Unveiling the General Intelligence Factor in Language Models: A Psychometric Approach

David Ilić
dilic3320rn@raf.rs

***Abstract*—This study uncovers the factor of general intelligence, or *g*, in language models, extending the psychometric theory traditionally applied to humans and certain animal species. Utilizing factor analysis on two extensive datasets - Open LLM Leaderboard with 1,232 models and General Language Understanding Evaluation (GLUE) Leaderboard with 88 models - we find compelling evidence for a unidimensional, highly stable *g* factor that accounts for 85% of the variance in model performance. The study also finds a moderate correlation of .48 between model size and *g*. The discovery of *g* in language models offers a unified metric for model evaluation and opens new avenues for more robust, *g*-based model ability assessment. These findings lay the foundation for understanding and future research on artificial general intelligence from a psychometric perspective and have practical implications for model evaluation and development.**

***Index Terms*—General Intelligence, G Factor, Language Models, Factor Analysis, Psychometrics.**

I. Introduction

A. The general intelligence factor in humans

In the early years of the 20th century, Charles Spearman proposed a hypothesis aiming to explain the observation that children's performance across unrelated school subjects was positively correlated. Spearman hypothesized that a single underlying trait, referred to as the "general intelligence factor" or *g* could account for this correlation [1]. To test his hypothesis and quantify *g*, Spearman developed a statistical method known as factor analysis [1] [2]. Since then, *g* has been found to be a robust and reliable construct, while the observed positive correlations between unobserved mental abilities (known as the positive manifold) have become one of the most replicated findings in differential psychology [3]. The *g* factor commonly explains more than 40% of the variance in cognitive ability tests in humans [4] and is an excellent predictor of various life outcomes [5].

While Spearman's concept of a general intelligence factor has stood the test of time, his two-factor model has been refined. Spearman's two-factor model posited that intelligence is composed of a general factor (*g*) and specific factors (*s*) unique to each test [6], displayed in Figure 1. The prevailing view in contemporary research is that *g* sits at the top of some type of hierarchical model, with several first-order factors below it [7], akin to the one displayed in Figure 2.

B. The general intelligence factor in non-human animals

The notion of general intelligence in non-human animals has been a subject of interest since the 1930s, shortly after

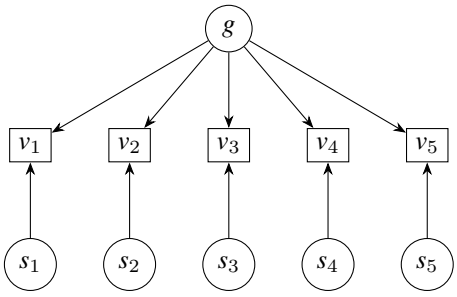

Fig. 1. Spearman's "two-factor" model of intelligence (errors omitted)

Spearman's concept of the *g* factor gained traction [8]. In terms of empirical evidence, rodents, particularly mice, have shown robust evidence for a *g* factor, explaining about 35% of the variance in cognitive performance [9]. In non-human primates, a meta-analysis covering over 4000 papers and 62 species of primates found that a single factor explained 47% of the variance, suggesting a presence of *g* similar to that in humans [10]. Beyond mammals, some bird species like bowerbirds have shown evidence for *g*, explaining over 44% of the variance [11]. In contrast, *g* may not exist in fish [12]. These findings suggest that the positive manifold is not unique to humans and sometimes exists in various non-human species, although the studies do often suffer from limitations such as small sample sizes and limited test diversity [3].

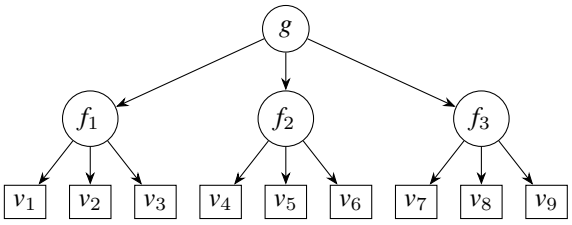

Fig. 2. Hierarchical model of intelligence (errors omitted)

C. The general intelligence factor in language models

Several theoretical and practical attempts have been made to measure general ability in language models and neural networks more generally [13] [14] [15]. Building on the foundational work in human and non-human intelligence, we extend the concept of the general intelligence factor to the realm of language models. Our central hypothesis posits the existence and strong invariance of a *g* factor in language

models, akin to that observed in humans and certain animal species. Furthermore, we hypothesize that the factor structure will have a hierarchical nature, with *g* sitting on top of lower-level factors. Finally, we hypothesize that there will be a significant positive correlation between model size and *g*. This hypothesis is based on the observation that increasing model size leads to lower test loss [16].

II. Methods

A. Factor Analyses

1) Open LLM Leaderboard:

Participants: We considered all 1232 models listed on the Open LLM Leaderboard [17] at the time of this study. This exceeds the usual sample size requirements for exploratory factor analysis [18]. These models, varying in architecture, training data, and other hyperparameters, have parameter counts ranging from 1 million to 180 billion, with a mean of 16.65 billion, a median of 13 billion, and a standard deviation of 19.3 billion. No additional selection criteria or preprocessing steps were applied.

Test Battery: The test battery used in this analysis initially consisted of 60 subtests, including ARC Challenge [19], Hellaswag [20], TruthfulQA [21], and 57 additional subtests derived from the MMLU dataset [22]. These subtests are designed to assess multiple dimensions of cognitive ability in language models.

Each subtest typically presents four answer choices, thereby establishing a 25% baseline accuracy, except for TruthfulQA, which occasionally offers more options. The 95% confidence interval for a model's performance under the model of random guessing was calculated to range approximately from 24% to 26%, with the specific range varying depending on the number of items in each subtest. Upon plotting the scores, it became evident that some subtests were too challenging for the models and were subsequently removed from the analysis. Specifically, only columns where 80% of the models scored at least 26% were retained. This refinement resulted in a final test battery of 22 subtests, which still included ARC Challenge, Hellaswag, TruthfulQA, and 19 MMLU subtests detailed in Table I. Importantly, the reduced set of subtests is still numerous and diverse enough for a stable extraction of *g* [7].

