A Fairness Analysis of Human and AI-Generated Student Reflection Summaries

Bhiman Kumar Baghel[1], **Arun Balajiee Lekshmi Narayanan**[2] and **Michael Miller Yoder**[1]
[1]Department of Computer Science, [2]Intelligent Systems Program
University of Pittsburgh, PA, USA
{bkb45, arl122, mmy29}@pitt.edu

Abstract

This study examines the fairness of human- and AI-generated summaries of student reflections in university STEM classes, focusing on potential gender biases. Using topic modeling, we first identify topics that are more prevalent in reflections from female students and others that are more common among male students. We then analyze whether human and AI-generated summaries reflect the concerns of students of any particular gender over others. Our analysis reveals that though human-generated and extractive AI summarization techniques do not show a clear bias, abstractive AI-generated summaries exhibit a bias towards male students. Pedagogical themes are over-represented from male reflections in these summaries, while concept-specific topics are under-represented from female reflections. This research contributes to a deeper understanding of AI-generated bias in educational contexts, highlighting the need for future work on mitigating these biases.

1 Introduction

Reflection is an effective metacognitive technique that promotes student learning (Baird et al., 1991; McNamara, 2011). Reflections can be used in a classroom setting to gather feedback from students on their comprehension and help both students and instructors identify topics of confusion. Given the substantial amount of reflection data in large classes, AI-based summarization techniques have been developed to summarize these reflections (Fan et al., 2015; Luo and Litman, 2015; Luo et al., 2016; Magooda and Litman, 2020). Automatic summarization (Hovy et al., 2006) is a popular NLP technique used to create or sample a smaller text that represents the most important or relevant information within the original content. This process inevitably involves decisions about which is the most important or relevant information.

AI bias is a well-discussed topic in recent years. Efforts to identify and mitigate bias in AI and NLP systems have been applied to tasks such as language modeling (Bolukbasi et al., 2016; Caliskan et al., 2017; Sun et al., 2019; Huang et al., 2020; Czarnowska et al., 2021; Field et al., 2021), coreference resolution (Rudinger et al., 2018; Cao and Daumé III, 2020), and machine translation (Savoldi et al., 2021). Specifically within NLP research for education, bias has been investigated in educational technologies like automated essay scoring (Amorim et al., 2018; Litman et al., 2021) and intelligent tutoring systems (Zhuhadar et al., 2016; Lin et al., 2023).

Reflection summarization is an important use case as it help instructors uncover student misconceptions, empowering them to adapt their instruction and create targeted learning opportunities that address knowledge gaps in subsequent lectures (Fan et al., 2017). Since the goal of reflection summarization is to save teaching staff time and reduce the need to read through so many reflections, biases in whose reflections are represented by the summaries can have a direct impact on whose concerns are addressed by teaching staff. This concern motivates our study to measure biases in summarization of student reflections.

Specifically, we scope our research to identifying if there are differences by student gender in a dataset of classroom reflections and if the summaries of these reflections exhibit bias toward any gender. We are particularly interested in representation from female reflections due to a history of exclusion of women in STEM classes (Brotman and Moore, 2008; Vincent-Ruz and Schunn, 2018). Unfortunately, we are only able to compare the representation of reflections students with those of male students within a gender binary, as we do not have sufficient data on the experiences on non-binary students, an important topic for future work.

Using the Structural Topic Model (STM; Roberts

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 60–77
August 16, 2024 ©2024 Association for Computational Linguistics

et al., 2014), we are able to model variation in topics within reflections along with the gender of the authors of these reflections. We also apply STM to measure how closely topics in summaries match those in reflections from male and female genders. We evaluate gender bias in several types of AI-generated summaries and contrast these with human-annotated summaries. We define our research questions as follows:

RQ1 What differences, if any, are there between reflections from male or female students?

RQ2 Are summaries biased towards any specific gender?

RQ3 If so, what is the nature of the gender bias in reflection summaries?

Using STM, we find subtle differences between reflections of male and female students, particularly a stronger emphasis on course logistics (such as projects) from female students. Measuring differences between summary topic distributions and those of male and female reflections, we find that AI abstractive summarization models exhibit bias toward reflections from male students, while summaries from humans and AI extractive models do not show a consistent bias. We find that AI abstractive summaries appear to under-represent specific course concepts that are brought up in reflections from female students, while over-representing pedagogical themes such as teamwork from male student reflections.

2 Related Work

Reflection Sumarization: We first review work on automatic summarization in the context of student reflections, the application area in which we investigate bias. Fan et al. (2015) and Zhong et al. (2024) independently observed that reflections can range from some phrases and sentences to multiple sentences. Luo and Litman (2015) argued that phrase-based summarizing is the most effective way to summarize student reflections as they are easy to read and browse as compared to abstractive or extractive summarization. They also introduced a notion of student coverage that gave importance to topics mentioned by most of the students. With these two motivations, they propose a student coverage-assisted phrase-based summarization algorithm.

Luo et al. (2016) improves upon the previous work by evaluating the phrases in their informativeness and alignment with the needs of the students. Magooda and Litman (2020) proposed a template-based data generation technique which, when used for training models, increases the model performance for abstractive summarization for low-resource data. We evaluate several of these summarization approaches for gender bias.

Bias in educational AI: A growing body of research has examined issues of bias and social justice in educational technologies. Shakir et al. (2022) discuss the relationships between intersectionality and student perspectives in academia, using simple but effective text mining approaches such as clustering that assists the qualitative analysis of the data. Roscoe et al. (2022), Madaio et al. (2022), and Baker and Hawn (2022) independently discuss the possibilities of injustices with and the development of fair AI systems in education. Dias et al. (2022) consider the need to take intersectionality into account when designing automated decision-making systems in computing education As discussed in Mayfield et al. (2019), there are potential improvements possible towards countermeasures for inherent biases in automated education assessment systems. Litman et al. (2021) conduct fairness evaluation of Automated Essay Scoring (AES) used for grading essays. They concluded that different AES models exhibit different types of biases, spanning students' gender, race, and socioeconomic status.

Bias in summarization: Huang et al. (2023) examined bias in opinion summarization through the perspective of opinion diversity. This work is analogous to ours, as biases in summaries of online reviews relate to student reflections as "reviews" of the course material. Like our work, they also generate a summary of the source texts. However, unlike us, an overall stance score was relevant, and they had access to pre-computed topic-specific tweet clusters that are utilized in combination with the opinion diversity / similarity to finally detect the stance taken by the source text or document.

Dash et al. (2019) showed that existing summarization algorithms often represent socially salient user groups very differently compared to their distributions in the original data. In our work, we focus on the salient differences in the topic distribution by student gender. Liu et al. (2024) develop methods to explicitly preserve author perspective ("bias") in news summarization.

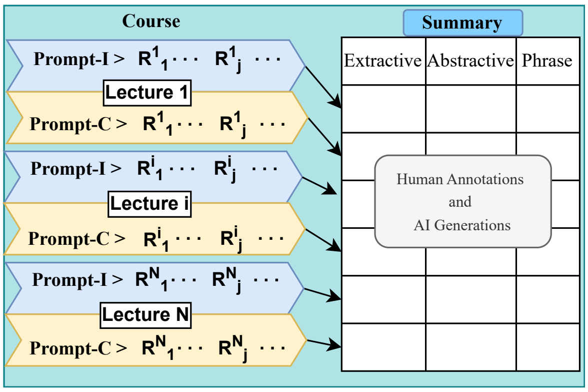
Course
Summary
Extractive
Abstractive
Phrase
Prompt-I > R^1_1 ··· R^1_j ···
Lecture 1
Prompt-C > R^1_1 ··· R^1_j ···
Prompt-I > R^i_1 ··· R^i_j ···
Lecture i
Prompt-C > R^i_1 ··· R^i_j ···
Prompt-I > R^N_1 ··· R^N_j ···
Lecture N
Prompt-C > R^N_1 ··· R^N_j ···
Human Annotations
and
AI Generations

Figure 1: ReflectSumm Structure. Each lecture of a course has two prompts (I & C) asking interesting and confusing things of the lecture (see section 3) from students. Provided reflections are summarized by human annotators and AI techniques.

