専門ドメインにおける LLM の文脈理解の分析

Analysis of LLM contextual understanding in specific domain.

緒方 陸[*1] 岡野 将大[*1] 大久保 順一[*1] 藤井 純一郎[*1]
Riku Ogata Masahiro Okano Junichi Okubo Junichiro Fujii

[*1] 八千代エンジニヤリング株式会社
Yachiyo Engineering Co., Ltd.

Large Language Models (LLMs) have low accuracy for long-tail knowledge, and solutions include Retrieval Augmented Generation (RAG) and fine tuning. On the other hand, there are reports that it is not practical due to factors such as differences in the evaluation task. In practical applications, more complex tasks can be solved and it is necessary to understand the context. However, it is not clear how LLMs internally processes texts and captures the context in specific domains such as civil engineering, which is a long-tail domain. We believe that clarifying how LLMs capture context in specific domains will contribute to improving for long-tail knowledge. In this study, we first identified uncertainty in response generation in both the general and specific domains, and then clarified the features of LLM internal processing processes in a specific domain by analyzing the variation of entropy in the intermediate representation of each layer of the LLM.

1. はじめに

大規模言語モデル（LLM）は long-tail の知識に対して回答精度が低く [Kandpal 2023]，この問題への対応として RAG（Retrieval Augmented Generation）や Fine-tuning がある．一方で既往の 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り [Zhou 2024]，その原因として対象ドメインの学習不足が挙げられる．

ここで，[Allen-Zhu 2024]は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．一方同手法は架空の事実に基づく合成データで評価しており，現実のデータとは複雑さなどの観点で大きく異なる．また，事前学習データに含まれる設問に関連するドキュメント数など，知識に関する事前学習データ頻度と QA 精度の関係を間接的に評価し，long-tail に関する事実ほど正答率が低いことを示した例もある（[Kandpal 2023], [Sun 2024]）．これらは単純な用語頻度よりも文章全体の文脈を踏まえた情報頻度の精度への寄与を示唆しており，LLM がテキストをどのように処理し，文脈を捉えるかを明らかにすることは LLM の long-tail の知識に対する性能向上に必要と考える．

本研究では，long-tail の知識である専門ドメインにおける LLM 内部処理プロセスの特徴を明らかにすることを目的とする．そのアプローチとして，一般／専門ドメインにおける中間表現のエントロピー変動を通じた LLM 内部処理過程の分析を行った．両ドメインの LLM 内部処理の違いから専門ドメインの特徴を明らかにすることで，専門ドメインの精度改善へ貢献可能と考える．

2. 関連研究

LLM の内部機構解明に向けた分析は，これまでに様々な方法で取り組まれている． Feed Forward 層のニューロンを分析した例や[Geva 2022]，Sparse AutoEncoder (SAE)を用いて LLM のスパースな潜在表現を分析した例[Elhage 2022]，Logit Lens を用いて LLM 各層のトークン予測を分析した例[Belrose 2023]などがある．

連絡先：緒方陸，八千代エンジニヤリング株式会社，東京都台東区浅草橋 5-20-8 CS タワー 技術創発研究所，TEL：03-5822-6395，Mail：rk-ogata@yachiyo-eng.co.jp

ここで，本稿では専門ドメインにおける LLM 内部の処理プロセスの特徴を明らかにすることを目的としており，個別のニューロンに着目した分析は本目的に合致しないと考える．また，SAE を専門ドメインに適用する場合，学習頻度の低さから適切な潜在表現が得られない可能性がある．[Belrose 2023]は Logit Lens は信頼できる解釈ができていない可能性を指摘しており，Tuned Lens を提案している．一方 Tuned Lens は学習が必要であり，本稿で対象とするデータ整備が進んでいない土木分野においては適用が難しい．