None of the subtests have been modified. The evaluation metric for all subtests is normalized accuracy, except for TruthfulQA, where the *mc2* option is used. This is consistent with the methodology employed by the Open LLM Leaderboard.

2) General Language Understanding Evaluation Leaderboard:

Participants: A total of 88 models listed on the General Language Understanding Evaluation (GLUE) Leaderboard were included in this analysis [23]. This sample size satisfies the recommended sample size to indicator (N:p) ratio of at least 5:1 [24]. While a minimum sample size of 50 is often considered the lower bound for factor analyses [18], samples smaller than 100 are generally characterized as "poor" [25]. However, the "strength" of the data can in certain circumstances compensate for a smaller sample size. Specifically, high communalities, low cross-loadings, and high factor loadings contribute to data strength [26]. Additionally, when a factor is indicated by at least 6 variables, proper convergence is achieved 99.6% of the time with samples as small as 50 [27]. In cases of high communalities (>0.6) along with a high number of indicators per factor, the sample sizes below 100 can still yield accurate estimations of population parameters [27]. Given these considerations, we assert that our sample size of 88 models is adequate for the objectives of the present study.

No additional selection criteria were imposed, and data pertaining to model sizes were not available for this subset of models. All listed models on the GLUE Leaderboard were included in the analysis.

Test Battery: The test battery consists of 10 subtests from the GLUE leaderboard which are detailed in Table III. Subtest quantity and diversity were judged to be sufficient for robust extraction of *g* [7].

3) Analysis Plan: The same analysis was applied to both datasets. To ensure a robust and comprehensive extraction of *g*, we employed common factor analysis, as it avoids the risk of inflating the variance explained by the first component, a limitation often associated with principal component analysis [28]. Still, to verify invariance across methods, the *g* derived from FA was correlated with *g* derived from PCA. Spearman's correlation coefficient was used to compute the correlation matrix.

Before proceeding with the analysis, we verified the appropriateness of our dataset for factor analysis by employing the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy [29] and Bartlett's test of sphericity [30].

The determination of the number of factors to retain was a multi-criteria process. We employed Kaiser's criterion method [31], scree plot analysis [32], Horn's parallel analysis [33], and minimum average partial (MAP) [34]. These methods were used to ensure a robust decision-making process. Additionally, the variance explained by each factor and the coherence and interpretability of the extracted factors were also considered.

For the factor extraction process, principal axis factoring was employed, aligning with common practices in the field [7]. In cases where a single factor was extracted, no rotation was performed. However, if multiple factors were extracted, rotations were carried out using primarily the promax method, with oblimin rotation used to verify the invariance of salient loadings under different rotation methods [35, p. 348]. Oblique rotations were chosen because orthogonal rotations like varimax mathematically preclude the extraction of *g* [28]. The threshold for salient loadings was set at .33.

B. Reliability Analyses

1) A singular g: To evaluate the existence of a singular general intelligence factor, a computational methodology was employed. The goal was to verify the reliability and stability of the extracted *g*-factor across different sets of randomly selected subtests. The dataset used in this analysis was derived from

the Open LLM Leaderboard, encompassing all 1,232 models considered. Below is a detailed breakdown of the steps taken:

Two separate test batteries were formulated by first randomly sampling 20 subtests without replacement from the complete test battery, which contained ARC Challenge, Hellaswag, TruthfulQA, and the 60 MMLU subtests. These sampled subtests were then equally divided into two disjoint sets, forming the first and second test batteries. Each battery therefore contained 10 subtests and was mutually exclusive with the other.

Principal axis factoring was used to conduct factor analysis on both batteries separately. Principal axis factoring was used as opposed to maximum likelihood because it is more robust to non-normality [36].

The scores of the models on the *g*-factors derived from the two disjoint test batteries were subsequently correlated. The Pearson correlation coefficient was computed to quantify the similarity and thus the reliability of the *g*-factors across different subsets of subtests.

This entire procedure was iteratively executed 100 times, each time employing a new random partitioning of the subtests into two test batteries. The resulting distribution of correlation coefficients was summarized using measures such as mean, median, and standard deviation to assess the overall reliability and stability of the extracted *g*-factor.

2) Reliability of g*-loadings:* To evaluate the reliability of *g*-loadings, we employed a similar methodology to the one used in the previous analysis.

Two separate test batteries were formulated by first randomly sampling 19 subtests without replacement from the complete test battery. From the 19 subtests, one was randomly selected to be the target and included in both batteries. The remaining 18 subtests were equally divided into two disjoint sets, forming two test batteries of length 10 with exactly one subtest in common.

Principal axis factoring was used to conduct factor analysis on both batteries separately, and the *g*-loadings of the common subtest were extracted. As in the previous analysis, this procedure was executed 100 times, and the *g*-loadings of the common subtest were then correlated.

C. Correlation Between Model Size and g

To investigate the relationship between model size and *g*, we leveraged data from the Open LLM Leaderboard, including all 1,232 models in our sample. For summary statistics regarding model sizes, as well as the test battery and method used to derive *g*, readers are referred to Section A. Pearson correlation coefficient was used to quantify the relationship between model size and *g*.

III. Results

A. Factor Analyses

1) Open LLM Leaderboard: The dataset exhibited a minimal percentage of missing data (0.02%), which was addressed using mean imputation [36]. Univariate skewness and kurtosis were within acceptable ranges, but multivariate kurtosis indicated significant departures from normality.

Factorability: The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was calculated to be 0.98, which is considered "marvelous" according to Kaiser's criteria. Bartlett's test of sphericity yielded a $\chi^2(22) = 7249$ with $p < 0.0001$, confirming the factorability of the dataset. All values in the correlation matrix were statistically significant at the .0001 level and the average off-diagonal correlation was $r = .84$, displayed as a heatmap in Figure 3.

