Masked and Fair?
Identifying and Mitigating Gender Cues in Academic Recommendation Letters with Interpretable NLP

Anonymous ACL submission

Abstract

Letters of recommendation (LoRs) can carry patterns of gendered language that can inadvertently influence downstream decisions, e.g. in hiring and admissions. In this work, we investigate the extent that Transformer-based Large Language Models (LLMs) can infer the gender of applicants in academic LoRs after explicit identifiers like names and pronouns are de-gendered. When fine-tuning three LLMs (DistilBERT, RoBERTa, and Llama 2) to classify the gender of anonymized and de-gendered LoRs, we find significant gender leakage evidenced by up to 68% classification accuracy. Using text interpretation methods, TF-IDF and SHAP, we demonstrate that certain linguistic patterns are strong proxies for gender, e.g. "*emotional*" and "*humanitarian*" are commonly associated with LoRs for female applicants. As an experiment in creating truly gender-neutral LoRs, we remove these implicit gender cues and observed a drop of up to 7% accuracy and 4% macro F_1 score on re-training the classifiers. However, applicant gender prediction still remains better than chance. Our findings highlight that LoRs contain gender-identifying cues that are hard to remove and may activate bias in decision-making. While technical solutions may be a concrete step toward fairer academic and professional evaluations, future work is needed to ensure gender-agnostic LoR review.

1 Introduction

Letters of recommendation (LoRs) remain one of the most influential yet least structured components of academic and professional evaluation. While evaluators ostensibly focus on merit-based content, a growing body of work (Isaac et al., 2009; Rice and Barth, 2016; Dastin, 2018; Alexander, 2022) demonstrates the presence of seemingly innocuous linguistic cues in application materials that can indicate gender and systematically sway perceptions of applicant competence, leadership, and fit. Detecting such patterns is therefore essential both for understanding implicit bias and for engineering AI-supported professional evaluation pipelines that can be safely and fairly deployed in real-world selection processes.

In an academic admissions context, this work explores this challenge along two fronts. First, casting gender identification as a supervised text classification task, we investigate the presence of implicit gender cues in LoRs by comparing how much gender signal is carried by explicitly gendered applicant-focused language versus the broader narrative. Specifically, we fine-tune several pre-trained LLMs (Sanh et al., 2020; Liu et al., 2019; Touvron et al., 2023b) on a corpus of 8,992 LoRs submitted to a medical residency program. These LLM text classifiers are trained on the raw text from anonymized LoRs, then subsequently on text with explicit gender identifiers (pronouns, titles, kinship terms, etc.) replaced with fixed gender alternatives. From this experiment, we find that LLMs achieve above-chance performance on classifying gender of applicants in explicitly de-gendered LoRs (68% accuracy and 60% macro F_1), suggesting the presence of implicit gender cues in LoRs.

Second, we apply common interpretability methods (TF-IDF and SHAP) to identify linguistic patterns associated with model decisions (i.e. possible implicit cues for applicant gender). Upon removing these implicit gender indicators, we then re-train models and re-evaluate their performance on the sanitized text. Significant drops in performance (i.e., to 61% accuracy and 56% macro F_1) indicate a partially successful obscuration of gender, but model predictions remain above chance. Complementing aggregate metrics, we deploy SHAP value decompositions to visualize which tokens the classifier leans on before and after de-gendering. These explanations audit the fairness of model decisions and may suggest iterative rules of refinement.

Collectively, this study yields (i) a repro-

ducible discriminator for gender inference from real human-written LoRs, (ii) a quantitative assessment of tokens associated with such inference, and (iii) an interpretable de-gendering strategy that can provide quantitative evaluation, and partial mitigation, of pervasive implicit gender cues in LoRs.

2 Related Works

Prior work has long gathered empirical evidence of significant gender bias in professional contexts, such as hiring decisions (Isaac et al., 2009; Koch et al., 2015; Rice and Barth, 2016; Hoover et al., 2019; Keck and Tang, 2020). In the context of open-source software development, Imtiaz et al. (2019) found that women's GitHub pull requests were, on average, accepted more frequently than men's, unless the contributor's gender was publicly identifiable, at which point acceptance rates fell significantly. The authors inferred that higher rejection stems not from inferior code quality, but from bias activated by visible gender markers. Similarly, Simon et al. (2023) found systematic differences in patterns of language used in LinkedIn profiles by gender. AI language model-based text classifiers are effective tools to expose correlations between text data and various class labels and categories (Schwartz et al., 2017; Gururangan et al., 2018; Poliak et al., 2018; Niven and Kao, 2019). As such, they can identify implicit gender cues in application materials, and when used as hiring tools, they can leverage such cues to inadvertently perpetuate historical hiring bias reflected in training data (Dastin, 2018). Highly relevant to our work, Liu et al. (2022) used a large language model (LLM) to assess gender bias in human-written feedback for medical students, finding that terms related to family and children were more likely to be used in evaluating female students. These findings support our concern that human-written text can encode latent gender information undetected by naïve anonymization.

Meanwhile, NLP technologies like LLMs themselves often learn and reproduce societal stereotypes present in text corpora used to pre-train them, exacerbating concerns of bias. Prior work has found evidence of gender and other social bias across broad domains and tasks, including in distributional semantic representations (Bolukbasi et al., 2016; Zhao et al., 2019), coreference resolution systems (Zhao et al., 2019; Rudinger et al., 2018), text classifier decisions (De-Arteaga et al., 2019; Jentzsch and Turan, 2022), and LLM-generated text (Bolukbasi et al., 2016; Kotek et al., 2023; Wan et al., 2023; Dhingra et al., 2023; Soundararajan and Delany, 2024; Wu and Ebling, 2024; Mirza et al., 2025). In turn, growing attention has focused on the behavior of LLMs in high-stakes selection contexts (Hickman et al., 2024; Phillips and Robie, 2024; Henkel et al., 2024; Leong et al., 2024; Li et al., 2024b,a; Wang et al., 2024; Karvonen and Marks, 2025), developing resources to support safe and fair application of LLMs in such areas.

