英語中心の大規模言語モデルの言語横断汎化能力

謝 素春 [1]　佐々木 翔大 [3]　Yunmeng Li[1]　坂田 将樹 [1,2]
赤間 怜奈 [1,2]　鈴木 潤 [1,2]
[1] 東北大学　[2] 理化学研究所　[3] 株式会社サイバーエージェント
{xie.suchun.p7, li.yunmeng.r1, sakata.masaki.s5}@dc.tohoku.ac.jp
sasaki_shota@cyberagent.co.jp　{akama, jun.suzuki}@tohoku.ac.jp

概要

LLaMA などに代表される大規模言語モデルは，事前学習データの大半が英語で構成されているにも関わらず，英語以外の言語についても文章の生成が可能であることが観察されている．本研究では，英語を中心に学習された大規模言語モデルに対して，英語データのみを用いて微調整学習 (fine-tuning) を実施した場合でも，英語以外の言語におけるタスク性能が向上する言語横断汎化能力に関して分析を行う．具体的には，英語の微調整学習前後でモデルの異言語間の内部表現に対する分析の結果，英語の微調整学習は言語非依存のタスクの解き方を学習していること，また，事前学習した言語間の類似度より性能汎化を実現していること示す．

1　はじめに

言語横断汎化 (cross-lingual generalization) とは，モデルがあるソース言語でのタスク学習を通じて，ソース言語と異なる言語でも同じタスクをこなせるようになることである．言語横断汎化能力の実現は特にデータの少ない低資源言語のタスク性能の改善を期待できるため，重要視されている [16, 15, 12]．

昨今，大規模言語モデル（LLM）は顕著な性能向上を遂げており [3]，BLOOM [21] を始めとする多言語 LLM の言語横断汎化に関する研究も盛んである [8, 12]．．一方で，LLaMA [18] のような事前学習データの大半が英語で構成される大規模言語モデル（以下，**「英語中心の LLM」**と言う）においても，英語以外の言語，例えば中国語なども生成可能なことが観察されている [23]．また，事前学習に明示的に含まれない言語に対する言語横断汎化能力を示唆する結果も報告されている [26]．しかしながら，英語中心の LLM の有する言語横断汎化能力に関して分析した研究はまだ限られている．英語のみでの微調整学習による言語横断汎化性能の解明はより効率的，かつ強力な LLM の構築につながる重要な知見と考えられる．

本研究では，まず事前実験として英語中心の LLM の言語横断汎化性能を多様なタスクにおいて観察した上で，**（1）英語中心の LLM の言語横断汎化能力はどこから来たのか**，**（2）言語横断汎化性能の言語間のばらつきをもたらす要因は何か**，という 2 つの研究課題に関して分析を行う．課題（1）に関しては，英語のみを使用した対照学習によって多言語モデルの多言語表現の質が向上したという先行研究での報告 [19] に基づき，「英語中心の LLM においても，英語で微調整学習することが，多言語タスクをうまく解くためのより良い多言語表現を獲得することに繋がる」という仮説を立て検証する．具体的には，英語での微調整学習実験で汎化性能を確認した後，微調整学習がモデルの多言語表現能力にどのような影響を与えるかを焦点に当て，この過程で生じる言語間の類似度変化を分析する．さらに課題（2）に関して，言語間の類似度と性能汎化の関係性を調査するために，二者の相関性に対して分析を行う．

分析の結果，モデルは事前学習で得られた言語間のアラインメントを手がかりに英語データによる微調整学習ではタスクそのものの解き方を学習することで言語横断汎化能力を獲得していると推測できる．

2　関連研究

英語中心の LLM に対する言語横断汎化に関する研究は，言語横断汎化能力に対する調査と分析に分けられる．

言語横断汎化能力調査　言語横断汎化能力に対する調査において，Bandarkar ら [2] は英語中心の LLM と従来の事前学習されたモデルとの比較を中心に，LLM に対して文脈内学習（in-context learning）手法

