Towards Robust Comparisons of NLP Models: A Case Study

Vicente Ivan Sanchez Carmona and Shanshan Jiang and Bin Dong
Ricoh Software Research Center (Beijing) Co., Ltd
{Vicente.Carmona, Shanshan.Jiang, Bin.Dong}@cn.ricoh.com

Abstract

Comparing the test scores of different NLP models across downstream datasets to determine which model leads to the most accurate results is the ultimate step in any experimental work. Doing so via a single mean score may not accurately quantify the *real* capabilities of the models. Previous works have proposed diverse statistical tests to improve the comparison of NLP models; however, a key statistical phenomenon remains understudied: variability in test scores. We propose a type of regression analysis which better explains this phenomenon by isolating the effect of both nuisance factors (such as random seeds) and datasets from the effects of the models' capabilities. We showcase our approach via a case study of some of the most popular biomedical NLP models: after isolating nuisance factors and datasets, our results show that the difference between BioLinkBERT and MSR BiomedBERT is, actually, 7 times smaller than previously reported.

1 Introduction

Proper comparison of NLP models is a cornerstone area of research in the NLP field. Comparing the efficacy of different models via an average test score across downstream datasets has been shown to be an oversimplistic evaluation that may not properly take into account nuisance factors affecting the scores such as noise, randomness, or hyperparameter values, and therefore such average scores may not reflect the true capability of the proposed models (Søgaard, 2013; Dror et al., 2019; Reimers and Gurevych, 2018). Previous efforts have improved the rigour of the comparison of NLP models (and in general the comparison of classifiers) by proposing statistical tests that aim to account for these nuisance factors in order to better see the real efficacy of the models (Demšar, 2006; Zhong et al., 2021; Dror et al., 2017).

However, an important statistical phenomenon inherent to model comparison has been understudied: explaining the variability in test scores. To exemplify this phenomenon, let us revisit the case study proposed by Sanchez Carmona et al. (2024). Figure 1 shows the test score distributions of a baseline model and 3 of the most popular biomedical models from the BLURB benchmark (Gu et al., 2021).[1] These distributions are formed by fine-tuning each model with varying hyperparameter values (random seeds, learning rates, batch sizes, number of epochs) on a set of 13 downstream datasets across different tasks.

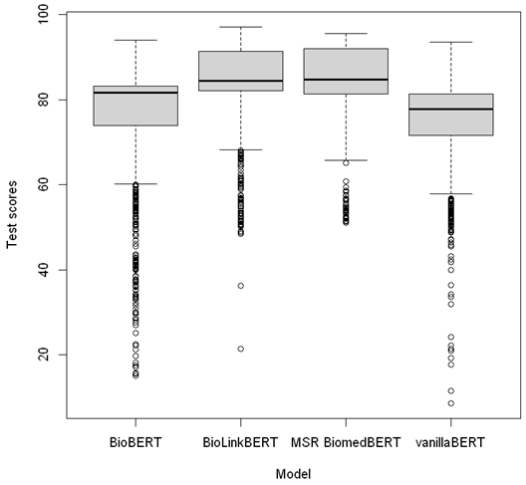

Figure 1: Boxplot of test score distributions of biomedical models across the BLURB benchmark's datasets.

In the BLURB leaderboard, BioLinkBERT is positioned as a top model due to its average test score. However, from Figure 1 we observe both a very similar performance to that of MSR BiomedBERT and a huge variability that ranges from low to high test scores. This variability, also present in the other models' distributions, and the models' average scores are derived not only from the true differences in the models' abilities to solve the datasets, but are also derived from differences in the datasets (different datasets may have different

[1] https://microsoft.github.io/BLURB/

Proceedings of the 31st International Conference on Computational Linguistics, pages 4973–4979
January 19–24, 2025.

difficulty) and from fine-tuning nuisance factors such as random seeds, or learning rates, among others. So, how can we disentangle and isolate the effects of the models' capabilities from effects of both datasets and nuisance factors to see the real mean test scores due to the ability of the models while explaining the variability observed? This problem remains as an open challenge, and while the work of Sanchez Carmona et al. (2024) is the work closest to addressing this problem it has some drawbacks.

