Beyond Binary Gender Labels: Revealing Gender Biases in LLMs through Gender-Neutral Name Predictions

Zhiwen You[1*], **HaeJin Lee**[1*], **Shubhanshu Mishra**[2],
Sullam Jeoung[1], **Apratim Mishra**[1], **Jinseok Kim**[3], **Jana Diesner**[1,4]
[1] University of Illinois Urbana-Champaign [2] https://shubhanshu.com
[3] University of Michigan - Ann Arbor [4] Technical University of Munich
[1] {zhiweny2, haejin2, sjeoung2, apratim3}@illinois.edu
[2] mishra@shubhanshu.com [3] jinseokk@umich.edu [4] jana.diesner@tum.de

Abstract

Name-based gender prediction has traditionally categorized individuals as either female or male based on their names, using a binary classification system. That binary approach can be problematic in the cases of gender-neutral names that do not align with any one gender, among other reasons. Relying solely on binary gender categories without recognizing gender-neutral names can reduce the inclusiveness of gender prediction tasks. We introduce an additional gender category, i.e., "neutral", to study and address potential gender biases in Large Language Models (LLMs). We evaluate the performance of several foundational and large language models in predicting gender based on first names only. Additionally, we investigate the impact of adding birth years to enhance the accuracy of gender prediction, accounting for shifting associations between names and genders over time. Our findings indicate that most LLMs identify male and female names with high accuracy (over 80%) but struggle with gender-neutral names (under 40%), and the accuracy of gender prediction is higher for English-based first names than non-English names. The experimental results show that incorporating the birth year does not improve the overall accuracy of gender prediction, especially for names with evolving gender associations. We recommend using caution when applying LLMs for gender identification in downstream tasks, particularly when dealing with non-binary gender labels[1].

1 Introduction

Name-based gender prediction is the task of identifying the most likely gender label for a given name. This task, while not reflective of the true gender identify of the individual, is often useful for aggregate downstream analysis and as a demographic feature for predictive models. Prior work has utilized name-based gender prediction to investigate gender bias in scientific productivity, citation practices, information extraction systems, personalized marketing, content recommendation, targeted advertising, gender-based sentiment analysis, and social network analysis (Diesner and Carley, 2009; Ross et al., 2022; Jentzsch and Turan, 2022; Teich et al., 2022; Liu et al., 2023; Larivière et al., 2013; Mishra et al., 2020, 2018; VanHelene et al., 2024). Most prior work has utilized computational tools (e.g., Genderize.io[2], Namsor[3], Gender API[4], or machine learning (ML) models) or datasets (e.g., US SSN) to assign probabilities of a name (along with other features like demographics, time) likely to be a male or a female. Since name-based gender is used both as a feature in downstream systems and an indicator of demographic representation, it can lead to both measurement bias and representational bias as identified in the framework proposed by

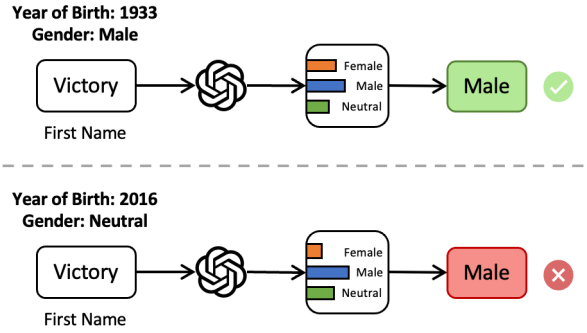


Figure 1: Example of an LLM predicting different gender labels over time for the same first name. "Victory" was labeled Male in 1933, and the LLM predicted it correctly. However, by 2016, the name had become predominantly gender-neutral, but the LLM still incorrectly predicted it as Male.

[*] Equal Contribution.
[1] Our code is available at https://github.com/zhiwenyou103/Beyond-Binary-Gender-Labels.
[2] https://genderize.io/
[3] https://namsor.app/
[4] https://gender-api.com/

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 255–268

Suresh and Guttag (2021).

A prevalent challenge in contexts utilizing inferred gender is the practice of treating gender as a binary construct, strictly categorizing names as either male or female (Chatterjee and Werner, 2021; Pilkina and Lovakov, 2022). This reliance on binary labels likely stems from historical and societal norms that often only recognize these two categories. Binary representations can reinforce existing gender biases and exclude non-binary and gender-diverse individuals, hindering their representation and understanding (Krstovski et al., 2023; Dinh et al., 2023; Mishra et al., 2018) in algorithm design and data annotation. The presence of gender-neutral names, as defined by Barry III and Harper (2014), further complicates this issue. These names, frequently assigned to both genders, contradict the binary classification system, leading to potential inaccuracies and misrepresentations in data and processes reliant on gender predictions.

This study aims to answer the following research questions to examine one of many aspects of gender biases in LLMs concerning gender prediction, especially for gender-neutral names and gender labels that change over time (Figure 1).

RQ1. How does the performance of autoregressive LLMs versus fine-tuned foundation language models compare when predicting gender categories (i.e., female, male, and neutral) given first names?

RQ2. How does adding the birth year impact gender prediction accuracy?

