専門ドメインを対象とした事前学習データと精度の関係分析

緒方 陸[1] 岡野 将大[1] 大曽根 宏幸[2] 大久保 順一[1] 藤井 純一郎[1]
[1]八千代エンジニヤリング株式会社 [2]インダストリアル・ドリーム株式会社
{rk-ogata, ms-okano, jn-okubo, jn-fujii}@yachiyo-eng.co.jp
{osone}@industrial-dream.com

概要

大規模言語モデル(LLM)は long-tail の知識に対して精度が低く，過去には事前学習データに含まれる知識に関する頻度が高いと精度も高くなる関係を示した例がある．一方，実用上はより複雑なタスクを解くことがあり，単純な頻度よりもいかに文脈を捉えるかが重要となる可能性もある．そのための基礎的な分析として，本研究では次の二つを実施した．1) 事前学習データに含まれる単語やトークンの頻度を算出し，精度の関係を分析した．2) 各層の埋め込み表現を用いて定性的な分析を行った．その結果，今回対象としたタスクでは頻度と精度の関係は確認できなかったものの，LLM は文章全体で文脈を捉えずに特定のトークンに着目している可能性を示した．

1 はじめに

大規模言語モデル（LLM）は long-tail の知識に対して精度が低いことが知られており[1]，long-tail の知識への対応として RAG（Retrieval Augmented Generation）や Fine-tuning がある[2, 3]．一方医療ドメインなどにおいては，既往の LLM や RAG は評価タスクの違いやドメイン知識が無いなどの要因により実用に値しないという報告もある[4]．また，LLM が外部知識を参照する際に Knowledge Conflicts（LLM が暗黙的にパラメータ内に保有する知識と外部知識の衝突）があった場合，LLM は popularity の高い知識を優先するとされている[5]．

土木分野など，データが充分に整備されていない分野においては，既往の LLM は実用に耐えうる性能ではなく，RAG や Fine tuning を検討する必要がある．一方で，これらの手法を検討するにあたってどの程度のデータが必要か，またどの程度事前学習で学習されていれば Knowledge Conflicts が生じにくいかなどは明らかになっていない．これらが明らかになれば，実運用においてもデータ整備や LLM の学習に対し，戦略的な投資が可能となる．

本研究では，上記を明らかにすることを目指した基礎的分析の位置付けとする．専門ドメインとして土木分野を対象に，技術士一次試験(建設部門)の過去問題[6]を用いた．このタスクに対し，LLM は単純な事前学習の用語やトークン頻度に伴い適切な回答を生成できるようになるか，または単純な頻度よりも文脈を捉えることがより必要かを明らかにしたい．

本稿では上の問いに対し，事前学習データに含まれる用語やトークンの頻度を算出し，精度（正答率）との関係を分析した（3 章）．また文脈理解の観点から，各層の埋め込み表現を用いて定性的な分析を実施した（4 章）．

2 既往研究

2.1 頻度と精度の関係

事前学習データと精度の関係を調査した例として，Kandpal ら[1]は，大規模なオープンデータセットに含まれるエンティティを DBpedia/Wikidata ID にリンクさせ，QA ペアのエンティティに関連するドキュメント数を算出した．このドキュメント数と QA 精度の関係から，long-tail の知識に関する質問は精度が悪化することを示している．Sun ら[7]は Knowledge graph から生成した QA データを用いて popularity score と精度の関係を分析し，Tail に関する事実ほど正答率が低いことを示した．ただし，これらは直接事前学習データと精度の関係を示したものではなく，事前学習データがどの程度精度に影響しているかは不明である．

また，Chang ら[8]は架空の事実から作成したデータセットを使用して知識獲得のダイナミクスを分析した．long-tail の知識の学習が難しいのは，知識の出現頻度が一定の閾値（learnability threshold）を下回ると忘却が進むことが原因であると述べている．Allen-Zhu and Li [9]は（名前，属性，値）タプルから

