大規模言語モデル内における文法回路と計算回路の重なり

木迫 璃玖 [1]　栗林 樹生 [2]　笹野 遼平 [1]
[1] 名古屋大学　[2] MBZUAI
kisako.riku.n3@s.mail.nagoya-u.ac.jp
tatsuki.kuribayashi@mbzuai.ac.ae　sasano@i.nagoya-u.ac.jp

概要

大規模言語モデル（LLM）は，文法性判断能力といった狭義の言語能力を超えて，計算や社会的推論といった言語を通して行われる様々な処理能力も獲得している．人間を対象とする神経科学領域では，このような形式的言語能力と機能的言語能力が脳内の共通の領域で行われているのかが議論・検証されてきた．そのような視点から，本研究では，LLM 内部における両能力に対応する回路の重なりを，ニューラルモデルの回路分析手法を用いて調査する．形式的言語課題として文法性判断や構文解析課題を，機能的言語能力として計算課題に焦点を当て，統制された実験の結果から，言語モデル内部において，計算を司る回路自体は形式的な言語処理とは分離された回路として埋め込まれていることが示唆された．

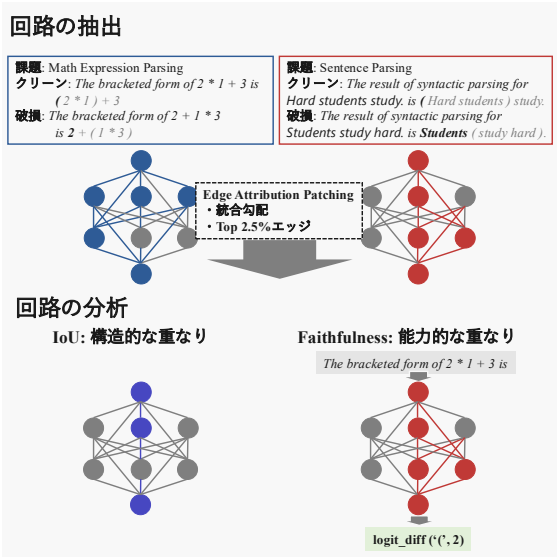


図 1　回路分析の概観．各課題において，正解となる次の単語が異なるペアデータを生成し，回路分析手法を用いてペア条件間のロジット差に寄与する課題の固有の回路を同定する．次にこれらの課題間の回路の重複部分を分析する．

1　はじめに

言語モデルは，文法的に正しい文章の生成といった狭義の言語能力（形式的言語能力）を超えて，論理推論のような必ずしも自然言語を必要としない能力（機能的言語能力）までも発揮している [1]．このような一見領域の異なる能力が同一のモデル内においてどのように組み込まれているかといった問いは，科学的な関心から，言語モデル解釈可能性の分野で近年扱われている [2, 3]．また神経科学分野では，人間の脳処理に関しても，機能局在性の観点から同様の問いが提起されており，脳機能イメージング技術や神経活動計測技術等を駆使して研究されている．例えば fMRI を用いたいくつかの研究では，文法性に対して反応する脳領域が，計算や推論，音楽等に反応する領域とは異なっていることが主張されている [4, 5, 6]．

本研究では数量を機能的言語能力の例として扱い，計算を含む入力が，計算を含まない入力と比較してどのようにモデル内部で処理されているかを分析する．両領域が干渉しているかもしれないと考える動機としては，例えば (The cute cat) sleeps と (2*3)+1，The cat (deeply sleeps) と 2+(3*1) といった対比のように，文・数式はともに階層構造を持つことが挙げられる．また経験的に，計算式やコードで事前学習したモデルが，自然言語処理課題で高い性能を示すことも知られており，両処理能力が干渉している可能性が示唆されている [7, 8, 9]．