Together, these studies establish the persistence of gender cues in professional and meritocratic artifacts and the potential of NLP and language technologies to perpetuate social biases. Building on this foundation, we demonstrate the presence of implicit gender cues in letters of recommendation (LoRs) and contribute an interpretable de-gendering workflow that combines LLMs with interpretable feature engineering to quantify (but only partially mitigate) gender leakage in LoRs.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a corpus of letters of recommendation (LoRs), where $x_i \in X$ denotes the i-th document and $y_i \in \{0, 1\}$ encodes the self-identified gender of the applicant (**0** for female, **1** for male). Our goal is twofold:

1. learn a mapping $f_\theta : X \to \{0, 1\}$ that predicts the gender from x as $\hat{y} = f_\theta(x)$, and
2. quantify features or tokens $v \in V$ derived from x using a function $\phi : V \to \mathbb{R}$ and suppress the lexical evidence that enables such prediction $\hat{y}$.

Formally, we view f_θ as a text-based *gender classifier* with parameters θ built atop a pre-trained Transformer encoder (Vaswani et al., 2017) and optimized via. *task-specific fine-tuning*. Also, features in our case are considered as tokens derived from the LoR texts.

3.2 Training a Gender Classifier

3.2.1 Model Architecture

We experiment with three language model variants on the LoR dataset to build our gender classifier.

DistilBERT. A six-layer student network distilled from BERT-BASE ($\sim$66 M parameters vs. 110 M) that preserves $\geq$ 95% of the original performance while being 40% smaller and 60% faster (Sanh et al., 2020).

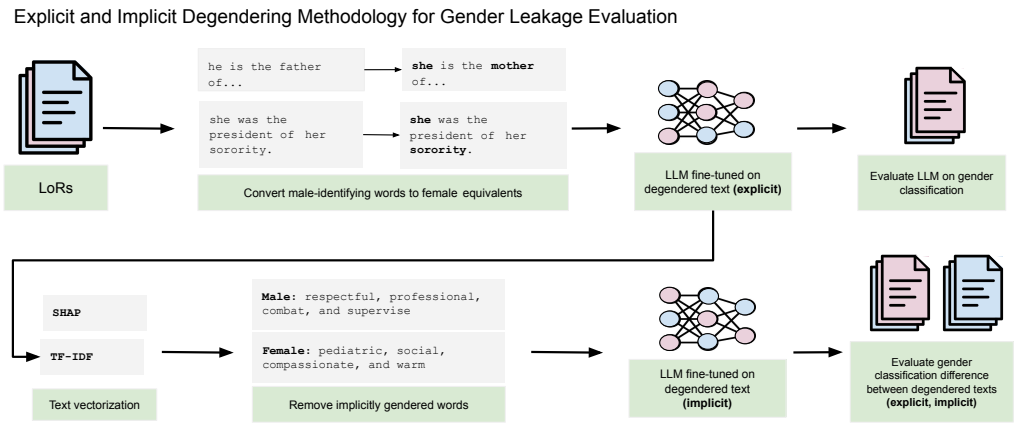


Figure 1: Illustration of the De-Gendering workflow showing the steps for initial Explicit De-Gendering and training followed by Implicit De-Gendering, fine-tuning and re-evaluation

RoBERTa. A robustly optimized BERT derivative trained with larger mini-batches, dynamic masking, and the removal of the next-sentence objective, yielding superior downstream accuracy (Liu et al., 2019).

Llama 2. A 7B parameter decoder-only model pre-trained on 2T tokens (Touvron et al., 2023a), fine-tuned using parameter-efficient LoRA adapters (Hu et al., 2022) inserted into attention and feed-forward layers. This allows efficient classification on limited GPU memory while preserving the core weights.

For all models, the final hidden state associated with the canonical `[CLS]` token ($\mathbf{h}_i \in \mathbb{R}^d$) is passed through a trainable affine head $\hat{y}_i = \sigma(\mathbf{w}^\top \mathbf{h}_i)$ with parameters $\mathbf{w}$ where σ denotes the logistic function and d is the dimension of the hidden state, predicting the gender category.

3.2.2 Data Processing

Our initial data processing pipeline includes a regex-based token matching filter, replacing all gender-identifying tokens (names, titles, pronouns, and kinship terms) with special tokens or their female counterparts. The resultant filtered texts were used for training and evaluating the baseline classifier. Upon building a baseline classifier, each LoR was further subjected to an automatic *de-gendering* filter g_ϕ such that $\tilde{X} = g_\phi(X)$, based on the the gender predictability factor ϕ as shown in Figure. 1. The dataset was randomly partitioned into $\mathcal{D}_{\text{train}}$:$\mathcal{D}_{\text{val}}$:$\mathcal{D}_{\text{test}} = 80 : 10 : 10$. Tokenization follows the standard BERT based tokenization scheme with a maximum sequence length of L=512.

3.2.3 Evaluation Metrics

We report accuracy, precision, recall, and macro-averaged F_1 score on $\mathcal{D}_{\text{test}}$. TF-IDF and SHAP values are qualitatively inspected to validate the efficacy of the de-gendering filter.

3.3 Quantifying Implicitly Gendered Tokens

SHAP : To inspect residual gender leakage, we employ SHAP *(SHapley Additive exPlanations)* (Lundberg and Lee, 2017). Formally, let $f : \mathbb{R}^n \to \mathbb{R}$ denote a predictive model, and let $x \in \mathbb{R}^n$ represent an instance with n features. The goal is to decompose the output $f(x)$ as a linear combination of feature contributions such that $f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$, where ϕ_0 is the base value (expected output over the dataset) and ϕ_i represents the marginal contribution of feature i to the deviation from the base. The SHAP value ϕ_i for a feature i is computed by taking the average marginal contribution of that feature over all possible subsets $S \subseteq \{1, \ldots, n\} \setminus \{i\}$, defined as: $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_{S(x_S)} \right]$ where $f_S(x_S)$ denotes the model trained (or approximated) using only features in subset S, and x_S is the projection of x onto S. Here, ϕ_i denotes the SHAP value associated with token v_i, quantifying how much that token or word contributes to the model's deviation from the base prediction over the dataset. Positive SHAP values indicate that the word pushes the prediction toward a particular gender, while negative values push it away. For instance, high-magnitude SHAP values associated with occupational terms ("*nurse*", "*engineer*") reveal how the model associates language features with gender.

TF-IDF : By computing the TF-IDF (Term Frequency–Inverse Document Frequency) scores of words across documents labeled by gender, one can determine which terms are most characteristic or discriminative of each gender class. TF-IDF boosts terms that are frequent in a document but rare across the corpus. If certain terms consistently have higher TF-IDF scores in documents of a particular gender, then those terms clearly show a strong contributing factor influencing the prediction towards that particular gender.