生成した合成データを用いて学習頻度と性能の関係を調査し，学習時に獲得したい知識が 1000 回曝されることで知識を格納できるとしている．しかし，これらは架空の事実に基づく合成データで評価したものであり，現実のデータとは複雑さなどの観点で大きく異なる可能性がある．

2.2 LLM は文脈を捉えているか

実用においては[9]などのようにルールベースで作成された合成データとは異なり，問題の回答にはより複雑な推論が必要な可能性がある．この場合，用語やトークンの単純な頻度よりもテキストの文脈を捉えることが必要であると考える．

Wang ら[10]は知識を問うタスク（TriviaQA），推論タスク（GSM8K），翻訳タスク（WMT）などを対象に，Generalization（汎化能力）と Memorization（暗記能力）のどちらの影響が大きいかを調査した．同著者らによると TriviaQA では Memorization が，GSM8K や WMT では Generalization の影響が大きいと述べている．また Memorization についても短文よりも長文の記憶が有用としており，概念や文脈を捉えることが有効に働く可能性を示唆している．

Ethayarajh [11]や Ju ら[12]によると，言語モデルはネットワークの後半の層で文脈を捉えると報告している．[11]は言語モデルの埋め込みを用いて各層の単語表現を分析し，より深い層で文脈を捉えていることを示した．[12]はプロービングタスクにより層ごとのトークンの情報量を評価し，LLM が深い層で文脈を捉えていることを示している．

3 頻度に関する分析

本稿では事前学習データの単語やトークンを直接扱い，かつ現実世界のデータを対象とするべく，専門ドメインの一つである土木分野を対象とし，事前学習データの単語やトークンの頻度と QA タスクの精度の関係を分析した．

3.1 モデル

使用するモデルとして，事前学習データが公開されていることが必要である．よって本研究では学習データもオープンなモデルである llm-jp-3-13b[i]を採用した．また，今回は日本語を対象とするため，分析対象の事前学習データは Wikipedia，Common Crawl (level2), WARP/PDF (e0 / e0.2), WARP/HTML, Kaken とした[ii].

3.2 QA タスク

土木分野のタスクとして技術士一次試験（建設部門）の過去問題[6]を採用した．各設問は 5 つの選択肢から数字を選び回答する多肢選択形式であり，対象年度は本検討開始時に過去問題が公開されていた平成 23 年度から令和 5 年度までの 14 年分（令和元年度は再試験があったため 2 年分）とした．参考として付表 1 に設問のサンプルを示す.

llm-jp-3-13b は画像入力非対応であるため, 設問に画像を含むものは除外した．合計 433 問のうち 3 問を Few-shot の例題として使用し, 残りを評価対象とした．なお，評価は選択肢をランダムに並び替えた設問で 3 回試行した結果の平均正答率を精度とした.

表 1 に QA タスク評価結果を示す．参考として llm-jp-3-1.8b および ChatGPT の結果も載せている．表より，llm-jp-3-1.8b はランダムと同程度の精度であったが, 今回採用するモデルである llm-jp-3-13b はそれよりもやや高い精度となった．ChatGPT は 8 割程度の正答率であった.

表 1　QA タスク評価結果

Model	Accuracy
llm-jp-3-1.8b	0.19
llm-jp-3-13b	0.24
chatgpt-4o-latest	0.79

3.3 頻度と精度の関係

頻度の計算方法として，土木用語の頻度で計算（3.3.1 節）とトークン頻度で計算（3.3.2 節）の二つを採用した.

3.3.1 土木用語の頻度で計算

土木用語は設問文中で重要なキーワードとなるため，事前学習データのテキスト中に含まれる土木用語の頻度が精度に影響する可能性がある.

技術士一次試験の過去問題から ChatGPT を用いてキーワード抽出し，テクリスキーワード[13]に含まれる用語，全 843 語を土木用語として定義した．なお，抽出された土木用語には専門用語のほか，一般にも使用される用語も含まれ，その頻度はオーダーレベルで異なる（図 1）.

