NIHR | Policy Research Unit Adult social care

Evaluating gender bias in Large Language Models in long-term care

Sam Rickman

March 2025

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

University of Kent

Abstract

Background: Large language models (LLMs) are being used to reduce the administrative burden in long-term care by automatically generating and summarising case notes. However, LLMs can reproduce bias in their training data. This study evaluates gender bias in summaries of long-term care records generated with two state-of-the-art, open-source LLMs released in 2024: Meta's Llama 3 and Google Gemma.

Methods: Gender-swapped versions were created of long-term care records for 617 older people from a London local authority. Summaries of male and female versions were generated with Llama 3 and Gemma, as well as benchmark models from Meta and Google released in 2019: T5 and BART. Counterfactual bias was quantified through sentiment analysis alongside an evaluation of word frequency and thematic patterns.

Results: The benchmark models exhibited some variation in output on the basis of gender. Llama 3 showed no gender-based differences across any metrics. Gemma displayed the most significant gender-based differences. Male summaries focus more on physical and mental health issues. Language used for men was more direct, with women's needs downplayed more often than men's.

Conclusions: Care services are allocated on the basis of need. If women's health issues are underemphasised, this may lead to gender-based disparities in service receipt. LLMs may offer substantial benefits in easing administrative burden. However, the findings highlight the variation in state-of-the-art LLMs, and the need for evaluation of bias. The methods in this paper provide a practical framework for quantitative evaluation of gender bias in LLMs. The code is available on GitHub.

Introduction

In the US and UK, large language models (LLMs) are being used to generate care documentation by summarising audio transcripts of care interventions or distilling extensive free text case notes into short summaries [1, 2, 3]. The case for such tools is compelling. Documentation is the most time-consuming task in health and long-term care [4, 5, 6]. Additionally, electronic care records often span decades, making it impractical for practitioners to review all the information. In some cases, avoidable harm has occurred where workers were unaware of important details in their records [7]. By automatically generating or summarising records, LLMs have the potential to reduce costs without cutting services, improve access to relevant information, and free up time spent on documentation.

There is political will to expand such technologies in health and care. The 2023 US Executive Order issued by President Biden sought to promote the "deployment of… generative AI-enabled technologies in healthcare", and established a Health and Human Services (HHS) Artificial Intelligence (AI) Task Force [8]. The Spring 2024 UK budget stated that LLMs will be used to increase the time clinicians can spend with patients and unlock an annual productivity benefit of £500 million - £850 million ($643 million - $1.1 billion USD) [9]. The European Union (EU) Artificial Intelligence (AI) Act provides a framework for the introduction of such products, though it also mandates significant regulatory oversight [10, 11].

LLMs can produce accurate summaries of healthcare records and even outperform humans [12]. High quality, relevant documentation is associated with lower cognitive burden, reduction in errors, and improved quality of care [13, 14, 15]. However, while accuracy is a necessary condition for the use of such models, it is not sufficient. LLMs can reproduce bias that appears in the data on which they are trained [16, 17]. Furthermore, variation in tone and style of accurate content may affect the decision-making of care practitioners [18].

This paper measures the gender bias in Meta's Llama 3 [19] and Google Gemma [20], two state-of-the-art, open-source LLMs released in 2024. Summaries of care records from individual-level, long-term care case notes in a London local authority were generated using each model. Lightweight models created in 2019, Google's T5 [21] and Meta's BART [22], were used as benchmarks. It has been established that these lightweight models exhibit gender bias, and that larger, more complex models may magnify bias found in training data [23, 24]. The aim is to determine

whether the gender bias in the state-of-the-art models differs from that observed in the earlier models when summarising long-term care notes.

Three questions are addressed in this study. Firstly, whether there are measurable, gender-based differences in summaries of long-term care case notes generated by state-of-the-art, open-source LLMs. Secondly, if so, whether there is measurable inclusion bias [25], where different topics are included in summaries for men and women, or linguistic bias [17], where the same topics are discussed using different language. Finally, the implications for care practice of gender-based differences are considered.

Materials and methods

Data

Pseudonymised records were extracted from a local authority adult social care case recording system in England, recorded between 2010 and 2020. Ethical approval was obtained for the use of the data. Texts about men and women were selected, and gender-swapped versions were created using Llama 3 as outlined in *Analysis and data pre-processing*. Summaries of each pair of texts were then generated, and the male and female versions of the output were compared in three ways. Firstly, sentiment analysis was applied to determine whether any model generates consistently more negative sentiment. Secondly, the inclusion bias [25] of certain topics was measured by comparing the frequency of terms related to domains such as health and physical appearance in summaries for each gender. Finally, linguistic bias [17] was assessed by comparing the frequencies of words appearing in the output generated by each model.

Table 1: Examples of paired sentences used as input to summarisation models

Original	Gender swapped
Mrs Smith is an 87 year old, white British woman with reduced mobility. She cannot mobilise independently at home in her one-bedroom flat.	Mr Smith is an 87 year old, white British man with reduced mobility. He cannot mobilise independently at home in his one-bedroom flat.
Mrs Jones is an older lady who has been diagnosed with dementia of Alzheimer's disease and has poor short term memory.	Mr Jones is an older gentleman who has been diagnosed with dementia of Alzheimer's disease and has poor short term memory.

Conceptual framework: counterfactual fairness

To assess bias, this paper uses the framework of counterfactual fairness defined in Kusner et al. [26], that a machine learning model is fair towards an individual if its output is the same in the actual world and a counterfactual world where the individual's circumstances are identical, except for a demographic change such as gender, race or sexual orientation.

More formally, a predictor $\hat{Y}$ is *counterfactually fair* if, for any individual with observed attributes $A = a$ (protected attribute) and $X = x$ (remaining attributes), and for any other possible value a' of A, Equation (0.1) holds.

$$P\left(\hat{Y}_{A\leftarrow a} = y \mid A = a, X = x\right) = P\left(\hat{Y}_{A\leftarrow a'} = y \mid A = a, X = x\right), \quad \text{for all } y. \tag{0.1}$$

Where:

- $P(\hat{Y}_{A\leftarrow a} = y \mid A = a, X = x)$ is the probability that the prediction $\hat{Y} = y$, given that the individual actually has attribute $A = a$ and characteristics $X = x$.
- $P(\hat{Y}_{A\leftarrow a'} = y \mid A = a, X = x)$ is the probability that the prediction $\hat{Y} = y$, if, counterfactually, the protected attribute A were set to a', while keeping all else the same.

This definition was originally designed for outputs ($\hat{Y}$) that are straightforward to compare, such as insurance premiums or predicted risk of offending. The output of LLMs are sequences of high-dimensional vectors of varying length. Direct comparisons between them in vector space may be challenging to implement or interpret. Instead, the approach taken here is to analyse differences in textual content of the model output, as outlined below.

Comparison of sentiment output

Three widely used, pre-trained sentiment analysis metrics were selected. Firstly, SiEBERT, a general-purpose sentiment analysis model [27] based on the RoBERTa language model [28], fine-tuned on 15 datasets of reviews and social media text, was used. This binary model predicts whether sentences are positive or negative in sentiment. As there are degrees of positive and negative sentiment, a popular sentiment analysis model based on DistilBERT [29, 30], which produces continuous sentiment scores, was also utilised. Finally, a metric specifically focused on measuring prejudices against different demographics was sought. Regard [31], which was designed to evaluate gender bias, was employed. A mixed regression model was applied for each of the sentiment metrics, where the summarisation model was included as a random effect, clustered by document ID as a random intercept, as specified in Equation (0.2).

$$\begin{aligned} \text{sentiment}_{ij} = \; & \beta_0 \; + \; \beta_1^\top \, \mathbf{model}_j \; + \; \beta_2 \, \text{gender}_j \\ & + \; \beta_3^\top (\mathbf{model}_j \times \text{gender}_j) \; + \; \beta_4^\top \, \mathbf{max\_tokens}_j \\ & + \; u_{0i} \; + \; \mathbf{u}_{1i}^\top \mathbf{model}_j \; + \; \epsilon_{ij} \end{aligned} \tag{0.2}$$

The dataset consists of 29,616 rows, representing 617 documents, each with 48 possible combinations of gender (2 levels), maximum token length (6 levels), and summarisation model (4 levels).

Where:

- sentiment_{ij} is the outcome (a numeric score) for observation j in document i.
- $\mathbf{model}_j$ is a vector of dummy variables indicating which model (Gemma, Llama 3, T5) level applies to row j, with BART as the reference level.
- gender_j is binary variable with 0 indicating female and 1 male.

- $\mathbf{model}_j \times \text{gender}_j$ is the interaction effect between gender and LLM.
- $\mathbf{max\_tokens}_j$ is a vector of dummy variables for the max_tokens factor (75, 100, 150, 300 or None), with length 50 as the reference level.
- u_{0i} and u_{1i} together define random intercepts for document-level i sentiment for the four LLMs. u_{0i} is the random intercept for the reference-level LLM (BART), and u_{1i} represent differences between random intercepts for each of the other models and the random intercept for BART.
- ϵ_{ij} is the residual error term, which is assumed to be $\mathcal{N}(0, \sigma^2)$.

Data was also available for the age, gender and ethnicity of each individual. However, inclusion of these variables in the model led to very similar results, and a Likelihood Ratio test indicated that they did not improve the model. An alternative specification including an interaction between max_tokens and gender was tested, but a likelihood ratio test indicated that this interaction did not significantly improve the explanatory power of the model. For the sake of parsimony, these models are not included in the output in the Results section. For robustness, estimates were bootstrapped, and a variance-structured mixed effects model, a Generalised Estimating Equations (GEE) model, a robust linear mixed model, and a separate linear model for each language model were fitted. Details of this are included in the Appendix.

Inclusion bias: comparison of themes

A sample of original documents was examined to identify common themes across texts. Four themes were identified: physical health, mental health, physical appearance, and subjective language. To aid in the interpretation of differences in output, lists of words related to each theme were created. Llama 3 and Gemma were used to systematically scan the original texts for phrases associated with each theme. For instance, the models were prompted to identify all subjective language (such as "dirty," "excessive," and "rude") in the original texts. A comprehensive list of terms was generated, which was manually refined to remove irrelevant entries, resulting in focused lists of terms. This process was repeated for each theme. The lists are included in the Appendix.

The total frequency of each term in the summaries generated by each model for male and female subjects was counted. As the original texts used all terms an equal number of times for each gender, any differences in the summaries were attributable to the summarisation models. The total counts of these terms in the summaries were compared, and χ^2 tests were used to determine if the differences

were statistically significant. The p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method [32].

Linguistic bias: word frequency analysis

To analyse linguistic bias, frequencies of individual words were compared at two levels: overall counts and document-level. Firstly, word counts were aggregated across all documents for each LLM, and the frequency of each word between male and female summaries was compared. A χ^2 test was used to determine if differences in overall counts were statistically significant except if counts of fewer than 5 were observed for either gender, where Fisher's exact test was used instead. Again, p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method [32]. For document-level analysis, regression was performed on the word counts. For each word, a table of all documents in which it appeared was created, and a Poisson regression was run, where the dependent variable was the word count, and the independent variables were document ID, gender, and the maximum number of tokens, as specified in Equation (0.3).

$$\log(\mathbb{E}[\text{count}_{ij} \mid \mathbf{X}_{ij}]) = \beta_0 + \beta_1 \,\text{gender}_j + \beta_2^\top \,\mathbf{max\_tokens}_j + \beta_3^\top \,\mathbf{doc\_id}_j \tag{0.3}$$

Where:

- $\log(\mathbb{E}[\text{count}_{ij} \mid \mathbf{X}_{ij}])$ is the log of the expected value of the count of each specific word for row j in document i, given a vector of explanatory variables $\mathbf{X}_{ij}$.
- gender_j is binary variable with 0 indicating female and 1 male.
- $\mathbf{max\_tokens}_j$ is a vector of dummy variables for the max_tokens factor (75, 100, 150, 300 or None), with length 50 as the reference level.
- $\mathbf{doc\_id}_j$ is a vector of dummy variables identifying document i on row j. This allows the model to account for the fact that words will be expected to appear a different number of times in each document. The document-level coefficients are not of interest and are not included in the results.

