Using language models to detect and reduce gender bias in university forum messages

Gianina Salomo-López, Cristóbal Alcázar, Roberto Barceló,
Camilo Carvajal Reyes, Darinka Radovic, Felipe Tobar

Introduction

Gender bias refers to the systematic and unequal treatment based on an individual's gender [1], or the preference or prejudice towards one gender over another [2]. Gender biases can be led by humans or autonomous systems, and although their occurrence in decision-making is usually unintentional, they have profound consequences on societal interactions. These biases are particularly detrimental in educational settings, where they can reinforce stereotypes, influence student performance and engagement, and perpetuate systemic inequalities. In this regard, a source of concern is Artificial Intelligence (AI) systems and machine learning (ML) models that use natural language processing (NLP) techniques, as they convey challenges and opportunities. While using gender-biased datasets on model training has known harmful consequences [3], AI/ML's unparalleled ability for text processing makes them an attractive resource for detecting and mitigating gender bias from natural language. Our focus is to address gender bias in educational contexts using AI, a crucial step to fostering inclusive learning environments and promoting equitable opportunities for all learners.

The detection, quantification, and mitigation of gender bias in various elements of NLP tasks have been extensively studied in the literature during the last decade. For instance, [4] measured gender biases in general-purpose language models, while others have focused on the emergence and control of gender biases in NLP tasks, such as machine translation [5], text generation [6] and coreference resolution [7]. Efforts have also been made towards producing labelled datasets to facilitate the detection of gender bias in English texts [8], and to detect and classify misogynistic content or behaviours in English, Spanish, and Italian texts [9]. Despite these advances, handling gender biases across AI-driven NLP applications

Gianina Salomo-López, Darinka Radovic, Cristóbal Alcázar and Roberto Barceló are with Universidad de Chile.
Felipe Tobar and Camilo Carvajal Reyes are with the Department of Mathematics and I-X, Imperial College London.
(Corresponding author: Felipe Tobar, e-mail: f.tobar@imperial.ac.uk)

in compliance with Diversity, Equality & Inclusion (DEI) values is still an open problem in the general case.

Most works at the interface of AI and gender studies focus on understanding and eradicating biases in models that interact with humans and assist in decision-making. In this work, however, we address a task that has received considerably less attention: to develop AI systems to detect and mitigate biases in human-generated content. The methodological focus on our work is on *large language models* (LLMs), which are large-scale neural networks that manipulate natural language to perform tasks such as text generation, summarisation, and in general answering questions on broad topics. Being a ML model, training LLMs requires massive datasets, which, in practice, are constructed by aggregating news, literary works and even online public discussion forums. Even though these raw, unprocessed, datasets feature gender biases [7], [8] that are transferred to the trained models, we claim that LLMs are a promising tool to assist in the correction of gender biases in text. Our hypothesis follows from the outstanding capacity of language models to follow instructions, i.e., *prompts*, to perform data-processing tasks of general interest. One of these tasks has been to maximise human satisfaction to the LLMs own answers, as done by GPT4 [10]. We conjecture that if an LLM can be instructed to modify text to fulfil such an ambitious objective, it can be instructed to identify and remove gender biases from human-generated text, and even to explain how it has done so. LLMs have evolved beyond simple text generation towards tasks that involve complex reasoning that are executed through agentic systems. These systems can be implemented as workflows, where LLMs and other tools build agents that are dynamically orchestrated by the same model [11]. LLMs' ability to manage their own decision-making process, while responding to environmental feedback, justifies their use in the context of mitigating gender bias in text.

Scope and contributions. To honour the standing promise of AI's potential for social good, we develop a Generative AI framework to detect and remove gender bias in human-generated texts using language models. The proposed methods evaluate whether gendered terms and phrases are unjustifiably biased, requiring an in-context assessment. In this way, our proposal is tested through a case study using texts corresponding to formal communications in the School of Physical and Mathematical Sciences at the Universidad de Chile (Spanish acronym FCFM). Our specific contributions are:

- A dataset of parallel sentences, where gender-biased sentences are scrapped from (publicly available) official communications within FCFM and bias-free versions of each sentence are created by annotated by a team of human experts. This set can be used to address a gender bias mitigation as a Text Style Transfer (TST) task [12], or to validate gender-bias removal tools.
- An LLM agent designed to augment the dataset above with explanations of the detected biases and how they were mitigated.

- A fine-tuned Language Model (LM) for gender-bias mitigation in Spanish.
- An LLM-based agent for detecting and explaining gender biases in text from a few-shot-learning rationale using the augmented dataset mentioned above.
- Qualitative and quantitative evaluations of the proposed methodologies (fine-tuned LM and LLM-based agent) using existing metrics in the literature, as well as an agent termed LLM-judge built specifically for this purpose.

Recognising that performing gender debiasing through supervised learning is rather ambitious, we approach this problem both from a few-shot-learning and a fine-tuning perspective, leveraging large and moderately-sized LMs. These two approaches require resources of different nature: fine-tuning can be achieved with in-house hardware at a minor computing cost, while few-shot-learning requires access to closed-models through a paid API. We evaluated both pipelines, where the fine-tuned local models showed a superior overall performance than the LLM for this task, although the latter showed a greater sensitivity. Our experimental results help to better understand the performance of the proposed approaches to gender bias mitigation in human-generated text, and thus pave the way to exploiting the potential of LLMs to advance Diversity, Equality, and Inclusion (DEI) values in STEM education. Furthermore, our results validate this alternative approach to data (text) processing and the lessons learnt are expected to catalyse the adoption of agents in more general data processing challenges. Lastly, the proposed methodology provides a conceptual language-agnostic debiasing procedure for human-written text that may present gender biases. We, in particular, test this claim with the challenging case of Spanish text in university communications.

Task formulation and context

Gender biases

Various studies within the realms of NLP, AI, and learning systems have identified and employed different models for categorising gender biases. These taxonomies discriminate between: i) biases arising when gender-neutral terms are syntactically referenced by gender-exclusive pronouns, and ii) biases following sexism and societal stereotypes, such as occupational biases and gendered roles [13]. For instance, [13], [14] proposed that structural gender biases occur when bias can be traced to specific grammatical constructions that reinforce gender assumptions in a gender-neutral context. They also identify contextual biases, such as societal and behavioural stereotypes. From a different perspective, [15] presents a general model for bias classification considering their potential harm: Harms of allocation, referring to unfairly assigned opportunities or resources by algorithms; and Harms of representation, pertaining

to discriminatory depictions filtered by algorithms. These diverse typologies can aid in identifying and modifying biases in texts.

Recasting gender bias removal as a learning problem

Pre-LLM methods for gender-bias detection and mitigation are mainly of two classes: i) based on rules [16, Sec. 4.2.1], or ii) based on statistical discrepancies [17]. The former operates by replacing tokens (e.g., *chairman* to *chairperson*) and thus requires a manually pre-defined dictionary of equivalent terms. The latter addresses gender bias detection via computing distances/divergences between probability distributions akin to *anomaly detection*. This makes bias mitigation in text rather challenging, since the text structure is destroyed by adopting a bag of words representation required to compute divergences between the (empirical) distributions. Both approaches are limited in the sense that they do not directly include the context of the documents, in fact, languages with grammatical gender involve sentences where a gender is implied by a subject. It is not just the presence of a gendered word, but the context and the subjects involved what determine the existence of a gender bias. This is analogous to other, more complex examples of gender bias, most of which are even harder to tackle with non-LLM algorithms.

Early deep learning methods address the mitigation of gender bias in texts by means of Text Style Transfer (TST) [12], which involves rewriting a sentence in a new style (e.g., without gender bias) while preserving its semantic content. TST tasks are usually addressed in a supervised manner [18], [19], where training is achieved using *parallel data*, i.e., sentences in the source style (e.g., biased) paired with their corresponding sentences in the target style (e.g., bias-free). Within NLP, TST is a sequence-to-sequence task (Seq2Seq), called so because it estimates an output sequence given an input sequence. Seq2Seq, and in particular TST, are usually addressed using encoder-decoder deep neural networks [20], [21] and recently transformers [22].

