Building Interpretable Natural Language Processing Models for Organizational Research

Mengqiao (MQ) Liu[1*],

Tianjun Sun[2*],

Feng Guo[3],

and Hanyi Min[4]

[1]Amazon.com Services, Inc.

[2]Rice University

[3]University of Tennessee at Chattanooga

[4]University of Illinois Urbana-Champaign

[This is a working paper and has not yet been peer-reviewed.]

Author Notes:

*Mengqiao (MQ) Liu and Tianjun Sun contributed equally to this paper.

Part of this paper was submitted for consideration in a symposium at the 2025 Annual Conference of the Society for Industrial and Organizational Psychology.

Correspondence concerning this article should be addressed to Tianjun Sun, Department of Psychological Sciences, Rice University, MS-25, 468 Sewall Hall, 6100 Main Street, Houston, Texas 77005 USA. Email: tianjunsun@rice.edu

Abstract

In the present work, we introduce and demonstrate the steps involved in using interpretable natural language processing (NLP) to predict decision-making in organizational contexts. Interpretability is important for scientific advancement, bias detection and mitigation, user trust, and ethical and legal compliance. In this study, we discuss why NLP interpretability is important and introduce a variety of interpretability methods that can be applied throughout the model development process. Using assessment center data from a recent machine learning competition organized by SIOP, we also showcase how we can develop and interpret NLP models of various intrinsic interpretability (logistic regression, XGBoost, DistilBERT, and GPT o1) through both global and local methods. This paper serves as an overview, guideline, and hands-on tutorial for applying interpretable NLP methods in organizational research, offering practical steps and considerations for researchers and practitioners. By combining conceptual discussions with practical applications, our work underscores the potential of interpretable NLP techniques in realistic organizational settings.

Keywords: explainable artificial intelligence, natural language processing, interpretable machine learning, organizational applications, tutorial

Building Interpretable Natural Language Processing Models for Organizational Research

Natural language processing (NLP) models provide a powerful way to help us uncover patterns and derive insights from text, yet these models are seen as “black boxes” by many – they provide relatively accurate predictions but often lack clear logic or explanations behind them. Organizational scientists have begun exploring and experimenting with using machine learning (ML) and NLP techniques on text (e.g., Campion et al., 2016; Speer, 2018; Putka et al., 2023), calling for a need to better understand and explain them. In this study, we aim to answer the following key questions: (a) *What is NLP interpretability, and why is it important in organizational research?* (b) *How can we build interpretable NLP models?* We hope this paper serves as a starting point for establishing responsible and best practices in organizational research toward interpretable text-based models.

We define a few technical terms here (see **Table 1** for a “cheat sheet” of related terminologies). ML is a subfield of computer science that aims to construct computer programs that can learn and improve with experience automatically (Mitchell, 1997). We commonly use ML models for two types of tasks, namely discriminative modeling and generative modeling. Discriminative modeling is used to optimize some predictions based on data (Hastie et al., 2009), such as using interview transcripts to predict personality. Discriminative modeling includes *classification* (for *categorical* criterion variables) and *regression* (for *continuous* criterion variables). Generative modeling can produce content such as text or images. NLP is a discipline that aims to program computers that can automatically process and learn human natural language data (Manning, 1999). NLP models often leverage ML techniques to process, represent, and analyze text data, and they can also handle classification or regression tasks. In **Table 2**, we summarize common NLP use cases for both classification and regression. In this paper, we focus on interpreting supervised NLP models, given their relevance in predictive modeling in organizational research and science.

In recent years, we have seen an initial set of publications by organizational psychologists using NLP techniques to address a variety of organizational topics, such as streamlining job analysis (Putka et al., 2023), assessing competencies (e.g., Campion et al., 2016) and personalities (e.g., Hickman et al., 2022), and scoring narrative performance reviews (e.g., Speer, 2018). In 2016, *Psychological Methods* produced a special issue on Big Data in Psychology, with papers providing overviews and tutorials on how big data techniques can be used in research relevant to organizational psychologists (e.g., Chen & Wojcik, 2016; Landers et al., 2016). In 2018, *Organizational Research Methods* introduced a special issue: “Ad Hoc” Feature Topic Big Data and Modern Data Analytics, featuring papers discussing how various big data techniques can be leveraged to advance organizational science and practices. In 2020, Personnel Psychology had a special issue call for papers that used machine learning and artificial intelligence to advance organizational research and practice (Campion & Campion, 2023; Woo et al., 2024). These research and applications benefit from the conceptual and methodological advances in the fields of computer science and artificial intelligence and showcase how such techniques can be applied to address organizational issues. NLP models and AI-driven chatbots are also increasingly being applied in organizational research for assessing individual differences, such as personality (Fan et al., 2023; Sun et al., 2024) and vocational interests (Chu et al., 2024). These applications have demonstrated promise, but the effectiveness and accuracy of chatbot-inferred traits still require further examination (Yuan et al., 2024).

Despite the increasing interest and applications of NLP for organizational research, we have seen little on model interpretability or explainability addressed in the context of organizational science. Henninger et al. (2023) showed various interpretability techniques for two types of ML models, namely random forests and neural networks, based on simulated, *tabular* data. However, to our knowledge, no published studies have been done on interpreting text-based, NLP models in the context of organizational research. We believe there are three main reasons for the lack of literature on this topic. First, the field of NLP is fast evolving, and so is the interpretability of these models. Although some standardization and systematic approaches exist in interpretability across different ML models (Doshi-Velez & Kim, 2017; Molnar, 2022; Rudin, 2019), advances such as large language models (LLMs) call for a reexamination of how interpretability operationalizes.

Second, as a field, organizational science is still relatively new to NLP techniques, where the published literature is, quite appropriately, mainly focused on illustrating specific use cases of applying NLP models to solve research or practical problems (e.g., assessing applicant competencies, Campion et al., 2016).

Third, issues surrounding model interpretability are not typical concerns for organizational scientists, given the “traditional” models (e.g., ordinary least square regression models, structural equation models) used for organizational science are, in most cases, intrinsically interpretable (i.e., easily explainable given their relatively simple and intuitive logic and structure). Therefore, seeking additional model interpretability might be a “blind spot” for organizational researchers. That said, given the increasing adoption of NLP models, organizational researchers who wish to leverage these more complex modeling techniques will find themselves needing to explain these sometimes not-so-intuitive methods.

In the sections to follow, we plan to (a) illustrate the importance of NLP interpretability and review methods and procedures to establish interpretations and (b) present a hands-on tutorial on building several explainable NLP models. We contextualize our entire discussion in organizational science and practice because interpretations of interpretability (pun intended) can vary across domains and users, i.e., interpretable to *whom*? (Lipton, 2017). Therefore, we will focus on addressing NLP interpretability issues most relevant to organizational researchers and present solutions that can be easily adopted by this targeted audience.

What is NLP Interpretability, and Why is it Important?

Interpretability can be broadly defined as "*the ability to explain or to present in understandable terms to a human*" (Doshi-Velez & Kim, 2017). Here, we use the terms interpretability and explainability interchangeably. Breaking this definition down further, "the *ability* to explain or to present…" posits that interpretability is not a dichotomous concept but a continuum; it is inaccurate to say model A is either interpretable or uninterpretable. In addition, "…*understandable* terms…" suggests that interpretability is inherently a qualitative concept; it is difficult to impose quantifiable measures on interpretability *as a whole*, although there are ways to derive quantifiable measures as *evidence* to support or enhance interpretability; this is discussed in detail in the next section. One can *probably* say model A is more interpretable than model B, but *never* model A is *n* times as interpretable as model B. Lastly, "… *to a human*" infers that interpretability is in the eye of the beholder, and here, we contextualize it as understandable to organizational researchers and stakeholders (e.g., business leaders, recruiters, candidates). Although not specified in this definition, we argue that interpretable NLP should *not* solely rely on deriving post-hoc explanations after a model has been developed. Instead, building interpretable NLP requires a systematic, theory- and data-driven process throughout the pre-modeling, modeling, and post-modeling stages.

Importance of NLP Interpretability

There is a growing interest in NLP interpretability; Google identified over 129k search results and 23.7k publications from 2020 to 2024 on this topic. The need for interpretability arises when the objectives of NLP models cannot fully address the needs in organizational research and practice. In a discriminative modeling task, the goal is to mathematically optimize a prediction goal, such as "maximize the prediction accuracy of personality based on interview transcript," while ensuring such prediction will generalize in *new* data. It does so by first *training* a model on a dataset with the predictors (e.g., interview transcript) and outcome (e.g., personality scores), a process that identifies a set of model parameter estimates that yields the closest prediction to the outcome. Then this trained model is *tested* in an independent testing dataset that has not been used in the model training process to see if the derived model parameters would generalize in an unseen sample. To that end, NLP can serve many areas of organizational science and practice by producing powerful, and sometimes superior prediction tools compared to traditional models. However, maximizing prediction is not the sole purpose of organizational research; we also hope to advance the fundamental knowledge of organizational science and guide best organizational practices. Therefore, we need to understand the question, "Why/how did the model make its decision?" in addition to what is predicted, especially considering the complexity of many NLP systems.

We argue there are four reasons why the interpretability of the methodology is critical for the advancement of organizational science and practice. First, a primary goal of organizational science is to advance our *scientific understanding* of critical issues relevant to the productivity and well-being of organizations and their employees. Whether a researcher is studying an organizational phenomenon for general scientific advancement or diving deep to understand a specific organizational problem or application at hand, it is essential to understand the "why".

For instance, an organizational researcher wishing to understand employee turnover can collect data around the employee (e.g., performance review narrative, work samples, etc.), feed them into a complex NLP model, and reach a high prediction accuracy of .90. However, this high prediction metric does very little in advancing our scientific understanding on this topic without understanding *why* employees choose to leave a company, i.e., what are the most important predictors of turnover. We also have unknown confidence whether this prediction would generalize beyond the sample at hand – What if a specific variable that is unique to a single organization was a main predictor for turnover in the data? What if this specific variable is not stationary? In such cases, the learned model will unlikely generalize to future samples. It is important to point out that, answering the "why" question is relevant regardless of whether a deductive or an inductive approach is used; deriving explanations from NLP models is crucial to (in)validating or informing hypotheses, with the ultimate goal of enriching theories and human knowledge. As Lipton (2017) points out, interpretable models allow researchers to better understand the "why" behind the predictions, thereby increasing the likelihood that these predictions will generalize across various settings and samples. In the context of NLP models, examining the data input-model output relationship helps identify whether the model is relying on robust, generalizable features or simply overfitting to idiosyncratic patterns in the data. By focusing on interpretability, we can scrutinize how changes in input data influence predictions, guiding adjustments to improve the model's adaptability to different organizational contexts. This process ultimately supports the development of more generalizable models, thereby enhancing their utility for researchers and practitioners in various organizational scenarios.

Second*,* seeking NLP interpretability is important in *detecting and mitigating biases*. Organizational science is an applied discipline, and collectively, our research has a significant impact on real people in real organizations. While NLP does not make subjective predictions *per se*, NLP models *will* learn and mimic the patterns of relationships that exist in the data and, therefore, inherit and sometimes even amplify existing biases that the data contains. For instance, research has shown that NLP models trained on Google News articles yielded significant gender biases (e.g., associating "nurse" with "female" and "captain" with "male"; Bolukbasi et al., 2016). It is important to note that, such biases were not intentionally built into the language models, but rather biases that the researchers did not consider a priori and only became aware after the fact. When left unexamined and untreated, the negative consequences of such biases can be detrimental to organizational practices. For instance, if an organizational researcher were to leverage this model to build a job recommendation system without understanding how the system derives its decisions, this researcher would further exaggerate gender biases in occupations without even knowing it. Behavioral science is known for relying heavily on data from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) samples (Henrich et al., 2010), which heightens the necessity to be wary of the potential biases that NLP models might inherit. Understanding how an NLP model makes predictions is crucial to detect biases that exist in data *and* algorithms and mitigate these biases before they turn into discriminations.

Third, being able to explain NLP systems in an understandable manner is crucial in facilitating *trust and adoption* from stakeholders and consumers of these systems. Drawing from the organizational justice literature (Greenberg, 1990; Colquitt et al., 2013), merely knowing the decision outcomes (distributive justice; e.g., whether one is hired or not) is only one piece of the puzzle. A sense of justice and fairness is fostered when individuals also perceive the decision-making process as consistent, accurate, and bias-free (procedural justice), as well as communicate the rationales behind the decisions in a respectful manner (interactional justice, which includes informational justice and interpersonal justice). We argue that procedural justice and interactional justice are critical to consider when dealing with NLP systems, given their "black box" reputation. Therefore, organizational decision makers need proper explanations to understand whether the model addresses the problem at hand, to ensure that the model predictions are accurate and fair, and to balance the impact of the model with other organizational objectives, such as diversity and inclusion.

On the other hand, end users of NLP systems, namely those impacted by predictions from the models, such as candidates and employees, will be looking for explanations on how they are being evaluated by the models, how the decisions are made, and whether they are being treated fairly by the algorithms. In addition, organizational researchers are also becoming consumers of NLP. Through this research, we also hope to promote trust and adoption from organizational researchers by unpacking and explaining NLP models, as well as demonstrating how they can do the same for their research and practice. Interpretability is often a prerequisite for trust, especially when deploying models in organizational research and practice. According to Ribeiro et al. (2016a), a model's ability to provide understandable and transparent explanations for its predictions significantly influences users' trust in its outcomes. In organizational contexts, decision-makers, such as business leaders, recruiters, and employees, are more likely to trust and adopt an NLP model if they can comprehend how it arrives at its conclusions. Without a clear rationale, even highly accurate models may be viewed with skepticism, limiting their practical use. Interpretability helps demystify the model's decision-making process, allowing stakeholders to evaluate whether the model aligns with their understanding of the underlying phenomena. This transparency not only fosters trust but also facilitates more informed decision-making. When users understand why a model makes specific predictions, they can better assess its reliability, identify potential biases, and apply its insights with greater confidence, making interpretability a crucial step toward achieving widespread acceptance of AI-driven tools in organizational settings.

Fourth, the interpretability of NLP models is necessary for *regulatory compliance*. The rise of artificial intelligence applications comes with increasing regulations on how algorithms can be used to make decisions that impact people, and that individuals have a right to be given explanations for algorithmic decisions (i.e., right to explanation). The European Union General Data Protection Regulation (GDPR; The European Parliament and the Council of the European Union, 2016) suggests that individuals who are "significantly" affected by algorithmic decisions have the right not to be subject to such decisions and to obtain an explanation (note that interpretations on "right to explanation" are heavily debated; e.g., Goodman & Faxman, 2017; Wachter et al., 2016). According to the updated Principles for the Validation of Personnel Selection Procedures (Society for Industrial Organizational Psychology, 2018) and Standards for Educational and Psychological Testing (2014), the use of automated scoring algorithms should be supported by theoretical and methodological bases to establish a clear linkage between the resulting scores and the criterion constructs of interest.

