"Women do not have heart attacks!" Gender Biases in Automatically Generated Clinical Cases in French

Fanny Ducel, Nicolas Hiebel, Olivier Ferret, Karën Fort, Aurélie Névéol

"Women do not have heart attacks!" Gender Biases in Automatically Generated Clinical Cases in French

Fanny Ducel[1], Nicolas Hiebel[1], Olivier Ferret[2], Karën Fort[3], Aurélie Névéol[1]
[1]Université Paris-Saclay, CNRS, LISN, Orsay, France
[2]Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
[3]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
Correspondence: fanny.ducel@universite-paris-saclay.fr

Abstract

Healthcare professionals increasingly include Language Models (LMs) in clinical practice. However, LMs have been shown to exhibit and amplify stereotypical biases that can cause life-threatening harm in a medical context. This study aims to evaluate gender biases in automatically generated clinical cases in French, on ten disorders. Using seven LMs fine-tuned for clinical case generation and an automatic linguistic gender detection tool, we measure the associations between disorders and gender. We unveil that LMs over-generate cases describing male patients, creating synthetic corpora that are not consistent with documented prevalence for these disorders. For instance, when prompts do not specify a gender, LMs generate eight times more clinical cases describing male (vs. female patients) for heart attack. We discuss the ideal synthetic clinical case corpus and establish that explicitly mentioning demographic information in generation instructions appears to be the fairest strategy. In conclusion, we argue that the presence of gender biases in synthetic text raises concerns about LM-induced harm, especially for women and transgender people.

1 Introduction

Healthcare is a field that is particularly exposed to societal and stereotypical biases. Studies document the presence and impact of biases carried out by healthcare professionals and clinical research (FitzGerald and Hurst, 2017). Biases are linked to patients' gender (Dwass, 2019), race (Williams and Wyatt, 2015), weight (Lawrence et al., 2021), sexual orientation (Albuquerque et al., 2016), gender identity (Drabish and Theeke, 2022), age (Chrisler et al., 2016), or socio-economic status (Arpey et al., 2017). These biases lead to underdiagnosis, serial misdiagnosis, and mistreatment of some disorders for patient categories that do not fit the "stereotypical patient" profile for the disorder in question.

Language models (LMs) are increasingly used in healthcare for preconsultation, clinical decision support, medication counseling, and jargon simplification, but also to help recruit patients for clinical trials or to train students in medical schools (Yang et al., 2023). LMs may also be used for synthetic text generation to support secondary use of health data. In this context, synthetic clinical cases are a source of shareable corpora that preserve patient confidentiality. However, the LMs used to produce synthetic texts contain, propagate, and even amplify stereotypical biases (Kirk et al., 2021). Thus, they create a risk of reinforcing stereotypes and harms. Generated biases could even feed health professionals' stereotypes (Adam et al., 2022) as biased systems influence users durably after exposure (Vicente and Matute, 2023). Real-world biases in data are one source of biases in LMs (Hovy and Prabhumoye, 2021) that can be reinforced by model biases.

In this study, we aim to characterize the presence of gender biases in generated clinical cases in French, addressing 10 common disorders – including some notable for gender stereotypes and biases. We distinguish between societal biases (e.g., women's heart conditions are often disregarded or mistaken for anxiety (Banks, 2008)) and medical prevalence (e.g., prostate cancer is more often found in men *vs.* women because the majority of people with a prostate are men).

Our contributions are the following:

- The fine-tuning of 7 LMs for controlled generation of synthetic clinical cases in French.
- A method leveraging morpho-syntactic gender markers for the automatic extraction of patient gender information from clinical cases.
- An evaluation of gender bias in 21,000 synthetic clinical cases addressing 10 disorders with various male/female prevalence.

- Recommendations for mitigation of biases in clinical case generation and an easy-to-implement mitigation strategy: the use of gendered prompts.

All data and code (including fine-tuned models, gender detection system, generated cases, and manually annotated cases) are freely available.[1] We find that LMs generate uneven proportions of masculine and feminine cases, favoring male patients. Further, these disparities do not mirror real-world gender prevalence: the proportion of masculine is increased for all disorders except prostate cancer. Interestingly, models generate fewer texts with male patients for stereotypically[2] feminine disorders (breast or ovarian cancer) when prompted to, than texts with female patients for stereotypically masculine disorders (prostate cancer). Finally, we discuss what an ideal, unbiased LM would output, taking into account real-world biases.

2 Related Work

2.1 Use of LMs in healthcare

LMs are increasingly used in healthcare to assist with a range of text-processing tasks including text classification, information extraction, or decision support (Hager et al., 2024), and specialized models were developed for the medical domain (Luo et al., 2022; Chen et al., 2023; García-Ferrero et al., 2024; Labrak et al., 2024b; Li et al., 2024).

Many applications aim at assisting health professionals with summarization tasks such as writing discharge summaries (Xu et al., 2024) and clinical notes (Nair et al., 2023; Ben Abacha et al., 2023a,b). Generative LMs are also used for patient interactions (Chowdhury et al., 2023).

In a more secondary use of healthcare data, LMs are used to create, augment, and share medical corpora to compensate for the limited availability of such resources (Ive et al., 2020; Amin-Nejad et al., 2020). Meoni et al. (2024) and Boulanger et al. (2024) use clinical keywords to guide the generation of synthetic clinical documents in English and French, while Hiebel et al. (2023) showed that automatically generated clinical cases in French can be used as an alternative to real clinical cases for training clinical entity recognition models with comparable performance.

2.2 Stereotypical biases

Biases in Natural Language Processing tools are defined as "skewed and undesirable association[s] in language representations which ha[ve] the potential to cause representational or allocational harms" (Barocas et al., 2017). Allocational harms occur when "[systems] allocate or withhold certain groups an opportunity or a resource", whereas representational harms occur when the depiction of members of certain groups is discriminatory and "reinforce subordination" (Barocas et al., 2017). As our focus is on stereotypical biases, they originate from "beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton and von Hippel, 1996), i.e., stereotypes.

In the context of LMs, a model can be defined as biased if it exhibits different, uneven behavior when one variable changes in the prompt (e.g., patient's gender), and/or if it does not replicate real-world data (e.g., gender prevalence for a given disorder).