Mainly due to its unprecedented ability to process text, transformer-based LLMs have become the *de facto* building block in a plethora of modern AI systems. Their generality, stemming from the pre-training stages combined with task-specific fine tuning (possibly using human feedback), has positioned them as the go-to method for a wide range of complex learning tasks. In fact, they have even been used for text-style transfer via zero-shot learning, that is, without requiring any datapoints [23]. However, we claim that the capabilities of LLMs have been underexplored in our particular written-content moderation task. Therefore, we conjecture that, with appropriate prompting, LLMs can be used within agents that detect and correct gender-biased text in a zero- or few-shot learning setting.

Universidad de Chile's School of Engineering (FCFM)

To provide context for our case study, we present a brief overview of FCFM and its gender distribution among both students and academic staff. FCFM's six-year undergraduate programmes host over 6,000 undergraduate students which take part in a two-year common plan leading to 13 different degrees. These programmes span the fields of Science, Technology, Engineering and Mathematics (STEM, [24]), with graduates populating the national industry, public sector, or continuing postgraduate programmes. Historically, FCFM has exhibited an underrepresentation of women. Approximately a decade ago, female students comprised less than 20% of the undergraduate programme, with female academics making up less than 15%. Over the past decade, increasing concerns regarding female participation and gender policies have emerged. In 2014, quotas for women were implemented in the undergraduate programme and academic recruitment processes. Subsequently, in 2018, establishing a Gender and Diversity Unit further institutionalised inclusion and gender concerns, consequently enhancing female participation. By 2023, the proportion of female undergraduate students had surpassed 30%, with female academics constituting 22% of the faculty. Against this backdrop, this case study endeavours to analyse a formal communication channel. Assessing and intervening language use in institutional educational settings is vital for fostering inclusivity and equity. Language not only reflects but also shapes societal norms; thus, the consistent use of gendered terms can inadvertently reinforce stereotypes and marginalise non-binary individuals. In written texts and institutional forums gendered language may signal bias or a lack of awareness, potentially alienating members of the educational community. Conversely, adopting gender-neutral language can promote a sense of belonging and demonstrates an institution's commitment to diversity and inclusion.

A DATASET OF PARALLEL SENTENCES

A. Data acquisition and annotation

We built a corpus from messages posted in the *"Novedades"* ("News") page in FCFM's student forum *U-Cursos*.[1] This public-access page contains posts about institutional events and activities in which students and the community might be interested. Built via web scraping, our corpus comprises two sets described in Table I.

For each reference sentence, gender-biased and gender-unbiased versions of the same sentence were manually constructed by a group of 10 experts trained in producing gender-neutral texts. The experts were also asked to indicate whether the original sentence had gender bias or not; this indication was intended for posterior statistical analyses and played no role in the subsequent fine-tuning of the LM or the design of

[1]See `www.u-cursos.cl/uchile/4/novedades_institucion`

TABLE I: Description of the two subsets of scrapped data. Each message was split into sentences using the `sent_tokenize` method from the `NLTK` package.

	Purpose	Acquisition period	Messages	Sentences
C	Training & validation	25 October 2007 to 7 August 2023	8,868	31,195
C_{test}	Testing	23 August 2023 to 6 December 2023	200	853

the agent. The experts were instructed to discard cases where gender-biased and gender-unbiased versions were impossible or deemed to be unrealistic, meaningless, or too synthetic. This included dates, places and other gender-free entities, such as sentences which do not refer to people. In order to standardise the construction of these versions, a manual was created for the expert annotators,[2] explaining different definitions of gender biases (relevant to the interests of the FCFM) and indicating how to complete the data sheet. As part of their training, the content of the manual was explained to each expert individually, and they were asked to test a preliminary subset of 20 sentences.

Lastly, the gender-biased and gender-unbiased sentences in C were organised in a parallel format suitable for supervised learning tasks such as TST. Specifically, for each pair of unbiased-biased sentences, the biased sentence was considered as an *input* denoted d_{in}, while the unbiased one as the output denoted d_{out}. For sentences where biasing/unbiasing was unfeasible (referred to as *unable-to-bias*) the original sentence was used for both d_{in} and d_{out}, since in that case a debiasing operation should leave the sentence invariant. For the sentences in C_{test}, each original sentence was considered as input d_{in} with multiple possible outpus d_{out}. For notational simplicity, we refer to the parallel subsets as C and C_{test} as well.

B. Exploratory analysis

Across both subsets C and C_{test}, about two thirds of the sentences were unable-to-bias, that is, they did not admit biasing/debiasing. Fig. 1 (left) shows the distribution of these sentences for the the training and validation sets, which were a 70-30 split of C, as well as the test set C_{test}. Observe that the proportion of unable-to-bias sentences is roughly invariant to the subset. We also investigated the proportion of gender-biased sentences in time, with respect to all sentences that admitted biasing/debiasing. Fig. 1 (right) shows that, until the year 2015, more than 80% of the sentences exhibited gender bias, and this proportion has decayed consistently since that year. This reveals the effect of the DEI policy in place at FCFM but also poses a challenge in our setting, since the occurrence of gender biases varies in time. In order to test the strength of our methodology in a setting as realistic as possible, we decided not to

[2]Permanent link to the annotators' manual: `Link`

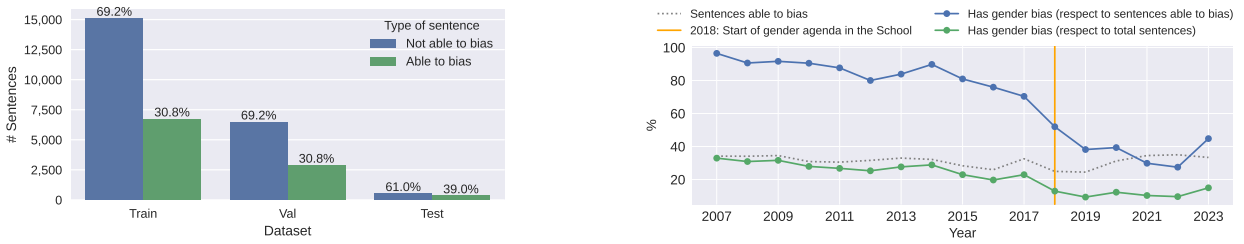

Fig. 1: Exploratory analysis of the constructed dataset. **Left:** Number of paired sentences in each subset, note that about 30%-40% of the sentences data admitted biasing/debiasing, irrespective of their subset. **Right:** Proportion of original sentences with gender bias compared against all sentences (green) and sentences that can be biased (blue). Observe the decrease in this proportion since around 2015.

modify this *temporal distribution shift* in the occurrence of gender biasedness. This means that, according to Table I, the training and test sets have different proportions of gender-biased examples and thus our method is expected to perform despite such disparity.

C. Dataset augmentation with an LLM

The dataset described above was augmented with brief explanations about how the biasing/debiasing was achieved, these explanations were then used by the debiasing and evaluation agents presented in the next sections. As these explanations had to be succinct and consistent in terms of length and style, they were also produced by LLMs. The role of the agents is illustrated in Figure 2 and described as follows.

- **Descriptor Agent:** Receives a biased sentence and its (expert) unbiased version to i) produce a description of the performed debiasing, and ii) identify the critical biased and unbiased words in the sentences. This agent generates four candidate outputs, as configured in the API function call, where the sampler generates different responses for different seeds.
- **Evaluator Agent:** Receives the (4) outputs of the Descriptor Agent and the original biased-unbiased pair of sentences, to determine the *best* description of the unbiasing from the 4 candidates.

