.14†	5744	looking options matches	0.99	0.08†	98	approach question okay need	0.82	-0.29†	162	need figure option best answer	0.91	0.00†	59
let	0.79	1.53	65410	options need	0.79	0.14†	981	let check options	0.94	0.08†	560	based passage provided let	0.81	-0.29†	146	need trying figure option correct	0.91	0.00†	72
case	0.75	0.14†	7906	answer wait	0.76	0.14†	3515	looking options let	0.93	0.08†	470	need trying figure correct	0.78	-0.29†	204	trying figure correct answer question	0.91	0.00†	203
listed	**-7.79**	**0.14†**	**3162**	**options wrong**	**-5.69**	**0.14†**	**735**	**incorrect options wrong**	**-5.73**	**0.08†**	**150**	think conclude correct answer	-6.37	-0.29†	183	problem incorrect options wrong choose	-4.01	0.00†	50
perhaps	-3.28	-2.17	70727	**okay determine**	**-5.67**	**0.14†**	**41**	**conclude correct answer**	**-4.87**	**0.08†**	**341**	think ve exhausted possibilities	-5.05	-0.29†	262	options perhaps correct answer listed	-3.91	0.00†	100
finally	**-3.07**	**0.14†**	**533**	**need determine**	**-3.35**	**0.14†**	**371**	**mistake alternatively perhaps**	**-4.38**	**0.08†**	**257**	option perhaps question incorrect	-3.59	-0.29†	74	correct answer isn option perhaps	-3.57	0.00†	69
determine	**-2.83**	**0.14†**	**1588**	**answer listed**	**-3.24**	**0.14†**	**1870**	**choose options perhaps**	**-4.36**	**0.08†**	**231**	option perhaps problem incorrect	-3.30	-0.29†	137	think ve exhausted possibilities correct	-3.22	0.00†	70
ll	**-2.74**	**0.14†**	**4579**	**equals finally**	**-2.85**	**0.14†**	**15**	**let try problem**	**-4.32**	**0.08†**	**54**	problem wait maybe problem	-3.01	-0.29†	105	option perhaps correct answer listed	-3.18	0.00†	247
consider	**-1.55**	**0.14†**	**5270**	**answer isn**	**-2.73**	**0.14†**	**1034**	**left frac right**	**-3.74**	**0.08†**	**16**	finally ll divide sides	-2.92	-0.29†	5	think conclude correct answer options	-2.98	0.00†	23
vote	**-1.12**	**0.14†**	**699**	**option perhaps**	**-2.57**	**0.14†**	**6005**	**answer option perhaps**	**-3.45**	**0.08†**	**954**	option perhaps answer listed	-2.86	-0.29†	164	correct answer isn options perhaps	-2.78	0.00†	61
mistake	**-1.06**	**0.14†**	**10804**	**ll subtract**	**-2.54**	**0.14†**	**84**	**think ve exhausted**	**-3.21**	**0.08†**	**334**	isn option perhaps answer	-2.84	-0.29†	227	number numbers divide leaving remainder	-2.72	0.00†	5
total	**-0.91**	**0.14†**	**28064**	**think ve**	**-2.23**	**0.14†**	**1206**	**okay need determine**	**-3.03**	**0.08†**	**137**	problem incorrect options wrong	-2.80	-0.29†	76	sides wait let double check	-2.57	0.00†	5
wrong	**-0.90**	**0.14†**	**4756**	**result think**	**-1.96**	**0.14†**	**488**	**need greatest common**	**-2.90**	**0.08†**	**8**	okay need determine total	-2.42	-0.29†	35	think conclude correct answer isn	-2.54	0.00†	33