NOTE: In the context of this research, we are only interested in studying the likelihood of a name being identified as `Male`, `Female`, and `Neutral`. As highlighted in Yee et al. (2021), predictive models cannot be accurate about demographic attributes, and it is best to rely on individual responses to assign sensitive demographic attributes e.g. gender, however, they can be useful at the aggregate level, which is the focus of this work.

2 Related Work

In the gender prediction task, models are trained to predict or classify gender labels based on various input features, such as first or last names, country information, behavioral data, or textual content from social media activity (Liu and Ruths, 2013; Tang et al., 2011; To et al., 2020). Consequently, the accuracy of gender prediction can impact the validity of research findings and derived implication, such as policies. In other words, inaccurate gender prediction can distort results and lead to misunderstandings of gender-related biases. Moreover, the reliance on binary gender categorizations constrains the nuanced understanding of bias and the representation of individuals. Therefore, ensuring accurate and unbiased gender prediction is essential as it can impact the fairness and effectiveness of downstream applications.

Previous studies found prevalent biases in NLP-based gender prediction using gender-predicting software tools (Misa, 2022; Alexopoulos et al., 2023), which failed to appropriately capture the fact that gender exists on a non-binary scale. While most studies of bias in gender prediction relied on binary gender labels (Teich et al., 2022; Liu et al., 2023), some studies have gone beyond binary labels by introducing an additional category for names that were not strictly associated with either female or male genders (Larivière et al., 2013; Mishra et al., 2018; Pinheiro et al., 2022). For instance, Krstovski et al. (2023) categorized names that appeared as both female and male as "gender ambiguous". Additionally, most prior work on gender prediction used names as the only input feature (Jia and Zhao, 2019; Hu et al., 2021; Pham and Nguyen, 2023), while others such as Blevins and Mullen (2015) and Misa (2022) inferred the gender of first names using historical datasets with multilpe features.

Recent advances in deep learning (DL) have produced pre-trained language models like BERT (Devlin et al., 2019), CharBERT (Ma et al., 2020), and RoBERTa (Liu et al., 2019), which have been widely used for gender prediction. For example, Hu et al. (2021) found that using the user's name achieved higher gender prediction accuracy than using other features (e.g., website page views and clicks) in both ML and DL models, while Jia and Zhao (2019) and Pham and Nguyen (2023) demonstrated the effectiveness of BERT-based models for gender prediction for Japanese and Chinese names. Despite these developments, few studies focused on gender prediction using autoregressive models like ChatGPT (OpenAI, 2024a) and Llama 2 (Touvron et al., 2023). The increasing application of LLMs for gender prediction (Kotek et al., 2023; Rhue et al., 2024) underscores the need to evaluate the limitations of LLMs, particularly for gender-neutral names. For example, Michelle et al. (2023) used a prompting approach with ChatGPT to predict the gender of Olympic athletes, showing ChatGPT performed at least as well as common com-

mercial tools (i.e., Gender-API and Namsor) and often outperforms them on a binary gender scale. In this paper, we conducted experiments beyond prior approaches by introducing the gender-neutral label and using three Social Security Administration (SSA) baby name datasets to investigate gender biases by predicting non-binary gender labels.

3 Experiments

This section discussed the datasets, pre-processing, experimental design, and how we compared various models for name-based gender prediction.

3.1 Data

Dataset Pre-processing. We re-used three datasets of first names of children: one from the SSA of the US[5], one from the province of Alberta, Canada[6], and one from France[7]. Each dataset included first names, gender (female or male), and birth year. To identify and associate the gender label for each name, we counted how often each name appeared with its associated gender labels (i.e., female or male) and year of birth for a specific year. For example, if the name "Harry" appeared five times as female and 15 times as male in a specific year, we calculated the gender ratios for that year as 25% female and 75% male. Using these ratios, we labeled the first names with the associated gender labels according to the following rule-set: if a first name was at least 10% female and 10% male representation in a given year, we labeled the name as neutral. For first names with at least 85% female representation, we labeled the names as female gender label. Similarly, for the first names with at least 85% male, we labeled the names as male.

Due to the scarcity of gender-neutral names in our relabeled datasets from the 1900s, we needed to balance the number of names by gender to ensure fair comparisons in our experiments. We achieved this by sampling an equal number of female, male, and neutral names each year in the relabeled datasets. Specifically, we randomly selected 300 names per gender for each year from 1914 to 2022 from the US SSA dataset. In the Canada SSA dataset, where gender-neutral names were rare before 2000 (less than five first names per year) but increased in recent years (after 2010), we sampled 273 names per gender for each year from 2013 to 2020. Similarly, the France SSA dataset had few gender-neutral names in the early 1900s. Therefore, we selected 32 names per gender for each year from 1908 to 2022. Additional details on the dataset statistics can be found in Appendix A. We used these balanced datasets for all the experiments in Table 2.