We used the LLM Gemini [25] for the Descriptor and Evaluator agents (see top row of Table II) with system prompts directly in Spanish, as this provided the best results. The corresponding system prompts are presented as follows.

TABLE II: LLMs used for Data Augmentation and Debiser agentic systems

System	Agent	Model	Provider	Reference
Dataset Augmentation	Descriptor & Evaluator	gemini-flash-1.5	Google	[25]
Debiaser	Detector & Neutraliser	claude-3-5-haiku-20241022	Anthropic	[26]
Debiaser	Critic	gpt-4o	OpenAI	[27]

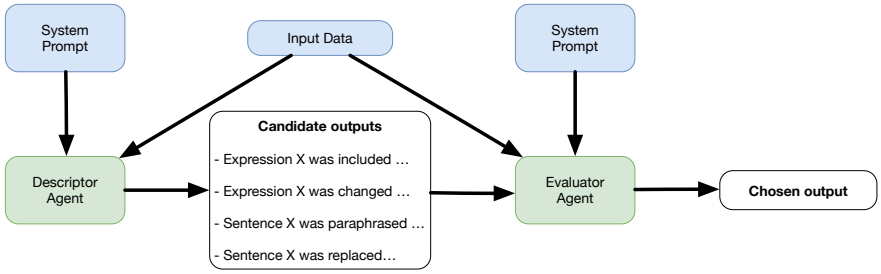

Fig. 2: Data augmentation agents. The **Descriptor** is instructed to provide possible explanations of the undertaken debiasing process for a pair of sentences. Then, the **Evaluator** chooses the the *best* explanation (as defined in the system prompt).

System prompt for Descriptor Agent

You are an expert in annotating gender bias in Spanish texts. Your task is to show how an original phrase with gender bias differs from a bias-free version provided by a human expert. You will be given two sentences:

Original phrase: `{original}`

Bias-free phrase: `{unbiased}`

Your responsibilities:

1) **Describe the differences:** Clearly and briefly identify how the original phrase was modified to remove gender bias. Provide a concise explanation enclosed in double angular brackets, for example: `<<brief explanation of the differences>>`.
2) **List biased terms:** Enumerate the specific words in the original phrase that express gender bias. These may include pronouns, adjectives, nouns, or participles. Return them in a list enclosed in double angular brackets, for example: `>>[list of biased terms {{insert pronouns, adjectives, nouns, or participles}}]<<`.
3) **List unbiased terms:** Enumerate the specific words in the bias-free phrase that were used to mitigate gender bias. These may include pronouns, adjectives, nouns, or participles. Return them in a list enclosed between a hyphen and angular brackets, following this example: `->[list of unbiased terms {{insert pronouns, adjectives, nouns, or participles}}]<-`.

Formatting requirements:

- Ensure that you respond in the correct format for each answer. The terms must be correctly enclosed in their corresponding formats, with no extra spaces or symbols.
- Keep the description of the differences as brief as possible, focusing on how the phrase was neutralised. Brevity and clarity are the most important aspects. It is imperative that they be short and concise.
- Do not add additional comments or explanations beyond the required output.

System prompt for Evaluator Agent

You are a validation and quality control assistant. Your task is to evaluate several candidate responses generated by another agent that has performed gender bias annotation on a Spanish text. All these candidate responses are based on the same input:

Original phrase: `{original}`

Bias-free phrase: `{unbiased}`

Each candidate response includes:

- A description of the differences between the biased phrase and the bias-free phrase, enclosed within `<<...>>`.
- A list of biased terms found in the original phrase, enclosed within `>>[...]<<`.
- A list of the new unbiased terms found in the bias-free phrase, enclosed within `->[...]<-`.

Your responsibilities:

1) **Evaluate Accuracy and Completeness:**
   - Compare all candidate responses with the original phrase and the bias-free phrase.
   - Determine which one best meets the following criteria:
     - Correctly identifies the modifications made to remove gender bias.
     - Provides a concise, clear, and correct description of those differences.
     - Accurately lists only the biased terms present in the original phrase.
2) **Select the Best Candidate:**
   - Among all candidates, choose the best response. This should be the one that most faithfully follows the original instructions, is accurate in its statements, and is clearly structured according to the specified format.
3) **No Additional Comments:**
   - Return only the best response in the exact format in which it was presented.
   - Do not add explanations, reasoning, or additional comments.
4) **Verify the Format:**
   - When responding, ensure that each answer follows the correct format.
   - Make sure that the terms are properly enclosed in their corresponding formats, with no extra spaces or symbols.
   - If they are not correctly formatted, you are authorised to fix only the formatting.

Candidate responses: `{candidates_str}`

We provide four examples of pairs annotated by the experts and the explanation of the LLM agents (**Descriptor** and **Evaluator**) in Table III. Critical words identified by the agents are highlighted in red (biased) and green (unbiased) each case. The table also provides an explanation in English for each

TABLE III: Agent's justification for the gender debiasing performed by expert (human) annotators. An explanation in English is provided for each example in bold font.

Example I	Original Sentence	Estimad@s: Se extiende plazo de postulación a Ayudas Estudiantiles Mineduc y CAE para estudiantes **nuevos** y **antiguos** hasta el 27 de marzo.
	Corrected Sentence	Estimad@s: Se extiende plazo de postulación a Ayudas Estudiantiles Mineduc y CAE para estudiantes **nuevas/os** y **antiguas/os** hasta el 27 de marzo.
	Agent's justification	Se reemplazó "estudiantes" por "estudiantes nuevas/os y antiguas/os" para incluir explícitamente ambos géneros.
	Explanation (English)	**The agent identified the inclusion of both genders in the change from (*estudiantes nuevos y antiguos*) to (*estudiantes nuevas/os y antiguas/os*)**
Example II	Original Sentence	**Los** estudiantes que quieran participar como **anfitriones**, guías de grupos o en la logística durante los días 18, 19 o 20 de octubre, recibirán de regalo una polera conmemorativa y un recuerdo de la Facultad, también colaciones.
	Corrected Sentence	**Las/os** estudiantes que quieran participar como **anfitrionas/es**, guías de grupos o en la logística durante los días 18, 19 o 20 de octubre, recibirán de regalo una polera conmemorativa y un recuerdo de la Facultad, también colaciones.
	Agent's justification	Se añadieron las formas femeninas "anfitrionas" y "guías" a "anfitriones" y "guías", respectivamente, y se incluyó "Las/os" antes de "estudiantes" para asegurar la inclusión de ambos géneros.
	Explanation (English)	**The agent identified the inclusion of both genders in the change from (*Los*) to (*Las/os*) and from (*anfitriones*) to (*anfitrionas/es*)**
Example III	Original Sentence	Los trabajos que realizan **los colaboradores** son muy variados, pueden hacer tours, recibir colegios invitados, apoyar en charlas o talleres, etc.
	Corrected Sentence	Los trabajos que realizan **las/os colaboradoras/es** son muy variados, pueden hacer tours, recibir colegios invitados, apoyar en charlas o talleres, etc.
	Agent's justification	Se incluyeron las formas femeninas y masculinas de los sustantivos y el artículo "las/os" para asegurar la inclusión de colaboradoras y colaboradores.
	Explanation (English)	**The agent identified that changing (*los colaboradores*) to (*las/os colaboradoras/es*) makes the noun phrase inclusive of both feminine and masculine forms.**
Example IV	Original Sentence	**Los ganadores** pasarán a formar parte de una red mundial de instituciones dedicadas a "Empoderar a las personas".
	Corrected Sentence	**Quienes ganen** pasarán a formar parte de una red mundial de instituciones dedicadas a "Empoderar a las personas".
	Agent's justification	Se reemplazó el sustantivo "Los ganadores" por la oración "Quienes ganen", eliminando la connotación masculina implícita en el sustantivo plural masculino.
	Explanation (English)	**The agent identified that replacing (*Los ganadores*) with (*Quienes ganen*) removes the masculine plural noun and replaces it with a gender neutral reference.**

example.