れた特徴は信頼度を低下させる一方で正解率を向上させ，赤下線で示された特徴はその逆の傾向を示している．赤下線の *n*-gram には，perhaps, maybe, seems, intended answer, mistake など，不確実性や再思考の可能性を示す表現が多く見られる．これらは推論過程における慎重さや丁寧さのような態度を言語的に表現しており，モデルの信頼度を高める方向に作用する一方で，推論が冗長化したり判断が曖昧になったりすることで，正解率を低下させる傾向を持つ可能性がある．これに対し，**青太字**の *n*-gram には，divide, subtract, set equation, calculate, total number など，問題を構造化し，明示的な操作として処理する表現が多く含まれる．これらは推論モデルが内部で行っている段階的な推論をそのまま言語化したものであり，論理的一貫性や正解率の向上に寄与していると考えられる．一方で，こうした表現は推論の迷いや検討過程をほとんど含まず，結論志向で断定的になりやすいため，モデルが出力する信頼度は相対的に低くなる傾向がある．

6　おわりに

本研究では，推論モデルにおいて，その推論過程に現れる言語表現と，モデルが最終的に示す信頼度との関係を分析した．その結果，信頼度は推論結果の正確さそのものよりも，推論過程がどのような言語表現であるかに強く影響されることが明らかになった．特に，不確実性や再検討の可能性を示唆する言語表現は，推論が慎重に行われている印象を与えることで信頼度を高める．また問題を構造化し，明示的な演算や操作として推論を進める言語表現は，論理的一貫性や正解率の向上に寄与することが示された一方で，これらの表現は検討や迷いといった推論過程をほとんど含まないため，結果としてモデルが出力する信頼度は相対的に低くなる傾向がある．これらは，推論過程に含まれる言語表現が，単に内部状態を反映する副次的な産物ではなく，モデル自身の信頼度形成に独立した影響を及ぼす重要な要因であることを示唆する結果となった．

参考文献

[1] 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.

[2] 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.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[4] 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.

[5] 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.

[6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, pp. 22199–22213, 2022.

[7] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. **Transactions on Machine Learning Research**, 2022. Survey Certification.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. **Nature**, Vol. 645, No. 8081, pp. 633–638, 2025.

[9] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. **arXiv preprint arXiv:2504.21318**, 2025.

[10] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. **arXiv preprint arXiv:2501.09686**, 2025.

[11] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. **arXiv preprint arXiv:2505.00949**, 2025.

[12] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. **arXiv preprint arXiv:2506.09038**, 2025.

[13] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In **The Twelfth International Conference on Learning Representations**, 2024.

[14] Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 3769–3793, Suzhou, China, November 2025. Association for Computational Linguistics.

[15] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In **Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery**, pp. 6107–6117, 2025.

[16] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the NAACL (Volume 1: Long Papers)**, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[17] Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 2025.

[18] Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. Calibrating reasoning in language models with internal consistency. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Vol. 58, No. 1, pp. 267–288, 1996.

[20] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 20286–20332, Suzhou, China, November 2025. Association for Computational Linguistics.

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint arXiv:2402.03300**, 2024.

[22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[23] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[24] Minhyuk Kim, Seungyoon Lee, and Heuiseok Lim. TORSO: Template-oriented reasoning towards general tasks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 16821–16829, Suzhou, China, November 2025. Association for Computational Linguistics.

[25] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. **TMLR**, 2022.

[26] Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 20090–20111, Vienna, Austria, July 2025. Association for Computational Linguistics.

[27] Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, and Junichi Tsujii, editors, **Proceedings of the 24th Workshop on Biomedical Language Processing**, pp. 1–17, Viena, Austria, August 2025. Association for Computational Linguistics.

[28] Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examining the necessity and accuracy of CoT step-by-step. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the NAACL 2025**, pp. 2586–2606, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[29] Victor Chahuneau, Kevin Gimpel, Bryan R. Routledge, Lily Scherlis, and Noah A. Smith. Word salad: Relating food prices and descriptions. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, **Proceedings of the 2012 EMNLP**, pp. 1357–1367, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[30] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. **arXiv preprint arXiv:1503.02531**, 2015.

[31] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. **arXiv preprint arXiv:2412.08905**, 2024.

[32] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Bowen Arbus, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. **arXiv preprint arXiv:2508.10925**, 2025.

[33] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the NAACL**, pp. 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[34] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **ICLR**, 2021.