Many research efforts show that LMs encapsulate and amplify stereotypical biases, both upstream – in the internal representations of the model (Choenni et al., 2021; Cao et al., 2022), and downstream – in real-world applications (Kirk et al., 2021; Wan et al., 2023; Kumar et al., 2024). Our work is contributing to recent studies exploring biases in the health sector (Kim et al., 2023; Zhang et al., 2024; Zack et al., 2024). However, to our knowledge, our study is the first to focus on biases in clinical case generation and more generally, on biases in medical texts in the French language.

3 Experimental Setup

3.1 Language models

This work focuses on French and 7 auto-regressive LMs of various sizes, from 4 families: BioMistral-7b-SLERP (Labrak et al., 2024b), BLOOM-1b1, BLOOM-7b1 (Le Scao et al., 2023), vigogne-2-7b, vigogne-2-13b (Huang, 2023), Llama-3.1-8B, and Llama-3.1-8B-Instruct (AI@Meta, 2024).

- BioMistral models are a series of open-source Mistral-based LMs, further pre-trained on medical corpora from PubMed Central Open

[1] See: https://github.com/FannyDucel/ClinicalCaseBias/

[2] Throughout the paper, we refer to gender and not sex. Thus, when we mention that some disorders, e.g., prostate cancer, is "a (more) masculine disorder", we mean to say that the majority of patients are men (because most people with a prostate are men) but acknowledge that some (transgender) women, intersex and non-binary people can also suffer from it (since they also have a prostate). However, we only focus on the two binary genders as in the medical corpora we used, only these two genders are mentioned.

	E3C	CAS	DIAMED	Total
Clinical cases (#)	1,069	646	333	2,048
Mean tokens	354.6	393.4	363.8	368.3
Mean constraints	25.2	27.2	24.3	25.7
% of feminine	51	42	51	48
% of masculine	48	58	47	51
% of undefined	1	0	2	1

Table 1: Statistics of the training corpus.

Access, primarily in English (98.75% of the corpus). It is the only model in our study that is adapted to the medical domain.

- BLOOM is an open-source, multilingual LM developed by the BigScience collective. It was trained on 46 languages, including French.
- Vigogne-2-Instruct is a "LLaMA model fine-tuned to follow French instructions".
- Llama-3.1 models are multilingual LMs trained on 8 languages, including French. The *-Instruct* version of the model was optimized to follow instructions in a dialogue setting.

We aimed to ensure all models had baseline training on clinical cases in French by fine-tuning them on our corpus. To our knowledge, it is not possible to have precise control over the fine-tuning conditions of API-based models for a comparable study. Therefore, we did not include any API-based models in this study.

3.2 Fine-tuning language models

Models are fine-tuned to generate clinical cases based on constraints that define clinical profiles. We build a corpus for instruction tuning of clinical cases in French, based on patients' demographic information and clinical elements. The corpus includes clinical cases from 3 different sources: CAS (Grabar et al., 2018) and DIAMED (Labrak et al., 2024a) – 2 French corpora of de-identified clinical cases – as well as the French documents of E3C (Magnini et al., 2020) – a multilingual corpus of de-identified clinical cases. Out-of-scope documents, i.e., nonclinical cases or forensic cases, are filtered out after manual inspection of the whole corpus.

We use the same methodology as Boulanger et al. (2024) to build an instruction corpus with a "clinical profile" comprising patients' demographic information and automatically extract clinical elements used as input (prompt) and the corresponding clinical case as expected output. More precisely, we manually annotate the gender and age of each patient when not available, and we automatically extract clinical elements using a BERT model fine-tuned on clinical annotations. From these annotations, we automatically select those that align with clinicians' definition of "salient" elements, primarily procedures and symptoms. An example of input is provided in Table 2.

Table 1 presents statistics about the training corpus. It consists of 2,048 clinical cases, each associated with a set of constraints. Even though the prevalence of male patients is higher (51%), there is no significant difference with a balanced gender distribution (exact binomial test, p-value = 0.2304).

We fine-tune each model on this instruction corpus using LoRA[3] trainable matrices on the *keys*, *queries*, and *values* of the transformer layers while the rest of the model is frozen (Hu et al., 2022).

3.3 Disorder selection and synthetic corpus generation

The fine-tuned LMs are then used for generating clinical cases in French for 10 different disorders: bladder cancer, breast cancer, colon cancer, COVID-19, depression, heart attack, osteoporosis, ovarian cancer, prostate cancer, and sickle-cell anaemia. These disorders were chosen because they evoke either societal biases and/or a medical gender imbalance. Specifically, ovarian cancer, breast cancer, depression, and osteoporosis are more feminine disorders whereas prostate cancer and bladder cancer are overwhelmingly present in male patients, and COVID-19, and heart attacks, slightly so. Colon cancer and sickle-cell anaemia seem neutral, as it affects an equivalent number of male and female patients (see Table 3). Further, some of these disorders are known to lead to the under-diagnosis of one gender over the other (see Section 6).

We find clinical cases associated with these disorders either in the 3 corpora described in Section 3.2 or by searching clinical cases in PubMed for the disorders that we did not find in the 3 corpora. Clinical cases in the test set were manually annotated with the salient elements as described in Boulanger et al. (2024) to ensure the relevance of the chosen constraints.

[3]For the fine-tuning with LoRa, we used the following parameters: r=8, alpha=16, and dropout=0.1.