4 Experiments

4.1 Classification

In our classification step, we begin by training a language model on the original LoR texts in which only the applicant names are anonymized. Since this version of the data contains explicit gender indicators such as pronouns (*he*, *she*) and titles (*Mr.*, *Mrs.*), we expect the model to predict applicant gender with near-perfect accuracy, serving as our baseline. The model used for this experiment was DistilBERT.

We then train a series of models on de-gendered versions of the text $\tilde{X}$, where all explicit gender-identifying tokens have been replaced with their female counterparts. This allows us to examine whether gender can still be inferred from more subtle linguistic patterns. For these experiments, we fine-tune transformer-based models, which includes DistilBERT, RoBERTa, and Llama 2, to evaluate their performance on the gender classification task in the absence of overt gender signals (Sanh et al., 2020).

4.1.1 Dataset

Our dataset consists of $8,992$ recommendation letters, each written on behalf of candidates applying to a major U.S. anesthesiology residency program. Of these, $2,787$ letters were written for female applicants and $6,205$ for male applicants, meaning approximately 31% of the dataset represents female applicants and 69% represents male applicants. To preserve anonymity, applicant names in the letters were replaced with fixed special tokens such as `FIRST_NAME`, `MIDDLE_NAME`, `LAST_NAME`, or `IDENTIFIER`.

4.1.2 Data Processing Pipeline

To de-gender the original dataset, i.e. to neutralize explicit gender-identifying tokens, we compiled a comprehensive list of gendered terms by aggregating entries from two publicly available sources: Bias-BERT and GN-GloVe (Jentzsch and Turan, 2022; Zhao et al., 2018). These lists include both obvious gender markers (e.g., *he*, *she*, *Mr.*, *Ms.*) and less immediately obvious terms with clear gender associations (e.g., *husband*, *father*, *brother*, *actor*, *actress*, *fraternity*, *sorority*). By excluding terms such as "*father*" or "*sorority*", we ensure that gender cannot be inferred from sentences like "*he is the father of...*" or "*she was the president of her sorority.*"

To ensure comprehensive coverage, we used regular expressions to capture variations of each term, including plural forms, verb tenses, contractions, punctuation, and positioning within a sentence. This allowed us to detect and replace forms such as "*she's*", "*husband.*", and "*mothers!*" with high precision. We then replaced any token appearing in the aggregated list with its counterpart from a single gender class to generate an **E**xplicitly **D**e-**G**endered (**EDG**) dataset. Specifically, all explicitly male-identifying terms were converted to their female equivalents. While we considered using neutral placeholders (e.g., *they*, *them*), this approach risked disrupting the grammatical structure of the letters and introducing unnatural linguistic artifacts. By consistently converting all gendered terms to female, we preserved grammatical fluency while preventing the model from relying on overt gender cues, thereby encouraging it to learn from subtler, implicit gender signals embedded in the language.

4.1.3 Training Setup

Using the EDG dataset, we trained DistilBERT, RoBERTa, and fine-tuned Llama 2 models. These Transformer-based pre-trained models were selected due to their state-of-the-art performance on a wide range of downstream classification tasks. For each model architecture, we conducted independent hyper-parameter sweeps. DistilBERT and RoBERTa were trained using HuggingFace's Trainer API, while Llama 2 used Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaption (LoRA). The parameters were optimized by Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$) and a linear learning-rate schedule. Unless otherwise noted, the training runs adopt the hyper-parameters mentioned in Appendix A and were executed on Intel Xeon Gold processors and NVIDIA A100 (40 GB) GPU hardware.

We first trained **DistilBERT** model on the original corpus (non-EDG) - retaining all gender markers - to establish an upper bound on gender learnability, using an $80 : 10 : 10$ stratified train–validation–test split, the same classification head, and only the first two encoder layers unfrozen. An identical configuration was then applied to the *de-gendered* corpus (EDG), this time keeping every sentence in each letter to preserve potential implicit gender cues. Finally, to assess model choice on the de-gendered data, we fine-tuned **RoBERTa** and **Llama 2** on the same splits with their individually tuned best settings. Further details of the hyper-parameters optimization are provided in the Appendix A.

4.2 Baseline Evaluation

We evaluated our trained models on a held-out test set that was not used during training. By measuring the classification performance, we assess how effective our initial matching-based de-gendering method is at neutralizing explicit gender signals, and whether the model can still infer gender from implicit cues. This helps us evaluate both the strength of the remaining bias in the text and the robustness of the model's predictions. To assess performance, we considered standard classification metrics, including accuracy, precision, recall, and F_1 score. Given the class imbalance in our dataset—with a larger proportion of letters written for male applicants-we placed particular emphasis on the macro F_1 score, which provides a more balanced evaluation by averaging the F_1 scores across both classes independently. This ensures that the model's performance on the minority class (female applicants) is not overshadowed by the majority class.

4.3 Selecting Implicitly Gendered Tokens

To interpret our classification results, we explored the linguistic artifacts and implicit signals the model may be using to identify gender, offering insight into the subtler ways in which gendered language cues can manifest in recommendation letters.