[i] https://huggingface.co/llm-jp/llm-jp-3-13b

[ii] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

定義した土木用語を対象に事前学習データを走査し，土木用語が完全一致する回数を算出した．なお本稿では，頻度算出においては類義語や誤字などは考慮しないこととした．各設問中に含まれる土木用語のうち，事前学習データに出現する最も頻度の高い／低い用語の頻度（(a), (b)），および設問中に含まれる土木用語の事前学習データに出現する頻度平均（(c)）の 3 パターンの頻度と QA 精度の関係をヒストグラムで可視化した．

(a)の頻度と精度の関係を図 2 に示す．(b)および(c)の図は付図 1 に整理した．図より低い頻度で高い精度を示し，[1]など既往研究において指摘されている高頻度と高精度の正の相関関係について，本研究では同様の傾向を確認するには至らなかった．

3.3.2 トークン頻度で計算

事前学習の際，各単語はトークン化されて LLM に入力されるため，土木用語の頻度ではなく，トークン頻度に影響を受ける可能性がある．

トークン頻度の計算は次の方法で行った．まず，事前学習データのテキストをトークナイザーによりトークン化し各トークンの頻度辞書を作成する．次に過去問題の問題文をトークン化し，先ほど作成した頻度辞書からトークン頻度を取得する．取得した頻度を用い，全トークン頻度の(a)合計値，(b)平均値，(c)中央値を算出し，これらを評価する頻度とした．

(a)の頻度と精度の関係を図 3 図 2 に示す．(b)および(c)の図は付図 2 に整理した．前節と同様に，既往研究と同様の傾向を確認するには至らなかった．

4 文脈に関する分析

前章で既往研究と同様の傾向を確認できなかったため，本タスクにおいては単純な頻度よりもテキストの文脈を捉えることが必要である可能性がある．

本研究で用いた llm-jp-3-13b は問題文の文脈を捉え切れていない可能性があると考え，これを明らかにするため各層の埋め込み表現の比較（4.1 節）と最終層各トークン埋め込み表現の比較（4.2 節），の二つの方法で分析を実施した．

4.1 各層の埋め込み表現の比較

[11]と同様に，各層の埋め込み表現を用い，設問文と設問文に関連する／しない事前学習データのテキスト（設問文中の土木用語を含む／含まないテキストと定義）の比較評価を行った．なお，事前学習データについて，全テキストを解析することが困難であったため，設問文中の土木用語を含む事前学習データのうち，無作為に 1000 件抽出したテキストを対象とした．

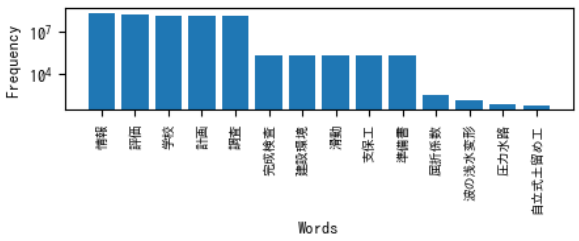
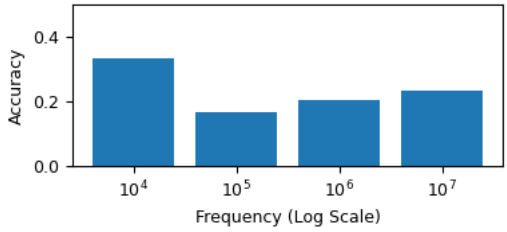
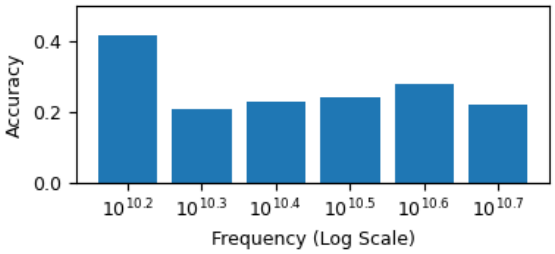