Occasionally, perfect separation occurred (i.e., words that never appeared for one gender), so Firth's penalised likelihood method of Poisson regression [33] was used to obtain reliable parameter estimates. In cases of overdispersion ($\frac{\sum(r_i^2)}{\text{df}_{\text{residual}}} >$

1.25), a negative binomial regression with the same independent variables was also run. As multiple comparisons were conducted, words were considered to appear significantly differently only if they were statistically significant in both the regression output and the Benjamini-Hochberg adjusted χ^2 test (adjusted $p < 0.05$).

Analysis and data pre-processing

Creating equivalent male and female texts

The data included free text records for 3046 older adults receiving care in a London local authority. Free text responses to the care needs assessment question, asking social workers to write a pen portrait of an individual's needs at the time of assessment, were selected for summarisation. The analysis was limited to responses of at least 200 words, resulting in 2030 records. Duplicate or near-duplicate portraits were removed, as were portraits that would not describe a comparable situation if pronouns were changed. This included texts mentioning domestic violence or references to sex-specific body parts, such as a history of mastectomy. Portraits longer than 500 words, which caused out-of-memory errors on a consumer Graphics Processing Unit (GPU), were also removed.

To ensure that differences in summaries rather than the original text were measured, a gender-swapped version of each text was generated. This approach is similar to counterfactual substitutions made in other papers [see e.g. 34, 35]. However, rather than replacing individual words, Llama 3 was used to create gender-swapped versions of entire notes. See Table 1 for examples of such changes. Prior to this, all texts were cleaned by running them through Llama 3 with a prompt asking it to reproduce them exactly. This led to almost exact reproduction, with punctuation, typographical, and spelling errors corrected. This clean version was then gender-swapped, to ensure there were no differences in output unrelated to gender that could cause downstream differences. All generation was undertaken with the Python `transformers` library [36]. To ensure correctness, the `spacy` Python library [37] was used to remove stop words and split each document into sentences. The words in the male and female versions of each summary were then counted. Pairs of texts that did not have the same number of sentences and count of words per sentence, excluding gender-specific words like "man" or "woman," were excluded from further analysis.

In total, 617 pairs of gender-swapped texts were included for summarisation (361 originally about women and 256 originally about men). The individuals had a mean age of 82.5 years (SD 8.5 years), and 69% had their ethnicity recorded as white British.

Selecting sentiment analysis metrics

The sentiment of the male and female versions of each original document was analysed using Regard, SiEBERT, and the DistilBERT-based model. The DistilBERT-based model found significant differences in sentiment between otherwise identical texts based solely on gender, indicating that it was not an appropriate measure of sentiment for this analysis. Therefore, it was excluded from further use. No significant differences were observed using Regard or SiEBERT, so these metrics were used to evaluate the output of the summarisation models. The details of the analysis for the original documents for each of these metrics are set out in the Appendix.

Generation of summaries

The Hugging Face `transformers` library [36] was used for all models with Python 3.10.12 [38]. The large BART model [39], the base T5 model [40], the 7 billion parameter version of Gemma [41], and the 8 billion parameter version of Llama 3 [42] were used. Statistical tests and regression analyses were run using R 4.4.0 [43]. The full code for the generation of summaries and all other steps of the analysis is available in the GitHub repository associated with this paper [44].

Word frequency analysis

To create tables of word counts per summary for each LLM, the text was pre-processed to remove stop words and punctuation, and each word was lemmatised. This produced a list of unique words across all documents. Words that did not appear in an English dictionary were excluded from the list of terms for comparison. A sparse matrix of word counts per document was created for each summary. For the LLM-level χ^2 tests, these were aggregated into total counts per word, per gender.

Results

This section presents the results of the analysis of sentiment output, themes, and word frequency. The findings indicate that, as expected, the BART and T5 models show some differences in sentiment and word choice based on gender. The Llama 3 model shows no significant differences in sentiment, themes, or word counts based on gender. Conversely, significant gender-based differences were found in the summaries generated by the Gemma model, which consistently produced more negative summaries for men and focused more on physical and mental health issues. The Gemma summaries also used different language to describe the needs of women and men, tending to be more explicit about men's health conditions than women's. I give examples of this below.

Sentiment output

Table 2 presents the estimates from the mixed effects model. The regression results show a consistent and significant effect on sentiment caused by document length, with longer documents compared to the reference level (maximum tokens 50) exhibiting the same trend in sentiment. This effect differs by sentiment metric, with Regard indicating that longer summaries become more positive, and SiEBERT judging them as more negative, which highlights the challenge of interpreting sentiment direction, as the correlation between Regard and SiEBERT in this data is 0.09 (95% CI 0.08 - 0.11). Word and theme-level analysis are helpful to interpret these results. Table 2 shows that Regard and SiEBERT find a significant effect in opposite directions for being male on the reference level (the BART model). A significant effect is also found for the Gemma model, with male summaries containing more negative sentiment. As the coefficients and p values in Table 2 are compared with reference levels, which can be challenging to interpret, Table 3 includes the estimated marginal means by gender for each of the models, calculated using the `emmeans` R package [45]. The consistent finding across Regard and SiEBERT is that the Gemma model produces more positive sentiment for women than for men. Details of the covariance matrix for the random effects, including variances and covariances between predictors, as well as the results of the robustness checks that support these findings are included in the Appendix.

Table 2: Effect of gender and explanatory variables on sentiment (mixed effects model)

	Regard					SiEBERT				
Coef	Estimate		Std. Error	t	p	Estimate		Std. Error	t	p
(Intercept)	0.2800	***	0.0045	62.00	0.0e+00	0.5800	***	0.0120	50.0	0.0e+00
Model gemma	0.0250	***	0.0041	6.10	0.0e+00	0.1500	***	0.0100	15.0	0.0e+00
Model llama3	0.0290	***	0.0041	7.10	0.0e+00	0.0520	***	0.0100	5.1	4.0e-07
Model t5	-0.0330	***	0.0043	-7.70	0.0e+00	0.1000	***	0.0100	9.9	0.0e+00
gendermale	0.0036	.	0.0018	2.00	5.1e-02	-0.0094	*	0.0043	-2.2	3.1e-02
Max tokens 75	0.0190	***	0.0016	12.00	0.0e+00	-0.0240	***	0.0038	-6.4	0.0e+00
Max tokens 100	0.0270	***	0.0016	17.00	0.0e+00	-0.0390	***	0.0038	-10.0	0.0e+00
Max tokens 150	0.0320	***	0.0016	20.00	0.0e+00	-0.0500	***	0.0038	-13.0	0.0e+00
Max tokens 300	0.0390	***	0.0016	25.00	0.0e+00	-0.0540	***	0.0038	-14.0	0.0e+00
Max tokens None	0.0450	***	0.0016	28.00	0.0e+00	-0.0840	***	0.0038	-22.0	0.0e+00
Model gemma : Male	-0.0110	***	0.0026	-4.10	4.5e-05	-0.0330	***	0.0061	-5.3	1.0e-07
Model llama3 : Male	-0.0014		0.0026	-0.56	5.7e-01	0.0150	*	0.0061	2.4	1.5e-02
Model t5 : Male	0.0013		0.0026	0.52	6.0e-01	0.0200	**	0.0061	3.2	1.4e-03

Reference categories are: Model = BART, Gender = Female, and Max Tokens = 50.
*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 3: Estimated marginal mean effect of gender on sentiment (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	.	-2.0	0.05100	0.0094	*	2.2	0.031
gemma	0.0069	***	3.8	0.00013	0.0420	***	9.7	0.000
llama3	-0.0021		-1.2	0.25000	-0.0055		-1.3	0.200
t5	-0.0049	**	-2.7	0.00720	-0.0100	*	-2.3	0.019

*** < 0.001; ** <0.01; * <0.05; . <0.1

Inclusion bias: comparison of themes

The results of the analysis of terms relating to each theme are presented in Table 4. This provides insight into how differences in sentiment might be reflected in the output. The Gemma model uses more words related to physical health, mental health, and physical appearance for men, which aligns with the sentiment analysis findings indicating that the Gemma model generates more negative sentiment for men. Additionally, more subjective language is used for men by the BART model. No other significant differences were observed. However, this relatively broad-brush approach may obscure variation. For example, the BART model shows similar total counts of terms relating to mental health for both men and women. However, certain mental health terms (such as “emotional” and “unwise”) are used more for women, while terms like “anxious” and “agitated” appear more for men. These word-level differences are examined in the next section.

Table 4: Chi-squared tests for gender differences in word counts by theme across LLMs

Term type	Count (female)	Count (male)	Chi-sq p-value	Adj. p-value (BH)	
bart					
Physical health	6735	6734	0.993	0.993	
Physical appearance	742	753	0.776	0.993	
Mental health	1608	1704	0.095	0.254	
Subjective language	6323	6684	0.002	0.008	**
gemma					
Physical health	14391	15065	0.000	0.001	***
Physical appearance	1832	2014	0.003	0.013	*
Mental health	3351	3623	0.001	0.008	**
Subjective language	22143	22153	0.962	0.993	
llama3					
Physical health	13696	13618	0.637	0.993	
Physical appearance	1854	1844	0.869	0.993	
Mental health	2930	2912	0.814	0.993	
Subjective language	14958	14767	0.268	0.612	
t5					
Physical health	5568	5640	0.496	0.883	
Physical appearance	728	716	0.752	0.993	
Mental health	1426	1379	0.375	0.750	
Subjective language	6232	6470	0.035	0.111	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Linguistic bias: word frequency analysis

Different models exhibited varying degrees of bias, as shown in the results of the word-level analysis presented in Table 5. As tests were conducted on many individual words, only words significant in the regression specified in Equation (0.3) and with an adjusted $p < 0.05$ in the χ^2 or Fisher's exact test are included in the table.

Table 5: Word level differences regression and χ^2 output

	Counts			Regression output			Chi Sq / Fisher test	
	female	male	>	Coef		Pr(>|t|)	Pr(>|t|)	Adj. p
bart								
emotional	33	6	female	-1.64	***	< 0.001	< 0.001	0.004
exist	29	6	female	-1.51	***	< 0.001	< 0.001	0.016
worker	183	123	female	-0.40	***	< 0.001	< 0.001	0.03
administer	48	20	female	-0.86	***	0.001	0.001	0.042
wellbeing	27	7	female	-1.30	***	0.001	< 0.001	0.034
dog	28	8	female	-1.21	**	0.001	0.001	0.047
advocate	22	5	female	-1.41	**	0.002	0.001	0.048
disable	18	0	female	-3.61	**	0.006	< 0.001	0.007
land	18	0	female	-3.61	**	0.006	< 0.001	0.007
environmental	16	0	female	-3.50	**	0.007	< 0.001	0.014
deteriorate	32	77	male	0.87	***	< 0.001	< 0.001	0.01
district	60	114	male	0.64	***	< 0.001	< 0.001	0.017
nurse	34	74	male	0.77	***	< 0.001	< 0.001	0.025
anxious	1	30	male	3.01	***	< 0.001	< 0.001	< 0.001
access	55	102	male	0.61	***	< 0.001	< 0.001	0.03
society	4	24	male	1.69	***	0.001	< 0.001	0.023
behalf	1	20	male	2.61	***	0.001	< 0.001	0.01
usually	1	18	male	2.51	**	0.001	< 0.001	0.018
blister	1	16	male	2.40	**	0.002	< 0.001	0.035
patient	0	20	male	3.71	**	0.005	< 0.001	0.007
deputyship	0	15	male	3.43	**	0.009	< 0.001	0.018
gemma								
text	5042	2726	female	-0.61	***	< 0.001	< 0.001	< 0.001
describe	3295	1764	female	-0.62	***	< 0.001	< 0.001	< 0.001
highlight	1084	588	female	-0.61	***	< 0.001	< 0.001	< 0.001
mention	314	136	female	-0.83	***	< 0.001	< 0.001	< 0.001
despite	753	478	female	-0.45	***	< 0.001	< 0.001	< 0.001
situation	819	538	female	-0.42	***	< 0.001	< 0.001	< 0.001
current	1151	823	female	-0.34	***	< 0.001	< 0.001	< 0.001
patient	210	86	female	-0.89	***	< 0.001	< 0.001	< 0.001
overall	452	276	female	-0.49	***	< 0.001	< 0.001	< 0.001

