The Language Proficiency Company
Learn. Assess. Certify.

AVANT RESEARCH · WHITE PAPER

Use and Validation of Avant's Proprietary Automated Scoring System

How Avant's LLM-based classifier scores STAMP Writing and Speaking.

Erica Sim, M.S. Computational Linguist, Avant
Victor D.O. Santos, Ph.D. Director of Assessment and Research, Avant
July 2026

Avant offers language proficiency assessments in over 150 languages and hundreds of thousands of test-takers take our assessments every year. Given the significant year-over-year growth in the number of clients who choose Avant, it is vital that we be able to maintain the high quality that our clients have come to know and expect of our assessments, and that we continue providing accurate and reliable assessment results in a timely manner.

Toward that goal, Avant has developed an in-house, proprietary tool that can automatically rate test-takers' responses in the Writing and Speaking sections of Avant STAMP, for a small number of languages. As we discuss below, the system is only used for a given language/skill combination (e.g., Spanish Writing) and specific STAMP levels (e.g. STAMP 4) once thorough validation of the accuracy and performance of our automated scoring tool has been established. We constantly and closely monitor the tool's accuracy and performance, and are able to immediately address any issues.

The implementation of our automated scoring tool provides several benefits to our clients, including: (a) the ability for Avant to provide test scores in a more timely manner than if only human raters were employed, (b) the fact that our scoring tool is 100% reliable and deterministic, always outputting the same score given a same response, (c) the fact that our tool often outweighs the accuracy of many human raters, and (d) the fact that it allows us to better monitor the performance and accuracy of our own human raters and better understand where they might need improvement. As will become clear by the end of this paper, Avant's human raters and our automated scoring system complement each other, leading to us being able to offer more efficient and accurate proficiency scores for our clients.

This paper outlines the difference between a generative large language model (LLM) and an LLM-based classifier (employed for our scoring system) as well as some of the characteristics of our automated scoring system.

The Difference Between a Generative Large-Language Model (LLM) and an LLM-based Classifier

A large language model (LLM) is a type of machine learning model that is trained on massive amounts of text data. They are huge models, allowing them to learn a great deal when trained on such data. They are initially trained to predict text given previous context and words, which gives them a kind of "understanding" of language based on the patterns of language usage in their training data. These models form the backbone of modern A.I. as they are both powerful and flexible, and can be fine-tuned, or further trained, on a specific task for a variety of purposes. This fine-tuning can also extend beyond only textual input.

One of the uses of LLMs commonly discussed in generative A.I. is a case where the model generates new text, images, or other content based on patterns it learned during training. This technology underlies chatbots like ChatGPT® and Claude®, as well as other applications like machine translation, image generation, automated transcription, and text summarization. However, there are many potential pitfalls when using a generative model. These models are non-deterministic, which means that they cannot reliably provide the same answer (output) when faced with the same prompt. Additionally, there are privacy concerns because the model can sometimes reproduce portions of text that were used in its training data.

Even though generative A.I. may be the most widely recognized application of LLMs, there are actually many other kinds of models based on this technology that are not generative; that is, they cannot create open-ended text (or images, etc.) based on their training data. These are models that are trained for a single, specific purpose instead of the general capabilities of something like ChatGPT®. The most common use case of this is classification, where the model assigns a label (or class) to each input. Some examples of this include spam filters, fraud detection, and, of course, automated scoring of written or spoken responses, which is one of the main purposes for which we employ such models at Avant.

The benefit of using an LLM as a base for classification/scoring is that the language understanding capabilities of the LLM from its original training data remain, while it is further trained on data specific to the task at hand. In our specific case, this data is hundreds of thousands of human-rated STAMP responses that have been de-identified and that are then converted to numerical values during the fine-tuning phase, not making it possible to link any given response to a specific test-taker. This task alone requires a large amount of high-quality curated data, but the model is still able to leverage the

enormous amount of data that went into training the base LLM. When fine-tuning the model for our proficiency classification task, the number of potential outputs is fixed (STAMP 1, STAMP 2, STAMP 3, and so on) and the model is incapable of returning any other output that is not a STAMP proficiency level. Additionally, the model becomes deterministic, always returning the same result (a specific STAMP level) if provided the same input (a test-taker response). This is a very important feature of the model, since it removes the possibility that any score awarded to a response is affected by the leniency or strictness of different raters.