STM for Bias Analysis: Structural Topic Modeling (STM; Roberts et al., 2014) has been used before to analyze text discourse with the goal of identifying biases with author metadata. In their work, Davidson and Bhattacharya (2020) use an STM approach to examine racial biases in a Twitter dataset. They are able to identify the interaction between prevalence of tweets with respect to the abusive nature of the tweets, and helps them identify biases with topic modeling by taking a multi–dimensional approach.

In another work, Zhang and Rayz (2022) examine the stereotypes embedded within the text of news articles using STM. Using a similar approach as Davidson and Bhattacharya (2020), the authors examine the gender stereotypes within the text across three dimensions: weak, medium and strong associations in interaction with male and female gender. STM allows them to discuss their results in terms of the detailed interaction between the two dimensions, whereby they can suggest conclusions such as "International Politics" that are historically in the "male" sphere of discussions being associated with the topics of articles written by male authors, while topics on "Music" being associated with articles written by female authors.

Villamor Martin et al. (2023) present a more meta–analytic approach to the use of STM in the context of identifying or detecting historical biases or stereotypes in the data. Since STM is a statistical, data–driven approach, the signals from the data indicate the general trend of associations of aspects such as demographic identities with topics in the text.

Similar to this prior work, we choose STM to identify biases in the discourse analysis with a dataset collected in an educational setting, associating topics with the binary gender of the author of student reflections.

3 Dataset Description

We selected ReflectSumm (Zhong et al., 2024), a benchmarking dataset for student reflection summarization, for analysis, since it contains student reflections, their summaries and student demographic information. It collects reflection and demographic information through the CourseMirror Application (Fan et al., 2015). The application prompted students with two types of reflection prompts: *Describe what you found most interesting in today's class (I)* and *Describe what was confusing or needed more details in today's class (C)*. Students who opted into the study have to answer these prompts at the end of every lecture. In this manner reflections were collected over the course of four semesters, from Fall 2020 to Spring 2022 in two large, public, American universities. Broadly, students who participated in this experiment were enrolled in courses belonging to 4 subject areas: *Computer Science (CS), Engineering (ENGR), Physics (PHY), and Computing Information (CMPINF)*. Demographic information was collected from students at the time of registering for the experiment. Table 1 shows the proportion of reflections across genders for each course. Due to insufficient data on non-binary and self-described genders, we performed our analysis within a gender binary of male and female and leave further analysis on reflections from non-binary students to future work. Other demographic information like race, ethnicity were also collected. This work focuses on gender, however, our methodology can be directly applied to other demographic information as well, another possibility for future work.

We compare bias among human- and AI-generated summaries. The human annotations were collected by Fan et al. (2015) and Zhong et al. (2024) by employing college students of appropriate subject background. We evaluate automatic summaries generated by Zhong et al. (2024) using various AI techniques ranging from classic machine learning to deep learning-based generative AI. Some of these models were also trained on human annotations.

Summaries were annotated or generated for each

Course	**Gender**				**#Reflections**	**#Lecture**
	Male	**Female**	**Prefer Not to Diclose**	**Prefer to Self Describe**		
CS	1526 (57.71%)	1178 (42.23%)	42 (1.5%)	43 (1.54%)	2789 (I:1434 C:1355)	79
ENGR	2330 (65.21%)	1155 (32.32%)	88 (2.4%)	0	3573 (I:1861 C:1714)	62
CMPINF	1272 (60%)	762 (35.96%)	52 (2.45%)	33 (1.55%)	2119 (I:1080 C:1068)	19
PHYS	5071 (47.49%)	5898 (53.11%)	129 (1.16%)	0	11098 (I:5618 C:5484)	57

Table 1: Reflection Distribution Across Genders

reflection prompt across all lectures, to mimic a scenario where teaching staff would like to view summaries of reflections for single lectures. The structure of the dataset can be viewed in Figure 1, where a course has multiple lectures with exactly two reflection prompts, I and C. Each prompt has multiple student reflections which are summarized. So each lecture has two summaries corresponding to each reflection prompt. For both human annotation and AI generation, three types of summaries were annotated or generated: extractive, phrase-level extractive, and abstractive. While creating human annotations, annotators were asked to extract five reflections and five phrases that best represent all student reflections for extractive and phrase-level extractive summaries respectively. They were also asked to write an abstractive summary to summarize the major points of student reflections.

In the case of automatic summarizing, we evaluate a selection of models presented by Zhong et al. (2024), including those that are fine-tuned on human annotations as well as those that use causal language models like ChatGPT in a zero- or few-shot setting[1]. Among these, we have selected the two best performing approaches, from findings by Zhong et al. (2024), to collect summaries for each summary type:

1. Extractive summary: MatchSum (Zhong et al., 2020) and GPT-reflect (Zhong et al., 2024). MatchSum uses a re-ranker, and follows a two-stage paradigm to achieve state-of-the-art extractive summarization. GPT-reflect uses ChatGPT (GPT-3.5 turbo) to generate zero-shot extractive summaries from reflections.

2. Abstractive summary: BART-Large (Lewis et al., 2020) and GPT-1-shot (Zhong et al., 2024). BART-large was fine tuned on human annotations. GPT-1-shot uses ChatGPT (GPT-3.5 turbo) prompted with a random summary and corresponding reflections set from the human annotations.

3. Phrase-level extractive summary: GPT-noun (Zhong et al., 2024) and GPT-noun-1-shot (Zhong et al., 2024). Both use ChatGPT (GPT-3.5 turbo) to extract 5 noun phrases from provided reflections where former is zero-shot and later is one-shot.

For our analysis, we remove reflections by students who do not disclose gender information, leave it blank, or self-describe their gender. We hope to analyze non-binary genders in future work with more data available. REFLECTSUMM has annotations and summarizations for all student reflections, including those who do not provide demographic information. We considered summaries where at least 80% of the reflections they summarize are from students who indicated male and female gender. This gave us a collection of 250 summaries.

4 Analysis Methodology

Our aim is to analyse any gender bias present in reflection summaries. In order to achieve this goal we apply topic modeling. Topic modeling learns a distribution over a set of topics for a given text document in an unsupervised fashion. We aim to capture what topics are reflected in summaries and measure their variance according to document metadata, in our case the gender of students who wrote the reflection. STM is designed for just this: to associate topics with document metadata. STM brings out the latent topics in a corpus of text and allows the use of additional covariates to alter the prior distribution used to estimate the latent topics better. This feature sets apart STM from other topic models like Latent Dirichlet Allocation (LDA) and BERTopic (Grootendorst, 2022).

We first used topic modeling to learn the topics present in the student reflections we have collected and provide the topic distribution for each reflection. We analyzed this topic distributions of reflection across genders to address **RQ1** and explore differences between reflections from male and female students.

[1] Details about annotation guidelines, model training and generation prompts are provide in Zhong et al. (2024).