Factor Extraction: Minimum average partial, scree plot analysis, and Kaiser's criterion all suggested the extraction of a single factor. In contrast, parallel analysis, whose details are displayed in Figure 4, suggested extracting two factors. Although correct most of the time, it is known that parallel analysis has a slight tendency to overfactor [37, p. 33].

As hypothesized, extraction of two factors resulted in symptoms of over-extraction. Specifically, the two extracted factors demonstrated a correlation of $r = .8$ ($p < .0001$; 95% CI $[.78, .82]$), suggesting possible discriminant validity issues. In other words, it is suggested that the factors do not measure different things [38]. The second factor displayed salient loadings on three distinct indicators but served as the primary loading exclusively for the TruthfulQA measure. Additionally, the second factor lacked theoretical coherence, as the indicators upon which it loaded were not conceptually related. When the factor structure was re-evaluated using oblimin rotation, the second factor was found to have only a single salient loading, without serving as the primary factor for that particular indicator. The first unrotated factor accounted for a substantial 85.4% of the total variance, and the inclusion of a second factor marginally increased the total explained variance to 87%.

For all the stated reasons the second factor was judged unlikely to replicate in future analyses, so the decision was made to proceed with a single factor solution.

Factor Loadings and Scores: The standardized loadings based on the correlation matrix are presented in Table I. The root mean square of the residuals (RMSR) was found to be 0.02, indicating a good model fit. Example factor scores are presented in Table II.

Finally, the *g*-factors derived from factor analysis and principal component analysis were found to be highly correlated ($r = .99; p < .0001$), indicating that *g* is highly invariant across methods of extraction.

2) General Language Understanding Evaluation Leaderboard: To address the issue of missing data, which constituted 0.23% of the dataset, mean imputation was employed [36]. The data demonstrated univariate skewness and kurtosis values that fell outside the generally accepted norms, indicating the non-normality of the data distribution.

Factorability: The KMO measure of sampling adequacy was recorded at 0.906, which is within the range labeled as "marvelous" [31]. Additionally, Bartlett's Test of Sphericity yielded $\chi^2(9) = 558$ with $p < 0.0001$. The average off-diagonal correlation was observed to be $r = .84$ ($p < .0001$

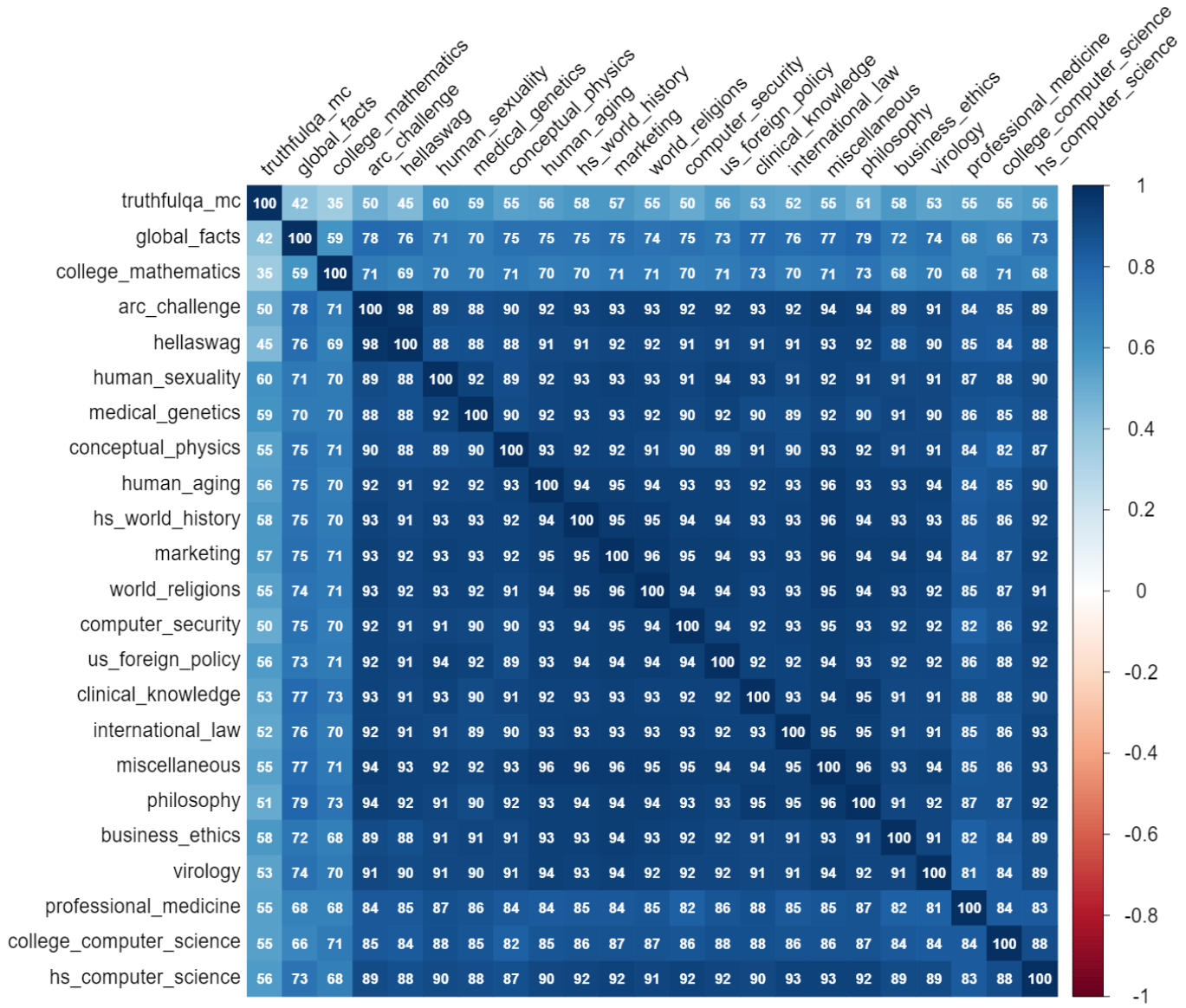

Fig. 3. Open LLM Leaderboard correlation matrix

for all values in the corr. matrix), which perfectly replicates the value from the Open LLM Leaderboard. The correlation matrix can be found in the Appendix.

