LLM 確信度推定の実証分析：Verbalized Confidence はどの条件で有効か？

石井愛 [1,3]　井之上直也 [1]　鈴木久美 [2]　関根聡 [2]
[1] 北陸先端科学技術大学院大学　[2] 国立情報学研究所　[3]BIPROGY 株式会社
ai.ishii@jaist.ac.jp　naoya-i@jaist.ac.jp　hisamis@nii.ac.jp　sekine@nii.ac.jp

概要

大規模言語モデル（LLM）の確信度推定において，モデルに確信度を自己申告させる Verbalized Confidence（VC）の有効性が報告されているが，その適用範囲は明確でない．本研究では，シングルホップ／マルチホップ QA タスクにおいて VC が較正（Calibration; 確信度と正答率の一致度）および識別（正誤の識別能力）を改善する条件を，出力の一貫性に基づくベースラインとの比較により検証した．その結果，シングルホップでは一部の VC 手法がベースラインを上回り，マルチホップでは Chain-of-Thought（CoT）が有効な場合がある一方，モデルによってはベースラインを下回る結果も観測された．さらに，公開情報に基づく便宜的分類としてポストトレーニングに強化学習（RL）を含むモデル群と含まないモデル群を比較すると，前者で VC による改善が見られやすい傾向があった．

1　はじめに

大規模言語モデル（LLM）の実サービス適用が進むにつれ，LLM が事実と矛盾する情報を生成する「事実性」の問題は依然として重要な課題となっている [1, 2]．この問題に対する解決策の一つが確信度推定であり，モデルの出力に対する確信度を定量化することを目的とする [3]．

確信度推定手法のうち，複数サンプルの一貫性に基づく手法は頑健な較正（Calibration）を示すことが報告されている．Farquhar ら [4] は Semantic Entropy が良好な較正を達成することを示した．一方，Tian ら [5] は，RLHF で学習されたモデルを用いたシングルホップ QA において，Verbalized Confidence（VC）が，複数サンプルの頻度に基づく Label probability（Label prob.）より良好な較正を示すケースを報告した．また，Podolak ら [6] は推論モデルの推論トークン数を制御する実験を行い，推論量を増やすと VC の識別能力が Semantic Entropy の水準に近づくことを報告した．筆者ら [7] は，マルチホップ QA における根拠レベルの確信度評価において，Label prob. が最も頑健であることを示した．これらの知見は，VC の有効性がタスクの複雑性（シングルホップ vs マルチホップ），確信度の粒度（回答レベル vs 根拠レベル），およびモデルの学習手法に依存する可能性を示唆している．

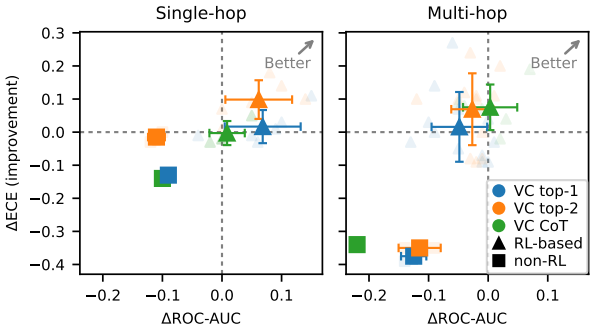

図 1　Label prob. を参照ベースラインとした各手法の相対性能（ΔROC-AUC, ΔECE；右上ほど良い）．ΔROC-AUC は ROC-AUC の差（method − Label prob.），ΔECE は ECE の改善量（Label prob. − method）．マーカーはモデル群（RL-based/non-RL）を示す．薄色は個別条件，濃色は平均，エラーバーは標準偏差．

そこで本研究では，複数生成の一貫性に基づく Label prob. をベースラインとして，回答レベルの確信度において VC が有効となる条件を，複数の LLM を用いシングルホップおよびマルチホップ QA タスクで検証する．検証には筆者らの先行研究 [7] の枠組み[1)]を用いる．図 1 に，Label prob. に対する各手法の相対性能を示す（詳細は §4 参照）．さらに，推論モデルを用いた予備的検証も行い，推論量の効果を比較する．