実験では，文法性判断課題から計算課題まで複数の課題を設定し，言語モデルの解釈可能性領域で用いられる回路分析手法を用いてモデル内部における回路の重なりを比較する (図 1)．とりわけ，計算課題や，同様の計算式を単にスペルアウトする課題，計算式の構文解析課題，計算式と同じ構造の自然言語文に対する構文解析課題など，入力の語彙分布や課題の複雑さを統制した課題を段階的に設定し，回

路の重なりを調査する．それぞれの課題で同定された回路の重なりから，計算を含む課題が，形式的言語処理とは分離された回路として埋め込まれていることが示された．したがって，言語モデルの場合において，形式的言語処理と計算は，近年の人間に対する研究と同様に，回路・領域として自ずと分離されている可能性が示唆される．

2 準備：言語モデルの回路分析

LLM が特定の課題を実行するための領域の特定は，モデルの解釈可能性の分野で取り組まれてきた．過去の研究では，パラメータ空間での局所化のためにバイナリマスクを学習する手法 [10, 11] や，モデルの中間層のニューロンを抑制・活性化する方法 [12] などが取られてきた．近年では，回路分析手法が用いられている [13, 14]．Transformer アーキテクチャ [15] の LLM における回路分析とは，モデルを，入力から始まり中間処理要素を経由して出力に至る有向グラフとしてみなし，特定の課題を解くために必要な部分グラフを特定する手法である．ここで，グラフのノードは，モデルのモジュール（個別のアテンションヘッドや，MLP），エッジはそれらの接続に対応し，回路とはその課題におけるモデルの動作特性を忠実に保持できる，必要最低限のサブグラフとなる [16].

2.1 分析手法

本研究では，回路分析手法の 1 つである EAP-IG を用いる [17]．EAP-IG では，各有向エッジ $i \to j$ に対して**間接効果スコア**を算出する．このスコアを計算するために，まず各課題においてペアとなる入力 $(x, \tilde{x})$ を構築する．ここで x はクリーンな入力，$\tilde{x}$ は破損した入力を指す．破損入力 $\tilde{\boldsymbol{x}}$ は，元の入力 $\boldsymbol{x}$ における正しい出力が誤りとなるように設計されている．例えば，正しい入力が *"The keys on the cabinet [are],"* である場合，破損入力は *"The key on the cabinet [is]."* のようにする．ノード i における対応する表現（入力の最終単語に対応）を $\boldsymbol{h}_i(x) \in \mathbb{R}^n$，トークン表現を $\boldsymbol{x} \in \mathbb{R}^m$ とそれぞれ表記する．微分可能な評価指標 $m(\cdot) : \mathbb{R}^m \to \mathbb{R}$（通常はロジット値）について考える．評価指標 m は，例えば主語動詞の数一致課題において，単数形と複数形の動詞間のロジット差として定義される：$m(\boldsymbol{x}) = \mathrm{logit}(\mathrm{are}|\boldsymbol{x}) - \mathrm{logit}(\mathrm{is}|\boldsymbol{x})$．入力トークン表現 $\boldsymbol{x}$ に対するノード i での勾配を $\nabla_{\mathbf{h}_i} m(\boldsymbol{x}) \in \mathbb{R}^n$ と表する．この勾配は，ノード i における補間経路 $\Delta\boldsymbol{h}_i = \boldsymbol{h}_i(x) - \boldsymbol{h}_i(\tilde{x}) \in \mathbb{R}^n$ に沿って積分される．経路 $\Delta\boldsymbol{h}_i$ を T ステップで離散化することで，間接効果スコアを以下のように近似する：

$$s_{i\to j} = \Delta\boldsymbol{h}_i \int_{\alpha=0}^{1} \nabla_{\boldsymbol{h}_j} m(\tilde{\boldsymbol{x}} + \alpha(\boldsymbol{x} - \tilde{\boldsymbol{x}})) \quad , \tag{1}$$

$$\approx \frac{\Delta\boldsymbol{h}_i}{T} \sum_{k=0}^{T-1} \nabla_{\boldsymbol{h}_j} m\Big(\tilde{\boldsymbol{x}} + \tfrac{k}{T}(\boldsymbol{x} - \tilde{\boldsymbol{x}})\Big) \quad , \tag{2}$$