Factor Extraction: All criteria unanimously indicated that a single factor should be extracted. The first unrotated factor accounted for 85% of the variance, again exactly replicating results from the Open LLM Leaderboard. Alternative one-factor and two-factor solutions were evaluated; however, the two-factor model was found to suffer from issues related to discriminant validity, most notably an inter-factor correlation of $r = .85$ ($p < .0001$; 95% CI $[.78, .9]$). As a result, a single-factor model was selected over the two-factor model.

Factor Loadings: The Root Mean Square of Residuals (RMSR) was calculated to be 0.03, signifying a good fit. Factor loadings for individual subtests are displayed in Table III.

B. Reliability Analyses

1) A singular g*:* As in the previous analysis involving the Open LLM Leaderboard, the dataset exhibited a minimal percentage of missing data (0.02%), which was addressed using mean imputation. The analysis yielded a distribution of Pearson correlation coefficients that measured the relationship between g-factors extracted from disjoint test batteries. The mean correlation coefficient was $r = .99$ (95% CI $[.96, 1]$) with a median of .99. The distribution of correlation coefficients is displayed in Figure 5.

2) Reliability of g*-loadings:* The issue of missing data was addressed in the same way as in the previous analysis. The

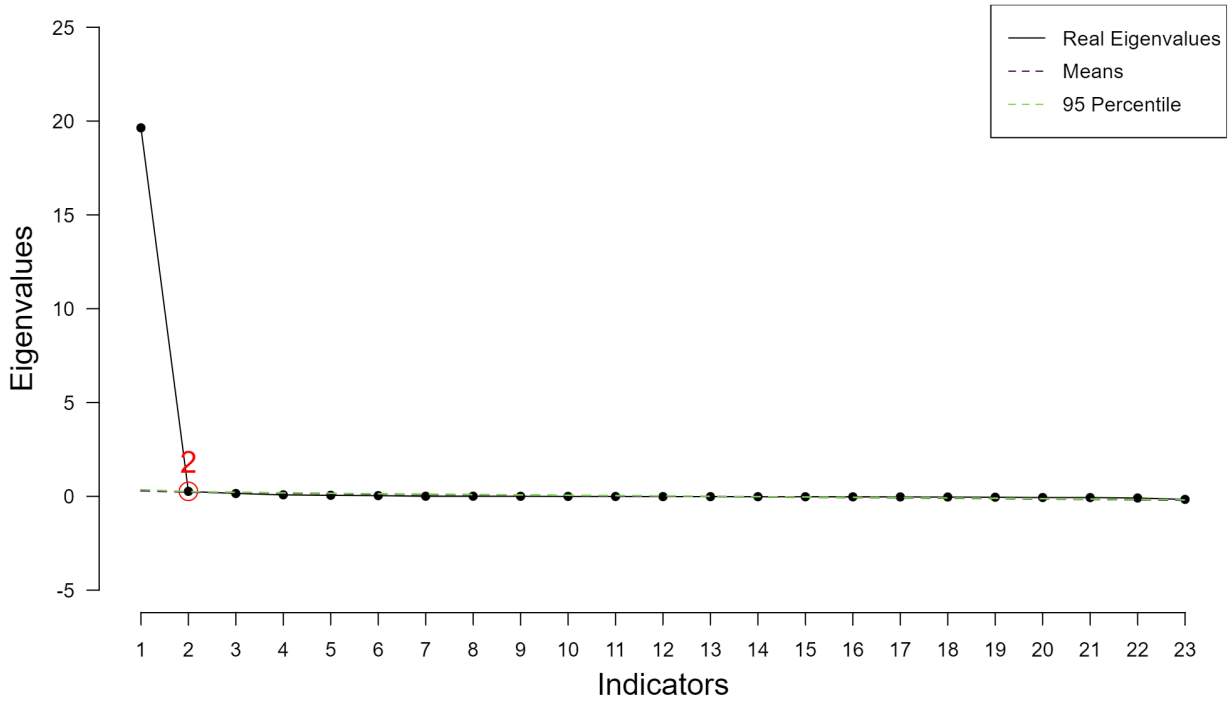

Fig. 4. Parallel analysis plot for the Open LLM Leaderboard

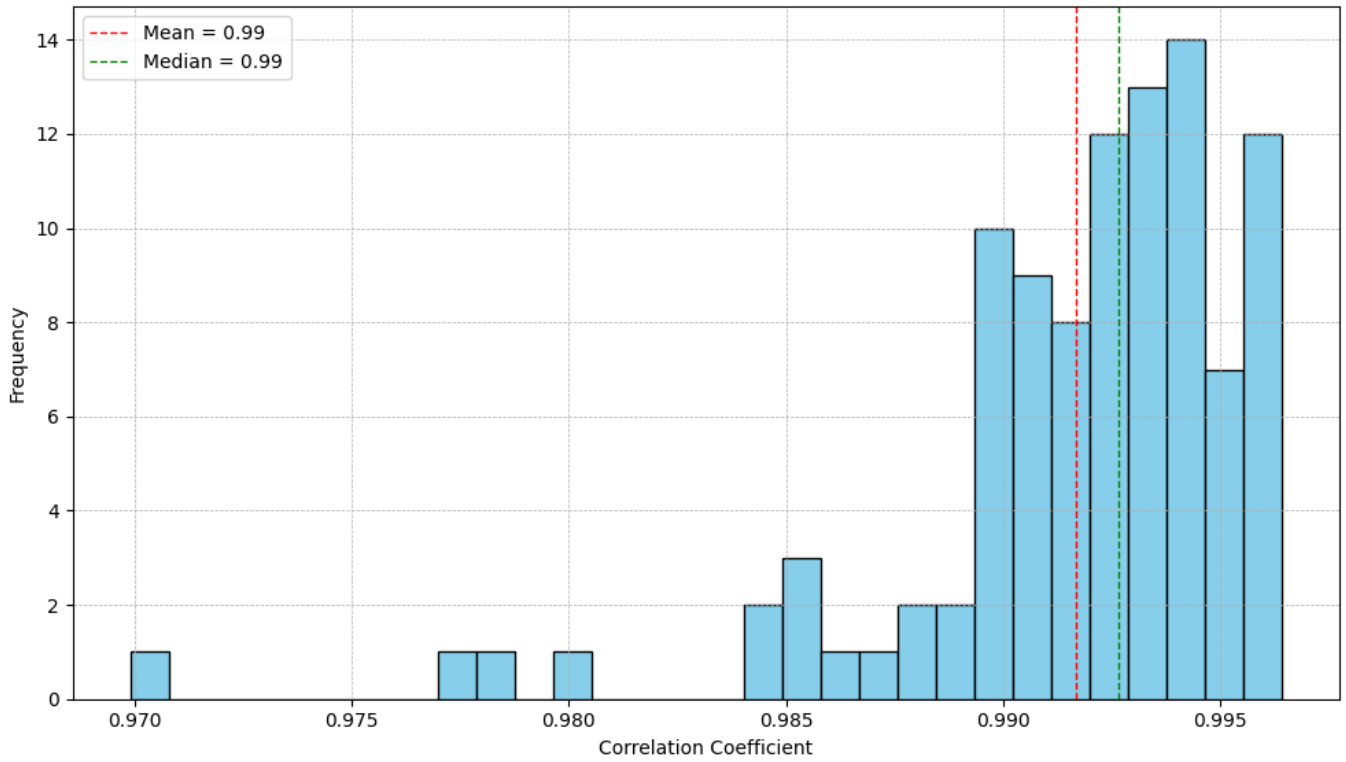

Fig. 5. Distribution of correlation coefficients between *g*-factors extracted from disjoint test batteries

TABLE I
DESCRIPTIVE STATISTICS AND PATTERN COEFFICIENTS FOR 1232 MODELS FROM THE OPEN LLM LEADERBOARD

	Descriptive Statistics		Factor Analysis	
Subtest	Mean	Std. dev.	g loading	h^2
ARC Challenge	48.1	14.0	0.96	0.92
Hellaswag	68.1	19.5	0.94	0.89
Business Ethics	43.0	15.7	0.95	0.90
Clinical Knowledge	44.2	16.3	0.96	0.93
College Computer Science	37.2	10.5	0.90	0.81
College Mathematics	30.9	5.6	0.74	0.54
Computer Security	49.8	19.4	0.96	0.92
Conceptual Physics	38.5	11.5	0.94	0.89
Global Facts	31.9	7.9	0.78	0.60
HS Computer Science	42.9	16.3	0.94	0.89
HS World History	53.2	23.3	0.97	0.94
Human Aging	48.0	18.3	0.97	0.93
Human Sexuality	46.9	19.6	0.95	0.91
International Law	53.4	21.4	0.96	0.92
Marketing	56.8	25.5	0.98	0.96
Medical Genetics	45.4	15.9	0.95	0.89