4.3.1 SHAP Analysis

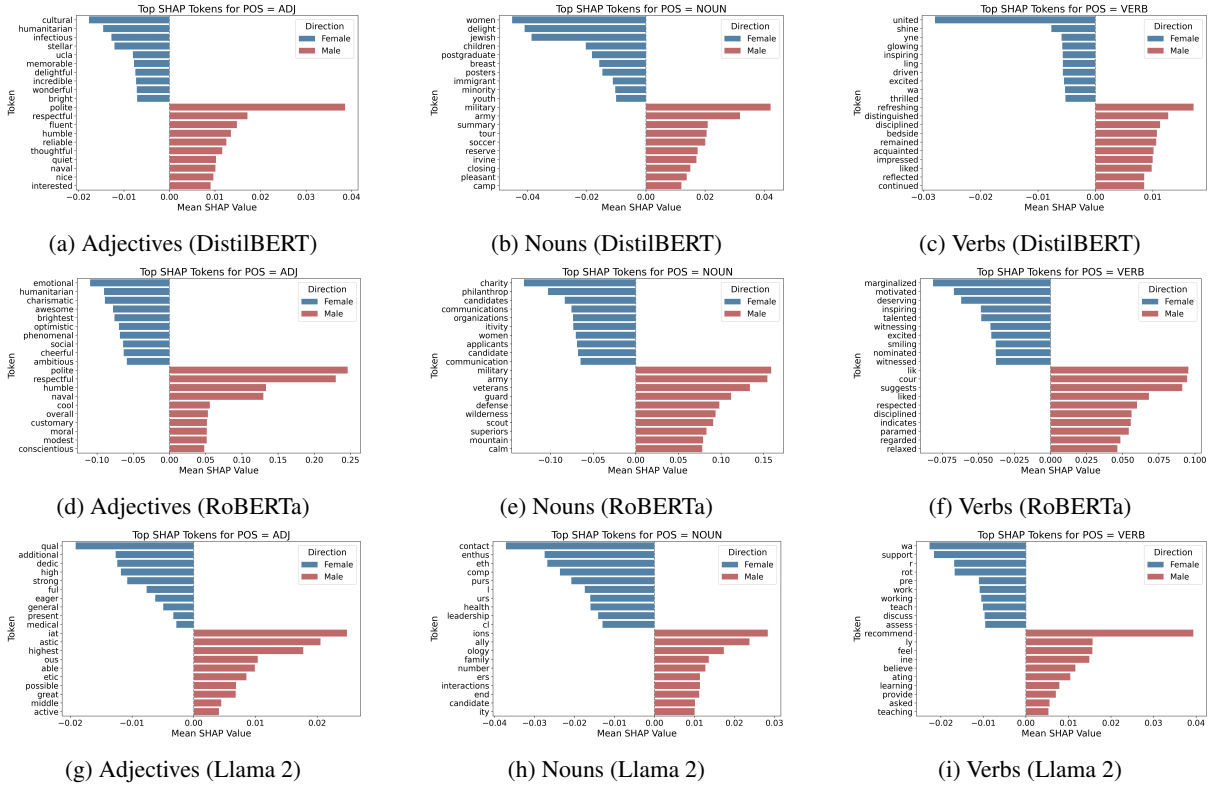
To better understand the sources of gender leakage within the de-gendered text, we used SHAP (Lundberg and Lee, 2017) values to interpret the model's predictions. SHAP assigns importance scores to individual tokens, allowing us to identify which words contributed most to the model's classification of an applicant as male or female. These influential tokens offer insight into the subtle linguistic cues the model relies on in the absence of explicit gender indicators. SHAP analysis was applied to all three of the models trained on our de-gendered dataset. For DistilBERT and RoBERTa, SHAP values were computed across the full set of recommendation letters. For Llama, we used a random sample of 100 letters due to the higher computational cost. We then extracted the tokens with the highest and lowest mean SHAP values across letters-associated with male and female predictions, respectively. These were grouped by part of speech to aid in our interpretation. To reduce noise, we only considered tokens that appeared at least 20 times across the dataset.

4.3.2 Interpretation using TF-IDF

In addition to model-based interpretation, we conducted a complementary analysis using TF-IDF. Rather than relying on model outputs, we applied the TF-IDF algorithm directly to the EDG dataset to identify terms that are most distinctive to male and female applicant letters, respectively. This approach provides a model-agnostic view of potential gendered patterns in the language used. We first aggregated all de-gendered recommendation letters for female applicants into a single document, and all de-gendered letters for male applicants into another. We then applied the TF-IDF algorithm to these two documents, computing a score for each token based on its frequency and uniqueness within its respective group. Tokens with higher TF-IDF scores are those that appear frequently and are particularly distinctive to one group, offering potential insight into the subtle linguistic patterns that may encode gender even after explicit identifiers have been neutralized. More details are provided in Appendix C.

4.4 Effect of Implicit Gender Tokens

After identifying implicitly gendered tokens using the TF-IDF and SHAP methods described in the previous sections, we selected the top 10 tokens for each part of speech (noun, verb, adjective), identified separately for each gender and for each method independently. This process produced two distinct datasets: one based on TF-IDF-selected tokens and another based on SHAP-selected tokens. In both cases, the selected tokens were altered in texts that had already been stripped of



(a) Adjectives (DistilBERT) (b) Nouns (DistilBERT) (c) Verbs (DistilBERT)

(d) Adjectives (RoBERTa) (e) Nouns (RoBERTa) (f) Verbs (RoBERTa)

(g) Adjectives (Llama 2) (h) Nouns (Llama 2) (i) Verbs (Llama 2)

Figure 2: Top 10 SHAP tokens with their corresponding values(+/-) for both genders (male/ female) grouped by part-of-speech category, identified by DistilBERT(last two encoder layers unfrozen), RoBERTa and Llama 2

Token	F → M Count	M → F Count	Absolute Difference
support	12	23.56	11.56
research	14	24.64	10.64
number	4	11.98	7.98
anesthesia	13	20.96	7.96
rotation	26	18.26	7.74
medicine	17	23.29	6.29
believe	8	14.2	6.2
leadership	3	9.02	6.02
professional	6	11.75	5.75
year	18	23.3	5.3

Table 1: Token-level gender prediction flips for TF-IDF derived tokens: comparison of top 10 tokens whose removal most frequently causes DistilBERT model prediction to flip from female to male or vice versa.

Token	F → M Count	M → F Count	Absolute Difference
liked	6	0.38	5.62
impressed	14	8.93	5.07
bedside	6	2.8	3.2
stellar	3	0.37	2.63
summary	1	3.16	2.16
thoughtful	4	1.97	2.03
acquainted	2	0	2.00
women	2	0	2.00
reliable	1	2.78	1.78
united	2	0.38	1.62

Table 2: Token-level gender prediction flips for SHAP derived tokens: comparison of top tokens whose removal most frequently causes DistilBERT model prediction to flip from female to male or vice versa.

explicit gender-identifying terms. For DistilBERT and RoBERTa, this involved substituting the tokens with the model's masked token; for Llama 2, which does not support masking, we used the unknown token instead. This intervention was applied consistently across both token sets, allowing us to evaluate the extent to which implicitly gendered language contributed to model performance, even in the absence of explicit gender cues. We then re-trained all three models on each of the new datasets (EDG w/o SHAP tokens and EDG w/o TF-IDF tokens).

4.4.1 Analyzing Prediction Flips from Token Removal

A deeper analysis of implicitly gendered tokens was conducted by focusing on a subset of recommendation letters that met two criteria: (1) they were correctly classified by the model trained on letters with only explicit gender-identifying tokens replaced (the EDG dataset), and (2) they were misclassified by the model trained on letters with both explicit and implicit tokens replaced. For this analysis, we used only the DistilBERT model due to its strong performance and computational efficiency.

