大規模言語モデル内部における言語と計算領域の区分

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, sasano@i}.nagoya-u.ac.jp
tatsuki.kuribayashi@mbzuai.ac.ae

概要

本研究では，大規模言語モデル（LLM）の内部において，言語非依存な思考の領域が一般的な言語領域よりもどれほど・どのように切り分けられているのかを調査する．特に算術計算問題をとりあげ，これらのデータの内部表現空間における分布を分析する．実験結果から，算術計算など言語一般とは異なるデータは入力層付近でただちに分離され，また計算式と算数文章題のように似た思考を要する問題同士ですら，どこかの層で同じ領域を占めることはないことが示唆された．すなわち，算術計算と言語の領域は分離されているものの，計算のための普遍的領域が存在するわけではなく，異なるタスク用に異なる数値計算領域が存在する，ある種冗長な様相であることが想定される．

1　はじめに

人間の脳において，言葉を流暢に扱う能力（言語）と，言語非依存な知識や推論，ないし社会性といった能力（思考）が，どれほど独立して機能しているのかという問いは，言語能力と思考能力は不可分かという言語（哲）学的な論争に端を発し [1, 2, 3]，主に神経科学領域において定量的に評価されてきた [4, 5, 6]．例えば，脳機能イメージングによる近年の観察によれば，言語の規則（語順など）に反応する言語領域と，（言語非依存に）計算や論理などに対応する思考領域は異なる，すなわち両者が独立しているとする主張がある [7, 8, 9]．またより広い文脈では，人間の言語処理が汎用的なネットワーク・計算で実現されているのか，言語特有なものなのかなど長らく議論されてきた [10, 11, 12]．このような議論を踏まえれば，一見人間らしい（あるいはそれ以上の）言語能力を見せる大規模言語モデル（LLM）の場合に，言語と思考がどの程度区分されているのかも当然関心の対象となる．この区分に対して言語モデルをどう捉えればよいかという抽象的な議論 [13] とは直交して，本研究では，言語と思考タスクによって実際に活性化する言語モデル内部の領域の違いを，ある種物理的なレベルで観察してみる．

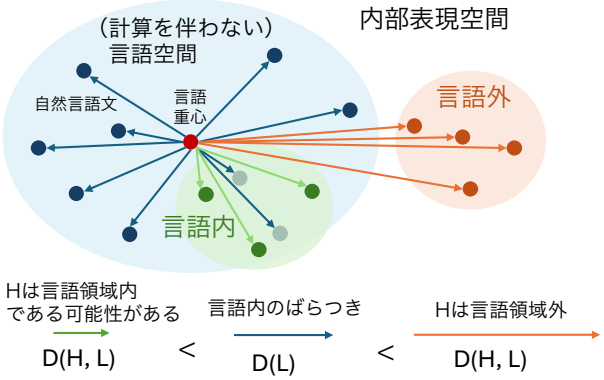

図 1　ある系列集合 H が言語集合 L からどれほど離れているかを調査したい．L の重心から $l \in L$ への平均距離よりも，$h \in H$ への平均距離のほうが長/短ければ，H と L の領域は別れている/いないとする．

我々の実験では，第一歩として，特に言語に非依存な思考能力として算術計算をとりあげ，言語モデルの内部において，算術計算を伴う入力が，それ以外の言語系列と比べて，どれだけ離れた領域で扱われているかを調査する．具体的に実験では，モデル内の各層において，計算を伴う入力が，そうでない言語系列集合のばらつきに対して，内部表現空間内でどれだけ平均的に引き離されているかを観察する（図 1）．少なくとも，これらの入力が，それ以外の言語系列に対して離れた場所でクラスタを形成するのであれば，算術計算に特化して活性化する領域がそこにあると考えてよいだろう．

実験の結果，算術計算データは，内部表現空間において 1 層目から直ちに区別されていた．一方で，例えば単純な計算式に答えるデータが占める領域と，算術文章題データの領域ですら，いずれの層でも離れていたり，言語理解が必要である算術文章題においても，言語データ一般からは 1 層目から離さ