を用いて性能評価を行った．そのほか，Ye ら [26] は英語を含む単言語での微調整学習を用いて，推論タスクにおいて多言語 LLM と英語中心の LLM の言語横断汎化能力を検証した．

言語横断汎化性能分析　言語横断汎化能力に対する分析において，Xu ら [22] と Zhu ら [28] はモデルの言語横断汎化性能を向上するために，言語間のアライメントを強化する手法を提案した．また，Zhao ら [27] は微調整学習前後のパラメータに対して分析する手法を用いた．性能汎化調査の部分では，Ye ら [26] の研究においても分類タスクで英語での微調整学習による汎化性能を検証したが，彼らの研究と違い，本研究では分類タスクに加えて，英語中心の LLM が生成タスクにおける言語横断汎化能力の調査しする．さらに，モデルの言語横断汎化能力はどこから来たのかという本質的な問題に対して重点に分析を行う．

3　英語中心の LLM の言語横断汎化

事前実験として，英語中心の LLM の言語横断汎化性能を複数の言語と多様なタスクを用いて調査し，研究課題の全体像を把握する．具体的には，英語中心の LLM を**英語データのみ**で微調整学習した際に，英語以外の言語における汎化性能を測る．

モデル　英語中心の LLM として LLaMA-7B モデル [18] を使用し，指示文微調整学習 (instruction tuning) [20] を行った．LLM の効率的な微調整学習手法が多く提案されているが [9, 6, 14]，本研究では LoRA [9] を採用した．

データセット　分類タスクと生成タスクの 2 種類のタスクを用いた．分類タスクのデータセットは二値分類の言い換え判定タスクである PAWS-X [25] と三値分類のテキスト含意判定タスクの XNLI [4] を用いた．生成タスクのデータセットは要約データセット XL-Sum [7] を用いた．いずれのデータセットも，英語を含む多言語で構成されている．PAWS-X においては学習データ全件で学習を行い，XNLI と XL-Sum ではそれぞれ学習データセットの 10 万および 5 万件を用いた．プロンプトおよびパラメーター設定の詳細は表 4に記載した．

評価　英語の学習データで微調整学習したモデルを用いて，各言語ごとにテストデータで 0-shot 性能の評価を行った．詳細設定は表5に示した．　本研究では英語（en），フランス語（fr），ドイツ語（de），スペイン語（es），中国語（zh），日本語（ja）の 6 つの言語を対象とした．XNLI に日本語のデータは存在しないため，本研究では JNLI [11] の検証データで評価を行った．評価時は統一的に英語のプロンプトを使用した．

表 1　微調整学習前の事前学習済みモデル (LLaMA-7B) と微調整学習後のモデル (FT) の性能評価結果

	en	fr	de	es	zh	ja
PAWS-X	**Acc.**					
LLaMA-7B	0.54	0.54	0.51	0.54	0.55	0.56
PAWS-X FT	**0.95**	0.87	0.87	0.86	0.75	0.71
XNLI	**Acc.**					
LLaMA-7B	0.33	0.33	0.33	0.33	0.33	0.15
XNLI FT	**0.88**	0.80	0.79	0.77	0.60	0.56
XL-Sum	**ROUGE-L**					
LLaMA-7B	0.08	0.12	-	0.13	0.19	0.18
XL-Sum FT	**0.32**	0.27	-	0.22	0.18	0.21

結果　評価結果を表 1に示す．微調整学習前のモデル（LLaMA-7B）の結果から，微調整学習前のモデルは 3 つのタスクにおいてもランダム回答に近いことがわかる[1)]．一方，微調整学習を行った場合，分類タスク（PAWS-X，XNLI）において，英語を含む全言語での性能が顕著に向上していることが観測できた．分類タスクにおけるこのような観測は Ye ら [26] の報告と一貫している．さらに，微調整学習後の XL-Sum の結果は，ROUGE-L スコアが中国語以外の全言語に渡って向上した．これらの結果から，英語のみで微調整された英語中心の LLM である LLaMA が分類タスクだけでなく，生成タスクにおいても言語横断的な汎化能力を有することがわかった．