Methodology

We next describe the proposed methods to address gender debiasing. The first method is a fined-tuned language model (LM) trained on a TST task using the produced dataset which runs locally. The second method is an LLM-based agent that addresses the problem via a few-shot learning.

Method 1: A fine-tuned language model for gender debiasing

We considered the Falcon-7B-Instruct model[3], a variant of Falcon-7B [28] fine-tuned on chat and instruction data. Trained on large volumes of data, this decoder-type model is expected to address our particular Seq2Seq task. This is because each token in our dataset described above is generated conditionally on the previous ones: our sentences correspond to input documents along with the instruction `remove gender bias`. Preliminary tests revealed that the public version of Falcon-7B was not able to successfully identify and remove gender bias from our test sentences. Therefore, Falcon-7B was finetuned for our case study using the dataset produced by the human experts; we emphasise that the LLM-based augmentations described in the previous section were not used for finetuning Falcon-7B.

We trained Falcon-7B-Instruct using Parameter-Efficient Fine-Tuning (PEFT), implemented through the `peft` package[4]. This package allows for efficient adaptation of LMs to various tasks by adjusting only a small number of additional parameters while freezing the remaining ones, thereby significantly reducing computational and storage costs[5]. Within PEFT, we employed Low Rank Adaptation (LoRA) [29] alongside its computationally-efficient quantised form (QLoRA) [30] with 4-bit quantisation, range of 16, alpha of 32, dropout of 0.05, and the remaining hyperparameters kept at the default values of the `TrainingArguments` class in the `transformers`. We used LoRA in the `query_key_value` layer of the attention blocks, and QLoRA for the weights in the linear layers of the attention blocks and those of the decoder. As a result, out of the model's original 6,926,439,296 parameters, only 4,718,592 were trainable (0.07%). Furthermore, we used a learning rate of $2 \cdot 10^{-4}$, the cross entropy loss and a batch size of 16 examples (input-output sequences) over 5 epochs with a cosine scheduler, resulting on 20.2 hours of training using a single GPU NVIDIA GeForce RTX 2080 with 8 GB, an Intel Core i7-10700KF at 3.8 GHz and 16 GB of RAM.

To prepare the data for training, the tokeniser available for the model was considered with a maximum sequence length of 250. Regarding the prompt, since Falcon is a generative model, each d_{in} from C_{train} was used to produce the following instruction termed d'_{in}:

d'_{in} := <human>: ¿Puedes reescribir el siguiente texto sin sesgo de género? d_{in}. <assistant>: d_{out}

Here, the human instructs *"Can you rewrite the following text without gender bias?"* in Spanish. Likewise, both for C_{test} and for the inferences of the adjusted model, each sentence d_{in} was used for the

[3] `huggingface.co/tiiuae/falcon-7b-instruct`

[4] `github.com/huggingface/peft`

[5] `huggingface.co/docs/peft/index`

instruction given in the following box (d'_{in}). This way, the generated output $\hat{d}_{\text{out}}$ corresponds to the text generated following `<assistant>`:.

> d'_{in} := <human>: ¿Puedes reescribir el siguiente texto sin sesgo de género? d_{in}. <assistant>:

Method 2: Debiaser, an agentic system for gender bias detection and mitigation

We designed and implemented a multi-agent system for gender debiasing, termed Debiaser, leveraging two LLMs: Claude 3.5 Haiku [26] and GPT-4o mini [27]—see Table II. The agentic system employs independent agents working collaboratively to perform the debiasing task. Adopting a few-shot learning approach [31], each agent is provided with a description of their specific task (e.g., bias detection, bias analysis, or bias mitigation), an available set of tools, and only a few examples of biased-debiased sentences with an explanation of the debiasing procedure.

Agentic system workflow

Debiaser implements a decision-based control flow for analysing and neutralising gender-biased text in two stages: detection and mitigation. By orchestrating the capabilities of LLMs and identifying subtasks, the agents leverage specialised tools to provide guidance and structure the output at each step. The Debiaser's control flow is organised into four stages, ensuring a clear and logical progression from text input to bias correction.

1) **Initial bias detection:** The bias detector agent receives a text input (a possibly biased sentence) and uses the tool `gender_bias_classifier` to determine whether the text contains gender bias. This tool performs a binary classification of the text, assessing whether it contains gender bias or not.
    a) If the text is classified as unbiased, the process terminates and the original text is returned unchanged. In this case, no modifications are required from Debiaser.
    b) If the text is classified as biased, the process continues to the next stage for further processing.
2) **Bias analysis:** When bias is detected, the same agent conducts a detailed analysis of the text. This involves span-based classification [32], [33], to identify the specific biases involved and pinpoint the exact text spans where this occurs. This stage also uses the tool `gender_bias_classifier` and generates a structured output that includes:
    a) The type of bias (e.g., stereotyping, sexism),
    b) The bias span (the specific text segment where the bias occurs),
    c) A bias confidence score (ranging from 0 to 1) indicating the certainty of the detection.

3) **Bias mitigation:** A bias neutraliser agent then uses the output from the bias analysis stage and generates a bias-free version of the text using the tool `debiasing_tool`. The tool explicitly elicits and records the model's reasoning process using Chain-of-Thought prompting [34], [35]. This technique reveals the implicit decision-making process which the LLM uses when debiasing the text, a key component in the pursuit of transparency and interpretability.
4) **Self-reflection:** This is a final stage where the Debiaser critic agent reviews the debiased text and the rationale used for the corrections. Here, the agent assesses the effectiveness of its debiasing procedure and identifies potential areas for improvement. This way, the bias neutraliser agent can refine its debiasing strategies and improve the debiased text by calling the `debiasing_tool` again, but incorporating the feedback of the critic.

The modular architecture of the presented agentic system is flexible, enabling the workflow to incorporate additional functionalities if needed. For instance, we can introduce a *confidence threshold filter* to condition the debiasing on the confidence scores generated by the bias analysis tool. This way, the agentic system would only neutralise text when the bias score surpasses certain threshold. In addition, the self-reflection stage can be run iteratively, e.g., N times, to refine the debiasing process, or until the critic agent detects no room for improvement in the debiased text. This will allow the Debiaser to take corrective actions based on its own feedback, at the sacrifice of latency and cost for the additional LLM calls.

An illustrative example of Debiaser in action

Figure 3 shows a diagram of Debiaser's work flow for an example text input that demonstrates a common but subtle form of gender bias in Spanish communication:

> *"Les doy a <u>todos</u> la bienvenida a este nuevo semestre de primavera, y reitero mi compromiso, a través de mi equipo de Facultad y de Escuela, a prestar el apoyo que sea necesario para que este semestre podamos afrontarlo de mejor manera."*

While this welcoming message appears neutral, it employs the Spanish masculine generic *'todos'* (everyone), which can inadvertently exclude non-male individuals despite being grammatically standard. A common challenge of Romance languages where masculine forms traditionally serve as defaults for mixed-gender groups.