This allowed us to examine how implicitly biased tokens influenced model predictions when neutralized.

For each such token, we counted how often replacing it with the model's masked token (or unknown token in the case of Llama) caused the model's prediction to flip from female to male or from male to female. A flip in prediction direction suggests that the replaced token may have carried meaningful gender-associated information that influenced the model's original prediction.

To mitigate the effects of class imbalance, we performed random sub-sampling of the majority class within this subset, repeated the token-flip counting process across multiple runs, and averaged the results. This approach provided a more balanced view of the influence of individual tokens. Table 1 & 2 presents the TF-IDF and SHAP–derived tokens along with their corresponding flip counts respectively.

5 Results & Discussion

From Table 4, we see that when explicit gender markers were preserved, the DistilBERT baseline achieved an almost perfect macro $F_1 > 0.95$, confirming that surface cues such as "*he*" or professional titles like "*Ms.*" virtually guarantee correct gender classification. After converting all explicit gender-identifying tokens to their female equivalents (the EDG dataset) and retraining, the same model architecture still obtained a macro F_1 score of 0.6 with an overall accuracy of 0.68 on a held-out test set, indicating that subtler linguistic patterns continue to signal applicant gender. Alternative models performed worse: RoBERTa reached a macro F_1 score of 0.547, and the considerably larger Llama 2 yielded a comparable F_1 score of 0.561 while incurring far greater computational cost.

The classification results of re-evaluation with their comparison to the model trained on the text with only the explicit gender identifying tokens replaced, are also shown in Table 4. For the DistilBERT model, replacing the implicitly gendered tokens identified via SHAP and TF-IDF with a masked token leads to a drop in macro F_1 score by $\sim 5\%$ and $\sim 2\%$ respectively. This decline highlights the extent to which these tokens contributed to the model's ability to predict applicant gender in the absence of explicit gender-identifying terms.

5.1 Replacing Explicit Gender Identifying Tokens

When we replace all explicit gender-identifying tokens in the recommendation letters with their female equivalents, the performance of our baseline classifier drops significantly. This confirms that our de-gendering process is effective and substantially limits the model's ability to infer applicant gender from overt cues. However, the fact that the macro F_1 score remains above random chance suggests that the model trained on de-gendered text can still identify subtle, implicit gender signals. In other words, even after neutralizing explicit gendered language, the way writers describe male and female applicants still carries implicit bias that the model can detect.

5.2 Implicit Gender Signals

Using our SHAP and TF-IDF analyses, we identify tokens that may carry implicit gender signals within the text. The top 10 SHAP values of adjectives, nouns and verbs for male and female candidates are shown in Figure 2a, Figure 2b, and Figure 2c respectively. Tokens more commonly associated with female recommendation letters include words like "*humanitarian*", "*delightful*", "*wonderful*", and "*children*". In contrast, tokens more frequently linked to male recommendation letters include "*respectful*", "*military*", "*combat*", and "*humble*". These patterns suggest subtle differences in how male and female applicants are described, even after explicit gender markers have been obscured.

5.3 Replacing Implicit Gender Identifying Tokens

Across all three models, replacing the implicitly gendered tokens identified by SHAP and TF-IDF with the corresponding model's masked token (or the unknown token in the case of Llama 2) resulted in drops in macro F_1 scores. This decline indicates that, even in the absence of explicit gender identifiers, the models trained on the EDG dataset relied heavily on implicit gender cues to make their predictions.

A closer analysis of the tokens whose replacement caused prediction flips further illustrates this point. For example, the token "*leadership*", when replaced, caused the model to flip its prediction from male to female 9 times, compared to just 3 flips in the opposite direction. This suggests that the presence of "*leadership*" strongly contributes to

Model	Gender	Precision	Recall	F_1	Acc.	Macro Precision	Macro Recall	Macro F_1	Wtd. Precision	Wtd. Recall	Wtd. F_1
DistilBERT	Female	0.481	0.394	0.432	0.681	0.615	0.604	**0.606**	0.665	0.684	0.673
	Male	0.750	0.815	0.781							
RoBERTa	Female	0.384	0.333	0.357	0.627	0.551	0.547	0.547	0.614	0.627	0.620
	Male	0.717	0.760	0.738							
Llama 2	Female	0.395	0.444	0.418	0.612	0.560	0.563	0.561	0.623	0.612	0.615
	Male	0.725	0.682	0.703							

Table 3: Comparison of DistilBERT, RoBERTa, and Llama 2 performance on test EDG(Explicitly De-Gendered) dataset including individual class metrics (Precision, Recall and F_1) and aggregate metrics(Macro + Weighted Precision, Recall and F_1).

Dataset	DistilBERT				RoBERTa				Llama 2			
	Acc.	Macro P	Macro R	Macro F_1	Acc.	Macro P	Macro R	Macro F_1	Acc.	Macro P	Macro R	Macro F_1
Original (non-EDG)	0.999	1.000	0.996	0.999	–	–	–	–	–	–	–	–
EDG (baseline)	0.68	0.615	0.604	0.60	0.627	0.551	0.547	0.547	0.612	0.560	0.563	0.561
EDG w/o SHAP Tokens	0.63	0.550	0.550	0.55 (↓ 5.0%)	0.633	0.533	0.525	0.521 (↓ 2.6%)	0.690	0.345	0.500	0.408 (↓ 15.3%)
EDG w/o TF-IDF Tokens	0.68	0.600	0.580	0.58 (↓ 2.0%)	0.661	0.563	0.541	0.535 (↓ 1.2%)	0.685	0.344	0.497	0.407 (↓ 15.4%)

Table 4: Comparison of DistilBERT, RoBERTa, and Llama 2 performance across three datasets: EDG (Explicitly De-Gendered), EDG with top SHAP-identified gender tokens removed, and EDG with top TF-IDF-identified gender tokens removed showing improvements(lowering) in macro precision, recall, and F_1 scores

the model associating the letter with a male applicant. Similarly, tokens like "*rotation*" and "*liked*" more often flipped predictions from female to male when replaced, indicating that their presence is more commonly associated with female applicants.