Miscellaneous	53.5	22.7	0.98	0.96
Philosophy	46.8	18.5	0.97	0.95
Professional Medicine	43.2	15.0	0.89	0.78
US Foreign Policy	56.4	24.8	0.97	0.93
Virology	37.9	10.2	0.95	0.90
World Religions	55.0	23.0	0.97	0.94
TruthfulQA MC	44.8	6.8	0.57	0.32

resulting Pearson correlation coefficient was $r = .98$ ($p < .0001$; 95% CI $[.97, .99]$)., indicating that the *g*-loadings of the common subtest were highly reliable and stable across different test batteries.

C. Correlation Between Model Size and g

The Pearson correlation coefficient between model size and the extracted *g* factor was found to be $r = .48$ ($p < .0001$; 95% CI $[.44, .52]$). These findings suggest a moderate to strong positive relationship between the size of the model and general ability.

TABLE II
G-FACTOR SCORES FOR 25 LANGUAGE MODELS RANDOMLY SELECTED FROM THE OPEN LLM LEADERBOARD

Model	Standardized g score
psmathur/orca_mini_v3_70b	1.8045
chargoddard/MelangeC-70b	1.7931
migtissera/Synthia-70B	1.7113
TheBloke/guanaco-65B-HF	1.2832
concedo/Vicuzard-30B-Uncensored	1.0978
FelixChao/vicuna-33b-coder	0.9571
gaodrew/OpenOrca-Platypus2-13B-thera-1250	0.9245
l3utterfly/llama2-7b-layla	0.5796
YeungNLP/firefly-llama-13b	0.5439
Aspik101/Nous-Hermes-13b-pl-lora_unload	0.5283
bongchoi/test-llama2-7b	0.4601
togethercomputer/GPT-JT-Moderation-6B	-0.0685
LMFlow/Robin-7b-v2	-0.0928
digitous/Janin-R	-0.7907
sartmis1/starcoder-finetune-selfinstruct	-0.8366
Devio/test-1400	-1.1427
EleutherAI/pythia-1b-deduped	-1.2016
w601sxs/b1ade-1b	-1.2225
KoboldAI/fairseq-dense-6.7B	-1.2279
clibrain/Llama-2-ft-instruct-es	-1.2357
DataLinguistic/DataLinguistic-34B-V1.0	-1.2485
EleutherAI/pythia-410m-deduped	-1.2720
MBZUAI/lamini-neo-125m	-1.3155
aisquared/dlite-v2-124m	-1.3232
BreadAi/PM_modelV2	-1.4153

IV. DISCUSSION

A. Summary of Key Findings

This study has made several contributions to our understanding of general intelligence in language models. The key takeaways are:

1) The general intelligence factor, referred to as *g*, has been found and successfully replicated across populations, test batteries, and methods.
2) The general intelligence factor accounts for 85% of the variance in ability for language models.
3) The uniqueness of *g* has been confirmed, as it correlates at $r = .99$ across different test batteries.
4) The g-loading of a subtest has been found to remain largely invariant between different test batteries, correlating at $r = .98$.