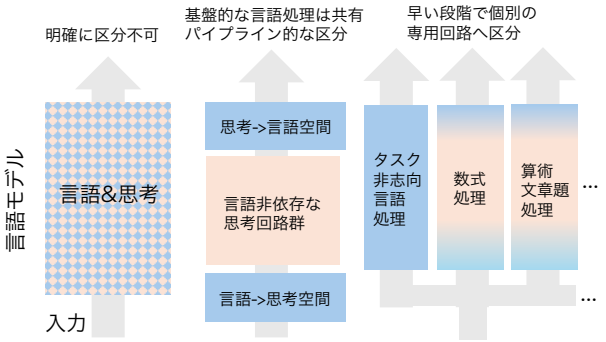

図 2　言語非依存な思考（本論文では算術計算）の領域が，言語モデル内でどう区分されている/いないかについて，例えば上のような可能性が考えられるだろう．

れていたりと，似たタスクに共通の基盤的な言語処理や計算モジュールが割り当てられるパイプライン的な様子は観察できなかった．すなわち入力は，図 2 右のように，タスクごとに細分化された異なる領域に直ちに振り分けられ，結果的に言語と計算も区分されて見えるが，似た問題同士の領域が重なることはないという点で，ある種冗長な内部機序である可能性がある．

2　実験設定

評価指標　ある系列集合 S_1 に対して，もう一つの系列集合 S_2 が，モデル内部でどれほど異なる領域で扱われているかを定量化する．例えば，S_1 に属する系列が特定の言語非依存な能力（計算など）を伴う刺激であり，S_2 がそうでない系列であるとする．S_1 に対応するモデルの内部領域と S_2 の領域が異なるのであれば，少なくともその空間では，両者は区別して処理されていると仮定したい．具体的には，系列集合 $S_\mathrm{x} = \{s_{(\mathrm{x},1)}, s_{(\mathrm{x},2)}, \cdots, s_{(\mathrm{x},n)}\}$ に含まれる各系列をモデルに入力し，それぞれの最終トークン[1]に対応する内部表象を言語モデルの l 層から取得する．S_x に対応する l 層の表象集合を $\boldsymbol{H}^l_\mathrm{x} = \{\boldsymbol{h}^l_{(\mathrm{x},1)}, \boldsymbol{h}^l_{(\mathrm{x},2)} \cdots, \boldsymbol{h}^l_{(\mathrm{x},n)}\}$ と表す．ある $\boldsymbol{H}_C$ が $\boldsymbol{H}_L$ からどれほど離れているか $d(\boldsymbol{H}_C, \boldsymbol{H}_L)$ を以下のように定量化する．

$$d(\boldsymbol{H}_C, \boldsymbol{H}_L) = \frac{D(\boldsymbol{H}_C, \boldsymbol{H}_L)}{D(\boldsymbol{H}_L)} \tag{1}$$

$$D(\boldsymbol{H}_C, \boldsymbol{H}_L) = \frac{\sum_{\boldsymbol{h}_C \in H_C} \left\| \boldsymbol{h}_C - \bar{\boldsymbol{H}}_L \right\|}{|H_C|} \tag{2}$$

$$D(\boldsymbol{H}_L) = \frac{\sum_{\boldsymbol{h}_L \in H_L} \left\| \boldsymbol{h}_L - \bar{\boldsymbol{H}}_L \right\|}{|H_L|} \tag{3}$$

1)　少なくとも系列の最初の方の部分系列では，どのような処理が求められているか曖昧な場合が想定されるため．

表 1　実験データ例．英語以外の例は割愛する．

データ	言語	数量	計算	例
言語データ	✓			How do you view the nature of the world we live in?
計算式			✓	$3 * 1 - 2 =$?
スペルアウト式	✓		✓	three times one minus two equals?
数量知識	✓	✓		What is the atomic number of hydrogen?
数量知識+計算	✓	✓	✓	{the number of fingers displayed in a peace sign} $-1 =$?
GSM8K	✓	✓	✓	A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

