Multilevel Analysis of Biomedical Domain Adaptation of Llama 2: What Matters the Most? A Case Study

V. Ivan Sanchez Carmona[1] and **Shanshan Jiang**[1] and **Takeshi Suzuki**[2] and **Bin Dong**[1]
[1]Ricoh Software Research Center (Beijing) Co., Ltd
[2]Ricoh
{Vicente.Carmona, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com
takeshi.suzuki@jp.ricoh.com

Abstract

Domain adaptation of Large Language Models (LLMs) leads to models better suited for a particular domain by capturing patterns from domain text which leads to improvements in downstream tasks. To the naked eye, these improvements are visible; however, the patterns are not so. How can we know which patterns and how much they contribute to changes in downstream scores? Through a Multilevel Analysis we discover and quantify the effect of text patterns on downstream scores of domain-adapted Llama 2 for the task of sentence similarity (BIOSSES dataset). We show that text patterns from PubMed abstracts such as clear writing and simplicity, as well as the amount of biomedical information, are the key for improving downstream scores. Also, we show how another factor not usually quantified contributes equally to downstream scores: choice of hyperparameters for both domain adaptation and fine-tuning.

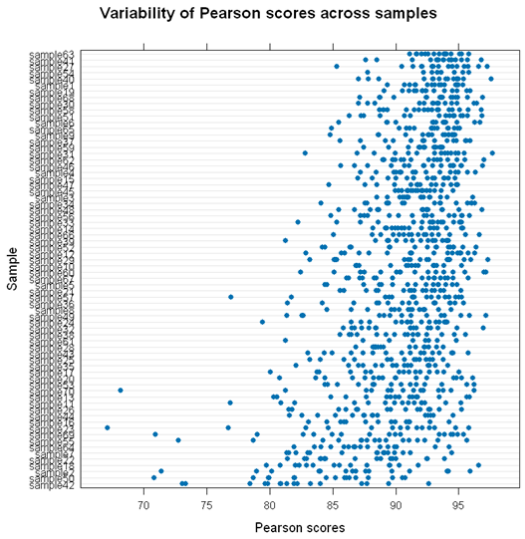

Figure 1: Variations in downstream score depending on the choice of both sample used for DAPT and hyperparameters. Each row represents a sample; each dot a Pearson score from a fine-tuned model (BIOSSES downstream dataset).

1 Introduction

Domain Adaptive Pretraining (DAPT) is an effective method to adapt a model to a particular domain via continual pretraining (Gururangan et al., 2020; Rietzler et al., 2020), with BioBERT (Lee et al., 2019) being a successful case in the biomedical domain. From our side, we have widely used DAPT not only to adapt LLMs to biomedicine, but as a part of a bigger pipeline to create customer-oriented, intelligent agents which can faithfully recover domain knowledge from both their parameters and external databases. However, domain adapting Llama 2 (Touvron et al., 2023) brought us a puzzle: huge variability on downstream scores.

Given both the size of Llama 2 and GPU memory restrictions, the domain adaptation of Llama 2 was restricted to a sample (subset) of PubMed abstracts. After domain-adapting and fine-tuning Llama 2 on PubMed abstracts and on the BIOSSES dataset (Soğancıoğlu et al., 2017), respectively, we obtained highly different downstream scores depending on the choice of the sample used for DAPT, as displayed in Fig. 1. This figure shows a huge variability in Pearson (downstream) scores: from 67 points up to an almost perfect score of 98 points. Surely, hyperparameter choice for both DAPT and fine-tuning has an impact on the scores, but, by comparing the patterns of score variation across samples used for DAPT we observe that this variation does not seem to be explained only by the choice of hyperparameter values. Clearly, data used for DAPT has also an impact on the scores.

Thus, we asked: What features from the samples used for DAPT impact on the downstream scores and to what extent? And to what extent is the impact of the choice of hyperparameter values? We hypothesized that text patterns, such as sentence length, syntactic dependencies, or text complexity,

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 449–456

among others, impact on the downstream scores. In order to analyze the importance of each text feature, and the effect of hyperparameters, we used Multilevel Modeling, a regression model widely used in the Social Sciences to explain social phenomena such as the effect of both school and student features on the students' performance scores.

