文脈内学習における文脈内事例の寄与度推定

葉夢宇 [1] 栗林樹生 [2] 小林悟郎 [1,3] 鈴木潤 [1,3]
[1] 東北大学 [2]MBZUAI [3] 理化学研究所
{ye.mengyu.s1, goro.koba}@dc.tohoku.ac.jp
tatsuki.kuribayashi@mbzuai.ac.ae jun.suzuki@tohoku.ac.jp

概要

モデルの出力を説明する一つの方針として，出力に寄与した訓練事例の提示が考えられる．近年成功を収めている文脈内学習では，通常の学習と異なり訓練事例が入力に含まれるため，入力に対する寄与度推定法をこの目的で適用できる可能性がある．本研究では，重要な事例が明らかになるよう設計した人工タスクを用いて，6 種類の代表的な解釈手法の適用可能性を検証する．実験により，Gradient Norm 以外の解釈手法は，文脈内学習における寄与事例推定に向いていないことが経験的に明らかになった．

1 はじめに

ニューラルモデルの出力を説明する主なアプローチとして，(i) どの訓練事例が出力に寄与したのか [1, 2] と (ii) 入力のどこが出力に寄与したのか [3, 4] が研究されてきた．近年の大規模言語モデルに対してもそれぞれのアプローチで研究が行われている [5, 6, 7]．大規模言語モデルは少数の具体例（**少数事例**）と共に解きたい事例（目標事例と呼ぶ）を入力してタスクを解かせる**文脈内学習（in-context learning）** [8] が成功しているが，少数事例のうちどれが出力に影響を与えたのかという**寄与度**の推定方法について研究は限られている．

文脈内学習の場合は訓練事例（少数事例）が入力に文脈として与えられるため，入力の寄与を考える (ii) の道具立てを (i) の訓練事例の解釈に適用することが考えられる．ただし，このような技術適用がうまく機能するかは非自明である．入力中の各単語の寄与度を推定する際，典型的には目標事例とモデル出力のペア (x, y) の二項関係（目標事例 x のどこが y に寄与したか）が考慮されてきた．一方で文脈内学習では，少数事例 $(x_1, y_1), (x_2, y_2) \cdots (x_{n-1}, y_{n-1})$ および目標事例 x_n とモデルの出力 y の間の高次の関係を推定する必要があるため，二項関係をもとに提案された手法が機能する保証がない．また，今回は特定の入力トークンの寄与ではなく，事例単位での寄与を推定するため，どのようにトークン単位の寄与を集約すれば良いのかといった論点もある．

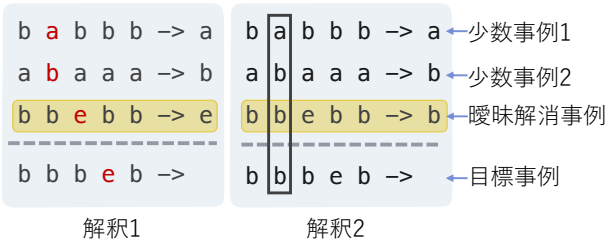

図 1 本研究に用いたタスクの例．曖昧性解消事例がない場合，入力中の唯一異なる記号を出力する問題なのか（解釈 1），特定の位置の記号を出力する問題なのか（解釈 2）解釈が絞られない．従って曖昧性解消事例は目標事例に正答するために考慮しないといけない事例であり，既存の寄与度解釈手法が曖昧性解消事例に最も高い寄与度を与えられるかどうかを検証する．

本研究では，少数事例の内，目標事例の回答に必要な事例が自明である人工的な問題設定を導入し，既存の 6 つの入力寄与度推定手法について，文脈内事例の寄与度推定能力を評価する．検証の結果，既存のベンチマークでうまく解釈できていた特定の推定法がこの問題設定においてうまく機能しないことや，逆にこれまで推定能力が低いとされてきた Gradient Norm が，文脈内学習の解釈において顕著に高い性能を示すことが経験的に明らかになった．