我々の実験では，$\boldsymbol{H}_L$ は計算を伴わない言語系列とし，$\boldsymbol{H}_C$ には計算を伴う様々な条件の系列を採用してみる．$D(\boldsymbol{H}_L)$ は $\boldsymbol{H}_L$ 内の表現がどれほどばらつくかを示しており，計算を伴わない言語データで求めている．$D(\boldsymbol{H}_C, \boldsymbol{H}_L)$ と $D(\boldsymbol{H}_L)$ の比をとることで，$\boldsymbol{H}_L$ 内のばらつきに対して，$\boldsymbol{H}_C$ が何倍離れているかという解釈ができ（図 1），1 を超えるなら $\boldsymbol{H}_C$ は $\boldsymbol{H}_L$ の領域外にあるとする．[2]なお，逆は必ずしも言えない点で（例えば $\boldsymbol{H}_L$ 領域内に $\boldsymbol{H}_C$ が隔離された小島を作っている可能性がある）やや厳しい条件であるが，実験結果では層横断的に言語領域と計算領域間でこの値が 1 を超えることを示す．

言語モデル　Gemma-2-9b-it [14]，Llama-3.1-8B-Instruction [15]，Qwen2.5-7B-Instruct [16] の 3 つのモデルを使用し，モデル横断的な傾向を観察する．

3　実験

3.1　設定

大きく分けて 6 種類のデータについて（表 1），それらが LLM の内部表現空間で，言語データ一般からどれほど離れた領域にあるかを調べる．

言語データ　数量計算を伴わない，一般的な言語刺激として，高品質な多言語データである MADLAD コーパス [17] から抽出した自然言語文を用いる．英語・中国語・日本語・ロシア語・アラビア語それぞれ 100 文，計 500 文を用いた．これらの一般的な言語入力に対して，計算を伴うデータがどのように特別な扱いを受けるかが焦点である．[3]

計算式　三項の加算，減算，掛け算を含む計算式（例：$2 * 2 - 3 =$?）を用いた．答えが 1 から 10 まで

2)　なお，$\boldsymbol{H}_L$ の重心を起点にばらつき・距離を考えており，$d(\boldsymbol{H}_C, \boldsymbol{H}_L)$ は $\boldsymbol{H}_C, \boldsymbol{H}_L$ に対して非対称であり，厳密には距離ではない．

3)　言語と思考の対立における「言語」は文法などのメタ言語的知識のことを意識し，非文に反応する領域との差分などをとることが多いが，今回の設定では扱いが非自明であるため，非文との差分などは考慮していない．

の自然数となる問題をそれぞれ 10 問，合計 100 問作成した．

スペルアウト計算式　計算問題を英語と中国語にそれぞれスペルアウトしたもの（例：two times two minus three equals?）も導入する．[4] 言語と計算が真に分離して処理されているのであれば，仮に表層が比較的自然言語に近くても，一般的な自然言語文とは異なる数式のネットワークで処理されるであろう．

数量知識データ　上記の言語データと計算式では出力しようとするトークンの種類が異なるため，特に後半層で過剰に両者が分離して見えるかもしれない懸念がある．理想的には，言語処理を伴い，かつ出力が数字であるというベースラインも導入したい．そこで，回答が数字となる計算を伴わない自然言語疑問文（例：`What is the atomic number of hydrogen?`）も異なるベースラインとして用いた．言語非依存な処理の傾向を調べるため，英語・中国語で各言語 100 問ずつの合計 200 問作成した．[5] このような問題の処理には，まず文から質問の意味を理解し，問われている数量知識にアクセスする段階的な処理が求められ，直観的には知識にアクセスする前の段階では一般的な自然言語文と近い領域で処理されることが想像される．また計算は伴わないため，もしこのデータが計算問題と近い領域に現れた場合，その領域は計算ではなく単に数量一般に紐づいている可能性がある．