6 Conclusion

In this paper, we investigated the presence of gendered language in academic LoRs for a medical residency program. Despite the replacement of explicit gender identifiers, such as names and gendered pronouns, with their female equivalents, our results demonstrated that LLMs could achieve above-chance accuracy in applicant gender classification. Our TF-IDF and SHAP analysis showed that specific adjectives, verbs, and nouns served as implicit indicators of gender and contributed heavily to this performance, and classification performance dropped sharply when replacing them with the model's masked token (or the unknown token in the case of Llama 2).

These findings raise important questions about the role of AI in professional evaluation and recruitment. Dastin (2018) highlight the risk of using naïve strategies like erasing explicit cues of gender or other sensitive attributes to prevent bias in AI recruitment solutions, especially where training data already reflects historical bias in recruitment. Our work supports this concern by revealing persisting implicit cues of gender in LoRs.

Limitations

We discuss some broader concerns around degendering in Section 6. However, another limitation of our approach for neutralizing implicit gender cues is that the semantic integrity and evaluative content of letters may not be preserved. This approach replaces nouns, verbs, and adjectives with masked (or unknown) tokens; thus the resulting letters may contain incomplete and ungrammatical sentences. As such, this approach is more appropriate for machine evaluation of letters, as human evaluators may have difficulties understanding some parts of the letters.

Additionally, we found that SHAP occasionally highlights words that are split into subtokens during tokenization. Although SHAP applies heuristics to collapse these subtokens back into full words for easier interpretation, this collapsing does not always align precisely with how the model processes inputs internally. This misalignment can lead to cases where attribution scores are assigned to fragments rather than complete tokens, introducing some ambiguity into the interpretability analysis. We note this as a limitation of our approach.

More broadly, we recognize the limitations inherent in a purely technical approach to addressing bias in the evaluation and selection of candidates. Recruitment methods that attempt to neutralize *any* cues (explicit or implicit) for applicant attributes such as gender may trivialize candidate identities as attributes that can simply be switched off. This may prevent the consideration of important information about the experiences of candidates from marginal-

ized backgrounds, and ultimately may serve to bypass or outsource diversity, equity, and inclusion efforts that should instead occur within organizations (Drage and Mackereth, 2022; Tilmes, 2022).

In fact, our work highlights the challenge of truly removing traces of identities like gender in professional evaluations; even after neutralizing implicit cues of gender, our LLM-based gender classifiers achieve above-chance accuracy, suggesting that gender signals remain. Moreover, while neutralizing these cues may prevent gender identification from LoRs, this may also mask positive qualities and experiences of candidates. As such, while we believe that de-gendering strategies like those explored in this work may be one component of fairer AI-supported evaluation of LoRs (e.g., as a flag to reviewers to be aware of the amount of gendered language in a LoR), this should be paired with increased investment in human evaluators and institutional change, e.g., implicit bias training.

Ethical Considerations

This work analyzes academic letters of recommendation, which contain identifiable and possibly sensitive human subjects data. As such, the collection of these data was IRB-approved, and all applicant names were anonymized before analysis.

References

Charlotte S Alexander. 2022. Text mining for bias: A recommendation letter experiment. *American Business Law Journal*, 59(1):5–59.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *Preprint*, arXiv:2307.00101.

Eleanor Drage and Kerry Mackereth. 2022. Does ai debias recruitment? race, gender, and ai's "eradication of difference". *Philosophy & technology*, 35(4):89.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Owen Henkel, Libby Hills, Adam Boxer, Bill Roberts, and Zach Levonian. 2024. Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 300–304.

Louis Hickman, Patrick D Dunlop, and Jasper Leo Wolf. 2024. The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, 32(4):499–511.

Ann E Hoover, Tay Hack, Amber L Garcia, Wind Goodfriend, and Meara M Habashi. 2019. Powerless men and agentic women: Gender bias in hiring decisions. *Sex Roles*, 80(11):667–680.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nasif Imtiaz, Justin Middleton, Joymallya Chakraborty, Neill Robson, Gina Bai, and Emerson Murphy-Hill. 2019. Investigating the effects of gender bias on github. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 700–711. IEEE.

Carol Isaac, Barbara Lee, and Molly L Carnes. 2009. Interventions that affect gender bias in hiring: A systematic review. *Academic Medicine*, 84:1440–1446.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task.

Adam Karvonen and Samuel Marks. 2025. Robustly improving llm fairness in realistic settings via interpretability. *Preprint*, arXiv:2506.10922.

Steffen Keck and Wenjie Tang. 2020. When "decoy effect" meets gender bias: The role of choice set composition in hiring decisions. *Journal of Behavioral Decision Making*, 33(2):240–254.

Amanda J Koch, Susan D D'Mello, and Paul R Sackett. 2015. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of applied psychology*, 100(1):128.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Chee Wee Leong, Navaneeth Jawahar, Vinay Basheerabad, Torsten Wörtwein, Andrew Emerson, and Guy Sivan. 2024. Combining generative and discriminative ai for high-stakes interview practice. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pages 94–96.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Xinyue Li, Zhenpeng Chen, Jie M Zhang, Yiling Lou, Tianlin Li, Weisong Sun, Yang Liu, and Xuanzhe Liu. 2024b. Benchmarking bias in large language models during role-playing. *arXiv preprint arXiv:2411.00585*.

Emmy Liu, Michael Henry Tessler, Nicole Dubosh, Katherine Hiller, and Roger Levy. 2022. Assessing group-level gender bias in professional evaluations: The case of medical student end-of-shift feedback. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 86–93, Seattle, Washington. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Imran Mirza, Akbar Anbar Jafari, Cagri Ozcinar, and Gholamreza Anbarjafari. 2025. Quantifying gender bias in large language models using information-theoretic and statistical analysis. *Information*, 16(5).

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830.

Jane Phillips and Chet Robie. 2024. Hacking the perfect score on high-stakes personality assessments with generative ai. *Personality and Individual Differences*, 231:112840.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Lindsay Rice and Joan M. Barth. 2016. Hiring decisions: The effect of evaluator gender and gender stereotype characteristics on the evaluation of job applicants. *Gender Issues*, 33:1–21.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Vivian Simon, Neta Rabin, and Hila Chalutz-Ben Gal. 2023. Utilizing data driven methods to identify gender bias in linkedin profiles. *Information Processing & Management*, 60(5):103423.

Shweta Soundararajan and Sarah Jane Delany. 2024. Investigating gender bias in large language models

through text generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 410–424, Trento. Association for Computational Linguistics.