In the LLM model used at Avant for automated scoring of responses, converting a base LLM into a proficiency classifier removes the text generation ability of the model and replaces it with the classification aspect.

The result is a model that understands language in much the same way as an LLM does but which is designed to perform only a single, well-defined task: rate responses at specific STAMP levels, with only STAMP levels for which the model's accuracy is comparable (or higher) than the accuracy of human raters being deployed and authorized for use. Therefore, Avant takes a very conservative approach when releasing our automated scoring tool for any STAMP level for a given language/skill combination.

Selective Training and Validation of our Automated Scoring Model

Our model, developed and validated in-house by Avant's Research team, uses a very large number of curated ratings by trained Avant human raters to fine-tune the base LLM, which converts it into a classification model that can only output STAMP proficiency levels. As part of this process, the responses are converted into numerical representations and used as part of an aggregated dataset. We trained multiple candidate models before choosing the model that best emulated how our human raters score the responses. We compared the results of different candidate models using different base LLMs, different sizes of models, different lengths of training, and more. Determining which of these models is best requires thorough testing of each candidate model on a completely different set of real-world STAMP responses (referred to as the 'test set') from those seen by the model during the training phase (called the "training set").

After selecting the most promising model, we conduct further validation to ensure that the model's accuracy meets (or surpasses) the standards that Avant has in place for our human raters. As part of the model's validation process, we compare the model's rating accuracy and misclassifications with those of our human raters on a large, common dataset of STAMP responses. In this analysis, we calculate accuracy both for the model overall (across all STAMP levels) as well as for each STAMP level. We also examine how far off any misclassification errors are from the true score since this allows us to only use our automated scoring tool for STAMP levels where we have validated that the model's accuracy for that specific level is at a highly acceptable level.

We only use the model for STAMP levels where the scoring model performs at least as well as our combined pool of human raters for that language, with the model's accuracy actually often surpassing the accuracy of an average human rater. If the model returns a STAMP level Avant's Research team has not validated for use or if the model is not confident enough in the STAMP level it has assigned to a test-taker's response, the response is instead immediately sent to a human rater.

The validation process does not stop once a model has been put into live use. We keep ongoing tabs on the model's accuracy on live data and continuously monitor the operational model's performance, checking for any potential issues. If we see that the accuracy for any validated STAMP level has dropped below our rating standards, we immediately disable that level for automated grading and route those responses to human raters. We also periodically re-run our validation on larger amounts of data so that we can detect any changes in performance over time and confirm that the model remains accurate as our prompts and responses evolve.

Once a specific STAMP level in our automated scoring tool is turned off for a given language/skill combination, it can only be turned on again once the Research team has investigated the model's performance for that level and been able to optimize its accuracy to a point that once again meets Avant's rating standards.

Going Beyond a traditional NLP Approach for Automatically Scoring Written and Spoken Responses in STAMP tests

When creating a classification model for the automatic scoring of written or spoken responses, such as the one used at Avant, two different approaches usually take center stage. One is a decades-old approach using traditional natural language processing (NLP), in which the linguistic features to be extracted from a given response are selected by researchers ahead of time and then calculated for each response. Then, they are fed into the classification model that will score the response in terms of its proficiency level. The other, much more recent approach discussed above uses a large-language model (LLM) as a base, which is subsequently trained on real ratings and responses holistically, with the model being free to select which aspects of the responses it deems important to focus on when building its understanding of language proficiency.