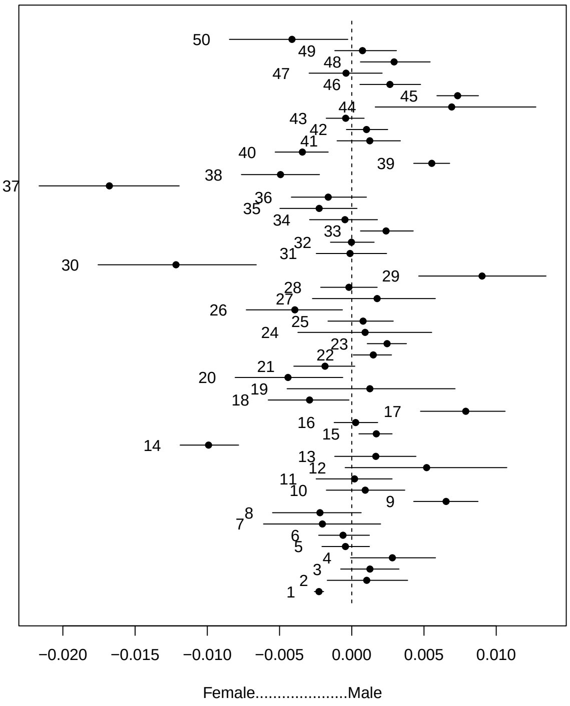

Figure 2: Topical difference between Genders

Topic 37: interest, found, thought, cool - Gender: Female
1: I found the derivations the most interesting part of today's class.
2: I found it most interesting looking at the enzymes in action in the video we had to watch. It was cool to see the stains disappear.
Topic 30: learn, engin, failur, super - Gender: Female
1: I found it interesting to see how engineering errors have caused major problems. I think that it is important for students to learn about how ethics and preventative measures should be taken into consideration when starting to design a project.
2: The most interesting I learned in class today was that various companies in the past tried to name themselves to be at the top of the alphabet.
Topic 14: question, top, hat, breakout- Gender: Female
1: the second conceptual top hat question
2: the second to last top hat question
Topic 38: project, new, design, present - Gender: Female
1: How presentations will work- will the final presentation be recorded?
2: the design project and scoping out a location for our problem
Topic 29: also, know, frequenc, didnt, unclear - Gender: Male
1: Today, it was confusing knowing how to interpret the frequency, wavelength, and time from the sinusoidal equations. It was also a bit unclear how to know nodes vs antinodes.
Topic 17: valu, determin, flux, compar - Gender: Male
1: The picture representing high k value and low k value
2: key and none key application.
Topic 45: one, exact, anoth, constant - Gender: Male
1: The color bands caused by thin films.
2: It was interesting that intervention can cause more harm than good. Another interesting thing would be the commons not working out due to human negligence.
Topic 44: current, direct, move, wire, electron - Gender: Male
1: What sort of chemical reactions happen in the batteries, and how does that lead to a moving current.
2: I found it confusing that both current density and electric field are in the opposite direction of the flow of electrons.

Table 2: Top reflections for four most associated topics with reflections of each gender

To address **RQ2** on gender bias in summaries, we apply our learned topic model, trained to represent topics in the reflections, to estimate a topic distribution within summaries. We then compute the distance between the topic distribution of summaries and the reflections they summarize in corresponding lectures.

To address **RQ3** on the nature of any bias in summaries, we examine topics in the summaries that are over-represented from the reflections of some groups while under-represented from the reflections of other groups. In order to identify the nature of bias, we compute a discrepancy measure between the topic probabilities in summaries and the genders involved.

5 RQ1: What differences, if any, are there between reflections from male or female students?

5.1 Learning the Topic Model

We trained STM model using its implementation in R (Roberts et al., 2019). It takes documents (individual student reflection), metadata of interest (gender) and number of topics as input. Along with gender, we have also provided course name and the prompt type (I or C) as metadata to control for their possible confounding influence. We allow the topic prevalence to vary by gender, type of prompt, and course name, as well as interactions among these covariates. Interesting (I) and confusing (C) are included as covariates since they also affect the content of the reflections and could act as confounding variables. Course is added to control for potential confounding effects of having different gender distributions in different courses. We used the approach of (Mimno and Lee, 2014) to select the optimal number of topics for this corpus (built-in to the R implementation). We choose number of topics as the mean ten runs of this approach (50) and then trained the STM model using it.

5.2 Analysis and Results

STM, as in LDA, represents topics as a probability distribution over words and documents as a probability distribution over topics. Figure 5 (Appendix A.1) shows the topics identified by our learned topic model, sorted according to highest proportion in the documents. To help characterize these topics, the top four words with the highest FREX score (Roberts et al., 2019) are presented. Pre-processing is performed before topic modeling by stemming these words to reduce the sparsity of the vocabulary. For example, the top words for *Topic 44: current, direct, move, wire, electron* seem to relate to electric current, which is a concept in the physics subject.

We have learned our topic model on student reflections and we have an intuition of what topics are present in those reflections. We can examine this topic model to address **RQ1** on differences between reflections from male and female students. STM provides a tool to estimate a regression model predicting the learned topic proportion from document metadata, which we use to examine the association between topics and particular genders. Associations between gender and the prevalence of learned topics are presented in Fig. 2. Topics that are further left in this figure are more inclined towards female. For example, *Topic 37: interest, found, thought, cool* and *Topic 30: learn, engin, love, failur* is associated with female student reflections. Similarly, topics shown further right are more inclined towards male. For example, *Topic 29: also, know, frequenc, didnt, unclear* and *Topic 17: valu, determin, flux, compar* are associated with male student reflections. Topics which are at around the center, near zero value, are not strongly associated with any particular gender. With this analysis we confirm that there are gender specific topical differences in student reflection because most of the topics are either side of the zero-center line and there exist topics which are at extreme left or right of the graph.[2]

To characterize differences in topics strongly associated with male and female student reflections, we examine their top words and highly probable documents. Highly probable documents, i.e. student reflections, for four most associated topics with each gender are shown in Table 2. This analysis will provide us with better insight into the topics and the contexts in which they appear.

Overall, we found only subtle differences in male and female reflections in terms of their ways of answering prompts and in different focuses of concern. Along with rather trivial differences in how female students answered the questions (including elements of the prompt), reflections from female students were more likely to emphasize the logistics of courses, such as projects and presentations. Reflections from male students brought up being unclear, but largely focused on specific course concepts.

The topic most strongly associated with reflections from female students, *Topic 37: interest, found, thought, cool* conveys that female students tended to explicitly use the words ‘found’ and ‘interesting’ to react to lectures. This could indicate relating their learning to themselves, but more practically indicates being more likely to copy parts of the prompt (I) in their reflections (e.g., “I found it interesting that...”). The second-most female-oriented topic, *Topic 30: learn, engin, failur, super*, also shows this tendency toward explicitly noting their own learning (or simply responding to the prompt in complete sentences). The content focus of this topic was on engineering failures that students found interesting.

In contrast, top words and documents for most male-oriented topics seem directly related to course concepts, such as frequency of waves (physics) shown in *Topic 29: also, know, frequenc, didnt, unclear*. Similarly *Topic 44: current, direct, move, wire, electron* refers to electric currents, another concept in physics. They also mention being unclear about those concepts. Examining topic associations with the prompt being interesting or confusing (see Appendix A.2, Fig. 6), we see a slight tendency for topics associated with male reflections to be associated with the confusing prompt (especially Topic 44), whereas topics associated with female reflections are more evenly balanced in their associations with both prompts (I and C).

6 RQ2: Are summaries biased towards any specific gender?

To measure how closely summaries represent the reflections of male or female students, we estimate the distance in topics captured in summaries from those presented in reflections from both genders.

6.1 Computing Summary and Reflection Distance

To see how representative summaries are of topics brought up in male and female reflections, we estimate topic distributions for summaries and calculate distances between topic distributions for summaries and reflections from both genders.

[2] We also plot topic gender associations mediated by prompt type (I or C), available in the Appendix A.2. These have similar topic associations as Fig. 2.

Distance Metric	Human Annotation			AI Generation					
	Extractive	Abstractive	Phrase	Extractive		Abstractive		Phrase	
cosine difference	**F***	**F***	F	(MatchSum)	**F***	(BART-large)	**M***	(GPT-noun)	F
				(GPT-reflect)	**F***	(GPT-1shot)	**M***	(GPT-noun-1shot)	**M***
jsd	F	F	F	(MatchSum)	F	(BART-large)	**M***	(GPT-noun)	M
				(GPT-reflect)	F	(GPT-1shot)	**M***	(GPT-noun-1shot)	**M***
hellinger	F	F	F	(MatchSum)	F	(BART-large)	**M***	(GPT-noun)	F
				(GPT-reflect)	**F***	(GPT-1shot)	**M***	(GPT-noun-1shot)	M
earthmover	**M***	**M***	**M***	(MatchSum)	**M***	(BART-large)	**M***	(GPT-noun)	**M***
				(GPT-reflect)	**M***	(GPT-1shot)	**M***	(GPT-noun-1shot)	**M***

Table 3: Inclination of Summary towards Gender. If the average distance of male reflections from the corresponding summary is less than the distance to female reflections, this is marked as M. Otherwise, it is marked F. **M*** or **F*** indicate the differences were significant by a *t*-test $p < 0.05$, hence biased towards that gender. Appendix A.3.