TABLE III
DESCRIPTIVE STATISTICS AND PATTERN COEFFICIENTS FOR 88 MODELS FROM THE GLUE LEADERBOARD

	Desc. Stats.		Factor Analysis	
Subtest *	M	SD	g	h^2
CoLA [39]	59.4	11.1	0.96	0.92
SST2 [40]	94.9	2.1	0.95	0.90
MRPC [41]	86.8	3.7	0.91	0.82
STSB [42]	87.5	4.9	0.96	0.93
Quora Question Pairs [41]	88.4	7.2	0.86	0.73
MultiNLI Matched [43]	87.0	3.9	0.98	0.96
MultiNLI Mismatched [43]	86.0	5.1	0.95	0.91
Recognizing Textual Entailment [44]	77.1	11.1	0.90	0.81
Winograd NLI [45]	71.9	14.4	0.78	0.61
Question NLI [46]	92.7	3.2	0.94	0.89

* Abbreviations: CoLA: Corpus of Linguistic Acceptability, SST2: Stanford Sentiment Treebank 2, MRPC: Microsoft Research Paraphrase Corpus, STSB: Semantic Textual Similarity Benchmark

5) A moderate correlation $r = .48$ has been found between the general intelligence factor and model size.

B. Hypothesis Testing

The findings strongly support our initial hypothesis that a general intelligence factor exists in language models. However, they contradict our apriori specified factor structure, which proposed multiple first-order factors with g as a second-order factor. Instead, the data suggest that a single factor accounts for the observed variance. While this is intriguing, we remain cautious and open to the possibility that a higher-order model may emerge in future studies. Additionally, as hypothesized, the findings indicate that this general intelligence factor remains highly consistent. The hypothesis stating a positive correlation between model size and general intelligence was also confirmed.

C. Relationship to Previous Research

To the best of our knowledge, this study represents the first effort to identify and quantify g in any non-biological system. While no previous studies have extracted the general intelligence factor from language models, our identification of the universally positive correlations between model abilities and g is consistent with research on human and some research on animal intelligence [47] [3] [48]. Furthermore, as in humans, the g-factor in language models is highly invariant [49] [28]. But, unlike in humans, g in language models appears to be a first-order factor [48]. Additionally, our findings on the correlation of $r = .48$ between model size and g are in agreement with previous research on the positive ($r = .24$) correlation between brain size and intelligence in humans [50]. Interestingly, the g accounts for 85% of the variation in language models but only accounts for 40% in humans, so the so-called "size" of the positive manifold and g is greater for language models [4]. This could be due to a lack of measurement error unique to psychometrics on language model populations, because no model was disproportionately trained on a specific ability, or some other reason. The analyzed subtests also have significantly higher loadings (average .92) than subtests usually do in humans (average .6), which further supports the claim that g is "stronger" in language models [49].

D. Practical Implications

The findings of this study offer several practical implications. First, the discovery of g in language models establishes a unified metric for comparing the capabilities of different models. This also allows for an objective ranking of various tests based on their g-loading, which in turn provides a standardized measure of test relevance for specific populations of language models. Furthermore, the identification of a general intelligence factor reduces the computational resources required for accurate model evaluation. Previously, assessing a model's performance necessitated running it through a broad array of tests, like the MMLU benchmark. With the introduction of a g factor, it becomes possible to evaluate a model's general capabilities by subjecting it to a limited set of tests and extrapolating its general intelligence from those results. This not only simplifies the evaluation process but also makes it more resource-efficient. We also advocate for focusing on improvements in g as the primary metric for evaluating advancements in language models. This emphasis on g is crucial for two reasons. First, the current landscape allows researchers to "SotA-hack" or otherwise selectively administer a variety of tests until a statistically significant result is achieved, potentially leading to inflated claims of model performance [51]. Second, improvements in specific abilities may not necessarily translate to enhancements in general intelligence, with such a phenomenon being observed in humans in the form of the Flynn Effect [52].

E. Limitations

While this study provides valuable findings about general intelligence in language models, several limitations must be acknowledged. Firstly, the sample size for the GLUE Leaderboard was relatively small ($n = 88$), which could potentially impact the robustness of the factor analysis results, despite meeting the bare minimum criteria. Secondly, although the exploratory data indicate a unidimensional model of intelligence, the study does not definitively confirm the factor structure. Future research may reveal the existence of a higher-order model.

F. Future Directions

Several avenues for future research are evident:

1) Confirming the true factor structure of intelligence in language models—whether it is a one-factor model, higher-order model, or something else entirely.
2) Is g present only in models trained on language or also in models trained on other modalities?
3) Investigating what factors, other than model size, explain variations in g.
4) Identifying tests with high g-loadings that are either challenging to train for or where training is easily detectable.
5) Exploring the impact of fine-tuning or RLHF on a model's general ability.
6) Examining the relationship between a model's general ability and its scores on measures of bias.

In conclusion, this study has laid the groundwork for understanding general intelligence in language models from a psychometric perspective, offering both theoretical insights and practical applications. The findings open up new questions and directions for future research.

V. Reproducibility

The data used in this study is sourced from the Open LLM Leaderboard and the GLUE Leaderboard at the time of this research. For transparency and ease of replication, the data and code used for this analysis are publicly available at GitHub.

VI. Conclusion

In this study, we successfully identified and quantified a general intelligence factor, termed g, in language models, corroborating its existence across different datasets, test batteries, and extraction methods. Our findings not only offer a unified metric for evaluating language models but also pave the way for more efficient, g-based assessments. Despite certain limitations such as the small sample size in the GLUE Leaderboard, the study represents a foundational step in understanding artificial general intelligence from a psychometric standpoint. Future research may focus on confirming the true factor structure of intelligence in language models, exploring other influencing factors, and investigating the implications for bias and fine-tuning.