本研究の貢献は以下の 3 点である．

1. **VC の有効条件の明確化**: RL-based（GPT-4.1 系，

1) https://github.com/aiishii/finegrained_conf

Llama-4）と non-RL（Phi-4）を比較し，VC が Label prob. を上回る条件を ΔECE・ΔROC-AUC の形で整理した．

2. **タスク複雑性と CoT の関係の整理**: シングルホップでは CoT の効果が限定的である一方，マルチホップでは CoT が較正・識別を改善する場合があることを確認した．ただし改善はモデルに依存し，逆効果となる場合も観測された．
3. **VC の限界条件の明確化**: VC はモデルやタスクによっては過信を招き識別性能が低下する場合があり，手法選択では較正（ECE）と識別（ROC-AUC）の両面から評価する必要があることを示した．

2 手法

2.1 確信度取得方法

本研究で比較する確信度推定手法を表 1 に示す．Token prob. は商用 API で取得が制限される場合がある．Label prob. は複数サンプリングを要し計算コストは高いが，API 制約の影響を受けにくい．VC は少ない追加トークンで確信度を得られる一方，高確信側に偏る傾向が報告されている [8, 9]．

表 1 評価対象の確信度推定手法

手法	説明
Label prob.	N=10 サンプル生成し，回答をクラスタリングした出現頻度を確信度とする
Token prob.	回答部分のトークン生成確率の幾何平均
VC top-1	回答と確信度（0–1）を同時出力
VC top-2	回答候補 2 件と各確信度を同時出力
VC CoT	CoT の後に回答と確信度を出力

Label prob. は Wang ら [10] の self-consistency に基づき，temperature=0.7, top-p=0.95 で N=10 回サンプリングし，正規化後に同一回答をクラスタリングして頻度を確信度とする[2]．

VC は Tian ら [5] のプロンプト設計に従い，回答と確信度を同時出力する top-1，top-2，および CoT を評価する．top-2 では上位 2 候補と各確信度を出力させ，最上位候補の確信度を評価に用いた．なお，出力される確信度は確率分布として正規化される保証はない．

プロンプトは筆者ら [7] の文面を踏襲した[3]．

2.2 評価指標

以下の指標を用いて，較正と識別の両面から評価を行う．

較正指標: Expected Calibration Error（**ECE**; [11]）は，予測確信度と実際の正答率の平均絶対差であり，サンプルを確信度に基づき 10 ビン（等間隔）のビンに分割して計算する．Brier Score（**BS**; [12]）は，予測確率と二値の正誤ラベル間の平均二乗誤差である．いずれも値が小さいほど較正が良好である．

識別指標: ROC-AUC[13] は正誤を二値分類として識別する能力を測定する．Selective-AUC[14] は，確信度の高い順にサンプルを選別した際の精度-カバレッジ曲線下面積であり，高確信サンプルの選別精度を評価する．いずれも値が大きいほど良好である．

3 実験設定

3.1 データセット

表 2 に示す各データセットから所定件数を抽出して用いた．マルチホップ QA の平均推論ステップ数は 2WikiMultiHopQA で 2.42，JEMHopQA で 2.06 である．

表 2 データセット

データセット	言語	タイプ	サンプル数
SciQ[15]	EN	シングルホップ	300
2WikiMultiHopQA[16]	EN	マルチホップ	300
JEMHopQA[17]	JP	マルチホップ	1,000

3.2 モデル

評価に用いるモデルを以下に示す．

- **GPT–4.1 系** [18] (ver. 2025–04–14; Dense)
- **Llama-4-Maverick-17B-128E-Instruct-FP8** [19] (SFT + instruction-tuned Mixture-of-Experts with 128 experts)[4]
- **Phi–4** [20] (14B SFT-trained Dense)[5]

モデル群の定義（作業仮説）：本稿では，ポストトレーニングにおける RL ベースの選好最適化（例：RLHF 等）の有無に着目し，公開情報に基づく便宜

2） 回答文字列は，数字・記号（全角/半角など）を正規化し，括弧内表現と句読点を除去し，空白を正規化して小文字化する．日本語の漢数字は算用数字に変換する．正規化後に一致する文字列は同一として統合する．