First, we infer topic distributions present in summaries for both human annotations and AI generations using our topic model learned from student reflections. To describe this process more formally, let S^i be the topic distribution in dimension T (the number of topics) for summary i and $R^i = R^i_1, ..., R^i_j, ...$ be a list of topic distributions in T for reflections associated with S^i. R^i is also a collection of male and female reflection topic distributions which can be denoted as $R^i = R^i_M + R^i_F$. Here R^i_M is a list of R^i_j reflection topic distributions where j belongs to male. Similarly, R^i_F is a list of reflection topic distributions for female.

To inform our analysis of potential bias (**RQ2**), we aim to calculate how close each summary's topic distribution is to the topic distributions of different genders. Ideally summaries would represent topics present in reflections from both male and female students equally. A summary's closeness to a particular gender's reflection with respect to other genders would indicate bias towards that gender. To analyse this closeness we computed the distance $D^i_M = S^i - R^i_M$ and $D^i_F = S^i - R^i_F$[3]. Similarly, distances are calculated for all $i \epsilon N$ summaries from their matched reflections, where N is the total number of summaries (250 as discussed in section 3). An average of these distances is calculated for each gender as $AvgD_M = \sum_{i=1}^{N} D^i_M / N_{DM}$ and $AvgD_F = \sum_{i=1}^{N} D^i_F / N_{DF}$, where N_{DM} are count of distances (D^i_M) for male reflections and similarly N_{DF} for female reflections. $AvgD_{M/F}$ signifies the average distance between summary topic distributions and their corresponding reflection topic distributions as per gender. A smaller value among these two averages would indicate summaries on average being closer to the gender with lower average distance.

6.2 Analysis and Results

We evaluate distances between summaries and reflections across four different distance metrics to see if any such differences we find are robust to the choice of metric. We select metrics that are symmetric and commonly used to measure distance across probability distributions such as our topic distributions (Chung et al., 1989). We apply the following four distance metrics - (1) Cosine difference (1 - cosine similarity), (2) Jensen-Shannon Divergence (jsd), (3) Hellinger Distance and (4) Earth Mover's Distance. We calculated the average distances for human annotations and AI generations across all three summary types using the previously described procedure for both genders. Table 3 shows a comprehensive view of our experiment results. Here, the value of each cell is the result of comparison between $AvgD_M$ and $AvgD_F$. If $AvgD_M < AvgD_F$, then the summaries on average are closer to male reflections, which is signified as 'M'. If the above condition is not true, then the summaries on average are closer to male reflections, which is signified as 'F'.

In order to check the significance of the mean distances we find from summaries to male and female reflections, we drawn out 1000 completely random samples $RandomD_M$ and $RandomD_F$ from a concatenated list of $D^1_M + ... + D^i_M + ... + D^n_M$ and $D^1_F + ... + D^i_F + ... + D^n_F$ respectively for each gender. We performed a Student's t-test with $RandomD_M$ and $RandomD_F$ and identified the human annotation and AI generations whose p-value is < 0.05. This signifies that those summaries are significantly skewed toward one gender over another. The significant ones are marked as '**M***' or '**F***' (considering the closeness result as

[3] D^i_M and D^i_F are list of distances between summary's topic distribution and specific gender's reflection topic distribution.

mentioned above) in Table 3.

Our experimentation results (Table 3) shows that results are mixed and inconclusive across distance measures for human annotations. When we shift our focus towards AI generations, we see different result patterns across all summary types.

Starting with the most consistent one, we see AI models generating abstractive summaries are consistently biased towards male reflections. BART-large (Zhong et al., 2024) was fine tuned on human annotations where as GPT-1shot (Zhong et al., 2024) was provided with a random summary and corresponding reflections set from the human annotations. This results contrast with the result for human annotations which were mixed.

However, in the case of AI models generating extractive summaries - MatchSum (Zhong et al., 2020) which is also trained on human annotations, results are also mixed. Another extractive summarization AI model examined is GPT-reflect. Although, it has no relation with human annotations, it follows the similar pattern except for Hellinger distance. For phrase-based extractive summarization models, GPT-noun and GPT-noun-1shot (Zhong et al., 2024) are similar in the sense that both are asked to extract 5 noun phrases and dissimilar in the sense that former is zero-shot and later is one-shot. It is interesting to note that GPT-noun toggled regarding closeness to a particular gender but when provided with an random example it became consistently closer towards male and significant as well indicating bias. Among all these observations, a unique observation of consistent bias towards male, irrespective of human annotations or AI generations, can be seen for Earth Mover's distance. To go deeper into what this observed biases entails, we need to understand the nature of the bias which we have shown in the next section.

7 RQ3: If so, what is the nature of the gender bias in reflection summaries?

7.1 Computing Discrepancy

For fairness, we want the topic distributions in summaries to equally represent both genders. However, our analysis investigating RQ2 found that abstractive AI summaries are biased toward representing male reflections. To dig deeper, we want to find which topics in particular contribute to this bias; we want to analyze how discrepant the topics are in the favored gender with respect to the disfavored gender. To measure this, we