また別の手法で LLM の文脈理解について分析した例として，[Ju 2024]は Layer-wise Probing により LLM の各層で context knowledge（文脈を踏まえた知識）をどのようにエンコードしているかを調査した．同論文ではネットワーク前半の層で entity に関連する事実をエンコードし，後半の層で self-attention により文脈を捉えたエンコードが行われるとしている．[Chen 2024]は QA タスクにおいて Llama2-Chat-7B の情報量の大きい層（32 層中 26 層目）を対象とし，エントロピーベースの指標を用いて分析を実施した．正答の場合はエントロピーが低く，誤答の場合はエントロピーが高くなることを示している．しかし、専門ドメインと一般ドメインを比較したもので，かつ日本語を対象とした事例は著者らが確認した限りでは存在しない．

本稿ではまず，一般ドメインと専門ドメインにおける LLM の性能差の原因として回答生成における不確実性に着目し，これを確認した（3.1 節）．次に一般ドメイン／専門ドメインの上記不確実性の違いの原因を探るべく，LLM の中間表現のエントロピーの変動を分析することで（3.2 節），専門ドメインの LLM 内部処理過程の特徴を明らかにした（4 章）．最後に，実験結果に基づき考察を行い，現状の LLM の課題について整理した（4 章）．

3. 実験

3.1 各ドメインの評価タスクと使用したモデル

一般ドメインの評価タスクとして JCommonSenseQA (jcommon) [Kurihara 22] を，土木ドメインの評価タスクとして技術士一次試験（建設部門）の過去問 (gijutsushi) [日本技術士会 2025] を採用した．また，jcommon と比較して gijutsushi の方が設問文が長いことから，この差を軽減するため，jcommon の質

問文を GPT-4o により拡張した jcommon-long も一般ドメインの評価タスクとして加えた．各タスクの設問文長を表 1 に示す．なお，いずれも 5 つの選択肢から数値で回答する多肢選択形式であり，技術士一次試験の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）としている．

使用したモデルとして，オープンかつ日本語を多く学習している llm-jp-3-13b-instruct[1]を採用した．各タスクの評価結果を表 2 に示す．出力は数値のみとし，選択肢をランダムに入れ替えて 3 回試行した結果の平均を Accuracy Score とした．次節ではこのスコア別（0, 1/3, 2/3, 3/3）に分析を行う．参考として GPT-4o の結果も載せている．表より，llm-jp-3-13b-instruct は jcommon や jcommon-long と比較して gijutsushi では大きくスコアを落としている．一方 GPT-4o は 8 割程度の正答率であった．

また，各タスクの設問から GPT-4o によりキーワード抽出を行い，llm-jp-3-13b-instruct の事前学習データに含まれるのキーワード頻度の確認を行った．今回は日本語を対象としているため，事前学習データのうち，Wikipedia，Common Crawl (level2)，WARP/PDF (e0 / e0.2)，WARP/HTML，Kaken を分析対象とした[2]．両分布は重なっているものの，jcommon の分布は頻度が大きい方に，gijutsushi の分布は頻度が小さい方に偏っている．jcommon-long は両者の中間の特徴を示した．

3.2 分析手法と結果

一般ドメインと専門ドメインにおける LLM の回答生成確率の違いを確認し，その原因分析として中間表現の比較を実施した．

(1) 一般ドメインと専門ドメインの生成確率

表 2 のスコアおよび図 1 の違いから，専門ドメインでは学習不足に伴い回答生成の不確実性が高い可能性がある．回答生成の不確実性の評価として生成確率を用い，両ドメインの比較を実施した．0.05 刻みの生成確率の密度分布を図 2 に示す．

図より，jcommon および jcommon-long は正答率が高いほど生成確率も高い位置に分布している．一方 gijutsushi ではいずれのスコアにおいても 0.2 付近にピークが存在する．また，gijutsushi のスコアが 3/3 のケースにおいては，jcommon と jcommon-long 同様高い生成確率となる設問も確認した．