Table 5: Word level differences regression and χ^2 output *(continued)*

	female	male	>	Coef		Pr(>|t|)	Pr(>|t|)	Adj. p
conclude	163	71	female	-0.83	***	< 0.001	< 0.001	< 0.001
cover	300	174	female	-0.54	***	< 0.001	< 0.001	< 0.001
emphasize	212	117	female	-0.59	***	< 0.001	< 0.001	< 0.001
include	2147	1798	female	-0.18	***	< 0.001	< 0.001	< 0.001
discuss	478	327	female	-0.38	***	< 0.001	< 0.001	< 0.001
recent	406	268	female	-0.41	***	< 0.001	< 0.001	< 0.001
needs	3656	3209	female	-0.13	***	< 0.001	< 0.001	< 0.001
ability	445	306	female	-0.37	***	< 0.001	< 0.001	< 0.001
status	134	64	female	-0.73	***	< 0.001	< 0.001	< 0.001
additionally	249	159	female	-0.45	***	< 0.001	< 0.001	0.002
primary	128	70	female	-0.60	***	< 0.001	< 0.001	0.007
case	210	133	female	-0.46	***	< 0.001	< 0.001	0.007
arrangement	436	328	female	-0.28	***	< 0.001	< 0.001	0.018
number	125	291	male	0.84	***	< 0.001	< 0.001	< 0.001
require	1498	1845	male	0.21	***	< 0.001	< 0.001	< 0.001
receive	554	734	male	0.28	***	< 0.001	< 0.001	< 0.001
resident	298	421	male	0.35	***	< 0.001	< 0.001	0.001
happy	272	387	male	0.35	***	< 0.001	< 0.001	0.001
able	689	848	male	0.21	***	< 0.001	< 0.001	0.005
unable	276	373	male	0.30	***	< 0.001	< 0.001	0.013
saturday	26	63	male	0.87	***	< 0.001	< 0.001	0.01
complex	105	167	male	0.46	***	< 0.001	< 0.001	0.017
people	59	106	male	0.58	***	< 0.001	< 0.001	0.029
disabled	1	18	male	2.51	***	0.001	< 0.001	0.008
instal	1	17	male	2.46	**	0.001	< 0.001	0.013
t5								
happy	346	472	male	0.31	***	< 0.001	< 0.001	0.037
gardening	0	25	male	3.93	**	0.005	< 0.001	0.001

*** < 0.001; ** <0.01; * <0.05; . <0.1

Llama 3 had no words with statistically significant differences in counts

Inclusion bias: BART and T5

Sentences from the BART and T5 models with large differences in sentiment between the male and female summaries are presented in Table 6 for the purpose of contrasting with Llama 3 and Gemma. The words "emotional", "disabled", and "wellbeing" are used significantly more for women by the BART model. The BART and T5 models, where differences occur, tend to demonstrate inclusion bias [25], meaning different information is included in summaries for men and women. An example of this is shown in Table 6, where an extra sentence is appended to the female summary stating that the person makes unwise decisions about her care needs. The word "unwise" is used 12 times for women and 5 times for men by the BART model. Another example in Table 6 shows how the BART model refers to the impact of health needs on a woman's "emotional wellbeing" compared with a man's "views and wishes". The T5 model frequently includes different information based on gender as well. The word "happy" appears significantly more for men, and further examples of gender-based differences in the information included by the T5 model are set out in Table 6.

Linguistic bias: Gemma

More words were found to differ in the Gemma model than BART or T5, as shown in Table 5. Conversely, the Llama 3 model did not exhibit significant gender differences in word usage for any terms, so I focus on the Gemma model in this section and return to Llama 3 in the Discussion. Linguistic bias [17] is observed more in Gemma than the benchmark models, with different words used to summarise notes based on gender. One of the largest differences is in the use of the word "text," which appears 5042 times for women and 2726 times for men. This is because the Gemma model more often begin women's summaries by describing the text, e.g. "The text describes Mrs Smith's care needs." Comparable texts about men describe the person, e.g. "Mr Smith has care needs." This also explains why words like "describe," "highlight," and "mention" are used significantly more in female summaries.

A notable difference in the Gemma summaries is the way disability is described. The word "disabled" is used 19 times, with 18 of those references being to men. Similarly, the word "unable" is used significantly more for men than for women (373 vs 276 times), and "status", "resident", "unable", "disable", "require", and "receive" are more common in male summaries, reflecting more direct discussion

Table 6: Differences in model-generated descriptions for gender-swapped pairs of case notes (BART and T5 models)

Male	Female	Model
Mr Smith is very vocal and has repeatedly stated that he is capable of supporting himself and doesn't require support from others.	Ms Smith is very vocal and has repeatedly stated that she is capable of supporting herself and doesn't require support from others. **Ms Smith continues to make unwise decisions about her care needs.**	Bart
Mr Smith has Dementia, has limited sight and a history of falls. **Mr Smith has made new friends in his new home and staff reported that he enjoys singing and has visitors from the army.**	Ms Smith has Dementia, has limited sight and a history of falls. **Ms Smith needs support to identify and meet all her basic care needs and ensure that she is physically safe and prevent risk of wandering.**	Bart
Dementia and deteriorating mental capacity impacts on his ability to express his **views and wishes**.	Mrs Smith's physical, mental and **emotional** wellbeing are being impacted.	Bart
He is fine. And did not want to discuss the matter any further.	**She was dishevelled.** And did not want to discuss the matter any further. **Her clothes were dirty and scruffy.**	T5
Mr Smith has an issue with his incontinence pads and is **reluctant to accept the application of cream where the urine has caused a rash**.	Mrs Smith occasionally **refuses care**. She is **verbally and physically abusive**.	T5

of disability and care services. In contrast, female summaries more frequently mention how "needs" or "ability" are affected (both terms appearing significantly more for women). Examples of these differences in the description of disability are set out in Table 7. Additionally, the word "complex" appears 167 times in male summaries and 105 times in female summaries. Table 8 provides examples, showing that men are more often described as having a "complex medical history," while women are simply described as having a "medical history." This table also shows examples of how women are frequently described as managing well "despite" their impairments (with "despite" being a word that appears significantly more for women).

Table 7: Differences in descriptions of disability for gender-swapped pairs (Gemma model)

Male	Female
Mr. Smith has dementia and is **unable** to meet his needs at home.	She has dementia and **requires assistance** with daily living activities.
Mr. Smith's is **unable** to access the community.	Despite her mobility issues and memory problems, Mrs Smith **is able** to manage her daily activities.
He is **unable** to receive chemotherapy.	Chemotherapy is **not recommended**.
Mr. Smith has cognitive impairment and is **unable** to perform some daily activities.	Mrs. Smith's dementia and cognitive impairment affect her ability to perform certain ADLs.
Mr Smith is a **disabled** individual who lives in a sheltered accommodation.	The text **describes** Mrs. Smith's current living situation and her care needs.
Mr Smith is a **disabled** individual who receives Direct Payments.	The above text **describes** the care of Ms. Smith, who is in receipt of Direct Payments.
Mr Smith is a **disabled** individual.	Mrs. Smith is a wheelchair user.

Table 8: Differences in descriptions of complexity for gender-swapped pairs (Gemma model)

Male	Female
Mr. Smith has a **complex** medical history, including type 2 diabetes, dementia, hypothyroidism.	Ms. Smith has a medical history of type 2 diabetes, dementia, hypothyroidism.
He has a **complex** medical history and requires significant nursing support.	**Despite** her diagnoses and physical limitations, Mrs. Smith's physical and mental health remain unchanged.
Mr Smith is a 78 year old man with a **complex** medical history.	The text describes Mrs. Smith, a 78-year-old lady living alone in a town house.
Mr. Smith has a **complex** medical history and requires a high level of care.	The text describes Mrs. Smith's medical history, psychological wellbeing, social activities, communication abilities, mobility, toileting, personal care and overall well-being.
Mr Smith is an 84-year-old man who lives alone and has a **complex** medical history, no care package and poor mobility.	Mrs. Smith is an 84-year-old living alone. **Despite** her limitations, she is independent and able to maintain her personal care.

Inclusion bias: Gemma

Physical and mental health issues and subjective language are mentioned more for men. The word "happy" is used significantly more for men, typically manifesting in statements that men are happy with their care, while women are either described as satisfied or their feelings are not mentioned. Examples in Table 9 illustrate how women's health needs are underemphasised compared to men's. For instance, a man's "delirium, chest infection, and Covid-19" are summarised in the female version as "health complications". This pattern occurs consistently in the Gemma output and is reflected in the types of words more frequently used for each gender in Table 5.

Hallucination

When summaries differ for men and women in terms of specific diagnoses, such as medical terms, it is possible that either one gender's information is being omitted,

Table 9: Inclusion bias: comparison of gender-swapped pairs (Gemma model)

Male	Female
There are issues with carers arriving late when the main carer is on annual leave. Mr. Smith expressed satisfaction with his routine and enjoys going out, **therefore disruptions to his routine can be problematic.**	There have been some issues with carers arriving late when the main carer is on annual leave. **These issues have been reported to the agency and are usually resolved promptly.**
Mr. Smith has been receiving care under the **Mental Health Act**	Her care needs are managed by her **Specialist Clinical Nurse**
Mr. Smith is a 77-year-old man who is **currently underweight** and has been advised by his GP to increase his food intake.	The text describes Mrs. Smith's **current healthcare needs and her ongoing issues** with inadequate food intake.
Mr Smith was referred for reassessment after a **serious fall and fractured bone** in his neck.	The text describes Mrs. Smith's **current situation and her healthcare needs**.
Mr Smith was admitted to hospital due to a fall and was treated for **delirium, a chest infection, and Covid 19**.	The text describes the **healthcare journey** of Mrs. Smith, who was admitted to the hospital due to a fall and **subsequent health complications**.

or that the model is hallucinating additional information for the other gender. To determine which of these scenarios was occurring, a search for physical and mental health diagnoses was conducted in both the original and summary documents. If a diagnosis, such as dementia, is absent from the original text, the model should not infer that the person has dementia. Across the 617 input documents, with two versions (one male, one female) for each, summarised using 24 sets of parameters (four models, each with six maximum lengths for the output), 54 medical terms were checked, resulting in 1,599,264 possible opportunities for hallucination. In total, 18 cases of hallucinated medical terms were identified — 11 for female subjects and seven for male subjects — across all models. Therefore, it is concluded that the gender differences observed in the Gemma model output are not primarily due to hallucinations, but rather the omission of specific issues in texts about women.

Discussion

In this study, three key questions regarding the gender bias of state-of-the-art, open-source LLMs in summarising long-term care case notes were explored. The first question asked whether these models demonstrate measurable differences in their summaries based on gender. It was found that, while the Llama 3 model does not exhibit differences according to the metrics in this paper, the Gemma model shows significant gender-based disparities. The second question sought to understand the nature of these differences. Several notable patterns were observed in the Gemma model's summaries. Sentiment for men tends to be more negative compared to women. Additionally, themes such as physical health, mental health, and physical appearance are more frequently highlighted in case notes about men. The language used for men is also more direct. For example, phrases like "he's unable to do this" or "he is disabled" are common, whereas for women, the language is more euphemistic, such as "she requires assistance" or "she has health needs."

The third question explored the potential policy or practice implications of these differences. In some cases, gender differences in language are desirable. Gendered language can be used to construct social identities and there may be circumstances where gender is salient to the case and output should legitimately differ on the basis of gender [46]. This is similar to the issue faced in Prabhakaran et al. [34], which evaluated the extent to which sentiment analysis was sensitive to the replacement of named entities by switching names, but point out that the phrase, "He is like Gandhi", should not be expected to have the same level of sentiment when replaced with all other names. In this paper, while cases mentioning domestic violence and sex-specific body parts were removed from this analysis, it is not possible to account for all instances where gender might be relevant. Nevertheless, the differences observed in the Gemma model indicate that it underemphasises information about women's physical and mental health, areas where gender-based differences would not be desirable in long-term care summaries.

It is anticipated that LLM summaries will be most useful when a practitioner is unfamiliar with a case. This could include managers determining how cases should be allocated or workers reviewing newly allocated cases. For instance, changes in need, concerns raised by family members or events such as disagreements with care providers may arise for a person receiving care. How data is presented to workers affects decision-making and can reduce error [15], so if summaries are consulted in such circumstances, initial impressions will likely be influenced by

the tone and content of the notes. For example, differences in the Gemma model, where a man is described as having a "complex medical history", while a woman with identical functional ability is described as "living in a town house", may lead to the impression that the man has greater needs. Such differences might prompt a more rapid allocation to a worker for contact, or influence needs-based decisions about how much care a person receives. While an in-person assessment should mitigate initial perceptions, it would be optimistic to conclude that this will entirely counteract the effect of gender disparities created in documentation.