First Names	Gender 1 (year)	Gender 2 (year)	Gender 3 (year)
Arlie	Male (1971)	Neutral (1980)	-
Hasani	Neutral (1983)	Male (2000)	-
Neer	Male (2014)	Neutral (2018)	-
CARMEL	Neutral (1920)	Male (1951)	-
FIDELE	Neutral (1918)	Female (1945)	-
Morley	Female (2013)	Neutral (2015)	Female (2017)
Victory	Male (1933)	Female (2000)	Neutral (2016)
Carmin	Male (1924)	Neutral (1958)	Female (2021)

Table 1: Examples of first names that were labeled as different genders over the years.

Dynamic gender label datasets. We observed that each balanced SSA dataset included first names labeled with different genders over the years, as shown in Table 1. For example, `Victory` was recorded as a male name in 1933, a female name in 2000, and as a gender-neutral name in 2016 (Figure 1). To further analyze the gender prediction performance of LLMs on first names with varying gender labels over time, we created a dynamic gender label dataset for each country. We selected first names with dynamic gender labels (i.e. names for which the gender association changes over time) from the test set of each balanced SSA dataset. The dynamic gender label datasets were used in the experiments of Table 3. The distribution of these dynamic gender labels is detailed in Appendix A.

3.2 Gender Prediction Models

We compared several pre-trained foundation language models with a classification head to predict the gender of first names as a multi-class classification task. Additionally, we conducted LLM-based 0-shot and 5-shot experiments to evaluate the performance of LLMs as gender classifiers.

Foundation Language Models. We fine-tuned three widely used foundation language models, i.e., BERT, RoBERTa, and CharBERT, as baselines for name-based gender prediction under the same experimental settings to conduct gender prediction. Model tuning hyper-parameters are detailed in Appendix B.

Large Language Models. We aimed to identify the potential gender bias of LLMs in predicting gender labels given first names (plus birth year).

[5] https://www.ssa.gov/oact/babynames/limits.html

[6] https://ouvert.canada.ca/data/dataset

[7] https://www.insee.fr/fr/statistiques/7633685?sommaire=7635552

		First Name				First Name + Year				
Datasets	Models	Male	Female	Neutral	Acc.	Male	Female	Neutral	Acc.	Avg.
US SSA	BERT	84.46	**89.30**	**90.55**	**88.10**	**86.64**	**90.98**	91.13	**89.58**	**88.84**
	RoBERTa	83.76	87.80	90.00	87.19	85.05	88.53	90.95	88.18	87.69
	CharRoBERTa	**84.62**	88.81	88.99	87.47	83.55	88.59	**91.96**	88.03	87.75
	GPT-3.5	91.62	**96.70**	15.99	68.10	94.68	**96.30**	14.37	**68.45**	68.28
	Llama 2	1.93	6.42	**99.66**	36.00	16.48	36.97	**90.37**	47.94	41.97
	Llama 3	**94.80**	94.83	13.03	67.55	95.29	95.26	6.09	65.55	66.55
	Mixtral-8x7B	64.62	85.81	53.30	67.91	61.38	78.44	56.42	65.41	66.66
	Claude 3 Haiku	91.50	93.67	30.00	**71.72**	**96.30**	93.46	6.97	65.58	**68.65**
Canada SSA	BERT	70.98	73.21	**82.14**	**75.45**	**74.11**	74.55	74.11	**75.15**	**75.30**
	RoBERTa	**72.77**	75.00	73.66	73.81	67.86	75.00	**76.34**	73.07	73.44
	CharRoBERTa	71.43	**76.34**	71.88	73.21	69.20	**76.34**	74.11	73.21	73.21
	GPT-3.5	82.14	**86.61**	27.68	65.48	**83.93**	83.93	28.12	65.33	65.41
	Llama 2	1.79	11.16	**100.00**	37.65	0.45	9.82	**100.00**	36.76	37.21
	Llama 3	**87.05**	84.38	21.43	64.29	76.79	86.16	28.57	63.84	64.07
	Mixtral-8x7B	50.45	69.64	68.30	62.80	35.27	46.43	90.62	57.44	60.12
	Claude 3 Haiku	78.12	80.80	57.59	**72.17**	77.68	**86.16**	32.59	**65.48**	**68.83**
France SSA	BERT	82.17	**84.57**	**93.04**	86.59	82.39	84.78	92.61	86.59	86.59
	RoBERTa	**85.22**	84.13	90.87	**86.74**	81.52	**86.09**	**93.04**	86.88	**86.81**
	CharRoBERTa	84.35	80.43	91.30	85.36	**83.04**	83.04	91.96	86.01	85.69
	GPT-3.5	89.35	**95.65**	8.91	64.64	92.61	**96.74**	8.26	**65.87**	65.26
	Llama 2	1.96	15.22	**91.52**	36.23	32.39	55.43	**71.96**	53.26	44.75
	Llama 3	**91.52**	94.57	7.17	64.42	92.39	95.87	6.52	64.93	64.68
	Mixtral-8x7B	71.96	88.70	38.04	**66.23**	68.26	83.26	39.35	63.62	64.93
	Claude 3 Haiku	89.13	93.91	13.70	65.58	**96.75**	94.78	4.57	65.36	**65.47**

Table 2: Experimental results for applying foundation language models and LLMs to the test sets of three balanced SSA datasets. We assessed gender prediction performance by calculating an accuracy score for each gender. Acc. represents the overall accuracy across genders. BERT, RoBERTa, and CharRoBERTa were fine-tuned using the training set of each SSA dataset. In contrast, we applied 0-shot prompting to evaluate other LLMs using the test sets.