2 問題設定

少数事例のうち，出力に寄与する事例が明確に定まる問題設定を導入する．具体的には，特定の事例（曖昧性解消事例と呼ぶ）を抜いた場合に，問題の定義が曖昧になる設定を考える．例えば図 1 は記号列が入力されて 1 つの記号を出力とするタスクであり，曖昧性解消事例が存在しない場合，入力中で唯一異なる記号を出力するべきなのか（解釈 1），2 番目に出現した記号を出力するべきなのか（解釈

2）が曖昧になる．したがって目標事例に正答した場合，曖昧性解消事例が出力に必要なことは，タスクの定義上明確になっている．また，単に目標事例と文字列的な意味で重複度の高い事例を取得する方法では，適切な解釈ができないことを確認しており，表層的な手がかりのみでは解釈できない困難な設定である．[1)]

具体的に，各事例は入力 x と正解 y の組 (x, y) で定義される．入力 x は長さ 5 の文字列 $n^k mn^{4-k}$ からなる．ただし，$n, m \in \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}, \mathsf{e}\}$，$m \neq n$，$k \in \{1, 2, 3\}$ である（例：bbbeb）．次に入力 x を出力 y に対応付ける関数 $f : x \mapsto y$ を二種類考える．ひとつは，入力中の唯一異なるアルファベットを出力する関数（解釈 1）であり，もうひとつは，入力 x 中の両端を除いた特定の位置のアルファベットを出力する関数（解釈 2）である（図 1）．ここで，k が同一である事例 (x, y) を 2 つと，それらとは異なる k をもつ事例（曖昧性解消事例）を 1 つ抽出し，少数事例組 $s = [(x_1, y_1), (x_2, y_2), (x_3, y_3)]$ を構成する．ただし，3 事例の内，どれを曖昧性解消事例とするかは無作為に決める．また各少数事例組 s_i 内の全ての事例について，前述した 2 つの関数のうちどちらかが一貫して適用されている．[2)] 各少数事例組 s_i に対して，どの 3 事例とも合致しない k をもつ目標事例 (x_4, y_4) を設定し，x_4 から y_4 を回答できるのか調査する．ただし，x_4 から y_4 への変換は，対応する少数事例組と同一のものが使用されており，モデルは少数事例組から用いられている規則を推測する必要がある．着目すべき点として，少数事例組 s 内の曖昧性解消事例を除くと，関数 f が一意に絞られず（図 1），目標事例に正答することが困難になる．したがって，モデルが目標事例に正解した場合，曖昧性解消事例に対する寄与度が最も高く割り振られるべきであり，各解釈手法がそのような傾向を示すか調査する．

3 入力寄与度推定手法

本研究では，出力に対して入力のどこが寄与したかを推定する手法（入力寄与度推定手法）として代表的な 6 種類を対象とする．

1） 今回の実験において，編集距離に基づく寄与事例推定法では，チャンスレートレベルでしか望んだ事例が取得できないことを確認している．

2） 解釈 2 が選ばれた場合，t は少数事例組と対応する目標事例内で一貫する．

3.1 ベース手法

Gradient Norm： Gradient Norm [9, 10] は，モデル出力（ロジットなど）に対する各入力要素の**勾配**を計算し，そのノルムに基づき寄与度を計算する．言語モデルでは，出力（単語）y_t に対する各入力単語 x_i の勾配は以下のように計算される：

$$g(x_i) = \nabla_{x_i} q(y_t \mid \boldsymbol{x}) \tag{1}$$

ここで $q(y_t \mid \boldsymbol{x})$ はモデルに入力単語埋め込み列 $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots \boldsymbol{x}_n]$ を与えた時の出力単語 y_t へのロジットである．既存研究 [4] に従い，この勾配の L1 ノルムを計算することでスカラー値の寄与度を得る：

$$S_{GN}(x_i) = \|g(x_i)\|_{L1} \tag{2}$$

Gradient × Input： Gradient × Input [11, 12] は勾配の L1 ノルムを計算する Gradient Norm とは異なり，勾配と入力埋め込み $\boldsymbol{x}_i$ の内積によって寄与度を計算する：

