Reliable and Efficient Amortized Model-based Evaluation

Sang Truong [1] **Yuheng Tu** [2] **Percy Liang** [1] **Bo Li** [3 4] **Sanmi Koyejo** [1 3]

Abstract

reserved

1. Introduction

Multidimensional evaluation of abilities has a positive correlation with the performance of upstream tasks. This hypothesis predicates on the assumption that a given test upstream task γ is a construction of either orthogonal or related abilities. In this sense, the relative performance of an upstream task can be represented as a vector of abilities $\vec{\gamma} = (\theta_1, \theta_2, \ldots, \theta_n)$ whose magnitude is a measure of the success probability of the model when performing general day-to-day tasks.

Sang: It's often helpful to outline 2-3 main contributions of the analysis in the introduction. For example: In summary, our contributions are:

- We conduct a multi-dimensional factor analysis to show that there is substantial overlap between abilities tested across datasets.
- 2nd contributions
- 3rd contributions

2. Related Work

Reserved

3. Preliminary

This experiment uses answers from HELM set, which is collected from 183 test-takers on 22 datasets. This creates a response matrix with size 183×78712. Each dataset is assumed to test a small subset of predetermined skills ($n \leq 3$) based on their description.

There should be a significant overlap of skills for every dataset (specialised upstream task) as is present with the common consensus for the generalisation of many standardised exams. To verify this claim, I have conducted some preliminary data analysis.

[1]Stanford University [2]University of California, Berkeley [3]Virtue AI [4]University of Illinois Urbana-Champaign. Correspondence to: Sang Truong <sttruong@cs.stanford.edu>.

3.1. Correlation metrics

The probability of correct answers shows significant correlations, as is present in Fig.1. This prompted the first Principal Component Analysis (PCA) on the dataset. I started with the first number of dimensions that begins to show a significant variance ($\sigma^2 \geq 95\%$), arriving at $\dim = 7$. Further analysis of PC loadings has led to a reasonable interpretation of each PC based on influential datasets.

- PC1: Foundational vs. Expert Reasoning
- PC2: General Academic Aptitude
- PC3: Thai-Specific Proficiency
- PC4: Thai Benchmark Conflict
- PC5: Safety vs. Specificity Trade-off
- PC6: Abstract vs. Applied Language Reasoning
- PC7: MMLU Specialization

I only considered datasets with significant influence ($|\text{load}| \geq 0.4$) for interpretation. Note that these are only suggested names for easy reference; further analysis is needed to understand specific PCs. The full loading result is attached in Appendix A of this report. This resulted in 6 meaningful PCs (PC1 didn't have a significant-load dataset).

3.2. Factor analysis

I also performed an additional factor analysis to verify the reliability of my PCA. I performed orthogonal rotation on different number of dimensions and found that both 7 and 9 number of PC yields the most AUC and Pearson Correlation, very close to the 2-PL model. However, when testing for strongest loaders, the results in each dimension are identical, indicating the data may have a strong global factor that overshadows more subtle, distinct latent traits.

4. Method

4.1. Reserved

5. Experiments

5.1. Linear experiment

From the conclusions of aforementioned preliminary experiments, there are 6 usable PCs (PC2-PC7) and the origi-

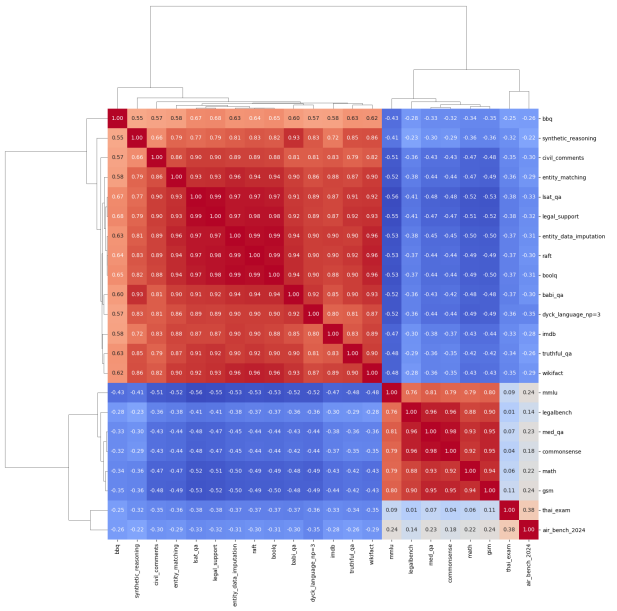

Figure 1: Clustered Correlation Matrix Heatmap for the probability of correct in each dataset.

nal dataset has a dominating latent ability across all scenarios. To fit the ability model for each scenario, I created a Q-matrix which is a one-hot encoding matrix of the scenarios according to the ability dimensions. To fit the model, I used Multidimensional Item Response Theory (MIRT) [1] on the test set.

$$p(y = 1|\theta, a, d) = \sigma(a \cdot \theta - d)$$

To verify if the dimensions are correlated with the accuracy of measured tasks, I conducted a report on the correlation of sub-scores with the original datasets **??**. The sub-scores are highly correlated with some original datasets, meaning they reflect similar skills or domains. When tested with the fit of thetas in 6 dimensions, the abilities displayed a weaker correlation with the original datasets. This may be explained by the strong general factor (g-factor) in the dataset. I performed an additional analysis taking into account that factor by appending an additional ability dimension which is shared by all datasets. This is called a Bi-factor model which separates a strong general factor and very specific ones. In any case, the correlation matrix showed significant noise. I also tested all of this approach with a larger question pool ($n \geq 40000$) but the model introduced even more noise.