In the United States, several states have enacted new legislation governing the use of artificial intelligence in employment practices. For example, Illinois was one of the first to introduce the *Artificial Intelligence Video Interview Act*, which requires applicants to provide informed consent, including explanations of how the algorithm works when AI is used for analyzing video interviews. Since then, other states have followed suit with more comprehensive regulations. *New York City's Automated Employment Decision Tools (AEDT) law*, which took effect on July 5, 2023, imposes strict requirements on employers and employment agencies using automated tools to screen candidates. The law mandates annual bias audits of AI-driven decision-making tools, transparency with job applicants, and the disclosure of the specific qualifications or characteristics used by the algorithm in assessments. Similarly, *California's proposed Workplace Technology Accountability Act* aims to regulate the use of surveillance and decision-making technologies, including AI, within the workplace to safeguard employees' rights. At the federal level, the *Artificial Intelligence (AI) in Hiring Act of 2023* has been introduced in Congress, seeking to establish guidelines for fairness, transparency, and accountability in the use of AI-based employment decision-making tools. These recent developments highlight the increasing regulatory focus on ensuring the ethical use of AI in employment practices, underscoring the need for organizations to adopt interpretable and bias-free AI systems. As companies continue to adopt NLP into their organizational practices, we should only expect to see more regulations requiring interpretations.

Given the importance of NLP interpretability in organizational research and practices, we then attempt to answer the question, "*How can we build interpretable natural language processing models?*" by first introducing the different *sources of interpretability evidence* for NLP and then illustrating how one can establish interpretability evidence.

NLP Interpretability: Sources and Methods

The interpretability of an NLP system should be seen as a unitary concept with different sources of evidence that, taken as a whole, support an overall understanding of such a system. The different sources of evidence can and should be drawn from the various development stages of an NLP system, including pre-modeling, modeling, and post-modeling. It is difficult to rank the importance or relevance of these different interpretability evidence out of context; they should be judged for the specific use case of the NLP system (see **Table 3** for a summary of the sources of interpretability evidence and sample use cases). We suggest that organizational researchers and practitioners should (a) identify all relevant sources of interpretability evidence based on the use case of interest and (b) establish sufficient evidence for each of the identified sources.

In the following paragraphs, we will discuss (a) data interpretability in the pre-modeling stage, including data properties and exploratory data analysis; (b) model-intrinsic interpretability in the modeling stage, including simulatability, decomposability, and algorithmic transparency; and (c) post-hoc interpretability in the post-modeling stage, including global and local explanation techniques.

Pre-modeling Stage: Data Interpretability

Interpretability at the *pre-modeling* stage mainly concerns *data interpretability*, i.e., the extent to which the input data are interpretable. Data interpretability can be examined *prior to* the model development process to gain a thorough understanding of the data and examined *after* models are developed to understand how varying data input affects the model output. Data interpretability can be derived and evaluated based on (a) data properties, (b) exploratory data analysis, and (c) data input-model output relationship.

Data properties. The nature of the data impacts how intrinsically interpretable they are. We outline four properties or factors to consider when evaluating data interpretability. The first is whether data is construct-relevant or construct-ambiguous. In the world of organizational research, we are used to dealing with variables that are derived from and therefore mapped onto psychological constructs (e.g., “Am the life of the party.” to the personality trait of extraversion; International Personality Item Pool; Goldberg, 1999). However, in the NLP and data science literature, variables or features are sometimes construct-ambiguous. For instance, when email content data is used to develop an NLP model for spam detection (see Spirin & Han, 2012 for a review on web spam detection), it is difficult to directly link each word/phrase from the email content to a specific underlying construct. The second property to consider is whether the format or type of the data is easily understandable intuitively. One can argue that features from text data (e.g., words from emails) or numeric data (e.g., ratings on job performance) can be more understandable compared to those from audio or image data. Note that this is different from construct relevance: Pixels from medical imaging data can be directly linked to a particular type of gene mutation (construct-relevant) yet difficult for a human to understand. The third property that impacts data interpretability is the amount of data, both the number of variables (*k*) and the sample size (*N*), as "bigger" data can be harder to digest. The fourth data property to consider is the level of sparsity (or data density), which corresponds to the proportion of empty (vs. non-empty) cells in a dataset. A sparse data matrix with a lower number of *k* might be more interpretable but could create downstream problems for modeling.

Exploratory data analysis. Leveraging exploratory data analysis (EDA) techniques (Hartwig & Dearing, 1979; Tukey, 1977), organizational researchers can discover vital information about the data. EDA techniques leverage both descriptive statistics and data visualization methods to provide data information on three levels: (a) univariate (individual variable), (b) bivariate (relationship between two variables), and (c) multivariate (relationship among three or more variables). We encourage the readers to follow the best practices of EDA (Tukey, 1977) in the organization science literature (Aguinis & Edwards, 2014) and provide high-level results and insights from these analyses to aid in better interpretability of the data.

Data input–model output relationship. Understanding the relationship between data input and model output is fundamental for building interpretable NLP models. In NLP applications, input data characteristics—such as text length, vocabulary diversity, and syntactic complexity—can significantly impact the model's predictions. Understanding the input-output relationship allows researchers and practitioners to make more informed decisions about the model application, evaluate model biases, and ensure that the predictions align with theoretical and practical knowledge in organizational science.

Modeling Stage: Model Intrinsic Interpretability

Some models used in NLP systems have *intrinsic interpretability* due to their simple structures, such as sparse (i.e., small number of features) linear models (see **Figure 1** for common models by intrinsic interpretability). Here, intrinsic interpretability refers to the level of interpretability achieved through understanding the mechanisms by which the model works, i.e., "*how does the model work?*" (Lipton, 2017, p.4). Three key factors determine model intrinsic interpretability: (a) *decomposability*: Can one decompose and understand each part of the trained model (e.g., model parameter)? (b) *simulatability*: Can one simulate the entire trained model and calculate model predictions? (c) *optimization transparency*: Can one understand and replicate the optimization or learning process of the model?

Decomposability. Decomposability corresponds to how intuitively explainable each component of the model (e.g., parameter) is, such as the decision-making points in decision trees. It is worth noting that, decomposability also relies on data interpretability; it requires the input data themselves to be individually interpretable. For instance, if a feature in itself is ambiguous (e.g., pixels from a medical image) or heavily engineered (e.g., an interaction term between income and education and GPA), it will be difficult to derive explanations regardless of how "simple" the model structure is. This is an example where data interpretability and model interpretability go hand in hand.

Simulatability. Simulatability denotes that an interpretable model is a simple model, where a human can work through the calculations of the entire model with reasonable effort. In other words, with model parameters and input data, one can reasonably compute and produce predictions. Therefore, simulatability relies on both the size of the model (i.e., the number of model parameters) and the computation required to make predictions. A linear regression model is typically deemed transparent and interpretable given its high simulatability; one can relatively easily compute the predicted outcome value provided the intercept, coefficient(s), and feature value(s).

Optimization transparency. Optimization transparency refers to how transparent or visible the model optimization method is to those are use, govern, and are affected by them. A high level of algorithmic transparency requires a fundamental understanding of how the model works, as well as making this information accessible. For instance, it is easy to follow, describe, and replicate how a linear model converges in the learning process. The lack of algorithmic transparency is heavily discussed and debated in the area of deep learning (i.e., a family of methods based on artificial neural networks inspired by the network of the human brain) and LLMs, with researchers even equating it to “alchemy” (Rahimi et al., 2017). Deep learning transparency is often lacking due to inadequate knowledge of how those complex models work, reluctance to share intellectual properties that can generate great commercial values, and increased “randomness” (e.g., random dropout, a technique used to prevent model overfitting where some number of layer outputs are randomly “dropped out” in the training process; Srivastava et al., 2014).

Taken together, model intrinsic interpretability takes into consideration simulatability, decomposability, and algorithmic transparency. In practice, model intrinsic interpretability impacts, sometimes even dictates, decisions on model selection. Arguments have been made that, in high-stakes decision-making situations (e.g., healthcare and criminal justice), complex models that lack intrinsic interpretability should be avoided altogether (Rudin, 2019). In organizational practice, high-stakes situations such as pre-hiring selection warrant a thorough and careful examination on model intrinsic interpretability against other factors (e.g., model prediction accuracy); sometimes, organizational scientists *might* want to adopt a simpler model with slightly worse prediction performance, for the purpose that it can be easily explained and understood by the various stakeholder audiences (e.g., candidates, business leaders, recruiters, legal and compliance experts).

Post-modeling Stage: Model Post-hoc Interpretability

Model post-hoc interpretability, as the term implies, is derived by applying methods that analyze the model after training, i.e., "*What else can the model tell me*?" (Lipton, 2017, p.4). Post-hoc explanations are especially important to obtain when the models lack intrinsic interpretability, although they can also be supplied with simpler models when needed (e.g., to provide additional information to a layperson). NLP model interpretation methods can be *model-specific* (can only be used to interpret a specific model) or *model-agnostic* (can be used to interpret any model). One example of model-specific interpretation is the coefficients in a linear regression model. In this case, the model has high intrinsic interpretability, so one can interpret the model results by directly evaluating the direction and magnitude of the feature weights. In other cases, more complex models (with lower intrinsic interpretability) do not have a simple structure that allows a direct examination of the feature weights; they require additional model post-hoc interpretations to help understand what the models have learned.

When interpreting model results, one can approach from a *global* or *local* perspective, with the former focusing on understanding the general patterns of the model behavior as a whole and the latter focusing on understanding how a specific individual prediction is derived. In an example of predicting turnover from employee data, global interpretation methods will shed light on how the entire model makes its prediction on the likelihood of turnover (e.g., highlighting the most important predictors), whereas local interpretation methods will be able to explain why employee A has a probability of 0.79 for leaving the company in a year. As illustrated in **Table 3**, model global and local interpretability differ in their intent, and one might find either or both relevant depending on the use case of the machine learning system. In the following paragraphs, we will describe a few methods and tools that can be used to derive global and local interpretations from trained NLP models. More derivation details of these methods can be found in the **Appendix**.

Model post-hoc global interpretability. Model post-hoc global interpretability is determined by the extent to which predictions from a trained NLP model can be interpreted as a whole. Obtaining global interpretations is essential across multiple use cases of NLP in organizational research, including advancing scientific understanding (e.g., understanding the "why" and "how"), detecting and mitigating biases (e.g., detecting model systematic biases towards certain populations), and meeting regulatory compliance (e.g., demonstrating construct relevance). We describe three methods to establish model-agnostic global interpretations, including permutation feature importance, partial dependence plots (PDP), and global surrogate model.

Permutation feature importance. First introduced by Breiman (2001) for random forests, permutation feature importance measures the importance of the model features by randomly changing the values on each feature and measuring how such changes impact model performance. The rationale is, if a feature is important, then randomly changing its values will make the model predictions less accurate (higher error); if a feature is unimportant, then model performance should not be impacted much regardless of the values for this feature. This is based on the fact that NLP models rely on important features to make predictions. Permutation feature importance can be applied to *any* supervised learning model for feature evaluation and model interpretation. Permutation importance can be calculated using either the training set or the testing set. Using the training set will generate feature importance metrics that are more reflective of the trained model (i.e., what actually are the most important predictors in the trained model), whereas using the testing set will shed more light on which features contribute more to the generalizability of model (i.e., what features will remain important in predicting the outcome in a new, unseen testing data).

Partial dependence plots (PDPs). Leveraging the intuitive nature of visualization, PDPs are graphical representations of the marginal impact of chosen features on the outcome (Hastie et al., 2009). Due to human perception limitations, only one-way PDPs between one feature and one outcome, and two-way PDPs between two features and one outcome, are typically done. PDPs operate under the independence assumption, namely the chosen feature(s) are independent from the complement features in the model, which is often violated in practice. Nevertheless, PDPs are easy to generate and can help offer intuitive interpretations to even individuals who are machine learning novices.

Global surrogate model. As the term implies, a surrogate model is an approximation model, presumably one with high intrinsic interpretability, trained to mimic the predictions of a complicated, less-interpretable model. For example, you can train a linear regression model on the predictions from RoBERTa (Robustly Optimized BERT Pretraining Approach; Liu, 2019). If the linear regression model can replicate the predictions from RoBERTa well, then you can draw conclusions with some confidence about the RoBERTa model by interpreting the surrogate linear regression model. It is important to point out that, a global surrogate is built on the *predictions* of another model ($\hat{Y}$, or predicted Y), not the actual true outcomes (actual Y). Therefore, conclusions from global surrogate models are limited to the to-be-explained, original model, not the data.

To summarize, the three model-agnostic methods outlined above all target model post-hoc global interpretability by offering explanations on predictions from a trained NLP model. We recommend using permutation feature importance and/or PDPs if the goal is to interpret a model *via its features*, namely to better understand a given trained model by understanding the impact of the features on the predicted outcomes from the trained model. On the other hand, if one intends to replicate the predictions from a given trained model by using a more intrinsically explainable model, developing a global surrogate model is the way to go.

Model post-hoc local interpretability. Model post-hoc local interpretability is concerned with how individual predictions can be explained. Different from model post-hoc global interpretability that focuses holistically on all predictions from a trained model, we are "zooming in" here to try to explain how a single prediction is derived from a trained model. Therefore, while global interpretability is at the model level and does not vary across individual predictions, local interpretability is at the prediction level and is expected to vary across predictions. Establishing model post-hoc local interpretability is important for detecting and mitigating biases (e.g., in-depth analysis on individual cases showing biases), as well as fostering user trust and adoption (e.g., providing rationale for predictions to individual consumers). We outline two methods to establish model-agnostic local interpretations, namely local surrogate model and the Shapley value/SHAP (SHapley Additive exPlanation).

Local surrogate model. A local surrogate model is an approximation model used to explain individual predictions of a model. For instance, you can train a linear regression model on a single, local prediction from Llama2 (Large Language Model Meta AI 2; Touvron et al., 2023). If the linear regression model can replicate this prediction from Llama2 well, then you can leverage the interpretability of the linear regression model to explain this prediction from Llama2. Distinct from a global surrogate model that approximates the entire predictions of a trained model, a local surrogate model focuses on a single prediction at a time and tries to replicate this prediction via a more intrinsically interpretable surrogate model.

The most widely adopted implementation of local surrogate models is Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al., 2016b), a model-agnostic method that aims to identify a highly interpretable model that is locally faithful to the original model. To ensure both interpretability and local fidelity (i.e., the extent to a local prediction is approximated), one must (a) ensure high interpretability of the surrogate model g (i.e., keeping model complexity $\Omega(g)$ low), and (b) minimize loss $\mathcal{L}(f, g, \pi_x)$, which measures how unfaithful (i.e., distant) a chosen surrogate model g is approximating the original model f in predicting samples around a chosen instance x. In practice, LIME only optimizes part (b) local fidelity, while the researcher needs to determine the surrogate model g a priori.