図 1　本稿で定義した土木用語の頻度サンプル



(a)　最も頻度の高い用語の頻度

図 2　土木用語頻度と QA タスク精度の関係



(a)　全トークン頻度の合計値

図 3　トークン頻度と QA タスク精度の関係

図 4 に結果を示す．上段の図は問題文と設問文に関連するテキストの，各層埋め込み表現のコサイン類似度を示す．エラーバーは事前学習データの各テキストによる数値範囲を示している．図より 4 層目で最大値を示し，層が深くなるにつれて類似度はなだらかに減少し，おおよそ 0.7〜0.8 の範囲で推移している．また，下段の図は問題文と設問に関連しないテキストの中間表現のコサイン類似度を示しているが，下段の図も上段の図と同様の傾向を示している．[11]や[12]が示す通り言語モデルはネットワークの深い層で文脈を捉えると仮定すると，文脈が類似したテキストの深い層の埋め込み表現は類似度が高くなると推察する．一方今回のタスクにおいて，llm-jp-3-13b は関連する／しないに関わらず埋め込み表現は同様の傾向を示しており，回答に有効な文脈は

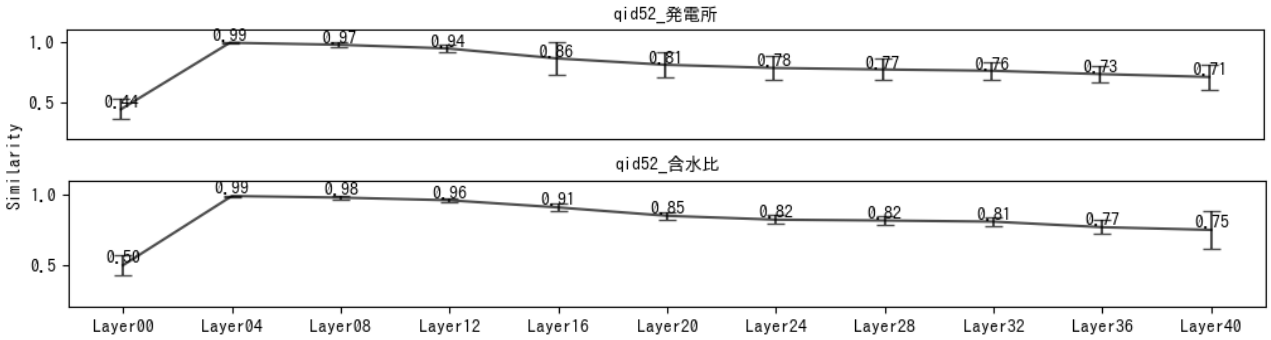


図 4　問題文と事前学習データの中間表現の類似度（上段：正答率 100%の設問，下段：設問文中に含まれない用語に関連するテキストとの比較）

捉えられていないことを示唆する結果となった．

4.2 最終層各トークン埋め込み表現の分析

前節では llm-jp-3-13b は文脈を捉えていないとしたものの，図 4 の上段の設問は 3 回の試行で正答率が 100%であった．前節は各層の埋め込み表現の全トークン平均化による評価を行ったが，平均化により情報が圧縮された可能性がある．よって本節では各トークンの個別の埋め込みを可視化し，定性的な分析を行う．

図 4 の上段で設問文との類似度が最大（①）・最小（②）のテキスト，および下段で設問文との類似度が最大のテキスト（③），また参考として ChatGPT に生成させた対象設問の解説文（④）を最終層の各トークンの埋め込み表現を UMAP [14]により可視化した結果を図 5 に示す．星印は各トークンの平均を示している．なお，図 4 における設問文と各テキストの類似度は，①，③，および④は 0.9 程度，②は 0.36 であった．設問文は火力発電所に関する内容だが，①は太陽光発電所の建設工事に関する内容，②は発電量などの数値データであった．また，③は多摩丘陵の地質に関する説明文であった．

図より，概ね各テキストごとに（色別に）クラスターができているが，②や④など，同じテキスト中でも一部のトークンが異なるクラスターに属するものもある．ここで，前節ではベクトルの数値をトークン全体で平均していた．一方で図は同じテキスト中でも性質の異なるトークンの存在を示唆している．これらは平均と大きく異なる可能性があり，回答に有効なトークンのみで設問に関する文脈を捉え，テキスト全体を使用していない可能性を示唆している．