	Prompt (input)	Generated clinical case (output)
FR	Écris un cas clinique pour le profil patient suivant : Age : 45 ; Il doit contenir les éléments suivants : douleurs périombilicale intenses ; vomissements ; empâtement sus-ombilical ; anémie à 9g/dl d'hémoglobine ; hypo albuminémie à 28g/l [...]	Un **homme âgé** de 45 ans, s'est **présenté** pour douleurs périombilicale intenses avec des vomissements spontanés, évoluant depuis 1 an. L'examen clinique a trouvé **un patient** maigre, alerte et stable sur le plan hémodynamique, avec un empâtement sus-ombilical et une perte d'appétit. Les examens de laboratoire ont montré une anémie à 9g/dl d'hémoglobine et une hypo albuminémie à 28g/l. L'immunoglobuline anti-transglutaminase a été positive. [...]
EN	Write a clinical case for the following patient profile: Age : 45 ; The following elements must be included: intense periumbilical pain; vomiting; supraumbilical swelling; anemia at 9g/dl hemoglobin; hypoalbuminemia at 28g/l [...]	A **45-year-old man** presented with intense periumbilical pain with spontaneous vomiting, which **he** has been experiencing for the past year. Clinical examination revealed a lean, alert, and hemodynamically stable patient with supraumbilical swelling and loss of appetite. Lab tests showed anemia at 9g/dl hemoglobin and hypo albuminemia at 28g/l. Anti-transglutaminase immunoglobulin was positive. [...]

Table 2: Truncated neutral prompt and matching generated clinical case (the EN text is a translation of the original FR text using DeepL and manual post-edition) from Llama-3.1-8B-Instruct, for colon cancer. Note that the prompt does not contain patient gender although gender information is present in the generated text (corresponding words are bolded – describing a male patient in this text). Clinical constraints supplied in the prompt are underlined.

During generation, the same decoding parameters are used for every model: a standard value of 0.9 for top-p and a temperature of 1 (in order to use the natural token distribution of the models).

We generate samples for each of the test set triplets (model, disorder, gender) and filter out generated texts containing multiple clinical cases or repetitions of 4-grams more than 15 times (the number was set heuristically to identify loops) until 100 valid texts are obtained for each triplet.

The resulting corpus contains 21,000 synthetic clinical cases (10 disorders $\times$ 7 LMs $\times$ 3 gender views $\times$ 100 generations), including $1/3$ obtained with a masculine prompt, $1/3$ obtained with a feminine prompt, and $1/3$ obtained with a gender neutral prompt. A sample neutral prompt and generated clinical case are provided in Table 2.

3.4 Automatic gender bias detection

To automatically detect gender biases, we adapt a system developed by Ducel et al. (2024). This hybrid system relies on linguistic rules, French lexical resources, and machine-learning techniques[4] to classify a text as "masculine", "feminine", "neutral", or "ambiguous" according to the morpho-syntactic gender that is the most present in the text, i.e., in the present study, the gender of the described patient. With the use of linguistic clues, we detect gender, as opposed to biological sex or assigned-at-birth sex. Biological features are not taken into account (e.g., genitals are not attributed to one gender over the other), which allows for greater accuracy and inclusiveness as biology does not define gender. As this approach is based on gender inflections, the presented work could be easily adapted to other inflected languages (e.g. Spanish, Italian, German, Hindi...).

For characterizing gender biases, we rely on two measures defined in (Ducel et al., 2024): Gender Gap and Gender Shift. Gender Gap is the difference between proportions of documents annotated as masculine (p^m), and as feminine (p^f): $GenderGap = p^m - p^f$. Gender Shift (GS) is used to analyze gendered prompts. When a generated text is labeled as ambiguous or as the gender opposite to the prompted one, the generated text has a GS of 1. If the generated text is labeled as neutral, or as the same gender as the prompted one, the GS is of 0. Then, the number of positive GS is divided by the total number of generated texts.

To evaluate the performance of gender detection, one author manually annotated a sample of 350 generated clinical cases. To ensure corpus representativity, 5 texts per disorder and per model

[4]The system uses SpaCy (Honnibal and Johnson, 2015) with its French transformer pipeline, which is based on CamemBERT (Martin et al., 2020), French Sequoia (Candito and Seddah, 2012) and Universal Dependencies (Candito et al., 2014).

were randomly selected (5 texts × 10 disorders × 7 LMs). The overall accuracy was 98% (7 errors/350 texts[5]). This performance is higher than the 92.8% accuracy reported by Ducel et al. (2024). The text genre and the use of the third person singular result in more gender markers (e.g., in French the noun "patient" is gendered so each occurrence carries gender information). In the absence of a formal evaluation of clinical case quality, the manual examination of the 350 cases suggests that the texts are grammatically reasonable and consistent with the clinical profile of the prompt. Besides, the manual evaluation of clinical cases presents some specific difficulties as the texts are long and technical. Moreover, exposure to content about death and disorder can be psychologically and emotionally challenging.

4 Analysis of Gender Biases in Synthetic Clinical Cases

4.1 LMs default to male patients

As mentioned in Section 2.2, bias can be defined as an imbalance between genders. According to one such definition, an ideal, unbiased, LM should follow gender information provided in the prompt and default to random when no gender information is provided. Thus, we analyze bias in LMs by comparing gender distributions in generated text with prompted gender information (see Figure 1).

Feminine and masculine prompts lead to equivalent proportions of expected genders (resp. 87.9 and 87.5%). This implies that 12% of texts generated from a gendered prompt are not consistent with the prompted gender. These tendencies to override the gender supplied in the prompt differ depending on the disorder and the LM (see Sections 5.1 and 5.2).

However, gender distributions are noticeably more uneven with neutral prompts: 1.9 times more clinical cases about male vs. female patients, with a Gender Gap of 30.9 (vs. expected 0). In other words, LMs seem to default to male patients when not explicitly prompted with gender. This behavior is not consistent with the gender distribution observed in the clinical case training corpus, suggesting that LMs are biased towards generating descriptions of male rather than female patients. These observations can be connected to the concept of "masculine defaults" (Cheryan and Markus, 2020), as masculine is "regarded as standard, normal, neutral", and feminine is viewed as specific, mainly associated with disorders that are related to majoritarily feminine organs (see Section 5.1).

[5]The classification report is provided in Appendix A.

Disorder	Men (%)	Women (%)
Prostate cancer	100	0
Bladder cancer	80.2	19.8
COVID-19	62.2	37.8
Heart attack	60	40
Colon cancer	50.2	49.5
Sickle-cell anaemia	46.9	53.1
Depression	33	66
Osteoporosis	30	70
Breast cancer	0.7	99.3
Ovarian cancer	0	100

Table 3: Estimated real-world gender prevalence from a French hospital information system or from the literature when too few patients were represented.