Shapley values. Named after Lloyd Shapley, a Nobel Prize-winning economist, the Shapley value is a solution concept in a coalition game that represents the distribution of total gains to the players (Shapley, 1951; 1953). In other words, the Shapley value aims to “fairly” distribute the total payoff from cooperation among players based on each player’s importance to the success of the cooperation. Applying it to explain learned models, the Shapley value provides a mathematical solution to distribute the contribution to a single prediction among the features of a trained model. In this context, a model prediction is a payout, a feature is a player, and a coalition is a set of features "collaborating" to make a prediction. Four properties inherited from the Shapley value apply in model explainability: (a) *Efficiency* (or *Pareto Optimality*): The sum of Shapley values equals the grand coalitions (sets) of all features; (b) *Symmetry* (or *Equal Treatment*): If two features are substitutes (i.e., they contribute equally to all possible coalitions), then their Shapley values are equal; (c) *Linearity* (or *Additivity*): In an ensemble model (e.g., random forest), a feature's total Shapley value equals the linear combination of its Shapley values across models; and (d) *Dummy* (or *Null*): A feature's Shapley value is zero (i.e., null) if coalitions containing this feature never change.

Combining global and local interpretability: SHAP. Lundberg and Lee (2017) proposed a unified framework for interpreting machine learning predictions: SHAP (SHapley Additive exPlanation). Lundberg and Lee posit that the best explanation of a model should be the model itself; when a model (f) is complex, a simpler, more interpretable explanation model (g) can be created to approximate the original model. With this, they defined the class of *additive feature attribution methods* as having an explanation model that is a linear function of binary variables, which allowed them to connect global interpretability methods (e.g., LIME) and local interpretability methods (e.g., Shapley values) that satisfy this definition. As such, SHAP enables one to derive consistent local and global explanations (since they are all based on Shapley values fundamentally), an advantage over using separate global and local interpretability methods outlined above.

On local interpretability, SHAP introduces a novel method, Kernal SHAP, that combines linear LIME and Shapley values to improve the sample efficiency of model-agnostic estimations of SHAP values (compared to computing Shapley values). On global interpretability, SHAP leverages the collective Shapley values to provide insights on both feature importance and feature dependence. For feature importance, SHAP averages the absolute Shapley values per feature across the data to provide an estimation of feature importance globally (an alternative to permutation feature importance). Combining feature importance with feature effects, SHAP also provides *SHAP Summary Plot*, a visualization that depicts both the magnitude and direction of the features' impact on model output. For feature dependence, SHAP offers *SHAP Dependence Plots*, an alternative to PDPs, that visualizes the marginal impact of a chosen feature on the outcome.

Conclusion on NLP Interpretability

As we discussed above, sources of interpretability evidence of an NLP system can be derived from (a) data interpretability (pre-modeling and post-modeling stage), (b) model intrinsic interpretability (modeling stage), and (c) model post-hoc interpretability (post-modeling stage). Based on the research question, the data, and the chosen model(s), we recommend that organizational researchers should identify and establish the relevant interpretability evidence from multiple sources outlined above. In the next section, we present a hands-on tutorial contextualized in the organizational research domain and illustrate how organizational researchers can build and explain NLP models.

NLP Model Interpretability: A Tutorial

In this tutorial, we use explainable NLP to predict decision-making based on responses to a series of assessment center (AC) simulations. We chose decision-making as the focal research topic because it is a critical competency in organizational settings and has been extensively studied in the context of personnel selection and leadership development. The use of NLP for decision-making prediction addresses the need to automate the assessment of behaviorally anchored performance dimensions, making the process more scalable and cost-effective (Campion et al., 2016; Speer, 2018). We utilized open-sourced data from the SIOP 2023 Machine Learning Competition to promote open science; readers who are interested can access the dataset to replicate the tutorial results and explore further applications. In this tutorial example, candidate responses to AC exercises are text-based and exhibit varying degrees of complexity. This variability necessitates a model that not only accurately predicts decision-making scores but also provides interpretable insights into how specific input features influence these predictions.

Method

Sample. The data for this study comes from the SIOP 2023 Machine Learning Competition, which focused on predicting rater scores for text-based responses to AC simulations. The total sample size for this study comprises 1,466 responses in the training dataset and 487 responses in the development dataset. We have focused solely on analyzing the training dataset for this tutorial. These responses were gathered from real job applicants who participated in a series of in-basket email exercises as part of a leadership simulation. The absence of explicit demographic details in the dataset limits the ability to describe the participants’ age, gender, or other demographic characteristics. Nevertheless, the responses represent a diverse range of real-world decision-making scenarios encountered in organizational settings, making the dataset valuable for developing and evaluating NLP models.

Measures. The measures in this dataset assess multiple facets of decision-making competency through *six* key indicators. Raters evaluated candidates' responses to a variety of exercises using the following criteria: (1) *Chooses Appropriate Action*: The extent to which the candidate selects suitable actions in response to given scenarios; (2) *Commits to Action*: The candidate's ability to confidently follow through with decisions; (3) *Gathers Information*: The thoroughness of information collection to inform decision-making; (4) *Identifies Issues and Opportunities*: The candidate's skill in recognizing key problems and potential opportunities within a situation; (5) *Interprets Information*: The effectiveness with which the candidate analyzes and makes sense of available information; and (6) *Involves Others*: The candidate's inclination to involve relevant stakeholders in the decision-making process. In addition, there is *Overall Decision-Making*, a holistic score that reflects the candidate's overall decision-making effectiveness across the exercises was reported, and this score uses a feedback reporting scale of Need for development (1-2), Proficient (3-5), and Strong (6-7). Our tutorial focuses on the rating task of the overall decision-making score. For illustration purposes, we merged these feedback options to create a binary classification task: scores of 1-2 for ineffective decision-making and scores of 3-7 for effective decision-making.

Overview of the Exercises. The AC simulations involved in-basket email exercises designed to mirror realistic workplace challenges. Examples of these scenarios include managing team conflicts, handling customer complaints, making strategic decisions, and mentoring. Candidates were tasked with responding to situations including (1) *Larry Hodges and Emily Carson disagree about SEQUENCE*: Handling conflict between colleagues over a work process; (2) *Request to move a team member with performance problems*: Deciding whether to transfer a team member due to performance issues; (3) *Kirkland plant*: Addressing concerns related to team members transferring from an older plant; (4) *Professional conduct*: Dealing with inappropriate team member behavior during a customer tour; (5) *SEQUENCE talk*: Managing worker complaints about increased effort without additional compensation; (6) *Team focus*: Responding to data showing higher error rates when the team lead can't support the group; (7) *Theft of company information*: Handling a situation where an employee is suspected of stealing confidential information; (8) *Turntable proposal*: Evaluating a new design proposal for production processes; (9) *Upgrades to robot software*: Deciding on IT's proposed software rollout plan; (10) *Victory lunch*: Managing a reward system for group leaders based on performance survey results; (11) *Customer satisfaction insights*: Reviewing and reacting to research on customer satisfaction; (12) *Promotion*: Making a decision about promoting a part-time employee to a full-time position; (13) *Weedler Contracting*: Addressing a situation where a customer is suspected of abusing company replacement policies; (14) *Eluto Caplanu*: Investigating an incident where an offensive message was left for an employee; (15) *Effective mentoring pays dividends*: Reviewing the impact of a mentoring program within the organization; and (16) *Bench strength*: Ensuring that representatives are signed up for mandatory training per company policy. These exercises provided a diverse range of scenarios, allowing raters to assess candidates' decision-making skills comprehensively across different organizational contexts. By using open-ended responses in realistic simulations, the dataset captures the complexity of naturally unfolding text in decision-making situations, providing a rich source for evaluating NLP models.

Model Descriptions and Intrinsic Interpretability.

To illustrate the practical application of NLP models in organizational research, we drew upon techniques explored in recent studies that leverage both classic and advanced NLP models for smarter people analytics (Guo et al., 2024). These models have been shown to improve measurement and prediction in personnel selection by integrating large language models with traditional assessment methods (Koenig et al., 2023). It is important to choose models that are suitable for answering the focal research question and the nature of our data. We focus on predicting decision-making using classification models. For the purposes of optimizing prediction and demonstrating NLP interpretability, we selected a few widely used classification models with varying levels of model intrinsic interpretability: logistic regression, eXtreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016), DistilBERT (Sanh et al., 2019), and GPT o1 (OpenAI, 2024b). Below, we describe these methods in detail.

Logistic Regression. Logistic regression (Kleinbaum & Klein, 2002) is a widely used statistical technique in organizational research. It is a generalized linear model for binary data where estimated values are probabilities of category membership (and thus range between 0 and 1). The method is efficient yet powerful. Particularly, it offers straightforward interpretability where the resulting model comes with weights associated with each feature.

XGBoost. A variant of gradient tree boosting system (or gradient boosted trees; Friedman, 2001), XGBoost is essentially a decision tree ensemble that combines the predictive power of many decision trees. Different from a random forest that builds trees independently and aggregates the results at the end of the process, XGBoost builds trees additively, where each tree focuses on correcting the errors coming from the previous tree (i.e., boosting), which overcomes the shortcomings of individual trees and further improves model prediction accuracy. Past gradient tree boosting systems usually requires heavy computation and is slow to train. XGBoost overcomes this issue by implementing a series of system optimizations (out-of-core computation, cache-aware and sparsity-aware learning), enabling it to scale and improve speed ten-fold. As a result, XGBoost is accurate and fast, making it a winner in many machine learning competitions (e.g., Kaggle competitions, KDDCup; Chen & Guestrin, 2016).

DistilBERT. Bidirectional encoder representations from transformers (BERT) is an open-source language model that has achieved high-level performance across numerous NLP tasks, such as text classification, question answering, and named entity recognition (BERT; Devlin, 2018). BERT has 340 million hyperparameters and is based on the Transformer architecture (Vaswani et al., 2017). The power of BERT came from pre-training the language model on an extensive corpus that included the entire Wikipedia and the BooksCorpus. Language model pre-training refers to the process of feeding a large amount of unlabeled text data to a language model to help the model "learn" the general language before it is applied to a specific NLP task. Once pre-trained, BERT can be fine-tuned for specific NLP applications with relatively small datasets, making it highly adaptable and efficient for a wide range of tasks.

DistilBERT is a smaller, faster, and more efficient variant of BERT, introduced by Sanh et al. (2019). It retains 97% of BERT's performance while using only 40% of its parameters, making it a more lightweight and accessible model. DistilBERT achieves this reduction in size through a process known as "knowledge distillation," where a smaller model (DistilBERT) learns to approximate the behavior of a larger model (BERT). It is based on the same Transformer architecture as BERT but is optimized for speed and memory efficiency, making it particularly suitable for scenarios requiring real-time processing or limited computational resources. Despite its compact size, DistilBERT maintains strong performance across numerous NLP tasks, including text classification, question answering, and named entity recognition. Its open-source nature and reduced computational demands have contributed to its widespread adoption in the NLP community.

GPT o1. GPT o1 is the latest LLM introduced by OpenAI in September 2024. It represents the new generation of LLMs that integrate reinforcement learning with human feedback (RLHF; Christiano et al., 2017) and chain-of-thought prompting (Wei et al., 2022) to enhance both model performance and interpretability. Reinforcement learning with human feedback allows the model to align more closely with human preferences by fine-tuning it through feedback on its responses. Chain-of-thought prompting enables the model to generate more coherent and detailed responses by breaking down complex reasoning tasks into a series of intermediate steps. Together, these techniques improve the model's ability to generate nuanced and interpretable outputs across a variety of NLP tasks.

Analyses

Before building the binary classification model to predict effective vs. ineffective decision-making, we concatenated all exercise responses for each respondent. Doing so allowed us to infer decision-making from a diverse set of situations throughout the AC. For the logistic regression and XGBoost models, we removed non-English characters, converted all text to lowercase, and removed stopwords. An advantage of the DistilBERT model is that it does not require data preprocessing, so we did not alter the raw text data before feeding it to the model.

Our analysis goal was twofold. First, in building the NLP models to answer our research question, we wanted to maximize the prediction of decision-making given a collection of simulation responses while ensuring such prediction would generalize to unseen data. Second, to demonstrate model explainability, we conducted additional analyses on each trained model to derive interpretations understandable to organizational researchers.

The model development process varied based on the nature of the model type. For logistic regression and XGBoost, we trained the model with our dataset. For DistilBERT, we fine-tuned the pre-trained DistilBERT model with our dataset. For GPT o1, we used zero-shot learning to derive predictions. To ensure model generalizability, we first split the data into 80:20, where 80% (training data, *n* = 1,172) was used for training the models and the rest 20% (testing data, *n* = 294) for evaluating the generalizability of the learned model parameters. Second, we trained logistic regression and XGBoost, and fine-tuned DistilBERT, on a dataset with the predictors (e.g., responses to AC exercises) and outcome (e.g., decision-making score). By training the models, we identified a set of parameter estimates for each model that yield the closest prediction to the outcome. We then evaluated the trained models in the 20% independent testing dataset to see if the derived model parameters would generalize. Trained models were evaluated based on four metrics: Area Under the Receiver Operating Characteristic Curve (ROC AUC), precision, recall, and (point-biserial) correlation. We included GPT o1 to showcase how organizational researchers can use simple prompts to understand the rationale behind an LLM's prediction. Specifically, we used a zero-shot prompt that instructs the model to classify effective vs ineffective decision-making. The prompt (**Table 4**) describes the role of the model as an expert AC assessor, introduces the AC design, details the six decision-making dimensions, and breaks down the task step by step.

Metrics for Model Evaluation

To evaluate the performance of our models, we used several commonly adopted metrics: ROC-AUC, precision, recall, and correlation. Each provides a different perspective on model accuracy and effectiveness.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve). The ROC-AUC metric measures the model's ability to distinguish between classes across all possible thresholds. The curve plots the true positive rate (sensitivity) against the false positive rate (1 -

specificity) at various threshold settings. The AUC (Area Under the Curve) summarizes the overall ability of the model to correctly classify positive and negative instances. An AUC of 0.5 indicates no discrimination (i.e., random guessing), while a value closer to 1 indicates better model performance. This metric is particularly useful for assessing the model's overall predictive power without relying on a specific threshold.

Precision. Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It reflects the model's ability to only identify relevant instances among the ones it predicts as positive. High precision indicates a low false positive rate, which is critical when the cost of false positives is high (e.g., incorrectly predicting that a candidate exhibits strong decision-making skills).

Recall (Sensitivity or True Positive Rate). Recall is the ratio of true positive predictions to the total number of actual positive instances. It measures the model's ability to identify all relevant cases. High recall indicates that the model is effective in capturing most of the positive instances, which is vital when the cost of false negatives is high (e.g., missing candidates who truly possess strong decision-making skills).

Correlation. Correlation measures the strength and direction of the linear relationship between the predicted scores and the actual scores. For this task, we adopted the point-biserial correlation to represent the relationship between a continuous variable and a binary variable. A high positive correlation (close to 1) suggests that the model's predictions align closely with the true scores, providing a straightforward and interpretable metric for evaluating models in the context of predicting continuous variables (e.g., overall decision-making scores).