(2) LLM の中間表現のエントロピーの比較

前節で gijutsushi は他のタスクと比較して LLM の回答生成における不確実性が高かった．この原因として情報の圧縮度が小さいと考え，これに着目する．ここで既往研究として Attention map を分析したものもあるが，Transformer Layer 内部の FFN 層や LN 層は特定の言語構成の強調や，冗長性を削減するとされる[Kobayashi 2024]．本稿では[Chen 2024]と同様の手法を採用し，中間表現として Transformer Layer 全体の出力を分析することでこれらの影響を考慮しないこととした．

一般ドメインおよび専門ドメインにおける LLM の各 Transformer Layer の中間表現のエントロピーの比較を実施する．本稿で定義するエントロピーH_{hidden}を以下に示す．

$$H_{hidden} = -\sum_i p_i \log(p_i + \epsilon) \tag{1}$$

$$p_i = \frac{e^{h_i}}{\sum_j e^{h_j}} \tag{2}$$

ここで，h_iは LLM の各層の中間表現，p_iは各次元の寄与度を確率的に表現した値，i, j は中間表現の次元を示す．ϵは対数計算の安定化項であり 10^{-9}を設定した．

結果を図 3 に示す．なお，横軸の 0 層目は Embedding Layer，1 層目以降は Transformer Layer の出力であり，色は各タスクのスコアを示す。図より，いずれのタスクも全体の傾向は同様で，0～2 層目でH_{hidden}が急激に低下し，その後 22 層目までは 0～0.05 付近で推移，その後 36～38 層目までは 0.2 付近で推移した後，0.1 程度まで低下している．詳細な考察は次節で行う．

表 1　QA タスク設問文長

Task	Question length		
	Min	Max	Mean
jcommon	12	103	31
jcommon-long	92	475	284
gijutsushi	50	614	325

表 2　QA タスク評価結果

Model	Accuracy Score		
	jcommon (n=1119)	jcommon-long (n=1119)	gijutsushi (n=436)
llm-jp-3-13b-instruct	0.81	0.88	0.26
gpt-4o-2024-08-06	0.97	0.95	0.78