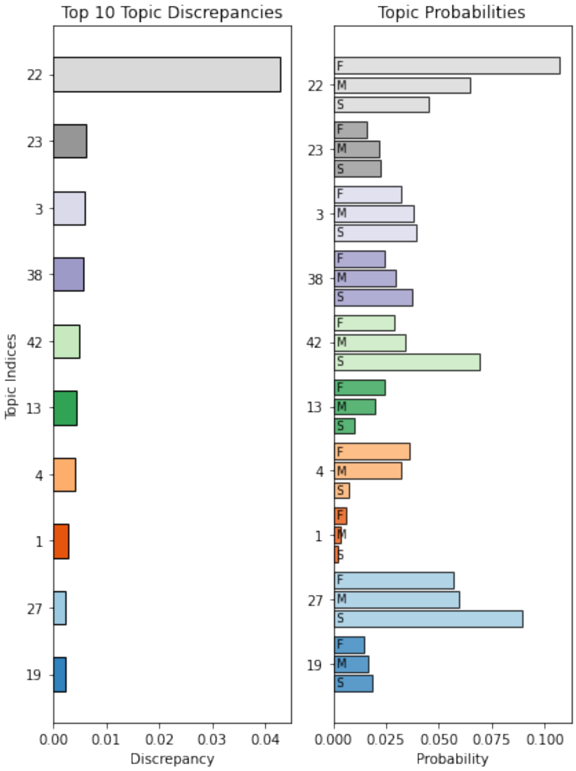

Figure 3: Top Discrepant Summary Topics - Bart-large

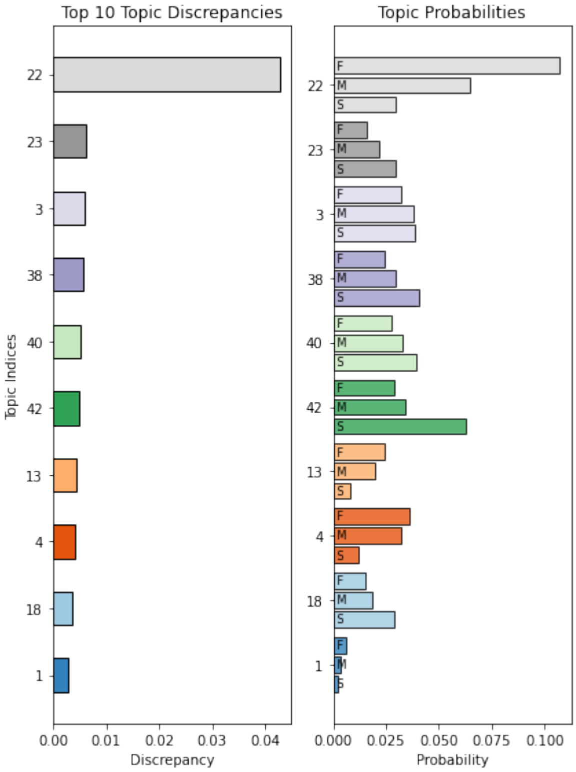

Figure 4: Top Discrepant Summary Topics - GPT-1shot

Topic 22: tree, binari, travers, search **- Under-representing Female**
1: how do you delete a black node vs. a red node from a red-black bst? 2: How to label and binary search tree. And the build tree method in the binary tree code
Topic 13: point, big, runtim, collis **- Under-representing Female**
1: I was confused about the Big O runtime details. I would love further explanation on how we can determine the estimated runtime. I would also like to know any tricks to more easily determine Big O. Additionally, I do not understand the difference between Big O, Little O, theta, and tilde. 2: BFS - how to keep track of what is seen/unseen
Topic 3: abl, group, team, meet **- Over-representing Male**
1: A04 and dividing work amongst team members 2: It was interesting to join groups and work together. It helped eliminate most confusion. And it was interesting to meet new people
Topic 42: class, today', assign, onlin **- Over-representing Male**
1: I think that the part that was most confusing today was what we were supposed to do for the in class assignment in class 2b 2: Due dates for assignment 10

Table 4: Top Reflections for Discrepant Topics

first computed the mean topic distribution for summaries $MeanS = 1/n{*}\sum_{i=1}^{N} S^i$ and both genders $MeanR_M = \sum_{i=1}^{N} R_M^i / \sum_{i=1}^{N} count(R_M^i)$ and $MeanR_F = \sum_{i=1}^{N} R_F^i / \sum_{i=1}^{N} count(R_F^i)$. We choose to analyze the most consistent one in terms of gender bias, i.e. the AI-generated abstractive summaries. Since, it is biased towards male gender, we compute $Discrepancy = |MeanS - MeanR_F| - |MeanS - MeanR_M|$. This computation will give us discrepancy, i.e. how skewed that topic was toward male or female students, for each of the T topics.

7.2 Analysis and Results

Our aim is to find out which male topics are being over-represented in biased summaries and which female topics are being under-represented. So, we extracted the top 10 topics [4] in decreasing order of discrepancy as shown in left part of Fig. 3 and Fig. 4. For each extracted topic we looked at its probability in $MeanS$, $MeanR_M$ and $MeanR_F$. Let those probabilities be $p(MeanS^t)$, $p(MeanR_M^t)$ and $p(MeanR_F^t)$, respectively, where $t \epsilon T$. These probabilities are shown in right part of Fig. 3 and Fig. 4. Now we compare these probabilities to figure out over-represented male topics and under-represented female topics. If $p(MeanS^t)$ is lowest among the three and $p(MeanR_M^t) < p(MeanR_F^t)$ then for topic t the summary is under-representing female reflections. On the flip side, if $p(MeanS^t)$ is highest among the three and $p(MeanR_M^t) > p(MeanR_F^t)$ then for topic t the summary is over-representing male reflections. It can be observed from Fig. 3 right part that 4 out of 10 topics *(22, 13, 4, 1)* are under representing female reflections and remaining 6 topics *(23, 3, 38, 42, 27, 19)* are over representing male reflections.

To understand these topics better we can look into their top words and reflections (described in section 5). Table 4 shows the details of two topics for each under-representing female and over-representing male categories. On analysis we discovered a common theme for both the categories. Topics that under-represented female referred to specific concepts like *Topic 22: tree, binari, travers, search, bst, black, red* where female students want to know about some functions of red-black tree and binary search tree - concepts belonging to computer science. Whereas, topics in summaries that over-represented male reflections were closer to a pedagogical theme instead of being related to concepts. For example, reflections from *Topic 3: abl, group, team, meet, teammat, everyon, work* in Table 4 shows that male students are talking about teamwork. *Topic 42: class, today', assign, onlin, brightspac, smooth part* also follows the same trend, where male students seem concerned about entire assignment or it's due date, instead of any specific concept or question in that assignment or where to focus in order to complete by due date.

Similar themes for both under-representing female and over-representing male topics were observed across all extracted discrepant topics for Bart-large and GPT-1shot models (top words and reflections are in Appendix A.4). It was also interesting that both Bart-large and GPT-1shot share 70% of top discrepant topics (see Fig. 3 and Fig. 4), providing evidence of convergent validity for our findings (both consistently biased towards male) and techniques for addressing **RQ2**.

With the discrepancy analysis we are able to find the nature of the bias answering **RQ3**. We have performed this analysis for AI-generated abstractive summaries, however, the same can be applied for other summary types, regardless whether how they were generated. It's important to mention that our work deals with identifying bias. A natural

[4] These are discrepant topics for summaries, not reflections.

followup question emerges about mitigating bias. To address this question one must find the reason for bias which in itself is a complex question to answer. Hence it can be formulated as future work.

8 Conclusion

In this work, we present the results from our fairness analysis of REFLECTSUMM (Zhong et al., 2024), a benchmark student reflection summarization dataset. We structured our analysis around three research questions: what topics differ between student reflections between male and female students (**RQ1**), are different types of summaries of those reflections biased toward any gender (**RQ2**), and if so, what is the nature of that bias (**RQ3**)?

We found slight topical differences between male and female reflections, such as female students being more likely to mention course logistics and refer explicitly to their own learning than male students. We also found that AI-generated abstractive summaries were biased towards male reflections, irrespective of whether the model was trained on human annotations or used generative causal model like ChatGPT. Human-generated summaries and extractive AI summaries did not exhibit consistent patterns of gender bias.

For the abstractive AI summaries, we found that topics with a pedagogical theme in the summaries are over-represented from male reflections while concept-specific topics were under-represented from female reflections. Such biases caution the use of popular LLM-based abstractive summarization techniques with educational reflection data.

This work could be extended to other educational datasets such as OULAD (Kuzilek et al., 2017), which has more demographic data, however there are not many student reflections available. Some issues with working with reflections data is hence the size and availability of these datasets. Our work could also be extended to analyze other demographic information present in REFLECTSUMM, such as race and ethnicity, as well as reflections from students identifying with genders outside of the gender binary.

We find STM to be a useful approach for analyzing bias in our case. Tracing where this bias could have originated in different training datasets with other tools (Feng et al., 2023) and across other abstractive summarization models would help illuminate possible sources of this bias.

9 Limitations

We have provided a basis framework for bias analysis. A deeper analysis on the basis of prompt or course is application specific and not performed as part of this work. However, it should be an natural extension for a complete analysis. Our analysis provide a birds eye view stating whether on average summaries are biased or not. Addition to this, a fined-grained analysis on individual summaries can be performed using our proposed techniques. STM finds all sorts of topics, those that are talking about content, logistics, etc. Other work may wish to filter the text first or otherwise specify the type of content they wish to investigate, such as comments about course content, learning style, learning technologies, classroom environment, etc.

10 Ethical Statement

It's of utmost importance to safeguard student demographic information from misuse. Safety measure have already been performed by Zhong et al. (2024). Their released version of the dataset contains no personal information like emails, first and last names and and other identification attributes. Our analysis is performed on this released version.

11 Bias Statement

By examining gender bias in summarization systems for student reflections, we are particularly concerned about the risk of allocational harm (Crawford, 2017; Lloyd, 2018; Blodgett et al., 2020). The intended use of such educational technologies is to summarize a potentially unwieldy number of reflections for teaching staff to understand student feedback about lectures and course content. If these summaries more closely represent the opinions and concerns of some groups while leave the comments of others unrepresented, teaching staff will only hear from and potentially adjust the class based on feedback from those groups.