$$S_{GI} = g(x_i) \cdot \boldsymbol{x}_i \tag{3}$$

Input Erasure： Input Erasure は入力から特定要素を消去することによる出力への影響から寄与度を測定する [13]．通常の入力単語埋め込み列 $\boldsymbol{x}$ と，特定単語の埋め込み $\boldsymbol{x}_i$ をゼロベクトルに置換した入力単語埋め込み列 $\boldsymbol{x}_{\neg i}$ の間での，出力単語 y_t へのロジットの差分によって単語 x_i の寄与度を計算する：

$$S_{IE}(x_i) = q(y_t \mid \boldsymbol{x}) - q(y_t \mid \boldsymbol{x}_{\neg i}) \tag{4}$$

3.2 対照的解釈手法

Yin と Neubig ら [4] は各入力単語の寄与度を計算する際に，実際にモデルが出力した単語 y_t と，それとは対立する他の単語（対照単語）y_f の間で比較を行う対照的な解釈手法を提案し，その有効性を報告している．例えば図 1 左においては，モデルはなぜ b（解釈 2）ではなく e（解釈 1）を出力したのかを説明するように寄与を計算する．

前述の 3 種類のベース手法それぞれに対して対照的解釈手法を考える．本研究では，対照単語 y_f は曖昧性解消事例によって定まる解釈と異なる解釈で回答したときの出力とする．つまり，各入力トークン x_i が誤った解釈に基づく回答 y_f を抑制し，正しい解釈の回答 y_t を促進することにどれほど寄与したかを計算する．

Contrastive Gradient Norm： Gradient Norm において，勾配計算の対象を出力単語へのロジット

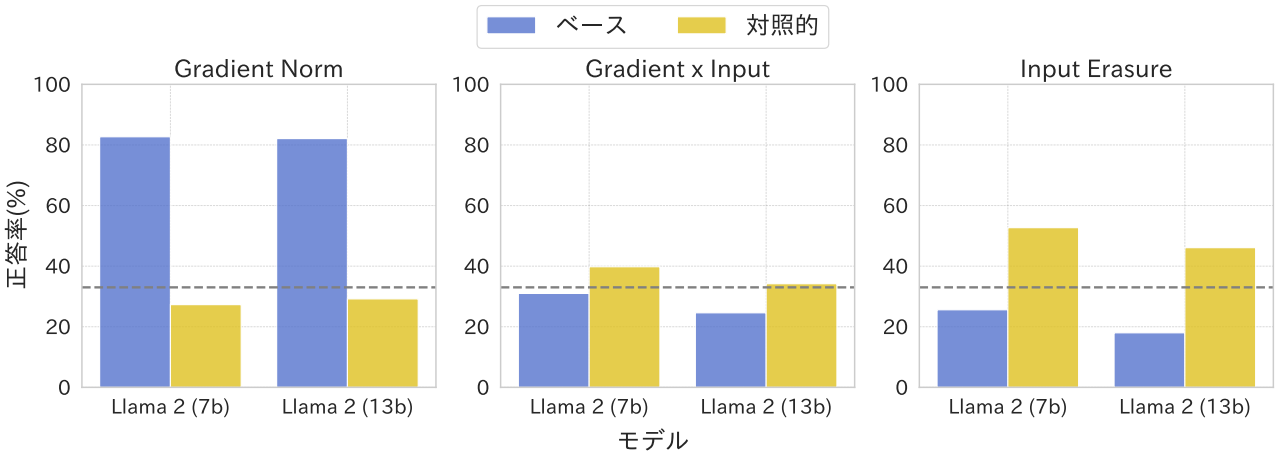

図 2　Llama 2 (7B) と Llama 2 (13B) における各解釈手法の正解率．青色はベース手法，黄色は対照的解釈手法を表す．破線はチャンスレートを表す．他の手法と比べて，ベースの Gradient Norm は顕著に高い性能を示す．

