GLOBAL LEADERSHIP ASSOCIATES PRESS

Validity and Reliability of AI Scoring of GLA's Developmental Psychometrics.

Authors:
Jonathan Frank – Global Leadership Associates, Chief Technologist ^
Asst. Prof. Kirill Veselkov – Intellligify Limited *

^ info@gla.global - for enquiries about the GLP, MWV and other GLA offerings
* office@intelligify.com - to whom correspondence about the technical aspects of this paper should be addressed.

In collaboration with and peer review from Global Leadership Associates and Intelligify Ltd.

Validity and Reliability of AI Scoring of GLA's Developmental Psychometrics.

Preface

In this paper we describe the way in which we developed an AI model to score the GLP, a leadership-focused psychometric. The accuracy of such a model is important since it forms feedback to an individual about their work, giving them clues as to how they can reach out and transform themselves and their situation rather than withdrawing from complexity in the world.

Since this AI scoring model was originally deployed in 2021, GLA has launched a derivative of the GLP, called MyWorldView (MWV). The scoring mechanisms for the two are identical, save for the granularity at which the scores are fed back to the user. As such, this paper should be taken to cover the automated scoring of both the original GLP psychometric and the newer MWV assessment.

Introduction

The Global Leadership Profile (GLP) is a sentence completion test, developed by Global Leadership Associates (GLA - https://gla.global) to analyse an individual's leadership stage of development: how they interpret and respond to problems, opportunities and relationships. The GLP is scored using a continuum of leadership styles rooted in adult development, or vertical development, theory.

An individual client takes the GLP or MWV by visiting a web page and completing 30 sentences in English. There are no constraints offered other than the opening 'stem' of each sentence, which the client uses as a prompt to complete the sentence. These completed sentences are each then scored, and the set of scores is combined into a single score - the "Total Protocol Rating" or TPR - for the client's profile.

Scoring the GLP is a specialist and complex task, requiring significant training and experience. GLA ensures that every scored GLP is supervised by a second scorer as a peer review. This ensures that any borderline scoring decisions benefit from a second opinion, reducing the likelihood of errors and leading to a reliable TPR. GLA has among its team many of the world's leading vertical analysts and scorers. This puts it in a unique position to use both its anonymised archives and its scoring expertise to develop a powerful AI scoring model which will operate without compromising accuracy.

State of the art

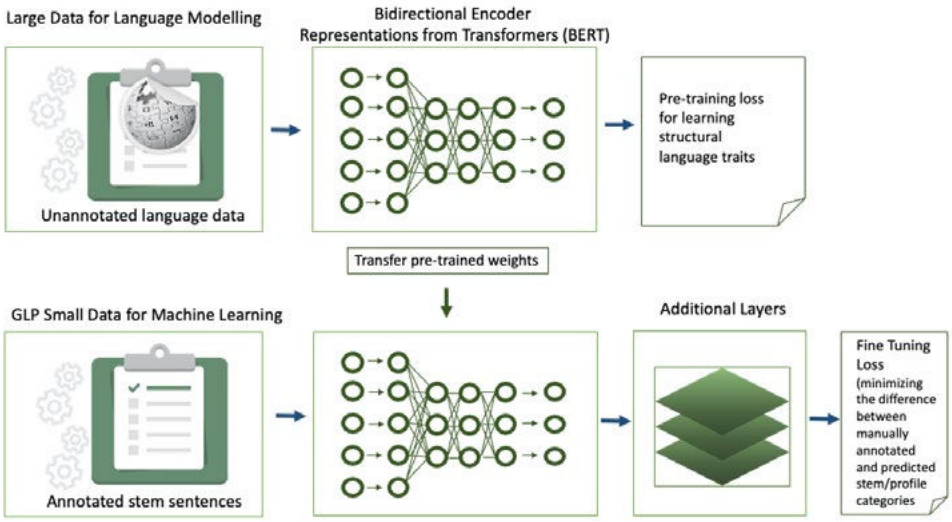
The process of scoring the GLP can be understood as a natural language processing task called "text classification". This process involves a trained expert assigning scores to text, including sentences or entire documents, in order to predict an individual's leadership qualities.

Automatic text classification is a method that can potentially reduce errors and make the time-consuming process of analysing the GLP more efficient. There are two main approaches to automatic text classification: rule-based methods and machine learning-based methods. Rule-based approaches to text classification involve using a set of predefined rules to assign texts to different scoring categories. These methods can be difficult to automate and typically require a thorough understanding of the subject matter being analysed. Machine learning-based methods, on the other hand, learn to classify text based on observations of data using expert scored examples as training data to learn the relationships between texts and their scores. Traditional machine learning approaches, such as Random Forest, Support Vector Machines, and Hidden Markov Models, often require domain-specific ("manual") feature engineering and thus do not usually perform well. Neural approaches, such as Recurrent Neural Nets and Convolutional Neural Nets, are able to overcome the limitations of using manually created features for text classification. However, they still need a large amount of pre-scored data to work properly. Recently, a new method called Pre-trained Language Models (PLMs) has emerged as a way to learn contextual text representations by predicting

words based on their context. PLMs have achieved the best performance in many natural language processing tasks, such as text categorization, even when there is only a small amount of scored target-domain data available.