Nicholas Tilmes. 2022. Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Inf. Technol.*, 24(2).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. JobFair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Guojun Wu and Sarah Ebling. 2024. Investigating ableism in LLMs through multi-turn conversation. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 202–210, Miami, Florida, USA. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *Preprint*, arXiv:1809.01496.

A Hyper-parameter Optimization

To maximize classification performance on de-gendered data, we conducted a limited grid search over batch size $\in \{4, 16, 32\}$, learning rate $\in [10^{-5}, 3.7\times10^{-5}]$, and weight decay $\in \{0.0, 0.03\}$ for both DistilBERT and RoBERTa. For Llama 2, we additionally tuned LoRA parameters (r, α, dropout). Validation performance was monitored using macro-averaged F_1. The resulting best settings are reported in Tables 5, 6 and 7.

Parameter	**Value**
Batch size (train / eval)	32 / 32
Epochs	10
Learning rate	3.7×10^{-5}
Weight decay	0.03

Table 5: Best DistilBERT fine-tuning hyper-parameters

Parameter	**Value**
Batch size (train / eval)	16 / 16
Epochs	6
Learning rate	2.0×10^{-5}
Weight decay	0.0

Table 6: Best RoBERTa fine-tuning hyper-parameters

Parameter	**Value**
Batch size (train / eval)	4 / 4
Epochs	5
Learning rate	3.0×10^{-5}
Weight decay	0.02
LoRA rank (r)	16
LoRA alpha	48
LoRA dropout	0.15

Table 7: Best Llama 2 fine-tuning hyper-parameters

B Topic Modeling

B.1 Explainability via Topic Modeling

In addition to token-level interpretability methods, we explored topic modeling as a way to uncover higher-level thematic patterns that may reflect gendered language in recommendation letters. Using this approach, we identified recurring topics across the corpus and analyzed which ones were most predictive of each gender. This allowed us to examine broader narrative trends and associations, offering a complementary perspective to the more granular insights provided by token-level analysis.

B.2 BERTopic

We used BERTopic to identify topics across our recommendation letters (Grootendorst, 2022). Each letter was first split into individual sentences, which were then embedded and clustered based on semantic similarity. In total, BERTopic extracted 251 distinct topics, capturing a broad range of recurring themes. To create a topic-level representation for each letter, we mapped the identified topics of individual sentences back to their originating letter. This resulted in a binary topic vector for each letter, where each entry indicates the presence (1) or absence (0) of a specific topic.

B.3 Topic-Only Classification

We created train and test splits using the topic vectors for each letter and trained a random forest classifier to predict gender based on these topic representations. Model performance was evaluated on the held-out test set, with results summarized in Table 8.

To better understand which topics were most predictive of gender, we used SHAP to interpret the random forest model. Figure 3 displays the top contributing topics and their relative impact on the model's predictions. Topic labels on the left were generated by aggregating all sentences assigned to each topic into a single document, then applying TF-IDF to extract the five most distinctive unigrams and bigrams. These are separated by underscores for readability. The number at the start of each label corresponds to the topic ID assigned by BERTopic.

Each dot in the plot represents a single recommendation letter. The horizontal position of the dot reflects the SHAP value, which quantifies how much that topic influenced the model's prediction for that letter. Positive SHAP values indicate a push toward predicting "*male*", while negative values indicate a push toward "*female*". The dot color reflects the feature value: since our input features are binary topic vectors, red dots represent letters in which the topic was present, and their position indicates the strength and direction of its influence. Blue dots represent letters where the topic was absent.

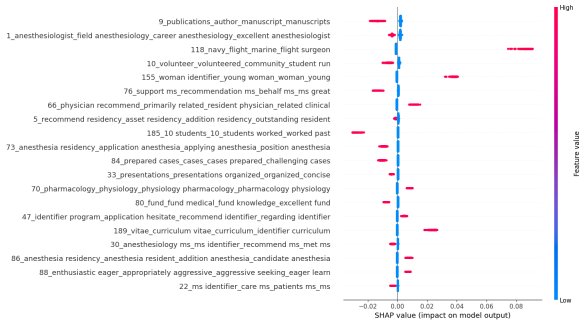


Figure 3: SHAP summary plot showing the most influential topics in predicting gender from recommendation letters using a Random Forest classifier trained on BERTopic vectors. Each row represents a topic, labeled with its BERTopic ID followed by the top five unigrams or bigrams (joined with underscores) extracted via TF-IDF. Each dot corresponds to a single letter; the horizontal position indicates the SHAP value, reflecting the topic's contribution to the prediction (positive values push toward "*male*" negative toward "*female*"). Red dots represent the presence of a topic in a letter, while blue dots represent its absence.

B.4 Topic and Text Embedding Infusion

In addition to using topic vectors alone to predict the applicant's gender, we also concatenated them with the contextualized embedding representations produced by our fine-tuned DistilBERT model during training. This fusion allowed the classifier to leverage both the high-level thematic structure captured through topic modeling and the nuanced contextual signals encoded by the transformer. We used an 80:10:10 train, validation, and test split, and trained the model using the same hyperparameters outlined in Table 5. Final evaluation results on the held-out test set are reported in Table 8.

C Analysis of TF-IDF Tokens

For each gender-specific document, the top 10 tokens were extracted for each part of speech (adjectives, verbs, and nouns). The top 10 adjectives for male and female letters are shown in Table 9 and Table 10, respectively. The top verbs are presented in Table 11 and Table 12, and the top nouns in Table 13 and Table 14.

Model	Gender	Precision	Recall	F_1	Acc.	Macro Precision	Macro Recall	Macro F_1	Wtd. Precision	Wtd. Recall	Wtd. F_1
Topic Vectors Only	Female	0.370	0.480	0.420	0.580	0.550	0.560	0.550	0.620	0.580	0.600
	Male	0.730	0.630	0.680							
Topic + DistilBERT Embeddings	Female	0.410	0.440	0.430	0.630	0.580	0.580	0.580	0.640	0.630	0.640
	Male	0.740	0.720	0.730							

Table 8: Classification performance of Random Forest models using (1) only topic vectors and (2) a combination of topic vectors and DistilBERT embeddings.