Metrics Choice. Given the nature of our task (i.e., predicting decision-making from text-based response), we used a combination of these metrics. ROC-AUC offers a broad view of model performance across thresholds, while precision and recall provide insight into the trade-offs between false positives and false negatives. Correlation is also particularly useful here, as it directly measures how well the model's predicted scores match the actual rater-assigned scores. Using these metrics in combination provides a comprehensive evaluation of the models, capturing both their classification performance (ROC-AUC, precision, recall) and their alignment with continuous outcomes (correlation). This multi-metric approach allows for a more nuanced understanding of model accuracy and interpretability in predicting decision-making competencies.

The second goal of the analysis was to demonstrate model explainability. Because model interpretability depends on both model intrinsic and model post-hoc interpretability, we tailored our interpretability analyses and illustrations to each of the trained models based on how the models work and how intrinsically interpretable they are. For instance, creating a global surrogate model can be an appropriate method to help interpret a trained XGBoost model due to its low model intrinsic interpretability, but it is unnecessary for a trained logistic regression model. We explain our rationale for choosing the interpretability methods for each trained model and walk through our model post-hoc interpretability analysis and results below.

We conducted all analyses in Python 3.10.12. We utilized the Optuna package for XGBoost and "distilbert-base-uncased" for BERT.

Model Results and Post-hoc Interpretability

Table 5 shows the model prediction results. Overall, the models adequately predicted the outcome: Logistic regression (ROC-AUC = 0.78, precision = 0.68, recall = 0.61, $r = 0.48$), XGBoost (ROC-AUC = 0.72, precision = 0.69, recall = 0.59, $r = 0.38$), and BERT (ROC-AUC =

0.71, precision = 0.64, recall = 0.60, *r* = 0.34). All the results are based on the holdout testing set to ensure generalizability.

Logistic regression interpretations. The structure of the logistic regression model makes interpreting results straightforward. To derive model global interpretations, we decomposed the various components of the learned logistic regression model and evaluated the direction and magnitude of most predictive words (see **Figure 2** for top 30 most important features). Across all the words, data (2.14), jj (1.57), and team (1.30) had the largest positive impact on predicting decision-making, whereas yes (-0.90), thank (-0.89), and meeting (-0.78) had the largest negative impact. The positive features are representative of the key decision-making behaviors that are being scored in the AC, such as gathering data, interpreting information, and involving others (JJ is the name of a key employee), whereas the negative features might indicate a lack of critical thinking (see examples below).

To derive model local interpretations for the logistic regression model, we selected two example cases: Case 28 with a high predicted decision-making of 0.94 and Case 741 with a low predicted decision-making of 0.24. Because logistic regression model generates intrinsically interpretable feature weights, we can directly interpret the highest weighted features of these individual cases to understand what is driving high or low prediction of decision-making. Specially, we used the global most positive and most negative features and multiplied the feature weights by the actual feature values from these individual cases. For Case 28, high positive values on global most positive features such as jj (0.35) and data (0.22) drove the high prediction (**Figure 3**), e.g., "*JJ, Thanks for the email and getting me up to speed with this project. I'd like to set up a meeting for later today to review your data and research*." On Case 741, high values on global most negative features such as thank (-0.27) and yes (-0.12) drove the low prediction

(**Figure 4**), "*Yes lets go forward with your plans and thank you for informing my team, i will touch basis with them when i get back from my 2 week vacation starting tomorrow*."

XGBoost interpretations. XGBoost is a gradient tree boosting system that combines the predictive power of many decision trees. XGBoost has lower model intrinsic interpretability compared to linear models like logistic regression. Below, we demonstrate how one can derive model post-hoc global and local interpretations based on the training dataset.

For model post-hoc global interpretability, we leveraged permutation feature importance, PDP, and global surrogate model. To create permutation feature importance, we randomly changed the values on each feature and measured how such changes impact model performance. As shown in **Figure 5**, the permutation importance plot showed that data (0.04), work (0.02), and team (0.02) were the most important features, such that permuting these features would to sizable increases in model prediction error (shown in the x-axis, average ROC-AUC increase = 0.03).

Next, we used PDPs to create intuitive, graphical representations of the marginal impact of these important features on predicting decision-making (see **Figure 6**). These figures demonstrate the positive impact that data and work have on predictive decision-making. These positive features are representative of the key decision-making behaviors in the AC, such as gathering data and involving others.

A logistic regression model was built as a global surrogate model to approximate the predictions of the trained XGBoost model. Specifically, we used the training data to train a logistic regression model to predict the predictions from the XGBoost model. The surrogate model showed adequate accuracy in replicating the XGBoost predictions: ROC-AUC = 0.77, precision = 0.68, recall = 0.64, *r* = 0.46. Therefore, we moved forward with interpreting the logistic regression model to understand the trained XGBoost model. As **Figure 7** demonstrates, data (2.55), team (1.83), and company (1.61) had the largest positive impact on predicting decision-making, whereas contact (-1.06) and yes (-0.91) had the largest negative impact.

Comparing results from the two global explainability methods, we noticed some overlap between the predictors (e.g., data) that can help explain the predictions, as expected. The positive features are representative of the key decision-making behaviors that are being scored in the AC, such as gathering data, interpreting information, and involving others e.g., "*I would like to meet with you and JJ before I leave today to determine what data were taken and what impact it will have on the company*". On the other hand, the negative features are prevalent in short responses that indicate a lack of decision-making behaviors and critical thinking. e.g., "*If there are any issues, continue contacting my team leaders and I am sure they will acomidate you. Thank you.*"

After establishing global interpretations of the XGBoost model, we then turned to model post-hoc local interpretability to understand how predictions on individual cases can be explained, using both SHAP and the local surrogate model. We picked two example cases: Case 844 with a relatively high predicted decision-making of 0.85, and Case 594 with a relatively low predicted decision-making of 0.17. Using SHAP, we computed how features in the trained XGBoost model contributed to the predictions of these two cases. **Figure 8** shows that, for Case 844, positive values on data (0.13), future (0.11), and plan (0.09) drove the high prediction on decision making, e.g., "*I alreadytold him not to remove data like that in the future, but it sounds like we have a very serious issue already. Obviously J.J. needs to be made aware of the gravity of this situation, but he is a key member of my team. Is it possible to gather more information regarding his situation?*" For Case 594 (**Figure 9**), the low prediction on decision making was mainly driven by negative values on features such as alex (-0.05) and company (-0.04), "*Alex, thankyou for bringing that to may attention, when I get back and on the 21st I have a Sequence* *meeting earily that morning and will discuss the issues the the whole group*". Comparing the local results with global permutation feature importance chart, we noticed some overlap between the predictors (e.g., data) that can help explain the predictions both globally and locally.

Another way to derive model local interpretability is via constructing a local surrogate model via LIME to explain individual predictions from a trained model. Take Case 844 as an example, we first created a perturbed sample by sampling the training data around Case 844. We then obtained predictions on the perturbed sample by using the trained XGBoost model. Based on the distance between each of the perturbed instances to the original case, we weighted the perturbed sample and trained a surrogate model on this weighted dataset. The resulting model is a surrogate for the XGBoost model that is locally faithful and only valid for the prediction on the individual case. Following this process, we trained a surrogate model for Case 844. **Figure 10** shows the features in the surrogate model and the features with the largest effects: data (0.10), future (0.09), and training (0.08) are among the most important features driving the explanation for the prediction in this case, "*I get the sense from her team members that she is struggling to develop in her new role, and I think this type of cross training and further interaction with the more experienced team leads would be beneficial*." Similarly, we trained another surrogate regression model for Case 594. **Figure 11** shows the features in the surrogate model and the features with the largest effects: meeting (0.01), alex (0.01), and shea (0.01) were among the most important features driving the explanation for the prediction in this case, e.g., "*Alex, thankyou for bringing that to may attention, when I get back and on the 21st I have a Sequence meeting earily that morning and will discuss the issues the the whole group.*"

DistilBERT interpretations. DistilBERT is a deep-learning model based on the Transformer architecture. Because tokens (words) are first transformed into embeddings, high dimensional vectors representing the tokens, before they are processed by the model, some explainability methods such as permutation feature importance and PDP are not optimal for understanding the BERT model family and are rarely used. Therefore, we trained a logistic regression as a global surrogate model to try to approximate the predictions of the fine-tuned DistilBERT model. To develop the surrogate model, we used the training data to train a logistic regression model to predict the predictions from the DistilBERT model. The logistic regression model showed adequate accuracy in replicating the predictions from the BERT model, ROC-AUC = 0.74, precision = 0.67, recall = 0.55, *r* = 0.41. Therefore, we moved forward with interpreting the logistic regression model to understand the DistilBERT model. As **Figure 12** demonstrates, data (1.53), work (1.50), and team (1.47) had the largest positive impact on predicting decision making, whereas thanks (-2.40), april (-1.34), and jamie (-1.26) had the largest negative impact.

After establishing global interpretations of the DistilBERT model, we then turned to model post-hoc local interpretability to understand how predictions on individual cases can be explained, using both SHAP and local surrogate model. We picked two example cases: Case 1167 with a relatively high predicted decision-making of 0.90 and Case 332 with a relatively low predicted decision-making of 0.25. Using SHAP, we computed how features in the trained DistilBERT model contributed to the predictions of these two cases. **Figure 13** shows that characters, such as *byy* (parts of a key employee's name Debby), *we*, and *need*, drove the high prediction on decision making, e.g., "*We need Debby to understand that she is valued member of the team and we want her and the team to be successful. I will schedule a meeting with Debby to discuss your concerns and get a better understanding of why absenteeism is a problem*." For Case 332 (**Figure 14**), the low prediction on decision-making was mainly driven by features such as you, thank, and great, e.g., "*Thank you I see something I can do. Thank You Jamie Pace*."

We also used LIME to derive local interpretability for Cases 1167 and 332. **Figure 15** shows the features in the surrogate model and the features with the largest effects: succeed, befitting, and early are among the most important features driving the explanation for the prediction on Case 1167, e.g., "*We need to be inclusive of all of our employees and willing to help them succeed*.", whereas thank, jamie, and BP07 (name of a meeting room) were among the most important features (**Figure 16**) driving the explanation for the prediction on Case 332, e.g., "*Great, I think room BP07 AT 12:00pm to 1:00pm is good . Thank You Jamie Pace*"

Data input–model output relationship. Data interpretability, the extent to which the input data are interpretable, is an important part of NLP interpretability. Data interpretability can be examined after models are developed to understand how varying data input affects the model output. By examining the data input–model output relationship, we can scrutinize how changes in input data influence predictions and make more informed decisions about the models.

To examine the data input–model output relationship, we first selected representative cases with high and low predicted probabilities. We then modified the original text input of each case, used the trained model to generate a new prediction on the case, and evaluated how such input modifications would change the probability prediction output. For example, as discussed earlier, Case 28 had high predicted decision-making scores from all three models, and "*data*" was a top feature driving this prediction. We manually replaced "*data*" with "*information*," a different word with similar meaning. In the logistic regression model, this substitution resulted in a decrease in predicted probability from 0.94 to 0.93. Similarly, in the XGBoost model, the probability dropped from 0.87 to 0.81. However, for the DistilBERT model, the predicted probability remained unchanged at 0.85. We suspect that the predicted decision-making scores dropped in logistic regression and XGBoost are because these two NLP models are based on bag-of-words, where replacing a highly weighted word such as "*data*" is essentially removing a feature from the model and directly impacts model prediction. However, pre-trained language models such as DistilBERT have garnered a better understanding of the language and the semantic relationships between different words, so swamping similar words like "*data*" and "*information*" would not impact model prediction significantly. This example illustrates that BERT-type models tend to be more robust than traditional bag-of-words-based models for grasping the underlying meaning of the language.

To test whether the DistilBERT model's prediction would be affected by a feature replacement with a different meaning, we substituted the name "*Debby*" with "*Dave*" in Case 28. As expected, this led to a sizable drop in the predicted probability, decreasing from 0.85 to 0.72. This finding demonstrates that BERT's performance can still be impacted when the substituted feature has a distinct semantic difference. Overall, these results can enhance our understanding of how different models interpret input features and highlight the varying degrees of sensitivity in their predictions when input was modified.

GPT o1. We included GPT o1 to show a simple illustration of how a prompt-based technique can be used to derive explanations from a chat-based LLM. To do so, we included detailed instructions in the prompt about how the model should predict decision-making, including setting up the role of the model as an expert AC assessor, introducing the AC design, detailing the six decision-making dimensions, and breaking down the prediction task step by step (see **Table 4**). This helps GPT o1 organize its thinking and rationale in a way that is theoretically sound. To generate model prediction explanations, we also included a sentence in the prompt to ask GPT o1 to share the rationale for the classification decision. We selected Case 300 with a high predicted probability of 0.90 and Case 640 with a low predicted probability of 0.25 by DistilBERT. For Case 300, GPT o1 classified the case as having effective decision-making. The model produced 17 reasoning steps, including evaluating response clarity, assessing responses, assessing core behaviors, and so on. In its rationale, it also highlights key behaviors within each decision-making dimension that the case exhibited, e.g., "*Seeks input from team members and other departments to find solutions collaboratively*." for the dimension of "*gathering information*". For Case 640, the rationale included behaviors that are counterproductive to decision-making, e.g., "*Abruptly decides to find a replacement for Pat Landis because Cory is buying him lunch, labeling it as 'unacceptable' without providing a clear rationale or following proper procedures*" for the dimension of "*choosing appropriate actions*."

Discussion

NLP interpretability is important for scientific advancement, bias detection and mitigation, user trust, and ethical and legal compliance. To address the need to better understand and explain NLP models in an organizational research context, this paper takes an initial step to illustrate what NLP interpretability is and how organizational researchers can build interpretable NLP modes.

Our discussion focuses on two areas. First, we discuss the implications of NLP interpretability for organizational research and offer some practical considerations. Second, we discuss the current study's limitations and future research directions.

Implications for Organizational Research and Practice

One important takeaway from this article is that NLP interpretability should be understood as a continuum based on different sources of evidence derived from the different stages of the NLP model development process, including pre-modeling (data properties and exploratory data analysis), modeling (model intrinsic interpretability), and post-modeling (global and local explanations). In other words, interpretable NLP should not solely rely on deriving post-hoc explanations after a model has already been developed, but instead requires a systematic, theory- and data-driven process throughout the pre-modeling, modeling, and post-modeling stages. Therefore, researchers and practitioners who are interested in using NLP to answer organizational questions should take into consideration interpretability issues during the study planning process, along with other factors such as the research question, data, high vs low-stakes nature of the study, and so on. For instance, if the main goal of the study is to maximize prediction using different NLP models in a research context, it is probably appropriate to use models with lower intrinsic interpretability and try to offer post-modeling explanations. In contrast, in high-stakes decision-making situations such as personnel selection, one might weigh intrinsic interpretability over prediction optimization and choose an NLP model that can be easily explained and understood by the stakeholders.