We used five widely used LLMs for experimentation: GPT-3.5[8] (OpenAI, 2024b), Llama 2[9] (Touvron et al., 2023), Llama 3[10] (AI@Meta, 2024), Mixtral-8x7B[11] (Jiang et al., 2024), and Claude 3 Haiku[12] (Anthropic, 2024). For more information about these models and the settings we used see Appendix B and Appendix C, respectively.

3.3 Results

RQ1: How does the performance of LLMs versus fine-tuned foundation language models compare in first-name gender prediction? Fine-tuned foundational language models predicted gender-neutral first names more accurately than LLMs under 0-shot prompting across all three datasets. As shown in Table 2, out of all models, BERT results in the highest average accuracy for the US and Canada dataset, while RoBERTa outperformed BERT on the France dataset. Claude 3 Haiku achieved the highest accuracy among the LLMs with 0-shot prompting on all three datasets. The Llama 2 model did best on identifying gender-neutral names (100% accuracy for Canada SSA, 99.66% for US SSA, and 91.52% for France SSA when using only first names as input). Llama 3 demonstrated a more balanced distribution of prediction performance across different gender categories, similar to other LLMs such as GPT-3.5, Mixtral-8x7B, and Claude 3 Haiku. However, most LLMs failed to predict gender-neutral first names in the France SSA dataset compared to the English-based datasets, with accuracies of 7.17% for Llama 3, 8.91% for GPT-3.5, and 13.7% for Claude 3 Haiku. To assess the performance of gender prediction in dynamic gender label datasets (see Table 3), we evaluated LLMs in 0-shot and 5-shot settings, using only first names as input. Most LLMs showed higher accuracy in gender prediction when provided with 5 labeled name-gender pairs through in-context learning compared to the 0-shot setting

[8] https://platform.openai.com/docs/models/gpt-3-5-turbo

[9] https://llama.meta.com/llama2/

[10] https://llama.meta.com/llama3/

[11] https://mistral.ai/news/mixtral-of-experts/

[12] https://www.anthropic.com/news/claude-3-haiku

		First Name				First Name + Year			
Datasets	Models	Male	Female	Neutral	Acc.	Male	Female	Neutral	Acc.
US SSA	GPT-3.5 (0-shot)	86.30	92.39	31.80	55.61	95.21	**93.66**	3.92	41.94
	Llama 2 (0-shot)	14.94	33.60	**94.23**	**63.94**	47.80	62.12	**66.70**	**61.04**
	Llama 3 (0-shot)	**92.53**	**93.19**	11.89	45.80	96.26	93.50	2.02	41.09
	Mixtral-8x7B (0-shot)	80.84	91.28	32.06	54.15	70.59	92.23	32.49	51.88
	Claude 3 Haiku (0-shot)	88.89	91.60	25.85	52.70	**96.74**	90.97	10.60	45.80
	GPT-3.5 (5-shot)	84.96	91.92	43.64	62.06	65.33	67.35	4.05	30.06
	Llama 2 (5-shot)	24.71	50.40	**86.17**	**64.46**	36.88	64.98	**68.94**	**59.93**
	Llama 3 (5-shot)	**92.82**	94.45	13.96	47.27	**93.77**	**95.72**	11.33	46.20
	Mixtral-8x7B (5-shot)	79.79	**95.09**	16.76	45.60	74.81	90.65	39.55	56.83
	Claude 3 Haiku (5-shot)	87.45	84.63	39.34	59.06	91.38	88.75	32.36	56.68
Canada SSA	GPT-3.5 (0-shot)	86.36	78.07	49.08	54.74	**97.27**	78.95	19.00	30.81
	Llama 2 (0-shot)	21.82	28.07	**98.62**	**86.01**	4.55	8.77	**99.82**	**83.87**
	Llama 3 (0-shot)	**92.73**	78.07	22.32	33.10	87.27	**84.21**	13.93	26.22
	Mixtral-8x7B (0-shot)	67.27	**78.95**	46.31	50.92	50.00	79.82	60.70	61.47
	Claude 3 Haiku (0-shot)	88.18	**78.95**	41.88	49.01	89.09	77.19	43.36	50.15
	GPT-3.5 (5-shot)	84.55	74.56	56.00	60.02	**97.27**	80.70	18.82	30.81
	Llama 2 (5-shot)	22.73	24.56	**97.42**	**84.79**	32.73	23.68	**87.27**	**77.14**
	Llama 3 (5-shot)	**91.82**	**79.82**	36.62	45.03	82.73	**85.96**	32.01	40.98
	Mixtral-8x7B (5-shot)	68.18	77.19	49.17	53.21	68.18	74.56	58.30	60.55
	Claude 3 Haiku (5-shot)	83.64	64.91	55.26	58.49	90.91	60.53	41.97	47.71
France SSA	GPT-3.5 (0-shot)	78.43	**98.31**	16.52	34.30	**90.20**	**98.31**	3.54	25.84
	Llama 2 (0-shot)	3.92	35.59	**89.38**	**72.61**	27.45	79.66	**74.93**	**70.16**
	Llama 3 (0-shot)	74.51	**98.31**	4.13	24.50	90.20	**98.31**	0.00	23.16
	Mixtral-8x7B (0-shot)	**82.35**	94.92	14.75	32.96	88.24	94.92	28.91	44.32
	Claude 3 Haiku (0-shot)	78.43	94.92	10.62	29.40	88.24	94.92	6.78	27.62
	GPT-3.5 (5-shot)	78.43	98.31	20.35	37.19	**98.04**	**100.00**	5.01	28.06
	Llama 2 (5-shot)	3.92	33.90	**88.20**	**71.49**	13.73	47.46	**91.15**	**76.61**
	Llama 3 (5-shot)	82.35	98.31	9.44	29.40	90.20	**100.00**	5.01	27.17
	Mixtral-8x7B (5-shot)	**88.24**	**100.00**	13.57	33.41	88.24	94.92	28.91	44.32
	Claude 3 Haiku (5-shot)	74.51	86.44	41.00	50.78	94.12	96.61	26.25	43.21