3） GitHub: aiishii/finegrained_conf, src/finegrained_conf/prompts/answer_prompts.py

4） Azure 内部バージョン 1; 2024/10/1 作成, 2025/5/7 更新

5） Azure 内部バージョン 7; 2024/10/1 作成, 2025/4/16 更新

的な分類としてモデルを 2 群に分けて分析する．なお，本分類は便宜的整理であり，因果的結論を主張するものではない．

RL-based（GPT-4.1 系，Llama-4）は，ポストトレーニングに RL ベースの選好最適化を含む（可能性が高い）モデル群である．

non-RL（Phi-4）は，RL ベースの選好最適化を含まない（または公開情報からは確認できない）モデル群である．

GPT-4.1 系は，GPT-4 が RLHF によって微調整されたことが報告されている一方で，GPT-4.1 の学習手法の詳細は公開資料では明記されていない．本稿では，GPT-4.1 が GPT-4 系列の後継として提供されていることを踏まえ，便宜上，同様の RLHF 等を含む可能性が高いと推定した．Llama-4 についても公開情報にポストトレーニングの詳細が明記されていないが，本稿では便宜上，RL-based（推定）に分類した．Phi-4 は Microsoft 社の技術報告 [20] において SFT のみが明記されている．

実験は Azure AI Foundry 経由で実施した．Label prob. は温度 0.7 で 10 回サンプリングし，VC/Token prob. は温度 0 で生成した．自動評価の設定（判定スキーマ，プロンプト，温度設定など）は筆者らの先行研究 [7] と同一である．なお，Phi-4 の 2WikiMultiHopQA は回答精度が著しく低く確信度推定が不安定であったため評価対象から除外し，GPT-4.1 を対象に追加した．

4 結果

表 3 に主要な評価結果を示す．その他の詳細は付録 表 4 を参照されたい．

シングルホップ QA: SciQ において，RL-based では VC top-2 が最良の ECE を達成した（GPT-4.1-mini: 0.10，Llama-4: 0.07）．VC CoT は効果が限定的であり，Tian ら [5] の知見と整合する．non-RL の Phi-4 では，Label prob. が ECE・ROC ともに最良であり，全ての VC 手法が Label prob. を下回った．Selective-AUC についても概ね ROC-AUC と同様の傾向が見られ，VC の有効性がモデル群に依存することが確認された（付録 表 4）．

マルチホップ QA: 2Wiki および JEMHop において，RL-based では VC CoT または VC top-2 により Label prob. から改善が見られた（Llama-4, 2Wiki: ECE 0.55 → 0.35）．Token prob. が取得可能な GPT-4.1 系では，Token prob. が ECE・ROC ともに最良となる場合があった（2Wiki, GPT-4.1: ECE 0.35, ROC 0.83；付録 表 4）．non-RL の Phi-4 では Label prob. が最良（ECE 0.12，ROC 0.79）であり，VC CoT は逆効果となった（ROC: 0.79 → 0.57）．

表 3 主要結果（ECE↓, ROC↑）．SciQ（シングルホップ），2Wiki（マルチホップ），JEMHop（マルチホップ）．各モデル，データセット，指標内で相対的な良否を色で表示（水色: 良好，橙色: 不良，無色: 中間）し，最良値は太字．表中の ROC は ROC-AUC の略記．