[1] https://www.psychometrics.cam.ac.uk/system/files/documents/multidimensional-item-response-theory.pdf

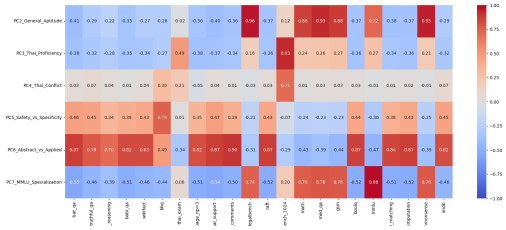

Figure 2: Correlation of Raw Sub-scores with original datasets.

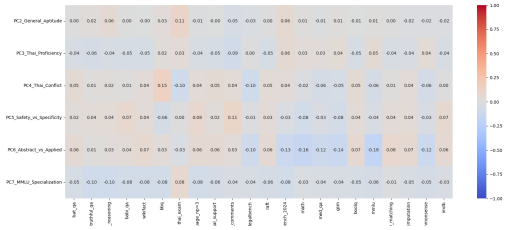

Figure 3: Correlation of γ dimensions on MIRT 2-PL with the original dataset.

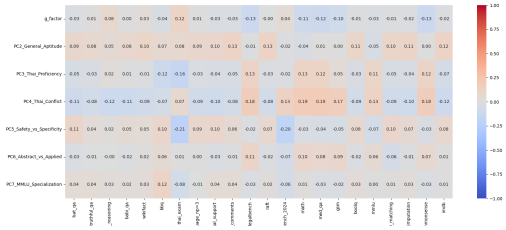

Figure 4: Correlation of γ dimensions on MIRT 2-PL (Bi-factor model) with the original dataset.

5.2. Non-linear experiment

Through the extensive survey, I have concluded that the regular, linear IRT model is insufficient to estimate the multidimensionality nature of the new γ. Another reasonable approach would be fitting the equation with a non-linear model, a Neural Network, which I find to have unanticipated success.

I reused the original IRT model to have a well-experimented reference. The new model is

$$p(y = 1|\theta, z) = \sigma(\mathcal{Q} \cdot \gamma - z)$$

Where $\gamma = (\theta_1, \theta_2, \ldots, \theta_6)$ is the 6 dimensional ability encoding from PCA. In the neural network, γ and z are treated as learnable embeddings; whereas $\mathcal{Q}$ is treated as a non-trainable Q-matrix buffer. Putting this model through roughly half the response matrix ($n \approx 45000$) and 5 epochs, I got improved results from the referenced experiment using the same measurement calculations.

Table 1: Neural IRT Rasch model results

Metric	Rasch Model	NeuralIRT
AUC Train	0.840	0.970
AUC Test	0.827	0.968
Pearson Correlation	0.999	0.999
Pearson Correlation	0.993	0.999

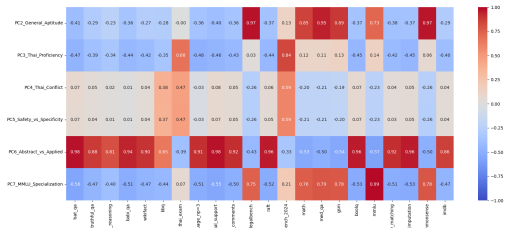

Figure 5: Correlations of γ dimensions on NeuralIRT to their upstream tasks.

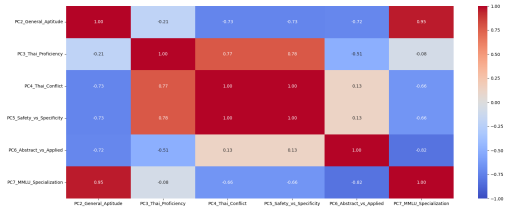

Figure 6: Correlation of γ dimensions on NeuralIRT.

The result shows a clear correlation in some dimensions with the success rate of their corresponding upstream tasks. Furthermore, an analysis of the interactions between abilities signals an influence from some dimensions on another. This suggests one of two things: there are inherent, shared ability factor (g-factor) that makes some ability yield greater correlations than others (like logical reasoning would imply the improved performance on English and Maths tests), or if we consider the hypothesis of ability dimensions are strongly orthogonal, the definition of sub-ability needs to be refined. In any case, with the presence of absolute correlation, there's still a need for a more thorough and theoretically sound definition of PC or even an alternative approach to dimensionality reduction procedure to ensure the utmost reliability and validity of the test.