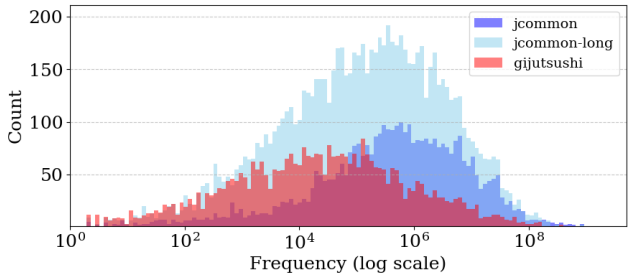

図 1　事前学習データにおける各データセットのキーワード頻度

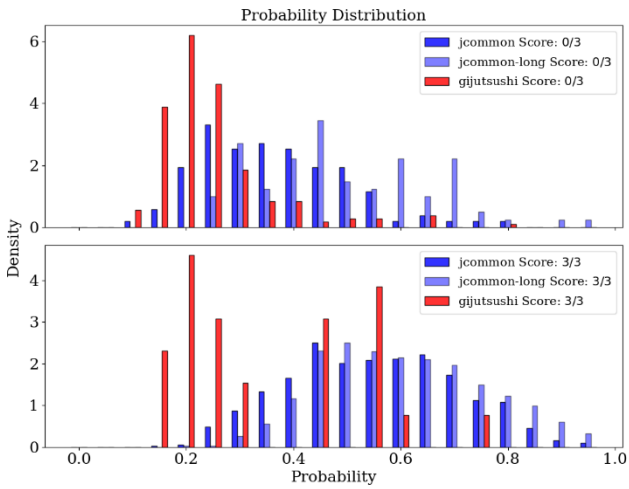

図 2　一般ドメインと専門ドメインにおける回答生成確率（上段：Accuracy 0%，下段：Accuracy 100%）

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

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

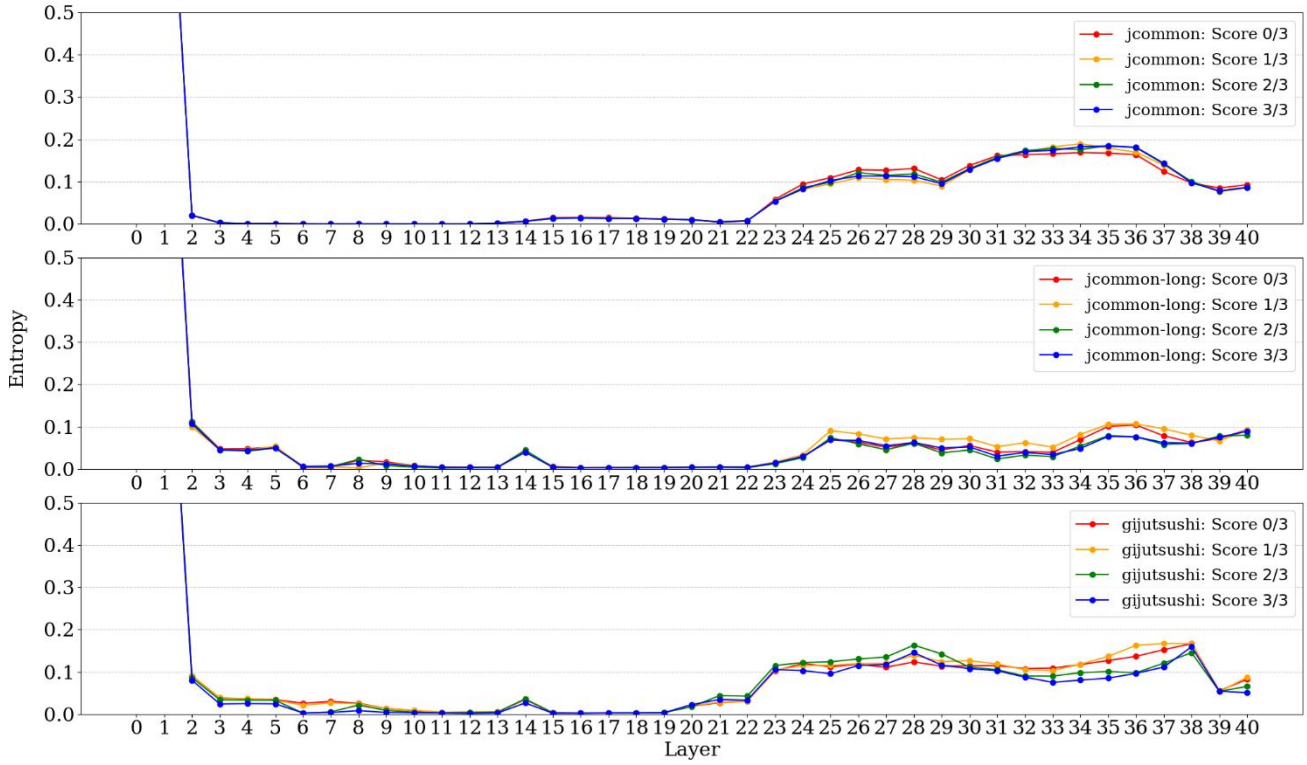

図 3　LLM 各層のエントロピー（上段：jcommon，中段：jcommon-long，下段：gijutsushi）

4. 考察

3.1 節の実験結果で生成確率が一般/専門ドメインで異なる原因について，3.2 節の実験結果から考察する．

gijutsushi の回答生成における不確実性が高い原因として，LLM が専門用語の語彙を獲得できなかったことと（4.1 節），文脈を捉えられなかったこと（4.2 節）を挙げる．また，これらを踏まえた結果から，LLM の専門ドメインにおける処理の特徴と課題を整理する（4.3 節）．以下では上記の原因も含め，各タスクにおける LLM の処理プロセスの違いについて考察する．