Our results are simple: clarity and simplicity of writing, and amount of biomedical information, are key text features from PubMed abstracts for improving downstream scores. Moreover, variation in scores is also largely due to the choice of hyperparameters, contributing approximately as equal as the text features. These results not only explain important features from domain adaptation but can also serve for designing better document sampling strategies and hyperparameter search methods. Moreover, the use of Multilevel Models (MLMs) is key for a deeper understanding of NLP models which we hope the NLP community will adopt.

2 Related Work

2.1 Analyses of Large Language Models

Different works have analyzed different aspects of LLMs. For example, some works studied the interplay between data and model abilities during the SFT (Supervised Fine-Tuning) phase showing the key impact of data on these abilities (Dong et al., 2024). Other works analyzed which training instances contributed to specific model predictions, or abilities learned, using methods such as gradient-based tracing-back (Koh and Liang, 2017; Garima et al., 2020; Akyurek et al., 2022) and machine unlearning (Jang et al., 2023; Eldan and Russinovich, 2023; Zhao et al., 2024). Furthermore, strategies have been proposed to improve both the quality and selection of pretraining data to optimize LLMs' training time, perplexity, or capabilities on downstream tasks (Lee et al., 2022; Rae et al., 2022; Tirumala et al., 2023; Nguyen et al., 2023). However, to our knowledge, ours is the first work analyzing the effect of text features from samples used for domain adaptation on downstream scores via Multilevel Analysis.

2.2 Multilevel Modeling

Multilevel Models (MLMs) are extensively used across fields in the Social Sciences to measure the effect of multi-level variables on an outcome. For example, in Education, MLMs can predict student performance while finding out the most important features from students (level-1) and schools (level-2) (Rasbash et al., 2010; Goldstein et al., 2007); also, MLMs are used to compare school effectiveness (Yang et al., 2002; Goldstein et al., 1993). In Epidemiology, MLMs have been used to 1) model the impact of personal-level risk factors on disease across populations (Weinmayr et al., 2016); 2) measure the effect of air pollution on cardiovascular disease (Forbes et al., 2009); and 3) estimate the risks of food constituents from different items for breast cancer (Witte et al., 1994).

3 Data and Multilevel Model

3.1 Multilevel Regression Analysis

In MLMs, the dependent variable (downstream scores) depends on a set of independent variables which can be at different levels in a hierarchy. We model our problem as a 2-level hierarchy where features from DAPT and fine-tuning, such as choice of hyperparameters, correspond to level-1 variables; and text features from the samples used for domain adaptation of Llama 2 correspond to level-2 variables. This choice of level-1 and level-2 variables is due to our design of the domain adaptation and fine-tuning of Llama 2: from each sample, by varying DAPT hyperparameter values, we obtain 2 domain-adapted models, and from each of these two models, by varying fine-tuning hyperparameters, we obtain 8 fine-tuned models; i.e., from each sample we obtain 16 fine-tuned models; from the perspective of Multilevel Analysis, we say that each sample is a group and the 16 fine-tuned models are grouped under this sample.[1]

Thus, variables at level 1 are indicator variables signaling the use of a specific combination of hyperparameters for DAPT and fine-tuning where values of random seed, batch size, among other hyperparameters, vary; and level-2 variables correspond to numeric and indicator variables capturing text features (Section 3.2). Then, a 2-level MLM (Hox et al., 2017) can be expressed as:

$$y = \beta_0 + u_{0j} + \beta_1 x_1 + ... + \beta_n x_n + \gamma_{1j} x_1 + ... + \gamma_{kj} x_k + e \quad (1)$$

where β_0 is the intercept and u_{0j} is a term called *random intercepts* which can be interpreted as a

[1]We chose MLM over simple linear regression since it is designed to deal with grouped (non-independent) instances while allowing for multi-level variables.

deviation in downstream score from the intercept according to each sample j; level-1 and level-2 variables are denoted by x_i and the terms β_i are called *fixed-effects* coefficients which represent the average effect of each variable (across all samples) on downstream scores (y); terms γ_{ij}, called *random coefficients or slopes*, are key terms in Multilevel Analysis and can be interpreted as an *adjusted* effect on the level-1 fixed-effects coefficients (β_i) according to each sample j;[2] e is the residual.

The advantage of modeling random coefficients lies in capturing differential contributions of each sample on downstream scores; that is, we expect different combinations of hyperparameters to have different effects, due to chance, on the scores depending on the choice of sample; thus, for each sample we can estimate the number of points (γ_{ij}) that a combination of hyperparameters deviates from the mean effect across all samples (β_i).