表 1　予備実験におけるモデルの正解率

モデル	3 事例	2 事例	
		曖昧性解消事例抜き	曖昧性を解消しない事例抜き
Llama 2 (7B)	66.2	47.2	65.4
Llama 2 (13B)	65.7	48.5	66.5

$q(y_t \mid \boldsymbol{x})$ から，出力単語と対照単語のロジットの差分 $q(y_t \mid \boldsymbol{x}) - q(y_f \mid \boldsymbol{x})$ にして寄与を計算する：

$$g^*(x_i) = \nabla_{x_i}\Big(q(y_t \mid \boldsymbol{x}) - q(y_f \mid \boldsymbol{x})\Big) \tag{5}$$

$$S^*_{GN}(x_i) = \|g^*(x_i)\|_{L1} \tag{6}$$

Contrastive Gradient × Input：　Gradient × Input でも同様に勾配計算の対象を変更して計算する：

$$S^*_{GI} = g^*(x_i) \cdot x_i \tag{7}$$

Contrastive Input Erasure：　Input Erasure でも出力単語 y_t と対照単語 y_f の差分に対して計算する：

$$S^*_{IE}(x_i) = \Big(q(y_t \mid \boldsymbol{x}) - q(y_t \mid \boldsymbol{x}_{\neg i})\Big) - \Big(q(y_f \mid \boldsymbol{x}) - q(y_f \mid \boldsymbol{x}_{\neg i})\Big) \tag{8}$$

4　実験設定

モデル：　実験では，Llama 2 のうち，パラメータサイズが 7B と 13B のものを評価対象とした[3]．

データセット：　実験用のデータセットは 600 問からなり，解釈 1 と解釈 2 の問題がそれぞれ 300 問ずつ占めている．なおそれぞれの問題において，少数事例 3 つのうちの曖昧性解消事例の位置は均等になっている．実験では，600 問のうちモデルが正答した問題のみ（Llama 2 (7B)：397 問，Llama 2 (13B)：394 問）を用いた．

予備実験：曖昧性解消事例の重要性　表 1 に少数事例から曖昧性解消事例を抜いた時のモデルの正答率の変化を示す．少数事例 3 つから曖昧性解消事例を抜くとどのモデルも正答率が大きく下がり，曖昧非解消事例を抜いても正答率がほぼ変わらないことを示した．すなわち，曖昧性解消事例は問題を正しく解くための必要不可欠であることを示唆した．なお，モデルの解釈別の正答率および曖昧性解消事例を抜いた時のモデルの回答傾向を付録 A に示す．

5　実験：分析手法の正解率

少数事例 e_1, e_2, e_3 に対し，各事例に含まれるトークンの寄与度の合計をその事例の寄与度と定義する．例えば Gradient Norm では，事例 e_n の寄与度は以下のように定義される：

$$S_{GN(e_n)} = \sum_{i=1}^{l_n} S_{GN}\Big(x_{(n,i)}\Big) \tag{9}$$

ここで，$n \in \{1, 2, 3\}$ は少数事例の番号であり，$x_{(n,i)}$ は各事例 e_n 内においての i 番目の入力トークンである．他の手法においても，同様に計算される．また，前述したように，各少数事例の長さは $l_n = 5$ で一定である．[4] この合計寄与度スコアが，曖昧性解消事例に最も高く付与された場合を正解とみなし，問題全体での平均的な正解率を報告する．

3）　2023 年 7 月 19 日に Meta AI から使用許可を得て https://huggingface.co/meta-llama/Llama-2-7b-hf および https://huggingface.co/meta-llama/Llama-2-13b-hf を用いた．

4）　なお事例の長さによるバイアスを排除するため，トークナイザによって自動的に先頭に追加される BOS トークン<s>は事例寄与度の和をとる段階で計算から除外した．

手法の正解率： 図 2 に三つのベース手法と三つの対照的解釈手法を検証した結果を示す．また，破線は 3 つある事例から 1 つをランダムに選択する時のチャンスレートを示す．多くの手法はチャンスレートである 33%に近い正解率に留まり，既存の解釈手法が文脈内学習をうまく解釈できていない可能性を示唆している．一方，最もシンプルなベース手法の一つである Gradient Norm は，80%以上の高い正解率で正しい解釈事例を特定できることが示された．