4.2 LMs amplify real-world prevalence

We established that LMs do not generate even numbers of feminine vs. masculine texts overall. Another approach and definition of bias is that the ideal LM should be representative of real-world gender prevalence in disorders. In this section, we use real-world statistics on gender distributions for the selected disorders (see Table 3) and compare them with the distributions in generated texts.

Figure 2 shows that, except for prostate cancer, synthetic texts have a higher proportion of male patients than real-world distributions. The increase of masculine sometimes leads to the inversion of the majority gender. According to real-world data, osteoporosis and depression affect more women than men but, in our study, lead to more masculine than feminine generations when the prompts are neutral. Bladder, ovarian, breast, and prostate cancers have relatively close generated and real gender prevalence.

In conclusion, LMs are also biased as they do not represent real-world statistics but amplify statistical differences by increasing the proportion of male patients. However, representing real-world statistics is not ideal as these statistics also reflect some stereotypes and human biases (see Section 6).

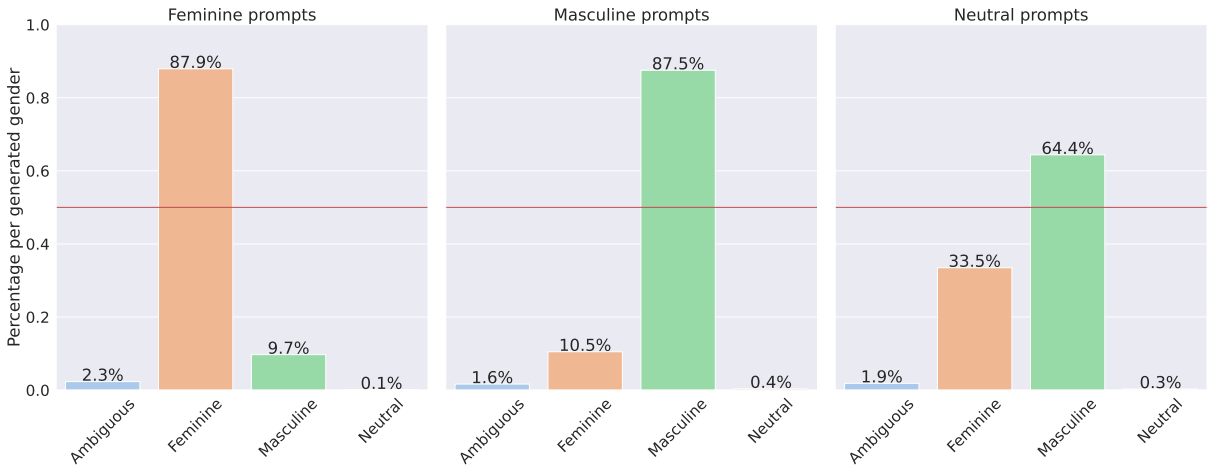

Figure 1: Distributions of generated gender w.r.t. the gender given in the prompt. The red line indicates 50%.

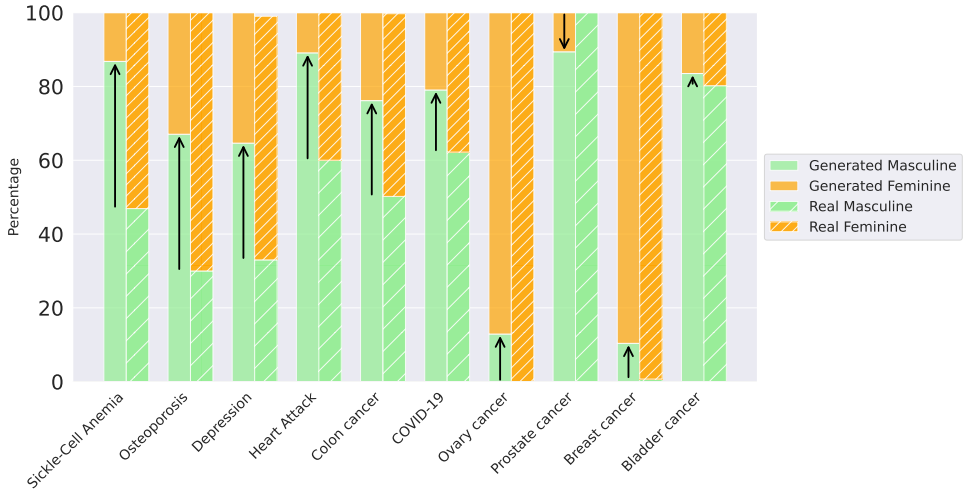

Figure 2: Gender distribution in generated cases (using neutral prompts) vs. real corpora. The arrows represent the distance between generated vs. real masculine distributions (i.e., an arrow going up means that the generated distribution of masculine is higher than real-world masculine distribution).

5 Further Analyses of Gender Biases

5.1 Disorders impact gender biases

Gender Gaps differ a lot depending on the prompted disorder (see Figure 3). With gendered prompts, all disorders but ovarian, breast, and prostate cancers exhibit rather low bias (at most 1.3x more masculine than feminine texts). Depression slightly leans towards feminine while, in ascending order of bias, osteoporosis, colon, bladder, heart attack, COVID-19, and sickle-cell anaemia are biased towards masculine. Amplified but similar trends can be observed with neutral prompts, with two exceptions: depression is biased towards masculine instead of feminine (1.8x more masculine vs. feminine) and, more surprisingly, heart attack is the most biased disorder, even slightly more than prostate cancer (8.1x more male vs. female patients). With neutral prompts, sickle-cell anaemia, bladder cancer, COVID-19, and colon cancer are also highly biased towards masculine: resp. 6.5x, 5x, 3.7x, and 3.2x more male patients.

Prostate cancer is highly biased in favor of masculine (8x more male vs. female patients with neutral prompts), even with gendered prompts (then, 1.5x more male patients). Similarly, ovarian and breast cancers are highly biased in favor of feminine, leading to respectively 6.75x and 8.15x more female than male patients with neutral prompts, 1.7x and 2x more female than male patients with gendered prompts. However, these 3 disorders are tied to an organ (prostate, breast, or ovary) that is mostly present in one gender over the other (see Section 4.2). The model is expected to default to one gender for these disorders when the prompt

Disorder	GS
Osteoporosis	0.065
Depression	0.073
Heart attack	0.077
Colon cancer	0.086
COVID-19	0.100
Bladder cancer	0.109
Sickle-cell anaemia	0.118
Prostate cancer	0.167
Ovarian cancer	0.170
Breast cancer	0.238

Table 4: Gender Shift (GS) per disorder (sorted).