3.2 Data for Multilevel Regression

To fit an MLM that predicts downstream scores based on both text features from samples used for domain adaptation and choice of hyperparameters for DAPT and fine-tuning, we extract text features from 70 samples[3] used for domain-adapting Llama 2 and use indicator variables to signal the use of a specific hyperparameter combination.

We obtained 1120 fine-tuned models but 16 were discarded since the outputs generated were outside the set of permissible outputs (a ranking score between 0 and 4 showing the degree of similarity between two input sentences), so we used 1104 models. Each fine-tuned model corresponds to an instance in our dataset for fitting MLMs. The dependent variable corresponds to the Pearson correlation score (scaled to [0-100] points) of a model's outputs with gold outputs from the BIOSSES validation set (as measured in the BLURB benchmark (Gu et al., 2021)). Independent variables correspond to features at levels 1 and 2 which we group in 8 groups according to their type. These text features are motivated by research on Education for predicting student writing performance since they have been shown to be strong predictors. We describe these features:

Hyperparameter choice: Level-1 indicator variables signaling the choice of hyperparameter combination used for both domain adaptation and fine-tuning. In Section 4.2 we call these variables DAPT_1 , DAPT_2 (2 combinations for DAPT) and HCF_1, ..., HCF_8 (8 combinations for fine-tuning).

Syntactic dependencies: We extracted 39 syntactic relations from the sentences in each sample via the Stanford dependency parser (Chen and Manning, 2014) and we used the frequency of each relation across the whole sample as a level-2 feature.

Terms overlaps: We hypothesized that overlap of information contained in the sample used for DAPT with information in the BIOSSES dataset may help to improve downstream scores; thus, we computed the frequency of overlapping terms. To do so, we computed frequencies of biomedical and non-biomedical terms at the unigram and bigram levels separately for the train and validation sets of the BIOSSES dataset (leading to eight level-2 numeric features) using the frequency metric of Kerz et al. (2021).

Biomedical information: We computed the ratio of biomedical unigrams to the total number of terms occurring in each sample used for DAPT as we hypothesized that the more the amount of biomedical terms in a sample the better the downstream scores; this led to a numeric level-2 feature.

Text complexity: We measured the linguistic complexity of the samples at 3 different levels: morphological, syntactical, and global, according to Kolmogorov metrics of complexity used in linguistics (Ehret and Szmrecsanyi, 2019), leading to three level-2 features. We hypothesized that complex texts may provide more information and thus better scores.

Average lengths: From each sample, we computed average lengths of both PubMed abstracts (in terms of words) and words (in terms of characters) resulting in two level-2 features.

Sample size: We hypothesized that the number of PubMed abstracts matter, so we tried two different sample sizes for DAPT: 25K and 50K, operationalized as a level-2 indicator variable.

Sampling method: We hypothesized that sampling contiguous abstracts, in terms of publication time, could improve scores since biomedical information tend to be more uniform; thus, we tried

[2] This only applies to level-1 variables, thus $k < n$ (Eq. 1).

[3] The suggested minimum number of groups is 50 (Maas and Hox, 2005).

two sampling methods: randomly and contiguously, which we operationalized as a level-2 variable.

4 Multilevel Analysis and Results

4.1 Fitting Multilevel Models

Goals: We have 2 goals (Harrell, 2015). First, finding which variables have a statistically-significant effect on downstream scores. And second, evaluating the predictive behavior of our Multilevel Model to unseen data via cross-validation.

Modeling Strategies: Our strategy is three-fold. First, we deal with the issue of *multicollinearity*, where it is difficult to assess the effect of variables when they are correlated, via a variant of the variable selection strategy from Yu et al. (2015). Second, we aim for a *parsimonious* model (Robson and Pevalin, 2016) that is simple enough, in the number of parameters, to be understood, yet complex enough to have low prediction error. Third, we perform suggested evaluations in the literature such as likelihood ratio tests (Brown, 2021), R-squared effects (Rights and Sterba, 2019), and cross-validation (Lindner et al., 2022).[4] Lastly, we note that we standardize (mean=0, std dev=1) all independent variables to allow for a head-to-head comparison of their impact on downstream scores.

The curse of multicollinearity: We found extreme cases of multicollinearity across most variables (Fig. 2). To alleviate this problem, we adjust the strategy of Yu et al. (2015): we first use lasso to eliminate non-essential variables; then, we discard redundant variables via variance decomposition proportions; and finally, we apply a backwards search to remove non-significant variables. To avoid introducing bias, we confirm our choice of deletion by measuring cross-validation error.