[35] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of ACL**, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[36] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[37] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on EMNLP-IJCNLP**, pp. 2391–2401, Hong Kong, China. Association for Computational Linguistics.

[38] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, **Proceedings of the 34th International Conference on Machine Learning**, Vol. 70 of **Proceedings of Machine Learning Research**, pp. 1321–1330. PMLR, 06–11 Aug 2017.

[39] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In **CVPRW**, 2019.

[40] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 29, 2015.

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, 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 EMNLP demo**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, Vol. 12, pp. 2825–2830, 2011.

[43] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. **Journal of the royal statistical society: Series B (Methodological)**, Vol. 36, No. 2, pp. 111–133, 1974.

A 付録

詳細な実験設定. 本研究の実験は Transformers ライブラリ [41] を用いて行った. 生成に際して, max_new_tokens を 32,768 (信頼度を生成する際は 10) に, top_p を 0.95, temperature を 0.6 に, seed を 42 に設定し, GPU は NVIDIA RTX 6000 Ada Generation を用いた. 表 2 に使用した詳細なモデル名を, 表 3 にそれぞれのモデルの QA に対する性能を記す.

表 2 実験で使用した詳細なモデル名.

モデル名	HuggingFace 上での名前
Qwen	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
Llama	deepseek-ai/DeepSeek-R1-Distill-Llama-8B
Phi-4	microsoft/Phi-4-reasoning
GPT-OSS	openai/gpt-oss-20b

詳細な回帰分析の設定. 本研究では scikit-learn [42] で実験を行い, 正則化係数 λ は交差検証 (検証数は 3) [43] により選択し, 検証データにおける平均二乗誤差が最小となる値を採用した. その他は標準の設定に従った.

正解率の回帰分析. 本研究では, テキスト集合 $\{x_i\}_{i=1}^N$ と対応する二値ラベル $\{y_i\}_{i=1}^N$ ($y_i \in \{0,1\}$) に対し, n-gram 頻度に基づく L1 正則化ロジスティック回帰モデルを用いて分類を行う. 各文書 x_i は, 5 回以上出現した n-gram を特徴とする二値ベクトル $\mathbf{x}_i \in \{0,1\}^p$ に変換され, 行列 $\mathbf{X} \in \{0,1\}^{N\times p}$ を構成する. 分類モデルは

$$P(y_i = 1 \mid \mathbf{x}_i) = \sigma(\mathbf{w}^\top \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}), \quad \sigma(z) = \frac{1}{1+e^{-z}} \tag{3}$$

で与えられ, 正則化係数 C は 3 分割交差検証により選択される. 学習後, 予測確率 $\hat{p}_i$ に基づき, $\hat{p}_i \geq 0.5$ のとき $\hat{y}_i = 1$ として予測ラベルを決定する.

回帰分析の評価. 回帰分析による予測がどれだけ正しいかについて, 信頼度は先行研究 [29] に倣って Mean Absolute Error (MAE; $\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^N |y_i - \hat{y}_i|$) および Mean Relative Error (MRE; $\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^N \left|\frac{y_i-\hat{y}_i}{y_i}\right|$) を用いて評価を行う. y_i は真の値, $\hat{y}_i$ は回帰モデルによる予測値, N は評価対象データ数である. MAE は絶対誤差の平均を測る指標であり, MRE は真の値に対する相対的な誤差を測る指標である. なお, MRE は $y_i = 0$ の場合に定義できないため, 本研究では $y_i \neq 0$ のデータのみを用いて計算した. 正解率に関しては, 正解・不正解の二値変数を目的変数とする Logistic 回帰を用い, モデル性能の評価には, 正解率 ($= \frac{1}{N}\sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i)$), ROC-AUC ($= (\{y_i, \hat{p}_i\}_{i=1}^N)$), 対数誤差 ($= -\frac{1}{N}\sum_{i=1}^N [y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)]$) の 3 つの指標を用いる. 正解率は予測ラベルの一致率を, ROC-AUC は判別性能を, 対数誤差は予測確率の適切さをそれぞれ評価する.

表 3 5 つの QA を解かせた結果 (正解率) とその平均.