GLA has uniquely created, tested and implemented transformer neural net architectures that are at the cutting edge of technology. GLA has demonstrated that these models based on PLMs (similar to those used by leading companies such as Google and Meta) are able to accurately learn the connections between stemmed sentences and categories related to leadership perspectives.

Since this AI model was deployed, ChatGPT and the subsequent slew of so-called "large language models" (LLMs) have taken the world by storm. While LLMs certainly present many interesting opportunities for other work with the GLP and indeed MyWorldView in the future, we see no need at present to redevelop a system that works so reliably. This model will remain as GLA's scoring engine for the foreseeable future.



The GLA end-to-end AI architecture is designed to process raw data through various stages, including sentence encoding and stacking network layers, in order to predict leadership qualities from unstructured text data.

Current practice (prior to deploying the AI scoring model) – dual scoring (peer review)

The scoring process is done by two experts, with the second expert serving as a moderator for the first expert's scores. Once the 30 GLP sentences are completed, they are each given a score on a scale of 3-9, which corresponds to seven stages of adult development. The combined set of 30 scored sentences is then scored as a whole, resulting in a Total Protocol Rating (TPR) on the same scale. However, the TPR scale also includes additional sub-steps, such as 5+ and 6- between the main levels of 5 and 6. These sub-steps also have a textual representation, with "Early" and "Late" prefixes, as shown in the table.

The individual stem scores are used to determine the TPR, possibly adjusted by the scorer using their overall judgment.

TPR / stem scores:	***Modifiers (TPR only):***
3: Opportunistic 4: Diplomat 5: Expert 6: Achiever 7: Redefining 8: Transforming 9: Alchemical	- : Early + : Late
	For example: 5+ is Late Expert 8- is Early Transforming 6 is Achiever

Accuracy of human scoring

Scoring accuracy is critically important to maintain credibility and 'face validity'. GLA has a reputation for highly accurate (and precise) scoring - partly delivered through well-trained and highly experienced scorers, and additionally through a process whereby the scoring of every profile is supervised by a second scorer.

However, even with two scorers one can still expect human biases in judgement. Also the scoring itself is a very strenuous and time-consuming process which limits the number of cases one expert can analyse given a limited amount of time as well as a limited number of experts available to provide a second opinion.

To address these limitations we aimed to develop an AI scorer of a comparable accuracy to a human expert which can be used as a primary scorer while still keeping a human expert in the loop to provide an independent opinion and/or a validation of the AI-generated score.

This AI development allows us to increase the throughput of GLP scoring without impacting its accuracy or precision. We developed and tested a fully-automated scoring engine, able to take a set of completed GLP sentence stems as input, and output a predicted score for each stem and a predicted TPR for the profile.

Research method

Available data

The training data consisted of 3,480 GLPs that were double-scored (by both a scorer and a peer reviewer), resulting in a total of 104,400 scored stems (30 stems per profile with scores ranging from 4 to 8). These profiles were scored by multiple scorers, but all of them adhered consistently to the same scoring manuals within the context of a single organisation, which placed a strong emphasis on data quality

Data pre-processing

Before using the data, it was pre-processed by eliminating double spaces, newline characters, tabs, and performing spelling corrections. Empty stems were not included in the data. (By default, the model scores empty stems as 4.)

GLP AI expert Architectural Design

The custom GLP "AI expert" neural network, which is based on pre-trained language models, evaluates GLP sentences for specific clients at different stages of adult development, and also calculates their overall protocol ratings (TPR). The network consists of two modules: a transformer neural module that predicts the scores for each GLP sentence, and an aggregator neural module that combines those predictions to determine the likelihood of each TPR leadership category using data from 30 GLP sentences.

The transformer module uses a pre-trained language model and additional deep learning layers to predict the probability of scores for each GLP sentence, ensuring overfitting is avoided. The pre-trained BERT "base" language model (trained on a large amount of English data including 800 million words from a book corpus and 2.5 billion words from Wikipedia) is used for this purpose. The model has 12 layers of transformer blocks, a hidden size of 768, and 12 self-attention heads, with 110 million trainable parameters. BERT is a transformer-based model that is pre-trained in a self-supervised way, using Masked Language Modeling and Next Sentence Prediction objectives, which allows it to learn a bidirectional representation of the sentence and an in-depth understanding of the English language, which are particularly suited for tasks such as scoring sentences. GLA enhanced the neural system by integrating BERT and adding deep learning layers to predict the scores of GLP sentences while avoiding overfitting. The module takes raw completed sentences as input, transforms them into a feature vector of dimension 768, and then passes them through classification layers to predict the probability of scores for each GLP sentence. These layers include a dense layer with 512 neurons and a ReLU activation function, which enables the learning of complex non-linear relationships between sentence vector embeddings and GLP scores. A dropout layer is then used for regularisation to prevent overfitting by randomly dropping 10% of neurons during training. The final dense layer employs a softmax activation function to output the probability of the sentence's score.