6. Conclusion, Limitations, Future Work

There are clear statistical correlations for the accuracy of some tasks to others, and also the accuracy of pre-defined skills to their related upstream task. Since we are expanding the ability estimation to account for a more complex, non-linear interaction, the need for a more capable approximation algorithm is the more ideal path to take. Going forward, I suggest further work in fundamental theoretical construct to support the reproducibility of the test and its results. For efficiency, I experimented with pre-existing HELM response matrix, but this doesn't fully represent all the aspects in which an LLM can be used on a day-to-day basis. For the sake of completeness and the possibility of better performance, I suggest performing the experiment of a more diverse conditions (different datasets, number of dimensions, samples, 2-PL, 3-PL models, etc.)

This experiment hasn't implemented the Computerised Adaptive Test suite which can be another hurdle to overcome.

References

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Baker, F. B. *The basics of item response theory*. ERIC, 2001.

Bock, R. D. and Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Brennan, R. L. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 1992.

Chalmers, R. P. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012. doi: 10.18637/jss.v048.i06. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i06.

Edgeworth, F. Y. The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3):599–635, 1888. ISSN 09528385. URL http://www.jstor.org/stable/2339898.

He, Y. and Chen, P. Optimal online calibration designs for item replenishment in adaptive testing. *psychometrika*, 85(1):35–55, 2020.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Lalor, J. P., Wu, H., Munkhdalai, T., and Yu, H. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, pp. 4711. NIH Public Access, 2018.

Lalor, J. P., Wu, H., and Yu, H. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, pp. 4240. NIH Public Access, 2019.

Liang, P. e. a. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Lord, F. M. *Applications of item response theory to practical testing problems*. Routledge, 1980.

Maia Polo, F., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Ostini, R. and Nering, M. *Polytomous Item Response Theory Models*. Polytomous Item Response Theory Models. SAGE Publications, 2006. ISBN 9780761930686. URL https://books.google.com.hk/books?id=wS8VEMtJ3UYC.

Perlitz, Y., Bandel, E., Gera, A., Arviv, O., Ein-Dor, L., Shnarch, E., Slonim, N., Shmueli-Scheuer, M., and Choshen, L. Efficient benchmarking (of language models). *arXiv preprint arXiv:2308.11696*, 2023.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., and Boyd-Graber, J. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346.

Ruan, Y., Maddison, C. J., and Hashimoto, T. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024.

Saranathan, G., Alam, M. P., Lim, J., Bhattacharya, S., Wong, S. Y., Foltin, M., and Xu, C. Dele: Data efficient llm evaluation. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, 1904. URL https://psycnet.apa.org/record/1926-00292-001.

Van der Linden, W. J., Glas, C. A., et al. *Computerized adaptive testing: Theory and practice*, volume 13. Springer, 2000.

Vania, C., Htut, P. M., Huang, W., Mungra, D., Pang, R. Y., Phang, J., Liu, H., Cho, K., and Bowman, S. R. Comparing test sets with item response theory. *arXiv preprint arXiv:2106.00840*, 2021.

Vivek, R., Ethayarajh, K., Yang, D., and Kiela, D. Anchor points: Benchmarking models with much fewer examples. *arXiv preprint arXiv:2309.08638*, 2023.

Wainer, H. and Mislevy, R. J. Item response theory, item calibration, and proficiency estimation. In *Computerized adaptive testing*, pp. 61–100. Routledge, 2000.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., and Goodman, N. Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*, 2020.

Xu, C., Saranathan, G., Alam, M. P., Shah, A., Lim, J., Wong, S. Y., Martin, F., and Bhattacharya, S. Data efficient evaluation of large language models and text-to-image models via adaptive sampling, 2024. URL https://arxiv.org/abs/2406.15527.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zheng, Y. *New methods of online calibration for item bank replenishment*. University of Illinois at Urbana-Champaign, 2014.

Table 2: Top 3 Positive and Negative Loadings Across Principal Components

PC1		PC2		PC3		PC4	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
legal_support	mmlu	legalbench	thai_exam	air_bench_2024	bbq	thai_exam	air_bench_2024
(0.256)	(-0.180)	(0.409)	(-0.141)	(0.741)	(-0.024)	(0.621)	(-0.502)
raft	gsm	commonsense	air_bench_2024	thai_exam	legalbench	bbq	mmlu
(0.256)	(-0.178)	(0.396)	(-0.010)	(0.637)	(-0.018)	(0.559)	(-0.123)
entity_data_imp	math	med_qa	bbq	synth_reasoning	dyck_lang_np=3	legalbench	entity_matching
(0.256)	(-0.175)	(0.395)	(0.059)	(0.105)	(-0.006)	(0.076)	(-0.074)

PC5		PC6		PC7	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
bbq	thai_exam	synth_reasoning	civil_comments	mmlu	legalbench
(0.767)	(-0.410)	(0.679)	(-0.459)	(0.916)	(-0.194)
air_bench_2024	dyck_lang_np=3	babi_qa	imdb	bbq	commonsense
(0.411)	(-0.149)	(0.303)	(-0.255)	(0.152)	(-0.172)
math	entity_matching	dyck_lang_np=3	lsat_qa	thai_exam	math
(0.043)	(-0.096)	(0.175)	(-0.186)	(0.069)	(-0.113)