LM	GS
vigogne-2-13b	0.093
Llama-3.1-8B	0.104
Llama-3.1-8B-Instruct	0.105
BioMistral-7b-SLERP	0.119
BLOOM-1b1	0.126
vigogne-2-7b	0.145
BLOOM-7b1	0.152

Table 5: Gender Shift (GS) per LM (sorted).

is neutral, and we argue that the implications and harms are different from other disorders.

Nonetheless, harms are still at stake when the gender of the prompt gets overridden because there is an underlying implication in the world representation put forth by the LM that the disorder cannot affect the specified gender. This is problematic and harmful as it can lead to underdiagnosis and mistreatment of patients of the overridden gender. The transgender community can also be especially impacted, as overriding the prompt results in misgendering the patient, e.g., always overriding feminine in cases of prostate cancer infers that a prostate cancer patient has to be a man. Table 4 indicates that in over 23.8% of generations about breast cancer, the gender of the prompt (when explicit) gets overridden (in the vast majority of cases, masculine prompts get overridden with feminine markers). This is the case for 17% of texts about ovarian cancer, and 16.7% of texts on prostate cancer.

5.2 LMs unevenly exhibit biases

Gender Gaps differ a lot based on whether or not the prompt contains a gender. With gendered prompts, most LMs exhibit slight biases toward feminine, whereas they all exhibit (higher) bias toward masculine with neutral prompts (see Figure 3). The bias towards feminine seems related to strong associations between ovarian/breast cancers and feminine (BLOOM-7b and vigogne-7b have the lowest Gaps with gendered prompts and the highest proportions of feminine texts for these disorders). These associations are so strong that they often lead to overriding masculine prompts (see Table 4).

Results on gendered prompts in Table 5 indicate a rather homogeneous behavior of LMs. We still notice that the BLOOM models exhibit higher Gender Shifts than the Llama models and that the size of vigogne leads to different levels of bias. Finally, the very small difference between the base Llama3.1 model and its *Instruct* version suggests that the combination of instruction and alignment does not have a significant influence here, which is a little bit surprising. More detailed figures are provided in Appendix B.

5.3 Quality of generated texts

Clinical case quality is assessed with two proxies: redundancy and constraints respect rate (CRR). The redundancy of a text is calculated as $1 - \frac{unique\_4-grams\#}{total\_4-grams\#}$ and allows us to estimate the potential presence of loops and duplicated texts in generations. More specifically, it was used to test if texts were more repetitive when they had a majority of feminine/masculine markers. Respected constraints are identified with exact and fuzzy string matching. We compute the proportion of respected constraints for each generated text, and the CRR corresponds to the mean of these proportions. The CRR is used to assess whether the generated texts follow the instructions equally depending on the generated patient's gender.

Table 6 indicates that feminine texts are as redundant as masculine texts while masculine texts seem more consistent with the constraints compared to feminine texts. However, we could not find any significant correlations between gender in generated texts and either redundancy or CRR. We can hypothesize that disorders impact the aforementioned CRR discrepancy as breast and ovarian cancers exhibit the lowest CRR (resp. 43.3% and 55.4%) as well as the highest association with feminine texts. In these cases, it is difficult to know whether the lower CRR for these two stereotypically feminine disorders is related to stereotype, or if it is due to other external reasons.

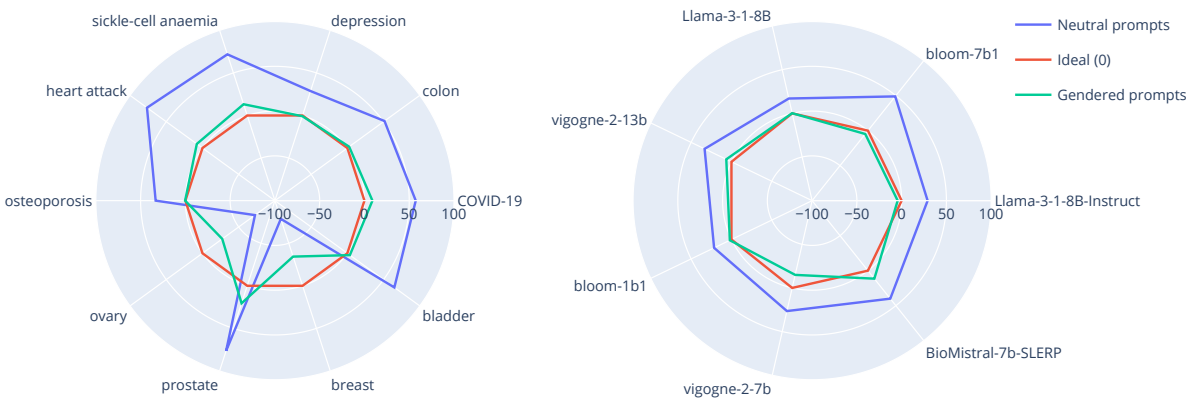

Figure 3: Gender Gaps in generated cases per disorder (left) and LM (right) using prompts with or without gender.

Output gender	Redundancy	CRR
Ambiguous	$0.080_{\pm 0.10}$	$0.650_{\pm 0.20}$
Feminine	$0.098_{\pm 0.12}$	$0.635_{\pm 0.18}$
Masculine	$0.098_{\pm 0.12}$	$0.681_{\pm 0.17}$
Neutral	$0.092_{\pm 0.10}$	$0.681_{\pm 0.16}$

Table 6: Redundancy and constraints respect rates (CRR) per generated gender. Standard deviations are indicated in subscripts.

Other, more specific proxies could be used to further analyze biases. With a naive approach based on keyword matching, we detected for example that 36.7% of texts with a majority of masculine markers result in the death of the patient whereas this number reaches 44.2% when the text is mostly feminine. However, the correlations do not seem significant.