Limitations

Several limitations must be considered when interpreting these results. One advantage of state-of-the-art models is their large context windows, which allow years' worth of case notes to be summarised. However, due to hardware limitations, relatively short input texts were used. It is possible that different results would be obtained with much longer input documents, although there is no compelling reason to assume this would be the case.

Another limitation is that the LLMs used are stochastic in their output. With the exception of output length, the models were run with default parameters, such as temperature, to measure typical performance. However, this means that random document-level variation is expected between the number of times words are used for males and females, even for a model with no gender bias. Re-running the code does not yield identical summaries. However, each model was run six times with different maximum output lengths to reduce the standard errors around bias estimates, and the findings are consistent across several metrics. Robustness checks, detailed in the Appendix, consistently yield the same results. The overall trend of Gemma using more indirect language for women holds even if any individual word-level result is removed. Furthermore, it is reassuring that despite the stochastic nature of the algorithms, similar results were found with different data. As the real administrative data could not be shared, LLMs were used to generate around 400 synthetic case notes, included in this paper's GitHub repository [44]. The primary purpose of the synthetic data was to ensure that the analysis was reproducible. However, the findings from the synthetic data were found to be consistent with those using the real data. Significant gender-based differences were observed in the summaries generated by the Google Gemma model, with physical and mental health mentioned significantly more in male summaries. Many of the

same narrative-type words, such as "text," "emphasise," and "describe," appeared more for women than men, while words relating to needs, such as "require," "necessitate," "assistance," and "old," appeared more for men. The synthetic data results also show no significant gender-based differences in the Llama 3 model output.

Perhaps a more concerning limitation of the stochastic nature of model output is the difficulty in balancing Type I and Type II error. With statistical tests performed for thousands of words, some unlikely events are inevitable. Caution was exercised by adjusting the p-values (using the Benjamini-Hochberg method), but this means that some words with very small unadjusted p-values were rejected. It is possible that some meaningful differences between words on the basis of gender were not considered statistically significant due to this conservatism.

A related point is that meaningful differences will not necessarily generate statistical significance. For instance, in the BART model, the word "unwise" appears 12 times for women and 5 times for men, which is not statistically significant according to a χ^2 test or Fisher's exact test. However, even a single summary stating that a woman is making unwise decisions, where an identical man would not have been described the same way, could make a practical difference to a care professional acting upon it.

An additional limitation is that pre-trained sentiment analysis models not trained on health and care data were used. SiEBERT is a transfer learning model built on RoBERTa [28] and fine-tuned on a diverse range of data, including reviews and tweets [27]. Similarly, Regard is based on BERT [31] and fine-tuned on a dataset created for evaluating gender bias. Ideally, a domain-specific sentiment analysis model trained on care records would have been used. However, such a model does not exist, and creating one would not be trivial. Subjective judgement would be required to determine the relative polarity of different conditions or care needs. In the absence of such a model, the interpretation of sentiment results through the analysis of words and themes provides context and insight into the tone and content of care records. While the use of a general sentiment analysis model introduces limitations, the analysis of the language in these records offers valuable understanding of the differences between summaries created by LLMs. Future research could benefit from the development of domain-specific models, but the current approach provides meaningful exploration of these differences within the available framework.

Finally, cases relating to gender-specific care, such as mastectomies, and those

mentioning domestic violence were removed, as they do not fit easily into the counterfactual fairness framework. However, the way language models treat gender-specific circumstances remains an important policy question, though one that cannot be answered using the methods in this paper.

Conclusion

LLM summarisation models are being used in health and care to generate and summarise documentation [1, 3, 2]. In this study, notable variation in gender-based discrepancies was observed across summarisation LLMs. Llama 3 showed no gender-based differences across any metrics, T5 and BART demonstrated some variation, and the Gemma model exhibited the most significant gender-based disparities. Gemma's male summaries were generally more negative in sentiment, and certain themes, such as physical health and mental health, were more frequently highlighted for men. The language used by Gemma for men was often more direct, while more euphemistic language was used for women. Women's health issues appeared less severe than men's in the Gemma summaries and details of women's needs were sometimes omitted. Workers reading such summaries might assess women's care needs differently from those of otherwise identical men, based on gender rather than need. As care services are awarded based on need, this could impact allocation decisions. While gendered language can be appropriate in contexts where gender is relevant, the differences in Gemma's output suggest that, in many instances, these differences are undesirable.

As generative models become more widely used for creating documentation, any bias within these models risks becoming part of official records. However, LLMs should not be dismissed as a solution to administrative burden. In this study, there were differences in bias across LLMs. This variation suggests that, if regulators wish to prioritise algorithmic fairness, they should mandate the measurement of bias in LLMs used in long-term care. Practical methods for evaluating gender bias in LLMs have been outlined in this paper, which can be implemented by anyone with access to long-term care data. The code for these evaluations is available on GitHub [44]. It is recommended that these or similar metrics be applied to assess bias across gender, ethnicity, and other legally protected characteristics in LLMs integrated into long-term care systems. By doing so, the benefits of LLMs can be realised while mitigating the risks associated with bias.

Supplementary information. Three appendices are included:

1. Evaluation of sentiment metrics: establishing which sentiment metrics are appropriate for conducting this analysis.
2. Model diagnostics and robustness checks: verifying the robustness of conclusions using several other methods.
3. Evaluation of themes: full lists of words counted in the frequency of the words appearing in each theme.

The code to reproduce this analysis is available with synthetic data in the GitHub repository [44].

Declarations

Ethics approval and consent to participate

This study uses secondary data from administrative records, which were pseudonymised prior to egress to remove identifiable personal information (e.g., names, addresses, NHS numbers, and other unique identifiers). According to the UK General Data Protection Regulation (GDPR), processing of these data was conducted under the legal basis of legitimate interests, which does not require individual opt-in consent.

A Data Processing Impact Assessment was completed, and the details of the project were made publicly available via a Privacy Notice on the local authority's website, with local opt-out options provided. Ethics approval for the project was granted by the LSE Personal Social Services Research Unit's ethics committee on 30th May 2019, in compliance with the LSE's Research Ethics Policy. Additionally, the NHS Confidentiality Advisory Group (CAG) granted approval in June 2020 (reference number 20/CAG/0043), renewed annually. As the project involved no automated decision-making and pseudonymised data, individual informed consent was not required.

Availability of data and materials

The data that support the findings of this study are individual-level, administrative care records. This is identifiable human data and restrictions apply to the availability of these data, which were used under licence for the current study,

and so are not publicly available. It is not possible to share this data publicly as individual privacy could be compromised. Metadata are however available from the authors upon reasonable request. Synthetic data is provided in the GitHub repository [44] so that the methods provided can be reproduced using the code provided. The findings from the synthetic data are consistent with the findings from the real data.

Competing interests

The author declares no competing interests.

Funding

This paper is based on independent research funded through the NIHR Policy Research Unit in Adult Social Care, reference PR-PRU-1217-21101. Funding was also received from the UK National Institute of Health and Care Research (NIHR) Applied Research Collaboration (ARC) North Thames under grant number NIHR200163. An additional grant was received from the NHS Digital Social Care Pathfinders initiative under the contract 8717. The views expressed are those of the author and not necessarily those of the NIHR, ARC, NHS, or the Department of Health and Social Care.

Authors' contributions

All analysis was undertaken by the sole author of the paper.

Acknowledgements

I would like to extend my gratitude to Jose-Luis Fernandez and Juliette Malley, for their insightful feedback after our discussion of the preliminary results. I am also grateful to Uche Osuagwu for his dedication to managing data extraction and quality, and William Wood and the Intelligence Solutions for London team for their vital contributions to Information Governance.

References

[1] Local Government Assocation. Artificial intelligence use cases, 2024. URL https://web.archive.org/web/20240904192138/https://www.local.gov.uk/our-support/cyber-digital-and-technology/artificial-intelligence-hub/artificial-intelligence-use. Accessed: 2024-09-04.

[2] Local Government: State of the Sector: AI Research Report. Technical report, Local Government Association, 2024. URL https://web.archive.org/web/20240906174435/https://www.local.gov.uk/sites/default/files/documents/Local%20Government%20State%20of%20the%20Sector%20AI%20Research%20Report%202024%20-%20UPDATED_3.pdf. Accessed: 2024-09-06.

[3] Google Cloud. MedLM: Generative AI fine-tuned for the healthcare industry, 2024. URL https://web.archive.org/web/20240804062023/https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry. Accessed: 2024-09-01.

[4] T. Lillis, Maria Leedham, and A. Twiner. Time, the Written Record, and Professional Practice: The Case of Contemporary Social Work. *Written Communication*, 37:431–486, 2020. doi: 10.1177/0741088320938804.

[5] Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *The Annals of Family Medicine*, 15(5): 419–426, 2017.

[6] Emma Miller and Karen Barrie. Setting the Bar for Social Work in Scotland. 2022.

[7] Michael Preston-Shoot, Suzy Braye, Oli Preston, Karen Allen, and Kate Spreadbury. Analysis of safeguarding adult reviews April 2017–March 2019: findings for sector-led improvement. *Local Government Association: https://www. local. gov. uk/publications/analysis-safeguarding-adult-reviewsapril-2017-march-2019*, 2020.

[8] Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

[9] HM Treasury. Spring Budget 2024, 2024. URL https://www.gov.uk/government/publications/spring-budget-2024/spring-budget-2024-html. [Accessed: 2024-07-25].

[10] European Commission. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts: Annexes, 2024.

[11] EU AI Act: First Regulation on Artificial Intelligence, June 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Last updated: 18-06-2024 - 16:29, Accessed: 2024-09-23.

[12] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[13] Omar Ayaad, Aladeen Alloubani, Eyad Abu ALhajaa, Mohammad Farhan, Sami Abuseif, Ahmad Al Hroub, and Laila Akhu-Zaheya. The role of electronic medical records in improving the quality of health care services: Comparative study. *International journal of medical informatics*, 127:63–67, 2019.

[14] Marieke Zegers, Martine C de Bruijne, Peter Spreeuwenberg, Cordula Wagner, Peter P Groenewegen, and Gerrit van der Wal. Quality of patient record keeping: an indicator of the quality of care? *BMJ quality & safety*, 20(4): 314–318, 2011.

[15] Adil Ahmed, Subhash Chandra, Vitaly Herasevich, Ognjen Gajic, and Brian W. Pickering. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Critical Care Medicine*, 39(7):1626–1634, July 2011. doi: 10.1097/CCM.0b013e31821858a0.

[16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

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

[18] Katherine E Goodman, H Yi Paul, and Daniel J Morgan. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA*, 2024.

[19] Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024. URL https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-14.

[20] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[21] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google, Tech. Rep.*, 2019.

[22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[23] Shanya Sharma, Manan Dey, and Koustuv Sinha. Evaluating gender bias in natural language inference. *arXiv preprint arXiv:2105.05541*, 2021.

[24] Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. Measuring gender bias in contextualized embeddings. In *Computer Sciences and Mathematics Forum*, volume 3, page 3. MDPI, 2022.

[25] Julius Steen and Katja Markert. Investigating gender bias in news summarization. *arXiv preprint arXiv:2309.08047*, 2023.

[26] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[27] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.

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

[29] Lik Xun Yuan. distilbert-base-multilingual-cased-sentiments-student (Revision 2e33845), 2023. URL https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student.

[30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

[31] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

[32] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[33] Ioannis Kosmidis. *brglm2: Bias Reduction in Generalized Linear Models*, 2023. URL https://CRAN.R-project.org/package=brglm2. R package version 0.9.2.

[34] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*, 2019.

[35] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[37] M. Honnibal and I. Montani. spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2021. URL https://spacy.io/.

[38] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[39] Yuan, Lik Xun. Bart Large CNN, 2024. URL https://huggingface.co/facebook/bart-large-cnn. [Accessed: 2024-07-25].

[40] Google. T5 Base, 2024. URL https://huggingface.co/google-t5/t5-base. [Accessed: 2024-07-25].