本稿の検討においては，どのようなトークンが正解の回答生成に貢献し，そのトークンに関連するデータが事前学習データにどの程度含まれているかは明らかにできていないため，今後の課題とする．

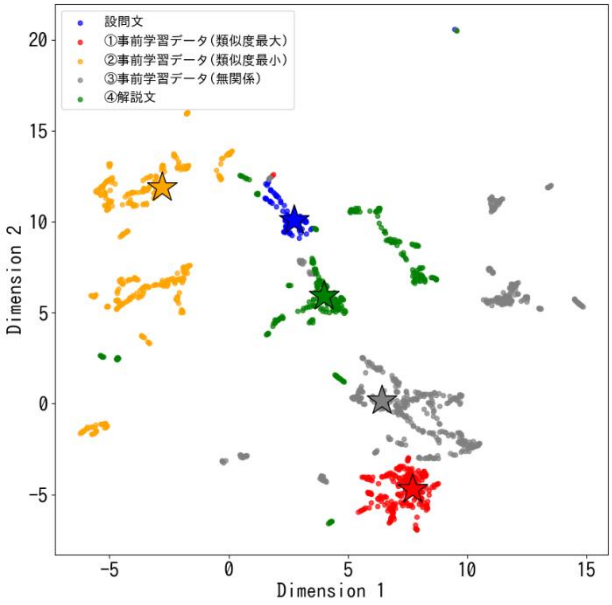


図 5　最終層の各トークン埋め込み表現の比較（星印：各トークン平均）

5 おわりに

本稿では LLM の事前学習データに出現する用語やトークン頻度とタスク精度の関係を分析した．結果から，既往研究と同様の頻度と精度の関係は確認できなかったものの，LLM は文脈を踏まえた回答に有効なトークンのみを使用している可能性を示した．

一方，今回使用したモデルは 13B クラスであり，性能が充分でなかった可能性がある．今後の課題として，より大きなモデルを使用して分析する必要があると考える．また埋め込み空間のトークンについて調査した例は多数あり（[15]など），文脈の評価方法についても検討の余地があると考える．

参考文献

[1] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, Colin RaffelLarge. Language Models Struggle to Learn Long-Tail Knowledge. Proceedings of the 40 th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[2] 箱石健太，一言正之，菅田大輔．土木分野における杉崎光一，全邦釘，阿部雅人．大規模言語モデルの動向と利活用に向けた検討，pp.220-230. AI・データサイエンス論文集 5 巻 3 号，2024.

[3] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, Ranveer Chandra. RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture, arXiv preprint arXiv: arXiv:2401.08406, 2024.

[4] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, David A. Clifton. A Survey of Large Language Models in Medicine, Progress, Application, and Challenge, arXiv preprint arXiv:2311.05112, 2024.

[5] Jian Xie and Kai Zhang and Jiangjie Chen and Renze Lou and Yu Su. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts, The Twelfth International Conference on Learning Representations, ICLR 2024 (Spotlight), 2024.

[6] 公益社団法人 日本技術士会．過去問題（第一次試験）（オンライン）（引用日: 2024 年 12 月 29 日.） https://www.engineer.or.jp/c_categories/index02021.html.

[7] Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, Xin Luna Dong, Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?, pp.311-325, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024.

[8] Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, Minjoon Seo. How Do Large Language Models Acquire Factual Knowledge During Pretraining? arXiv preprint arXiv: arXiv:2406.11813, 2024.

[9] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws, arXiv preprint arXiv:2404.05405, 2024. https://cthp.jacic.or.jp/jacic/doc/t02_12_jisseki.pdf.

[10] Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, William Yang Wang. Generalization vs. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. ICML 2024 FM-Wild Workshop, 2024.

[11] Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings, In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. ACL, pp.55–65, 2019.

[12] Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, Gongshen Liu. How Large Language Models Encode Context Knowledge? A Layer-Wise Probing Study, arXiv preprint arXiv:2402.16061, 2024.