LLaMA の事前学習では英語中心に，ラテン文字（en，fr，de，es）及びキリル文字（ru）を使用する言語を用いており [18]，日本語と中国語は事前学習，微調整学習のデータに**明示的には含まれていない**にも関わらず，日本語と中国語でも性能の向上が見られた．同時に，他の言語での性能は英語と比較してばらつきがあることが観察され，フランス語とドイツ語は一貫して中国語と日本語より高い性能を達成している．そこで，英語中心の LLM の言語横断汎化能力はどこから来たのか，汎化能力のばらつきはどの要素に影響されるのかという疑問が残される．

1)　例外として，要約タスクの XL-Sum では中国語と日本語の ROUGE スコアは比較的に高かったものの，出力結果に対する定性分析を行ったところ，その 2 つの言語の予測は単に入力文の最初の文を抜き出していることがわかった．

表 2 PAWS-X データセットにおける事前学習済みモデルとファインチューニング済みモデルの Mean-pooling および Last-token 埋め込みに対するコサイン類似度．”-P”は対訳文ペア，”-R”はランダム文ペアを表す．

	Mean-pooling Embedding						Last-token Embedding					
	微調整学習前		分類タスク		生成タスク		微調整学習前		分類タスク		生成タスク	
	Before-P	Before-R	Pawsx-P	Pawsx-R	Xlsum-P	Xlsum-R	Before-P	Before-R	Pawsx-P	Pawsx-R	Xlsum-P	Xlsum-R
PAWS-X Sentence 1												
en-fr	0.54	0.34	0.55	0.36	0.55	0.36	0.50	0.39	0.48	0.36	0.47	0.36
en-de	0.50	0.32	0.51	0.34	0.51	0.34	0.37	0.28	0.34	0.25	0.31	0.23
en-es	0.49	0.29	0.51	0.32	0.50	0.32	0.38	0.28	0.36	0.25	0.33	0.23
en-zh	0.26	0.17	0.26	0.18	0.28	0.20	0.31	0.25	0.32	0.25	0.33	0.27
en-ja	0.22	0.15	0.23	0.16	0.24	0.17	0.31	0.27	0.33	0.27	0.33	0.28

4 分析と考察

4.1 言語間埋め込み類似度の変化

本節では，英語のみで微調整学習する際，英語中心の LLM の言語横断汎化能力がどのような要因によって引き起こされているかについて分析と考察を行う．§3の実験では，微調整学習によって英語以外の言語での性能向上が見られ，モデルの言語横断汎化能力は微調整学習によるものであると推測される．また，多言語モデルの場合，英語のみを使用した対照学習によって多言語の埋め込み表現の質を向上させることが先行研究で報告されている [19]．これらを踏まえ，英語中心の LLM が言語横断汎化を実現する要因は，**英語で微調整学習することを通じて，多言語タスクをうまく解くためのより良い表現を獲得できたことにある，つまり，異言語で同じ意味を持つ対訳文の埋め込み表現の類似性が向上したことにある**，と仮説を立てる．この仮説を検証するために，英語のみでの微調整学習がモデルの内部の各言語表現に，どのような影響を与えるかを定量的に分析を行う．具体的には，微調整学習前後の対訳文の言語間類似度の変化を比較し，英語のみの微調整学習が他の言語への影響を調査する．なお，本研究では言語間の意味的類似度に焦点を当て，Li ら [13] の研究に従い，コサイン類似度を用いた言語間類似度を計算する．また，モデルが対訳文の意味を正しく理解できることを検証するため，対訳文ペアとランダム文ベアの類似度をそれぞれ計算する．