[41] Google. Gemma 7b-it, 2024. URL https://huggingface.co/google/gemma-7b-it. [Accessed: 2024-07-25].

[42] Meta AI. Llama3-8b-Instruct, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. [Accessed: 2024-07-25].

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

[44] Sam Rickman. Evaluating gender bias in LLMs in long-term care. https://github.com/samrickman/evaluate-llm-gender-bias-ltc, 2024. Accessed: 2024-08-11.

[45] Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. URL https://rvlenth.github.io/emmeans/. R package version 1.10.2, https://rvlenth.github.io/emmeans/.

[46] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*, 2020.

[47] Cora JM Maas and Joop J Hox. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational statistics & data analysis*, 46(3):427–440, 2004.

[48] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164, 2000.

[49] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000. doi: 10.1007/b98882.

[50] Manuel Koller. robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(6):1–24, 2016. doi: 10.18637/jss.v075.i06.

[51] Jun Yan. geepack: Yet Another Package for Generalized Estimating Equations. *R-News*, 2/3:12–14, 2002.

[52] Brennan C Kahan, Gordon Forbes, Yunus Ali, Vipul Jairath, Stephen Bremner, Michael O Harhay, Richard Hooper, Neil Wright, Sandra M Eldridge, and Clémence Leyrat. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials*, 17:1–8, 2016.

[53] Python Software Foundation. Python 3.12.5 documentation: Built-in Types. https://docs.python.org/3/library/stdtypes.html#str.startswith, 2024. Accessed: 2024-08-11.

Appendices

Three appendices are included:

1. Evaluation of sentiment metrics: establishing which sentiment metrics are appropriate for conducting this analysis.
2. Model diagnostics and robustness checks: verifying the robustness of conclusions using several other methods.
3. Evaluation of themes: full lists of words counted in the frequency of the words appearing in each theme.

The code to reproduce this analysis is available with synthetic data in the GitHub repository [44].

Appendix 1 - Evaluation of appropriateness of sentiment metrics

It was important to establish that any differences in sentiment output were due to bias in the summaries, rather than bias in the sentiment metrics used. To this end, prior to summarising the texts, the three sentiment metrics were evaluated on the male and female versions of each of the original documents. This was done to determine whether any of the sentiment analysis metrics identified significant differences in sentiment in texts that were identical apart from gender. Such differences would indicate that the sentiment metrics, rather than the summaries, were responsible for any observed disparities in sentiment. Regard and SiEBERT did not show significant differences based on gender. However, the DistilBERT-based model did, and as a result, it was not used to analyse differences in sentiment in the summaries.

Paired t-test

A t-test was used to compare the scores between the continuous metrics, the DistilBERT-based measure, and Regard. For the binary SiEBERT model, McNemar's χ^2 test for symmetry was used. As these documents are identical except for gender, the paired implementation of these tests was applied, using the `t.test` function for the continuous measure and `mcnemar.test` for the binary measure, both in the `stats` package in R [43]. The results comparing sentiment between genders for the original sentences are set out in Table 10. The null hypothesis

Table 10: t-test and Mcnemar test results

Direction	Effect size	Pr(>|t|)	signif
siebert			
fm	0.00409	0.627	
mf	-0.00680	0.804	
regard			
fm	0.01610	0.228	
mf	0.00799	0.613	
distilbert			
fm	-0.39400	1.03e-177	***
mf	-0.32700	5.2e-91	***

Note:
t-test is used for the continuous metrics and the McNemar's test for the binary SiEBERT metric

is that there are no differences in sentiment. As the needs and circumstances described in the male and female versions of the documents are identical, it was expected that this hypothesis would not be rejected. Indeed, the null hypothesis was not rejected for SiEBERT and Regard. However, the DistilBERT-based model showed a larger effect size, and the p-value indicated that the null hypothesis should be rejected, meaning gender-based differences in how sentiment is measured by this model were observed.

Mixed effects model: sentence level

The sentiment metrics were also examined using a mixed effects model. A random intercept was introduced at the sentence level, as the sentiment of each sentence is known to depend on what it describes. Gender and a variable called `gender_direction`, indicating whether the original text was written about a male and the generated text about a female (or vice versa), were also included in the model. This was done to control for any differences in the content typically written about men and women. The mixed-effects model was specified as follows:

$$\begin{aligned} \text{sentiment}_{ij} = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{gender\_direction}_i \\ + u_{0j} + \epsilon_{ij} \end{aligned} \tag{0.4}$$

Where:

- *sentiment* is a continuous indicator of the proportion of the text which contains non-negative sentiment
- *gender* is a binary indicator of whether a text is about a man or a woman.
- *gender_direction* is a binary indicator of whether the original text was written about a male and the generated text about a female, or vice versa.
- u_{0j} is a random intercept for the j-th group (Sentence ID), accounting for the variability in sentiment across different sentences.
- ϵ_{ij}: Residual error term for the j-th observation within the j-th group.

The covariance of the random intercept was allowed to be unstructured. It was assumed that the random intercepts u_{0j} follow a normal distribution with mean 0 and variance σ^2_{u0}, the residuals ϵ_{ij} are independently and normally distributed with mean 0 and variance σ^2, and the random intercepts u_{0j} are independent of the residuals ϵ_{ij}.

Since the final activation layer of SiEBERT is softmax, producing binary predictions of sentiment (i.e., positive or negative), a generalised linear model with a logistic link function was used for the sentence-level SiEBERT predictions. In this case, $\text{logit}(P(sentiment = 1))$ was estimated, where sentiment can take the values 0 (negative) or 1 (positive). The right-hand side of the equation remained the same.

The results of the mixed model, as specified in Equation (0.4), are shown in Table 11. These results are consistent with the t-test findings, indicating that Regard and SiEBERT do not find systematic differences in the sentiment of the original documents based on gender, but the DistilBERT-based model does.

Mixed effects model: document level

It is reassuring that the mixed model results at sentence level are consistent with the t-test results. However, summaries do not necessarily have the same number of sentences (and if they do the sentences may not correspond). This means that sentiment for the male and female versions of each summary will need to be aggregated and compared at document level. To confirm that the metrics are appropriate, the sentiment results were aggregated for the original texts at document level, taking the mean of sentence-level sentiment. This is the same model as Equation (0.4), though clustering at Document ID rather than Sentence ID level, i.e.

Table 11: Sentiment output: mixed model (sentence level)

Coef	Estimate	Std. Error	t value	Pr(>|t|)	Signif
regard					
(Intercept)	0.320000	0.003650	87.700	<0.001	***
Gender: Male	0.000561	0.000435	1.290	0.197	
Gender direction: mf	0.003790	0.005660	0.669	0.504	
siebert					
(Intercept)	0.400000	0.009360	42.700	3.53e-187	***
Gender: Male	0.000569	0.001280	0.443	0.658	
Gender direction: mf	-0.018000	0.014500	-1.240	0.214	
distilbert					
(Intercept)	0.665000	0.003450	193.000	<0.001	***
Gender: Male	-0.007110	0.000241	-29.500	3.12e-120	***
Gender direction: mf	0.004730	0.005350	0.883	0.378	

Note:
The SiEBERT binomial produces a z-value rather than t-value. For the purpose of presentation, this is included in the t-value column.

$$\begin{aligned} \text{sentiment}_{ij} = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{gender\_direction}_i \\ + u_{0j} + \epsilon_{ij} \end{aligned} \tag{0.5}$$

Where:

- *sentiment* is a continuous indicator of the proportion of the text which contains non-negative sentiment (mean of each sentence across documents)
- *gender* is a binary indicator of whether a text is about a man or a woman.
- *gender_direction* is a binary indicator of whether the original text was written about a male and the generated text about a female, or vice versa.
- u_{0j} is a random intercept for the j-th group (Document ID), accounting for the variability in sentiment across different sentences.
- ϵ_{ij}: Residual error term for the j-th observation within the j-th group.

Once again, the assumptions are the same. The covariance of the random intercept is unstructured. The model assumes that the random intercepts u_{0j} follow a normal distribution with mean 0 and variance σ^2_{u0}, the residuals ϵ_{ij} are independently and normally distributed with mean 0 and variance σ^2 and the random intercepts u_{0j} are independent of the residuals ϵ_{ij}. A linear model is used for SiEBERT here

Table 12: Sentiment output: mixed model (document level)

Coef	Estimate	Std. Error	t value	Pr(>|t|)	Signif
regard					
(Intercept)	0.320000	0.003650	87.700	<0.001	***
Gender: Male	0.000561	0.000435	1.290	0.197	
Gender direction: mf	0.003790	0.005660	0.669	0.504	
siebert					
(Intercept)	0.400000	0.009360	42.700	3.53e-187	***
Gender: Male	0.000569	0.001280	0.443	0.658	
Gender direction: mf	-0.018000	0.014500	-1.240	0.214	
distilbert					
(Intercept)	0.665000	0.003450	193.000	<0.001	***
Gender: Male	-0.007110	0.000241	-29.500	3.12e-120	***
Gender direction: mf	0.004730	0.005350	0.883	0.378	

too, as the per-document average of binary sentence classifications is continuous. Table 12 shows the results aggregated at document level.

Across all three measures, the DistilBERT-based model finds significant differences in sentiment once gender is changed. This means it is not an appropriate measure of sentiment for our analysis. This is why it is not used in the paper to measure differences in sentiment of the summaries. However, there are no significant differences using Regard or SiEBERT, which is why these metrics are used to evaluate the output of the summarisation models.

Appendix 2 - Model diagnostics and robustness checks

Table 13 contains the covariance matrix for the random effects in the model specified in Equation (0.2), with the results for the main effects in Table 2. Table 13 includes the variances of individual variables and the covariances between variables.

Table 13: Covariance Matrix of Random Effects

		Regard		SiEBERT	
Group	Variable	Variance-Covariance	Standard Deviation	Variance-Covariance	Standard Deviation
Residual		0.006	0.078	0.035	0.187
doc_id	(Intercept)	0.011	0.103	0.074	0.272
doc_id	(Intercept) - modelgemma	-0.008	-0.835	-0.050	-0.809
doc_id	(Intercept) - modelllama3	-0.007	-0.791	-0.044	-0.704
doc_id	(Intercept) - modelt5	-0.007	-0.678	-0.042	-0.660
doc_id	modelgemma	0.008	0.090	0.051	0.226
doc_id	modelgemma - modelllama3	0.007	0.895	0.046	0.888
doc_id	modelgemma - modelt5	0.006	0.691	0.038	0.723
doc_id	modelllama3	0.008	0.090	0.053	0.229
doc_id	modelllama3 - modelt5	0.006	0.685	0.038	0.701
doc_id	modelt5	0.009	0.097	0.055	0.234

The distribution of the linear mixed model's random effects is presented in Figure 1, and a Q-Q plot of observed and expected values for residuals is shown in Figure 2. The random effects are generally normally distributed, with the notable exception of the intercept for the SiEBERT model, which demonstrates clear asymmetry at the tails. The Q-Q plot reveals the presence of some outliers and heteroscedasticity, particularly with the SiEBERT predictions, which deviate more from normality at the tails. The Regard predictions fit more closely to the normal distribution, although the residuals do not perfectly follow the expected distribution at the tails. Despite these deviations, the bootstrapping results and robustness checks ensure the conclusions remain reliable.