We are particular concerned about gender bias in STEM courses due to a history of exclusion of female students from and within these courses (Clark Blickenstaff, 2005; Vincent-Ruz and Schunn, 2018). This history of bias and exclusion in university courses can contribute to fewer women in STEM professions and a potentially more hostile work environments (Arredondo et al., 2022). As education technologies are increasingly incorporated into such classes, they have the potential to further this bias and exclusion if not investigated properly. Our work is a step in this direction to measure gender bias for one such tool, automatic summarizations of student reflections.

12 Acknowledgement

We thank Diane J. Litman for inspiring this work, and also thank members of her lab who collected the dataset we use. We also thank Yang Zhong for providing continuous support with the dataset.

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681. The opinions expressed are those of the authors and do not represent the views of the NSF.

References

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

Patricia Arredondo, Marie L. Miville, Christina M. Capodilupo, and Tatiana Vera. 2022. *Women and the Challenge of STEM Professions: Thriving in a Chilly Climate*. International and Cultural Psychology. Springer International Publishing, Cham.

John Baird, Peter Fensham, Richard Gunstone, and Richard White. 1991. The importance of reflection in improving science teaching and learning. *Journal of Research in Science Teaching*, 28:163 – 182.

Ryan S Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41.

Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. ArXiv: 2005.14050 Issue: c.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Jennie S. Brotman and Felicia M. Moore. 2008. Girls and science: A review of four themes in the science education literature. *Journal of Research in Science Teaching*, 45(9):971–1002. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.20241.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yang Trista Cao and Hal Daumé III. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

J. K Chung, P. L Kannappan, C. T Ng, and P. K Sahoo. 1989. Measures of distance between probability distributions. *Journal of Mathematical Analysis and Applications*, 138(1):280–292.

Jacob Clark Blickenstaff. 2005. Women and science careers: leaky pipeline or gender filter? *Gender and Education*, 17(4):369–386. Publisher: Routledge _eprint: https://doi.org/10.1080/09540250500145072.

Kate Crawford. 2017. The Trouble with Bias - NIPS 2017 Keynote.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *CoRR*, abs/2106.14574.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Thomas Davidson and Debasmita Bhattacharya. 2020. Examining racial bias in an online abuse corpus with structural topic modeling. *arXiv preprint arXiv:2005.13041*.

Olivia Dias, Raechel Walker, and Cynthia Breazeal. 2022. Teaching an intersectional data analysis on affirmative action. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, pages 1238–1238.

Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. 2015. Coursemirror: Enhancing large classroom instructor-student interactions via mobile interfaces and natural language processing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, page 1473–1478, New York, NY, USA. Association for Computing Machinery.

Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. 2017. Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, page 363–374, New York, NY, USA. Association for Computing Machinery.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. *CoRR*, abs/2106.11410.

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

Eduard H Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *LREC*, volume 6, pages 604–611.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. *Scientific data*, 4(1):1–8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chien-Chang Lin, Anna YQ Huang, and Owen HT Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41.

Diane Litman, Haoran Zhang, Richard Correnti, Lindsay Clare Matsumura, and Elaine Wang. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In *Artificial Intelligence in Education*, pages 255–267, Cham. Springer International Publishing.

Yuhan Liu, Shangbin Feng, Xiaochuang Han, Vidhisha Balachandran, Chan Young Park, Sachin Kumar, and Yulia Tsvetkov. 2024. p^3sum: Preserving author's perspective in news summarization with diffusion language models.

Kirsten Lloyd. 2018. Bias Amplification in Artificial Intelligence Systems. ArXiv:1809.07842 [cs].

Wencan Luo and Diane Litman. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Lisbon, Portugal. Association for Computational Linguistics.

Wencan Luo, Fei Liu, and Diane Litman. 2016. An improved phrase-based approach to annotating and summarizing student course responses. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 53–63, Osaka, Japan. The COLING 2016 Organizing Committee.

Michael Madaio, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2022. Beyond "fairness": Structural (in) justice lenses on ai for education. In *The ethics of artificial intelligence in education*, pages 203–239. Routledge.

Ahmed Magooda and Diane Litman. 2020. Abstractive summarization for low resource data using domain transfer and data synthesis.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460, Florence, Italy. Association for Computational Linguistics.

Danielle S McNamara. 2011. Measuring deep, reflective comprehension and learning strategies: challenges and successes. *Metacognition and Learning*, 6:195–203.

David Mimno and Moontae Lee. 2014. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319–1328, Doha, Qatar. Association for Computational Linguistics.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2019. stm: An r package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

Rod D Roscoe, Shima Salehi, Nia Nixon, Marcelo Worsley, Chris Piech, and Rose Luckin. 2022. Inclusion and equity as a paradigm shift for artificial intelligence in education. In *Artificial Intelligence in STEM Education*, pages 359–374. CRC Press.

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.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Umair Shakir, Sarah Ovink, and Andrew Katz. 2022. Using natural language processing to explore undergraduate students' perspectives of social class, gender, and race. In *American Society for Engineering Education Annual Conference*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Marta Villamor Martin, David A Kirsch, and Fabian Prieto-Nañez. 2023. The promise of machine-learning-driven text analysis techniques for historical research: topic modeling and word embedding. *Management & Organizational History*, 18(1):81–96.

Paulette Vincent-Ruz and Christian D. Schunn. 2018. The nature of science identity and its role as the driver of student choices. *International Journal of STEM Education*, 5(1):48.

Damin Zhang and Julia Rayz. 2022. Examining stereotypes in news articles. In *The International FLAIRS Conference Proceedings*, volume 35.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Yang Zhong, Mohamed Elaraby, Diane Litman, Ahmed Ashraf Butt, and Muhsin Menekse. 2024. ReflectSumm: A benchmark for course reflection summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13819–13846, Torino, Italia. ELRA and ICCL.

Leyla Zhuhadar, Scarlett Marklin, Evelyn Thrasher, and Miltiadis D. Lytras. 2016. Is there a gender difference in interacting with intelligent tutoring system? can bayesian knowledge tracing and learning curve analysis models answer this question? *Computers in Human Behavior*, 61:198–204.

A Appendix

A.1 Top Topics

Fig. 5 shows all the topics learned by the topic model (as described in section 5) in decreasing order of expected topic proportions. For each topic, top four words ranked according to FREX score (Roberts et al., 2019) are also specifies which help in characterizing the topic.

A.2 Topical Analysis

Along with estimating a regression model to find association between topics and particular genders (as described in section 5) we also estimated a similar one to find associations between topics and prompts types (I and C) which is shown in Fig. 6. We also went a step deeper in our analysis and plot topic gender associations mediated by prompt type (I or C), as shown in Fig. 7 (mediator being prompt I) and Fig. 8 (mediator being prompt C). Similarly, we plot topic gender associations mediated by course, as shown in Fig. 9 (mediator being course *CS*), Fig. 10 (mediator being course *ENGR*), Fig. 11 (mediator being course *CMPINF*) and Fig. 12 (mediator being course *PHYS*).

A.3 Bias Analysis

Tables 5 shows details about the mean distance calculated between summaries and their corresponding reflections for each gender (as described in section 6). We performed a Student's t-test on a random sample of these computed distances. The $p - value$ for each corresponding test is also mentioned in the table. The significant ones whose $p - value$ is < 0.05, marked with $*p - value$.

A.4 Discrepant Topic Reflections

Table 6 and Table 7 show top words and top reflections for all the discrepant topics identified in section 7 for both abstractive summary generation AI models - BART-large and GPT-1shot.