4.1 語彙の獲得

図 3 より，全体の傾向としてはいずれのタスクも同様の傾向を示したものの，gijutsushi は 6〜7 層目でスコアが低い場合（0/3, 1/3）にH_{hidden}が高く出ている．

ここで，サブワード分割された単語が LLM のどの層で単語に変換されるかを調査した[Kaplan 2024]によると，サブワード分割される／されないに関わらず，多くは 3〜5 層目で単語に変換され，最大で 15 層目まで到達する token もあるとしている．また同著者らは，LLM は潜在的な語彙（latent vocabulary）を持ち，tokenizer の語彙に無い語彙を内部的に構成できるとする一方，表現できない未知語も存在すると述べている．

図 3 は，3〜20 層目前後まではいずれもH_{hidden}は低い値を示すが，設問文長が長い jcommon-long と gijutsushi は 3〜10 層目で jcommon と比較してやや高い値を示した．前者二つのタスクは設問文長が長い故に語彙数も多く，情報の圧縮に処理を多く要したと考える．また，gijutsushi は 6〜7 層目で低スコア／高スコアで差が出た．gijutsushi は土木分野の専門用語を多く含み，tokenizer の語彙に含まれない未知語が多い．[Kaplan 2024]の主張を踏まえると，LLM の latent vocabulary に新たに語彙を構成するために処理を多く要した，または専門用語が latent vocabulary で未知語と分類されたことが，gijutsushi の低スコアの設問における回答生成の不確実性が高い原因であると推察する．

4.2 文脈理解

図 3 より，設問文長の長い jcommon-long および gijutsushi では 35 層目前後で低スコアの場合（0/3, 1/3）の方が高スコアの場合（2/3, 3/3）よりもH_{hidden}が高い．

ここで，[Chen 2024]はネットワーク後半の層を用い，QA タスクで正答率の高い（低い）設問ほどエントロピーが低い（高い）ことを示した．また[Ju 2024]によると，LLM の後半の層では self-attention により文脈を捉えたエンコードが行われるとされる．

jcommon と比較して設問文長の長い jcommon-long および gijutsushi が 35 層目前後において低／高スコアでH_{hidden}の差が出た原因として，文脈を捉えて正答に資する特徴を獲得できなかった／獲得できたためと推察する．35 層目付近で低スコアの方がいずれのタスクもH_{hidden}が高く，情報圧縮度が小さい．これにより，回答生成の不確実性が高まったと考えた．

また，jcommon-long と gijutsushi を比較すると 35 層目付近では jcommon-long の方がH_{hidden}は低い値をとる．これは jcommon-long は gijutsushi と比較すると文脈を捉えることができ，回答生成における不確実性を軽減できたと推察する．対して jcommon のH_{hidden}はスコアに依らず 35 層目付近で gijutsushi と同程度のピークを迎えるが，これについては次節で考察する．

4.3 専門ドメインにおける処理の特徴と課題整理

上記の考察を踏まえると，図 3 より全体の傾向として 0〜2 層目で token の圧縮を行い，3〜20 層目付近までで語彙を獲得する．その後 35 層目前後までで文脈を捉え，それ以降でタスクに沿った形式に変換しているものと推察した．

ここで gijutsushi の 20〜25 層目，35〜38 層目に着目する．一般ドメイン（jcommon, jcommon-long）は 20〜22 層目でH_{hidden}が上昇していないのに対し，専門ドメイン（gijutsushi）では上昇