Traditional NLP approaches, which employ hand-crafted linguistic features, can offer a few obvious benefits. When designing these, the features are chosen ahead of time and are explicitly individually coded. This allows for a great deal of control over which features are used when scoring a response and which aspects of language are being measured and fed into the model. Additionally, this allows for a more explainable and transparent model, in which the value of each chosen feature can be examined for a given response. Although explainability and transparency of A.I. is an important concern and is one of the main strengths of such a traditional feature-based (NLP) approach, a highly interpretable system is not necessarily a more valid, accurate, or high-performing system.

It is quite often the case in language assessment that only looking at a limited number of pre-determined linguistic features limits the complexity of the model and prevents it from fully modeling the multidimensional nature of human raters' thought processes when scoring responses, even in cases where a robust scoring rubric, such as the one employed by Avant raters, is in place. Written and spoken language proficiency involves many interactions between vocabulary, grammar, and complexity, among other linguistic and communicative factors such as naturalness and fluency. In a traditional NLP framework, each of these interactions must be pre-determined and modeled explicitly. Calculating all of these individually quickly becomes computationally impractical since the model needs to be optimized and re-assessed every time a new feature is added.

In addition to computational constraints, there are other concerns when selecting which linguistic features to use for automated scoring of written and spoken responses. The more features that are added, the more the model has difficulty giving proper weight to each one. It may give some important features too low a weight when it is attempting to consider too many different features. Alternatively, it may consider features or interactions that are ultimately not very important to the final score. It will necessarily privilege the features that were chosen, while ignoring any patterns outside of those, which is another major shortcoming of traditional NLP models. Finally, the features that are important for lower-proficiency responses often differ from those used when rating higher-proficiency responses, which can prevent the model from being able to accurately rate more advanced language, which may involve many linguistic features that are not captured by the limited features employed by the model.

Features chosen in a traditional NLP framework are ultimately merely a proxy for measuring the construct and cannot always align with the subtleties of the rubric and linguistic knowledge that human raters use. It ignores the prompt/scenario upon which the student's response is based as well as the wider context of the response that goes beyond a specific sentence being analyzed for feature extraction. This larger context of the response is however, very important when rating, and is something that Avant's human raters are quite observant of. Without this wider context, the model can become less robust to new prompts or to changes in response quality, especially concerning vocabulary usage.

We have researched many different candidate models for our automated scoring tool, including many traditional NLP solutions. We have found that the results from merely NLP-based models did not meet the high rating accuracy standards that we hold our human raters to. Errors from the NLP tools used to calculate the values of the pre-selected features would easily propagate to other stages of the scoring system, significantly impacting the model's classification accuracy. Additionally, having to reduce the response to a small number of features did not allow the model to properly rate more

complex responses, and each linguistic feature we added proved to give greatly diminishing returns.

On the other hand, employing a solution that can rate in a single step, as the one currently used by Avant, prevents this problem. Instead of scoring by first calculating pre-determined features before feeding them into a model, our LLM-based model scores in a single step– thus reducing the chance of the type of propagation errors seen with traditional NLP-based models– and is trained directly on scores awarded by highly trained Avant raters and on the entire response. This significantly increases the linguistic context that the model can consider when evaluating the proficiency level demonstrated in any given response. This data-driven approach allows our model to automatically learn the features and patterns it will consider, out of a much higher number of candidate features than would be possible with a traditional NLP approach. It is able to automatically select model features that best match what human raters consider when scoring a response, and to optimize how much weight is assigned to each.

Our currently operational LLM-based classification model therefore looks at the entire context of the response, rather than a handful of predetermined features. During the extensive validation process of our LLM-based classifier, we have found that this approach leads to considerably higher accuracy when compared to a traditional NLP-based model, performing at par with, and often outperforming, a large portion of our trained Avant raters. Although it is harder to pinpoint the exact features that the model considers when rating a response, our use of an LLM-based model aligns with the increasingly standard practice across industries of using LLMs as a basis for classification models, given their higher accuracy.