Distance Metric	**Human Annotation**			**AI Generation**		
	Extractive	**Abstractive**	**Phrased**	**Extractive**	**Abstractive**	**Phrased**
1-cosine	M:0.448 F:0.442 *p-value: 0.01	M:0.474 F:0.473 *p-value: 0.03	M:0.470 F:0.468 p-value: 0.08	(MatchSum) M:0.453 F:0.444 *p-value:0.006	(BART-Large) M:0.551 F:0.555 *p-value:8.7e-05	(GPT-noun) M:0.486 F:0.483 *p-value:0.001
				(GPT-reflect) 0.46 F:0.45 *p-value:0.0002	(GPT-1shot) M:0.55 F:0.57 *p-value:1.5e-07	(GPT-noun-1shot) M:0.482 F:0.486 p-value:0.65
jsd	M:0.157 F:0.154 p-value: 0.96	M:0.169 F:0.165 p-value: 0.25	M:0.1480 F:0.1487 p-value: 0.81	(MatchSum) M:0.16 F:0.15 p-value:0.25	(BART-Large) M:0.18 F:0.20 *p-value:0.3.6e-07	(GPT-noun) M:0.16 F:0.59 p-value:0.98
				(GPT-reflect) M:0.16 F0.15 p-value:0.14	(GPT-1shot) M:0.18 F:0.20 *p-value:2.6e-08	(GPT-noun-1shot) M:0.15 F:0.16 *p-value:0.03
hellinger	M:0.40 F:0.39 p-value: 0.28	M:0.42 F:0.41 p-value: 0.6	M:0.389 F:0.388 p-value: 0.8	(MatchSum) M:0.408 F:0.402 p-value:0.25	(BART-Large) M:0.44 F:0.46 *p-value:1.8e-5	(GPT-noun) M:0.405 F:0.401 p-value:0.26
				(GPT-reflect) M:0.4 F0.39 *p-value:0.01	(GPT-1shot) M:0.44 F:0.47 *p-value:1.5e-6	(GPT-noun-1shot) M:0.39 F:0.40 p-value:0.09
earthmover	M:0.005 F:0.006 *p-value: 4.3e-6	M:0.005 F:0.006 *p-value: 1.1e-6	M:0.006 F:0.007 *p-value: 0.02	(MatchSum) M:0.0062 F:0.0068 *p-value:0.001	(BART-Large) M:0.006 F:0.007 *p-value:5.4e-11	(GPT-noun) M:0.0072 F:0.0077 *p-value:0.001
				(GPT-reflect) M:0.0061 F0.0066 *p-value:7.8e-5	(GPT-1shot) M:0.006 F:0.007 *p-value:5.1e-10	(GPT-noun-1shot) M:0.006 F:0.007 *p-value:1e-5

Table 5: Mean Difference between Reflection (for each gender) and Summary.

Topic 22: tree, binari, travers, search - Under-representing Female
1: how do you delete a black node vs. a red node from a red-black bst?
2: How to label and binary search tree. And the build tree method in the binary tree code
Topic 13: point, big, runtim, collis - Under-representing Female
1: I was confused about the Big O runtime details. I would love further explanation on how we can determine the estimated runtime. I would also like to know any tricks to more easily determine Big O. Additionally, I do not understand the difference between Big O, Little O, theta, and tilde.
2: BFS - how to keep track of what is seen/unseen
Topic 4: algorithm, abstract, network, prim - Under-representing Female
1: The most interesting part was the application of emojis stored in unicode as well as audio encodings in relation to MP3 players.
2: eager prims and lazy prim
Topic 1: forc, object, mass, resist - Under-representing Female
1: I think it's interesting that momentum can be conserved if no external forces are acting on an object.
2: linear momentum using center of mass, derivative of momentum

Table 6: Top Reflections for Discrepant Topics - Under-representing Female. Sorted in decreasing order of discrepancy.

Topic 23: figur, instruct, criteria, sourc - Over-representing Male
1: When printing a vector, I am able to display it in individual statements in the command window, with one fprintf statement. However, when I am attempting to display two vectors like I would with two values in a fprintf stament it does not work.
2: I found it interesting that there were 5 spots open for criteria but only 4 listed. Why bother with adding a blank row?
Topic 3: abl, group, team, meet - Over-representing Male
1: A04 and dividing work amongst team members
2: It was interesting to join groups and work together. It helped eliminate most confusion. And it was interesting to meet new people
Topic 38: project, new, design, present - Over-representing Male
1: Introduction of new design project
2: Taking a look at the new memo to see the new project
Topic 40: data, comput, regress, communic - Over-representing Male
1: The Data, Information, Knowledge, Wisdom debatability.
2: How data is raw and needs to be processed into information and that data can ultimately be turned into wisdom
Topic 42: class, today', assign, onlin - Over-representing Male
1: I think that the part that was most confusing today was what we were supposed to do for the in class assignment in class 2b
2: Due dates for assignment 10
Topic 18: need, detail, prototyp, suppos - Over-representing Male
1: I was confused on what to do on some places because I couldn't find the documents in brightspace.
2: The type of prototypes that we have to make by Monday for testing.
Topic 27: noth, everyth, explain, clear - Over-representing Male
1: Nothing, today went at a great pace
2: Nothing. You explained everything very well
Topic 19: sinc, multipl, put, main - Over-representing Male
1: I may need more clarity on prefix trees since they're kinda complicated especially when there are many nodes
2: The idea of communication between living cells was very interesting, but dwelling too much time on it may be off the mark for the scope of this class. Perhaps the idea of packet switching on the internet could be related to neurotransmitters or some other physical packet.

Table 7: Top Reflections for Discrepant Topics - Over-representing Male. Sorted in decreasing order of discrepancy.