		SciQ (EN)		2Wiki (EN)		JEMHop (JP)	
モデル	手法	ECE↓	ROC↑	ECE↓	ROC↑	ECE↓	ROC↑
RL-based							
GPT-4.1-mini	Label prob.	0.20	0.62	0.50	**0.68**	**0.22**	0.78
	VC top-1	0.23	0.68	0.50	0.59	0.31	0.77
	VC top-2	**0.10**	**0.76**	**0.40**	0.63	0.30	0.77
	VC CoT	0.22	0.62	0.41	0.63	0.24	**0.80**
Llama-4	Label prob.	0.21	0.59	0.55	0.62	0.38	**0.69**
	VC top-1	0.18	**0.69**	0.52	0.52	0.41	0.56
	VC top-2	**0.07**	**0.69**	**0.35**	0.64	**0.27**	0.67
	VC CoT	0.16	0.64	0.36	**0.69**	0.31	0.68
non-RL							
Phi-4	Label prob.	**0.14**	**0.79**	–	–	**0.12**	**0.79**
	VC top-1	0.27	0.70	–	–	0.48	0.68
	VC top-2	0.17	0.69	–	–	0.47	0.70
	VC CoT	0.28	0.67	–	–	0.46	0.57

モデル群別の傾向: 図 1 に示すとおり，シングルホップでは RL-based の VC top-2 が右上象限（Label prob. より優位）に位置する一方，non-RL では VC による改善が見られなかった．マルチホップでは手法間の優劣が条件により異なり，Token prob. が取得可能な場合はそれが有利だが，API 制約下では Label prob. が安定した性能を示した．ただし，non-RL は Phi-4 の 1 モデルに限られるため，この傾向の一般性には追加検証が必要である．

推論モデルにおける推論量の効果（予備的検証）: Podolak ら [6] による，推論モデルの推論トークン数を制御することで VC の識別能力が Semantic Entropy の水準に近づくという知見について，簡易的な設定で検証した．推論モデルとして DeepSeek-R1-0528[6)]（Azure AI Foundry 経由）を用い，Semantic Entropy と同様に複数生成の一貫性に基づく Label prob. を比較対象とした（詳細は付録 B 参照）．

図 2 に示すとおり，推論量（トークン予算）の増加に伴い ROC-AUC は概ね改善するが，中程度の予

6) Podolak らが使用した DeepSeek-R1-Distill-Qwen-32B は，DeepSeek-R1 を教師として蒸留した小型モデルである．本研究では DeepSeek-R1-0528（API 経由）を使用しており，モデル構造が異なるため結果が一致しない可能性がある．

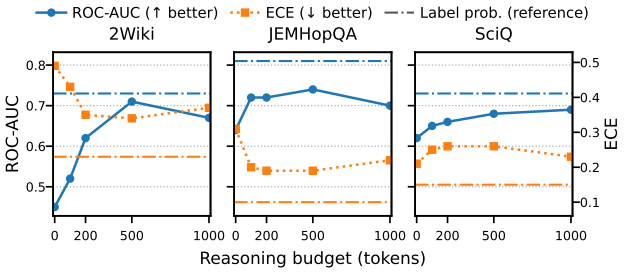

図 2 DeepSeek-R1 の推論量と確信度推定性能．実線が ROC-AUC（↑），点線が ECE（↓）．水平線は Label prob. の参照値を示す．推論量の増加に伴い ROC-AUC は改善傾向を示すが，ECE の改善は限定的である．

算（200–500 トークン）で頭打ちとなるデータセットもあった（2Wiki: 500 トークンで 0.71 → 1000 トークンで 0.67）．また，ECE の改善は限定的であり，推論量の増加が較正の改善に直結しないことを示唆する．

5 考察

本節では，観察された傾向の要因を議論する．

VC の効果とポストトレーニング手法: RL-based（GPT-4.1 系，Llama-4）ではシングルホップで VC による改善が見られた一方，non-RL（Phi-4）では改善が限定的であった．この差は，RL ベースの訓練過程で確信度表現が最適化されていることが一つの可能性として考えられるが，non-RL が Phi-4 のみであるため，一般化には追加検証が必要である．

top-2 の優位性: シングルホップにおいて VC top-2 が top-1 より優れた結果を示した．これは，上位 2 候補を出力させることでモデルに代替案の検討を促し，確信度が適切に分散して過信が抑制されるためと考えられる．実際，top-1 では確信度が高い値に集中する傾向が見られた一方，top-2 では分布がより広がる傾向が観察された．