している．また，いずれのドメインでも23層目からH_{hidden}が上昇しているが，gijutsushiは他に比べ勾配が急である．これは23層目が [Ju 2024]の主張の entity から context knowledge への変換を行う機能を持つ一方，gijutsushi は 22 層目までに情報を圧縮できず token から entity への変換が完了していないため，急な情報展開となり，H_{hidden}が上昇したものと推察した．さらに，gijutsushi は 38 層目にピークがあり，その後急激にH_{hidden}が減少している．これも上記と同様に context knowledge の変換が完了しておらず，タスクに沿った形式への変換により急激なH_{hidden}の低下が生じたものと考えた．対して jcommon のH_{hidden}はスコアに依らず 35 層目付近でピークを迎えた後，なだらかに減少している．これは gijutsushi と異なり context knowledge の変換がタスクに沿った形式へ適切に変換できているが，文脈が無いため高スコアの場合も回答が絞り切れず，低／高スコアで同程度のH_{hidden}になったと推察した．この結果は図 2 にも表れており，jcommon の方が jcommon-long よりも全体的に低い生成確率を示している．

ここで図 1 より，今回対象とした LLM は専門ドメインの方が一般ドメインよりもキーワード頻度が低く，関連するトークン頻度も低いと考える．以上より，事前学習データに含まれるトークン頻度が原因で，専門ドメインは一般ドメインと比較して全体的に entity や context knowledge のエンコードが後半の層へシフトし，結果として設問の文脈を捉え切れずに回答生成における不確実性が高まったと推察した．この問題への対応として，専門ドメインにおける LLM の課題を整理する．以下の課題は既往研究（[Kandpal 2023], [Tao 2024]）でも関連した内容について述べられており，本成果はそれらを裏付けるものとなった．また，クローズドモデルに対して以下を実施することは困難であるが，RAG における embedding model にも同様の効果があると考える．

- 専門ドメインの学習頻度を上げるため，専門ドメインの事前学習データを増加させる
- LLM の Transformer Layer を増やし，事前学習を行う
- tokenizer の vocabulary に専門用語を追加する

5. おわりに

本稿では，一般／専門ドメインを対象に LLM の回答生成の不確実性の違いを確認した．また，LLM の中間表現のエントロピーを指標として両者の違いが生じた原因を分析し，結果から，専門ドメインは一般ドメインよりも entity や context knowledge のエンコードが後半の層へシフトしており，これが専門ドメインにおける回答生成の不確実性を高める原因であると結論付けた．

一方，本稿で使用したモデルは llm-jp-3-13b-instruct であり，英語を多く学習したモデルや，モデルサイズなどにより異なる挙動を示す可能性がある．これらの検討は今後の課題とする．

参考文献

[Kandpal 2023] 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.

[Zhou 2024] 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.

[Allen-Zhu 2024] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models, Part 3.3, Knowledge Capacity Scaling Laws, arXiv preprint arXiv:2404.05405, 2024.

[Sun 2024] 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.

[Geva 2021] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg, Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, p.p.30–45, 2022.

[Elhage 2022] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022b. Toy models of superposition. Transformer Circuits Thread. https://transformer-circuits.pub/ 2022/toy_model/index.html, 2022.

[Belrose 2023] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112.

[Ju 2024] 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.

[Chen 2024] Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, Junxian He: In-Context Sharpness as Alerts: An Inner Representation Perspective for Hallucination Mitigation, Proceedings of the 41st International Conference on Machine Learning, PMLR 235:7553-7567, 2024.

[Kurihara 22] Kurihara, K., Kawahara, D., and Shibata, T.: JGLUE: Japanese General Language Understanding Evaluation, in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp.2957–2966, Marseille, France (2022), European Language Resources Association, 2022.

[日本技術士会 2025] 公益社団法人 日本技術士会: 過去問題（第一次試験）https://www.engineer.or.jp/c_categories/index02021.html.（2025 年 2 月 3 日アクセス）.

[Kobayashi 2024] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, Kentaro Inui: Analyzing Feed-Forward Blocks in Transformers through the Lens of Attention Maps, ICLR 2024 spotlight, 2024.

[Kaplan 2024] Guy Kaplan, Matanel Oren, Yuval Reif, Roy Schwartz: From Tokens to Words: On the Inner Lexicon of LLMs, ICLR 2025, 2025.

[Tao 2024] Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, Ngai Wong: Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies. NeurIPS 2024 poster, 2024.