Table 3: Gender prediction results of LLMs using dynamic gender label datasets under 0- and 5-shot settings. We report the gender prediction performance using accuracy for each gender. Acc. denotes the overall accuracy across genders. Appendix D and E provide the prompt templates and prompt robustness evaluation for LLMs.

across all datasets.

RQ2: How does adding the birth year impact gender prediction accuracy? The effectiveness of the input variation (i.e., first name + birth year) varied among different language models. Incorporating birth years as an additional input feature improved the prediction accuracy of foundational language models compared to the first-name-only setting (Table 2). However, most LLMs showed a decline in accuracy when birth years were added, particularly in predicting gender-neutral names. Despite this trend, Mixtral-8x7B consistently improved its prediction accuracy for gender-neutral names across all three datasets by adding birth year information. Similarly, the overall accuracy of Llama 2 increased, with improvements of 12% and 17% in the US and France SSA datasets, respectively.

Additionally, including birth years decreased the accuracy of predicting gender-neutral names in both 0- and 5-shot settings across all datasets (Table 3), except for the Mixtral-8x7B model, which increased the gender prediction accuracy by adding birth years. The accuracy of GPT-3.5 and Llama 3 in predicting gender-neutral names dropped when adding the birth year among all three datasets.

We observed varying trends in prediction accuracy over time across 5 LLMs (Figure 2). The accuracy of gender prediction using the US SSA dynamic gender label dataset has increased in recent years for most LLMs, including Llama3, Mixtral-8x7B, Claude 3 Haiku, and GPT-3.5. In particular, GPT-3.5 performed better without than with birth years, suggesting that incorporating recent birth year information in the US SSA dataset did not enhance predictive accuracy. The over-time results in Figure 2 indicated that most LLMs were better at predicting the genders of more recent first names. The over-time comparison of the other two datasets was provided in Appendix F.