In this paper, we present a type of regression analysis for robustly comparing NLP models: a Cross-classified Mixed Effects model (CCMEM), a model widely used in the Social Sciences (Nieuwenhuis et al., 2021; Rasbash et al., 2010), Psychology (Claus et al., 2020; Brown, 2021), and Health Sciences (Barker et al., 2020; Doedens et al., 2022) to analyze diverse social and health phenomena. In a nutshell, a CCMEM can not only disentangle and isolate the effects of datasets and nuisance factors from the effects of the capabilities of NLP models, it can also test for interactions between models and datasets while explaining variation on test scores. We compare our CCMEM to the approach from Sanchez Carmona et al. (2024) on the case study shown in Figure 1. We show that our model achieves smaller cross-validation error and improves by 4% the amount of variability explained. Moreover, our CCMEM shows that the difference in test scores between BioLinkBERT and MSR BiomedBERT is only 0.33 points–a difference 7 times smaller than previously reported.

2 Comparing NLP Models: Previous Works

2.1 Statistical Tests

Diverse statistical tests have been proposed that are able to control for the effect of some nuisance factors which can be confounded with the true effect of the models' capabilities. For example, parametric and non-parametric tests such as ANOVA (and variants) and the Friedman's test (Demšar, 2006; Yıldız et al., 2011; Rainio et al., 2024), as well as Bayesian tests (Corani et al., 2017); these tests compare differences in scores, in their variability, or in the models' performance ranking, and estimate the probability that these differences are not just by chance. Other works have provided improved statistical tests such as comparisons of score distributions (Reimers and Gurevych, 2018), comparisons at the instance level (Zhong et al., 2021), and methods adapted from other disciplines such as *Meta-analysis* (Søgaard, 2013) and *Almost Stochastic Dominance* (Dror et al., 2019).

However, we believe none of these approaches completely resolves the question we posed in Section 1. Some approaches account for the effects of nuisance factors by aggregating scores across them, but this does not isolate their effect from the models' effects. And while some works isolate the effect of some nuisance factors, or explain the variance of test scores, no single approach can isolate and estimate the effects of fine-tuning factors and datasets while explaining their variance in test scores for several NLP models at once.

2.2 Regression Analysis

Sanchez Carmona et al. (2024) proposed a regression analysis to predict the score of any model when given as input information about pretraining, fine-tuning hyperparameters, and the choice of downstream dataset. In this way, the learned coefficients represent the isolated effects of the input features on the test scores, and the effects of the models' contributions are compared. However, this analysis assumes a dependency, in the form of a statistical interaction, between models and datasets which if false would render incorrect results; and moreover, it forbids to estimate the variability due to models and datasets separately. Thus, we propose a regression analysis–a CCMEM–that does not make such an assumption and can estimate variation due to model and dataset choice separately.

3 Cross-classified Mixed Effects Model to Compare NLP Models

3.1 Mathematical Formulation

A CCMEM is a type of regression model which learns two types of effects in the form of coefficients or intercepts: *fixed* and *random*. Based on the work of Fielding and Goldstein (2006) and Garson (2020), we derive our mathematical formulation as follows:

$$y = \beta_0 + \sum_{fixed} \beta_i x_i + \sum_{random} \alpha_{ij} x_i + u_m + u_d + u_{md} + e \quad (1)$$

where

$$\alpha_{ij} \sim N(0, \sigma^2_{\alpha_{ij}})$$
$$u_m \sim N(0, \sigma^2_{model})$$
$$u_d \sim N(0, \sigma^2_{dataset})$$
$$u_{md} \sim N(0, \sigma^2_{model*dataset})$$
$$e \sim N(0, \sigma^2_e)$$