数量知識+計算データ　最後に計算問題の中に数量知識が入ったデータも導入した（例：{`the number of fingers displayed in a peace sign`}`-1=?`）．通常の計算や知識問題が解かれる領域とどのような関係になるのか興味深い．なお正答率（数の完全一致）は 55%程度であり，モデルはある程度問題の形式を理解できている．

GSM8K　数量計算を伴うベンチマークとしてよく用いられる GSM8K [18] も活用する．GMS8K は人手で作成された小学生向けの算数文章問題のデータセットであり，回答するには問題文を理解する能力と，四則演算の能力が必要になる．[6]

4）日本語スペルアウトについては，少なくとも小学校の教科書などを見ると，数字はアラビア数字で書く慣習が強いため，モデルの混乱を防ぐ意味で採用しなかった．

5）Wikipedia から数字に関する一般知識を抽出して手作業で作成し，多言語化の際には ChatGPT を用いた．

6）なお，計算問題文と数値知識問題文，計算+数量知識混合問題文，GSM8K に関しては `Please answer the following question by providing numbers alone as your answer: {` **問題文** `}` というプロンプトを与えて，答えの数字のみを回答するように促している．

3.2　評価

以下の 7 つの値を層ごとに計算する：d(計算, 言語データ)，d(数量知識, 言語データ)，d(数量知識+計算, 言語データ)，d(GSM8K, 言語データ)，d(数量知識, 計算)，d(数量知識+計算, 計算)，d(GSM8K, 計算)．言語データと比較している最初の 4 つについては，言語に対して算術計算が特殊な扱いを受けている場合，値が 1 より大きくなる（両者が遠ざかる）．残り 3 つについては，単純な計算データで同定された計算領域について，数値・計算を伴う他のデータにおいても再利用される普遍的領域なのか洞察を得るために観察する．例えば，計算データで得られた計算領域に，GSM8K データ集合が一度も現れないのであれば，同じ計算を扱っているにも関わらず，問題の形式・フォーマットによって，両タスク用に異なる計算領域があることが示唆される．

3.3　実験結果

Gemma の結果を図 3 に示す．0 層目は埋め込み表現に対応する．他モデルの結果は付録 A に記載するが，同様の傾向を示している．また，内部表象を主成分分析で可視化した結果を付録 B に記載する．

言語 vs. 計算式　図 3 の一番上の図は，言語データと計算式データ間の遠さを示している．参考として載せている「計算内のばらつき」を除き，いずれの組み合わせの場合も常に値が 1 を超えており，言語データ内でのばらつき以上に，離れた位置で計算データが表現されていることがわかる．なお，「計算内のばらつき」は計算式とスペルアウト計算式の集合に式 (3) を適用したあと，他の値と同様に言語データのばらつきで正規化している．この値は 1 より小さいことから，計算式データ群は言語データよりも散らばっておらず，言語データの外側で個別のクラスタを形成していることが示唆される．また，この傾向は 1 層目から観察されることから，数式が入力された時点で，ただちに一般的な自然言語文とは異なる扱いを受けていることがわかる．これはスペルアウトされた計算式でも概ね言えることであり，単に文字種の違いに対応しているわけではないことが示唆される．ただし，英語と数式スペルアウト（英）については，値が 1 を超えるものの，前半層で距離がやや近いことも分かる．