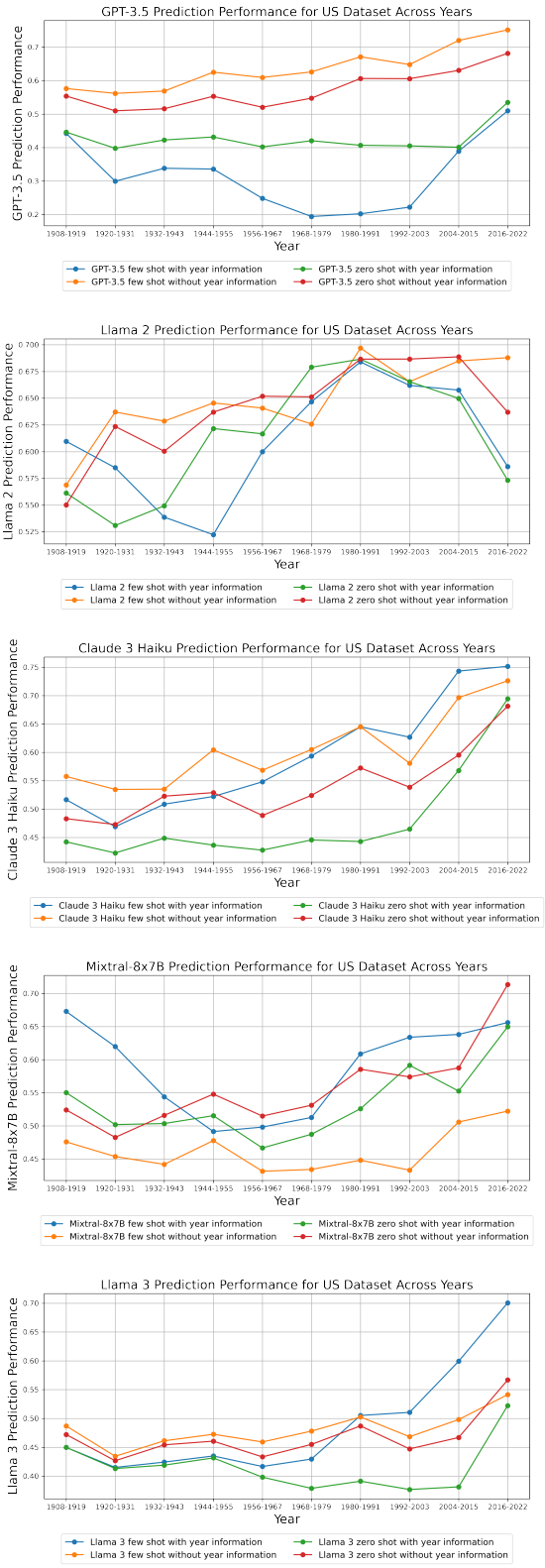


Figure 2: Temporal-level comparison of 5 LLMs using the US SSA dynamic gender label dataset given the results of Table 3. We report the overall accuracy of gender prediction for each year.

4 Discussion

LLMs are poor at accurately predicting gender. Gender bias occurs in LLMs when performing name-based gender predictions, which shows varying performance in predicting non-binary gender labels. Llama 2 categorizes nearly all names as neutral genders, with first names only as input. This tendency may result from Llama 2's training approach, which used reward modeling to promote more inclusive responses, where initial model outputs are adjusted based on human feedback to maximize inclusivenes and factual accuracy (Touvron et al., 2023). The rewarding process allows the model to better align with modern datasets' nuanced and inclusive expectations.

Including temporal information mostly degrades accuracy. When providing dynamic gender label datasets with birth year information, the gender-prediction performance of most LLMs decreased, especially for gender-neutral names. However, Mixtral-8x7B showed an increase in overall accuracy when birth years were added in 0- and 5-shot settings. We hypothesize that Mixtral-8x7B can better use temporal data as a reference for gender prediction because it is trained with more numerical information. Although Llama 2 outperformed other LLMs in predicting gender-neutral names, it exhibited biased prediction results, often classifying most names as gender-neutral. We assume Llama 2's Reinforcement Learning with Human Feedback (RLHF) approach (Touvron et al., 2023) guides the model to generate more inclusive responses. When Llama 2 is unsure about a name's gender, it may default to labeling it as neutral, potentially reducing prediction accuracy for gender-neutral names.

LLMs have worst performance on gender-neutral names. We also find that most tested LLMs have more difficulties in predicting gender-neutral first names than binary genders, which may stem from the training data of LLMs that primarily includes binary gender labels in the training documents (Touvron et al., 2023). Llama 3, in particular, performed poorly overall across all three datasets with different input variations (i.e., first names with or without birth years). As detailed in Appendix A, the datasets used for dynamically labeling genders were imbalanced, with gender-neutral names being the majority. Specifically, the total numbers of gendered names for the US, Canada, and France

SSA datasets were 3,996, 1,308, and 449, respectively, with around 58.1%, 82.9%, and 75.5% being gender-neutral. Consequently, Llama 3 underperformed in overall prediction accuracy compared to other LLMs due to its poor accuracy in predicting neutral genders despite performing better in predicting binary genders.

LLM performance is biased towards recent year patterns. Based on the over-time comparison of the US SSA dataset (Figure 2), we hypothesize that the improved prediction performance of LLMs for recent data can be attributed to the increased volume of training data from recent years. We assume that the training data of LLMs is unbalanced, predominantly consisting of recent data, potentially explaining the higher gender prediction accuracy of LLMs in recent years. The comparison of balanced SSA datasets and dynamic gender label datasets shown in Table 2 and Table 3 indicates that LLMs face challenges not only with predicting gender-neutral names but also with dynamically changing gender associations for the same names. This issue likely originates from the inherent limitations of the pre-training approach and data used in LLMs. These models tend to memorize training data, which lacks inferential capability, rather than adapting well to names with evolving gender labels over time. Overall, most LLMs better predict female names than male names, and the accuracy of gender prediction is higher for English-based first names in the US and Canada SSA datasets than in the France SSA.

Suggestions for practitioners As we have highlighted in this work, LLMs have a biased and inaccurate understanding of names and hence we should be careful about using them for gender inference related tasks, even at an aggregate level. Furthermore, when dealing with temporal and especially historical data, LLM's name-based gender understanding may be limited and hence their usage for aggregated data analysis is likely to lead to incorrect results.

5 Conclusion

This study underscores the limited performance of LLMs as classifiers in predicting gender-neutral names compared to binary genders and the challenges posed by the inherent biases in the datasets used to train LLMs, which may lead to unbalanced gender prediction results. By introducing a "neutral" category, we have taken a step towards more inclusive gender prediction. However, our findings revealed that LLMs may struggle recognizing gender-neutral names, especially for non-English first names. Despite efforts to enhance LLMs' predictive capabilities by including temporal data, there were no meaningful improvements in gender prediction accuracy, especially for gender-neutral names. This suggests a fundamental limitation of current LLMs and training datasets when adapting to the complexities of gender identities. In future studies, we plan to expand our work by using more inclusive gender categories (e.g., cisgender and transgender) to thoroughly assess gender bias in LLMs across various NLP downstream tasks, including sentiment analysis and coreference resolution.