Test scores are represented by y. The intercept is β_0. The summation of *fixed* effects corresponds to input variables x_i weighted by their coefficients β_i which are interpreted as the isolated mean effect of x_i on test scores. The summation of *random* effects can be interpreted as deviations from the fixed-effects coefficients of variables x_i according to a specific model, dataset, or interaction of these two; thus, the final effect of a variable x_i is equal to the sum of its fixed and random coefficients: $\beta_i + \alpha_{ij}$ according to model or dataset with index j. Terms u_m, u_d, and u_{md} correspond to random intercepts which are interpreted as the isolated effect that models, datasets, and their interaction have on test scores. All random terms are random variables assumed to be drawn from normal distributions and their corresponding variance (σ^2) is the amount of variation on test scores explained by these terms; in this way, we can explain the variability observed in Figure 1 while isolating the effects of nuisance factors. Term e is the residual. All fixed and random effects are parameters to be estimated.

3.2 A Working Example: Using a CCMEM

To predict the test score of, for example, BioBERT on the BioASQ dataset after fine-tuning for 15 epochs using random seed=20, batch size=16, learning rate=1e-05, one possible way to instantiate Equation 1 is as follows:

$$\begin{aligned} y = \beta_0 &+ \beta_1(seed\_20) + \beta_2(batch\_16) \\ &+ \beta_3(lr\_1e-5) + \beta_4(num\_epochs) \\ &+ \alpha_{1,BioBERT}(seed\_20) \\ &+ \alpha_{2,BioBERT}(batch\_16) \\ &+ \alpha_{3,BioBERT}(lr\_1e-5) \\ &+ \alpha_{4,BioBERT}(num\_epochs) \\ &+ u_{BioBERT} + u_{BioASQ} \\ &+ u_{BioBERT*BioASQ} \end{aligned} \quad (2)$$

β_0 represents the mean baseline score for any fine-tuned model. All variables x_i are binary indicators, except for num_epochs which is numeric. Thus, each fixed and random effect provides a specific contribution to the test score (y). As we see, the random intercept $u_{BioBERT}$ shows the effect of BioBERT, disentangled from any other effect, which when added to the intercept will result in the mean score of this model which we can use to compare against the other models. Furthermore, we can estimate the effect that each nuisance factor has on the test score by adding its fixed and random coefficients; for example, the random seed's effect accounts for $\beta_1 + \alpha_{1,BioBERT}$ points.[2] The logic behind this formulation is that by modeling test scores of fine-tuned models via the additive composition of nuisance factors, model choice, and dataset choice, we are able to see the amount of points that each of these factors contribute to those scores as shown by their learned coefficients.

4 Comparing NLP Models via CCMEM: Analyses and Results

We revisit the case study introduced in Section 1 and show results from our lowest-error CCMEM. All results and their statistical significance code: p=0 '***', p<0.001 '**', p<0.01 '*' are obtained via the statistical software R (see Appendix A).

4.1 Case Study

In the case study proposed by Sanchez Carmona et al. (2024), three biomedical models and one baseline are compared against each other on 13 downstream datasets via a regression analysis. The dataset generated to fit a regression model comprises 5154 instances where each instance corresponds to one fine-tuned model: the dependent variable is a test score on a downstream dataset; independent variables are a) the main pretraining contribution of each model (for BioBERT it is using Domain Adaptive Pretraining (DAPT), for MSR BiomedBERT it is pretraining BERT from scratch using biomedical documents, and for BioLinkBERT it is using the *Link* function), b) fine-tuning (nuisance) factors: random seed, learning rate, batch size, and number of epochs, and c) the choice of downstream dataset. We use the same dataset to fit our CCMEM.

[2]The subscript $BioBERT$ in the random coefficients $\alpha_{i,BioBERT}$ indicates that these will vary according to the choice of NLP model; thus, each coefficient will be adjusted according to each model. This modeling choice of coefficients follows the hypothesis that nuisance factors will have different effects for each model whether due to chance or due to the idiosyncrasies of each model.