曖昧性解消事例の分布の影響： モデルの出力にバイアスは見られなかったが，入力寄与度推定手法には大きなバイアスが存在することが観察された．図 3 には，Llama 2 (13B) で各手法が一番目と三番目の事例に最高寄与度を推定した事例数の割合が示されている．破線はチャンスレートを示す．ここで，Gradient Norm の対照的手法は，事例 1 に 100%で最高寄与度を推定したことが観察された．また，Input Erasure のベース手法は，事例 3 に全く最高寄与度を推定できなかったことも観察された．これは，曖昧性解消事例の分布が手法の解釈性に大きく影響を与える可能性を示唆している．また，この曖昧性解消事例の分布によるバイアスは，入力寄与度推定手法に依存する可能性も示されている．なお，モデルが正解した問題について，曖昧性解消位置の偏りはほとんどない（付録 A）．

6 関連研究

大規模言語モデルの解釈： パラメータサイズが小さい言語モデルの内部動作に関する研究は，多く行われている [14, 15]．また，言語モデルの入力から出力までのプロセスを解釈する試みも見られる [3]．しかし，大規模言語モデルの内部構造は著しく複雑であり，入力も長文化しているため，これまでの方法での分析の可能性は自明ではない．

本研究では，大規模言語モデルを既存の手法で解釈することを試みた．この研究の成果を通じて，既存の手法の適用範囲の解像度を高め，新たな手法が考えるきっかけとなることが期待される．

文脈内学習： 少数事例を用いた文脈内学習は，モデルのパラメータを更新することなく新たなタスクに対して高い性能をもたらすため，訓練コストが高い大規模言語モデルにおいて注目を集めている [8, 16, 17]．多くの既存研究は文脈内学習のメカニズムの解明に焦点を当てているが [18, 19, 20]，少数事例と出力との関連性についての分析は十分に行われていなかった．

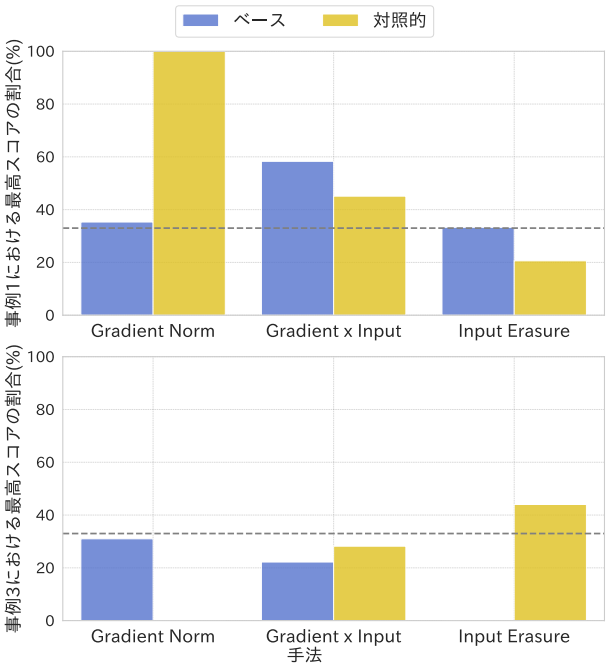

図 3 Llama 2 (13B) において，各手法が事例 1 と事例 3 に最高寄与度を推定した事例数の割合．上が事例 1，下が事例 3 の場合を示す．青色はベース手法，黄色は対照的解釈手法を表す．破線はチャンスレートを表す．割合が 0 か 100 という極端な事例を観察できる．