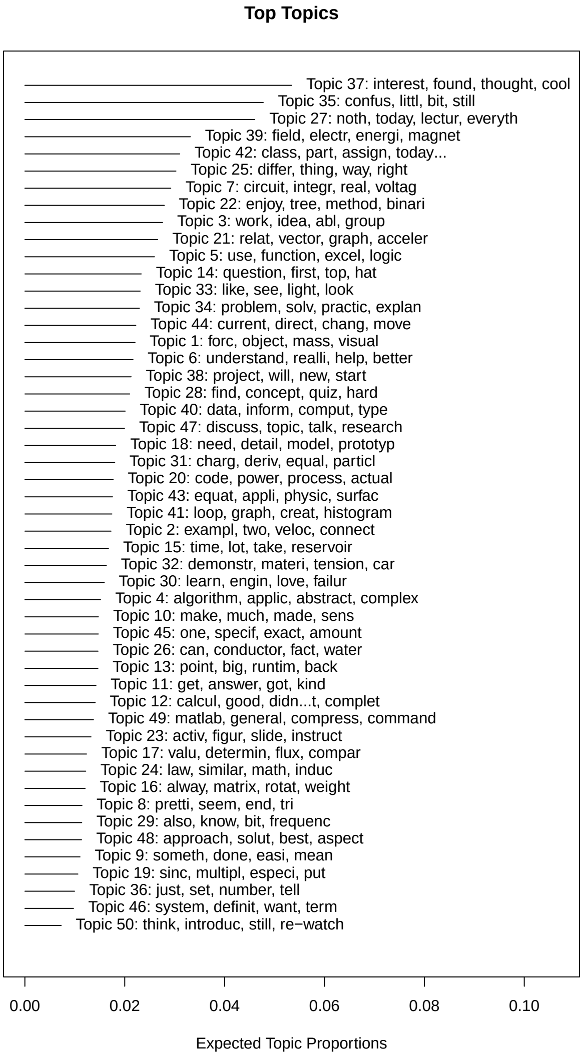

Figure 5: Topics sorted by FREX score

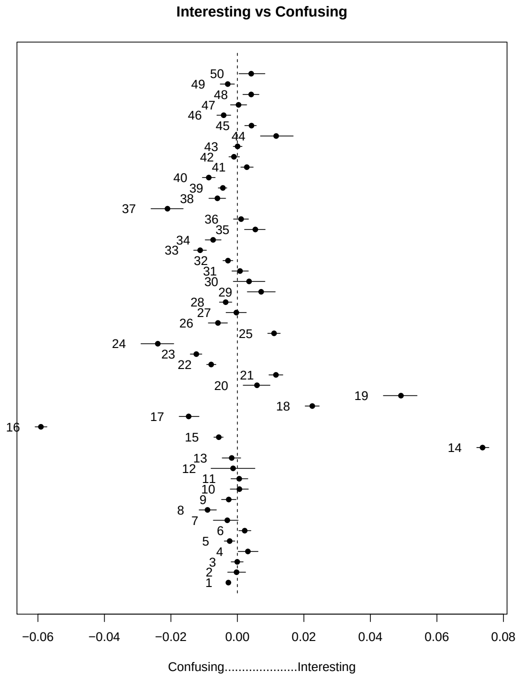

Figure 6: Topical Difference between Prompts

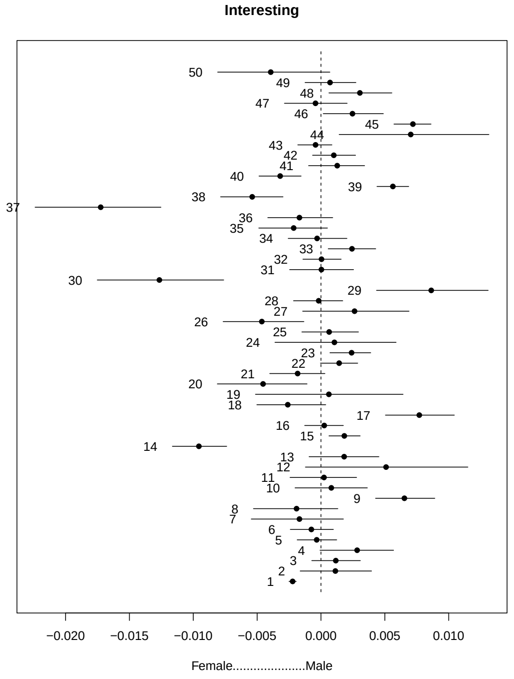

Figure 7: Topical Difference between Genders w.r.t. to Prompt (I)

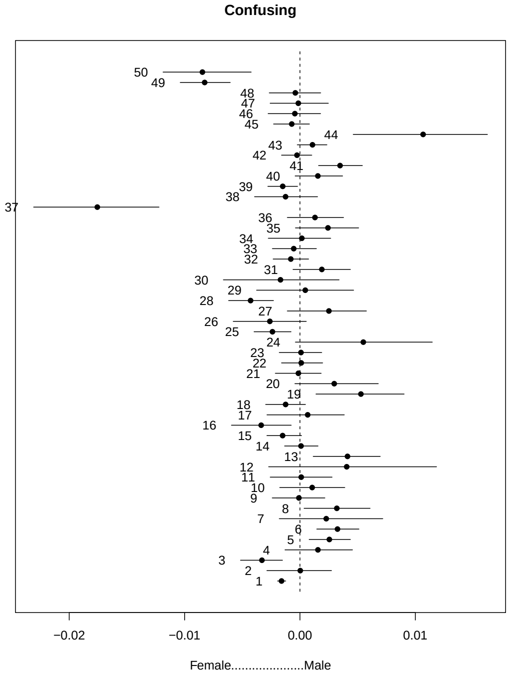

Figure 8: Topical Difference between Genders w.r.t. to Prompt (C)

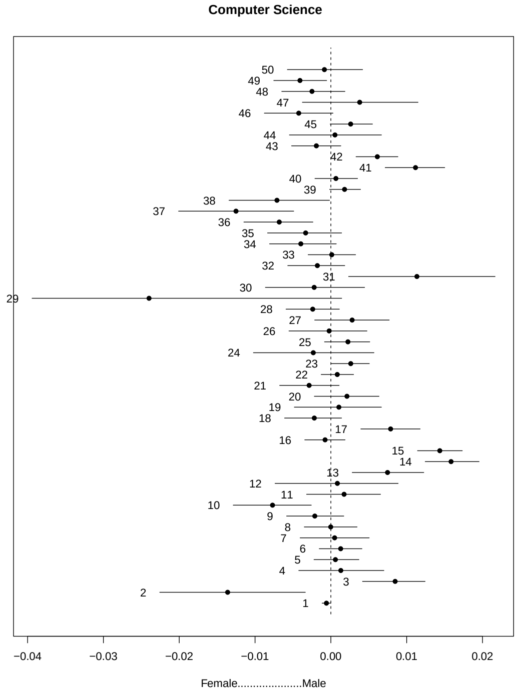

Figure 9: Topical Difference between Genders w.r.t. to Course (CS)

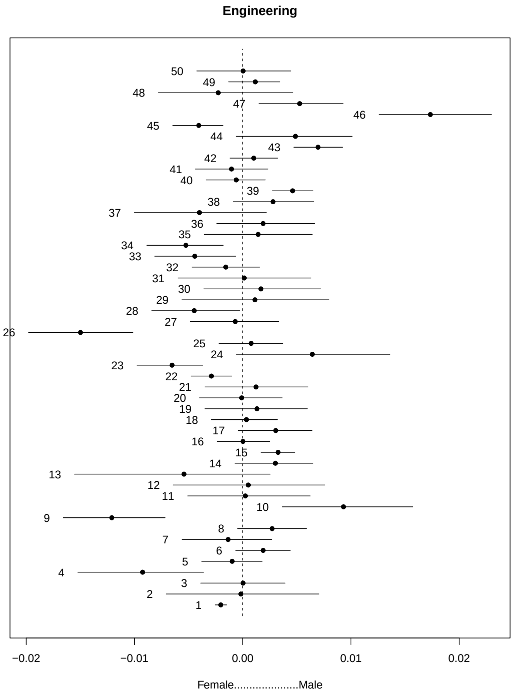

Figure 10: Topical Difference between Genders w.r.t. to Course (ENGR)

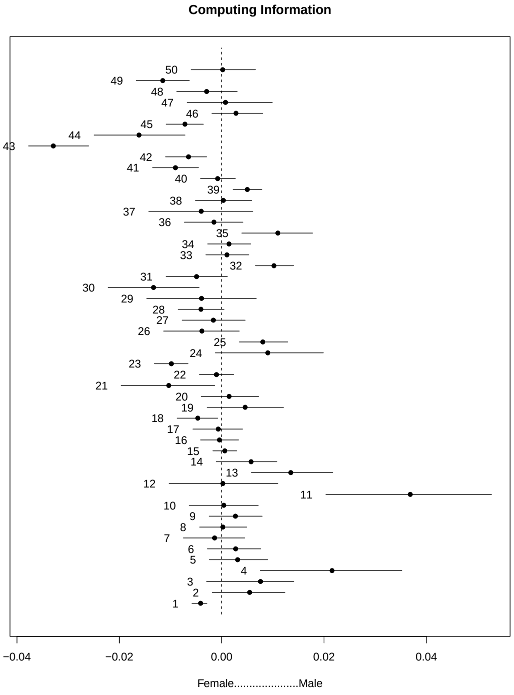

Figure 11: Topical Difference between Genders w.r.t. to Course (CMPINF)

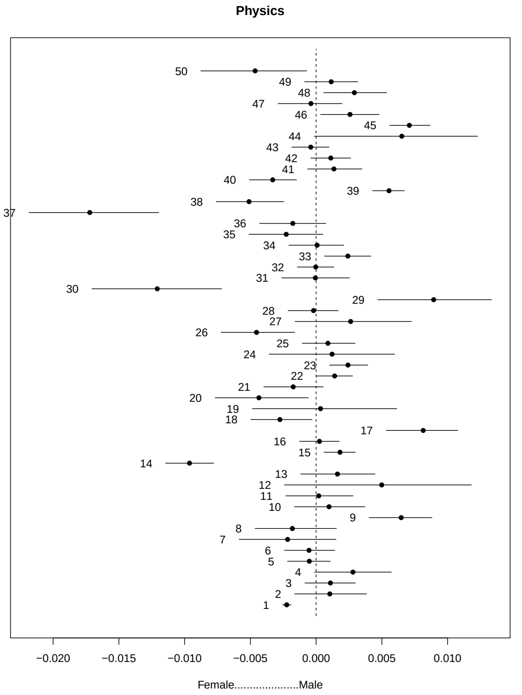

Figure 12: Topical Difference between Genders w.r.t. to Course (PHYS)