CoT の効果：シングルホップ vs マルチホップ: Tian ら [5] はシングルホップで CoT が較正を改善しないと報告した．本研究でもシングルホップでは同様の傾向を確認した一方，マルチホップでは異なる結果が得られた．マルチホップでは複数の推論ステップが必要であり，CoT によって推論過程を明示することで較正・識別が改善する傾向が観測された．この差異は，シングルホップでは推論が単純なため追加の言語化が確信度に寄与しにくい一方，マルチホップでは各ステップの不確実性が CoT を通じて確信度に反映されやすくなるためと考えられる．CoT が回答精度を改善することは Wei ら [21] 等で既知であるが，本研究の貢献は，CoT が確信度推定（較正・識別）にも影響すること，およびその効果がモデルやタスクによって正負いずれにもなり得ることを実証した点にある．

non-RL の Phi-4 ではマルチホップで CoT が逆効果となった．回答精度は 0.48 → 0.53 と向上したが，識別能力（ROC-AUC）は 0.79 → 0.57 と低下した．これは，CoT によって確信度が一様に高くなる過信が生じたためと考えられる．Tian ら [5] は RLHF が確信度表現に影響を与えることを報告しており，RL-based モデルでは「不確実なら低い確信度を出す」ことが RL の報酬信号を通じて学習されている可能性がある．一方，non-RL の Phi-4 ではこの能力が獲得されていない可能性があるが，この仮説の検証には RL の有無のみが異なるモデル対（例：ベースモデルと RLHF 後モデル）での比較が必要であり，今後の課題である．

実装上の示唆: Token prob. が取得可能な場合は有力な選択肢となる．API 制約がある場合は Label prob. が安定した性能を示す．VC は，シングルホップかつ RL-based では top-2 が有効であり，マルチホップでは CoT が有効な場合がある．ただし，モデルによっては過信を招くリスクがあるため，手法選択では較正（ECE）と識別（ROC-AUC）の両面から評価することが重要である．

6 おわりに

本研究は，VC の有効性がポストトレーニング手法，タスク複雑性，および CoT の使用に依存しうることを，複数モデル・複数 QA タスクで検証した．検証の結果，RL を含むモデルのシングルホップでは top-2 が ECE と ROC-AUC を同時に改善した．マルチホップでは CoT が有効な場合がある一方，モデルによっては過信を招き識別性能が低下する場合も観測された．これらの結果から，手法選択では較正（ECE）と識別（ROC-AUC）の両面から評価する必要がある．本研究の限界として，RL を含まないモデルの検証が Phi-4 のみに限られており，この傾向の一般性は追加モデルでの検証が必要である．今後は，RL なしの instruction-tuned モデルや RL の有無のみが異なるモデル対での比較を通じて，VC の有効条件とポストトレーニング手法の関係を明らかにしたい．また，マルチホップ QA 以外の複雑な推論タスクへの拡張も検討する．

謝辞

本研究は，JST 創発的研究支援事業 JPMJFR232K，および中島記念国際交流財団の助成を受けたものです．

参考文献

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, Vol. 43, No. 2, p. 1–55, January 2025.

[2] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

[3] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey, 2025.

[4] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. **Nature**, Vol. 630, No. 8017, pp. 625–630, 2024.

[5] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics.

[6] Jakub Podolak and Rajeev Verma. Read your own mind: Reasoning helps surface self-confidence signals in LLMs. In Bryan Eikema, Raúl Vázquez, Jonathan Berant, Marie-Catherine de Marneffe, Barbara Plank, Artem Shelmanov, Swabha Swayamdipta, Jörg Tiedemann, Chrysoula Zerva, and Wilker Aziz, editors, **Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)**, pp. 247–258, Suzhou, China, November 2025. Association for Computational Linguistics.

[7] Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. Fine-grained confidence estimation for spurious correctness detection in llms. In **Proceedings of the 14th International Joint Conference on Natural Language Processing and 4th Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)**, 2025.

[8] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 7752–7764, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2959–2979, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] 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**, September 2022.

[11] 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.

[12] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. **Monthly Weather Review**, Vol. 78, No. 1, pp. 1 – 3, 1950.