本研究の結果から，文脈内学習には既存の出力解釈手法がうまく機能しないことを示した．

7 おわりに

本稿は，大規模言語モデルが文脈内学習に既存の寄与度分析手法が適用できるかどうかを調査した．文脈内学習の際に使われている少数事例の中，重要な事例が明らかになるよう設計した人工タスクを用いて実験を行った結果，調査した 6 種類の手法のうち，5 種類がチャンスレート付近の正解率に留まり，文脈内学習の分析に不適切であることを観察した．また，調査した手法のうち，これまで推定能力が低いとされてきた Gradient Norm は 80%以上の正解率で妥当な事例を特定できた．今回の研究は非常に経験的なものであり，なぜこのような結果が得られたのかを理論的な理解を進める必要がある．今後は，文脈内学習における少数事例および目標事例とモデルの出力の間の高次関係を深掘りしていく．また，少数事例の数を増やし，曖昧性解消事例の出現位置によるバイアスを調べる方向性も興味深い．

謝辞

本研究は，JSPS 科研費 JP21H04901，JP22J21492，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research) の助成を受けて実施されたものである.

参考文献

[1] Xiaofei Sun, Diyi Yang, Xiaoya Li, and et al. Interpreting deep learning models in natural language processing: A review. **CoRR**, Vol. abs/2110.10470, , 2021.

[2] Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. Neuron-level interpretation of deep NLP models: A survey. **Trans. Assoc. Comput. Linguistics**, Vol. 10, pp. 1285–1303, 2022.

[3] Eric Wallace, Jens Tuyls, Junlin Wang, and et al. AllenNLP interpret: A framework for explaining predictions of NLP models. In Sebastian Padó and Ruihong Huang, 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): System Demonstrations**, pp. 7–12, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 184–198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[5] Roger B. Grosse, Juhan Bae, Cem Anil, and et al. Studying large language model generalization with influence functions. **CoRR**, Vol. abs/2308.03296, , 2023.

[6] Kelvin Guu, Albert Webson, Ellie Pavlick, and et al. Simfluence: Modeling the influence of individual training examples by simulating training runs. **CoRR**, Vol. abs/2303.08114, , 2023.

[7] Vivek Miglani, Oliver Aobo Yang, Aram H. Markosyan, Diego García-Olano, and Narine Kokhlikyan. Using captum to explain generative language models. **CoRR**, Vol. abs/2312.05491, , 2023.

[8] Tom B. Brown, et al. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[9] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, **2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings**, 2014.

[10] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 681–691, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. **CoRR**, Vol. abs/1605.01713, , 2016.

[12] Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of salient sentences from labelled documents. **CoRR**, Vol. abs/1412.6815, , 2014.

[13] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. **CoRR**, Vol. abs/1612.08220, , 2016.

[14] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Elena Voita, Rico Sennrich, and Ivan Titov. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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. 4396–4406, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Hugo Touvron, Louis Martin, Kevin Stone, and et al. Llama 2: Open foundation and fine-tuned chat models. **CoRR**, Vol. abs/2307.09288, , 2023.

[17] OpenAI. GPT-4 technical report. **CoRR**, Vol. abs/2303.08774, , 2023.

[18] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In **NeurIPS**, 2022.

[19] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In **The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023**. OpenReview.net, 2023.

[20] Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, and et al. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 35151–35174. PMLR, 2023.

A　モデルの正答率の詳細

表 2 にモデルの解釈別の正答率の詳細を示す．また，少数事例 3 つから曖昧性解消事例を抜いた時のモデルの出力傾向も示す．解釈 1 にでも解釈 2 にでもない属しない出力も存在するため，解釈 1 と解釈 2 の事例数の和が問題総数の 600 になっていない．

表 2　モデルの正解率

モデル	推論タイプ	正解率（%）	曖昧性解消事例 e_i の位置別正解率（%）			曖昧性解消事例抜きの解釈数
			e_i = 事例 1	e_i = 事例 2	e_i = 事例 3	
Llama 2 (7B)	解釈 1	56.3	59.0	56.0	54.0	300
	解釈 2	76.0	65.0	80.0	83.0	266
Llama 2 (13B)	解釈 1	53.3	47.0	59.0	54.0	282
	解釈 2	78.0	74.0	84.0	76.0	300