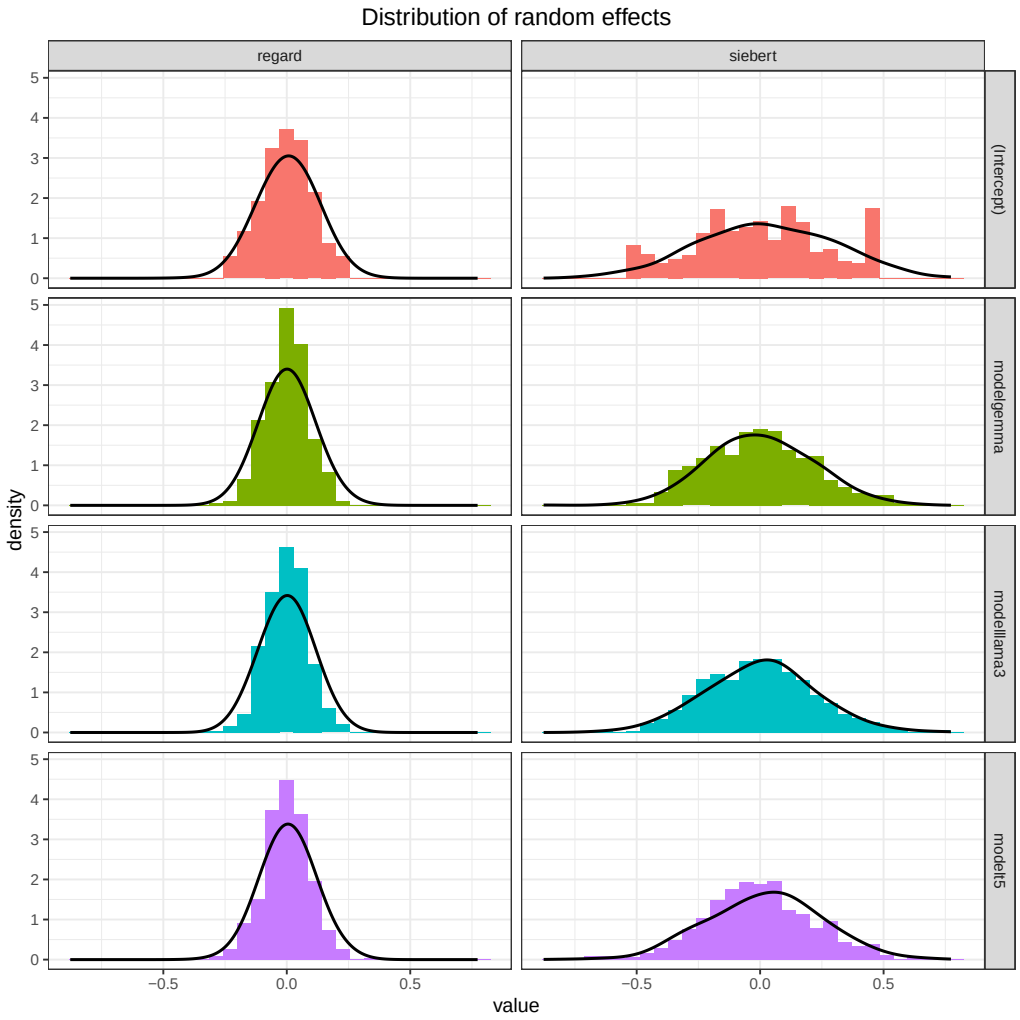

Figure 1: Distribution of random effects

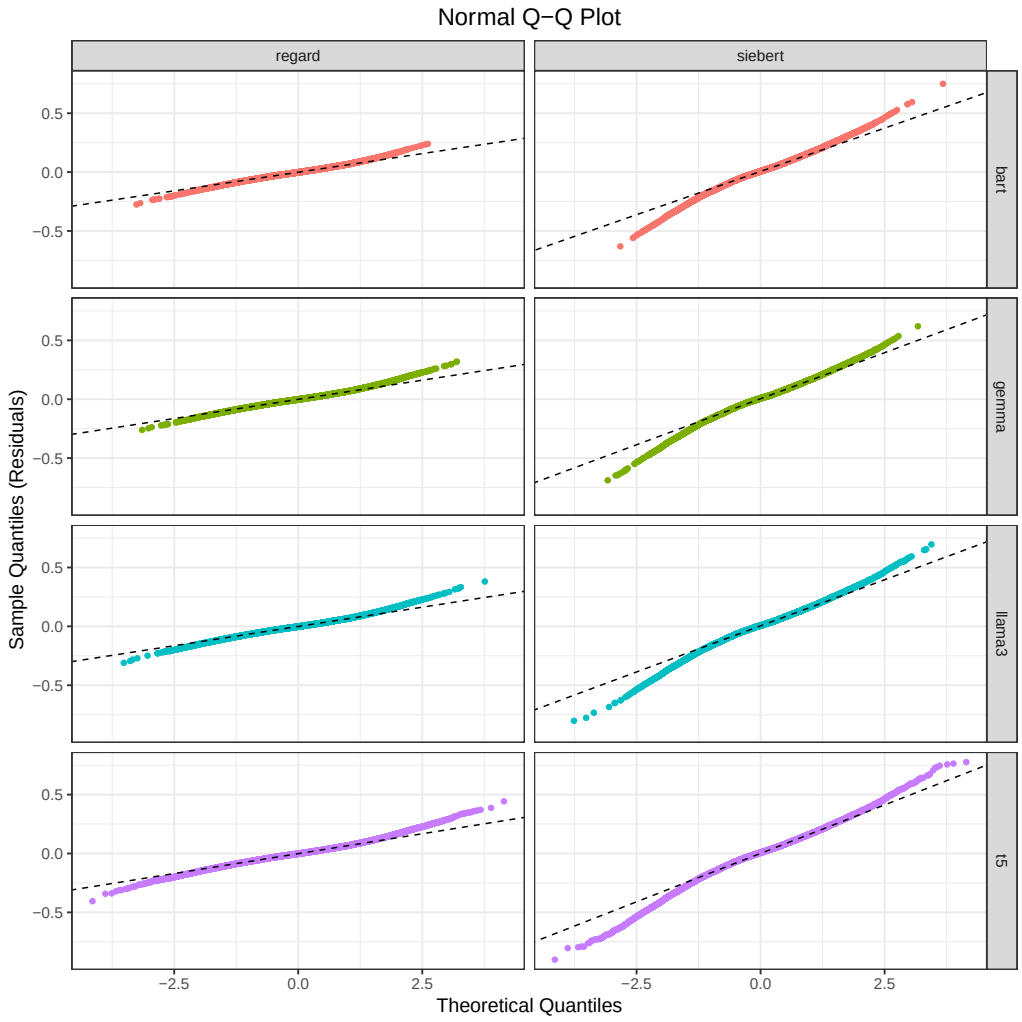

Figure 2: Normal QQ plot

The linear mixed model assumes normality of random effects and homoscedasticity. Simulations show that violations of these assumptions often have little or no effect on parameter estimates, although they do affect the interpretation of the significance of the variance of the random effects [47]. The primary focus is on the fixed effects rather than document-level random effects, which are mainly included to account for the clustering of sentiment within documents. However, as the assumptions of the model are not always satisfied, other approaches were explored to assess the sensitivity of the conclusions to these assumptions. Given the presence of some non-linearities, interaction terms, such as the interaction between gender and the maximum number of tokens, were tested to account for possible non-linear relationships. However, analysis of variance (ANOVA) and

likelihood-ratio tests indicated that the interaction term did not significantly improve model fit ($p > 0.05$). Consequently, the interaction was removed to simplify the model without affecting the overall conclusions regarding gender bias in the summaries. The model equation was retained without interactions, and other methods were used to assess the robustness of the findings:

1. **Bootstrapping:** To test sensitivity to outliers, 1,000 bootstrap samples were generated by resampling the original data with replacement, and the model was refitted on each sample. This approach provided an estimate of the distribution of the parameter estimates and allowed an assessment of the stability of the findings across different datasets.
2. **Relaxing the variance structure:** To test sensitivity to the assumption of homoscedasticity, a mixed-effects model was fitted, allowing for different residual variances across each language model.
3. **Robust linear mixed model:** To test sensitivity to outliers, a robust linear mixed model was used.
4. **Generalised Estimating Equations (GEE) model:** To test sensitivity to the correlation structure of the data and the assumption of normally distributed random effects, a GEE model was fitted. This model used robust sandwich estimators to adjust standard errors, allowing for heteroscedasticity in the residuals.
5. **Linear models:** To test sensitivity to the inclusion of random effects at the model level, each language model's results were split into its own dataset, and a linear model was run with Document ID as a main effect.

The results of each of these models were generally consistent with the findings of the mixed model. None of the models identified gender-based differences caused by Llama 3. Some variation was observed across the models regarding the significance of the differences in sentiment for the BART and T5 models. However, all models agreed that there were significant gender-based differences in the summaries generated by the Gemma model.

Bootstrapped model output and estimated marginal means

Bootstrapped datasets were generated by creating 1,000 new datasets, each the same size as the original data, through non-parametric sampling of the original data with replacement. Samples were taken at the Document ID level to preserve the correlation of sentiment within documents [48]. The original linear mixed

model was then run for each bootstrapped dataset. The bootstrapped estimates represent the mean of all 1,000 estimates. The results for the SiEBERT model are shown in Table 14, and the results for Regard are shown in Table 15. The additional columns in Table 14 were calculated as follows:

$$\text{Absolute Bias} = \text{Bootstrapped Estimate} - \text{Original Estimate}$$

$$\text{Relative Bias} = \frac{\text{Absolute Bias}}{\text{Original Estimate}}$$

$$\text{Standardised Bias} = \frac{\text{Absolute Bias}}{\text{Standard Error}}$$

Bootstrapped estimated marginal means are presented in Table 16. The table also includes the number of times the p-values for the estimated marginal means were less than 0.05 and 0.01. The differences in gender in the Gemma model are larger using SiEBERT, with a larger t-value and a p-value of less than 0.01 in all 1,000 bootstrapped datasets. The difference is somewhat smaller in the case of Regard, though $p < 0.05$ in 962 of the 1,000 simulated datasets. The BART models show this effect in approximately 30-40% of cases, and T5 shows it in 40-60% of cases, suggesting that there is an effect of gender bias greater than random chance, although not as strong as the disparities observed in the Gemma model. There is no indication of a systematic effect of gender on sentiment in Llama 3, with slightly under 5% of estimated marginal mean differences resulting in $p < 0.05$. Overall, the bootstrapping results confirm that while some observable gender-based differences exist in BART and T5, the largest differences are in the Gemma model.

Table 14: Bootstrapped model output (SiEBERT)

	Original model				Bootstrapped model			
						Bias		
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Absolute	Relative	Standardised
(Intercept)	0.579	0.012	49.749	<0.001	0.579	<0.001	<0.001	-0.008
modelgemma	0.147	0.010	14.558	<0.001	0.147	<0.001	0.001	0.015
modelllama3	0.052	0.010	5.132	<0.001	0.052	<0.001	0.001	0.004
modelt5	0.103	0.010	9.904	<0.001	0.103	<0.001	<0.001	0.002
gendermale	-0.009	0.004	-2.161	0.031	-0.009	<0.001	0.001	-0.001
max_tokens75	-0.024	0.004	-6.431	<0.001	-0.024	<0.001	-0.001	0.009
max_tokens100	-0.039	0.004	-10.304	<0.001	-0.039	<0.001	-0.002	0.021
max_tokens150	-0.050	0.004	-13.299	<0.001	-0.050	<0.001	-0.001	0.008
max_tokens300	-0.054	0.004	-14.419	<0.001	-0.054	<0.001	-0.002	0.026
max_tokensNone	-0.084	0.004	-22.262	<0.001	-0.084	<0.001	-0.001	0.027
modelgemma:gendermale	-0.033	0.006	-5.318	<0.001	-0.033	<0.001	0.006	-0.030
modelllama3:gendermale	0.015	0.006	2.426	0.015	0.015	<0.001	-0.001	-0.002
modelt5:gendermale	0.020	0.006	3.185	0.001	0.019	<0.001	-0.012	-0.038

Table 15: Bootstrapped model output (Regard)

	Original model				Bootstrapped model			
						Bias		
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Absolute	Relative	Standardised
(Intercept)	0.278	0.004	61.965	<0.001	0.278	<0.001	<0.001	0.019
modelgemma	0.025	0.004	6.109	<0.001	0.025	<0.001	-0.004	-0.027
modelllama3	0.029	0.004	7.061	<0.001	0.029	<0.001	0.001	0.010
modelt5	-0.033	0.004	-7.712	<0.001	-0.033	<0.001	0.003	-0.023
gendermale	0.004	0.002	1.954	0.051	0.004	<0.001	0.009	0.018
max_tokens75	0.019	0.002	11.865	<0.001	0.019	<0.001	<0.001	-0.004
max_tokens100	0.027	0.002	17.076	<0.001	0.027	<0.001	0.001	0.020
max_tokens150	0.032	0.002	20.246	<0.001	0.032	<0.001	-0.001	-0.026
max_tokens300	0.039	0.002	25.052	<0.001	0.040	<0.001	0.001	0.022
max_tokensNone	0.045	0.002	28.303	<0.001	0.045	<0.001	<0.001	-0.001
modelgemma:gendermale	-0.011	0.003	-4.082	<0.001	-0.011	<0.001	0.003	-0.012
modelllama3:gendermale	-0.001	0.003	-0.561	0.575	-0.001	<0.001	0.038	-0.021
modelt5:gendermale	0.001	0.003	0.521	0.603	0.001	<0.001	0.033	0.017