[13] Tom Fawcett. An introduction to roc analysis. **Pattern recognition letters**, Vol. 27, No. 8, pp. 861–874, 2006.

[14] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. **Advances in neural information processing systems**, Vol. 30, , 2017.

[15] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, **Proceedings of the 3rd Workshop on Noisy User-generated Text**, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[16] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[17] Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9515–9525, Torino, Italia, May 2024. ELRA and ICCL.

[18] OpenAI. Introducing GPT-4.1 in the API, June 2025. Announces the release of GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano, with major improvements in coding, instruction following, and long context handling.

[19] Meta AI. Llama 4 Maverick 17B-128E Instruct, 2025. Model release date: April 5, 2025. Llama 4 Maverick is a 17B parameter, 128-expert, natively multimodal large language model released under the Llama 4 Community License. Knowledge cutoff: August 2024.

[20] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.

A　結果詳細

表 4　全モデルの評価結果．各モデル，データセット，指標内で相対的な良否を色で表示（水色: 良好，橙色: 不良，無色: 中間）し，最良値は太字．ROC は ROC-AUC，Sel は Selective-AUC の略記．

		SciQ					2Wiki (EN)					JEMHop (JP)				
Model	Method	Acc	ECE↓	BS↓	ROC↑	Sel↑	Acc	ECE↓	BS↓	ROC↑	Sel↑	Acc	ECE↓	BS↓	ROC↑	Sel↑
GPT-4.1	Label prob.	–	–	–	–	–	0.41	0.45	0.40	0.74	0.57	–	–	–	–	–
	Token prob.	–	–	–	–	–	0.42	**0.35**	**0.31**	**0.83**	0.66	–	–	–	–	–
	VC top-1	–	–	–	–	–	0.44	0.48	0.45	0.75	0.61	–	–	–	–	–
	VC top-2	–	–	–	–	–	0.45	**0.35**	0.34	0.72	0.61	–	–	–	–	–
	VC CoT	–	–	–	–	–	0.56	0.37	0.37	0.70	**0.69**	–	–	–	–	–
GPT-4.1 -mini	Label prob.	0.75	0.20	0.21	0.62	0.81	0.32	0.50	0.46	**0.68**	0.41	0.56	0.22	0.23	0.78	0.73
	Token prob.	0.74	0.15	0.19	0.72	0.84	0.32	**0.40**	0.38	0.67	0.45	0.56	**0.16**	**0.19**	**0.83**	0.78
	VC top-1	0.74	0.23	0.24	0.68	0.83	0.36	0.50	0.47	0.59	0.45	0.58	0.31	0.31	0.77	0.78
	VC top-2	0.73	**0.10**	**0.18**	**0.76**	**0.87**	0.32	**0.40**	**0.36**	0.63	0.45	0.59	0.30	0.30	0.77	0.78
	VC CoT	0.75	0.22	0.23	0.62	0.80	0.44	0.41	0.41	0.63	**0.56**	0.63	0.24	0.25	0.80	**0.82**
GPT-4.1 -nano	Label prob.	0.66	0.24	0.26	0.66	0.75	–	–	–	–	–	0.45	0.33	0.31	0.74	0.61
	Token prob.	0.67	**0.11**	**0.22**	**0.67**	**0.77**	–	–	–	–	–	0.46	**0.26**	**0.27**	0.75	0.62
	VC top-1	0.68	0.25	0.27	0.62	0.73	–	–	–	–	–	0.45	0.38	0.38	0.69	0.60
	VC top-2	0.63	0.17	0.24	0.66	0.73	–	–	–	–	–	0.45	0.41	0.39	0.71	0.63
	VC CoT	0.67	0.27	0.28	0.64	0.75	–	–	–	–	–	0.49	0.29	0.29	**0.77**	**0.69**
Llama-4	Label prob.	0.75	0.21	0.22	0.59	0.77	0.33	0.55	0.52	0.62	0.39	0.51	0.38	0.36	**0.69**	0.61
	VC top-1	0.74	0.18	0.21	**0.69**	0.84	0.32	0.52	0.50	0.52	0.35	0.53	0.41	0.42	0.56	0.54
	VC top-2	0.74	**0.07**	**0.18**	**0.69**	**0.86**	0.34	**0.35**	**0.33**	0.64	0.42	0.50	**0.27**	**0.30**	0.67	0.65
	VC CoT	0.75	0.16	0.21	0.64	0.84	0.49	0.36	0.36	**0.69**	**0.63**	0.61	0.31	0.31	0.68	**0.70**
Phi-4	Label prob.	0.67	**0.14**	**0.19**	**0.79**	**0.83**	–	–	–	–	–	0.48	**0.12**	**0.21**	**0.79**	**0.67**
	VC top-1	0.69	0.27	0.28	0.70	0.80	–	–	–	–	–	0.49	0.48	0.47	0.68	0.59
	VC top-2	0.71	0.17	0.22	0.67	0.79	–	–	–	–	–	0.46	0.47	0.45	0.70	0.61
	VC CoT	0.68	0.28	0.28	0.69	0.79	–	–	–	–	–	0.53	0.46	0.45	0.57	0.55