References

[1] S. A. Mulaik, *Foundations of factor analysis*. CRC press, 2009.
[2] C. Spearman, *"General Intelligence" Objectively Determined and Measured*. Appleton-Century-Crofts, 1961.
[3] J. M. Burkart, M. N. Schubiger, and C. P. van Schaik, "The evolution of general intelligence," *Behavioral and Brain Sciences*, vol. 40, p. e195, 2017.
[4] R. Plomin, "The genetics of g in human and mouse," *Nature Reviews Neuroscience*, vol. 2, no. 2, pp. 136–141, 2001.
[5] I. J. Deary, "Intelligence," *Annual Review of Psychology*, vol. 62, pp. 453–482, 2011.
[6] M. Brunner, G. Nagy, and O. Wilhelm, "A tutorial on hierarchically structured constructs," *Journal of personality*, vol. 80, no. 4, pp. 796–846, 2012.
[7] C. L. Reeve and N. Blacksmith, "Identifying g: A review of current factor analytic practices in the science of mental abilities," *Intelligence*, vol. 37, no. 5, pp. 487–494, 2009.
[8] L. D. Matzel and B. Sauce, "Individual differences: Case studies of rodent and primate intelligence." *Journal of Experimental Psychology: Animal Learning and Cognition*, vol. 43, no. 4, p. 325, 2017.
[9] M. J. Galsworthy, J. L. Paya-Cano, L. Liu, S. Monleón, G. Gregoryan, C. Fernandes, L. C. Schalkwyk, and R. Plomin, "Assessing reliability, heritability and general cognitive ability in a battery of cognitive tasks for laboratory mice," *Behavior genetics*, vol. 35, pp. 675–692, 2005.
[10] S. M. Reader, Y. Hager, and K. N. Laland, "The evolution of primate general and cultural intelligence," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 366, no. 1567, pp. 1017–1027, 2011.
[11] J. Isden, C. Panayi, C. Dingle, and J. Madden, "Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success," *Animal Behaviour*, vol. 86, no. 4, pp. 829–838, 2013.
[12] M. Aellen, J. M. Burkart, and R. Bshary, "No evidence for general intelligence in a fish," *Ethology*, vol. 128, no. 5, pp. 424–436, 2022.
[13] F. Chollet, "On the measure of intelligence," *arXiv preprint arXiv:1911.01547*, 2019.
[14] N. J. Schaub and N. Hotaling, "Assessing intelligence in artificial neural networks," *arXiv preprint arXiv:2006.02909*, 2020.
[15] D. Barrett, F. Hill, A. Santoro, A. Morcos, and T. Lillicrap, "Measuring abstract reasoning in neural networks," in *International conference on machine learning*. PMLR, 2018, pp. 511–520.
[16] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
[17] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, and T. Wolf, "Open llm leaderboard," https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.
[18] J. C. de Winter*, D. Dodou*, and P. A. Wieringa, "Exploratory factor analysis with small sample sizes," *Multivariate behavioral research*, vol. 44, no. 2, pp. 147–181, 2009.
[19] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," 2018.
[20] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" 2019.
[21] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," 2022.
[22] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," 2021.
[23] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *GLUE*, 2019, in the Proceedings of ICLR.
[24] T. A. Kyriazos *et al.*, "Applied psychometrics: sample size and sample power considerations in factor analysis (efa, cfa) and sem in general," *Psychology*, vol. 9, no. 08, p. 2207, 2018.
[25] R. F. DeVellis and C. T. Thorpe, *Scale development: Theory and applications*. Sage publications, 2021.
[26] A. Costello and J. Osborne, "Best pr best practices in explor actices in exploratory factor analysis: Four or analysis: Four recommendations for getting the most from your analysis," *Pract. Assess. Res. Eval*, vol. 10, pp. 1–10, 2005.
[27] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, "Sample size in factor analysis." *Psychological methods*, vol. 4, no. 1, p. 84, 1999.
[28] A. R. Jensen and L.-J. Weng, "What is a good g?" pp. 231–258, 1994.
[29] H. F. Kaiser and J. Rice, "Little jiffy, mark iv," *Educational and psychological measurement*, vol. 34, no. 1, pp. 111–117, 1974.
[30] M. S. Bartlett, "Tests of significance in factor analysis." *British journal of psychology*, 1950.
[31] H. F. Kaiser, "An index of factorial simplicity," *psychometrika*, vol. 39, no. 1, pp. 31–36, 1974.
[32] R. B. Cattell, "The scree test for the number of factors," *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.
[33] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, pp. 179–185, 1965.
[34] W. F. Velicer, "Determining the number of components from the matrix of partial correlations," *Psychometrika*, vol. 41, pp. 321–327, 1976.
[35] R. Gorsuch, *Factor Analysis: Classic Edition*, ser. Psychology Press & Routledge Classic Editions. Taylor & Francis, 2014.
[36] M. W. Watkins, "Exploratory factor analysis: A guide to best practice," *Journal of Black Psychology*, vol. 44, no. 3, pp. 219–246, 2018.
[37] T. A. Brown, *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[38] T. A. Schmitt, D. A. Sass, W. Chappelle, and W. Thompson, "Selecting the "best" factor structure and moving measurement validation forward: An illustration," *Journal of personality assessment*, vol. 100, no. 4, pp. 345–362, 2018.

[39] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *arXiv preprint 1805.12471*, 2018.

[40] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of EMNLP*, 2013, pp. 1631–1642.

[41] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Proceedings of the International Workshop on Paraphrasing*, 2005.

[42] E. Agirre, L. M'arquez, and R. Wicentowski, Eds., *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007.

[43] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of NAACL-HLT*, 2018.

[44] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer, 2006, pp. 177–190.

[45] H. J. Levesque, E. Davis, and L. Morgenstern, "The Winograd schema challenge," in *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, vol. 46, 2011, p. 47.

[46] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[47] I. J. Deary and P. G. Caryl, "Neuroscience and human intelligence differences," *Trends in Neurosciences*, vol. 20, no. 8, pp. 365–371, 1997.