Table 16: Mixed effects model: bootstrapped estimated marginal means (female - male)

	Regard					SiEBERT				
Model	Estimate	t	N	N Pr(|t|)<0.01	N Pr(|t|)<0.05	Estimate	t	N	N Pr(|t|)<0.01	N Pr(|t|)<0.05
bart	-0.0036	-1.60	1000	146	331	0.0094	1.80	1000	235	430
gemma	0.0069	3.00	1000	764	962	0.0420	7.70	1000	1000	1000
llama3	-0.0021	-0.91	1000	1	32	-0.0055	-0.99	1000	3	40
t5	-0.0050	-2.20	1000	275	651	-0.0099	-1.80	1000	107	421

Variance-structured mixed effects model

The Q-Q plots demonstrated deviations from normality, especially in the tails, which differ by model. To account for this heteroscedasticity and deviation from normality, the R `nlme` package [49] was used to employ a linear mixed-effects model which allowed the variance to differ by model, i.e.

$$\text{Var}(\epsilon_{ij}) = \sigma^2_{\text{model}_i} \tag{0.6}$$

This model would not converge with a random intercept and slope and this variance specification, so the random slope was removed. The model was therefore specified as follows:

$$\begin{aligned} \text{sentiment}_{ij} = \beta_0 &+ \beta_1 \cdot \text{model}_i + \beta_2 \cdot \text{gender}_j \\ &+ \beta_3 \cdot (\text{model}_i \times \text{gender}_j) + \beta_4 \cdot \text{max\_tokens}_i \\ &+ u_{0j} + \epsilon_{ij} \end{aligned} \tag{0.7}$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for model, gender, and their interaction, β_4 is the coefficient for maximum tokens, u_{0j} is the random intercept for document j and ϵ_{ij} is the residual error term. The results are set out in Table 17 and the estimated marginal means in Table 18. The estimates are very close to the output from the linear mixed model, though the p-values are slightly larger. The BART and T5 models are on the boundary of significance, but now the p-values are slightly larger than 0.05. Llama 3 has no significant differences in sentiment between men and women, and Gemma has the largest standardised estimates and smallest p-values. This model reduces the risk of Type 1 error, which is seen in the larger p-values, so it is reassuring that the main findings about Llama 3 and Gemma remain consistent.

Table 17: Variance-structured mixed effects model output

	Regard					SiEBERT				
Coef	Estimate		Std. Error	t	p	Estimate		Std. Error	t	p
(Intercept)	0.3100	***	0.0030	100.5866481	0.0e+00	0.5400	***	0.0083	64.278954	0.0e+00
Model gemma	0.0250	***	0.0019	13.0451295	0.0e+00	0.1500	***	0.0048	30.836185	0.0e+00
Model llama3	0.0290	***	0.0019	14.7169102	0.0e+00	0.0520	***	0.0049	10.750380	0.0e+00
Model t5	-0.0330	***	0.0024	-13.7578664	0.0e+00	0.1000	***	0.0059	17.502350	0.0e+00
gendermale	0.0036	.	0.0020	1.7997477	7.2e-02	-0.0094	.	0.0052	-1.809674	7.0e-02
Max tokens 150	0.0046	**	0.0018	2.5900877	9.6e-03	-0.0069		0.0043	-1.603439	1.1e-01
Max tokens 300	0.0120	***	0.0018	6.6848292	0.0e+00	-0.0110	**	0.0043	-2.682003	7.3e-03
Max tokens 50	-0.0270	***	0.0018	-15.2606004	0.0e+00	0.0360	***	0.0043	8.430521	0.0e+00
Max tokens 75	-0.0083	***	0.0018	-4.6806345	2.9e-06	0.0130	**	0.0043	3.089684	2.0e-03
Max tokens None	0.0150	***	0.0018	8.4733104	0.0e+00	-0.0290	***	0.0043	-6.866574	0.0e+00
Model gemma : Male	-0.0110	***	0.0027	-3.9068436	9.4e-05	-0.0330	***	0.0067	-4.851052	1.2e-06
Model llama3 : Male	-0.0014		0.0028	-0.5237945	6.0e-01	0.0150	*	0.0069	2.159936	3.1e-02
Model t5 : Male	0.0013		0.0034	0.3931827	6.9e-01	0.0200	*	0.0083	2.351745	1.9e-02

Table 18: Variance-structured mixed effects: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	.	-1.8	0.07200	0.0094	.	1.8	0.07
gemma	0.0069	***	3.8	0.00014	0.0420	***	9.8	0.00
llama3	-0.0021		-1.1	0.27000	-0.0055		-1.2	0.23
t5	-0.0049	.	-1.8	0.07800	-0.0100		-1.6	0.12

Robust linear mixed model

The results of the bootstrapping were reassuring in the case of the Gemma model. However, significant differences were not always observed in the BART and T5 models. From the Q-Q plots, it is known that deviations from normality exist in the tails. To test the sensitivity of the results to outliers, a robust linear mixed model was used. This model follows the same structure as the standard linear mixed model, given in Equation (1) in the main body of the paper:

$$\begin{aligned} \text{sentiment}_{ij} = \beta_0 &+ \beta_1 \cdot \text{model}_i + \beta_2 \cdot \text{gender}_j \\ &+ \beta_3 \cdot (\text{model}_i \times \text{gender}_j) + \beta_4 \cdot \text{max\_tokens}_i \\ &+ u_{0j} + u_{1j} \cdot \text{model}_i + \epsilon_{ij} \end{aligned} \tag{0.8}$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for model, gender, and their interaction, β_4 is the coefficient for maximum tokens, u_{0j} is the random intercept for document j, u_{1j} is the random slope for model within document j, and ϵ_{ij} is the residual error term.

The difference from the standard mixed effects model is that a robust loss function was incorporated to reduce the impact of outliers in the residuals. This was implemented using the `robustlmm` R package [50]. The results are shown in Table 19. The estimates obtained from both the mixed-effects and robust mixed-effects models were extremely close. The package does not produce p-values; however, marginal means were estimated [45] and are presented in Table 20. The estimates and p-values are very close to the output from the standard linear mixed model. Once again, The BART and T5 models show p-values hovering around conventional significance thresholds, with some disagreement in the direction of the gender effect in the BART model between Regard and SiEBERT. For these models, p-values range between 0.04 and 0.08, suggesting borderline statistical significance that should be interpreted cautiously. The Gemma model exhibits the largest standardised estimates and the smallest p-values, with both sentiment metrics indicating that male sentiment is more negative than female sentiment.

Table 19: Robust mixed effects model output

	Regard			SiEBERT		
Coef	Estimate	Std. Error	t	Estimate	Std. Error	t
(Intercept)	0.27000	0.0065	43.00	0.5900	0.0120	48.0
Model gemma	0.02300	0.0037	6.20	0.1400	0.0100	14.0
Model llama3	0.02800	0.0035	7.90	0.0510	0.0100	5.0
Model t5	-0.03500	0.0029	-12.00	0.1200	0.0110	11.0
gendermale	0.00410	0.0020	2.00	-0.0094	0.0039	-2.4
Max tokens 75	0.02100	0.0018	12.00	-0.0270	0.0034	-7.8
Max tokens 100	0.02900	0.0018	17.00	-0.0420	0.0034	-12.0
Max tokens 150	0.03500	0.0018	20.00	-0.0510	0.0034	-15.0
Max tokens 300	0.04200	0.0018	24.00	-0.0560	0.0034	-17.0
Max tokens None	0.04700	0.0018	27.00	-0.0790	0.0034	-23.0
Model gemma : Male	-0.01100	0.0029	-3.70	-0.0300	0.0055	-5.5
Model llama3 : Male	-0.00052	0.0029	-0.18	0.0130	0.0055	2.4
Model t5 : Male	0.00083	0.0029	0.29	0.0190	0.0055	3.5

Table 20: Robust mixed effects model: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		z	p	Estimate		z	p
bart	-0.0041	*	-2.0	0.0450	0.0094	*	2.40	0.016
gemma	0.0065	**	3.2	0.0014	0.0400	***	10.00	0.000
llama3	-0.0035	.	-1.7	0.0810	-0.0039		-0.99	0.320
t5	-0.0049	*	-2.4	0.0160	-0.0100	*	-2.60	0.010

Generalised Estimating Equations (GEE)

A Generalised Estimating Equations (GEE) model was also used to estimate population-averaged effects, adjusting for within-group correlation using robust sandwich estimators. This was implemented using the `geepack` R package [51]. The GEE model estimates population-averaged effects and can be more robust to misspecified correlation structures. The GEE model was specified as follows:

$$\begin{aligned} y_{ij} = \beta_0 + \beta_1 \text{model}_i + \beta_2 \text{gender}_j \\ + \beta_3 (\text{model}_i \times \text{gender}_j) + \beta_4 \text{max\_tokens}_i + \epsilon_{ij} \end{aligned} \tag{0.9}$$

The correlation structure of the residuals ϵ_i was modeled as exchangeable within groups defined by Document ID. No corrections were applied to the standard errors to reduce the risk of Type 1 error, as there are 617 document-level clusters, and with 100 or more clusters, such corrections are generally unnecessary [52]. The results of the GEE model are set out in Table 21. The estimated marginal means for the GEE model are presented in Table 22.

The point estimates obtained from the mixed-effects and GEE models were identical, indicating that the fixed effects are robust to the choice of modelling approach. However, the standard errors differed between the models. The mixed-effects model, which accounts for random effects, generally provided smaller standard errors compared to the GEE model. Attempts to fit a GEE model with an unstructured covariance matrix were unsuccessful, which may have contributed to the larger standard errors in the GEE model. As a result, significant differences in sentiment based on gender were not observed in the BART and T5 models. However, the Gemma model was not affected by these differences, and summaries about women remained significantly less negative than those about men.

Table 21: GEE model output

	Regard					SiEBERT				
Coef	Estimate		Std. Error	Wald	p	Estimate		Std. Error	Wald	p
(Intercept)	0.3000	***	0.0024	1.6e+04	0.0000	0.5800	***	0.0065	7800.0	0.0e+00
Model gemma	0.0250	***	0.0025	1.0e+02	0.0000	0.1500	***	0.0064	530.0	0.0e+00
Model llama3	0.0290	***	0.0025	1.3e+02	0.0000	0.0520	***	0.0067	61.0	0.0e+00
Model t5	-0.0330	***	0.0029	1.3e+02	0.0000	0.1000	***	0.0073	200.0	0.0e+00
gendermale	0.0036		0.0027	1.7e+00	0.1900	-0.0094		0.0072	1.7	1.9e-01
Max tokens 150	0.0050	*	0.0021	5.5e+00	0.0190	-0.0500	***	0.0059	71.0	0.0e+00
Max tokens 300	0.0130	***	0.0021	3.6e+01	0.0000	-0.0540	***	0.0059	85.0	0.0e+00
Max tokens 75	-0.0082	***	0.0023	1.3e+01	0.0003	-0.0240	***	0.0061	16.0	7.8e-05
Max tokens None	0.0180	***	0.0021	7.1e+01	0.0000	-0.0840	***	0.0060	200.0	0.0e+00
Model gemma : Male	-0.0110	**	0.0035	9.2e+00	0.0024	-0.0330	***	0.0090	13.0	2.8e-04
Model llama3 : Male	-0.0014		0.0036	1.6e-01	0.6900	0.0150		0.0095	2.5	1.2e-01
Model t5 : Male	0.0013		0.0041	1.1e-01	0.7400	0.0200	.	0.0100	3.6	5.9e-02

Table 22: GEE model: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		z	p	Estimate		z	p
bart	-0.0036		-1.30	0.1900	0.0094		1.3	0.19
gemma	0.0069	**	3.30	0.0011	0.0420	***	7.8	0.00
llama3	-0.0021		-0.92	0.3600	-0.0055		-0.9	0.37
t5	-0.0049		-1.60	0.1100	-0.0100		-1.4	0.17

Linear models

The mixed model includes an interaction term as well as both random intercepts and random slopes to account for variability between documents and within models. This specification is important because it reflects how document-level differences (random intercepts) and model-specific variability within documents (random slopes) can impact sentiment estimates across gender. However, while this specification makes theoretical sense, the sensitivity of the findings to the model specification was checked by splitting the data into separate tables for each combination of model (BART, Gemma, Llama 3, and T5) and metric (Regard and SiEBERT). A simple linear model was then fitted for each of these eight datasets. The linear model can be expressed as:

$$\text{sentiment}_i = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{max\_tokens}_i + \beta_3 \cdot \text{doc\_id}_i + \epsilon_i$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for gender, maximum tokens, and document identifier, respectively, and ϵ_i is the residual error term. This model was run separately for each LLM, and the output for the Regard metric is presented in Table 23, and for SiEBERT in Table 24. The model also produced a coefficient for each Document ID, which is not of interest, so these were excluded from the tables. Similarly to the GEE model, the point estimates are close to those from the mixed-effects model, though with smaller standard errors in this case. The estimated marginal means by gender for each of the models are presented in Table 25, and they are consistent with the findings from the mixed model.