モデル	サイズ	MMLU	HellaSwag	Math	RACE	Cosmos	平均
DeepSeek-Qwen	7B	86.5	52.4	77.2	71.6	57.6	62.67
DeepSeek-Llama	8B	79.5	62.2	68.2	77.9	25.2	67.74
gpt-oss	20B	92.2	74.3	78.4	81.9	78.0	78.27
Phi-4	14B	94.5	73.7	79.2	86.8	79.1	79.54

表 4 回帰分析の評価結果. Logistic 回帰の評価には正解率, ROC-AUC, そして対数誤差を, Lasso 回帰の評価には MAE と MRE を用いている. λ は 3 節で記載した Lasso 回帰の正則化係数である.

モデル	n	正解率			信頼度					
		正解率	ROC AUC	対数誤差	生成手法			強制デコード手法		
					λ	MAE	MRE	λ	MAE	MRE
Llama	1	0.6980	0.6120	0.6305	0.0088	0.2606	0.6126	0.0037	0.0488	0.0694
	2	0.6985	0.5834	0.6556	0.0020	0.2538	0.5990	0.0003	0.0453	0.0635
	3	0.6878	0.5819	0.6890	0.0002	0.2404	0.5665	0.0002	0.0453	0.0637
	4	0.6992	0.6379	0.6836	0.0002	0.2565	0.6037	0.0000	0.0437	0.0611
	5	0.6976	0.6359	0.6879	0.0001	0.2713	0.6379	0.0000	0.0441	0.0615
Qwen	1	0.6747	0.6960	0.6213	0.1079	0.2777	0.6527	0.0021	0.0494	0.0681
	2	0.6499	0.6100	0.6744	0.0349	0.2805	0.6589	0.0027	0.0522	0.0718
	3	0.6545	0.5821	0.6892	0.0272	0.2807	0.6595	0.0001	0.0422	0.0574
	4	0.6554	0.6229	0.6813	0.0070	0.2807	0.6593	0.0000	0.0419	0.0567
	5	0.6544	0.6700	0.6824	0.0017	0.2777	0.6519	0.0001	0.0499	0.0681
Phi	1	0.8073	0.6549	0.5111	0.0187	0.1877	0.3995	0.0089	0.0058	0.0063
	2	0.7930	0.5320	0.5738	0.0011	0.1642	0.3477	0.0003	0.0053	0.0058
	3	0.7930	0.6088	0.6192	0.0003	0.1646	0.3456	0.0001	0.0053	0.0057
	4	0.7930	0.5000	0.5993	0.0002	0.1674	0.3481	0.0000	0.0053	0.0056
	5	0.7930	0.5000	0.6019	0.0001	0.1703	0.3523	0.0000	0.0055	0.0059
GPT	1	0.7988	0.5571	0.5802	0.1229	0.3487	0.8895	0.0216	0.0315	0.0459
	2	0.7964	0.5109	0.6227	0.0961	0.3601	0.9221	0.0041	0.0304	0.0447
	3	0.7917	0.5526	0.6768	0.0125	0.3601	0.9220	0.0010	0.0309	0.0451
	4	0.7866	0.5526	0.6842	0.0001	0.3113	0.7964	0.0001	0.0312	0.0453
	5	0.7796	0.5642	0.6906	0.0001	0.3275	0.8390	0.0000	0.0308	0.0450

表 4 に回帰分析の予測を評価した結果を記載する. 正解率評価においては, Phi- が最も高い性能を示し, Accuracy および ROC-AUC が他のモデルを一貫して上回った. また, 対数誤差も最小であり, n-gram に基づく言語特徴が正解・不正解の判別に有効であることが示された. 一方, Llama および Qwen は識別性能が相対的に低く, GPT-OSS はその中間的な性能を示した. 信頼度推定に関しては, Lasso 回帰による評価において, Phi-4 が最も低い誤差を示し, 特に強制デコード手法では MAE $\fallingdotseq$ 0.005, MRE $\fallingdotseq$ 0.005 と推定が正しいことがわかる.