The bias detector agent detects the presence of gender bias in the text and performs the span-based analysis to identify the biases. The `gender_bias_classifier` tool returns the following structured output:

- **Bias types:** `[GENERIC_PRONOUNS, EXCLUSIONARY_TERMS]`

- **Bias spans:** `["todos", "mi equipo de Facultad y de Escuela"]`
- **Bias scores:** `[0.85, 0.75]`

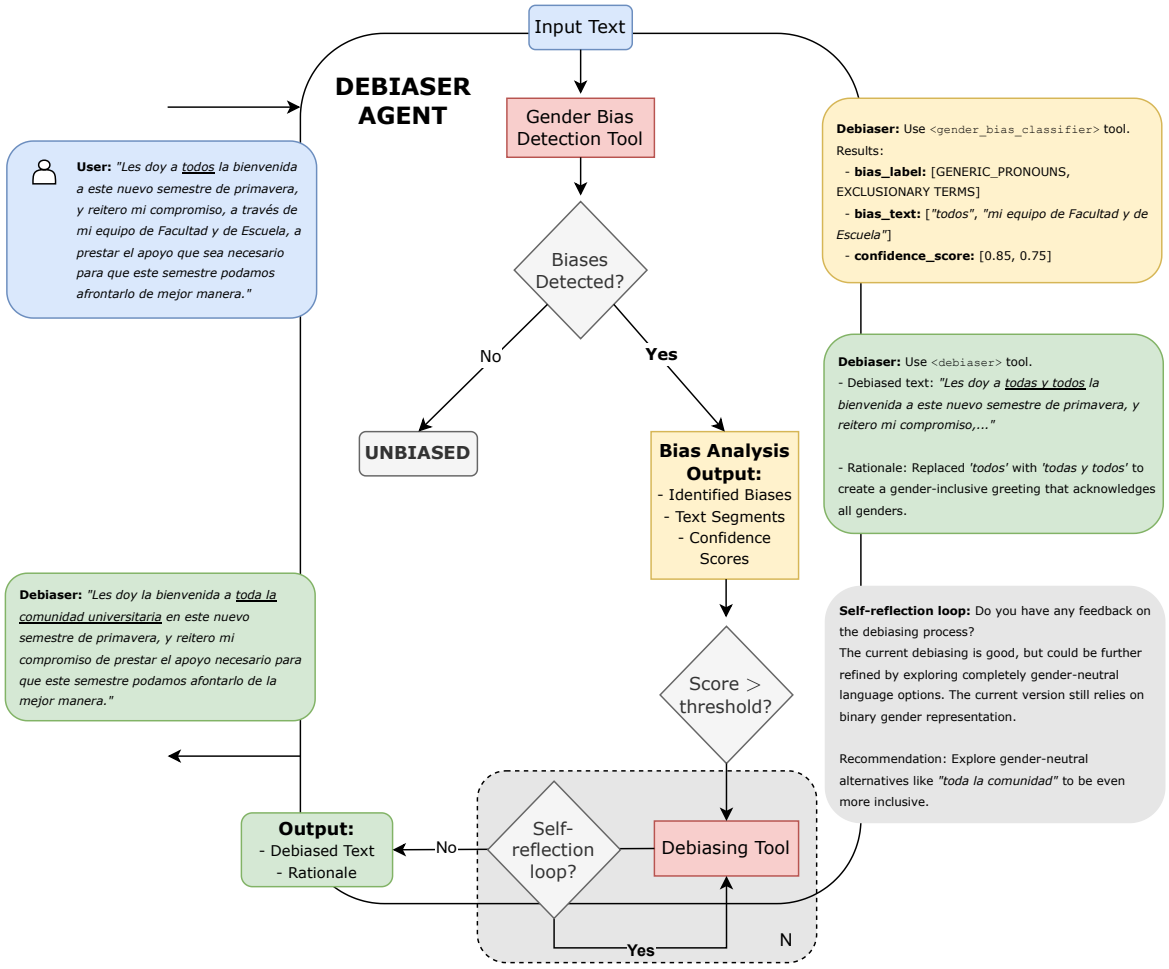

Fig. 3: Debiaser agent workflow with iterative refinement. Input text is analysed for gender bias with confidence scores. When biases are detected above threshold, a debiasing tool generates revised text with rationale. A self-reflection loop evaluates the output and suggests improvements, triggering re-invocation of the debiasing tool if needed. This process repeats up to N iterations until no further refinements are suggested, as shown in the example progression from *'todos'* to *'todas y todos'* to *'toda la comunidad universitaria'*

Then, the `debiasing_tool` based on the above analysis, generated this neutralised version of the text:

> *"Les doy a todas y todos la bienvenida a este nuevo semestre de primavera, y reitero mi compromiso, a través de mi equipo de Facultad y de Escuela, a prestar el apoyo que sea necesario para que este semestre podamos afrontarlo de mejor manera."*

The bias neutraliser agent addresses both detected issues: it replaces the masculine generic *"todos"* with the binary-inclusive *"todas y todos"* to explicitly acknowledge all genders, while maintaining the original institutional context and message structure. Notably, the detection of *"mi equipo de Facultad y de Escuela"*, which means "my team at FCFM", as `EXCLUSIONARY_TERMS` represents a potential false positive, as this phrase describes organizational structure rather than exclusionary language. This highlights the need for more nuanced bias classification to distinguish between genuinely problematic content and neutral institutional references.

However, as the Debiaser critic agent identified in the self-reflection loop, this initial debiasing, while effective, could be further refined:

> *"The current debiasing is good, but could be further refined by exploring completely gender-neutral language options. The current version still relies on binary gender representation.* ***Recommendation: Explore gender-neutral alternatives like 'toda la comunidad' to be even more inclusive****"*

Providing such feedback to the bias neutraliser agent enables it to move beyond binary gender representation and generate a more comprehensively inclusive revision. This iterative process ensures that the agent progresses from basic inclusion to sophisticated gender neutrality, balancing bias mitigation with linguistic naturalness while avoiding unnecessary removal of factual content. Therefore the bias neutraliser agent generates a more refined and contextually appropriate debiased text:

> *"Les doy la bienvenida a toda la comunidad universitaria en este nuevo semestre de primavera, y reitero mi compromiso de prestar el apoyo necesario para que este semestre podamos afrontarlo de la mejor manera.."*

Table IV provides examples of input-output sentences for the Debiaser agent, with the same agent's *rationale* as an explanation in English (see Fig. 3).

The role of the tools in the agentic system

Designing tools such as a calculator or an API [36] is —at least conceptually— very clear. However, the construction of tools with higher-level objectives, such as those for detecting and mitigating gender biases, is far from straightforward. In order for the LLM to effectively implement these tools, they should be defined in a clear and direct way.

For Debiaser, the tools play a crucial role: they focus on structuring the zero- or few-shot learning capabilities of the LLM [31], rather than providing additional information from external computation. Furthermore, Debiaser's tools orchestrate the in-context learning capabilities of the LLMs and providing pathways to solve the detection and mitigation of gender biases. We provide a summary of the

TABLE IV: Examples of Debiaser: input and output sentences with the agent's *rationale* as a means of explanation in English.

Input	**Output**	**Agent's Rationale (English)**
Invitamos a los profesores del departamento a la reunión mensual.	Invitamos al personal docente del departamento a la reunión mensual.	"Replaced the gendered phrase '*los profesores*' with '*personal docente*' to use a gender-neutral collective term that encompasses all educators."
Estimados estudiantes de Pregrado,Junto con saludar les invitamos a la charla abierta...	Estimada comunidad estudiantil de Pregrado, junto con saludar les invitamos a la charla abierta...	"Original phrase '*Estimados Estudiantes*' contained a Generic Pronouns bias. Replaced with '*comunidad estudiantil*' to create a more inclusive, gender-neutral salutation."
Se solicita a los alumnos entregar su informe antes del viernes.	Se solicita a la comunidad estudiantil entregar su informe antes del viernes.	"Original phrase '*los alumnos*' contains a gender-exclusive masculine generic term. Replaced with the gender-neutral collective pronoun '*comunidad estudiantil*' to ensure inclusive language while maintaining the original semantic meaning and communication intent."