ここで，各ステップ k において，対応する勾配（$\in \mathbb{R}^n$）と離散化された経路（$\in \mathbb{R}^n$）の内積を計算し，それらを累計する．上位スコアのエッジを選択することで回路 E を抽出する．実験では，全エッジの上位 2.5% を保持する．その後，特定の課題 A と B に対して，抽出された回路 E_A と E_B の重複度を 2 つの指標で測る．

2.2 評価指標

IoU (Intersection over Union) 構造的な回路の重なりを比較するために IoU を用いる．IoU は次の式で表される．

$$\mathrm{IoU}(A, B) = \frac{|E_A \cap E_B|}{|E_A \cup E_B|} \quad . \tag{3}$$

ここで E_A, E_B はそれぞれ課題 A, B の回路に含まれるエッジの集合である．

Faithfulness 構造的な重複がわずかであっても，回路の多義性によりある課題用に抽出された回路が別の課題にも適用可能となる場合がある（能力的な重なり）．ある課題 A を別の課題 B の回路でどの程度解けるかによっても，両課題の回路の重なりを定量化する．x^A と $\tilde{x}^A$ をそれぞれ課題 A のクリーン入力と破損入力とし，$E_B \subseteq \mathscr{E}$ を課題 B 向けに同定した回路とする．各有向エッジ $e \in \mathscr{E}$ について，$a_e(x)$ はモデルを入力 x で実行する際の活性化値とする．E_B を用いて課題 A に対して行う「課題横断型の回路適用」を以下のように定義する：モデルをクリーン入力 x^A で実行しつつ，E_B の外部にあるすべてのエッジの活性化を，破損入力 $\tilde{x}^A$ に対する実行時の活性化値で置き換え，E_B 内のエッジはそのまま（すなわちクリーンな状態）とする．具体的には，混合エッジ活性化を以下のように構築する：

$$a_e^{(A\leftarrow B)} = \begin{cases} a_e(x^A) & \text{if } e \in E_B \quad , \\ a_e(\tilde{x}^A) & \text{if } e \notin E_B \quad , \end{cases} \tag{4}$$

そして，この混合フォワードパス（エッジパッチ適用時）において課題 A を評価し，評価スコア $m_B^{A\text{-clean}}$ を得る．課題 A に対する標準のクリーン実行時ス

表 1　実験に用いたデータの例．出力に関して，実際は一単語目のみを出力しており，二単語目以降は灰色で示す．計算列に ✓ が着いているものを計算関連課題，ついていないものを言語処理課題として扱う．

課題名	入力	出力	分野	計算	出力種別
計算	The answer to '2 * 1 + 3 =' is	5	数式	✓	アラビア数字
計算（書き下し式）	The answer to 'two times one plus three equals' is	five	数式	✓	英単語（数字）
書き下し	The spellout version '2 * 1 + 3 =' is	two times one plus three	数式		英単語（数字）
数式化	The numerical formula for 'two times one plus three' is	2 * 1 + 3	数式		アラビア数字
計算式構文解析	The bracketed form of '2 * 1 + 3' is	(2 * 1) + 3	数式	(✓)	アラビア数字か "("
自然言語構文解析	The result of syntactic parsing for 'Hard students study.' is	(Hard students) study.	言語		英単語か "("
主語動詞の数一致	The keys on the cabinet	are	言語		英単語
性別代名詞の一致	Maria said that	she	言語		英単語
上位語判断	Roses are a type of	flower	言語		英単語
算術文章問題	Q: Janet's ducks lay 16 eggs per day... A: The answer is	18	両方	✓	アラビア数字

コアを $m^{\text{A-clean}}$，破損実行時スコアを $m^{\text{A-corr}}$ と表記する．それぞれのスコアはデータセット全体で平均され，最終的に Faithfulness は以下のように計算される：

$$\text{Faithfulness}(A, B) = \frac{m_B^{\text{A-clean}} - m^{\text{A-corr}}}{m^{\text{A-clean}} - m^{\text{A-corr}}} \quad . \tag{5}$$