Token	Female	Male	Diff
calm	0.021404	0.031890	0.010486
young	0.030375	0.040406	0.010310
medical	0.609714	0.619496	0.009782
good	0.144768	0.154044	0.009275
professional	0.064924	0.074169	0.009246
ethic	0.055420	0.064632	0.009211
long	0.057019	0.065654	0.008635
internal	0.056397	0.064164	0.007766
appropriate	0.033128	0.040831	0.007703
respectful	0.016786	0.024482	0.007696
great	0.198324	0.205860	0.007536
able	0.143258	0.150765	0.007507
critical	0.059950	0.067016	0.007066
humble	0.014122	0.021033	0.006911
interested	0.029487	0.035552	0.006065
right	0.038190	0.044025	0.005834
willing	0.018029	0.023801	0.005771
inpatient	0.024513	0.030145	0.005632
happy	0.030908	0.036446	0.005538
hard	0.035348	0.040746	0.005398

Table 9: Male Adjectives

Token	Female	Male	Diff
identifi	0.404463	0.382980	-0.021483
outstanding	0.126650	0.111211	-0.015439
clinical	0.320622	0.308854	-0.011768
pediatric	0.043608	0.033721	-0.009887
public	0.013766	0.007025	-0.006741
global	0.011546	0.005663	-0.005883
identifier	0.053733	0.048580	-0.005153
numerous	0.028154	0.023204	-0.004950
new	0.057019	0.052242	-0.004777
social	0.017408	0.012731	-0.004677
future	0.051513	0.046962	-0.004550
compassionate	0.036858	0.032529	-0.004329
academic	0.096187	0.092222	-0.003965
pre	0.018385	0.014434	-0.003951
bright	0.035260	0.031337	-0.003923
efficient	0.015187	0.011496	-0.003692
competitive	0.013056	0.009367	-0.003689
specific	0.015099	0.011751	-0.003347
warm	0.016075	0.012731	-0.003345
fantastic	0.013322	0.010048	-0.003274

Table 10: Female Adjectives

Token	Female	Male	Diff
show	0.105123	0.121443	0.016321
like	0.046116	0.058660	0.012544
feel	0.104389	0.114893	0.010504
believe	0.092022	0.100338	0.008316
learn	0.221565	0.228380	0.006815
display	0.034482	0.041047	0.006565
know	0.153859	0.160210	0.006351
spend	0.071375	0.077291	0.005916
supervise	0.021800	0.026783	0.004982
build	0.016140	0.020863	0.004723
enjoy	0.046221	0.050800	0.004579
ask	0.116652	0.120910	0.004258
benefit	0.005764	0.009801	0.004036
maintain	0.021171	0.024890	0.003719
begin	0.022534	0.026152	0.003618
attend	0.070117	0.073652	0.003535
write	0.226701	0.230223	0.003523
read	0.049574	0.052934	0.003360
answer	0.017713	0.021009	0.003296
require	0.049889	0.053128	0.003240

Table 11: Male Verbs

Token	Female	Male	Diff
take	0.133840	0.123918	-0.009922
complete	0.102293	0.093642	-0.008651
stand	0.038150	0.030276	-0.007874
support	0.052719	0.044977	-0.007741
organize	0.035530	0.027947	-0.007583
present	0.078292	0.071420	-0.006872
excel	0.065715	0.059096	-0.006619
match	0.034377	0.028432	-0.005945
care	0.054291	0.048568	-0.005723
recruit	0.029032	0.023483	-0.005549
include	0.115289	0.109896	-0.005393
hope	0.038779	0.033430	-0.005349
shadow	0.017608	0.012275	-0.005332
evaluate	0.028403	0.023144	-0.005259
try	0.021591	0.016448	-0.005143
run	0.020857	0.015963	-0.004894
waive	0.081960	0.077291	-0.004669
manage	0.036683	0.032217	-0.004466
encourage	0.014044	0.009655	-0.004389
reach	0.023896	0.019699	-0.004198

Table 12: Female Verbs

Token	Female	Male	Diff
medicine	0.151597	0.165628	0.014031
staff	0.047931	0.059773	0.011843
knowledge	0.114200	0.125307	0.011107
physician	0.063732	0.072574	0.008842
practice	0.024229	0.031521	0.007292
time	0.136849	0.143022	0.006173
anesthesia	0.105677	0.111843	0.006166
year	0.201395	0.207183	0.005789
number	0.028825	0.034538	0.005713
rotation	0.166823	0.172302	0.005479
demeanor	0.014796	0.020069	0.005273
base	0.019392	0.024618	0.005225
residency	0.225863	0.230636	0.004773
training	0.079150	0.083820	0.004670
week	0.046255	0.050836	0.004581
discussion	0.018770	0.023269	0.004499
question	0.089732	0.094151	0.004419
resident	0.153272	0.157651	0.004378
topic	0.017046	0.020983	0.003937
anesthesiologist	0.059422	0.063293	0.003871

Table 13: Male Nouns

Token	Female	Male	Diff
health	0.047691	0.032527	-0.015164
research	0.121909	0.108163	-0.013746
identifier	0.356199	0.342753	-0.013446
student	0.314541	0.305929	-0.008612
applicant	0.024851	0.019063	-0.005788
surgery	0.063780	0.058059	-0.005721
education	0.034571	0.029075	-0.005496
community	0.034380	0.029189	-0.005190
leadership	0.032991	0.028001	-0.004990
patient	0.342936	0.338182	-0.004754
child	0.014604	0.010012	-0.004592
clerkship	0.050947	0.046424	-0.004523
meeting	0.019105	0.014629	-0.004476
study	0.027293	0.022995	-0.004298
skill	0.141541	0.137261	-0.004280
passion	0.022792	0.018515	-0.004277
department	0.053820	0.049762	-0.004059
team	0.142882	0.138999	-0.003883
care	0.164429	0.160873	-0.003556
project	0.052240	0.048733	-0.003507

Table 14: Female Nouns

D Computational Frameworks

We used several standard software libraries in our experiments. All model loading and tokenization were performed using the HuggingFace Transformers library (Wolf et al., 2020), and model fine-tuning was carried out using PyTorch (Paszke et al., 2019). For evaluation, we used scikit-learn (Pedregosa et al., 2011), including its implementations of standard metrics such as accuracy and F_1 score. For model interpretability, we applied SHAP (Lundberg and Lee, 2017) to compute token-level attributions.

E Artifact Licenses

Our experiments leverage the **Llama2-7B-chat** large language model (Touvron et al., 2023a). This model was accessed via the Hugging Face `transformers` library and is distributed under the Llama 2 Community License. We acknowledge and adhere to the terms of this license for our research. Further details on Llama2 are available at Meta AI's official Llama webpage.