The architecture of the aggregator module combines dense layers, activation functions, and regularisation techniques to integrate the predicted scores of 30 sentences, generated by the transformer module, into a final TPR rating. The predicted scores are first reshaped through a Flatten layer into a one-dimensional array. The output is then fed into a Dense layer with 150 neurons, followed by a ReLU activation function to learn complex non-linear relationships between the scored sentences and overall TPR rating. A Dropout layer is employed as a regularisation technique to prevent overfitting by randomly dropping 10% of neurons during training. The model's capacity is further increased by using two additional Dense layers with 100 and 64 neurons respectively, both with a ReLU activation function. The final Dense layer employs a softmax activation function to output the probabilities for each TPR rating.

GLP AI expert tuning and optimisation

To improve the performance of GLP AI expert neural network for scoring GLP sentences and determining overall TPR ratings, we fine-tuned and optimised its parameters using a 5-fold, 5-recite cross-validation method. This method involved dividing the data into 5 equal parts, training the model on 4 of those parts and using the remaining part as the test data. GLA repeated this process 5 times, each time using

a different part as the test data. Additionally, we split the training data into 70% for training and 30% for validation. GLA used this 5-recite cross-validation technique to make our model more robust by repeating the process with different randomization of the data and different splits of the folds. GLA optimised several parameters, such as the sequence length, learning rate, number of epochs, and batch size, to make our model better at predicting scores for GLP sentences or overall TPR ratings compared to a well-trained human expert. GLA used a confusion matrix to evaluate the accuracy of our model by comparing the predicted output with the true output, and provided a detailed breakdown of the true positive, true negative, false positive, and false negative values.

GLP AI expert performance evaluation using previously collected data

In our cross-validation testing, the GLP AI scoring expert was able to predict scores for individual stems with a high level of accuracy, similar to that of a trained expert. The best results were achieved using a sequence length of 60, maximum of 5 epochs for the transformer module and 1000 epochs for the aggregator module, a learning rate of $5 \cdot 10^{-5}$, and a batch size of 16. The model's predictions were only slightly different from the expert's scores, which is typical among human experts.

We also compared the performance of models which were pre-trained uniquely for each of the 30 stems, and found that they had similar performance to the single general-use model. Given that the pre-trained models for individual stems were much larger in size (36Gb vs 1.2Gb) without adding any noticeable benefit to the prediction accuracy, the one-for-all transformer network was ultimately chosen.

Stem scores	4	5	6	7	8	else
Accuracy within a one step difference from the human expert's score	95%	99%	98%	99%	97%	80%

Table: The cross-validation summary of the AI expert scorer predictions of individuals stems within a one step difference from the human expert' score.

The GLP AI scoring predictive performance for the TPR rating within a one sub-step and one step difference from the human expert's rating was, respectively, 91.4±4.5 and 99.5±0.5 respectively.

TPR scores	5	5+	6-	6	6+	7-	7	7+	8-	8	8+
Accuracy within a one sub-step difference from the human expert's rating	98%	97%	98%	92%	92%	91%	90%	89%	90%	85%	83%
Accuracy within a one step difference from the human expert's rating	100%	100%	100%	100%	100%	100%	99%	99%	100%	97%	100%

Table: The cross-validation summary of the AI expert scorer predictions of TPRs within a one sub-step and one-step difference from the human expert rating.

Summary: GLP AI expert performance in real world settings

The performance of the GLP AI scoring expert in real-world scenarios was independently tested by comparing its predictions to those made by human experts on 951 independent profiles. The test took place on data which was generated between November 2021 and January 2023.

The assessment was carried out on both GLP and MWV scoring systems. The following comments however relate specifically to the MWV version, which combines adjacent "late" and "early" TPRs into a single "bridge" or transition category. Since the MWV scoring system is slightly less granular than the original GLP version, an increase in accuracy is to be expected, compared with the higher-precision GLP results from the cross-validation testing data shown above.

We found that in 99.9% of the cases, the AI expert predictions were a very close match with the human experts, showing a difference of no more than one step on the MWV scale of TPRs (for example, between Achiever and Achiever/Redefining Bridge); moreover in 76.0% of profiles, the AI prediction exactly matched the human expert TPR. This gives us confidence that the scoring system can be used in a general unsupervised manner with clients in MWV, as long as proper safeguards and ongoing monitoring are in place.

Overall, the GLP AI expert has learned from many years of human experience and expertise and has captured that knowledge during the training process to accurately predict the TPR of a set of completed GLP stems.

Authors:

Jonathan Frank – Global Leadership Associates, Chief Technologist

Jonathan is GLA's Chief Technologist. He has a maths degree from the University of Oxford, and his previous technology roles range from being a systems architect at IBM, to rescuing complex Salesforce implementations, to hands-on development of several full-stack systems and mobile apps.

Asst. Prof. Kirill Veselkov – Intellligify Limited

Kirill is co-founder of Intelligify, an AI innovation business. He is also an Assistant Professor at Imperial College London, and an Assistant Visiting Professor at Yale University.