4.2 Regression Results

We show the results of our best MLM: we show the features that have a significant impact on scores. For Tables 1 and 2, the statistical significance code is: p=0 '***', p<0.001 '**', p<0.01 '*'.

Biomedical information matters: In Table 1 we observe the standardized coefficient of Biomedical_info having a positive and statistically significant effect of 1.78 meaning that for every standard deviation increase in the frequency of biomedical terms in a sample used for DAPT, the downstream scores increase, in average, by 1.78 points.

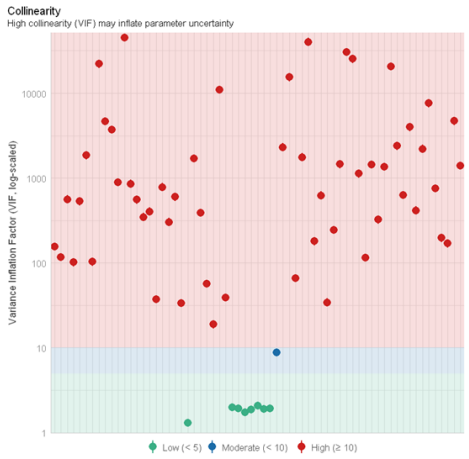

Figure 2: Multicollinearity of all independent variables according to Variance Inflation Factor scores. Scores bigger than 10 show severe cases of multicollinearity.

Text structure and clarity of writing matters: As we observe in Table 1, the syntactic dependency of parataxis has a negative effect on downstream scores: for every standard deviation increase in frequency, scores reduce by 1.92 points. Parataxis occurs when complex sentences are split into clauses separated by commas or semicolons without using any subordinating or coordinating conjunction to make their relationship clear (de Marneffe et al., 2021). Academic writing, as that in PubMed abstracts, often uses parataxis, for example, for reporting previous findings. If overused, parataxis can convey a sense of text unclarity. In contrast, adnominal clauses (acl) occur when the main nominal in a sentence is modified by a subordinate clause usually via clear connectors and in a specific order which conveys text clarity. As we observe in Table 1, acl is the only syntactic relation having a positive effect on downstream scores.

Simplicity matters: As shown in Table 1, two other dependencies have a negative effect on scores: mwe (multiword expressions) and cc.preconj. Simply stated, complex terms such as compounds (e.g. *USB cellphone charger*), proper names, fixed expressions (e.g. *as well as*), or preconjuncts (e.g. *both DNA and RNA*) (de Marneffe et al., 2021), which are common in academic writing, decrease scores. Moreover, longer words tend to substantially decrease scores, as captured by the feature Avg_word being, surprisingly, the feature with the biggest negative impact. Thus, a concise writing

[4] We compute 5-fold cross-validated RMSE (Root Mean Squared Error), averaged over 5 different random seeds.

Variable	Coeff. (β)	SE	t
Intercept	92.66***	0.28	325.58
DAPT_1	-1.00*	0.43	-2.30
HCF_1	-2.53***	0.38	-6.63
HCF_2	-2.49***	0.34	-7.20
HCF_3	-1.12**	0.35	-3.20
HCF_4	-1.12***	0.29	-3.75
HCF_5	-2.76***	0.37	-7.42
HCF_6	-2.68***	0.37	-7.25
acl	1.86**	0.59	3.15
mwe	-0.82**	0.28	-2.87
parataxis	-1.92**	0.56	-3.39
cc.preconj	-3.04***	0.72	-4.20
Biomedical_info	1.78*	0.71	2.49
Avg_word	-3.91***	0.80	-4.84

Table 1: Results of MLM: fixed-effects of variables at levels 1 and 2. Coeff: coefficient. SE: Standard Error. t: t-value (values truncated at the hundredths). DAPT_1 and HCF_i are indicator variables signaling the use of a specific combination of hyperparameters for DAPT and fine-tuning, respectively. We use DAPT_2 and HCF_8 as references to avoid perfect collinearity.

Variable	Variance	Std. Dev.
Intercepts	3.06***	1.75
HCF_1	3.90***	1.97
HCF_2	2.09*	1.44
HCF_3	2.34**	1.53
HCF_5	3.38**	1.84
HCF_6	3.28***	1.81
DAPT_1	11.03***	3.32

Table 2: Results of MLM: random-effects (random intercepts and random coefficients). We use DAPT_2 and HCF_8 as references to avoid perfect collinearity.

with less *idiomatic* and complex expressions is key for a better domain adaptation. However, it is unclear whether biomedical terms, which are often complex, may jeopardize the domain adaptation; thus, this phenomenon deserves a deeper analysis for future work.