In this paper, we illustrated different methods to provide model post-hoc interpretability evidence, including global (e.g., permutation feature importance, PDP, and global surrogate model) and local (e.g., SHAP, local surrogate model). The former focuses on understanding the general patterns of the model behavior as a whole, whereas the latter focuses on understanding how a specific individual prediction is derived. While global and local interpretability methods serve distinct purposes, they often converge meaningfully. Global methods, such as permutation feature importance and PDPs, provide an overarching view of how different features influence the model’s predictions as a whole. Conversely, local methods, like SHAP values and LIME, offer detailed insights into how individual predictions are derived. Despite these differences in focus, both types of interpretability methods can highlight consistent patterns in the model's decision-making process. For example, features identified as most important by global methods frequently emerge as key drivers in local interpretations. If a feature like "*data*" is of high importance globally, it is often found to influence specific predictions at the local level, as we have demonstrated through our examples above. This convergence is reassuring because it suggests that the model's decision-making process is coherent and aligns across different levels of analysis.

Additionally, the agreement between global and local methods can serve as a validation mechanism, increasing confidence in the model's interpretability and its underlying logic. By leveraging both global and local interpretability, researchers and practitioners gain a more comprehensive understanding of the model. Global methods help identify broad patterns, while local methods allow for case-by-case analysis, illustrating how the model applies these patterns in practice. Together, these methods create a multifaceted view of the model's behavior, enhancing its transparency and usability for various stakeholders in organizational research.

That said, we do not think it is necessary to include every single interpretability method in a particular study. Rather, researchers and practitioners can pick and choose specific methods based on the nature of their research. For instance, if the research focuses on building and advancing scientific theory, then the global interpretability of the model is a priority. In contrast, in a practical prediction scenario such as predicting job performance, one might want to leverage local interpretations to understand what is driving extremely high vs. low performers, as these individuals might make the biggest practical difference in return on investment.

It should not be forgotten that, besides using these interpretability methods, we, as the subject matter experts of organizational phenomena, can enhance the interpretability of NLP models by bringing in theory and job relatedness when designing and interpreting research studies or organizational applications. For instance, when developing personnel assessments using NLP, organizational researchers should be guided by theory and job-relatedness (Tippins et al., 2021), such as considering the theoretical basis of the assessment stimuli and the elicited predictors, as well as how they relate to the knowledge, skills, abilities, and other characteristics (KSAOs) required for the job. In addition, when interpreting results from NLP models, rather than stopping at merely identifying the variables that contribute the most to model predictions, we can offer theory-based insights as to what theoretical mechanisms are driving such variables to impact the criterion.

Study Limitations and Future Research Directions

In this study, we took an initial step toward illustrating interpretable NLP in organizational research. To that end, we focused on supervised NLP models as we believe this is one of the most commonly adopted NLP methodologies in organization research so far. However, other types of techniques, such as unsupervised learning (e.g., Tonidandel et al., 2022) or reinforcement learning (Kumwilaisak et al., 2022), also apply to organizational research. Therefore, future research can extend the current study by providing guidance and illustrating methods for interpreting unsupervised and reinforcement learning models.

Second, the field of NLP is advancing rapidly, with newer, more complex models emerging constantly. In the current study, we sampled a few NLP models on the spectrum of model intrinsic interpretability from high (logistic regression) to low (GPT o1) to illustrate how interpretability evidence can be established. While we cannot exhaust the different types of NLP models, future studies can extend our work by illustrating how interpretability can be established for other types of NLP models, such as large multimodal models (LMMs) like GPT-4 Vision

(OpenAI, 2024a) and sparse models like Switch Transformers (Fedus et al., 2022). These models present unique interpretability challenges due to their ability to process diverse data types (e.g., text, images) and their dynamic routing of information through vast parameter spaces.

Furthermore, the increasing complexity of LLMs has led to growing concerns about their application in organizational research. For example, recent studies have highlighted potential limitations and biases in LLMs that require careful consideration. Wang, Xiao et al. (2024) found that LLMs can exhibit cognitive biases, such as the representativeness heuristic, suggesting that their decision-making processes do not always align with human reasoning. Similarly, Wang et al. (2024) emphasize the need to understand rater effects when interpreting ratings derived from these models. Additionally, the debate continues on whether LLMs can fully replace human respondents in psychometric research, as Wang, Zou et al. (2024) argue that current LLMs still fall short in capturing the nuances of human responses. These findings underscore the importance of interpretability and the need for future research to address these challenges when employing LLMs in organizational contexts.

Future research should focus on developing more robust frameworks for AI applications in organizational research. Min et al. (2024) proposed a checklist to ensure transparency and trust in supervised machine learning studies, which could serve as a guideline for extending the interpretability of advanced NLP models. Additionally, exploring new methods, such as pseudo-factor analysis of language embedding similarity matrices (Guenole et al., 2024), could provide innovative ways to model latent constructs and further the interpretability of LLMs in the context of organizational assessments.

References

Aguinis, H., & Edwards, J. R. (2014). *Methodological Wishes for the Next Decade and How to Make Wishes Come True. Journal of Management Studies, 51*(1), 143–174. https://doi.org/10.1111/joms.12058

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of Advances in Neural Information Processing Systems 29 (NIPS2016)*. arXiv:1607.06520

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology*, *76*(4), 993-1009.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958-975.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458-474.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems 30*. arXiv:1706.03741

Chu, C., Sun, T., Zhang, B., & Rounds, J. (2024). Assessing vocational interests through chat: Development and validation of the career guidance chatbot (CGC-bot). https://doi.org/10.31234/osf.io/upx5q

Colquitt, J. A., Greenberg, J., & Zapata-Phelan, C. P. (2013). What is organizational justice? A historical overview. In *Handbook of Organizational Justice* (pp. 3-56). Psychology Press.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, (pp. 4171–4186). arXiv preprint arXiv:1810.04805*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

The European Parliament and the Council of the European Union (2016). The European Union General Data Protection Regulation.

Fan, J., Sun, T., Liu, J., Zhao, T., Glorioso, M., Chen, Z., Zhang, B., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology, 108*(8), 1277-1299. https://doi.org/10.1037/apl0001082

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1-39. arXiv:2101.03961

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, *16*(2), 399-432.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe, 7*, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Goodman & Faxman (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." https://www.aaai.org/ojs/index.php/aimagazine/article/view/2741

Guenole, N., D'Urso, E., Samo, A., & Sun, T. (2024). Pseudo Factor Analysis of Language Embedding Similarity Matrices: New Ways to Model Latent Constructs. https://doi.org/10.31234/osf.io/vf3se

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2024). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal, 34*(1), 39-54. https://doi.org/10.1111/1748-8583.12426

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis*. Sage.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.

Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. doi: 10.1037/met0000560

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323-1351. doi: 10.1037/apl0000695

Kleinbaum, D. G., & Klein, M. (2002). Logistic regresion for correlated data: GEE. *Logistic regression: A self-learning text*, 327-375.

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J. F., Lowery, S. B., Kitching, M., Nimbkar, A., Boyce, A. S., Sun, T., Guo, F., Min, H., Zhang, B., Lebanoff, L., & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, 76(4), 1061-1123. https://doi.org/10.1111/peps.12608

Kumwilaisak, W., Phikulngoen, S., Piriyataravet, J., Thatphithakkul, N., & Hansakunbuntheung, C. (2022). Adaptive call center workforce management with deep neural network and reinforcement learning. *IEEE Access*, *10*, 35712-35724.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475-492.

Lipton, Z. C. (2017). The doctor just won't accept that! *The 31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA. *arXiv preprint arXiv:1711.08037*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Manning, C. D. (1999). *Foundations of statistical natural language processing*. The MIT Press.

Min, H., Guo, F., Sun, T., Liu, M., & Oswald, F. (2024). Ensuring Transparency and Trust in Supervised Machine Learning Studies: A Checklist for Organizational Researchers. https://doi.org/10.31234/osf.io/vukxp

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Education.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/

OpenAI. (2024a). *GPT-4*. OpenAI. https://openai.com/index/gpt-4

OpenAI. (2024b). *GPT-o1*. OpenAI. https://openai.com/o1/

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, *38*(2), 385-410.

Rahimi, A., Benini, L., & Gupta, R. K. (2017). *From Variability Tolerance to Approximate Computing in Parallel Integrated Architectures and Accelerators*. Springer International Publishing.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should I trust you?: Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Sanh, V. (2019). DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Shapley, H. (1951). Proxima Centauri as a flare star. *Proceedings of the National Academy of Sciences*, *37*(1), 15-18.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, *39*(10), 1095-1100.

The Society for Industrial and Organizational Psychology. (2018). *Principles for the Validation and Use of Personnel Selection Procedures* in *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*(Supl 1), 2-97. https://doi.org/10.1017/iop.2018.195

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299-333.

Spirin, N., & Han, J. (2012). Survey on web spam detection: principles and algorithms. *ACM SIGKDD explorations newsletter*, *13*(2), 50-64.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15*(1), 1929-1958.

Sun, T., Roberts, B., Drasgow, F., & Zhou, M. X. (2024). Development and validation of an artificial intelligence chatbot to assess personality. https://doi.org/10.31234/osf.io/ahtr9

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. *Personnel Assessment and Decisions*, *7*(2), 1.

Tonidandel, S., Summerville, K. M., Gentry, W. A., & Young, S. F. (2022). Using structural topic modeling to gain insight into challenges faced by leaders. *The Leadership Quarterly*, *33*(5), 101576.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-wesley.

Vaswani, A. (2017). Attention is all you need. *in Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wachter, S., Mittelstadt, B., & Floridi, L. (2016). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

Wang, P., Myeong, H., & Oswald, F. L. (2024). On putting the horse (raters and criteria) before the cart (variance components in ratings). *Industrial and Organizational Psychology*, 1–5. https://doi.org/10.1017/iop.2024.16

Wang, P., Xiao, Z., Chen, H., & Oswald, F. L. (2024). Will the real Linda please stand up… To large language models? Examining the representativeness heuristic in LLMs. *The Conference on Language Modeling (COLM 2024)*. https://doi.org/10.48550/arxiv.2404.01461

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z., & Zhang, B. (2024). Not Yet: Large Language Models Cannot Replace Human Respondents for Psychometric Research. OSF. https://doi.org/10.31219/osf.io/rwy9b

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Woo, S. E., Tay, L., & Oswald, F. (2024). Artificial intelligence, machine learning, and big data: Improvements to the science of people at work and applications to practice. *Personnel Psychology*. https://doi.org/10.1111/peps.12643

Yuan, L. I., Sun, T., Dennis, A. R., & Zhou, M. X. (2024). Perception is reality? Understanding user perceptions of chatbot-inferred versus self-reported personality traits. C*omputers in Human Behavior: Artificial Humans*. https://doi.org/10.1016/j.chbah.2024.100057

Tables and Figures

Table 1.

A "Cheat Sheet" of AI-related Terminologies

	Definition	**Also Known As…**
Artificial Intelligence	The simulation of human intelligence processes by machines, particularly computer systems.	AI
Machine Learning	A subfield of computer science that aims to construct computer programs that can learn and improve with experience automatically.	ML
Supervised Learning	A type of machine learning where the model is trained on labeled data, meaning the output is known during training.	
Classification	A supervised learning task that involves predicting a categorical label for new data based on training data.	
Regression	A supervised learning task that involves predicting a continuous numerical value based on input features.	
Model	A mathematical representation of a process that can make predictions based on input data.	
Algorithm	A set of rules or steps used to perform calculations, process data, or make decisions in machine learning.	
Feature	An individual measurable property or characteristic used as input in a model.	Independent variable
Target/Label/Ground Truth	The output variable that the model is trying to predict, derived from the features. Each label corresponds to a specific instance in the training dataset and serves as the known correct answer during the learning process.	Dependent variable Outcome variable
Sparsity	A condition where most of the elements in a dataset or model are zero or missing, often leading to efficiency in storage.	
Model fit	A measure of how well a model's predictions align with actual observed data.	
Training	The process of using a dataset to teach a model to make predictions or decisions.	
Testing	The evaluation of a trained model's performance on unseen data to assess its accuracy and generalization ability.	
Cross-validation	A technique for assessing how the results of a statistical analysis will generalize to an independent dataset by partitioning the data into subsets.	

Table 2.

Descriptions of Common NLP Models

	Description	**Classification or Regression**	**Sample Use Case**
Linear Regression	Models a linear relationship between a continuous dependent variable and one or more independent variable(s).	Regression	Personality prediction
Logistic Regression	Models the probability of a binary dependent variable using a logistic function.	Classification	Customer churn prediction
K-Nearest Neighbors (K-NN)	A non-parametric pattern recognition method that makes predictions based on k closest data points.	Both	Product recommendation system
Naïve Bayes	Models the probability of a binary dependent variable using Bayes' theorem with strong (naïve) independence assumptions between the features.	Classification	Email spam detection
Decision Trees	A tree-like model that explicitly represents decision points going from observations to target values.	Both	Employee turnover prediction
Random Forest	An ensemble method that combines a multitude of decision trees.	Both	Sentiment analysis
Support Vector Machine (SVM)	A supervised machine learning algorithm primarily used for classification tasks. SVM finds the optimal hyperplane that best separates data points of different classes in a high-dimensional space.	Both	Job performance prediction from narratives
Gradient Boosted Trees	A machine learning technique that builds a predictive model by combining the outputs of several weak learners, typically decision trees, to create a strong learner.	Both	Credit scoring
Artificial Neural Networks	Consists of connected units or nodes called artificial neurons. ANN Receives signals from connected neurons, processes them, and sends a signal to other connected neurons.	Both	Speech recognition

Table 3.

Sources of Interpretability Evidence and Common Use Cases

Sources of Interpretability Evidence	**Data Interpretability**	**Model Intrinsic Interpretability**	**Model Post-Hoc Global Interpretability**	**Model Post-Hoc Local Interpretability**
Description	The extent to which the input data (variables) are interpretable.	The extent to which the machine learning model itself is interpretable based on its structure.	The extent to which predictions from machine learning can be interpreted as a whole.	The extent to which individual predictions from the machine learning model can be interpreted.
Stage	Pre-modeling	Modeling	Post-modeling	Post-modeling
Methods	Descriptive statistics; data visualization	Linear models	Permutation feature importance, partial dependence plots, feature importance	LIME, SHAP, local surrogate models
Importance in Scientific Understanding	High	Medium	High	Low
Importance in Detecting and Mitigating Biases	High	Medium	High	High
Importance in User Trust and Adoption	Medium	Medium	Medium	High
Importance in Regulatory Compliance	High	Medium	High	Medium

Table 4.