手順　ソース言語（英語）と各ターゲット言語（英語以外の言語）の対訳文ペアとランダム文ペアそれぞれに対して，微調整前と後のモデルを用いて文の埋め込み表現を獲得し，文ペアのコサイン類似度を計算する．LLM の文埋め込みを分析する際は，入力文の全トークンの隠れ表現の平均 (Mean-pooling embedding)[17]，または入力最後尾のトークンに対応する隠れ表現 (Last token embedding) を文埋め込み [10] とするのが一般的であるが，Decoder-only のモデルにおいて，どちらの手法がより適切な表現を獲得できるかは自明ではないため，本研究では両方の文埋め込みを用いる．

モデル　事前学習済みモデル（LLaMA-7B）と微調整学習した 2 つのモデル（PAWS-X FT，XL-Sum FT）を用いた．

データセット　対訳文データとして PAWS-X のテストデータを用いた．PAWS-X の各サンプルは 2 つの文を含むが，1 文目を用いた．各言語対のデータ数はそれぞれ 2000 件である．

結果　文ペアに対する類似度の平均値の結果を表 2に示す．表 2より，対訳文ペアのコサイン類似度はランダム文ペアより高く，対訳文ペアに対して内部でアラインメントが高く取れていることが示唆される．微調整学習前後の対訳文における類似度の変化は，我々の予想に反して小さい（0 から 3 ポイントの範囲内）ことがわかった．この観察と，微調整学習によって言語横断汎化が起きていることを踏まえると，**モデルは微調整学習を通じて多言語表現の質を向上させたわけではなく，事前学習時に既に獲得していた多言語表現に基づき，言語横断的なタスクを解く能力を学習した**ことが推察される．

また，対訳文データセットにおいて，分類と生成タスクで微調整学習したモデルは異なる傾向が観察される．Mean-pooling 文埋め込みからでは，微調整したモデルの言語間類似度は 2 つの分類データセットで 1 から 3 ポイントを変化した．Last-token 埋め込みでは，Mean-pooling 文埋め込みの類似度変化より大きく (1〜7 ポイント)，傾向が見られなかった．これは，微調整プロセスが文の末尾の処理において

表 3 英語中心の LLM を英語データのみで微調整学習をした際の性能と，微調整学習前の各言語間の埋め込みのコサイン類似度の Spearman 相関 (ρ)．分類タスクで微調整学習したモデルは Mean-pooling 文埋め込みとの相関係数が高く，生成タスクで学習したモデルは Last-token 文埋め込みとの相関係数が高いことを表す．

	Mean-pooling		Last-token	
	Spearman's ρ	p 値	Spearman's ρ	p 値
PAWS-X Sentence 1				
PAWS-X	**0.99**	0.01	0.70	0.19
XNLI	**0.97**	0.01	0.75	0.14
XL-Sum	0.80	0.20	**0.97**	0.03
PUD				
PAWS-X	**0.97**	0.01	0.88	0.05
NLI	**0.93**	0.07	0.81	0.19
XL-Sum	0.81	0.19	**0.94**	0.06

特定の言語表現の偏りをもたらしている可能性を示唆している．

4.2 言語間類似度と汎化能力の関係

表 2に基づくと，微調整学習前の英語とフランス語（en-fr），やスペイン語（en-es）間のコサイン類似度は英語と中国語（en-zh），日本語のコサイン類似度（en-ja）よりも一貫して高くなり，汎化性能のばらつきと同じ傾向を示している．LLaMA のような英語中心のモデルでは，他言語の学習データの割合は極めて不均衡という問題があるが，これまでの多言語モデルの研究では，言語間のアラインメントがモデルの言語横断能力に寄与してると主張されている [5, 24]．そこで，英語のみで学習したタスクの解き方は，事前学習の段階で構築した言語間表現の類似度を通じて，汎化したのではないかを推測される．