`gender_bias_classifier` tool below, which includes its name, a detailed description of its functionalities, the expected input scheme and the fields to be produced as output. Notice that the tool guides the LLM in detecting gender biases in a step-by-step fashion, forcing the decision-making procedure to be transparent and clear. As a consequence, the decisions made by the agent can be interpreted by a human which can, for instance, verify the quality of the corrections, as shown in the example in Figure 3. We emphasise that, given appropriate definitions and categorisations, the proposed multi-agent decision flow can be repurposed for other languages and datasets.

Tool definition: Gender Bias Classifier

Name: `gender_bias_classifier`

Description: "Identify (if any) one or more of the following gender biases in the text:

- `GENERIC_PRONOUNS`: Generic pronouns bias is the use of gender-specific pronouns when a gender-neutral pronoun would be more appropriate.
- `STEREOTYPING_BIAS`: Stereotyping bias is the use of stereotypes to make assumptions about a person's abilities, interests, or characteristics.
- `SEXISM`: Sexism can be defined as discrimination, stereotyping, or prejudice based on one's sex.
- `EXCLUSIONARY_TERMS`: Exclusionary terms bias is the use of terms that exclude or marginalise a particular gender, often by using male-oriented terms as the default.
- `SEMANTIC_BIAS`: Semantic bias is the use of words or phrases that have a gendered connotation, which can reinforce stereotypes or biases.
- `UNBIASED`: No gender bias detected. Don't return None if no bias is detected."

Input Schema:

```
{
    "properties": {
        "bias_label": {
            "description": "List of biases labels detected within the text.",
            "title": "bias_label",
            "type": "array"
        },
        "bias_text": {
            "description": "A list with the specific parts of the text that trigger
                a gender bias detection specified in bias_label. If no bias is
                detected, the text is considered unbiased and use None.",
            "title": "bias_text",
            "type": "array"
        },
        "score_label": {
            "description": "A list with the classification score of the bias_label
                detected in the text, ranging from 0.0 to 1.0.",
            "title": "score_label",
            "type": "array"
        }
    },
    "required": ["bias_label", "bias_text", "score_label"],
    "title": "MultiLabelGenderBiasClassifier"
}
```

Evaluation

The proposed methods (fine-tuned LM and agent LLM) were quantitatively evaluated for bias identification and mitigation using the constructed dataset from communications within FCFM. We consider classical reference-based metrics and an alternative assessment procedure using LLMs.

Bias detection and sensitivity

We first evaluated the ability of both approaches to identify a sentence as biased. In the case of the agent this is precisely the output of the first stage (initial bias detection), while for the fine-tuned LM it is safe to assume that no bias was found if and only if the output is equal to the input. Table V shows the detection percentage, where the ground truth is given by the expert annotations in the dataset.

TABLE V: Bias detection performance for each proposed method: Number of sentences and percentage of detected sentences for each method and input class (biased, unbiased and unable-to-bias).

		Fine-tuned LM		LLM Agent	
subset	number of sentences	detected	not detected	detected	not detected
biased	58	37 (63.8%)	21 (36.2%)	47 (81.0%)	11 (19.0%)
unbiased	271	50 (18.5%)	221 (81.5%)	208 (76.8%)	63 (23.2%)
unable-to-bias	453	19 (4.2%)	434 (95.8%)	136 (30.0%)	317 (70%)

Notice that the LLM agent is considerably more sensitive: its 81% of correct detection in cases with bias is in sharp contrast with the 76.8% of cases where a bias was observed that was not reported by the annotators. Conversely, the fine-tuned LM failed to make modifications in 21 out of 58 biased inputs but exhibited a large percentage of correct negatives at 95.8% and 81.5% for unable-to-bias and unbiased inputs respectively. These figures suggest that the fine-tuned LM tends to under-identify biases, while the agent LLM tends to over-identify them. Generally speaking, the applicability of each model will depend on the sensitivities needed. For instance, a model that suggests changes in real time may benefit less from a less sensitive model. Moreover, a false-positive bias detection can be dealt with in the debiasing stage, while a false-negative means that the sentence is not analysed. In the rest of this section, we assess the modifications introduced by both systems to the sentences and whether they change their conveyed meaning. These figures will also be used to construct an aggregate performance metric at the end of this section.

Reference-based evaluation for sentences with detected bias

When reference text is available, previous approaches to Seq2Seq have relied on metrics such as Bilingual Evaluation Understudy (BLEU) metric [37] or Recall-Oriented Understudy for Gisting Evaluation

(ROUGE) [38] to compare the generation of an automated system to said reference. These metrics assess the proportion of shared tokens between the candidate and the target text, obtaining a similarity score between the two. Even though these metrics fail at capturing semantic similarities beyond exact word matching, they are appropriate to some extent. We expect that a successful gender-debiasing method can modify only tokens that represent gender bias, while keeping the rest of the message intact. However, since the annotated reference and system output are expected to have several common tokens, we need a method to assess whether the system is becoming closer to the references or not.

To address this challenge, we consider BLEU on the subset C_{test} defined earlier. Since in the context of gender bias mitigation, the overlap of n-grams between d_{in} and d_{out} is significant, we propose a novel performance metric termed dBLEU, given by:

$$\text{dBLEU}(d_{\text{in}}, d_{\text{out}}, \hat{d}_{\text{out}}) = \text{BLEU}(d_{\text{out}}, \hat{d}_{\text{out}}) - \text{BLEU}(d_{\text{out}}, d_{\text{in}}). \tag{1}$$

This way, we obtain a differential, or *relative*, BLEU that avoids the inherent overlapping of tokens in our setting. We expect positive dBLEU values for biased d_{in}, meaning that the similarity between d_{out} and $\hat{d}_{\text{out}}$ is higher than that between d_{out} and d_{in}. As an alternative to dBLEU, our preliminary experiments also considered a differential version of BERTScore [39], however, the results of this metric were virtually identical to dBLEU and are thus omitted in our experimental evaluation.

TABLE VI: Performance of proposed methods using dBLEU on sentences where bias was detected: Fine-tuned LM (37 sentences) and LLM Agent (47 sentences). Results are deemed successful or unsuccessful depending on the sign of dBLEU.

	Fine-tuned LM (37)		LLM Agent (47)	
Metric	unsuccessful	successful	unsuccessful	successful
dBLEU	9	22	39	8

Table VI shows dBLEU for both methods, where tokens were identified using NTLK's `word_tokenize`. For the fine-tuned LM, 22 outputs were closer to the reference as desired, while 9 outputs were closer to the input instead. Notice that there were 6 undefined cases for the fine-tuned LM where dBLEU is equal to zero, this means that the sentences are equal *under the tokenisation* and thus subtle changes such as single-letter replacement are potentially ignored. The LLM agent only reported 8 correct outputs and 39 incorrect ones according to dBLEU, that is, with positive and negative dBLEU respectively. Upon inspecting the debiased sentences, our understanding is that the changes made by the LLM agent corrector are stronger than the LM and thus result in negative values for dBLEU, however, they are not necessarily

incorrect. Therefore, we complement our assessment with an LLM-based evaluation, termed LLM-judge, presented in the following section.

Agentic-based evaluation: LLM-judge

LLMs are increasingly used to evaluate text generation tasks, thanks to their ability to process and identify complex language patterns. Via prompting, LLMs can be provided with instructions that help humans to assist in subjective tasks in a scalable and hopefully consistent way. In this case, we need tools that can complement the role of reference-based metrics. In particular, we argue that LLMs can be used to identify words that have changed to address gender bias, and also whether the semantics of the original message are maintained. Therefore, we instruct two LLM *judges* built on GPT4 [10]: the first one, termed *debiasing*, evaluates whether biased sentences were successfully debiased, while the second one, termed *semantics*, evaluates if sentences that need not be modified in fact maintained their semantics. We present their (summarised) prompts as follows.

LLM-judge: debiasing

You will be given a `input_text` and `system_output` pair, both in Spanish.

Your task is to assess whether the `system_output` text has the same meaning as `input_text`, but without gender bias.