5.4 Environmental impact of the experiments

Our experiments were conducted using an NVIDIA A100 SXM4 80GB GPU for fine-tuning all the models, except for BLOOM-1b1 for which we used an NVIDIA Tesla V100 SXM2 32GB GPU. NVIDIA A100 SXM4 80GB GPUs were used for inference, with VLLM (Kwon et al., 2023) to accelerate the computation. Finally, the gender detection was done on an NVIDIA GeForce GTX 1080 Ti. Using MLCA (Morand et al., 2024), we estimate that the gender detection system emitted 0.23kgCO_2e, fine-tuning the models emitted 0.49kgCO_2e, and the generation process emitted 1.15kgCO_2e, for a total of 1.87kgCO_2e.

6 Discussion: What Is the Ideal LM Like?

We evidenced two types of biases in LMs: they neither produce balanced gender distributions nor replicate real-world gender prevalence. However, we discuss here what an ideal LM would be like, and the flaws of both "ideal outputs".

First, an ideal LM that is neutral or gender-balanced could be considered inefficient or irrelevant as some disorders exhibit strong gender prevalence because of biological attributes.

Then, we argue that an LM that replicates real-world gender prevalence is not a viable option either because real-world statistics and estimated gender prevalence are biased as well. Healthcare professionals have their own stereotypes and biases, especially harming women and other disadvantaged groups (Dwass, 2019). Some disorders are known to be stereotypically associated with a gender, to the point that it leads to misdiagnosis, mistreatment, and is a taboo for patients of the opposite gender. It is the case for osteoporosis and depression in men (Rinonapoli et al., 2021; Van de Velde et al., 2010), and heart attack in women (Banks, 2008). Thus, replicating real-world statistics could reinforce these issues for patients of the opposite gender by conveying the idea that it is not possible that they suffer from this disorder. These problems result in possibly incorrect real-world statistics that underestimate the proportion of the minority, misdi-

agnosed gender. Further, it is often unclear if these statistics take into account gender or (assigned-at-birth) sex. It is problematic for transgender people, as we do not know how/if they are taken into account. As a consequence, prostate cancer is often overlooked in transgender women (Deebel et al., 2017). Finally, even when the gender prevalence is faithful to reality, socio-economic factors play a role, so the prevalence is not entirely biological: social and economic insecurity, as well as exposure to psychological and physical violence, increase the risk of depression, and women are much more affected by these situations than men (Van de Velde et al., 2010; Gresy et al., 2020). Discrimination also impacts the health of socially disadvantaged populations: harassment and violence have repercussions on people's health (Jahnke et al., 2019), and the wage gap between men and women plays a role in the gendered disparity of mood disorders (Platt et al., 2016). Moreover, men can also be impacted by biases, e.g. men suffering from breast cancer may face stigmatization, as well as a later diagnosis than women (Robinson et al., 2008; Midding et al., 2018).

Based on the results of the study and these facts, we recommend that demographic bias must be taken into account when constructing a synthetic corpus of clinical cases. Including demographic information in generation constraints appears to be the best strategy, as it allows for control over the demographic distribution of patients, and leads to less biased distributions of digital patients (see the differences between gendered and neutral prompts in Figure 3).

However, the reported results and the harms they could lead to provide clear and reliable evidence of one noteworthy potential limitation of using LMs in clinical contexts and should be taken into account by all stakeholders. This study reinforces the argument that the integration of LMs in medical practice requires caution and must be closely supervised by medical experts. Even in secondary applications of health data, such as pedagogical purposes for which clinical cases are commonly used, rigorous oversight is necessary. Exposure to biased clinical cases could create or reinforce (implicit) gender stereotypical biases in future healthcare professionals' practice and harm their patients.

7 Conclusion

In this study, we fine-tuned LMs to generate clinical cases in French, on 10 disorders. We found out that LMs generate more mentions of male than female patients, increasing the real-world proportion of men affected by a given disorder. However, gender distributions vary a lot depending on the prompted disorder. All disorders but ovarian and breast cancers are tilted towards masculine generations, especially prostate cancer, heart attack, and sickle-cell anemia. In other words, unless disorders are massively feminine and related to mostly feminine organs, LMs mostly generate male patients. More surprisingly, heart attack is the disorder that is the most associated with male patients when the prompts are neutral. All LMs exhibit biases, but in different proportions: with neutral prompts, BLOOM-7b shows the highest Gender Gap, whereas, with gendered prompts, vigogne-2-13b is the most biased.

We can conclude that by under-representing female patients, LMs can harm women. It is a form of representational harm. Moreover, by strongly associating gender with some disorders, LMs can participate in misdiagnosing and mistreating patients who do not match the stereotypical gender of a given disorder. LMs can also lead to misgendering transgender people and reinforce biological essentialism by shifting the gender of the prompt, especially for disorders associated with sex-specific organs. These can cause allocational harms.

Perspectives. We believe that stereotypical biases should be further studied, especially in the context of healthcare. Possible continuations of this work include looking into the possible gaps between medical consistency among genders and extending the scope of studied bias types. We intend to fine-tune LMs on a NER task to detect other demographic information (nationality, skin color, socioeconomic status...) and study intersectionality, as well as the impact of explicitly mentioning other demographic information in the prompts. We also plan to study the impact of fine-tuning, and its possible bias amplification.

Limitations

Our bias evaluations do not take into account the medical consistency or plausibility of the clinical cases, due to the lack of medical experts among authors and the amount of texts (it is not possible

to review them all manually). However, it would be relevant to find a method to assess the semantic and medical information, as we were already able to spot some inconsistencies (e.g., a "virgin" woman who just gave birth, or a person consulting for a "burn" that occurred while "fishing parrots"). Note that the manual annotation mentioned in Section 3.4 also served as a sanity check, and apart from these few examples of obvious logical failures, the rest of the annotated corpus seemed sound from a linguistic and common knowledge perspective (all medical knowledge aside). Further, the main goal of this study is to evaluate associations between genders and disorders, and these associations are still present, even when the text is not semantically/medically flawless.

Besides, as the gender detection system is imperfect, it sometimes labels texts as neutral instead of masculine, which leads to an underestimation of masculine generations.