言語 vs. 数量を伴う問題　図 3 の 2 つ目の図も同様に，言語とそれ以外のデータの遠さを示してい

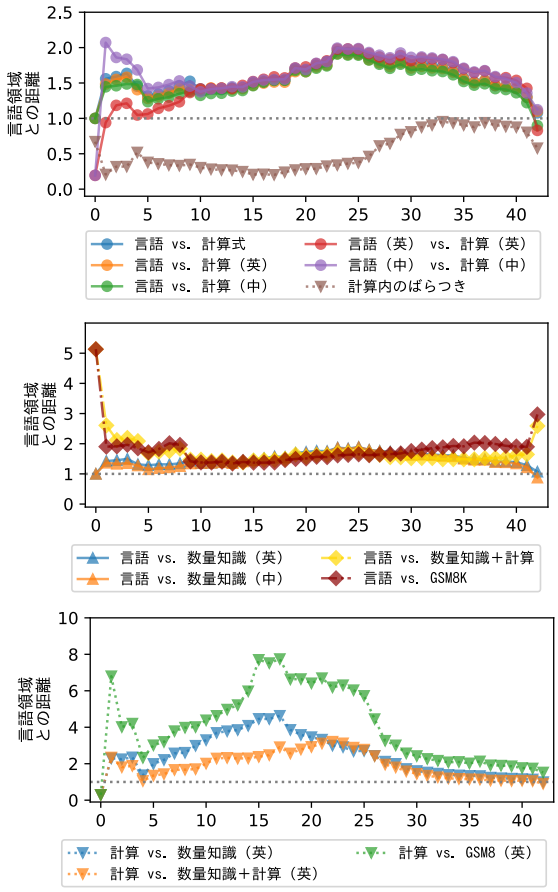

図 3　実験結果．凡例にて（英）などの表記がない限り，対応するデータを全て用いている．例えば，最も上の図の「言語 vs. 計算式」では，言語は英語・中国語・日本語・ロシア語・アラビア語を混ぜて用いている．

る．上の図と同様，レイヤを通して値は 1 を超えている．計算を伴う数量知識＋計算や GSM8K と比べ，計算を伴わない数量知識問題は前半層で比較的言語に近いことから，この比較では，前半層で計算に特化して反応する領域があることが示唆される．興味深い点として，GSM8K のような文章問題を解くためには，初めに文章の理解（一般的な言語処理）が必要であり，その意味で特に前半層では一般言語領域に近づいても良いはずであるが，そのような傾向は観察されなかった．したがって GSM8K も，算術文章題と判明した時点で，ただちにその問題に対応する領域に移されているようである．

計算 vs. 数量を伴う問題　最後に，図 3 の 3 つ目の図は，数式で同定された計算領域と，数量・計算が関わるデータ（数量知識，数量知識+計算，GSM8K）との距離を示している．[7]計算を伴わない数量知識問題が計算領域と離れていることから，計算データで同定した領域は単に数字を扱う話題に紐づいているわけではないことが分かる．一方で，数量知識＋計算と GSM8K は，問題を解く過程で算術計算を伴うはずである．それにもかかわらず，層を通して，計算領域からは遠い領域でこれらのデータが処理されている．したがって，計算を伴う様々な問題に対して共通に活用される計算用の絶対的な領域があるわけではなく，異なる形式の計算問題に対しては，異なる領域で数量計算が行われている可能性が示唆された．

7）　ここでは，数式データのばらつきで距離を正規化している．すなわち，計算領域の重心からの散らばりに対して，対するデータが何倍遠くにあるかを示している．

4　おわりに

以上を踏まえると，計算式，算術文章題など計算を伴う問題が入力されると，それらは 1 層目からただちに一般的な言語データから引き離され，図 2 の一番右のように処理されていることが示唆される．言い換えれば，図 2 の真ん中のように，特定の層が特定の思考（計算など）の枠割を果たしたり，入力非依存に使い回される言語・計算モジュールがあったりするのではなく，似た部分問題（計算）を解くためのモジュールが，冗長に複数存在する可能性が示唆される．これは少なくとも，言語学分野において一般的に考えられている言語と思考の分業のされ方とは異なるだろう．また，既存研究では数量がどのように表現されているかといったことがしばしば調査されている [19, 20] が，本研究ではどの（どれぐらい広い）領域を使って計算を行っているかの分析は行っていない．