GPT o1 Prompt

Define the role	You are an expert assessment center assessor. You are evaluating responses to assessment center simulations to determine a job candidate's level of decision-making skills. The assessment center consists of various in-basket exercises simulating on-the-job scenarios. These simulations aim to elicit behaviors associated with decision-making. You goal is to classify whether or not the candidate's responses demonstrate effective or ineffective decision-making.
Define the task	You will receive the job candidate's responses to the simulations enclosed in <responses></responses > XML tags. You will not receive the exercises; they are the same for every job candidate.
Specify the task details	To do this task, you should: 1. Review the candidate's responses. Extract the six types of behaviors that demonstrate effective decision-making, including chooses appropriate actions, commits to action, gathers information, identifies issues and opportunities, interprets information, and involves others. 2. Classify whether the candidate's responses as a whole demonstrate effective (1) or ineffective (0) decision-making. Effective decision-making means the six decision-making behaviors were displayed or excellently displayed. Ineffective decision-making means the six decision-making behaviors were not adequately displayed or the displayed behaviors were counterproductive. 3. Output the classification score 0 or 1. Share your rationale.
Specify the format details	<responses> {CANDIDATE_RESPONSES} </responses>

Table 5.

NLP Model Prediction Results

Model	ROC-AUC	Precision	Recall	r
Logistic Regression	0.78	0.68	0.61	0.48
XGBoost	0.72	0.69	0.59	0.38
BERT	0.71	0.64	0.60	0.34

Figure 1.

Common NLP Models by Intrinsic Interpretability

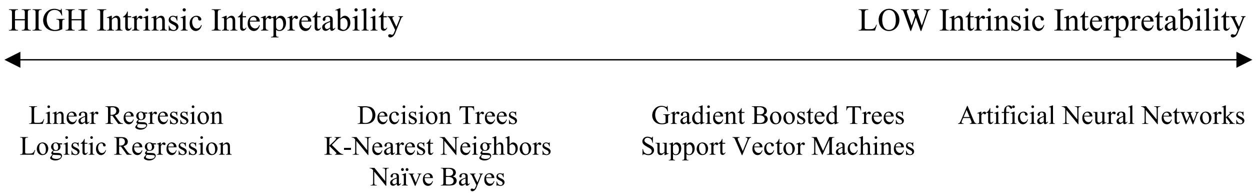


Figure 2.

Feature Importance for Logistic Regression

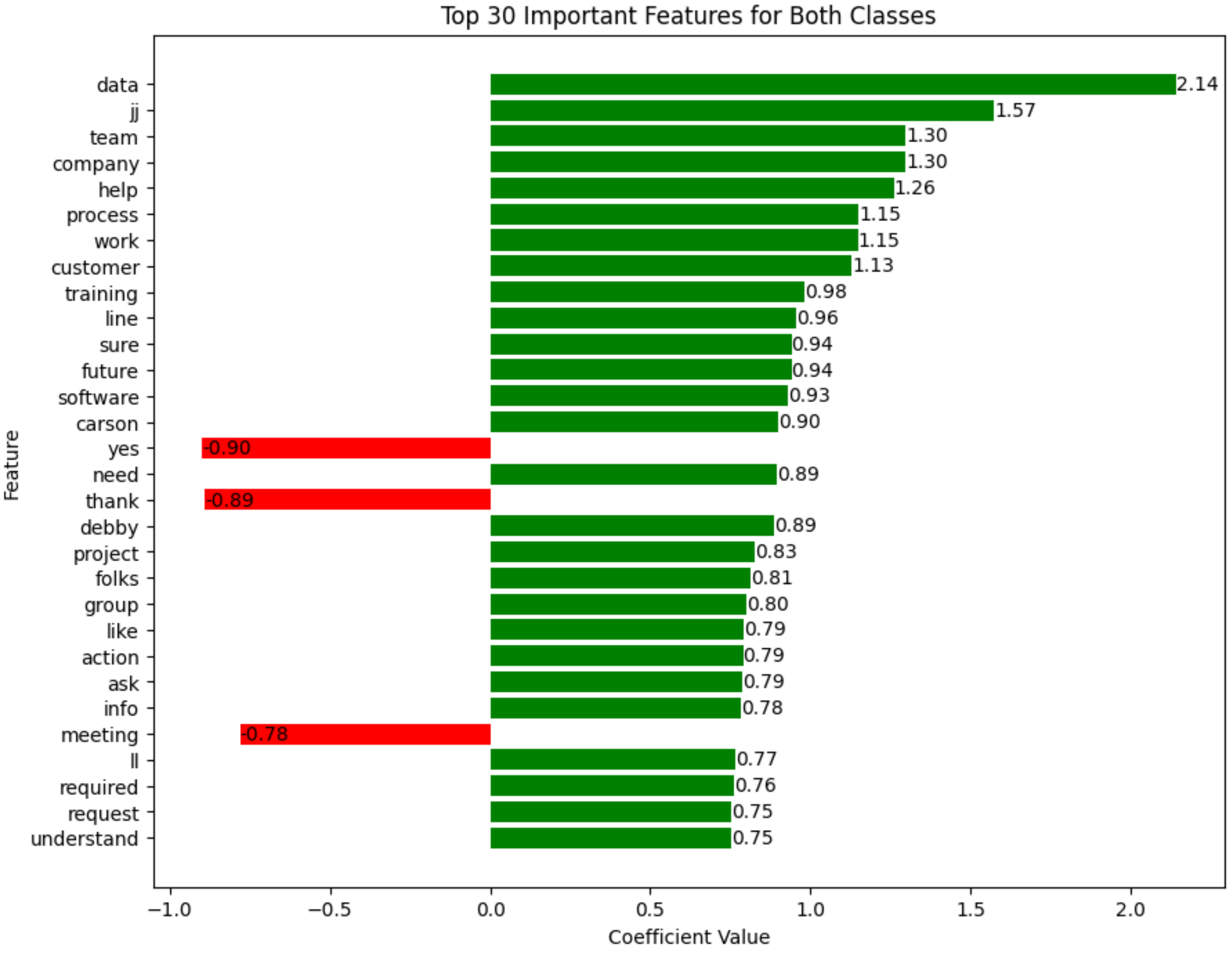


Figure 3.

Key Features Driving High Prediction in Case 28

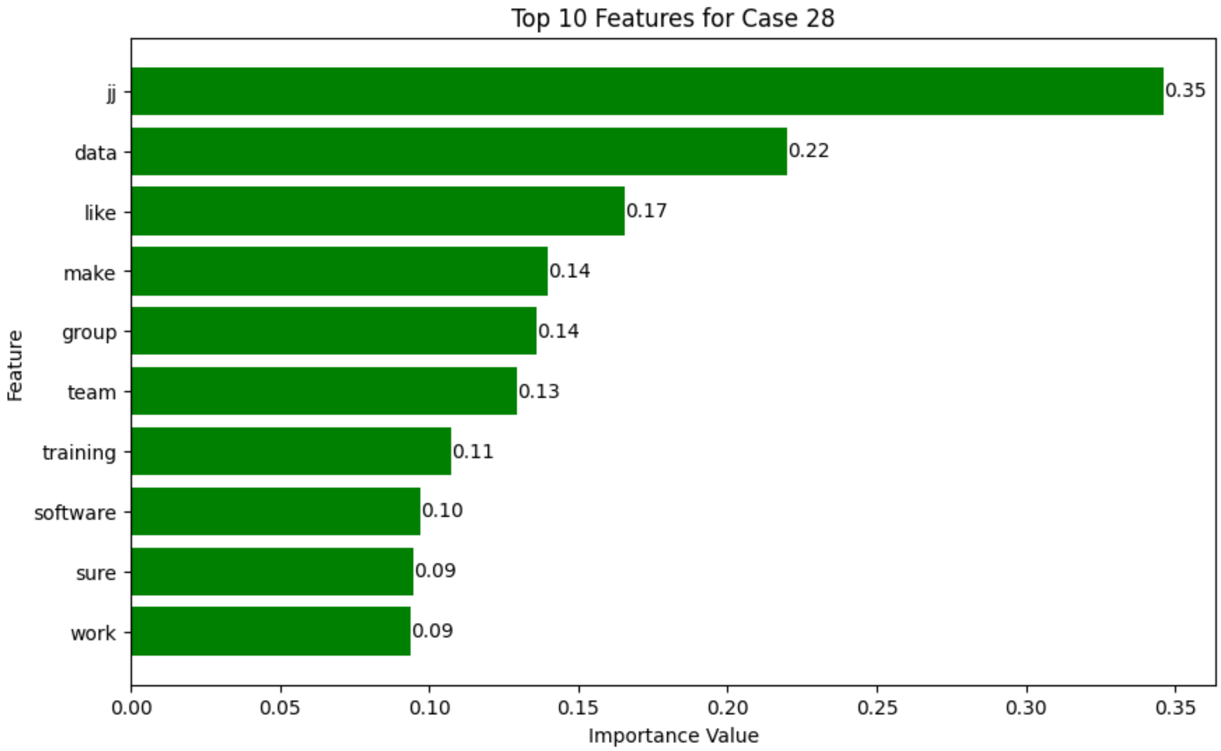


Figure 4.

Key Features Driving Low Prediction in Case 741

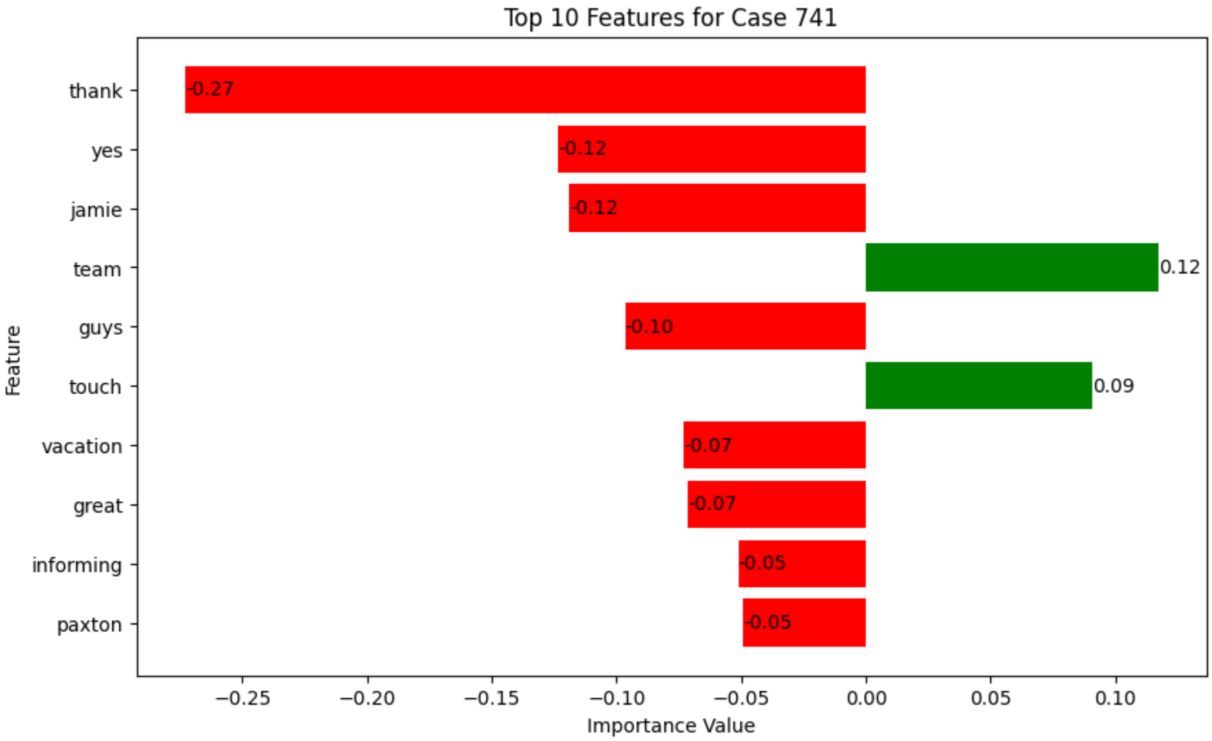


Figure 5.

Permutation Feature Importance for XGBoost Model

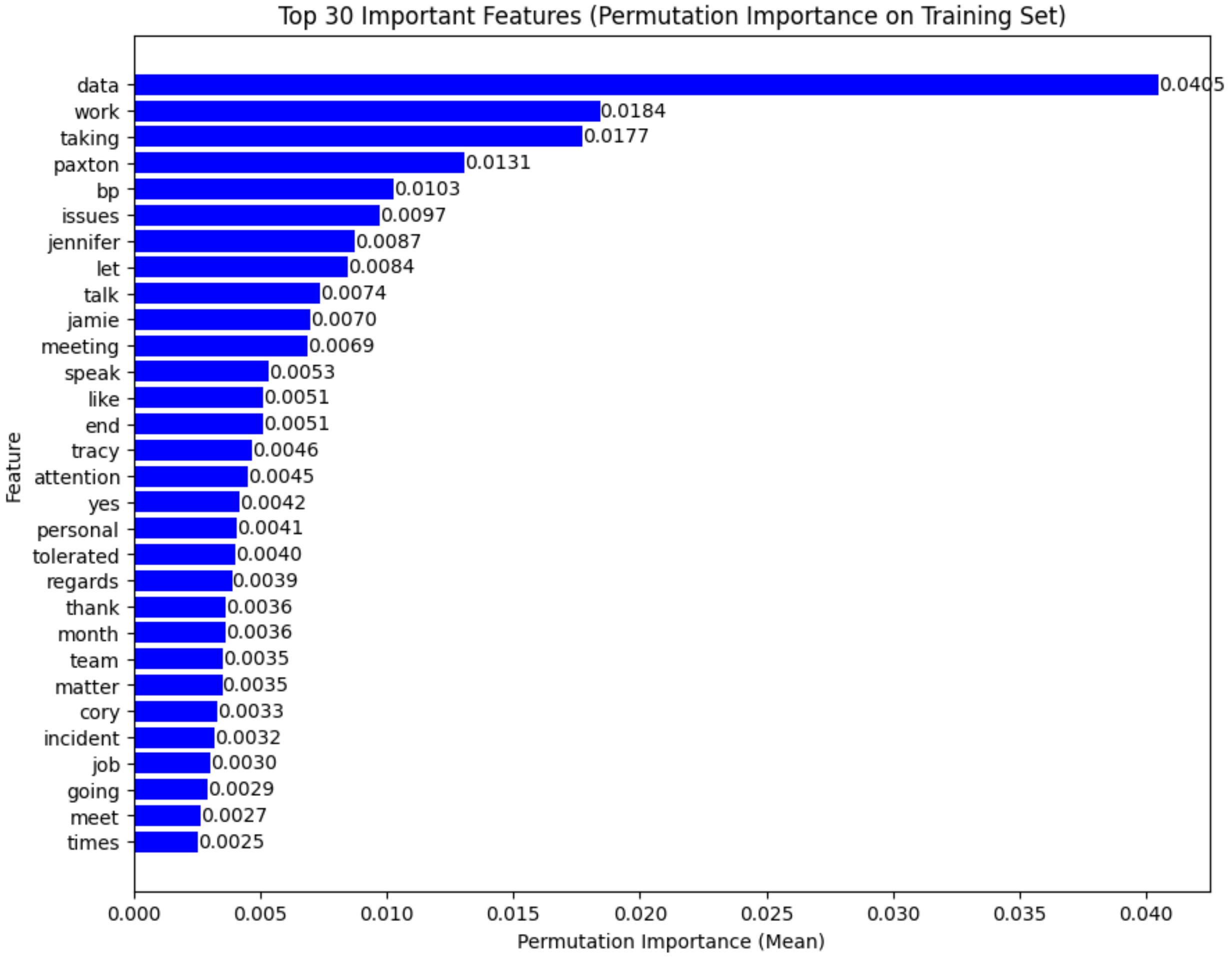


Figure 6.