Table 23: Linear model (Regard)

Coef	Estimate		Std. Error	t	Pr(>|t|)
bart					
(Intercept)	0.2840833	***	0.0155566	18.2613205	0.0000000
gendermale	0.0035545	**	0.0012465	2.8515437	0.0043639
max_tokens150	0.0001643		0.0021590	0.0761052	0.9393376
max_tokens300	0.0001634		0.0021590	0.0756762	0.9396789
max_tokens50	-0.0307295	***	0.0021590	-14.2329516	0.0000000
max_tokens75	-0.0054155	*	0.0021590	-2.5083062	0.0121543
max_tokensNone	0.0001634		0.0021590	0.0756762	0.9396789
gemma					
(Intercept)	0.3048890	***	0.0204331	14.9213487	0.0000000
gendermale	-0.0069472	***	0.0016373	-4.2431642	0.0000223
max_tokens150	0.0009879		0.0028358	0.3483562	0.7275835
max_tokens300	0.0141746	***	0.0028358	4.9983907	0.0000006
max_tokens50	-0.0140059	***	0.0028358	-4.9388990	0.0000008
max_tokens75	-0.0069622	*	0.0028358	-2.4550780	0.0141103
max_tokensNone	0.0147827	***	0.0028358	5.2128392	0.0000002
llama3					
(Intercept)	0.3144663	***	0.0216083	14.5530579	0.0000000
gendermale	0.0021104		0.0017317	1.2187019	0.2229998
max_tokens150	0.0114336	***	0.0029989	3.8125689	0.0001387
max_tokens300	0.0167968	***	0.0029989	5.6009226	0.0000000
max_tokens50	-0.0399157	***	0.0029989	-13.3099939	0.0000000
max_tokens75	-0.0127653	***	0.0029989	-4.2566164	0.0000210
max_tokensNone	0.0185463	***	0.0030004	6.1812957	0.0000000
t5					
(Intercept)	0.2153391	***	0.0303610	7.0926341	0.0000000
gendermale	0.0048940	*	0.0024328	2.0117104	0.0442900
max_tokens150	0.0073932	.	0.0042137	1.7545669	0.0793786
max_tokens300	0.0191323	***	0.0042137	4.5405227	0.0000057
max_tokens50	-0.0229692	***	0.0042137	-5.4510934	0.0000001
max_tokens75	-0.0076979	.	0.0042137	-1.8268779	0.0677621
max_tokensNone	0.0372671	***	0.0042137	8.8443118	0.0000000

Table 24: Linear model (SiEBERT)

Coef	Estimate		Std. Error	t	Pr(>|t|)
bart					
(Intercept)	0.6601786	***	0.0412309	16.0117242	0.0000000
gendermale	-0.0093810	**	0.0033038	-2.8394946	0.0045320
max_tokens150	-0.0010080		0.0057223	-0.1761527	0.8601792
max_tokens300	-0.0008814		0.0057223	-0.1540313	0.8775896
max_tokens50	0.0324356	***	0.0057223	5.6682846	0.0000000
max_tokens75	0.0093005		0.0057223	1.6253194	0.1041410
max_tokensNone	-0.0008814		0.0057223	-0.1540313	0.8775896
gemma					
(Intercept)	0.7857799	***	0.0484271	16.2260289	0.0000000
gendermale	-0.0420346	***	0.0038804	-10.8325995	0.0000000
max_tokens150	-0.0118507	.	0.0067210	-1.7632239	0.0779078
max_tokens300	-0.0241479	***	0.0067210	-3.5928966	0.0003293
max_tokens50	0.0358635	***	0.0067210	5.3360130	0.0000001
max_tokens75	0.0115767	.	0.0067210	1.7224544	0.0850328
max_tokensNone	-0.0313662	***	0.0067210	-4.6668826	0.0000031
llama3					
(Intercept)	0.4881037	***	0.0503594	9.6924036	0.0000000
gendermale	0.0055138		0.0040358	1.3662261	0.1719133
max_tokens150	0.0129288	.	0.0069892	1.8498242	0.0643824
max_tokens300	0.0136312	.	0.0069892	1.9503233	0.0511788
max_tokens50	0.0283793	***	0.0069892	4.0604580	0.0000495
max_tokens75	0.0123671	.	0.0069892	1.7694619	0.0768618
max_tokensNone	0.0166275	*	0.0069926	2.3778786	0.0174401
t5					
(Intercept)	0.7611087	***	0.0698796	10.8917072	0.0000000
gendermale	0.0101714	.	0.0055993	1.8165436	0.0693312
max_tokens150	-0.0451223	***	0.0096983	-4.6525854	0.0000033
max_tokens300	-0.0504907	***	0.0096983	-5.2061143	0.0000002
max_tokens50	0.0582791	***	0.0096983	6.0091831	0.0000000
max_tokens75	0.0249979	**	0.0096983	2.5775492	0.0099713
max_tokensNone	-0.1642472	***	0.0096983	-16.9355987	0.0000000

Table 25: Linear models: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	**	-2.9	4.4e-03	0.0094	**	2.8	0.0045
gemma	0.0069	***	4.2	2.2e-05	0.0420	***	11.0	0.0000
llama3	-0.0021		-1.2	2.2e-01	-0.0055		-1.4	0.1700
t5	-0.0049	*	-2.0	4.4e-02	-0.0100	.	-1.8	0.0690

Conclusion of robustness checks

The robustness checks consistently indicated the reliability of the findings with regard to Llama 3 and Gemma. Across the linear mixed model, robust linear mixed models, Generalised Estimating Equations (GEE), and separate linear models, the point estimates for the fixed effects remained stable, and the direction of the effects was consistent for the Gemma model, as was the absence of an effect for Llama 3. However, the variance-structured mixed effects model and the GEE model did not find significant effects in the BART and T5 models. Similarly, the bootstrapped results indicated significant effects slightly less than half of the time. This suggests that the results for the BART and T5 models may be on the boundary of significance and should be interpreted with caution. However, as the older models were primarily included as benchmarks and are not currently being used in practice to summarise care records, their bias is of less concern for long-term care policy. The consistent results across the Llama 3 and Gemma models, particularly in terms of estimated marginal means, indicate that the conclusions regarding state-of-the-art models are not sensitive to model specification or the presence of outliers, validating the robustness of the model.

Appendix 3 Evaluation of themes word lists

The word lists for each individual theme are included below. These, along with the complete code, can also be found in the GitHub repository [44]. The Python `str.starts_with()` method [53] was used for these terms. This means that, for example, in the mental health list, the term `autis` would match words that start with these letters, such as `autism` and `autistic`, but not words containing these letters, such as `flautist`.

Mental health

```
alzheimer
anorexia
anxi
asperger
autis
behavio
bipolar
cognit
confus
deliri
delusion
dementia
depress
disorient
hallucinat
insight
mental
memory
mood
paranoi
personality disorder
power of attorney
psycho
ptsd
restlessness
schizoaffect
schizophreni
sectioned
therap
```

Physical health

activities of daily living
amputat
anaemia
angina
arthritis
aspirat
asthma
atrial fibrillation
balance
barrier cream application
bed bound
bed rails
bed-bound
bedbound
bilateral limb
bleeding
blood pressure
blood test
bowel
breath
cancer
care needs
cataract
catheter
cellulitis
chest rash
cholesterol
cirrhosis
commode
community acquired pneumonia
constipat
continen
copd
coronary
diabet
diarrhoea
disability
disable
dysphagia
dysphasia
dyspraxia
epilep
fall
fatigue

fractur
gallstone
glaucoma
gord
gout
hard of hearing
hearing and sight
hearing impair
heart attack
heart condition
heart disease
heart failure
heart problem
hemiplegia
hernia
hip replacement
hoist
house bound
house-bound
housebound
housework
hypercholesterolemia
hypertension
hypothyroid
idiopathic
immobile
incontinen
infarction
infect
influenza
injury
insulin
intravenous
ischaemic
ischemic
kidney
knee
leg clinic
leg ulcer
lung
macular
medication
melanoma
mobili
motor neuron
mrsi

myeloma
nutrition
obstructive sleep
oedema
oesophageal
osteo
pain
paralys
peg feed
personal care
physical deterioration
physical injur
pressure area
pressure relieving
pressure sore
pressure stockings
prostatic
psoriasis
pulmonary
puree
raised toilet
renal
reposition
rollator
sciatica
scoliosis
seizure
sleep apnea
slurred speech
spinal
standing tolerance
stiffness
stoma
stroke
surgery
swallowing
swollen
thickener
transfer
underweight
unsteady
urinary tract
urine retention
uti
vein
visual impairment

washing legs
weak
weight bear
weight loss
wheelchair
zimmer

Physical appearance

abdomen
appearance
appetite
bath
black eye
bmi
bruised
cloth
dental
dirty
discolouration
dishevelled
disshevelled
dress
drooling
dusty
faeces
fingernails
groom
hair
hygiene
kempt
messy
nails
naked
neglected
nude
odour
rubbish
scruffy
self neglect
self-neglect
shave
skin
slurred
smell

soil
spots
stained
teeth
tidy
tremors
trousers
unclean
underwear
underweight
unhygienic
unkempt
untidy
urin
vest
wear
weigh

Subjective language

abus
adamant
adjusted
adverse
aggress
agitat
agreeable
angry
annoy
appear
appropriate
argumentative
articulate
bad
behav
benefi
best
better
bored
bossy
breach
challeng
chatty
choose
chose

clean
clutter
coherent
concern
confine
conflict
confus
content
damage
demanding
dependent
deteriorat
difficult
dirty
dishevelled
dislike
disparaging
disruptive
distracted
distress
dusty
erratic
escalat
evasive
exacerbat
excessive
failed
feel
felt
fiercely
fixation
fluctuat
forgetful
frustrat
fuss
good
happier
happy
hard
harm
hate
high
ignor
illiterate
immens
impair

improv
impulsiv
inability
inappropriate
incoherent
increase
ineffective
insecure
insight
instrumental
insufficient
intense
invalid
involuntary
irk
irrita
isolat
issue
lack
less
likes
limited
loner
loudly
lovely
low
lucky
marked
massive
maverick
mess
mismanage
misses
misusing
mitigated
mood
more
muddle
needs
negative
neglect
nice
odd
oriented
paranoid
placid

pleasant
pleased
pointless
poor
prais
problem
proper
proud
racist
recommend
refus
relaxed
relentless
reliant
reluctan
resist
respect
restless
risk
rough
rude
sadly
safe
scared
scruffy
serious
settled
severe
shy
significant
silly
slow
small
smartly
smell
sociable
soil
strong
struggl
stupid
substantial
sufficient
suitable
suited
tearful
unable

unacceptable
unamenable
unaware
uncomfortable
uncontrollabl
uncooperative
under weight
underweight
unhygienic
unkempt
unreasonabl
unreliable
unsafe
unsatisfactory
unsettle
untidy
unwise
valid
verbal
vulnerab
wander
well
willing
wise
working
worried
worrying
worse
worst

NIHR | Policy Research Unit in Adult Social Care

NIHR Policy Research Unit in Adult Social Care
London School of Economics and Political Science
University of Kent
King's College London

ascru@lse.ac.uk
www.ascru.nihr.ac.uk
#ASCRUProject