3　実験設定

3.1　モデル

本節では，Llama-3.1-8B [18] および Qwen3-8B [8] の実験結果について報告する．さらに，いくつかの小規模モデルについても分析を行っており，その結果は付録 A に記載している．これらの結果は，モデル間における結果の一般性を裏付けるものである．

3.2　課題

本研究で使用する課題の一覧を表 1 に示す．

計算　3 項の加算，減算，掛け算を含む計算式を用いた．答えは自然数となるように調整した．

計算（書き下し式）　英語で書き下した式（*one times one plus three is?*）に対して計算を行う．これは**計算**課題で得られた回路が単にアラビア数字に反応するものでないことを確認するために用いる．

書き下し　言語課題として，計算課題と同じ計算式を計算を伴わずに単に英語でスペルアウトする．（例：2*1+3 → *two times one plus three*）

数式化　書き下し課題の逆であり，英語の計算式をアラビア数字を用いた計算式に変換する課題である．これは**書き下し**課題で得られた回路が，単純に数字を生成するための回路ではないことを確認するために利用する．

計算式構文解析　3 項の加算，減算，掛け算を含む計算式に対して，計算順序に応じて括弧を付与する課題を設定した．例えば，2*1+3 は (2*1)+3 となる．

自然言語構文解析　自然言語文の処理と計算処理の重なりを調査する研究 [19] を参考に，3 単語からなる文に対して構文解析を行い括弧を付与する課題を設定した．一方では後方単語を動詞・副詞として結合し，他方では前方単語を形容詞・名詞として結合する．例えば，*Students study hard.* → *Students (study hard.)*，*Hard students study.* → *(Hard students) study.* となる．

また，数式処理とは直接関係しない形式的言語能力課題として，以下も導入する．

主語動詞の数一致　統語能力を測る一般的な課題の一つとして，主語と動詞の数一致課題を採用した．この課題では，モデルに *"The keys on the cabinet"* のような入力を与え，*"is"* ではなく *"are"* が出力されることを期待する．

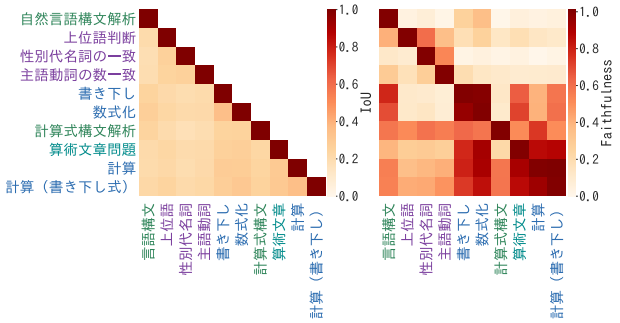
性別代名詞の一致　性別代名詞の一致課題である．モデルに *"Maria said that"* のような性別が明らかな文脈を入力し，正しい性別の代名詞である *"she"* が選択されることを期待する．

上位語判断　意味的能力を測る課題として，上位語の理解を評価する課題も導入した．モデルに *"Roses are a type of"* のような入力を与え，*rose* の上位概念である *flower* を出力できるか測定する．

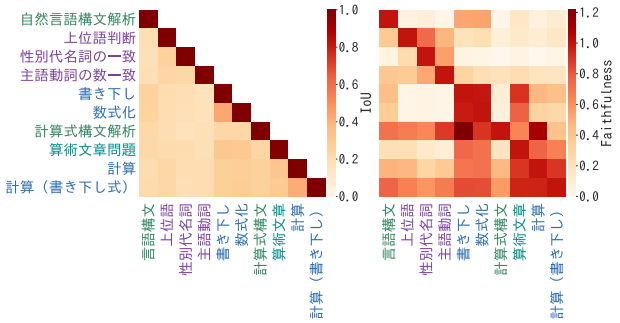
算術文章問題　また，言語処理と計算の両方が必要な課題として，算術文章問題データセットである GSM-Plus [20] も利用した．