言語間表現の類似度はモデルの言語横断汎化能力との関係を分析するために，表 1で得られた英語以外の 5 つ言語でのモデル性能と，前節で観察された事前学習済みモデルから取得した文埋め込みの言語間類似度との Spearman の順位相関係数を計算した．

結果 結果を表 3に示す． 表3より，Spearman の相関係数はいずれも 0.7 より高く，言語間の類似度と言語横断汎化能力との間に強い相関関係が存在することを示している．また，PAWS-X のような分類タスクで微調整学習したモデルでは，Last-token 文埋め込みの相関と比べて (相関係数 0.70，p 値: 0.19)，Mean-pooling 文埋め込みの相関は高い値を示している (相関係数 0.99; p 値: 0.01)．これは分類タスクで微調整したモデルでは，平均的な文脈情報がモデルの言語横断汎化能力に重要であることを示唆してる．その一方で，生成タスク（XL-Sum）で学習したモデルでは，Mean-pooling 文埋め込みの相関より，Last-token 文埋め込みの相関が高い傾向が見られる．

考察 以上の結果から，モデルの多言語汎化性能は事前学習によって獲得した表現に対する言語間の類似度と強く関連していることが明らかとなった．すなわち，多言語汎化性能のばらつきは，言語間類似度のばらつきの影響を受けてることが示唆される．また，前節で観察された言語間の類似度変化より，英語で学習したタスクの解き方は，事前学習で学習した英語と各言語間のアラインメントにより，他の言語にも適用されることが考えられる．そのほか，Mean-pooling 文埋め込み類似度と Last-token の文埋め込み類似度をそれぞれいつ使うかを明らかにした．分類タスクで微調整学習したモデルでは，ソース言語との Mean-pooling 文埋め込み類似度が性能汎化の鍵となり，生成タスクで微調整したモデルでは，Last-token の文埋め込み類似度が言語横断汎化能力に重要であると分かった．ただし，事前学習の言語間類似度から言語横断汎化能力を完全に解釈するのは不可能であり，言語間の構造など他の要素からの影響を今後の課題としたい．

5 おわりに

本研究では，英語中心の LLM に対して英語のみでの微調整学習を行うことで，同一タスクにおける言語横断汎化性能を有することを示した．それに基づいて，微調整学習前後でモデルの異言語間の内部表現の類似度比較より，微調整学習は言語非依存のタスクの解き方を学習している可能性を示した．さらに，相関分析により，言語間の類似度と言語横断汎化能力の間には強い相関関係が存在し，事前学習で得られた言語間の類似度が微調整学習による多言語性能の向上に重要な役割を果たしていることが示唆された．この一連の知見は，英語中心の LLM の微調整学習と言語横断汎化能力に関する新たな洞察を提供する．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research），JST 次世代研究者挑戦的研究プログラム JPMJSP2114，JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2102，JSPS 科研費 JP22K17943 の支援を受けたものです．また，本研究に助言を下さった栗田宙人氏，小林悟郎氏，横井祥氏に感謝致します．

参考文献

[1]S. H. Bach, V. Sanh, Z.-X. Yong, A. Webson, C. Raffel, N. V. Nayak, A. Sharma, T. Kim, M. S. Bari, T. Fevry, Z. Alyafeai, M. Dey, A. Santilli, Z. Sun, S. Ben-David, C. Xu, G. Chhablani, H. Wang, J. A. Fries, M. S. Al-shaibani, S. Sharma, U. Thakker, K. Almubarak, X. Tang, X. Tang, M. T.-J. Jiang, and A. M. Rush. Promptsource: An integrated development environment and repository for natural language prompts. In **arXiv:2202.01279**, 2022.

[2]L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In **arXiv:2308.16884**, 2023.

[3]H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. In **arXiv:2210.11416**, 2022.

[4]A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2018.

[5]A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)**, 2020.

[6]T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[7]T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, **Findings of the Association for Computational Linguistics (ACL-IJCNLP)**, 2021.