[48] J. B. Carroll, *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, 1993, no. 1.

[49] W. Johnson, T. J. Bouchard Jr, R. F. Krueger, M. McGue, and I. I. Gottesman, "Just one g: Consistent results from three test batteries," *Intelligence*, vol. 32, no. 1, pp. 95–107, 2004.

[50] J. Pietschnig, L. Penke, J. M. Wicherts, M. Zeiler, and M. Voracek, "Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean?" *Neuroscience & Biobehavioral Reviews*, vol. 57, pp. 411–432, 2015.

[51] O. Gencoglu, M. van Gils, E. Guldogan, C. Morikawa, M. Süzen, M. Gruber, J. Leinonen, and H. Huttunen, "Hark side of deep learning– from grad student descent to automated machine learning," *arXiv preprint arXiv:1904.07633*, 2019.

[52] J. te Nijenhuis, "The flynn effect, group differences, and g loadings," *Personality and individual differences*, vol. 55, no. 3, pp. 224–228, 2013.

Appendix

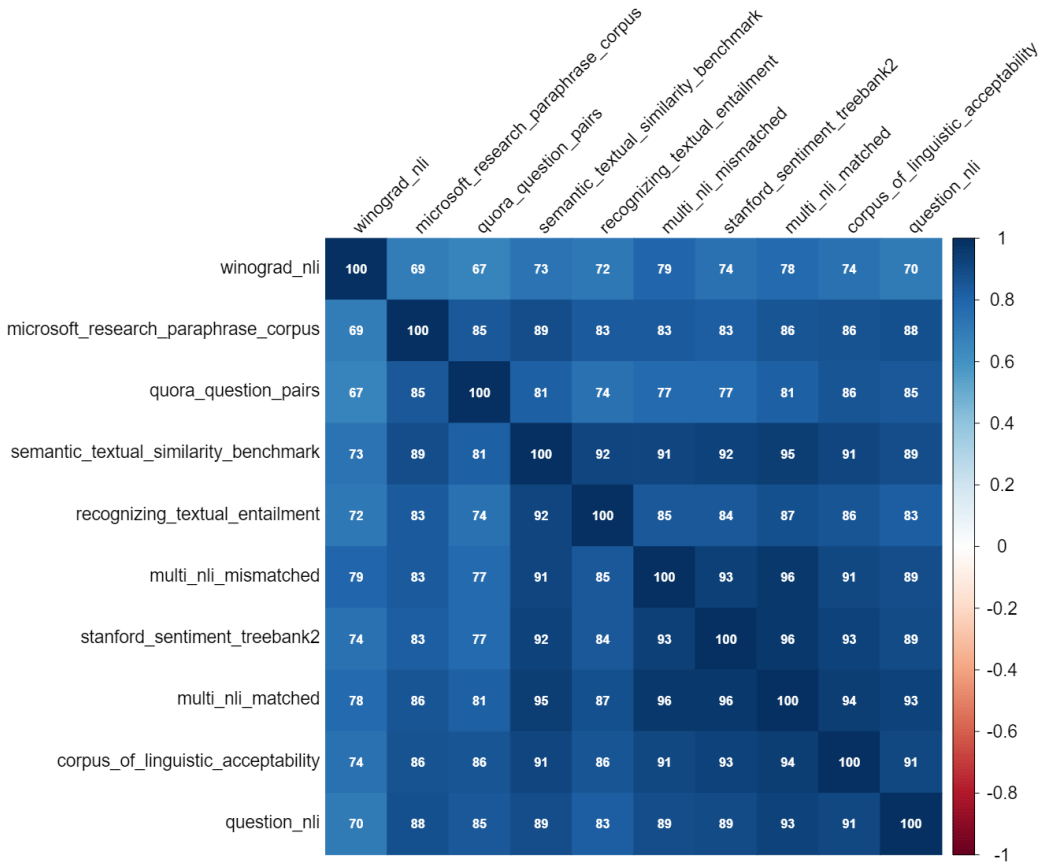

Fig. 6. GLUE Leaderboard correlation matrix

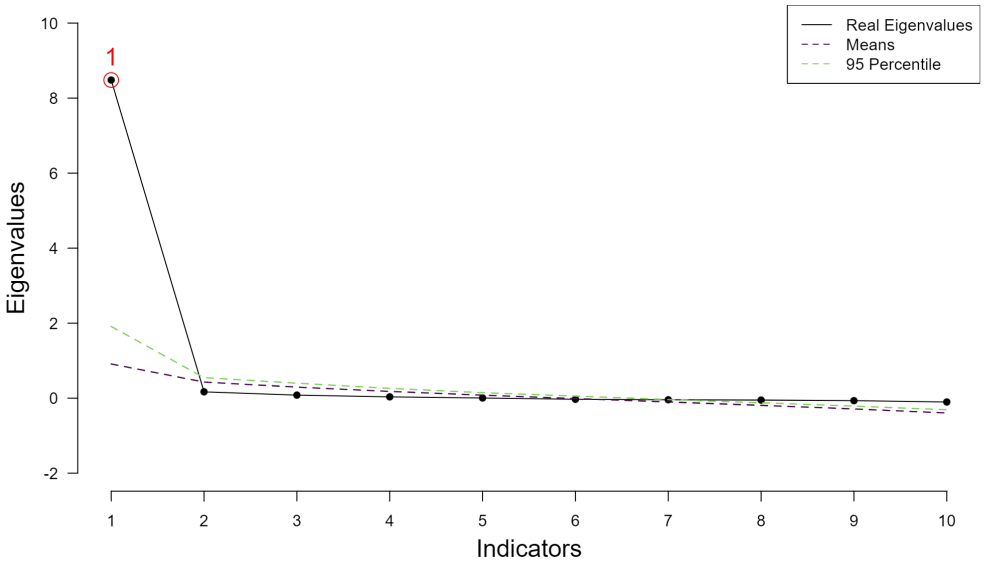

Fig. 7. Parallel analysis plot for the GLUE Leaderboard