4　実験結果

図 2 に，課題間の IoU と Faithfulness を示す．表 1 の「計算」の有無（✓）に依存して，回路が離れているのかが焦点である．



(a) Qwen3-8B



(b) Llama-3.1-8B

図 2　左: IoU（構造的な重なり）．右: 課題間 Faithfulness（能力的な重なり）．各行のスコアは対角要素で正規化されている．

IoU　図 2（左）に課題間の IoU を示す．値は異なる課題ペア間で顕著な差異は見られないものの，計算課題の間では比較的高いスコアを示し，互いに類似した傾向が見られる．

Faithfulness　図 2（右）には，課題間の Faithfulness の結果を示している．この図は，列に示された課題が行に示された課題によって解かれた場合の性能を示しており，行ベクトルは通常，特定の課題の回路がどの課題を解決するかを表す特性と見なされる [2]．顕著なパターンとして，例えば，計算関連課題間と，文字列変換課題間（書き下しおよび数式化）では，互いに高い Faithfulness スコアを示している．さらに，(i) 計算関連課題間のスコア，および (ii) 言語処理課題と計算関連課題間のスコアの 2 つのグループ間でスコアを比較した．Mann-Whitney U 検定の結果，計算関連課題間のスコアは，言語処理課題と計算関連課題間のスコアよりも有意に高いことが示された（$p < 0.05$）．（モデルごとの検定結果を付録 B に示す．）また，IoU 値は低いものの，課題間 Faithfulness が比較的高いケースも観察されており，これは特定のエッジが多義性や *super weight* [21] のような特性を持つ多くの課題において極めて重要であることを示唆している．

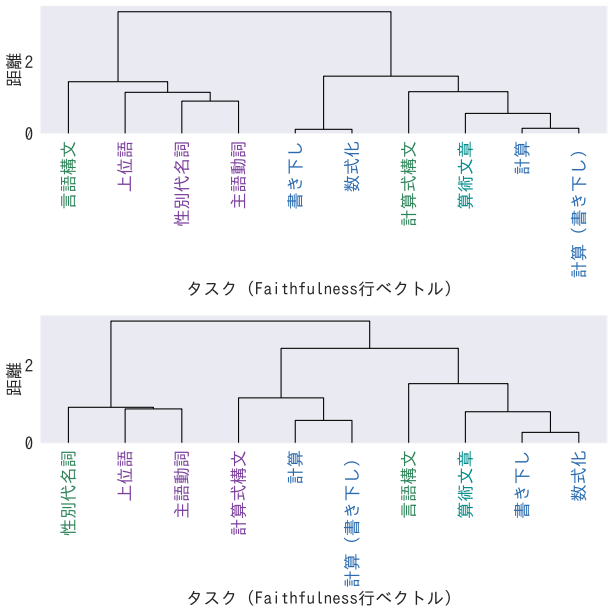


図 3　課題間 Faithfulness の階層的クラスタリングの結果．上: Qwen3-8B．下: Llama-3.1-8B．

階層的クラスタリング　Qwen3-8B および Llama-3.1-8B について，Faithfulness 行列の行ベクトルを用いて階層的クラスタリングを実施した（図 3）．その結果，計算関連課題と言語課題が明確に分離されたクラスタを形成していることが分かる．例えば，計算課題と書き下し課題は入力が同じであるが，クラスタリング結果では確かに互いに離れている．また，本質的に類似した課題である計算式構文解析と言語構文解析課題は，異なるクラスタに属しており，これは算術処理が数学式の解析といった言語処理課題においても言語回路を利用しないことを示している．この結果は，算術処理が独自の回路によって実行され，他の言語処理とは分離されていることを支持し，近年の人間を対象とした神経科学研究の知見とも一致する．

5　まとめ

本研究では，回路解析手法を用いて，大規模言語モデル（LLMs）内部における文法処理回路と算術処理回路の重なり合いについて調査した．計算関連課題は，言語処理課題と比較して相互間の重なり合い度合いが高いことが明らかになった．これは，LLMs における算術処理が，形式言語処理とは異なる独立した回路機構によって支えられていることを示唆しており，近年の人間を対象とした神経科学研究で報告されている示唆と類似する．

参考文献

[1] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. **Trends in Cognitive Sciences**, Vol. 28, No. 6, pp. 517–540, 2024.

[2] Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. Are formal and functional linguistic mechanisms dissociated in language models? **Computational Linguistics**, pp. 1–40, 09 2025.

[3] Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 7035–7052, Singapore, December 2023. Association for Computational Linguistics.

[4] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: Defining rois functionally in individual subjects. **Journal of Neurophysiology**, Vol. 104, No. 2, pp. 1177–1194, 2010. PMID: 20410363.

[5] Jennifer Hu, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. **Cereb. Cortex**, Vol. 33, No. 8, pp. 4384–4404, April 2023.

[6] Evelina Fedorenko, Michael K. Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. **Proceedings of the National Academy of Sciences**, Vol. 108, No. 39, pp. 16428–16433, 2011.

[7] Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code or not to code? exploring impact of code in pre-training. In **The Thirteenth International Conference on Learning Representations**, 2025.

[8] Qwen Team. Qwen3 technical report, 2025.

[9] Jackson Petty, Sjoerd van Steenkiste, and Tal Linzen. How does code pretraining affect language model task performance? **Transactions on Machine Learning Research**.

[10] Ruizhe Chen, Tianxiang Hu, Yang Feng, and Zuozhu Liu. Learnable privacy neurons localization 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 2: Short Papers)**, pp. 256–264, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[11] Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. Masking as an efficient alternative to finetuning for pretrained language models. 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. 2226–2241, Online, November 2020. Association for Computational Linguistics.

[12] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024.

[14] Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen, editors, **Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, pp. 407–416, Miami, Florida, US, November 2024. Association for Computational Linguistics.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[16] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In **The Eleventh International Conference on Learning Representations**, 2023.

[17] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In **First Conference on Language Modeling**, 2024.

[18] Llama Team AI @ Meta. The llama 3 herd of models, 2024.

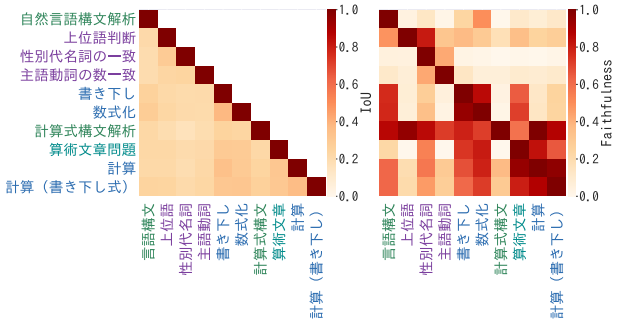
[19] Tomoya Nakai and Kazuo Okanoya. Neural evidence of cross-domain structural interaction between language and arithmetic. **Scientific Reports**, Vol. 8, No. 1, p. 12873, 08 2018.

[20] Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. 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. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

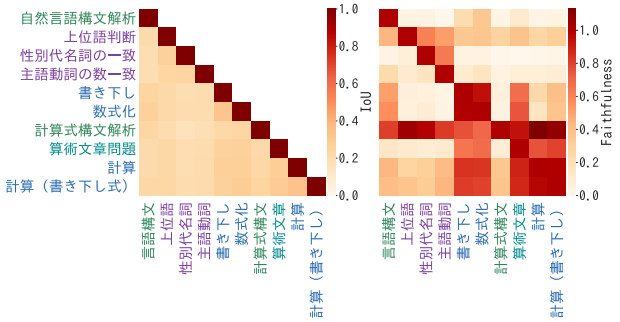
[21] Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. The super weight in large language models, 2025.

表 2　Mann-Whitney U 検定の結果.