実験手法については改善の余地が様々ある．例えば，内部空間における向きの重要性を考慮すれば得られる洞察は変わるかもしれない．またプロービングなどで領域間の境界を同定するほうがよいかもしれない [21]．また言語領域については，文法性判断タスクに紐づけて同定したほうが自然かもしれない [22, 23]．おそらく，モデルの言語と思考が区分されているかに答える，より洗練された実験設定としては，ある言語モデルを，言語に流暢だが思考タスクができないモデルと，言語に流暢ではないが思考タスクが解けるモデルに容易に分解できるか，またはそのような介入は実現するかといった，モデルの分解 [24] や枝刈 [25] に取り組むと良いのかもしれない．

参考文献

[1] Donald Davidson. Thought and talk. 1975.

[2] N Geschwind. The organization of language and the brain. **Science**, Vol. 170, No. 3961, pp. 940–944, November 1970.

[3] Peter Carruthers. The cognitive functions of language. **Behav. Brain Sci.**, Vol. 25, No. 6, pp. 657–674, December 2002.

[4] 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.

[5] Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. **J. Neurophysiol.**, Vol. 112, No. 5, pp. 1105–1118, September 2014.

[6] Evelina Fedorenko, Anna A Ivanova, and Tamar I Regev. The language network as a natural kind within the broader landscape of the human brain. **Nature Reviews Neuroscience**, pp. 1–24, 2024.

[7] 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.

[8] 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.

[9] Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. Language is primarily a tool for communication rather than thought. **Nature**, Vol. 630, No. 8017, pp. 575–586, Jun 2024.

[10] Carl Wernicke. The aphasic symptom-complex: a psychological study on an anatomical basis. **Archives of Neurology**, Vol. 22, No. 3, pp. 280–282, 1970.

[11] Cory Shain, Idan A Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. **bioRxiv**, p. 2021.09.18.460917, September 2021.

[12] Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. **Neuropsychologia**, Vol. 138, p. 107307, February 2020.

[13] 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**, 2024.

[14] Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

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

[16] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.

[17] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset, 2023.

[18] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[19] Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. **arXiv preprint arXiv:2411.04986**, 2024.

[20] Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric attributes 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. 175–195, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[21] Guillaume Alain. Understanding intermediate layers using linear classifier probes. **arXiv preprint arXiv:1610.01644**, 2016.

[22] Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large 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. 6228–6247, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[23] Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. Interpretability of language models via task spaces. 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. 4522–4538, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[24] Anonymous. Extend model merging from fine-tuned to pre-trained large language models via weight disentanglement. In **Submitted to The Thirteenth International Conference on Learning Representations**, 2024. under review.

[25] Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**.