For this, take into account the following guidelines:

A text presents gender bias if it has one or more of the following types:

1) Use of (non-)generic pronoun: This refers to the tendency to use pronouns or other linguistic forms that imply a specific gender, often masculine, when referring to a group of people in general or to a person whose gender is not known. This can lead to the invisibility or underrepresentation of women and other genders in language, reinforcing gender perceptions and inequalities in society. To construct unbiased versions, one can: make use of typographic strategies, such as: /, (), @; not refer to persons but to actions; or make use of general references, such as "the people", "the community", "the student body", etc. This type of bias is the most common.
2) Occupations: Corresponds to the bias that occurs when occupations or professions are stereotypically associated with a gender. It does not correspond to bias if a specific person is being identified.
3) Emotions: Corresponds to bias when stereotypically associating an emotion or feeling to a specific gender.
4) Social and behavioral stereotyping: Social stereotypes represent traditional gender roles that reflect social norms. Behavioral bias refers to phrases that contain attributes and traits that are generalised towards a person or gender.
5) Adjectives/Personality: Refers to the biased use of certain adjectives to describe behavior or personality distinctively for women and men.
6) Sexism: Subordinate position of women. Refers to attitudes, beliefs, and systems that perpetuate gender inequality and place women in a position of inferiority or subordination to men.

The "Bias outcome" answer needs to be one of the following options:

- (X) "Unsuccesful debiasing": `input_text` was biased in terms of gender and this bias has NOT been corrected in `system_output`.
- (Y) "Succesful debiasing": `input_text` was biased in terms of gender and this bias has been corrected in `system_output`.

Both bias and semantics outputs can be justified. Provide your answer as follows:

—

Bias outcome: (your choice between (X) and (Y))

Justification::: —

LLM-Judge: semantics

You are an expert evaluator tasked with assessing whether the output of an automated system alters the semantic meaning of a given input text. Additionally, you must determine whether the output introduces gender bias, following the provided gender bias guidelines.

Given:

- Input Text (human-written)
- Output Text (automated system-generated)
- Gender Bias Guidelines (provided as an argument)

Your task is to determine whether the output text negatively changes the input text based on two factors:

1) **Semantic Preservation**: Check if the output faithfully retains the meaning of the input. Any major changes, omissions, or additions that alter meaning should be considered a negative change.
2) **Gender Bias**: Analyze the output against the provided gender bias guidelines. If the output introduces bias that was absent in the input, this counts as a negative change.

The answer needs to be one of the following options:

- (X) "Negative modification": The output either (a) significantly alters the semantic meaning of the input or (b) introduces gender bias.
- (Y) "No negative modifications": The meaning of the input is preserved without introducing gender bias.

Both bias and semantics outputs can be justified. Provide your answer as follows:

—

Bias outcome: (your choice between (X) and (Y))

Justification::: —

LLM-judges were provided with examples of successful and incomplete debiasing based on explanations from the augmented training dataset, and qualitatively validated using the augmented dataset. Table VII shows the evaluations of both models and all sentences recognised as needing debiasing: LLM-judge *debiasing* was used for the the biased subset, and LLM-judge *semantics* for the unbiased and unable-to-bias subsets. After a detailed inspection of the LLM-judges justifications and of input-output pairs, we could identify patterns in both models' limitations. On the one hand, several biased sentences that were deemed unsuccessfully debiased by the bias judge correspond to corrections that only addressed part of the bias present in the input text, particularly for the LLM agent. In some cases, the fine-tuned model also missed parts of the input or slightly modified the text in a way that did not incur any meaningful change. Despite these limitations in performance, both models were able to correct bias in several cases. On the other hand, the semantics judge found that modifications by the fine-tuned LM to unbiased text

were mostly insignificant, like changes in punctuation. The LLM agent, however, had a tendency to prefer certain unbiased forms of writing in Spanish and would unnecessarily change the text accordingly. While the majority of these modifications maintained the message, sometimes they were deemed to change the exact meaning of specific words by the judge.

TABLE VII: Performance of proposed methods using an LLM-judges for sentences recognised as needing debiasing by each model.

		Fine-tuned LM		LLM Agent	
subset	LLM-judge	unsuccessful	successful	unsuccessful	successful
biased	*debiasing*	17	20	19	28
unbiased	*semantics*	7	43	64	144
unable-to-bias	*semantics*	3	16	63	73

Overall evaluation

By aggregating the subset-based assessments in Tables V and VII, we quantify the performance of the proposed methods across the entire dataset. To this end, we define **overall performance** as the number of successfully processed sentences divided by the total number of analysed sentences, i.e., $\frac{CU+MU+CD}{T}$, where CU is the number of sentences correctly identified as unbiased or unbiasable, MU are sentences modified but having their semantics maintained, and CD are sentences correctly debiased. The first term in the numerator comes from Table V and the other two terms from the LLM-judges in Table VII. Performance figures for both methods are shown in Table VIII. Results clearly show that the fine-tuned LM succeeds for the majority of cases.

TABLE VIII: Overall performance of the proposed methods

	Fine-tuned LM	LLM agent
overall performance	93.9%	80%

Discussion

Overall contribution. Building upon recent advances for detecting and mitigating linguistic bias in AI, this work presents a novel framework specifically designed for detecting and correcting gender-biased text in university communications. We validated our framework experimentally through a case study at the School of Engineering (FCFM) at Universidad de Chile with a challenging Spanish corpus. Our

methodological contribution comprises the production of an *ad hoc* purpose-built dataset; a fine-tuned language model for gender debiasing; agent-based pipelines for augmenting the dataset, few-shot debiasing and performance quantification; and the definition of dedicated performance indicators to experimentally validate our work. This work aligns with emerging research demonstrating the potential of AI to identify and address biases in language—an essential step towards fostering inclusive academic and institutional environments, especially in educational contexts. The modularity of these tools facilitates its application to other datasets and linguistic contexts.

Dataset constructed. The parallel-sentence Spanish corpus for gender bias mitigation was built by first extracting the text from public forums at Universidad de Chile, and then curating it by expert annotators to produce biased/unbiased versions of the original sentences. This provides a reliable resource for fine-tuning a language model for gender bias mitigation in Spanish as a text-style transfer task. This dataset also allowed us to perform a statistical analysis to confirm a decrease in the frequency of gender bias in the communications of the FCFM "News" portal over time since 2015. The dataset is available the project's repository: `github.com/GianniCatBug/spanish-gender-debias`.

Performance and impact. Our experimental assessment confirms that the proposed models exhibit an overall performance of 80% (LLM agent) and 93.9% (fine-tuned LM), while their ability to recognise a biased sentence as such is 81% and 63.8% for the agent and LM respectively. The number of successful corrections is, however, not promising across both methods and subsets of sentences, as shown in Table VII. Despite the limitations of the fine-tuned and agentic methods, our study validates the proposed approach as a proof-of-concept for real-time mitigation of gender bias in Spanish academic communications. In particular, the deployment strategy is key to determine if the higher sensitivity of the LLM is a desirable feature. While each method shows its own advantage, we particularly highlight the performance of the local and light-weight solution posed by the fine-tuned LM. Indeed, the action of not changing requires an advanced contextual understanding of modifications. In this sense, the capacity showed by both models was remarkable, particularly for a local solution. To conclude, our results pave the way for further improvements regarding the format of the dataset and the definition of the agentic tools.

Future work. Additional research could extend our approach by combining the fine-tuning procedure and LLM agent. Indeed, the annotated parallel sentences may be passed to the LLM to (auto) adjust its weights, or the explanatory capabilities of the agent could be exploited but with the corrections carried out by the local LM. Although this would result in higher computational (and thus financial) costs, we believe it could lead to a more refined model. In this context, another direction worth exploring is how the contextualised embeddings change in the models pre and post fine-tuning using the parallel dataset. Furthermore, the confidence score of the agent can be used to achieve further alignment with

human criteria as well as to understand the limitations of LLM-based approaches. In this sense, future investigations can also be devoted to determine to which extent the biases present in the LLMs, perhaps inherited from their original training procedure, affect the debiasing of human-generated and how this can be minimised.