Finally, this study was only conducted on French, and even if the methodology could easily be applicable to other inflected languages, further experiments are required in order to see if the results are similar in other linguistic and cultural contexts.

Acknowledgments

This work has received funding from the French "Agence Nationale de la Recherche" through grants InExtenso - ANR-23-IAS1-0004 and CODEINE - ANR-20-CE23-0026-01. The authors also thank Bastien Rance at HEGP for his clinical insights. Fine-tuning of models and inference were performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011014538R1), and gender detection was carried out using the Grid'5000 testbed (Inria, CNRS, RENATER, ...).

References

Hammaad Adam, Aparna Balagopalan, Emily Alsentzer, Fotini Christia, and Marzyeh Ghassemi. 2022. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine*, 2(1):149.

AI@Meta. 2024. Llama 3 model card.

Grayce Alencar Albuquerque, Glauberto da Silva Quirino, Francisco Winter dos Santos Figueiredo, Laércio da Silva Paiva, Luiz Carlos de Abreu, Vitor Engrácia Valenti, Vânia Barbosa do Nascimento, Erika da Silva Maciel, Fernando Rodrigues Peixoto Quaresma, Fernando Adami, et al. 2016. Sexual diversity and homophobia in health care services: perceptions of homosexual and bisexual population in the cross-cultural theory. *Open Journal of Nursing*, 6(06):470.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Nicholas C Arpey, Anne H Gaglioti, and Marcy E Rosenbaum. 2017. How socioeconomic status affects patient perceptions of health care: a qualitative study. *Journal of primary care & community health*, 8(3):169–175.

Angela D. Banks. 2008. Women and heart disease: Missed opportunities. *Journal of Midwifery & Women's Health*, 53(5):430–439.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *SIGCIS conference paper*, Philadelphia, Pennsylvania, USA.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Hugo Boulanger, Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2024. Using structured health information for controlled generation of clinical cases in French. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 172–184, Mexico City, Mexico. Association for Computational Linguistics.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric de la Clergerie. 2014. Deep syntax annotation of the sequoia French treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2298–2305, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a

parser lexical domain adaptation method) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, 127(6):1022.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. 2023. Can large language models safely address patient questions following cataract surgery? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 131–137, Toronto, Canada. Association for Computational Linguistics.

Joan C Chrisler, Angela Barney, and Brigida Palatino. 2016. Ageism can be hazardous to women's health: Ageism, sexism, and stereotypes of older women in the healthcare system. *Journal of Social Issues*, 72(1):86–104.

Nicholas A. Deebel, Jacqueline P. Morin, Riccardo Autorino, Randy Vince, Baruch Grob, and Lance J. Hampton. 2017. Prostate cancer in transgender women: Incidence, etiopathogenesis, and management challenges. *Urology*, 110:166–171.

Kerry Drabish and Laurie A Theeke. 2022. Health impact of stigma, discrimination, prejudice, and bias experienced by transgender people: a systematic review of quantitative studies. *Issues in mental health nursing*, 43(2):111–118.

Fanny Ducel, Aurélie Névéol, and Karën Fort. 2024. "you'll be a nurse, my son!" automatically assessing gender biases in autoregressive language models in french and italian. *Language Resources and Evaluation*, pages 1–29.

Emily Dwass. 2019. *Diagnosis Female: How Medical Bias Endangers Women's Health*. Rowman & Littlefield.

Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18:1–18.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. Cas: French corpus with clinical cases. In *LOUHI 2018-The Ninth International Workshop on Health Text Mining and Information Analysis*, pages 1–7.

Brigitte Gresy, Emmanuelle Piet, Catherine Vidal, and Muriel Salle. 2020. Prendre en compte le sexe et le genre pour mieux soigner. *Haut Conseil à l'Egalité entre les femmes et les hommes.[Internet]*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, pages 1–10.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology*, 47(1):237–271. PMID: 15012482.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

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

Bofeng Huang. 2023. Vigogne: French instruction-following and chat models. https://github.com/bofenghuang/vigogne.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3.

Sara A Jahnke, Christopher K Haddock, Nattinee Jitnarin, Christopher M Kaipust, Brittany S Hollerbach, and Walker SC Poston. 2019. The prevalence and health impacts of frequent work discrimination and harassment among women firefighters in the us fire service. *BioMed research international*, 2019(1):6740207.

Michelle Kim, Junghwan Kim, and Kristen Johnson. 2023. Race, gender, and age biases in biomedical masked language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11806–11815, Toronto, Canada. Association for Computational Linguistics.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624, Virtual-only conference. Curran Associates, Inc.

Abhishek Kumar, Sarfaroz Yunusov, and Ali Emami. 2024. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 375–392, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickael Rouvier, Pacome Constant Dit Beaufils, Natalia Grabar, Béatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-Antoine Gourraud, and Richard Dufour. 2024a. DrBenchmark: A large language understanding evaluation benchmark for French biomedical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5376–5390, Torino, Italia. ELRA and ICCL.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024b. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Blake J Lawrence, Deborah Kerr, Christina M Pollard, Mary Theophilus, Elise Alexander, Darren Haywood, and Moira O'Connor. 2021. Weight bias among health care professionals: a systematic review and meta-analysis. *Obesity*, 29(11):1802–1812.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Rumeng Li, Xun Wang, and Hong Yu. 2024. LlamaCare: An instruction fine-tuned large language model for clinical NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10632–10641, Torino, Italia. ELRA and ICCL.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Simon Meoni, Éric De la Clergerie, and Théo Ryffel. 2024. Generating synthetic documents with clinical keywords: A privacy-sensitive methodology. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 115–123, Torino, Italia. ELRA and ICCL.

Evamarie Midding, Sarah Maria Halbach, Christoph Kowalski, Rainer Weber, Rachel Würstlein, and Nicole Ernstmann. 2018. Men with a "woman's disease": Stigmatization of male breast cancer patients—a mixed methods analysis. *American Journal of Men's Health*, 12(6):2194–2207. PMID: 30222029.

Clément Morand, Aurélie Névéol, and Anne-Laure Ligozat. 2024. MLCA: a tool for Machine Learning Life Cycle Assessment. In *2024 International Conference on ICT for Sustainability (ICT4S)*, Stockholm, Sweden.

Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217, Toronto, Canada. Association for Computational Linguistics.

Jonathan Platt, Seth Prins, Lisa Bates, and Katherine Keyes. 2016. Unequal depression for equal work? how the wage gap explains gendered disparities in mood disorders. *Social Science & Medicine*, 149:1–8.

Giuseppe Rinonapoli, Carmelinda Ruggiero, Luigi Meccariello, Michele Bisaccia, Paolo Ceccarini, and Auro Caraffa. 2021. Osteoporosis in men: A review of an underestimated bone condition. *International Journal of Molecular Sciences*, 22(4).

John Robinson, Kenneth Metoyer, and Neil Bhayani. 2008. Breast cancer in men: A need for psychological intervention. *Journal of clinical psychology in medical settings*, 15:134–9.

Sarah Van de Velde, Piet Bracke, and Katia Levecque. 2010. Gender differences in depression in 23 european countries. cross-national variation in the gender gap in depression. *Social science & medicine*, 71(2):305–313.

Lucía Vicente and Helena Matute. 2023. Humans inherit artificial intelligence biases. *Scientific Reports*, 13(1):15737.

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

David R Williams and Ronald Wyatt. 2015. Racial bias in health care and health: challenges and opportunities. *Jama*, 314(6):555–556.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024. Climb: A benchmark of clinical bias in large language models. *Preprint*, arXiv:2407.05250.

A Performance of the Gender Detection System

	Prec.	Recall	F1-score	N
Ambiguous	0.3333	1.0000	0.5000	1
Feminine	1.0000	0.9684	0.9839	158
Masculine	0.9793	0.9895	0.9844	191
Neutral	0.0000	0.0000	0.0000	0
Accuracy			0.9800	350
Macro avg	0.5782	0.7395	0.6171	350
Weighted avg	0.9868	0.9800	0.9828	350

Table 7: Classification report of the gender detection system.

B Bias per Disorder and per LM

We provide on Figure 4 detailed figures that illustrate the proportions of generated gender per disorder and per LM, when the prompts are neutral.

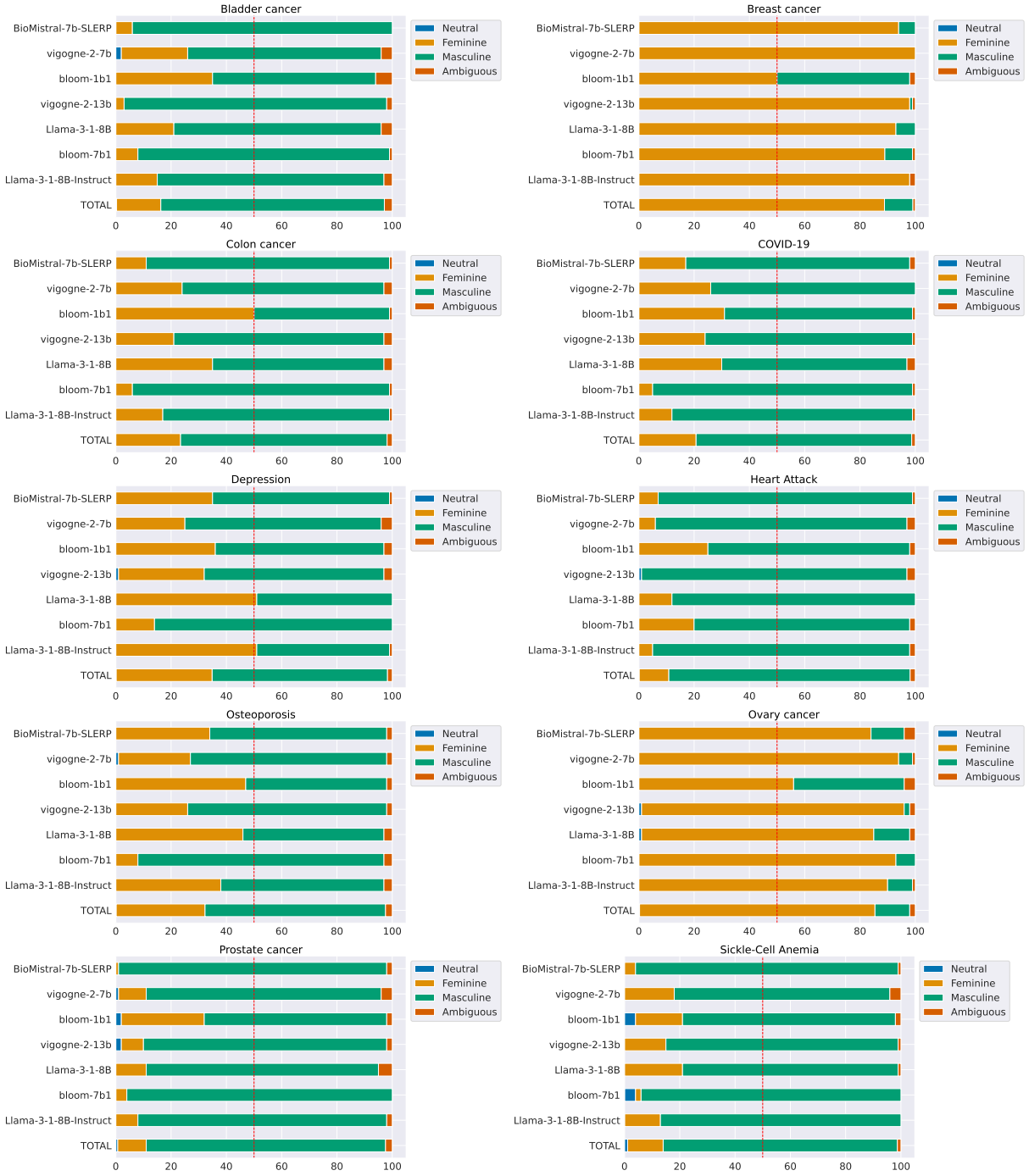

Figure 4: Proportions of generated gender per disorder and per LM with neutral prompts. The red dotted line indicates 50% (even distributions).