[8]E. Hlavnova and S. Ruder. Empowering cross-lingual behavioral testing of NLP models with typological features. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2023.

[9]E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations (ICLR)**, 2022.

[10]T. Jiang, S. Huang, Z. Luan, D. Wang, and F. Zhuang. Scaling sentence embeddings with large language models, 2023.

[11]K. Kurihara, D. Kawahara, and T. Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)**, 2022.

[12]J. Li, H. Zhou, S. Huang, S. Cheng, and J. Chen. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. In **arXiv.2305.15083**, 2023.

[13]T. Li and K. Murray. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In **Findings of the Association for Computational Linguistics (ACL)**, 2023.

[14]Y. Li, Y. Yu, C. Liang, P. He, N. Karampatziakis, W. Chen, and T. Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models, 2023.

[15]X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, and X. Li. Few-shot learning with multilingual generative language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, 2022.

[16]N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel. Crosslingual generalization through multitask finetuning. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[17]D. Nikolaev and S. Padó. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In **Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**, 2023.

[18]H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models. In **arXiv:2302.13971**, 2023.

[19]Y. Wang, A. Wu, and G. Neubig. English contrastive learning can learn universal cross-lingual sentence embeddings. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, 2022.

[20]J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. In **arXiv:2109.01652**, 2022.

[21]B. Workshop and et al. Bloom: A 176b-parameter open-access multilingual language model. In **arXiv.2211.05100**, 2023.

[22]N. Xu, Q. Zhang, J. Ye, M. Zhang, and X. Huang. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In **Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP2023 Findings)**, 2023.

[23]N. Xuanfan and L. Piji. A systematic evaluation of large language models for natural. In **Proceedings of the 22nd Chinese National Conference on Computational Linguistics**, 2023.

[24]H. Yang, H. Chen, H. Zhou, and L. Li. Enhancing cross-lingual transfer by manifold mixup. In **The Tenth International Conference on Learning Representations (ICLR)**, 2022.

[25]Y. Yang, Y. Zhang, C. Tar, and J. Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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)**, 2019.

[26]J. Ye, X. Tao, and L. Kong. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. In **arXiv:2306.06688**, 2023.

[27]J. Zhao, Z. Zhang, Y. Ma, Q. Zhang, T. Gui, L. Gao, and X. Huang. Unveiling a core linguistic region in large language models. In **arXiv:2310.14928**, 2023.

[28]W. Zhu, Y. Lv, Q. Dong, F. Yuan, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li. Extrapolating large language models to non-english by aligning languages. In **arXiv:2308.04948**, 2023.

A　パラメータ設定

A.1　学習時のパラーメータ

学習時のパラーメータ設定を表 4に示す．

表 4　英語のデータのみを用いた微調整学習時のハイパーパラメータ設定

ハイパーパラメータ	設定値
Model	LLaMa-7B
Epoch	6
Learning rate	2e-5
Learning rate scheduler	Cosine
LoRA rank	8
LoRA alpha	32

A.2　評価時のパラーメータ

評価時のパラーメータ設定を表 5に示す．

表 5　微調整学習したモデルを用いた性能評価時のパラーメータ設定

パラメータ	設定値
Temperature	0.1
Top_p	0.75
Top_k	1
Num_beams (XL-Sum)	4
Max_new_tokens (PAWS-X &XNLI)	1
Max_new_tokens (XL-Sum)	100

B　プロンプト設定

本研究で用いたプロンプト設定は PromptSource [1] のテンプレートに基づいたものである．詳細を表 6に示す．

表 6　プロンプト設定

Datasets	**Prompt**
PAWS-X	Sentence 1: {sentence 1} Sentence 2: {sentence 2} Question: Do Sentence 1 and Sentence 2 express the same meaning? Only answer with Yes or No. The answer is
XNLI	Suppose {premise} Can we infer that {hypothesis}? Yes, no, or maybe?
XL-Sum	Write one sentence to summarize the given document. Document is: {Input paragraph} Summarize: