Discover Artificial Intelligence

RESEARCH Open Access

Streamlining admission with LOR insights: AI-Based leadership assessment in online master's program

Meryem Yilmaz Soylu[1*], Adrian Gallard[1], Jeonghyun Lee[1], Gayane Grigoryan[1], Rushil Desai[1] and Stephen Harmon[1]

*Correspondence:
Meryem Yilmaz Soylu
meryem@gatech.edu
[1]College of Lifetime Learning, Georgia Institute of Technology, Atlanta, USA

Abstract

Letters of recommendation (LORs) provide valuable insights into candidates' capabilities and experiences beyond standardized test scores. However, reviewing these text-heavy materials is time-consuming and labor-intensive. To address this challenge and support the admission committee in providing feedback for students' professional growth, our study introduces LORI: LOR Insights, a novel AI-based detection tool for assessing leadership skills in LORs submitted by online master's program applicants. By employing natural language processing and leveraging large language models by using RoBERTa and LLAMA, we seek to identify leadership attributes such as teamwork, communication, and innovation. Our latest RoBERTa model achieves a weighted F1 score of 91.6%, a precision of 92.4%, and a recall of 91.6%, showing a strong level of consistency in our test data. With the growing importance of leadership skills in the STEM sector, integrating LORI—a tool designed with cutting-edge AI models—into the graduate admissions process is crucial for accurately assessing applicants' leadership capabilities. This approach not only streamlines the admissions process but also automates and ensures a more comprehensive evaluation of candidates' capabilities.

Keywords Leadership, 21st-century skills, Durable skills, Natural language processing, Machine learning, Large language models, Graduate education, Holistic admission

1 Introduction

Since the outbreak of the COVID-19 pandemic, it has become clear that various challenges to personal and national economic stability, coupled with rapid advancements in technology and infrastructure, are significantly changing our work and lifestyle dynamics. As a result, these changes are influencing our educational priorities. The increasing need for top-notch education is going beyond conventional school environments and geographical borders, leading to the emergence of online learning platforms that cater to all educational levels and are accessible to learners across the globe. Notably, numerous institutions have recently introduced online graduate degree programs spanning diverse

fields. However, while these programs address the growing demand–supply gap, the mere acquisition of subject matter expertise falls short of adequately equipping individuals to navigate and excel in our rapidly evolving societal landscape. As ongoing transformations continue, it's crucial to develop adaptable skills that can help individuals thrive. These skills are generally referred to as 21st-century skills (21CS), which were acknowledged by researchers (e.g., [10, 66, 75]), educational institutions (e.g. [1, 60, 85],), and economic organizations (e.g., [6, 32]).

Among 21CS, leadership is a highly valued skill in both professional and academic settings. It is critical for higher education institutions to identify and nurture students exhibiting robust leadership qualities. This is particularly crucial for prospective graduate students, as demonstrating a degree of leadership aptitude is essential in showcasing their potential for future advancement. However, there are almost no standardized methods available to evaluate leadership skills during the graduate student admission process. Typically, an applicant's suitability for a program is evaluated through standardized tests (e.g., GRE) and written application documents such as essays, statements of purpose, or letters of recommendation (LORs). Among these, LORs provide valuable insights from external perspectives regarding applicants' experiences and leadership abilities. Yet, manually scrutinizing these letters to assess such competencies demands significant time and resources.

To address this challenge, we propose the development of an AI-driven tool capable of analyzing LORs submitted for an online master's program (OMP) application, with the objective of identifying indicators of leadership. Specifically, this study is guided by the following research objectives: (1) to develop an AI-based tool capable of detecting leadership-related content in graduate applicants' letters of recommendation (LORs); (2) to design a scalable NLP and LLM-based pipeline for extracting and verifying leadership attributes (communication, teamwork, innovation) from LORs; and (3) to evaluate the performance, reliability, and transparency of the tool through systematic analysis and reporting of model performance metrics. Through this work, we aim to contribute a transparent, reproducible, and practical tool for improving holistic graduate admissions review.

2 Related work

2.1 LORs in the admission process

Holistic admissions or "whole-file" review is the consideration of the "broad range of candidate qualities, including non-cognitive or personal attributes when reviewing applications for admissions" ([57] p. 1). In holistic admissions, LORs play a significant role, as they offer unique insights into an applicant's personal and professional characteristics and qualities that extend beyond traditional academic metrics like GPA and test scores. This approach helps graduate programs foster diversity by considering a broader range of candidate qualities, aligning with the principles of the Council of Graduate Schools. LORs provide narratives that offer depth to an application, reflecting personal attributes such as leadership, professionalism, and adaptability [83] and are frequently a factor in final admissions decisions [90].

However, despite their importance, LORs are subject to criticism due to their unstandardized nature [22, 51, 63], the variation in the context of the writer [21, 76], and bias from the writer, the reader, or both [5, 22, 42, 90, 96], which can perpetuate inequality.

A recent study of over 31,000 LORs identified content differences based on gender, race, and intersections of both, although these factors beyond GPA and test scores were not predictive of admission outcomes [22]. Additionally, Kim et al. [58] applied advanced natural language processing to examine over 600,000 counselor recommendation letters, finding notable disparities in length and content tied to race, socioeconomic status, and school type, emphasizing the importance of context-sensitive evaluations in the admissions process.

These findings highlight the complexities of selective admissions. Despite inherent biases, LORs remain valuable in the admissions process as they provide crucial insights into applicants' intellectual engagement, creativity, and potential, helping admissions committees differentiate between candidates with similar academic credentials [15]. This encourages the development of tools that allow for deeper analysis of LORs to better support admission officers.

2.2 Leadership skills in graduate school and beyond

Today, most admissions officers report that their institutions use holistic review in their admissions process [9, 47]. This approach allows graduate programs to assess various applicant qualities, including academic preparedness, demonstrated interest in a specific field, research experience, and 21CS—alternatively referred to as soft, non-cognitive, durable or lifetime skills, such as leadership and perseverance [38, 77, 82, 86, 87].

Among these skills, leadership development is recognized as a critical objective across all disciplines, especially in STEM fields. Studies show that the most effective leaders not only master technical expertise but also excel in professional skills like communication and collaboration [4, 24]. Globally, business leaders and executives often prioritize leadership and talent development programs, recognizing that individuals with strong leadership abilities are essential for ensuring smooth project execution and the timely completion of tasks [24, 67]. For graduate students in the sciences, technical proficiency is a given, while those who possess leadership training are increasingly sought after by employers [12, 79].

Given the significance of leadership, possessing these skills has become highly advantageous for applicants seeking acceptance into graduate-level programs. Leadership capabilities demonstrate a candidate's ability to collaborate effectively, take initiative, communicate clearly, and solve complex problems, all of which highlight their potential for success in the rigorous academic and professional environments of graduate education [97]. Moreover, these attributes suggest a candidate's readiness to assume leadership roles within both academic and professional communities, qualities that are highly valued for future career success [19].

Additionally, research suggests that alignment between applicants' goals and program objectives, along with their demonstrated competencies in 21CS, significantly influences admissions decisions [108]. Among these skills, leadership has emerged as a key predictor of not only enrollment but also retention and overall success in graduate programs [37, 64]. As a result, higher education institutions actively seek evidence of these qualities in application materials, including LORs [52, 63, 97, 100].

2.3 Leveraging NLP to review LORs

Examining LORs like text-heavy application materials is a time-consuming and labor-intensive task. However, recent advancements in technology have led to the development of various artificial intelligence (AI) tools capable of analyzing different attributes of applicants efficiently. One notable application is Natural Language Processing (NLP), a specialized application of machine learning (ML) tailored for interpreting natural language data. NLP techniques use a combination of statistical, ML, and deep learning approaches to understand, interpret, and categorize text based on its content, context, and structure [49].

The strength of NLP lies in its ability to transform unstructured human language into structured data that can be analyzed, interpreted, and applied across various contexts. NLP techniques allow for the efficient processing of vast amounts of text data, automating tasks that would otherwise require significant manual effort [43]. Advanced models, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT), are capable of interpreting ambiguous language, understanding idiomatic expressions, and capturing the nuanced meanings of words within their context. This makes NLP especially powerful for tasks like sentiment analysis, machine translation, and text classification [26].

Language is often ambiguous, meaning that the same word or phrase can have different meanings depending on the context. NLP systems are particularly skilled at resolving this ambiguity by identifying the correct meaning based on the surrounding text. This ability allows NLP to effectively interpret homonyms (words with multiple meanings), metaphors, and other complex language structures. These skills are especially valuable for tasks like answering questions and translating text between languages [45, 89]. In addition, NLP models can be tailored to specific fields, such as law, medicine, or technical areas. By fine-tuning these models for a particular domain, they become more capable of understanding the unique vocabulary, structure, and nuances found in specialized texts, leading to more accurate and relevant analysis [8].

NLP has also been explored in the context of education to automate and enhance the analysis of text-heavy educational data to derive insights into improving teaching and learning outcomes. For instance, Authors (2022) contributed to the understanding of cognitive presence in online learning environments by building an ML model that classifies students' discussion forum posts into phases of cognitive presence. By applying a BERT model, their study achieved 92.5% accuracy in predicting cognitive presence. Similarly, Dornauer et al. [28] developed a German-language cognitive presence classifier for online discussions using linguistic analysis tools, such as Linguistic Inquiry and Word Count (LIWC), and additional learning traces, such as file attachments and course glossary terms. In a recent study [88] using a dataset of 2,500 survey comments from biomedical science courses, the authors showed that GPT-4 can achieve human-level performance across various tasks such as classification, extraction, thematic analysis, and sentiment analysis by leveraging effective prompting.

To date, NLP has been utilized to evaluate students' performance in their application materials, notably LORs, for various post-graduate programs, including admission to graduate school [48, 109], adaptive behavioral compliance [53], and predicting neurosurgical residency outcomes [84]. These studies highlighted the important role of LORs in providing crucial insights into applicants' characteristics and backgrounds,

which significantly influence admission decisions and subsequent performance in graduate programs. Considering the growing importance of leadership skills in the STEM workforce, integrating NLP methods into the admission process for graduate education programs becomes imperative to assess applicants' leadership competencies accurately. This approach not only makes the admissions process more efficient but also allows for a deeper assessment of candidates' capabilities.

In recent years, LLMs such as GPT, LLAMA, and BERT-based variants have significantly advanced the field of natural language understanding, enabling applications ranging from text summarization and classification to open-domain dialogue and question answering [13, 26, 104]. While LLMs have demonstrated strong capabilities in conversational tutoring systems and educational chatbots [53, 91], our use of LLMs in this study serves a distinct purpose. Rather than supporting instructional interactions, we leverage LLMs within a structured pipeline to extract, verify, and summarize leadership-related phrases in LORs.

This distinction is important, as it clarifies the scope of our study. Accordingly, our literature review has focused on work related to educational NLP, leadership detection, and automated analysis of LORs. Nonetheless, we acknowledge that the broader educational potential of LLM-based tutoring systems remains an exciting adjacent domain, particularly as these systems evolve to support personalized learning and formative feedback in other instructional contexts.

3 Methodology

In this section, we describe the design and implementation of our study in detail. We begin with the leadership annotation schema used to guide data labeling. We then present our data collection and preprocessing methods, followed by the iterative development of ML and LLM components. We also report on model evaluation and validation procedures to ensure analytical rigor and transparency throughout the pipeline.

3.1 Leadership annotation schema

Reports from The Chronicle of Higher Education and the World Economic Forum emphasize essential 21CS such as leadership, critical thinking, communication, and teamwork [16, 27]. These skills are increasingly in demand, with organizations urged to prioritize their development [46, 50]. Leadership development alone accounts for nearly US$50 billion in global investments annually [23, 59]. Employees who excel in communication, teamwork, and intercultural competence contribute to organizational productivity and retention, and their participation in cross-functional teams further strengthens leadership capabilities [3]. Particularly in today's rapidly evolving STEM industries, effective leadership is critical to driving innovation and growth [3, 56, 70, 110].

Our comprehensive review of leadership training practices in graduate education revealed a wide range of skills incorporated into these programs [29, 68]. Despite the variety of skills covered, the most consistently emphasized were effective communication, teamwork, and innovation [29, 68, 101] (Fig. 1).

Effective communication, as detailed in various studies [29, 68], relies on strong listening and comprehension skills, whether in speaking or writing. A key component of this is active listening, which involves paraphrasing the speaker's words, encouraging further elaboration, providing feedback, and ensuring the message is accurately understood.

Fig. 1 Conceptual framework of the leadership in this study

Empathy is also crucial, as it requires receptiveness to others' values and emotions, as well as openly sharing one's own thoughts. When human annotators analyzed the sample of LORs, they looked for language that indicated active listening, the ability to adapt communication to diverse audiences, and strategies for overcoming common communication barriers.

In addition to communication, teamwork is also key to success, as no one can succeed in isolation. Interdepartmental and inter-organizational relationships rely heavily on collaboration [40]. Successful collaboration requires openness to diverse perspectives, teamwork in developing plans, and coordinated efforts in execution [35, 61]. As such, human annotators looked for LOR language that highlighted team-building, collaborative work, and the use of tools and platforms to facilitate teamwork.

Finally, innovation lies at the heart of STEM disciplines [68]. It involves questioning the status quo, observing details, and connecting seemingly unrelated concepts. Innovation also requires collaboration with diverse individuals to gain fresh perspectives and experiment with new ideas [3]. Accordingly, human annotators sought language that reflected the ability to spot opportunities for innovation, generate and test new ideas through rapid prototyping and user feedback, manage risks, navigate uncertainties, and embrace failure as an essential part of the innovation process.

3.2 Data collection and processing

3.2.1 Data source

Data used in this study was gathered from the application packages submitted to the OMP offered by a technology-focused public research university in the U.S. The program, which is designed to improve learners' knowledge of big data analytics techniques

through a one-to-two-year program, received more than 10,000 applications as of Spring 2023. The OMP requires the submission of at least three LORs during the application.

Three distinct datasets were prepared for this project. To begin, we required an ample dataset comprising sentences from multiple LORs that were accurately annotated for leadership skills. To obtain this dataset, we employed a Python script to extract individual sentences from the random sample of LORs. Initially, an expert manually annotated sets of LORs from 25 randomly chosen students. Upon analysis of these annotations, we discovered that the dataset was imbalanced with a much larger number of non-leadership sentences than leadership sentences. After generating an initial model that utilized BERT, we applied BERT to a portion of the unlabeled dataset to help our team locate more sentences containing leadership to generate a balanced dataset. By examining the predicted leadership labels and having our expert review and determine which annotations were correct, we were able to include additional leadership sentences to our dataset.

This process resulted in 1,048 sentences from LORs corresponding to 120 unique applicant IDs. These applicants were randomly selected from the entire pool of individuals who applied, regardless of whether they were admitted to the program. The sample of LORs included recommendations written by the applicants' former or current managers, instructors, and colleagues, with the letters varying in format—some were lengthy and detailed, while others were shorter and more informal. This initial set of annotated sentences comprised the first dataset.

These annotated sentences are used to train the weak-supervision models, which utilize datasets where only a portion of the data is manually labeled. This approach leverages a combination of labeled and unlabeled data, making it more cost-effective and efficient compared to fully supervised learning, where all data must be manually labeled [94, 114, 115]. This data was divided into 943 lines of training data and 105 lines of validation data. In the final model run, the 1048-lined dataset was divided into a second dataset of two equal parts comprising a validation set of 524 and a test set of 524 lines of data. The final datasets were created to provide a larger pool of data for validation and testing for the final model.

The second dataset refers to the processed weakly-labeled dataset produced after running the weak-supervision pipeline. Any overlapping student IDs from the first dataset were removed from the unlabeled dataset. Using weak-supervision techniques, we created over 250,000 lines of data, forming the foundation for training a subsequent weakly-supervised model. Initially, the raw data contained 15,293 unique student IDs and 39,465 distinct LORs. Ultimately, the data for training the final model was machine-annotated, while the previously human-annotated dataset served as a benchmark for validating and testing the weakly-supervised model.

A separate group of LORs from a set of students was pulled from the unlabeled dataset (unique to the sentences of the previous dataset) to form a third dataset to check the inter-rater operability between humans and the ML model. Two experts analyzed these sentences using a library of phrases and keywords associated with leadership skills, including teamwork, communication, and innovation (Author, 2024). The sentences were then labeled with "1" if the leadership skill was present and "0" if not. Based on the predicted label for leadership, the human coders' inter-rater reliability, measured using

Cohen's Kappa, was 0.65, indicating a substantial level of agreement among the raters [62, 65, 102].

3.2.2 Preprocessing

Preprocessing steps were conducted at both the weak-supervision pipeline development and model training stages to ensure data quality and enhance performance. These steps included handling outliers, generating numeric features with the Spacy library, and using regex for text pattern matching and word separation functions. The Spacy library was used for NLP tasks such as tokenization and feature extraction, while regex helped identify and manage specific text patterns during data cleaning.

Outliers within the unlabeled dataset were determined based on the distribution of sentence length. This distribution was then broken down into interquartile ranges, and the dataset was reduced to contain only sentences within the Q1 and Q3 ranges, which contained the middle interquartile range of data. This was done to prevent incomplete and run-on sentences from occurring within the dataset.

The generation of numeric features helps to improve the training and performance of the Random Forest model by providing structured, quantifiable representations of the text data. By breaking down the text into components such as verbs, adjectives, and nouns, the model can more effectively understand and differentiate between key characteristics of each sentence [31]. Numeric features were generated for the training of the Random Forest model within the weak-supervision pipeline. All but 1 of the 119 numeric features were extrapolated by using the Spacy library to break down the subcomponents of the text data within each sentence. These numeric features included the number of verbs, adjectives, nouns, etc. These features were then normalized to maintain a similar scale across all features. The character length of a sentence was generated as a separate function outside of Spacy. By converting the text into numeric subcomponents, we enable the model to interpret and analyze the data effectively. Essentially, this process distills the sentences into a structured format that captures linguistic patterns, allowing the Random Forest model to operate on the underlying structures of the English language.

To process the text itself, we implemented a regex function to keep only Alphanumeric characters and a function to correct occurrences of words becoming conjoined to previous words using a Python package called Word Ninja. We set a default threshold of 6 characters within the function based on our examination of the character length distribution from all tokens in the human-annotated dataset and some trial-and-error evaluations over a select subset of sentences directly related to the issue of conjoined words.

3.3 Machine learning model

Our approach to ML development was intentionally iterative and progressive to ensure robust model accuracy and gradual complexity in design [39, 112]. Starting with simpler NLP models, such as Bag-of-Words and n-gram models, provided essential baselines. These models allowed us to evaluate performance with low computational requirements, making it easier to identify areas for improvement before scaling up to more complex frameworks [74, 103]. Additionally, this stepwise progression helped us establish foundational insights, enabling better comparisons and refinements as we introduced advanced models, aligning with best practices in ML development [20]. We explored the

use of SetFit as well as Random Forest models utilizing extracted numeric values from the text using Spacy. However, given the complexity of pulling leadership qualities from the LOR sentences, we eventually turned to Transformer-based models starting with the original BERT [95]. In training BERT, we discovered both a greater level of performance and a bottleneck pertaining to the availability of data as stated in the literature (e.g. [80]).

Our approach aimed to creatively and pragmatically enhance model performance by experimenting with BERT-based frameworks and optimizing data utilization. Initially, we generated synthetic data to increase data diversity and volume, especially to balance the minority label,'leadership.' However, this attempt yielded limited success, as the synthetic samples did not sufficiently improve model performance [34]. We then experimented with integrating a Generative Adversarial Network (GAN)-BERT framework, which combines BERT with GANs to address data scarcity issues, but this also resulted in suboptimal outcomes for our dataset [115]. In response, we turned to larger, more robust iterations of BERT, specifically using RoBERTa, which is designed to improve upon BERT's language masking and training efficiency through a more extensive pretraining process [69]. RoBERTa demonstrated significant improvements over previous attempts, aligning well with the specific task. Nonetheless, we continued to explore further enhancements in pursuit of even greater performance. By iteratively refining our approach and leveraging the larger model's capabilities, we gained deeper insights into model fine-tuning and the limits of data augmentation strategies [26].

Understanding the data bottleneck, we decided to integrate Weak Supervision techniques to create a larger pool of data from our extensive set of unlabeled data. Weak Supervision, which involves labeling data with potentially noisy annotations from multiple sources, is a widely used approach for leveraging large amounts of unlabeled data when manual labeling is costly and time-consuming [93]. Though we anticipated that a weakly supervised dataset would contain some noise, we hypothesized that the increased volume of examples could enhance the model's ability to generalize by exposing it to a broader range of data patterns [114].

To implement this, we developed a custom script to generate a weakly supervised dataset. This approach allowed us to apply labeling functions and heuristics to approximate labels for unlabeled instances, maximizing the utility of our available data while balancing potential noise with the benefits of increased data diversity. Previous studies have shown that, despite some noise, weakly supervised datasets can significantly improve model performance by approximating real-world data distributions, which makes models more resilient and robust to variation [7, 115]. By adopting Weak Supervision, we aimed to create a more robust dataset that would support further model fine-tuning and contribute to a better-performing final model.

During the development of the weak-supervision pipeline, confidence thresholds of 0.7 were established for both the Sentence Transformers for Few-shot Learning (SetFit) [106] & Robustly Optimized BERT Approach (RoBERTa) [69] models. Increasing the threshold beyond this level led to decreased performance of the models. After some trial and error, a threshold of 0.7 was determined to be effective at maintaining consistent output from the models as well as preventing the models from contributing to the pipeline on sentences where they perceived lower confidence in determining the correct label. SetFit & RoBERTa have the most extensive coverage over the unlabeled dataset by far, which led to the implementation of thresholds as a potential safeguard against undue

influence over the other contributing labeling functions. The Random Forest model was initially set to have a threshold of 0.8, but due to insufficient coverage of the unlabeled dataset, the threshold was ultimately left out of the process.

The final ML model was generated using the resulting weakly supervised dataset from the previously mentioned process. RoBERTa was implemented for the final model due to its robust pre-training data having proven effective for our use case. Our dataset contains over 250 k rows of weakly labeled leadership sentences. Initially, the model was trained on data subsets at intervals of 5 k, 25 k, 50 k, & 100 k. With each increase in data, the performance of the final model improved. We used the entire dataset to train the model to achieve strong performance in leadership classification within the LORs.

3.4 LLM model

Our RoBERTa model analyzes the data to extract leadership-related sentences from the LORs. Building on these results, we aimed to further enrich the extracted insights. However, due to the constraints of limited annotated data and the need for deeper analysis, we integrated LLMs to augment the application's capabilities. This addition allowed us to leverage the advanced contextual understanding of LLMs to capture more nuanced details and provide a comprehensive analysis. The LLAMA model was utilized as a predictive model with no interactivity from the user (e.g. not implementing chatbot functionality). The prompt was designed to instruct the model to produce 1) A summary of the identified leadership sentences, 2) Extracted phrases from the leadership sentences that align with leadership, 3) Assign the sublabels of Teamwork, Communication, and/or Innovation per each individual leadership sentence.

Since an LLM is trained on an extremely large dataset, instead of reasoning about the task at hand, it is widely known that LLMs heavily focus on extracting relevant information from the data it was trained on [55]. However, current literature on this topic suggests that there are methods through prompt engineering to get the LLM to demonstrate and apply reasoning skills [91].

Our preliminary findings indicate that the simple approach of trusting an LLM to extract the correct phrases is not the best way to tackle this problem, as it extracts many irrelevant phrases. Recognizing the possibility of unpredictable outputs from LLMs, we implemented constraints to the generated content produced by the LLM using an external library called Guidance. Constraining the outputs of the generative language model provided the overall system with reliable consistency. This, in turn, facilitated our ability to create a pipeline from one output to another (having removed a large part of the unpredictability of the LLM outputs). Due to the lack of additional annotated data per the subcomponents of leadership (communication, teamwork, innovation), we decided to add verification and traceability components to our pipeline. These processes were implemented using Reasoning and Acting (ReAct), a general paradigm that combines reasoning and acting with LLMs with the added capability of utilizing external tools [113]. ReAct prompts LLMs to generate verbal reasoning traces and actions for a task. Essentially, it provides a way to trace the chain of thoughts or the cognitive process within the LLM, from initial reasoning to final action [113]. In addition to the traceability this framework provides, it also allows for the use of external tools outside the context of the LLM model. These external tools are chosen dynamically based on the decision-making of the LLM itself.

When prompted using the ReAct framework, as seen in Fig. 2, the LLM begins by generating a "Thought" related to the question, evaluating which action to take next. It then moves to the "Action" stage, where it selects and applies a predefined tool (located outside the LLM prompt). Following the use of the tool, the LLM enters the "Observation" stage, where it reports the information discovered. This process repeats iteratively until the LLM reaches a "Thought" that it has found the answer, followed by an "Action" to conclude the process and provide the final output.

To leverage ReAct, we built a separate pipeline with different prompts for each of the leadership skills we wanted to extract (teamwork, communication, and innovation). In each of these pipelines, we first used ReAct [113] practices to prompt the LLAMA2 [104] model to generate verbal reasoning traces and actions for the task at hand. This allowed the system to perform dynamic reasoning to create and adapt plans for acting to extract teamwork, communication, and innovation phrases. Table 1 demonstrates an example of how we utilized ReAct prompting to extract teamwork skills.

The ReAct framework provided a key advantage by enabling interaction with external tools and the environment, facilitating the retrieval and integration of additional information necessary for completing a given task. This functionality became particularly important in our work when refining and verifying the leadership phrases generated by the LLM.

Table 1 presents a worked example of this prompting process. We include this example to improve transparency into the pipeline's logic and operation; while presented in table

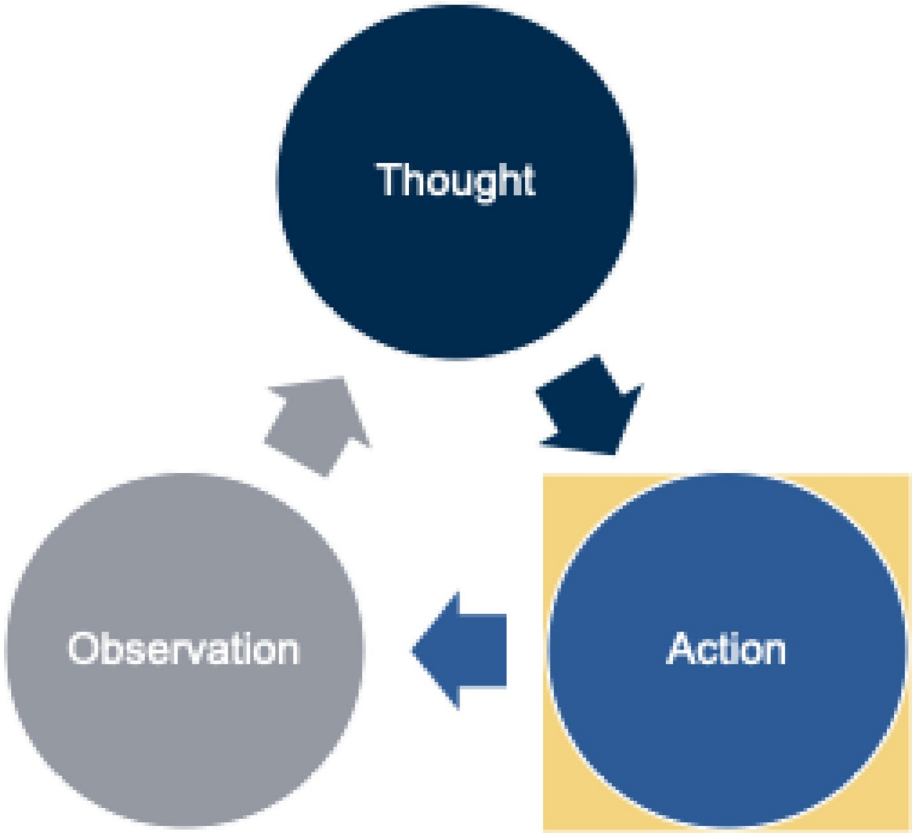

Fig. 2 ReAct flow

Table 1 Sample example of ReAct prompt

Step	Content
Example 1	He is an excellent communicator and a skilled collaborator when working on teams
Thought 1	I should first extract phrases which contain skills related to teamwork
Action 1	Carry out Thought 1 to extract phrases and separate multiple phrases using a ";"
Observation 1	excellent communicator; skilled collaborator
Thought 2	Are all the extracted phrases actually related to teamwork skills? I should verify each of the extracted phrases
Action 2	verify_teamwork("excellent communicator; skilled collaborator")
Observation 2	excellent communicator is a teamwork phrase; skilled collaborator is a teamwork phrase
Thought 3	I now know the final answer
Final Answer	excellent communicator; skilled collaborator

format for clarity, this is not a data results table but a representation of prompt-driven reasoning and verification.

To take full advantage of this capability, we incorporated an additional instance of a separate LLM model. The purpose of this separate instance was to function as a verification mechanism. Importantly, this instance was isolated from the context of the main LLM, meaning it did not have access to the ongoing prompt and responses within the original LLM session. Instead, its role was exclusively to assess and verify the phrases extracted by the primary LLM during the initial stages of the process.

The verification LLM would receive only the extracted phrases as inputs, free from any contextual biases or incomplete information from the original task. This isolation allowed for a more objective assessment, reducing the risk of errors or inconsistencies being propagated through the pipeline. By utilizing this secondary LLM model in a verification capacity, we ensured that only validated and reliable phrases were considered as the final output of the process.

4 Findings

We report a comprehensive evaluation of the weakly supervised RoBERTa model and the LLM-based extraction and verification pipeline. Our analysis includes quantitative performance metrics—precision, recall, F1-score, and inter-rater reliability with human annotators—as well as an error analysis to characterize model limitations. We further present qualitative insights into system behavior using illustrative examples and token-level attribution visualizations to support interpretability and transparency.

We achieved strong performance from the weakly-supervised RoBERTa model, with results indicating high accuracy and reliability. Specifically, the model attained an F1-Score of 91.6%, supported by a precision of 92.4% and a recall of 91.6%, evaluated across 524 instances in the test dataset. These metrics suggest balanced performance, demonstrating the model's effectiveness in identifying relevant instances while maintaining a low error rate.

However, an error analysis revealed that the model currently generates more false positives than false negatives. This indicates that while the model is highly sensitive in detecting relevant phrases, it tends to occasionally misclassify non-relevant instances as positive. This is likely due to overlapping features between positive and non-positive examples in the dataset. Consequently, the model over-predicted the number of leadership sentences, resulting in inter-rater reliability scores of 40.4% and 35.2% for each annotator, respectively.

Ideally, Type I errors are more acceptable in this context, as they contribute to identifying leadership qualities. Observing more false positives could be due to the overall positive tone and context of the recommendation letter. This again confirms the general purpose of these letters, to support the candidate, even if specific skills or traits (like leadership) are not explicitly demonstrated. Also, LORI might associate positive sentiment or praise words (e.g., "excellent," "exceptional") with leadership, even if leadership is not actually implied. Refining the model to address its optimistic bias is an ongoing aspect of our research. Our ultimate goal is to align the model's inter-rater reliability scores with those of human-to-human Cohen's kappa metrics. The confusion matrix (Fig. 3a) further illustrates the model's performance across both classes, showing a substantial number of correctly identified true positives and true negatives (240 and 244 instances, respectively). This balance highlights the model's overall effectiveness while also pointing to opportunities for fine-tuning to reduce false positives in future iterations. Additionally, we will apply explainable artificial intelligence (XAI) techniques to better understand the inner workings of our model and features that contribute the most to the detection of leadership skills [2, 41].

Moreover, Fig. 3(b) presents the summary metric for the precision-recall curve. An average precision of 0.86 confirms the high performance indicated by the confusion matrix, demonstrating the classifier's capability to effectively distinguish between positive and negative samples.

Regarding LLM, one of the key components of our approach was the implementation of a verification layer using a secondary LLM. This verification LLM received only the extracted phrases as inputs, independent of any contextual information from the original task. By isolating these phrases from their broader context, the verification process mitigated potential biases and incomplete information, resulting in a more objective assessment. This strategy reduced the risk of errors or inconsistencies propagating through the pipeline, as only validated and reliable phrases were retained for the final output.

Additionally, the integration of the ReAct framework proved essential in facilitating this validation step. The framework's ability to interact with external systems allowed us to incorporate an additional LLM instance dedicated to verification, introducing a layer of independent scrutiny. This multi-step approach enhanced both the accuracy and reliability of the extracted phrases, as evidenced by the improved consistency and quality of

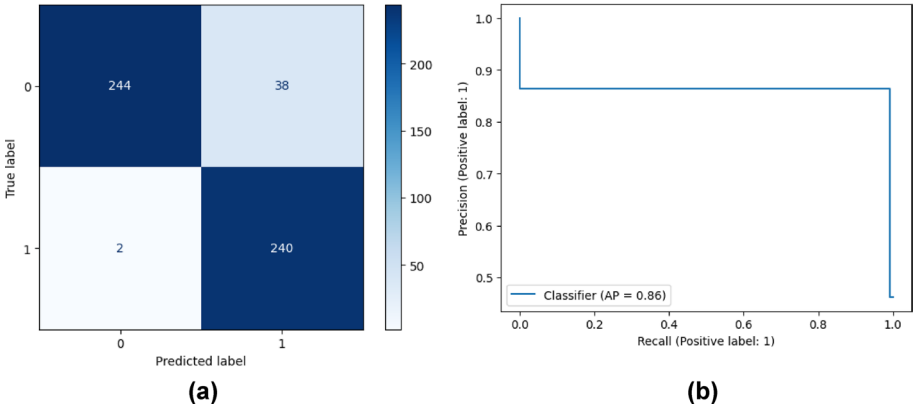

Fig. 3 **a** The model confusion matrix. **b** The precision recall curve

the final outputs. The validated phrases were then used in subsequent analyses, contributing to a more robust and credible set of findings.

To effectively present applicants' leadership attributes from LORs, we developed a minimum viable product (MVP) called LORI—an AI-driven web application prototype built with Streamlit in Python. As shown in Fig. 4, LORI integrates multiple ML models and AI techniques, working in tandem to extract and display meaningful insights from applicants' LORs. The application accepts a PDF file containing three LORs for a given student, converts the letters into images, and applies optical character recognition (OCR) to accurately interpret and process the text. To enable seamless integration, we created additional Python scripts allowing LORI to interact with both the RoBERTa model and the LLAMA2 model (7 billion parameter version). LLAMA2 (Large Language Model Meta AI) is the more advanced successor of LLAMA version 1, offering improved performance and longer context handling. LLAMA is a transformer decoder-based model that generates text by predicting one token at a time based on previous tokens [104]. LLAMA considers self-attention to understand the relationships between

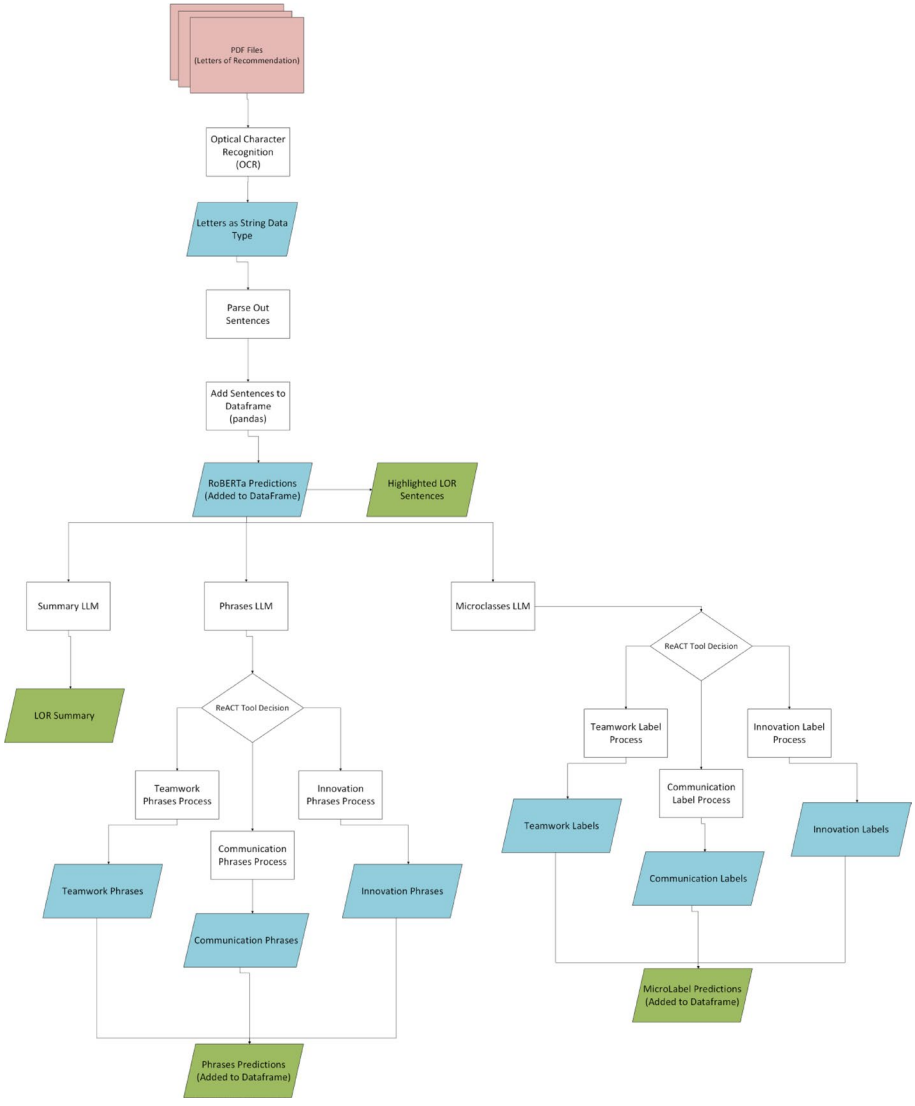

Fig. 4 Flowchart showing LORI's process from pdf file to outputs

the tokens and outputs a probability distribution over the vocabulary for the following words. LLAMA2 and RoBERTa integration allows LORI to leverage the strengths of LLAMA2—like contextual text generation and reasoning—in tandem with more traditional encoder models like RoBERTa. LLAMA2 is used as one of the language models backends to process and classify the presence of varying skills within the LORI using a few-shot prompts. The LOR PDF files are parsed and converted into text, which is processed by the RoBERTa model. The model's output is visualized through highlighted sections, indicating where leadership-related content is detected. These highlighted sentences are further analyzed using LLM pipelines for advanced information extraction, including phrase identification, detailed breakdown of leadership subcomponents, and an overarching summary of leadership qualities across all three LORs.

LORI demonstrates how the AI-based model performed on the tasks of detecting phrases of leadership attributes and tallying the instances of leadership-related phrases. As shown in Fig. 5, the LORI provides information about the number of leadership sentences detected across multiple LORs for an individual applicant. The user can select one of the collected LORs from the dropdown menu to view results associated with the selected letter. For each selected LOR, LORI shows the full text, highlighting specific sentences that contain the leadership phrases. LORI also captures the proportion of the highlighted sentences out of the total number of sentences.

Additionally, powered by the LLM, the Summary feature offers a concise summary of the applicant's leadership attributes based on the synthesis of the information gathered across the three different LORs. We provided the LLM with leadership phases and prompted it to generate a brief overview (approximately 100 words) of each applicant's leadership qualifications. The resulting summaries are presented in this section for admission officers to reference quickly.

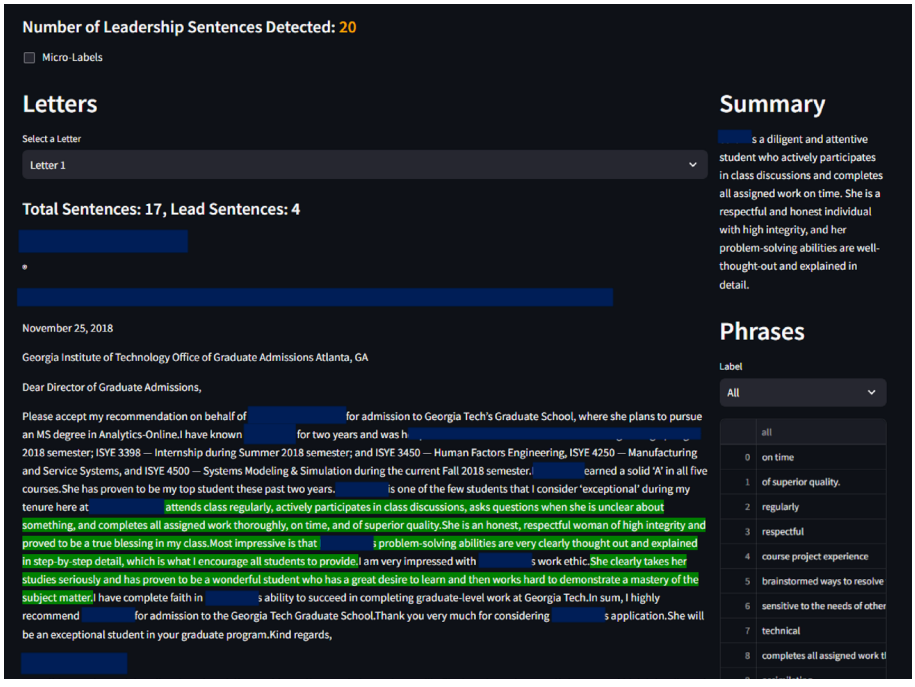

Fig. 5 A screenshot from the LORI MVP prototype, illustrating an example of the model results

Furthermore, LORI displays a bar chart that visualizes the distribution of specific attributes of leadership, including teamwork, communication, and innovation (i.e., micro-label), as illustrated in Fig. 6. These results exhibit the usefulness of LLM in capturing nuanced leadership skills by drilling down into deeper details beyond the initial classifications and dramatically minimizing the data processing for phrase extraction and summarizing.

5 Discussion

The RoBERTa model's overall performance on the test data was very promising and showcases the model's ability to produce a strong level of consistency in detecting leadership skills. However, we believe it is important to note that on the dataset designed to measure agreement between human annotators and the model, there was a larger pool of leadership sentences detected by the model than by the human annotators. This indicates a key point of concern: if the expert annotations are treated as proper labels for the context of the dataset, it shows that the model is biased towards positive identifications of leadership sentences. For the purposes of the LORI app, finding too many leadership skills is preferred over finding too few. However, this result likely indicates that there may be excessive noise within the weakly labeled dataset, suggesting the need for additional examination to further improve model performance. Other potential avenues for improvement include adjusting the model thresholds set for the weakly supervised dataset and possibly adding additional models to enhance the weak supervision pipeline.

Furthermore, the model's performance may be influenced by the inherent biases present in LORs [22, 42]. These biases can stem from various factors, such as the writer's perspective or the socioeconomic background of the applicant, potentially impacting how leadership attributes are described. Therefore, addressing such biases in model training and ensuring that the model generalizes well across different contexts are crucial for

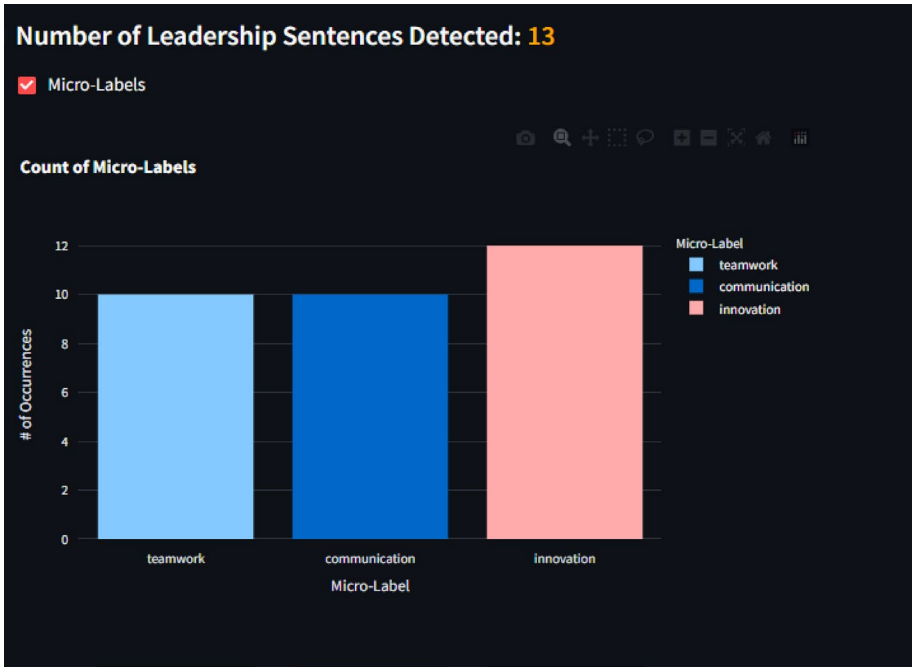

Fig. 6 A screenshot from the LORI MVP prototype, illustrating micro-label (teamwork, communication, innovation) results extracted from LORs

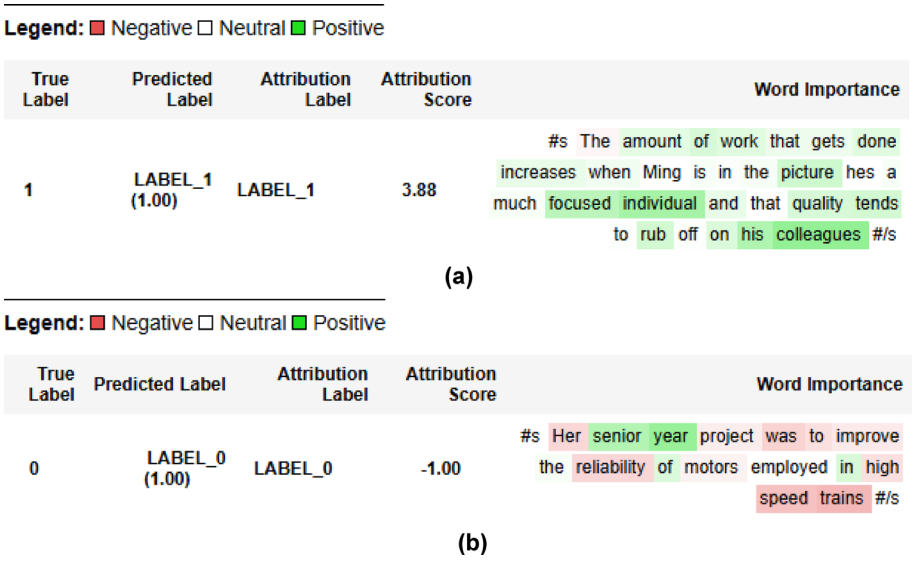

Fig. 7 **a** Example of feature attribution for a true positive prediction. **b** Example of feature attribution for a true negative prediction

future development. While the RoBERTa model has shown effectiveness in this study, future research could explore integrating more diverse training data or employing hybrid models. Specifically, it may be possible to fine-tune the RoBERTa model (or a different transformer-based model) using a publicly available dataset that has a close similarity to the topic of leadership skill detection, which could benefit the model through additional data for fine-tuning.

Another promising direction involves aligning the "importance scores" of tokens within a given sentence with human perceptions of word relevancy to leadership. ML models do not weigh tokens in a sentence the way humans do [95, 107]. Consequently, developing a model that closely aligns with human perceptions would likely not only perform more effectively but also more easily extract key terms directly from the documents rather than relying on an extended process later in the pipeline [17, 105]. At present, the model outputs an overall summary indicating the presence of leadership traits in each letter but does not specify which terms or phrases drive the classification. In the future, improving this output to highlight the most contributive tokens could clarify the aspects of the text that signal leadership qualities. For instance, Fig. 7 provides a Shapley Additive Explanation (SHAP) [73, 81] output displaying feature attributions, where each token or phrase is analyzed to show its respective influence on the final prediction, such as the attribution label. By highlighting the importance levels associated with individual tokens or phrases, these explanations offer insights into which features are most impactful, thereby enhancing the interpretability of the model's predictions.

As seen in Fig. 7(a), the tokens'amount of work that gets done increases,"picture,' and'focused' are particularly influential for this specific observation from the dataset. These tokens indicate elements that the model considers significant in shaping the prediction for this instance. The phrase "amount of work that gets done increases" likely signals an emphasis on productivity or effectiveness, while "picture" may indicate the presence of the individual in a team. The term "focused" suggests an orientation toward concentration or goal-driven behavior. Together, these influential tokens suggest that the model attributes importance to themes of leadership, contribution to teamwork, and

goal-driven, which cumulatively impact the final prediction outcome. Figure 7- (b) presents a case with a true negative prediction. By examining the various influential tokens, we can observe the reasons behind the negative label assignment given these attributions. In cases where the model predicts a false positive or false negative, we can observe the word attributions produced by the model to better understand how the model arrived at that conclusion, which in turn informs us, researchers, on additional considerations when developing and improving the model. Overall, these attributions will pave the way to understanding the nuanced factors that the model interprets as key drivers for the final prediction.

Moving forward, we plan to enhance the RoBERTa model's performance by employing Bayesian Optimization for hyperparameter tuning. Bayesian Optimization is an effective method for hyperparameter search, utilizing a probabilistic surrogate model to explore the parameter space efficiently with fewer evaluations [33, 99]. By iteratively converging on an optimal set of hyperparameters, this approach could significantly improve the model's predictive accuracy and generalizability [111].

Additionally, the LLM component of our tool is undergoing further investigation to assess its effectiveness in tasks such as summarization, phrase extraction, and micro-label categorization. LLMs, including GPT and BERT-based models, have demonstrated strong capabilities in generating high-quality text summaries and extracting meaningful phrases due to their advanced contextual understanding [14, 92]. However, their performance is sensitive to factors such as model size, architecture, and tuning parameters [26]. To address this, we aim to refine evaluation metrics that encompass both qualitative and quantitative dimensions. This will allow us to comprehensively assess the model's capabilities. These ongoing efforts will enhance the robustness and scalability of our tool, ensuring its effectiveness in real-world applications.

The LORI dashboard is a pivotal component of the AI-driven system designed to assess leadership qualities in applicants through their LORs. This dashboard offers a user-friendly interface that provides evaluators with clear, actionable insights into an applicant's leadership attributes, thereby streamlining the admissions process. Notably, the dashboard employs intuitive visualizations to highlight identified leadership qualities such as teamwork, communication, and innovation. This visual strategy allows admissions committees to quickly understand an applicant's strengths and areas for development. Users can also explore specific sections of the LORs to gain deeper context into how leadership traits are put forward. As a result, this interactivity helps admissions committees make well-informed decisions based on comprehensive data. The dashboard enables easy comparison between applicants by aggregating leadership sub-scores in teamwork, communication, and innovation and presenting them side by side. This helps identify standout candidates and supports a fair evaluation process.

While this study establishes the technical feasibility and strong performance of LORI, we also initiated preliminary user engagement by introducing the tool to several admissions teams and gathering initial feedback. These early conversations and feedback sessions provided insights that informed improvements to the tool's interface and clarified where LORI can provide the most value in the admissions process. Notably, admissions officers highlighted LORI's potential usefulness in *edge cases*—for example, when distinguishing between otherwise comparable applicants or when leadership qualities are not immediately obvious from other materials. While these initial interactions were

promising, a more comprehensive user-centered evaluation is planned as part of future work. Formal pilot studies and usability testing with admissions professionals will further assess the system's interpretability, utility, and practical integration into real-world admissions workflows, guiding future refinements to better support human-in-the-loop decision-making.

To further enhance the dashboard, additional text analytics features like word count, readability scores, and word clouds could be integrated. For example, readability scores can indicate the complexity and accessibility of the text, while word clouds provide a visual representation of the most frequent terms, offering a quick overview of key themes [25, 54]. Moreover, by setting a baseline for average metrics across the student population, the tool could enable comparative analysis of leadership skills in individual students. Research suggests that comparative analytics can provide meaningful insights, facilitating personalized feedback and helping educators identify unique student strengths and areas for improvement [11]. Such enhancements would contribute to a more thorough evaluation of student competencies and support targeted educational interventions.

It is essential to recognize that recommendation letters are generally selected and written by referees chosen specifically to reflect a positive assessment of an individual's skills, achievements, and personality. This selection process is inherently biased, as individuals tend to choose referees who are likely to portray them favorably [36, 78]. As a result, this leads to selection and representation biases, with referees more likely to highlight strengths while minimizing weaknesses, creating a skewed sample that is not representative of all possible opinions on an individual's character and abilities [18].

Gender differences further complicate these biases in LORs. Studies have shown that recommendation letters for men often emphasize accomplishments, leadership, and intellectual ability, whereas those for women tend to focus more on personal attributes like kindness or diligence, using more subjective, relationship-oriented language [30, 71, 72, 98]. This language bias can disadvantage women by aligning with stereotypical gender roles rather than objective measures of qualifications. Furthermore, omitted information bias [18] presents another concern. Given that LORs are typically brief, critical details may be omitted, either intentionally or unintentionally. This is problematic, as ML models only have access to the provided content, lacking insight into any significant missing information. For instance, a referee may omit notable achievements of a female applicant due to implicit gender biases, leading the model to undervalue her qualifications compared to male counterparts.

Therefore, if these biases are not addressed, they could result in systematic errors in candidate evaluation, favoring traits often highlighted in LORs for male applicants. To address these issues, adversarial training —a technique that exposes the model to specially designed examples to help it recognize and correct biased patterns—can help the model distinguish between biased and unbiased representations. Moreover, XAI techniques can be used to assist in identifying influential features that drive predictions, which will contribute to promoting a fairer assessment process [2, 44]. By implementing these strategies to mitigate the effects of gender biases, the evaluation of LORs can become more equitable.

Beyond issues of bias, the use of AI to automate the assessment of subjective materials such as LORs introduces important ethical considerations. While tools such as

LORI can enhance efficiency and provide valuable insights, they also raise new questions regarding data privacy, transparency, fairness, and the appropriate role of AI in high-stakes decision-making processes. LORI has been developed with a strong emphasis on data privacy and confidentiality; all LOR data is securely stored and processed in compliance with institutional review board (IRB) approval, and only anonymized data is used for model development and evaluation. Additionally, automating LOR assessment introduces the risk of inadvertently amplifying inequities if model outputs are influenced by variations in writing styles that correlate with applicants' socioeconomic status, cultural background, or native language. Ongoing efforts are needed to ensure fairness and mitigate these risks through diverse training data and explainability techniques. It is also essential to establish clear accountability frameworks, ensuring that admissions decisions remain the responsibility of human evaluators and that AI-generated insights are used transparently and judiciously. Finally, given the rapid evolution of writing practices—including the increasing use of AI-assisted writing tools—it will be critical to continuously monitor model performance and ensure that LORI remains robust to shifts in language use over time.

6 Conclusion

The increasing emphasis on leadership skills in graduate education underscores the need for innovative solutions. Developing and validating LORI to detect these skills benefits both admission committees and applicants by automating the review process, significantly reducing the time and effort required to evaluate application documents. This automation leads to more precise and efficient admissions decisions. As higher education shifts toward holistic reviews that prioritize a broader set of candidate qualities, LORI emerges as an essential tool to promote equity, efficiency, and depth in the admissions process—ultimately shaping a more competent, adaptable, and diverse future workforce.

As we continue to refine LORI, we also recognize the importance of addressing emerging ethical challenges, including the increasing use of AI-generated content in LORs. Future development will incorporate mechanisms for detecting such content and establishing clear guidelines to ensure that AI tools enhance, rather than compromise, the integrity of holistic admissions processes.

Beyond admissions, LORI's versatility extends to instructional settings, where it can analyze other educational data sources such as essays, peer evaluations, and project reflections. By integrating LORI into formative assessment, peer feedback, project-based learning, and virtual classrooms, institutions can enhance leadership development in a scalable, data-driven manner. As a formative assessment tool, LORI provides students with automated, targeted feedback on leadership development through essays, discussion posts, and reflections, allowing for early identification of strengths and areas for improvement while fostering personalized learning paths. In peer review processes, LORI facilitates structured feedback by analyzing evaluations, identifying leadership traits in student submissions, and encouraging self-reflection. Within project-based learning, it assesses leadership indicators in team reports and reflections, enabling faculty to track leadership growth and identify emerging student leaders.

Looking ahead, LORI's applications could expand further, including piloting in leadership-focused graduate courses, integrating into corporate training programs, and

adapting to discipline-specific leadership needs in STEM, business, humanities, and healthcare. By embedding LORI into diverse instructional contexts, institutions can foster leadership development in a more systematic and competency-driven manner, equipping students with the essential skills needed to thrive in dynamic professional environments.

Author contributions
The authors collaboratively contributed to the conceptualization, methodology, data analysis, and manuscript preparation for this study. Specifically, MYS: Led the overall research design and development of the LORI (LOR Insights) framework, led to drafting and revising the manuscript, and supervised the project execution. AG, RD: Conducted the machine learning model development, including fine-tuning RoBERTa and LLAMA models, and implemented the weak supervision pipeline. MYS, AG, JL, SH: Managed data collection and preprocessing, including annotation schema development and dataset balancing, and provided key insights into leadership skill evaluation criteria. AG: Designed and tested the LORI dashboard interface and visualizations. AG, RD: Conducted error analysis, evaluated model performance, and contributed to refining the verification process using LLMs and the ReAct framework. MYS, JL, GG, AG, SH: Led the literature review on leadership skills, holistic admissions, and biases in LORs, and provided substantial contributions to the discussion and conclusions. All authors reviewed and approved the final version of the manuscript. They collectively affirm their responsibility for the integrity and accuracy of the research.

Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of data and materials
The datasets generated and/or analyzed during this study are not publicly available to protect applicant confidentiality, but anonymized data can be obtained from the corresponding author upon reasonable request.

Declarations

Competing Interests
The authors declare no competing interests.

Ethical approval
This study was reviewed and approved by the Institutional Review Board at the Georgia Institute of Technology.

Received: 4 April 2025 / Accepted: 25 July 2025
Published online: 21 October 2025

References

1. ABET. (2019). *General criterion 3. Student outcomes from criteria for Accrediting Engineering Programs, 2018–2019*. https://www.abet.org/accreditation/accre ditation-criteria/criteria-for-accrediting- engineering-programs-2019–2020/#GC3
2. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60.
3. Akdere M, Hickman L, Kirchner M. Developing leadership competencies for STEM fields: The case of Purdue Polytechnic Leadership Academy. Adv Dev Hum Resour. 2019;21(1):49–71.
4. Akhtar A (2020) 20 soft skills every leader needs to be successful https://www.theladders.com/career-advice/20-soft-skills-every-leader-needs-to-be-successful/
5. Akos P, Kretchmar J. Gender and ethnic bias in letters of recommendation: considerations for school counselors. Profess School Counsel. 2016;20(1):1096–2409.
6. Ananiadou K, Claro M 21St century skills and competences for new millennium learners in OECD countries. OECD education working papers, no. 41. OECD Publishing (NJ1) (2009).
7. Bach SH, He B, Ratner A, Ré C. Learning the structure of generative models without labeled data. International Conference on Machine Learning (2017).
8. Bagheri A, Giachanou A, Mosteiro P, Verberne S (2023) Natural Language Processing and Text Mining (Turning Unstructured Data into Structured). In *Clinical Applications of Artificial Intelligence in Real-World Data* (pp. 69–93). Springer.
9. Bastedo MN, Bowman NA, Glasener KM, Kelly JL. What are we talking about when we talk about holistic review? Selective college admissions and its effects on low-SES students. J Higher Educ. 2018;89(5):782–805.
10. Bentur A, Zonnenshain A, Nave R, Dayan T (2019) Education of engineers in the 21st century: Paradigms, insights and implications to Israel. Samuel Neaman Institute for National Policy Research.
11. Bernacki ML, Greene MJ, Lobczowski NG. A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)? Educ Psychol Rev. 2021;33(4):1675–715.
12. Brookes R, Wong B, Ho S (2017) Why scientists should have leadership skills. Scientific American. https://blogs.scientificamerican.com/observations/why-scientists-should-have-leadership-skills.
13. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
14. Brown TB (2020) Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.
15. Butt ME. The role of letters of recommendation in perpetuating or challenging the social stratification of American secondary schools: a quantitative analysis of admission officer assessments in highly selective college admission. Discover Education. 2024;3(1):91.

16. Carlson S. The future of work: How colleges can prepare students for the jobs ahead. Chronicle of Higher Education; 2017.
17. Chancellor S. Toward practices for human-centered machine learning. Commun ACM. 2023;66(3):78–85.
18. Chapman BV, Rooney MK, Ludmir EB, De La Cruz D, Salcedo A, Pinnix CC, Das P, Jagsi R, Thomas CR, Holliday EB. Linguistic biases in letters of recommendation for radiation oncology residency applicants from 2015 to 2019. J Cancer Educ. 2022;37(4):965–72.
19. Chhinzer N, Russo AM. An exploration of employer perceptions of graduate student employability. Educ Train. 2017;60(1):104–20.
20. Chollet F. The future of deep learning. Future. 2017;8(2):5.
21. Clinedinst M, Koranteng A State of College Admission. National Association of College Admission Counseling (2017).
22. Dalal DK, Randall JG, Cheung HK, Gorman BC, Roch SG, Williams KJ. Is there bias in alternatives to standardized tests? An investigation into letters of recommendation. Int J Test. 2022;22(1):21–42.
23. Deloitte Consulting L, by Deloitte B, Deloitte B (2014) Global human capital trends 2014: Engaging the 21st-century workforce. In: Deloitte University Press.
24. Denecke D, Feaster K, Stone K. Professional development: Shaping effective programs for STEM graduate students. Council of Graduate Schools; 2017.
25. DePaolo CA, Wilkinson K. Get your head into the clouds: Using word clouds for analyzing qualitative assessment data. Springer; 2014.
26. Devlin J, Chang M-W, Lee K, Toutanova K Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota (2019).
27. Di Battista A, Grayling S, Hasselaar E Future of jobs report 2023 (2023).
28. Dornauer V, Netzer M, Kaczkó É, Norz L-M, Ammenwerth E. Automatic classification of online discussions and other learning traces to detect cognitive presence. Int J Artif Intell Educ. 2024;34(2):395–415.
29. Dowsett J, Lacey S Optimising online transversal skills delivery in STEM doctoral education. Irish Educational Studies, 1–19 (2023)
30. Dutt K, Pfaff DL, Bernstein AF, Dillard JS, Block CJ. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. Nat Geosci. 2016;9(11):805–8.
31. El-Morr C, Jammal M, Ali-Hassan H, El-Hallak W (2022). Machine Learning for Practical Decision Making. International Series in Operations Research and Management Science.
32. Forum, W. E. (2016). The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution. In: World Economic Forum Geneva.
33. Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint* arXiv:1807.02811.
34. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018),
35. García MG, López CB, Molina EC, Casas EE, Morales YAR. Development and evaluation of the team work skill in university contexts. Are virtual environments effective? Int J Educ Technol High Educ. 2016;13:1–11.
36. Gillis LG (2021). Room for improvement: Reducing the power imbalances in academic letters of reference.
37. Gomez D. Leadership behavior and its impact on student success and retention in online graduate education. Acad Educ Leadership Journal. 2013;17(2):13.
38. Gooch RM, Paris JH, Haviland SB, Sotelo J. (Non) cognitive Dissonance? A Stakeholder-Based Exploration of the Consideration of Graduate Admissions Applicants' Personal Skills and Qualities. Journal of College Access. 2024;9(1):25–42.
39. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
40. Graesser AC, Greiff S, Stadler M, Shubeck KT (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. In (Vol. 104, pp. 106134): Elsevier.
41. Grigoryan, G. (2024). *Explainable Artificial Intelligence: Methods and Evaluation* Old Dominion University].
42. Grimm LJ, Redmond RA, Campbell JC, Rosette AS. Gender and racial bias in radiology residency letters of recommendation. J Am Coll Radiol. 2020;17(1):64–71.
43. Gruetzemacher R (2022) The Power of Natural Language Processing. https://hbr.org/2022/04/the-power-of-natural-language-processing
44. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—Explainable artificial intelligence. Sci Robot. 2019;4(37):eaay7120.
45. Haber J, Poesio M. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. Comput Linguist. 2024;50(1):351–417.
46. Hackett A What are the Key Practices that STEM & Manufacturing based Companies are Deploying to Drive Improvements in the Diversity of their Workforce? (2015).
47. Haviland S, Paris J, Gooch R, Sotelo J What's now and what's next in graduate admissions policies: Results from the ETS/NAGAP 2022 admissions survey (ETS Research Notes). Educational testing service (2023).
48. Heilman M, Breyer FJ, Williams F, Klieger D, Flor M. Automated analysis of text in graduate school recommendations. ETS Res Rep Ser. 2015;2015(2):1–12.
49. Holdsworth J What is NLP (natural language processing)? IBM. https://www.ibm.com/topics/natural-language-processing (2024)
50. Hora MT. Beyond the skills gap: Preparing college students for life and work. Harvard Education Press; 2019.
51. Houser C, Lemmons K. Implicit bias in letters of recommendation for an undergraduate research internship. J Furth High Educ. 2018;42(5):585–95.
52. Hout M. Berkeley's comprehensive review method for making freshman admissions decisions: An assessment. Berkeley: University of California; 2005.
53. Jeon J, Lee S (2023) Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies, 1–20.
54. Kalmukov Y (2021) Using word clouds for fast identification of papers' subject domain and reviewers' competences. arXiv 2021; 26 dez
55. Kambhampati S. Can large language models reason and plan? Annals of the New York Academy of Sciences; 2024.

56. Karimi H, Pina A. Strategically addressing the soft skills gap among STEM undergraduates. J Res STEM Educ. 2021;7(1):21–46.
57. Kent JD, McCarthy MT (2016) Holistic review in graduate admissions. Council of Graduate Schools.
58. Kim BH, Park JJ, Lo P, Baker D, Wong N, Breen S, Truong H, Zheng J, Rosinger K, Poon OA Inequity and College Applications: Assessing Differences and Disparities in Letters of Recommendation from School Counselors with Natural Language Processing. EdWorkingPaper No. 24–953. Annenberg Institute for School Reform at Brown University (2024).
59. Kirchner MJ, Akdere M. Examining leadership development in the US Army within the human resource development context: Implications for security and defense strategies. Korean J Def Anal. 2014;26(3):351–69.
60. Knowledge DT (2012). Education for life and work.
61. Koh E, Hong H, Tan JP-L. Formatively assessing teamwork in technology-enabled twenty-first century classrooms: exploratory findings of a teamwork awareness programme in Singapore. Asia Pacific J Educ. 2018;38(1):129–44.
62. Kolesnyk A, Khairova N. Justification for the use of Cohen's Kappa statistic in experimental studies of NLP and text mining. Cybern Syst Anal. 2022;58(2):280–8.
63. Kuncel NR, Kochevar RJ, Ones DS. A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. Int J Sel Assess. 2014;22(1):101–7.
64. Kyoung Ro H, Lattuca LR, Alcott B. Who goes to graduate school? Engineers' math proficiency, college experience, and self-assessment of skills. J Eng Educ. 2017;106(1):98–122.
65. Landis JR, Koch GG. The measurement of observer agreement for categorical data. UK: International Biometric Society Stable; 2016.
66. Lavi R, Tal M, Dori YJ. Perceptions of STEM alumni and students on developing 21st century skills through methods of teaching and learning. Stud Educ Eval. 2021;70: 101002.
67. Lawrence E, Dunn MW, Weisfeld-Spolter S. Developing leadership potential in graduate students with assessment, self-awareness, reflection and coaching. J Manag Dev. 2018;37(8):634–51.
68. Lenhart C, Bouwma-Gearhart J, Keszler DA, Giordan J, Carter R, Dolgos M. STEM graduate students' development at the intersection of research, leadership, and innovation. J Coll Sci Teach. 2022;52(2):3–5.
69. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
70. Lnenicka M, Kopackova H, Machova R, Komarkova J. Big and open linked data analytics: a study on changing roles and skills in the higher educational process. Int J Educ Technol High Educ. 2020;17:1–30.
71. Madera JM, Hebl MR, Dial H, Martin R, Valian V. Raising doubt in letters of recommendation for academia: gender differences and their impact. J Bus Psychol. 2019;34:287–303.
72. Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. J Appl Psychol. 2009;94(6):1591.
73. Mangalathu S, Hwang S-H, Jeon J-S. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng Struct. 2020;219: 110927.
74. Manning CD, Raghavan P, Schütze H. An Introduction to Information Retrieval. Cambridge UP; 2009.
75. Marbach-Ad G, Eagan L, Thompson K. A discipline based teaching and learning center. Springer Publications; 2015.
76. McCarthy C, Van Horn Kerne V, Calfa NA, Lambert RG, Guzmán M. An exploration of school counselors' demands and resources: relationship to stress, biographic, and caseload characteristics. Profes School Counsel. 2010;13(3):2156759X1001300302.
77. Michel RS, Belur V, Naemi B, Kell HJ. Graduate admissions practices: A targeted review of the literature. ETS Res Rep Ser. 2019;2019(1):1–18.
78. Morgan WB, Elder KB, King EB. The emergence and reduction of bias in letters of recommendation. J Appl Soc Psychol. 2013;43(11):2297–306.
79. National Academies of Science, E, & Medicine, Graduate STEM education for the 21st century (2018).
80. Niu C, Li C, Ng V, Chen D, Ge J, Luo B An empirical comparison of pre-trained models of source code. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE) (2023),
81. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed. 2022;214: 106584.
82. Nye CD, Ryan AM. Improving graduate-school admissions by expanding rather than eliminating predictors. Perspect Psychol Sci. 2023;18(1):54–60.
83. Okahana H, Augustine R, Zhou E (2018). Master's admissions: Transparency, guidance, and training. Washington, DC: Council of Graduate Schools. Retrieved March, 16, 2019.
84. Ortiz AV, Feldman MJ, Yengo-Kahn AM, Roth SG, Dambrino RJ, Chitale RV, Chambless LB. Words matter: using natural language processing to predict neurosurgical residency match outcomes. J Neurosurg. 2022;1:1–8.
85. P21 (2019) P.f.s.C.L. Framework for 21st century learning definitions. 2019. https://static.battelleforkids.org/documents/p21/P21_Framework_DefinitionsBFK.pdf
86. Paris JH, Birnbaum M Selecting Students Graduate Admissions Criteria and Evaluation Methodologies Joseph H. Paris, Matthew Birnbaum, and Nicholas Dix. A Comprehensive Guide to Graduate Enrollment Management: Advancing Research and Practice, 72 (2024).
87. Paris JH, Birnbaum M, Dix N (2020) Selecting Students: Graduate Admissions Criteria and Evaluation Methodologies. In A Comprehensive Guide to Graduate Enrollment Management (pp. 72–91). Routledge.
88. Parker MJ, Anderson C, Stone C, Oh Y. A large language model approach to educational survey feedback analysis. Int J Artif Intell Educ. 2024;35(1):1–38.
89. Patwardhan N, Marrone S, Sansone C. Transformers in the real world: a survey on NLP applications. Information. 2023;14(4):242.
90. Posselt JR. Trust networks: a new perspective on pedigree and the ambiguities of admissions. Rev High Educ. 2018;41(4):497–521.
91. Qiao S, Ou Y, Zhang N, Chen X, Yao Y, Deng S, Tan C, Huang F, Chen H. Reasoning with language model prompting: A survey. arXiv preprint arXiv:2212.09597 (2022).
92. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.

93. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C (2017) Snorkel: Rapid training data creation with weak supervision. In: Proceedings of the VLDB endowment. International conference on very large data bases,
94. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J. 2020;29(2):709–30.
95. Rogers A, Kovaleva O, Rumshisky A. A Primer in BERTology: What We Know About How BERT Works. Trans Assoc Comput Linguist. 2021;8:842–66. https://doi.org/10.1162/tacl_a_00349.
96. Sagaria MAD. An exploratory model of filtering in administrative searches: toward counter-hegemonic discourses. J Higher Educ. 2002;73(6):677–710.
97. Sandlin M. An admissions/enrollment imperative for predicting student success. College and University. 2019;94(2):2–11.
98. Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles. 2007;57(7):509–14.
99. Snoek J, Larochelle H, Adams RP Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems 25 (2012).
100. Sternberg RJ. College admissions: beyond conventional testing. Change Mag Higher Learn. 2012;44(5):6–13.
101. Strubbe L, Bosinger M, Stauffer HL, Tarjan M (2022) The value of teaching leadership skills to STEM graduate students and postdocs. Leaders in effective and inclusive STEM: Twenty years of the Institute for Scientist & Engineer Educators.
102. Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method. 2011;11:145–63.
103. Tao C, Gao S, Li J, Feng Y, Zhao D, Yan R Learning to organize a bag of words into sentences with neural networks: An empirical study. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021).
104. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
105. Truong S, Koyejo S. Machine Learning From Human Preferences. The MIT Press; 2025.
106. Tunstall L, Reimers N, Jo UES, Bates L, Korat D, Wasserblat M, Pereg O (2022) Efficient Few-Shot Learning Without Prompts. arXiv preprint arXiv:2209.11055.
107. Wallace E, Feng S, Kandpal N, Gardner M, Singh S. Universal adversarial triggers for attacking and analyzing NLP. arXiv preprint arXiv:1908.07125 (2019)
108. Walpole M, Burton NW, Kanyi K, Jackenthal A. Selecting successful graduate students: in-depth interviews with GRE® users. ETS Res Rep Ser. 2002;2002(1):i–29.
109. Waters A, Miikkulainen R. Grade: machine learning support for graduate admissions. AI Mag. 2014;35(1):64–64.
110. Watt WM. Effective leadership education: developing a core curriculum for leadership studies. J Leadersh Educ. 2003;2(1):13–26.
111. Wu J, Poloczek M, Wilson AG, Frazier P Bayesian optimization with gradients. Adv Neural Inf Process Syst 30 (2017).
112. Xin D, Ma L, Song S, Parameswaran A How Developers Iterate on Machine Learning Workflows--A Survey of the Applied Machine Learning Literature. arXiv preprint arXiv:1803.10311 (2018).
113. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, CaoY React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022).
114. Zhou Z-H. A brief introduction to weakly supervised learning. Natl Sci Rev. 2018;5(1):44–53.
115. Zhu D, Shen X, Mosbach M, Stephan A, Klakow D Weaker than you think: A critical look at weakly supervised learning. arXiv preprint arXiv:2305.17442 (2023).

Publisher's Note