How much hyperparameters impact on scores? As we see in Table 1, different hyperparameter combinations lead to different results being combination HCF_5 the one with the biggest impact: whenever used, it leads to an average decrease of 2.76 points in scores across all samples. Moreover, in Table 2 we observe that the choice of sample adds a random effect to the fixed effect of most of the hyperparameter combinations; i.e., we can expect average variations of ($\pm$)[1.44-1.97] and ($\pm$) 3.32 points in the effects of fine-tuning and DAPT hyperparameters observed in Table 1, respectively.

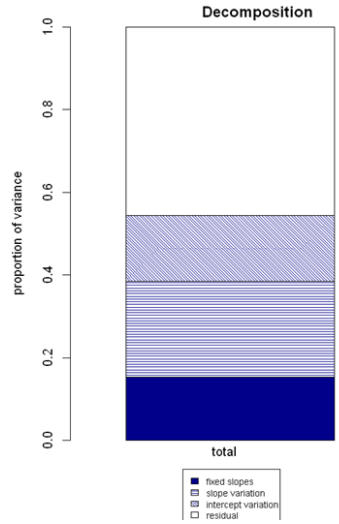

Figure 3: R-squared: Decomposition of variance across fixed and random effects.

Cross-validation error: Our MLM obtains an RMSE of only 4.02 points, which means that our model will deviate, in average, only by 4 points from expected Pearson scores on unseen data.

R-squared effects: Figure 3 shows that around 55% of the variation in downstream scores is accounted by fixed- and random-effects, where 15% is due to fixed-effects, 23% due to random-effects (slopes), and the rest (intercept variation) is due to other features from the samples that we were not able to identify. From Fig. 3 and Table 1, we estimate that the overall effect of hyperparameters on downstream scores is approx. equal to that of text features for domain adaptation described above.

5 Conclusions

How important is to analyze the data used for DAPT? Working *in the trenches* has allowed us to see the paramount importance that data plays on the right adjustment of LLMs to a target domain. From a customer-oriented perspective, DAPT plays a vital role for the correct adjustment of LLMs not only to parametric knowledge but also to human alignment via SFT and to databases via RAG (Retrieval Augmented Generation). Thus, as in a *snowball effect*, studying the factors that matter for biomedical DAPT –clarity and simplicity of writing as well as biomedical information– assures us to provide better adapted LLMs for customer applications.

Limitations

We note that our work has some limitations. For example, despite the low RMSE from our MLM (4% error) and even though we tried our best to propose a comprehensive set of variables that could fully explain the variation in downstream scores, we acknowledge (as observed in Fig. 3) that 45% of the variation in scores remains unexplained; that is, there are more variables at level 1 and level 2 that may have an impact on downstream scores. Furthermore, while we chose the most widely used type of MLM in the literature (2-level MLM), it is possible that other choice of MLM may be a better fit to our problem such as a 3-level model where at level-1 we define only hyperparameter combinations of DAPT, at level-2 we define hyperparameter combinations of fine-tuning, and at level-3 we define features from the samples; however, a model of this type requires a substantial increase in the number of both domain-adapted and fine-tuned models and thus of computing time.

Acknowledgments

We thank the anonymous reviewers for their thorough comments and to Zhang Yuming for sharing Llama 2 with us.

References

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Violet A. Brown. 2021. An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, 4(1):1–19.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. *ArXiv*.

Katharina Ehret and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in sla production data. *Second Language Research*, 35(1):23–45.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *ArXiv*.

Lindsay J. L. Forbes, Minal D. Patel, Alicja R. Rudnicka, Derek G. Cook, Tony Bush, John R. Stedman, Peter H. Whincup, David P. Strachan, and Ross H. Anderson. 2009. Chronic exposure to outdoor air pollution and markers of systemic inflammation. *Epidemiology*, 20(2):245–253.

Garima, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Harvey Goldstein, Simon Burgess, and Brendon McConnell. 2007. Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 170(4):941–954.

Harvey Goldstein, Jon Rasbash, Min Yang, Geoffrey Woodhouse, Huiqi Pan, Desmond Nuttall, and Sally Thomas. 1993. A multilevel analysis of school examination results. *Oxford Review of Education*, 19(4):425–433.