Partial Dependence Plot for Key Predictive Features in XGBoost Model

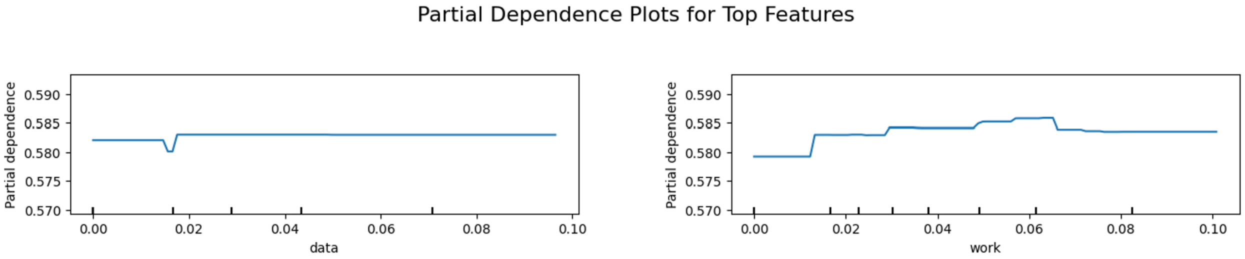


Figure 7.

Feature Impact in Logistic Regression Surrogate Model for XGBoost Predictions

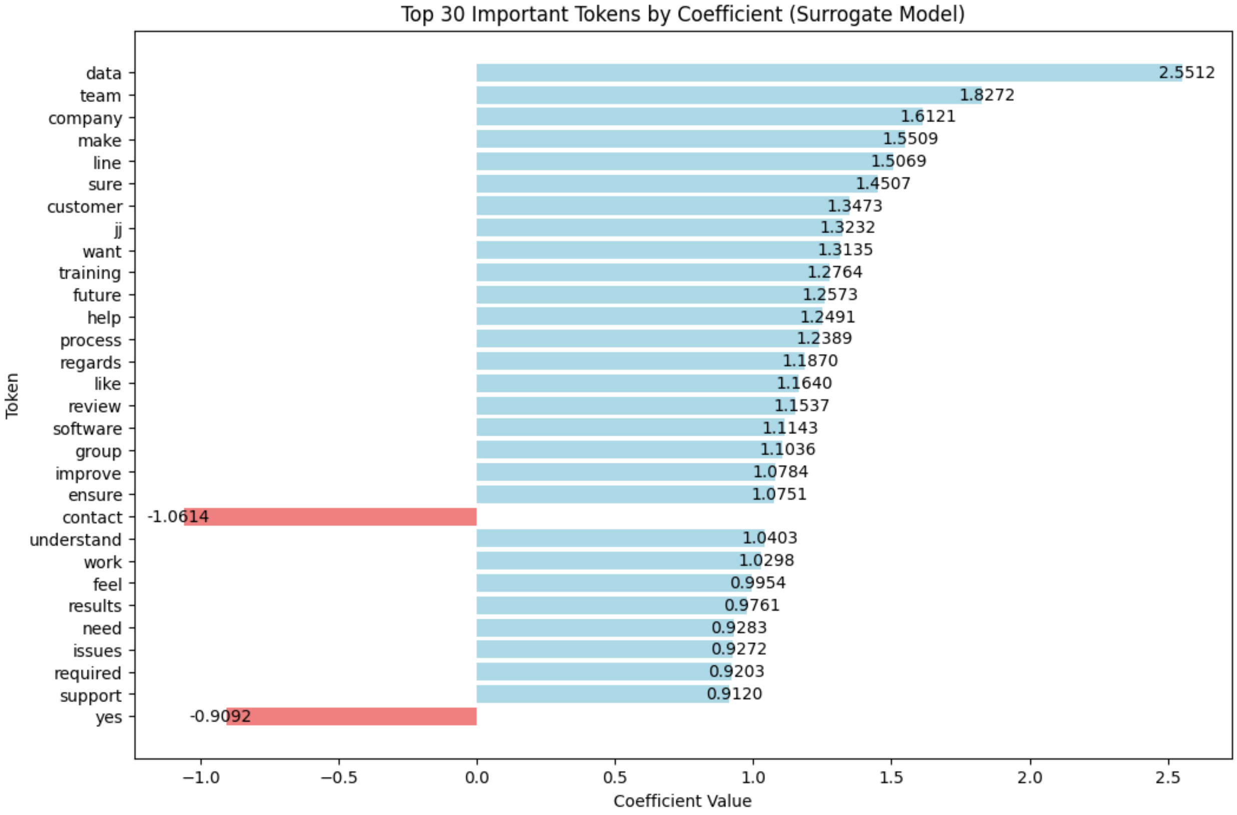


Figure 8.

SHAP Feature Contributions for High Prediction in Case 844 for XGBoost

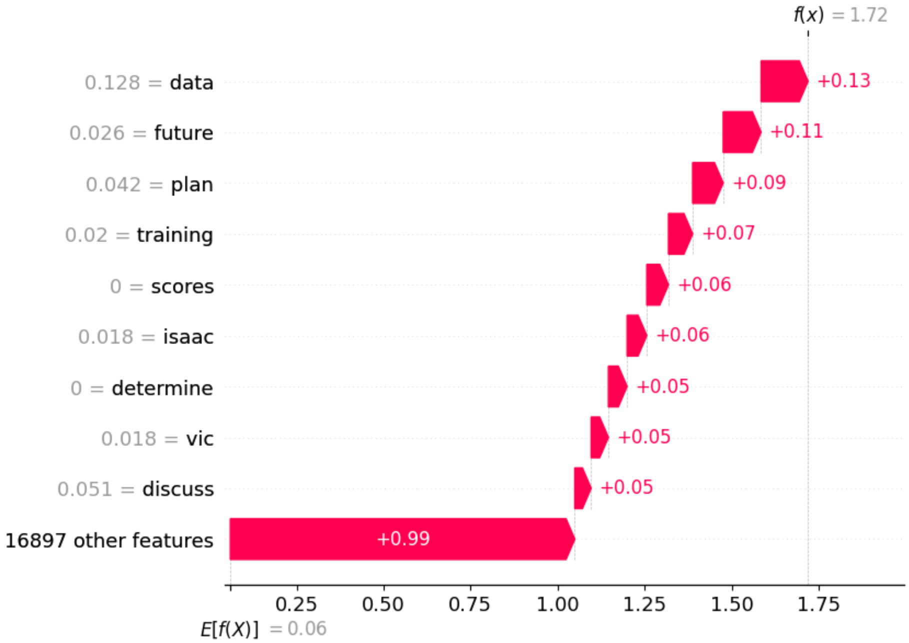


Figure 9.

SHAP Feature Contributions for Low Prediction in Case 594 for XGBoost

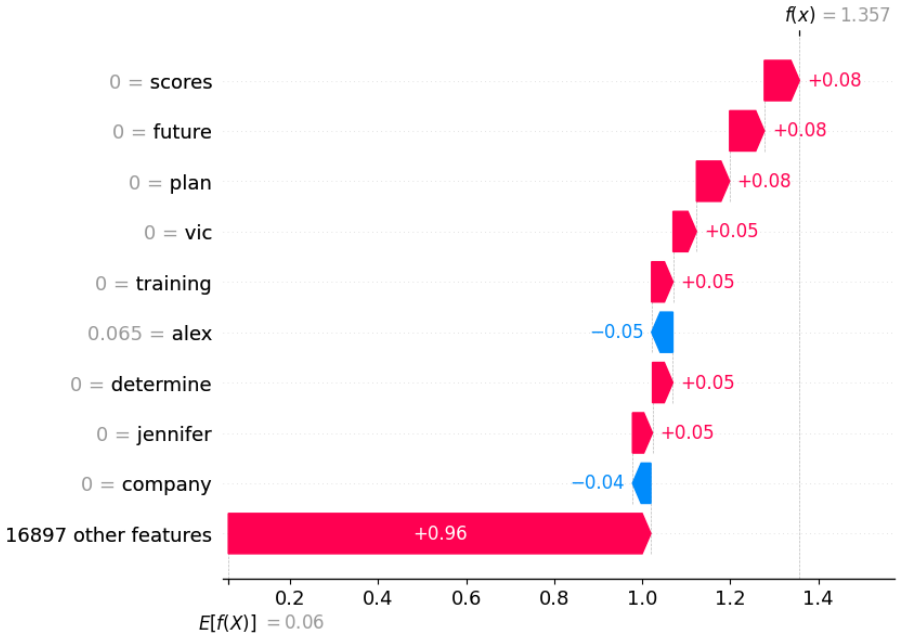


Figure 10.

LIME Surrogate Model Explanation for High Prediction in Case 844 for XGBoost

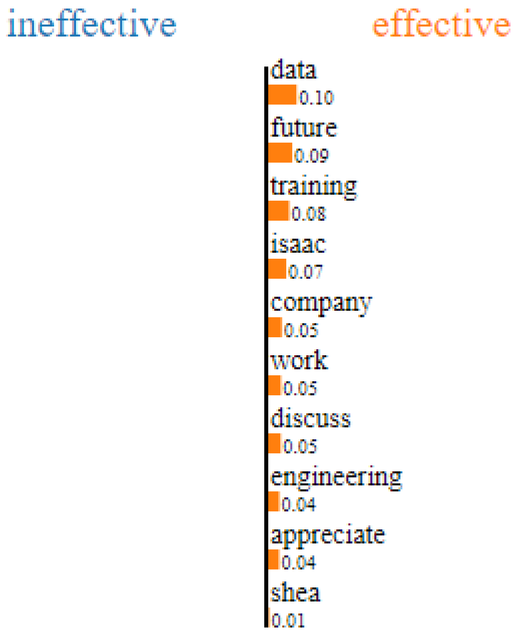


Figure 11.

LIME Surrogate Model Explanation for Low Prediction in Case 594 for XGBoost

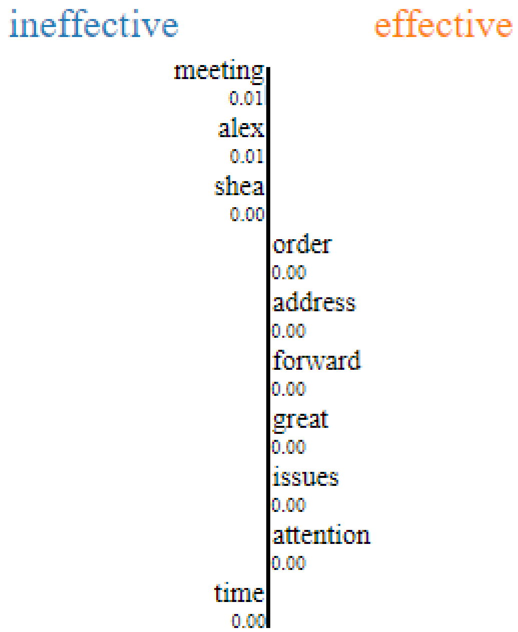


Figure 12.

Feature Impact in Logistic Regression Surrogate Model for DistilBERT Predictions

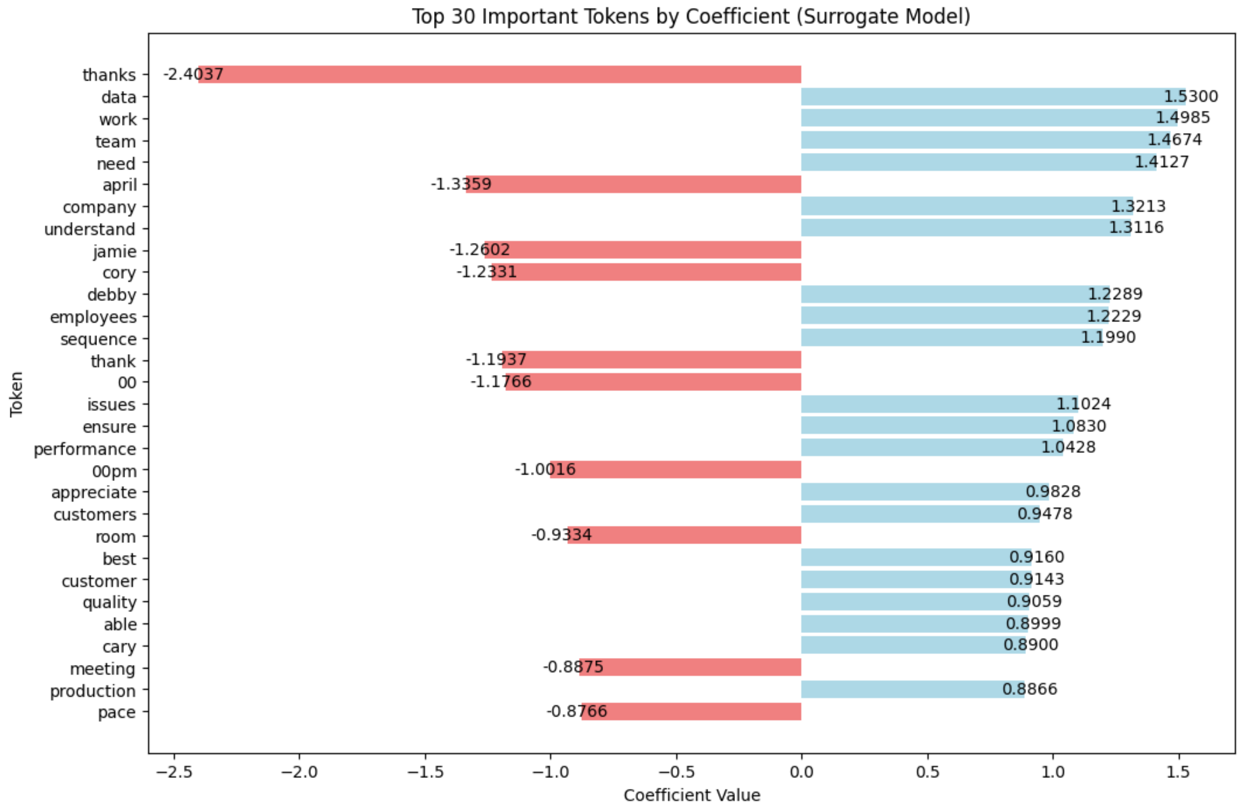


Figure 13.