Variable	Coeff. (β)	SE	t
Intercept	77.21***	2.94	26.24
seed_20	0.47	0.71	0.66
seed_47	0.21	0.88	0.24
lr_1e-5	-0.96**	0.36	-2.65
lr_2e-5	0.63	0.56	1.13
lr_3e-5	0.60	0.47	1.28
lr_4e-5	0.06	0.22	0.30
batch_16	0.51**	0.18	2.81
num_epochs	0.10**	0.03	2.94

Table 1: Fixed effects of intercept and nuisance factors. *Coeff*: coefficient. *SE*: Standard Error. *t*: t-value. Variables used as reference to avoid collinearity: seed_59, lr_5e-5, batch_32. Values truncated at the hundredths.

Variable	Coeff.	SE	t
DAPT	4.48**	1.38	3.25
Pretrain_from_scratch	6.99**	2.08	3.35
Link function	0.33	1.38	0.24

Table 2: Fixed effects of main pretraining contributions from each NLP model. *Coeff*: coefficient. *SE*: Standard Error. *t*: t-value. Values truncated at the hundredths.

4.2 Research Questions

We decompose our question posed in Section 1 into the following research questions that we consider to be fundamental for robustly comparing models.

1. Do fine-tuning factors play a role on the test scores?
2. How effective are the main contributions of each NLP model?
3. What portion of variability in tests scores is due to model choice, dataset choice, and fine-tuning factors?
4. Is there any interaction between the choices of model and dataset?
5. Which model is the most accurate, in average, in the BLURB benchmark?

4.3 Answers to Research Questions

Answer 1: Always isolate the effects of fine-tuning factors. In Table 1 we can see the fixed effects of nuisance factors on test scores; as noted, only 3 factors are statistically significant, i.e. their effect is consistent across all models and datasets: learning rate 1e-5, batch size 16, and number of epochs. We can interpret the effect of factor batch_16 as increasing scores by 0.51 points, in average, whenever used instead of using batch_32. Similarly, we interpret the coefficient of num_epochs as increasing scores, in average, by 0.1 points for each epoch added to the fine-tuning of a model. Moreover, we allowed random coefficients of these factors to be adjusted according to the choice of model and dataset; as shown in Table 3, the random effects of all fine-tuning factors (except for lr_4e-5) are statistically significant which means that they will have a different effect on the scores depending on the choice of model and dataset; for example, whenever random seed 47 is used for fine-tuning we can expect an average shift on test scores of ($\pm$) 6.15 points, a shift that we would wrongly attribute to the models capabilities.

Variable	Variance	Std. Dev.
model choice	9.79**	3.12
dataset choice	70.86***	8.41
model*dataset	19.81***	4.45
seed_20	23.41***	4.83
seed_47	37.89***	6.15
lr_1e-5	2.81***	1.67
lr_2e-5	9.91***	3.14
lr_3e-5	7.75***	2.78
lr_4e-5	0.00	0.00
batch_16	0.28*	0.53
num_epochs	10.27**	3.20
residual	27.67	5.26

Table 3: Random effects: intercepts and coefficients (vary with respect to model*dataset interaction). Variables seed_59, lr_5e-5, batch_32 used as references to avoid collinearity. Values truncated at the hundredths.

Answer 2: Not every model has the expected effect. To measure the effect of the main pretraining contribution from each model on test scores we added such factors to our cross-classified model to estimate their fixed-effect coefficients. Table 2 shows these effects; as noted, the contribution from BioBERT, namely DAPT, has a statistically significant mean effect of 4.48 points across all datasets; i.e. using DAPT increases test scores across all datasets, in average, by 4.48 points with respect to vanilla BERT. Similarly, pretraining BERT from scratch with biomedical documents (MSR BiomedBERT's contribution) improves scores, in average, by almost 7 points. However, BioLinkBERT's contribution, the Link function, is rather small (0.33 points) and not statistically significant which means

that this effect is not systematic across all datasets.