Andreas Groll. 2023. *glmmLasso: Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*. R package version 1.6.3.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jr. Frank E. Harrell. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, second edition. Springer Cham.

J. Hox, M. Moerbeek, and R. van de Schoot. 2017. *Multilevel Analysis: Techniques and Applications*, third edition. Routledge.

Muhammad Imdadullah, Muhammad Aslam, and Saima Altaf. 2016. mctest: An R Package for Detection of Collinearity among Regressors. *The R Journal*, 8(2):495–505.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1885–1894. JMLR.org.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Thomas Lindner, Jonas Puck, and Alain Verbeke. 2022. Beyond addressing multicollinearity: Robust quantitative analysis and machine learning in international business research. *Journal of International Business Studies*, 53(7).

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139.

Cora J. M. Maas and Joop J. Hox. 2005. Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for The Behavioral and Social Sciences*, 1(3):86–92.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *ArXiv*.

Ludvig Renbo Olsen and Hugh Benjamin Zachariae. 2023. *cvms: Cross-Validation for Model Selection*. R package version 1.6.0.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*.

Jon Rasbash, George Leckie, Rebecca Pillinger, and Jennifer Jenkins. 2010. Children's Educational Progress: Partitioning Family, School and Area Effects. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173(3):657–682.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

J. D. Rights and S. K. Sterba. 2019. Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological Methods*, 24(3):309–338.

Karen Robson and David Pevalin. 2016. *Multilevel Modeling in Plain Language*, first edition. SAGE Publications Ltd.

Deepayan Sarkar. 2008. *Lattice: Multivariate Data Visualization with R*. Springer, New York.

Mairead Shaw, Jason D. Rights, Sonya S. Sterba, and Jessica Kay Flake. 2022. r2mlm: An r package calculating r-squared measures for multilevel models. *Behavior Research Methods*, 55:1942–1964.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. In *Advances in Neural Information Processing Systems*, volume 36, pages 53983–53995. Curran Associates, Inc.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*.

Gudrun Weinmayr, Jens Dreyhaupt, Andrea Jaensch, Francesco Forastiere, and David P Strachan. 2016. Multilevel regression modelling to investigate variation in disease prevalence across locations. *International Journal of Epidemiology*, 46(1):336–347.

J. S. Witte, S. Greenland, R. W. Haile, and C. L. Bird. 1994. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*, 5(6):612–621.

Min Yang, Harvey Goldstein, William Browne, and Geoffrey Woodhouse. 2002. Multivariate Multilevel Analyses of Examination Results. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 165(1):137–153.

Han Yu, Shanhe Jiang, and Kenneth C. Land. 2015. Multicollinearity in hierarchical linear models. *Social Science Research*, 53:118–136.

Yang Zhao, LI DU, Xiao Ding, Kai Xiong, Zhouhao Sun, Jun Shi, Ting Liu, and Bing Qin. 2024. Deciphering the impact of pretraining data on large language models through machine unlearning. *ArXiv*.

A Appendix

A.1 Statistical Software

To fit MLMs we use the R-package *lmerTest* (Kuznetsova et al., 2017). To compute cross-validated RMSE we use the *cvms* package (Olsen and Zachariae, 2023), where for folds used as test data, we leave out all fine-tuned models from the samples selected for testing to avoid training and testing on models derived from the same sample. Furthermore, to estimate the proportion of explained variability in downstream scores we compute R-squared effects using the framework of Rights and Sterba (2019) via the R-package *r2mlm* (Shaw et al., 2022). We plot Figures 1 and 2 via the *lattice* (Sarkar, 2008) and *performance* (Lüdecke et al., 2021) packages in R, respectively. To fit lasso regression we use the *glmmLasso* (Groll, 2023) package in R; to compute variance decomposition proportions we use the *mctest* (Imdadullah et al., 2016) package in R. Likelihood ratio tests and backward search are implemented via the *lmerTest* package.

A.2 Training Features

We used a Titan RTX GPU (24GB of memory) for both domain adaptive pretraining and fine-tuning. We domain-adapted Llama 2 for 1 epoch with each sample. We fine-tuned each domain-adapted model with each hyperparameter combination for 30 epochs and kept models with the highest validation score. We used QLoRA (Dettmers et al., 2024) to be able to fit Llama 2 in GPU memory.