6 Bias Statement

Our study investigates gender bias in LLMs and fine-tuned foundation language models when predicting the gender of names by introducing a "neutral" category alongside the traditional binary classification of male and female gender labels. Traditionally, the binary gender classification system has not accounted for gender-neutral names. This exclusion arises from imbalanced training data and fixed representations of gender (i.e., female and male), causing LLMs to be prone to classify names into binary gender labels.

When using LLMs in name-based gender prediction tasks, they generally consider only two gender labels, thereby restricting the scope of gender-related analysis. This binary approach perpetuates potential biases in areas associated with fixed gender representations (Liu et al., 2023; Teich et al., 2022), e.g., how male and female authors express sentiment (Jentzsch and Turan, 2022) or how male and female researchers face different challenges in academia (VanHelene et al., 2024). However, this binary labeling of gender overlooks individuals with gender-neutral names, which could encompass both female and male identities, thereby missing valuable insights from a more inclusive perspective. Our work considers more inclusive gender labeling by examining the accuracy of gender-neutral name predictions using LLMs while also providing insights into factors that may lead to biased gender prediction results (i.e., poorer prediction for neutral names compared to binary names) in these models.

Limitations

Our study's limitations are as follows: (1) Our assessment was limited to specific countries, i.e., the US, Canada, and France, not considering a broad spectrum of countries and cultures, particularly in Asia and Africa. This limitation may affect the generalizability of our findings across different cultural and linguistic contexts. (2) The dataset preparation involved a subjective threshold to determine gender-neutral names, defined as names where the gender frequency for both males and females is greater than 10%. This choice may impact the reliability and consistency of the presented findings. (3) The prompt templates employed for interacting with LLMs were not optimized, which may lead to variations in results with different prompt formulations. This indicates a potential variability in LLMs' performance that could impact the robustness of our conclusions, as LLMs are sensitive to prompt design.

References

AI@Meta. 2024. Llama 3 model card.

Michelle Alexopoulos, Kelly Lyons, Kaushar Mahetaji, Marcus Emmanuel Barnes, and Rogan Gutwillinger. 2023. Gender inference: can chatgpt outperform common commercial tools? *arXiv preprint arXiv:2312.00805*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Herbert Barry III and Aylene S Harper. 2014. Unisex names for babies born in pennsylvania 1990–2010. *Names*, 62(1):13–22.

Cameron Blevins and Lincoln Mullen. 2015. Jane, john... leslie? a historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, 9(3).

Paula Chatterjee and Rachel M Werner. 2021. Gender disparity in citations in high-impact journal articles. *JAMA Network Open*, 4(7):e2114509–e2114509.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jana Diesner and Kathleen M Carley. 2009. He says, she says. pat says, tricia says. how much reference resolution matters for entity extraction, relation extraction, and social network analysis. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–8. IEEE.

Ly Dinh, Janina Sarol, Sullam Jeoung, and Jana Diesner. 2023. Are we projecting gender biases to ungendered things? differences in referring to female versus male named hurricanes in 33 years of news coverage. *Computational Communication Research*, 5(1):141.

Yifan Hu, Changwei Hu, Thanh Tran, Tejaswi Kasturi, Elizabeth Joseph, and Matt Gillingham. 2021. What's in a name?–gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4):1537–1563.

Sophie Jentzsch and Cigdem Turan. 2022. Gender bias in bert-measuring and analysing biases through sentiment rating in a realistic downstream classification task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199.

Jizheng Jia and Qiyang Zhao. 2019. Gender prediction based on chinese name. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 676–683. Springer.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Kriste Krstovski, Yao Lu, and Ye Xu. 2023. Inferring gender from name: a large scale performance evaluation study. *arXiv preprint arXiv:2308.12381*.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. 2013. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213.

Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213.

Fengyuan Liu, Petter Holme, Matteo Chiesa, Bedoor AlShebli, and Talal Rahwan. 2023. Gender inequality and self-publication are common among academic editors. *Nature human behaviour*, 7(3):353–364.

Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI spring symposium series*.

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

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. CharBERT: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexopoulos Michelle, Lyons Kelly, Mahetaji Kaushar, Barnes Marcus Emmanuel, and Gutwillinger Rogan. 2023. Gender inference: Can chatgpt outperform common commercial tools? In *Proceedings of the 33rd Annual International Conference on Computer Science and Software Engineering*, pages 161–166.

Thomas J Misa. 2022. Gender bias in big data analysis. *Information & Culture*, 57(3):283–306.

Shubhanshu Mishra, Brent D Fegley, Jana Diesner, and Vetle I Torvik. 2018. Self-citation is the hallmark of productive authors, of any gender. *PloS one*, 13(9):e0195773.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition. In *Proceedings of the KG-BIAS Workshop 2020 at AKBC 2020*.

OpenAI. 2024a. Chatgpt. https://openai.com/chatgpt/. Accessed: 2024-05-20.