Answer 3: Explaining variability in test scores. In Table 3, we can observe the amount of variance in test scores attributed to the choice of NLP model: $\sigma^2_{model} = 9.79$, which is the variance of the distribution from where u_m is drawn in Equation 1, and can be interpreted in two ways: 1) as a standard deviation: the expected amount of points (± 3.12) that test scores will vary due to model choice, and 2) as a measure of impact on test scores: this amount of variance in proportion to the total variance from all the random effects in Table 3 (i.e. adding all variances), results in 4.44% representing the approximate[3] amount of variability on scores due to differences in the models' capabilities.[4] On the other hand, the variability in test scores attributed to the choice of downstream dataset is $\sigma^2_{dataset} = 70.86$ –the largest variance– explaining 32.13% of such variability. In addition, most of the nuisance factors significantly contribute to variation in scores. For example, the proportion of variability due to random seed 47 is 17.18%, almost 4 times the variation due to model choice.

Answer 4: Different models perform different according to the dataset. An interaction term of the form $model * dataset$ indicates whether the effect of a model on test scores differs according to the choice of dataset. As observed in Table 3, the interaction between models and datasets is statistically significant, explaining almost 9% of the variability in test scores ($\sigma^2_{model*dataset} = 19.81$); this could mean that some datasets may be more difficult than others and some models perform better on some datasets than on others due to their specific characteristics. For example, while the combined effect of BioLinkBERT with BIOSSES data is $u_{BioLinkBERT*BIOSSES} = 9.58$ points, the effect with BC2GM data is $u_{BioLinkBERT*BC2GM} = -0.37$ points; clearly, BioLinkBERT is better suited for solving the former dataset.

Answer 5: There is a *winner* by a narrow margin. By estimating the random effect of each model (u_m) and adding it to the intercept, we obtain the following average scores across datasets after isolating the effects of all nuisance factors ($\beta_i + \alpha_{ij}$), datasets (u_d), and interactions (u_{md}): vanilla BERT: 73.22; BioBERT: 76.99; MSR BiomedBERT: 79.15; BioLinkBERT: 79.48. All models surpass vanilla BERT, and though BioLinkBERT is the best model, its difference with MSR BiomedBERT is only 0.33 points, which is, in fact, much smaller than reported in the literature: 2.23 points.[5] But, if the main contribution of BioLinkBERT is very small and not statistically significant, how come BioLinkBERT is the best model? That is because BioLinkBERT was pretrained from scratch in the same way as MSR BiomedBERT; thus, it takes the benefit of using MSR BiomedBERT's contribution, and the difference in mean score between these two models, 0.33 points, is exactly the points contributed by the Link function. We note that for a more detailed comparison between models on a dataset basis, interaction effects should be taken into account.

4.4 Evaluation of our Cross-classified Model

While the regression model from Sanchez Carmona et al. (2024) obtains a Mean Absolute Error of MAE=2.28 on cross-validation data and is able to explain $R^2 = 78.55\%$ of the variability in test scores, our CCMEM obtains a MAE=2.22 points while explaining $R^2 = 82.77\%$ of the variability; i.e. our CCMEM reduces cross-validation error and improves by 4.22% the variance explained.[6]

5 Conclusions

We presented a Cross-classified Mixed Effects model (CCMEM) which can robustly compare NLP models by isolating the effects of fine-tuning factors and datasets from the effects of the true models' capabilities while explaining the variability in test scores according to the choice of model, dataset, and fine-tuning factors. Our CCMEM estimated a more accurate picture of the mean scores of 3 of the most popular biomedical models in the BLURB benchmark showing a different picture than previously portrayed: differences in test scores previously obtained are not only due to models' capabilities but also due to datasets and fine-tuning factors; after isolating these factors, the scores difference between BioLinkBERT and MSR BiomedBERT becomes 7 times smaller than previously thought: only 0.33 points of difference.

[3]The exact proportion of variance follows an slightly more complex equation (Leckie et al., 2020).

[4]We note that the CCMEM does not explicitly include the fixed-effects of the pretraining contributions of the models since these are implicitly included in their random intercepts.