[13] 一般財団法人日本建設情報総合センター（JACIC）．業務キーワード（オンライン）（引用日: 2025 年 1 月 6 日.）

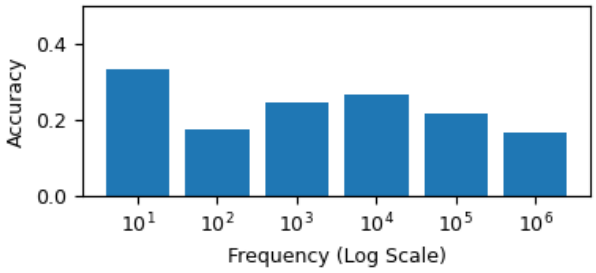
[14] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426, 2020.

[15] Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, Sadao Kurohashi. Beyond English-Centric LLMs: What Language Do Multilingual Language Models Think in? arXiv preprint arXiv:2408.10811, 2024.

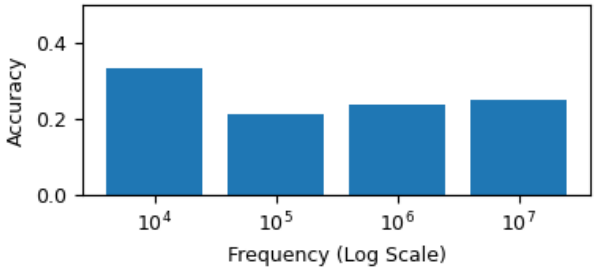
A 付録

付表 1　QA データセットサンプル

設問	選択肢	回答
土中の浸透と地下水に関する次の記述のうち、誤っているものはどれか。	1. 土中の水の流れに対しては、速度水頭は非常に小さいので無視することができ、全水頭は、圧力水頭と位置水頭の和で表される。 2. 飽和した多孔質媒体中の地下水の流量、速度は、ダルシーの法則に従う。 3. 地下水によって飽和されている透水性の良好な地盤を、帯水層という。 4. 土中の間隙において水が自由に上下に移動できる大気圧と等しい圧力を持つ地下水の表面を、地下水面という。 5. 下向きの浸透力によって、土中の有効応力が次第に減少してゼロになるような動水勾配を、限界動水勾配という。	5
長さ 2L[mm]の単純ばりの中央に集中荷重 P[N]が鉛直方向下向きに静的かつ弾性内で作用している。はりの断面二次モーメントは I[mm⁴]、ヤング率は E[N/mm²]であり、せん断変形は無視するものとする。この単純ばりの中央の鉛直方向たわみ δ [mm]として、正しいものは次のうちどれか。	1. PL³ / 48EI 2. PL³ / 24EI 3. PL³ / 12EI 4. PL³ / 6EI 5. PL³ / 3EI"	4

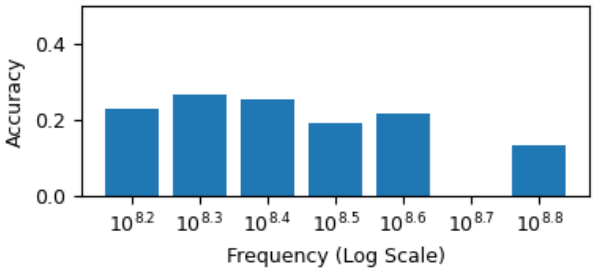


(b) 最も頻度の低い用語の頻度

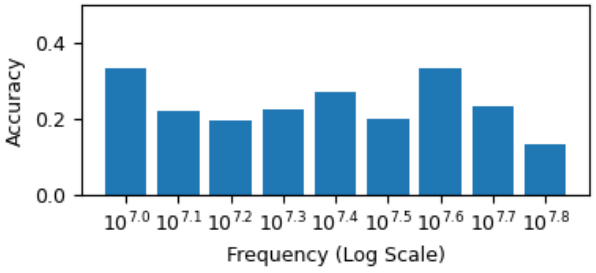


(c) 各用語頻度平均

付図 1　土木用語頻度と QA タスク精度の関係



(b) 全トークン頻度の平均値



(c) 全トークン頻度の中央値

付図 2　トークン頻度と QA タスク精度の関係