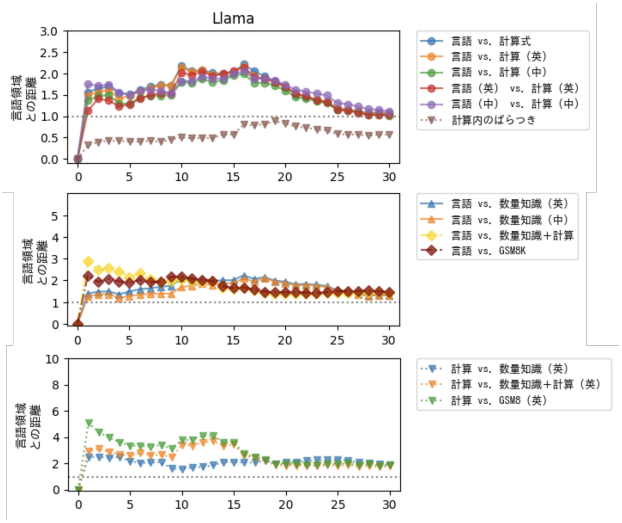

図 4　Llama 3.1 の結果

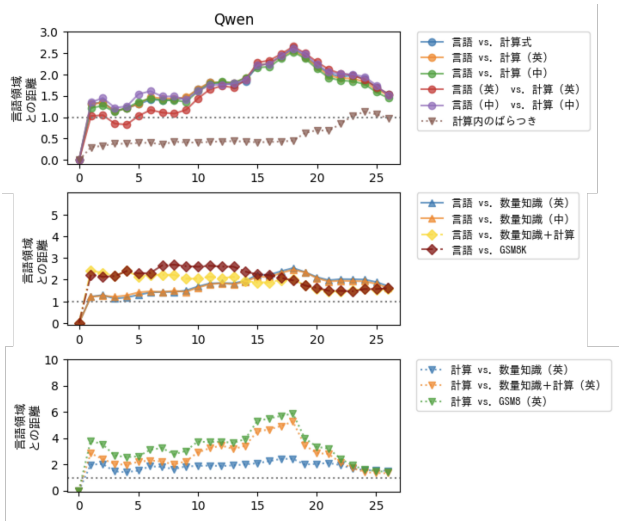

図 5　Qwen2.5 の結果

A　Llama および Qwen の結果

3 節で行った実験の，Llama の結果を図 4，Qwen の結果を図 5 に示す．Llama と Qwen の両モデルにおいて Gemma と似た結果を得た．従って，本実験で得られた LLM における言語と計算領域の区分性についてはモデル横断的に成り立つ可能性が示唆される．

B　主成分分析による可視化

3 節で行った実験で取得した内部表象を，主成分分析によって 2 次元まで次元削減してプロットした結果を図 6 に示す．モデルごとの散布図は上から下にかけて層が深く（出力方向側に）なっている．いずれのモデルも入力側の層ではバラバラだったものが，中間の層になると言語，計算，数量知識の 3 つのクラスターにまとまっていることが確認できる．

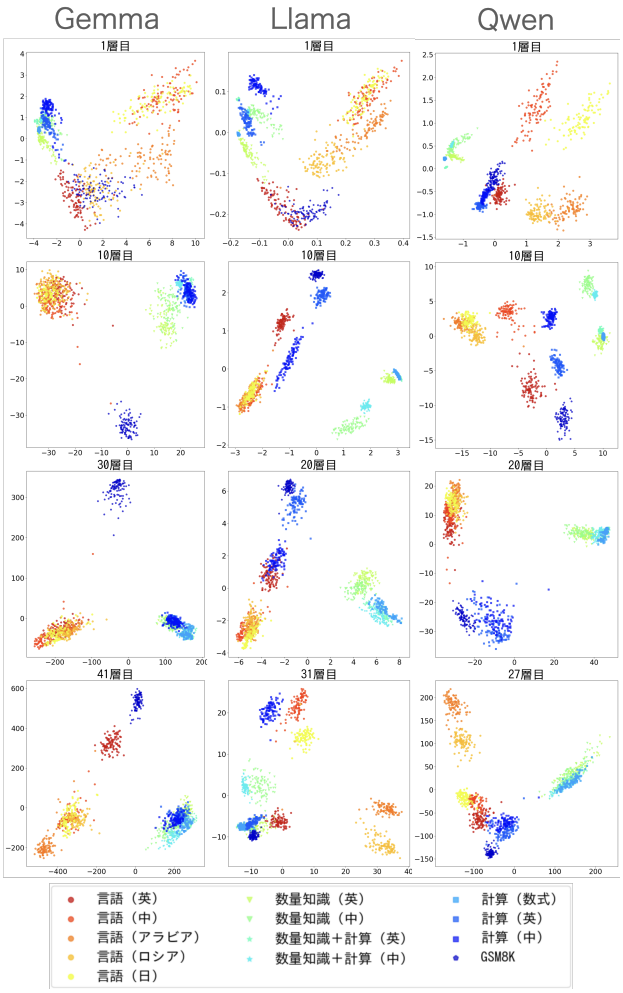

図 6　次元削減による内部表象の可視化結果．左から順に Gemma，Llama，Qwen となっている．各モデルは Gemma が 42 層，Llama が 32 層，Qwen が 28 層の構成になっている．