[5]https://microsoft.github.io/BLURB/leaderboard.html BioLinkBERT-Base's score compared against that of MSR BiomedBERT (uncased; abstracts).

[6]This variance includes also the variability explained by fine-tuning fixed-effects terms.

Limitations

Our approach has some limitations. First, this approach is better suited for comparing several NLP models since it needs to estimate variances of normal distributions. Second, we tested our approach on one case study; to fully appreciate its usefulness it is advisable to use it on more case studies with other types of pretrained models such as LLama or Mistral, which we leave for future work. Third, the proportion of variation in the test scores explained by our CCMEM is 82.77%, which means that 17.23% of the variation goes unexplained; i.e. there seem to be more factors contributing to this variation that we could not identify; nevertheless, future works can propose more factors and test with a CCMEM whether they play a role on test scores.

References

Kathryn M. Barker, Erin C. Dunn, Tracy K. Richmond, Sarah Ahmed, Matthew Hawrilenko, and Clare R. Evans. 2020. Cross-classified multilevel models (ccmm) in health research: A systematic review of published empirical studies and recommendations for best practices. *SSM - Population Health*, 12:100661.

Kamil Bartoń. 2023. *MuMIn: Multi-Model Inference*. R package version 1.47.5.

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.

Anna M. Claus, Matthias G. Arend, Christian L. Burk, Christoph Kiefer, and Bettina S. Wiese. 2020. Cross-classified models in i/o psychology. *Journal of Vocational Behavior*, 120:103447.

Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. 2017. Statistical comparison of classifiers through bayesian hierarchical modelling. *Machine Learning*, 106:1817–1837.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Paul Doedens, Gerben ter Riet, Lindy-Lou Boyette, Corine Latour, Lieuwe de Haan, and Jos Twisk. 2022. Cross-classified multilevel models improved standard error estimates of covariates in clinical outcomes – a simulation study. *Journal of Clinical Epidemiology*, 145:39–46.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Antony Fielding and Harvey Goldstein. 2006. Cross-classified and multiple membership structures in multilevel models: An introduction and review. *Research Report RR791*. Available online at https://dera.ioe.ac.uk/id/eprint/6469/1/RR791.pdf.

G. David Garson. 2020. *Multilevel Modeling: Applications in STATA®, IBM® SPSS®, SAS®, R, & HLM™*. SAGE Publications, Inc., Thousand Oaks, California.

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).

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.

George Leckie, William J. Browne, Harvey Goldstein, Juan Merlo, and Peter C. Austin. 2020. Partitioning variation in multilevel models for count data. *Psychological Methods*, 25(6):787–801.

Jaap Nieuwenhuis, Tom Kleinepier, and Maarten van Ham. 2021. The role of exposure to neighborhood and school poverty in understanding educational attainment. *Journal of Youth and Adolescence*, 50:872–892.

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

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(6086).

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.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *Preprint*, arXiv:1803.09578.

Vicente Sanchez Carmona, Shanshan Jiang, and Bin Dong. 2024. A multilevel analysis of PubMed-only BERT-based biomedical models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 105–110, Mexico City, Mexico. Association for Computational Linguistics.

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611, Atlanta, Georgia. Association for Computational Linguistics.

Olcay Taner Yıldız, Özlem Aslan, and Ethem Alpaydın. 2011. Multivariate statistical tests for comparing classification algorithms. In *Learning and Intelligent Optimization*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. Are larger pretrained language models uniformly better? comparing performance at the instance level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Online. Association for Computational Linguistics.

A Appendix

A.1 Statistical Software

We use the R statistical framework to carry out all analyses. In particular, we use the *lmerTest* package (Kuznetsova et al., 2017) to train CCMEMs and to obtain statistical significance of both fixed and random terms; we train our CCMEMs using REML (restricted maximum likelihood); statistical significance of random terms is obtained via likelihood-ratio tests; we use the *cvms* package (Olsen and Zachariae, 2023) to compute cross-validation error (MAE); we use the *MuMIn* package (Bartoń, 2023) to compute R^2 effects.