モデル	Faithfulness（計算–言語 vs 計算–計算）		IoU（計算–言語 vs 計算–計算）	
	統計量（n, 平均）	p	統計量（n, 平均）	p
Llama-3.1-8B	n=(48, 12); mean (0.407, 0.706)	3.03×10^{-3}	n=(24, 6); mean (0.223, 0.299)	2.55×10^{-3}
Qwen3-1.7B	n=(48, 12); mean (0.356, 0.673)	1.34×10^{-3}	n=(24, 6); mean (0.244, 0.288)	2.85×10^{-2}
Qwen3-4B	n=(48, 12); mean (0.369, 0.769)	3.98×10^{-4}	n=(24, 6); mean (0.225, 0.268)	1.06×10^{-2}
Qwen3-8B	n=(48, 12); mean (0.379, 0.716)	3.48×10^{-4}	n=(24, 6); mean (0.236, 0.283)	1.43×10^{-2}



(a) Qwen3-1.7B



(b) Qwen3-4B

図 4　左: IoU（構造的な重なり）．右: 課題間 Faithfulness（能力的な重なり）．各行のスコアは対角要素で正規化されている．

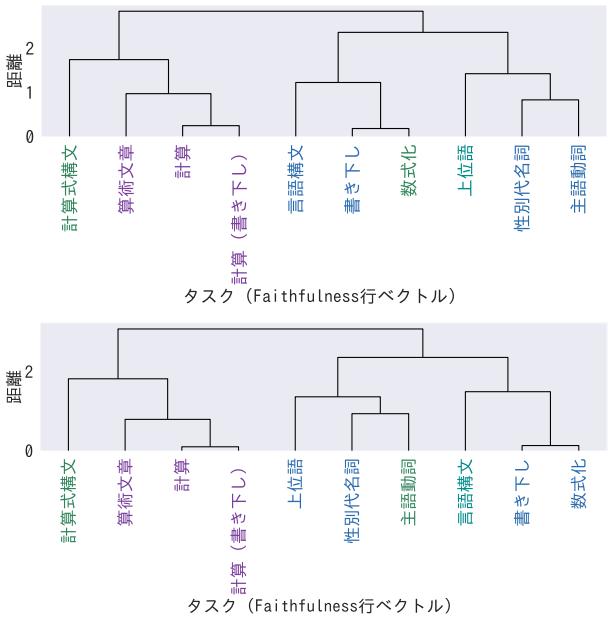


図 5　課題間 Faithfulness の階層的クラスタリングの結果．上: Qwen3-1.7B．下: Qwen3-4B.

A　他モデルの結果

Qwen3-1.7B，Qwen3-4B の IoU と Faithfulness の結果を図 4 に，Faithfulness の階層的クラスタリングの結果を図 5 に示す．モデルサイズが小さい 1.7B，4B サイズのモデルでも同じモデルファミリーの上位モデルである Qwen3-8B や，モデルファミリーの異なる Llama-3.1-8B と同じ傾向を示した．これらの小規模モデルの結果は，得られた知見の一般性を裏付けるものである．

B　Mann-Whitney U 検定の結果

計算関連課題が計算–言語課題ペアと比較して，より強い回路関連性を示すかを検証する．課題を**計算–計算**ペアと**計算–言語**ペアの 2 つのグループに分類する．ここで，計算関連課題は計算，計算（書き下し），計算式構文解析，算術文章問題の 4 つで構成され，それ以外の全ての課題を言語処理課題とする．課題間の Faithfulness については，$A \neq B$ を満たすすべての順序付きペア (A, B) を考慮する．IoU については，重複を避けるため，一意の非順序付きペアのみを使用する．その後，片側 Mann-Whitney U 検定を実施し，帰無仮説 H_0 として「算術–言語 ≤ 算術–算術」を設定した．表 2 に示すように，全てのモデルにおいて，**計算–計算**ペアでは Faithfulness と IoU の値が一貫して高い結果が得られた．