References

[1] K. Stanczak and I. Augenstein, "A survey on gender bias in natural language processing," *arXiv preprint arXiv:2112.14168*, 2021.

[2] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proceedings of the national academy of sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.

[3] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *arXiv preprint arXiv:1809.01496*, 2018.

[4] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[5] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," *arXiv preprint arXiv:1906.00591*, 2019.

[6] C. Borchers, D. S. Gala, B. Gilburt, E. Oravkin, W. Bounsi, Y. M. Asano, and H. R. Kirk, "Looking for a handsome carpenter! debiasing gpt-3 job advertisements," *arXiv preprint arXiv:2205.11374*, 2022.

[7] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint arXiv:1804.06876*, 2018.

[8] J. Doughman and W. Khreich, "Gender bias in text: Labeled datasets and lexicons," *arXiv preprint arXiv:2201.08675*, 2022.

[9] E. Fersini, P. Rosso, M. Anzovino *et al.*, "Overview of the task on automatic misogyny identification at ibereval 2018." *Ibereval@ sepln*, vol. 2150, pp. 214–228, 2018.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, and F. L. e. a. Aleman, "Gpt-4 technical report," 2023.

[11] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[12] M. Toshevska and S. Gievska, "A review of text style transfer using deep learning," *IEEE Transactions on Artificial Intelligence*, 2021.

[13] Y. Hitti, E. Jang, I. Moreno, and C. Pelletier, "Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 8–17.

[14] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, and Z. Berjawi, "Gender bias in text: Origin, taxonomy, and implications," in *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 2021, pp. 34–44.

[15] K. Crawford, "The trouble with gender bias," 2017. [Online]. Available: https://www.youtube.com/watch?v=fMym_BKWQzk

[16] E. Vanmassenhove, C. Emmery, and D. Shterionov, "NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Association for Computational Linguistics, Nov. 2021, pp. 8940–8948. [Online]. Available: https://aclanthology.org/2021.emnlp-main.704/

[17] F. Jourdan, T. T. Kaninku, N. Asher, J.-M. Loubes, and L. Risser, "How optimal transport can tackle gender biases in multi-class neural network classifiers for job recommendations," *Algorithms*, vol. 16, no. 3, 2023. [Online]. Available: https://www.mdpi.com/1999-4893/16/3/174

[18] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence-to-sequence models," *arXiv preprint arXiv:1707.01161*, 2017.

[19] S. Rao and J. Tetreault, "Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer," *arXiv preprint arXiv:1803.06535*, 2018.

[20] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Computational Linguistics*, vol. 48, no. 1, pp. 155–205, 2022.

[21] Z. Hu, R. K.-W. Lee, C. C. Aggarwal, and A. Zhang, "Text style transfer: A review and experimental evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 24, no. 1, pp. 14–45, 2022.

[22] X. Ma, M. Sap, H. Rashkin, and Y. Choi, "Powertransformer: Unsupervised controllable revision for biased language correction," in *EMNLP*, 2020.

[23] E. Reif, D. Ippolito, A. Yuan, A. Coenen, C. Callison-Burch, and J. Wei, "A recipe for arbitrary text style transfer with large language models," *arXiv preprint arXiv:2109.03910*, 2021.

[24] R. W. Bybee, "What is stem education?" *Science*, vol. 329, no. 5995, pp. 996–996, 2010. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1194998

[25] G. Team, P. Georgiev, V. I. Lei, R. Burnell, and L. B. et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530

[26] Anthropic, "The claude 3 model family: Opus, sonnet, haiku," 2024, accessed: 2024-10-22. [Online]. Available: https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf

[27] OpenAI, "Gpt-4o mini: advancing cost-efficient intelligence," 2024, accessed: 2024-07-18. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[28] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, and G. Penedo, "Falcon-40B: an open large language model with state-of-the-art performance," 2023 (accessed 8/5/2025). [Online]. Available: https://huggingface.co/tiiuae/falcon-7b-instruct

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[30] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[32] M. G. Sohrab and M. Miwa, "Deep exhaustive model for nested named entity recognition," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2843–2849. [Online]. Available: https://aclanthology.org/D18-1309/

[33] N. T. H. Nguyen, M. Miwa, and S. Ananiadou, "Span-based named entity recognition by generating and compressing information," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1984–1996. [Online]. Available: https://aclanthology.org/2023.eacl-main.146/

[34] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[35] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023. [Online]. Available: https://arxiv.org/abs/2205.11916

[36] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, "Webgpt: Browser-assisted question-answering with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2112.09332

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[38] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Association for Computational Linguistics, 2004, pp. 74–81.

[39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

Biographies

Gianina Salomó-López is a Data Scientist with a Master's degree in Data Science from Universidad de Chile, and a Diploma in Statistics from Pontificia Universidad Católica de Chile. She has worked in various areas of data science and programming, applying machine learning models and data analysis in the tech sector. Additionally, she has been a lecturer and content creator at Desafío Latam. Her interests focus on machine learning, deep learning, and natural language processing.

Cristóbal Alcázar is a Data Scientist with a Master's degree in Data Science from Universidad de Chile. His research interests lie in diffusion models and how to adapt them, specifically using reinforcement learning. In addition, he is a co-founder of Vendie a conversational e-commerce solution. He has professional experience working in Fintech and macroeconomic statistics at the Central Bank of Chile. Prior studies include a BSc in business administration and a master's in finance.

Roberto Barceló is an Electrical Engineer and Data Scientist with a Masters in Data Science from Universidad de Chile. Currently a data analyst at Solver Ingenieros, he develops machine learning models focused on industrial process optimization. His primary interests lie in generative models, representation learning, and reinforcement learning, specifically in understanding model representations and leveraging this learned knowledge for broader applications.

Camilo Carvajal Reyes is a Mathematics PhD student at Imperial College London. He holds a Masters degree in Data Science and a BSE in Mathematical Engineering from Universidad de Chile, as well as a Diplôme d'Ingénieur from CentraleSupélec, France with a mention in Research. He has been a teaching assistant for courses including Machine Learning, Deep Generative Models and Stochastic Simulation Laboratory. His research interests are ethics and safeness in generative models and mathematical modelling of language and images. Additionally, he is the cofounder AEDIA, a student initiative for ethics in artificial intelligence at the University of Chile.

Darinka Radovic is the Deputy Director of the Diversity and Gender Office and a researcher at the Center for Mathematical Modeling, both part of the Faculty of Physical and Mathematical Sciences at the University of Chile. She completed her PhD in Education at the University of Manchester in the United Kingdom. Her research and professional work have been focused on school and university education, primarily addressing gender issues in STEM (Science, Technology, Engineering, and Mathematics). This includes research on affirmative policies for women's participation in STEM fields, estimating educational gaps, and developing identities related to education and professions. In 2024, she was the coordinator of the Diploma in Gender Perspective Integration in STEM at the University of Chile.

Felipe Tobar Felipe Tobar is an Associate Professor in Machine Learning at the Department of Mathematics and I-X, at Imperial College London. Previously, he was an Associate Professor at Universidad de Chile and the Director of the Initiative for Data and Artificial Intelligence of the same Institution. Felipe was a postdoc at the Machine Learning Group, University of Cambridge, during 2015 and received a PhD in Signal Processing from Imperial College London in 2014. Felipe's research interests lie in the interface between Machine Learning and Statistical Signal Processing, including approximate inference, Bayesian nonparametrics, spectral estimation, optimal transport, diffusion models and Gaussian processes.