B　DeepSeek-R1 予備的検証の詳細

§4 で示した図 2 の DeepSeek-R1-0528 における確信度推定性能の詳細値を表 5 に示す．また，推論量を制御する実装の詳細を以下に記す．本検証は Podolak ら [6] の推論トークン予算を制御する設定を，API 利用環境で可能な範囲で近似したものである．

実装の詳細（budgeted_N）：`budgeted_N` では推論出力をストリーミング受信し，目標予算 N 到達時点で打ち切り後，回答と確信度を出力させる二段階方式（Turn 0，Turn 1）を用いた．Turn 0 では `stream=True` で推論用プロンプトを投げ，受信した出力を逐次連結してトークナイザにより推論トークン数を推定する．推定値が目標予算 N に到達した時点でクライアント側で受信を打ち切り，得られた推論断片を会話履歴として保持する．自然終了（モデルが自発的に停止）した場合はそのまま採用する．Turn 1 では Turn 0 の断片を履歴に含めた上で最終出力のみを要求し，`[OUTPUT]` の先頭出力を `prefix` で強制することで，`<think>`等の推論出力が混入する失敗を抑制した．最終出力は `[/OUTPUT]` で停止させた（`stop=["[/OUTPUT]"]`）．温度は Turn 0/Turn 1 ともに `temperature=0.0` とした．

予算値と実測トークン：ストリーミングのチャンク単位で打ち切るため，実際の推論トークン数は目標予算 N と厳密一致しない（表に実際のトークン数を併記）．また，`baseline`（推論なし）は Turn 0 を省略し，`[OUTPUT]` のみを 1 回で出力させるベースラインとして実装した．

表 5　DeepSeek-R1: 推論量と確信度推定性能．最良値は太字．ROC は ROC-AUC，Sel は Selective-AUC の略記．

データ	手法	トークン	ECE↓	ROC↑	Sel↑
2Wiki	baseline	0	0.49	0.45	0.29
	budgeted_100	99	0.43	0.52	0.36
	budgeted_200	199	0.35	0.62	0.56
	budgeted_500	497	0.34	0.71	0.62
	budgeted_1000	906	0.37	0.67	0.63
	Label prob.	3,368	**0.23**	**0.73**	**0.68**
JEMHop	baseline	0	0.31	0.64	0.71
	budgeted_100	100	0.20	0.72	0.79
	budgeted_200	200	0.19	0.72	0.80
	budgeted_500	500	0.19	0.74	0.85
	budgeted_1000	929	0.22	0.70	0.82
	Label prob.	7,714	**0.10**	**0.81**	**0.89**
SciQ	baseline	0	0.21	0.62	0.82
	budgeted_100	99	0.25	0.65	0.81
	budgeted_200	199	0.26	0.66	0.81
	budgeted_500	492	0.26	0.68	0.81
	budgeted_1000	786	0.23	0.69	0.83
	Label prob.	3,125	**0.15**	**0.73**	**0.86**