SHAP Feature Contributions for High Prediction in Case 1167 for DistilBERT

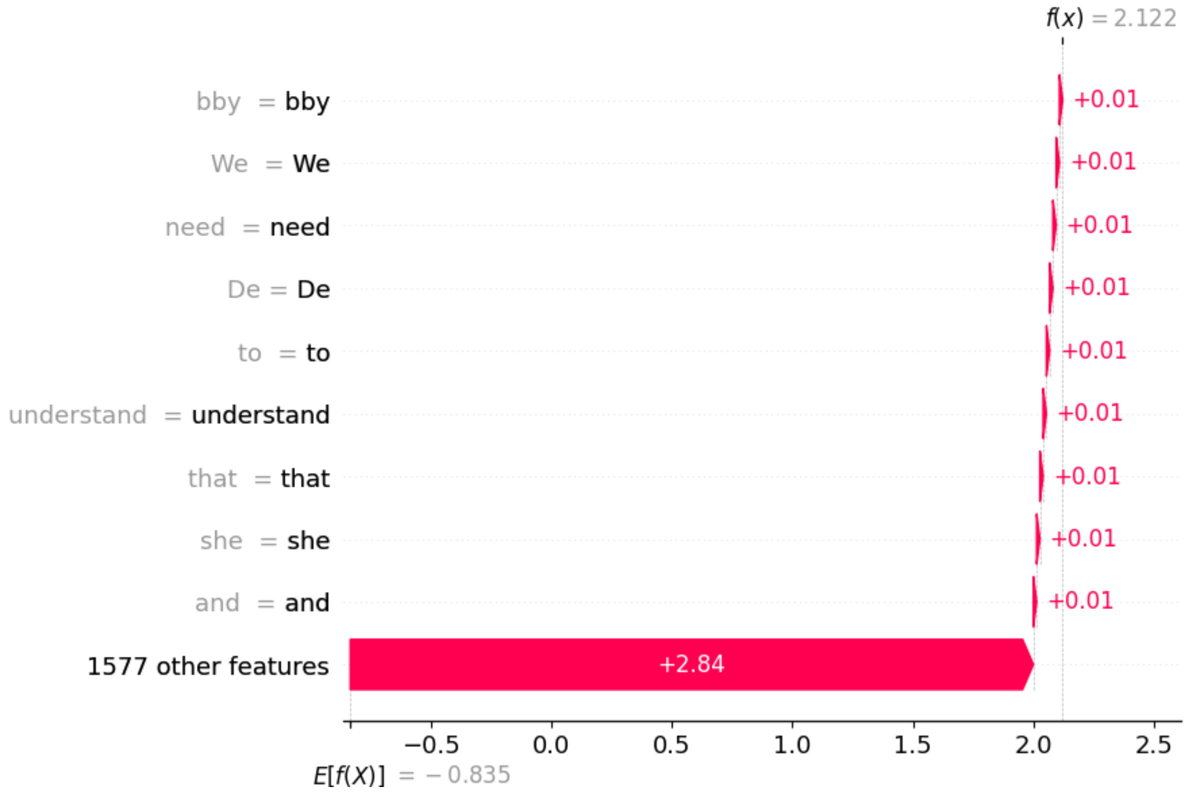


Figure 14.

SHAP Feature Contributions for Low Prediction in Case 332 for DistilBERT

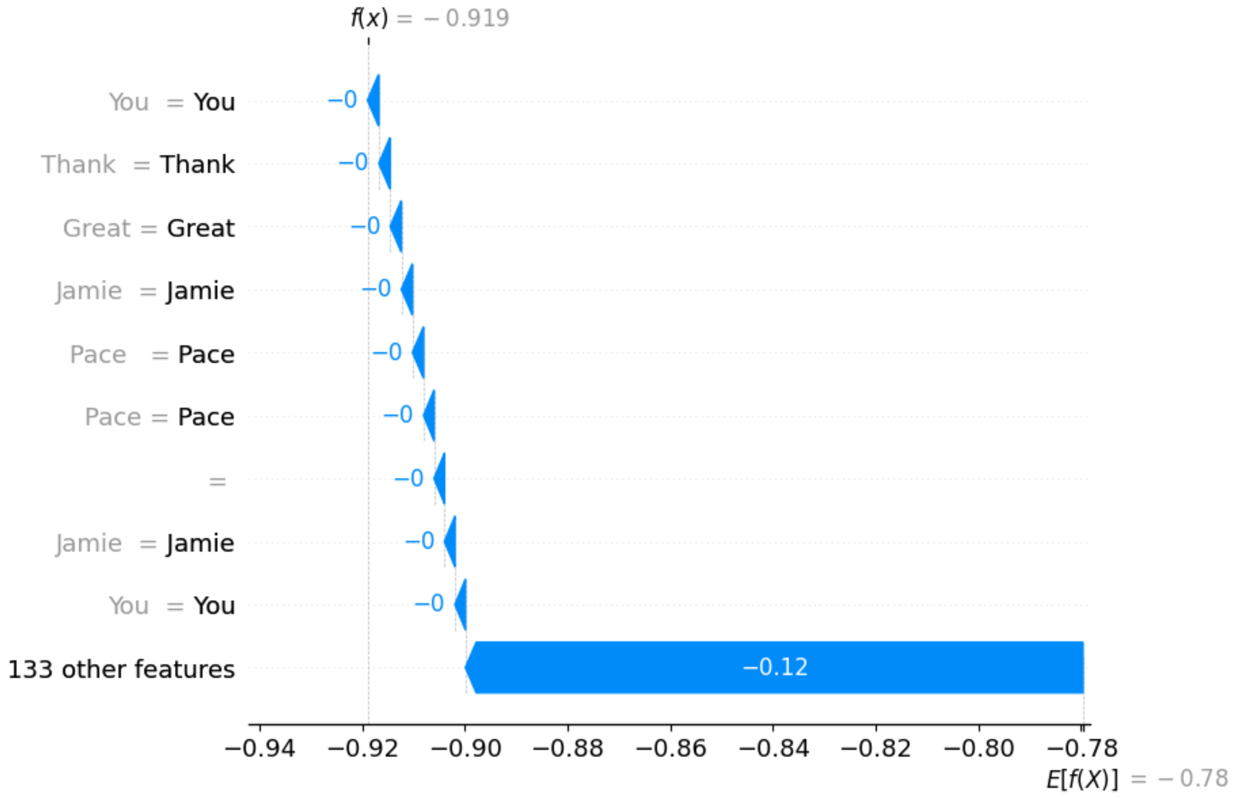


Figure 15

LIME Surrogate Model Explanation for High Prediction in Case 1167 for DistilBERT

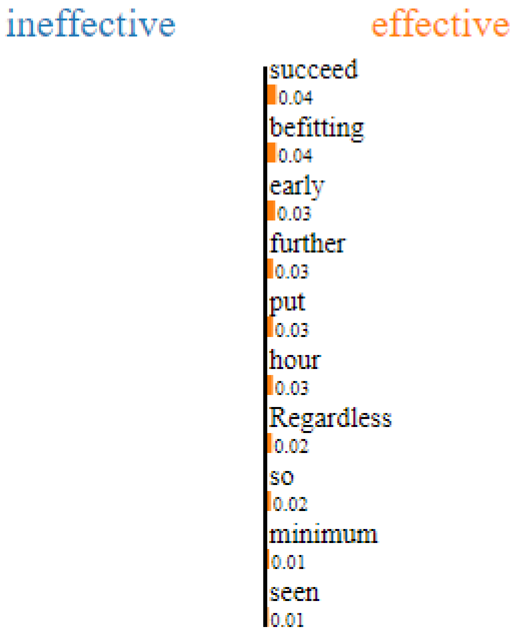
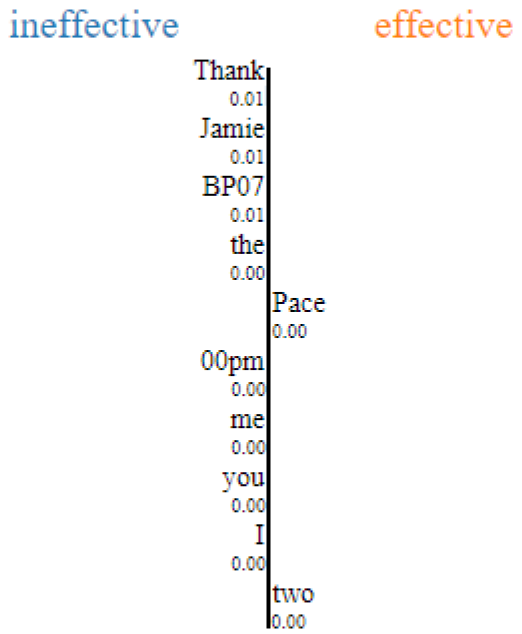


Figure 16.

LIME Surrogate Model Explanation for Low Prediction in Case 332 for DistilBERT



Appendix: Calculation Details

Permutation feature importance

Permutation feature importance on a given model (f) can be derived via the following steps:

1. The original model error (e_{orig}) is defined as the expected loss of a prediction model f with the original feature matrix X^{orig}

$$e^{orig}(f) \coloneqq \mathbb{E}L(f, (Y^{orig}, X^{orig})) \tag{1}$$

2. For each feature X_j , j ∈ $\{1, \dots, k\}$,

   a. The permutated model error (e^{perm}) is defined as the loss of model f with permutated X_j^{perm} replacing X_j^{orig} (resulting in the permutated feature matrix X^{perm})

$$e^{perm}(f) \coloneqq \mathbb{E}L(f, (Y^{orig}, X^{perm})) \tag{2}$$

   b. The importance of X_j on model f is defined as the difference between the permutated model error and the original model error (see Gregorutti, Michel, & Saint-Pierre, 2017)

$$I(X_1) \coloneqq e^{perm}(f) - e^{orig}(f) \tag{3}$$

   Higher $I(X_j)$ value indicates greater reliance of f on X_j (i.e., greater feature importance of X_j).

   c. Repeat steps a-b *n* times to compute an average feature importance of X_j ($\overline{I(X_j)}$):

$$\overline{I(X_j)} = \frac{1}{n}\sum_{i=1}^{n} I(X_j)_i \tag{4}$$

3. After obtaining permutation feature importance on all *k* features $\{\overline{I(X_1)}, \dots, \overline{I(X_k)}\}$, one can compare and determine relative feature importance in a machine learning model.

Partial dependence plot

Let *S* be a subset of *k* features, $S \in \{X_1, \dots, X_k\}$, and *C* be a complement to *S*, $S \cup C = \{X_1, \dots, X_k\}$. The partial dependence of the *S* features on model f (trained based on all *k* features) is defined as

$$f_{\mathrm{S}}(X_S) = \mathbb{E}_{X_C}(f(X_S, X_C)) \tag{5}$$

Given that there are multiple observations in the data on X_C, the partial dependence of a given variable needs to be computed via the following steps:

1. For each feature X_{Ci},

$$\widehat{f_{\mathrm{S}}(X_S)} = f(X_S, X_{Ci}) \tag{6}$$

2. Average the estimated partial dependence for all *k* features of X_C

$$\overline{f_{\mathrm{S}}(X_S)} = \frac{1}{n}\sum_{i=1}^{n}[f(X_S, X_{Ci})] \tag{7}$$

Global surrogate models

A high-level process of creating and evaluating a global surrogate model from an original machine learning model is outlined below:

1. Select a highly intrinsically interpretable model to be the global surrogate model (g^{sur}).
2. Select X to be the input feature set for developing the global surrogate model. X can be the entire training set, a subset of the training set, or a new dataset of the same distribution, depending on your use case.

3. Apply the trained original model (f^{ori}) to X to drive predictions $\hat{Y}^{ori}$. Note that you are not training a model, but merely applying a trained model to get predicted values from the original, to-be-explained model.
4. Train the surrogate model (g^{sur}) using X and $\hat{Y}^{ori}$.
5. Evaluate how well the predictions from the trained surrogate model ($\hat{Y}^{sur}$) replicate the predictions from the original model ($\hat{Y}^{ori}$) via metrics for measuring regression performance. Below are two sample metrics, root mean square error (RMSE) and R squared, where $\hat{y}_i^{ori}$ is the *i*th prediction from the original model, where $\hat{y}_i^{sur}$ is the *i*th prediction from the surrogate model:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^{sur} - \hat{y}_i^{ori})^2} \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i^{ori} - \hat{y}_i^{sur})^2}{\sum_{i=1}^{n}(\hat{y}_i^{ori} - \bar{\hat{y}}_i)^2} \tag{9}$$

6. If results from #5 show a good congruence (e.g., low RMSE or high R squared values) between the surrogate model and the original model, then you can interpret the surrogate model to understand the original model’s predictions. Ironically, when there is near-perfect congruence between the surrogate and the original model, you might consider replacing the not-so-interpretable original model with the surrogate model for parsimony purpose.

Local Interpretable Model-agnostic Explanations (LIME)

The explanation obtained by LIME ($\xi(x)$) is expressed by the following:

$$\xi(x) = \underset{g \subset G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{10}$$

The following steps depict how to obtain a local surrogate model via LIME:

1. Select a highly intrinsically interpretable model to be the local surrogate model (g^{sur}).
2. Select an instance x for which you want to have an explanation of its original machine learning model.
3. Sample perturbed (i.e., varied) instances around x. Data perturbation needs to be appropriate for the data type. For numeric features (most common seen in organizational science), LIME perturb them by sampling from a normal distribution and doing the inverse operation of mean-centering and scaling based on means and standard deviation in the training dataset.
4. Based on the original trained model (f^{orig}), obtain predictions from the perturbed samples from (3).
5. Weight the perturbed samples from (3) by their proximity to x and train the chosen local surrogate model (g^{sur}) on this weighted dataset.
6. Interpret the local surrogate model to understand the original model's prediction on x.

Shapley value

Mathematically, the Shapley value for player i in game v , with a set N (of n players), is a weighted sum of player i's marginal contribution to the coalitions S (i.e., the coalitions with player i minus the coalitions without player i) in the form of $[v(S \cup \{i\}) - v(S)]$. Taken together, the expression of the unique value of an n-person game (i.e., the expected marginal contribution of player i) is expressed as:

$$\phi_i(v) = \sum_{S \sqsubseteq N\backslash\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)] \tag{11}$$

In a trained model with F as the set of all features, the Shapley value for feature i in a prediction is a weighted sum of this feature i's marginal contribution to the prediction across all possible coalitions of features S ($S \sqsubseteq F$), i.e., the difference between a model trained with feature i present ($f_{S\cup\{i\}}$) and another model trained without feature i (f_S).

$$\phi_i = \sum_{S \sqsubseteq N\backslash\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}})(S \cup \{i\}) - f_S(x_S)] \tag{12}$$

In theory, calculation of the Shapley value on feature i requires iterating through feature sets with and without feature i across all possible sets of features. This is computationally cumbersome as the number of feature sets grows exponentially ($2^{|F|}$) as the number of features increases. To solve this problem, approximation methods have been proposed, such as Shapley Sampling Value (Štrumbelj & Kononenko, 2010, 2014). Shapley Sampling Value applies Monte-Carlo sampling to Equation (11) and approximate the effect of removing a feature from the model to estimate feature importance.