OpenAI. 2024b. Gpt-3.5. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2024-05-21.

Duong Tien Pham and Luan Thanh Nguyen. 2023. Gendec: A machine learning-based framework for gender detection from japanese names. *arXiv preprint arXiv:2311.11001*.

Marina Pilkina and Andrey Lovakov. 2022. Gender disparities in russian academia: A bibliometric analysis. *Scientometrics*, 127(6):3577–3591.

Henrique Pinheiro, Matt Durning, and David Campbell. 2022. Do women undertake interdisciplinary research more than men, and do self-citations bias observed differences? *Quantitative science studies*, 3(2):363–392.

Lauren Rhue, Sofie Goethals, and Arun Sundararajan. 2024. Evaluating llms for gender disparities in notable persons. *arXiv preprint arXiv:2403.09148*.

Matthew B Ross, Britta M Glennon, Raviv Murciano-Goroff, Enrico G Berkes, Bruce A Weinberg, and Julia I Lane. 2022. Women are credited less in science than men. *Nature*, 608(7921):135–145.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

Cong Tang, Keith Ross, Nitesh Saxena, and Ruichuan Chen. 2011. What's in a name: A study of names, gender inference, and gender behavior in facebook. In *Database Systems for Adanced Applications: 16th International Conference, DASFAA 2011, International Workshops: GDB, SIM3, FlashDB, SNSMW, DaMEN, DQIS, Hong Kong, China, April 22-25, 2011. Proceedings 16*, pages 344–356. Springer.

Erin G Teich, Jason Z Kim, Christopher W Lynn, Samantha C Simon, Andrei A Klishin, Karol P Szymula, Pragya Srivastava, Lee C Bassett, Perry Zurn, Jordan D Dworkin, et al. 2022. Citation inequity and gendered citation practices in contemporary physics. *Nature Physics*, 18(10):1161–1170.

Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2020. Gender prediction based on vietnamese names with machine learning techniques. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 55–60.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alexander D VanHelene, Ishaani Khatri, C Beau Hilton, Sanjay Mishra, Ece D Gamsiz Uzun, and Jeremy Warner. 2024. Inferring gender from first names: Comparing the accuracy of genderize, gender api, and the gender r package on authors of diverse nationality. *medRxiv*, pages 2024–01.

Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Datasets	# Names	Year span	Train	Val	Test	Overall
US SSA	300	1914 - 2022	78480	9810	9810	98100
Canada SSA	273	2013 - 2020	5232	648	672	6552
France SSA	32	1908 - 2022	8625	1035	1380	11040

Table 4: Statistics of balanced SSA datasets. # Names represent the number of names per gender per year.

Datasets	# Neutral	# Male	# Female
US SSA	2321	1044	631
Canada SSA	1084	110	114
France SSA	339	59	51

Table 5: Statistics of dynamic gender label datasets.

A Dataset Statistics

Overall training and testing dataset statistics were reported in Table 4. We split the train/val/test sets into 80%/10%/10% of the data. We found that gender-neutral names have increased in both the US and Canada SSA datasets over time and surged in more recent years (i.e., after 2000).

Dataset statistics of dynamic gender labels extracted from the three datasets' test sets are reported in Table 5. Note that the Canada SSA dataset only contained 63 first names whose gender labels changed over time in the test set and 50 in the validation set, which was insufficient for evaluating LLMs' performance in dynamic gender prediction. Therefore, we used the training set to extract the names with dynamic gender labels for the Canada SSA dataset.

B Experimental Settings

In foundation language model fine-tuning, we set the maximum length of the tokenizer to 32 across all three models since the results won't change with an increase in the maximum input length. We fine-tuned foundation language models through 7 epochs, and the batch size for either training or validation was 128. We set the warm-up ratio to 0.1 and the learning rate toas 2e-5. The foundation language models included BERT (`bert-base-cased`), RoBERTa (`roberta-base`), and CharRoBERTa. We chose the cased models because they are case-sensitive and can distinguish names such as "huntley" and "Huntley".

For the model settings of LLMs, we applied GPT-3.5 (`gpt-3.5-turbo-instruct`), Llama 2 (`meta/llama-2-70b-chat`), Llama 3 (`meta/meta-llama-3-70b-instruct`), Mixtral-8x7B (`mixtral-8x7b-instruct-v0.1`), and Claude 3 Haiku (`claude-3-haiku-20240307`) for name gender prediction tasks.

C LLMs for Gender Prediction

We applied the 5 LLMs for name-based gender prediction using three country-level SSA datasets.

GPT-3.5. GPT-3.5 is an autoregressive generation model developed by OpenAI (OpenAI, 2024b). The model (`gpt-3.5-turbo-instruct`) has been tuned through an instruction-tuning technique and aims to generate human-preferred responses.

Llama 2. Llama 2 is a collection of open-source chat models developed by Meta, ranging from 7 to 70B parameters (Touvron et al., 2023). It was trained on 2 trillion tokens of publicly available data and tuned through over one million new human-annotated examples. We applied `llama-2-chat` for our experiments.

Llama 3. Following Llama 2, Llama 3 is a series of pre-trained and instruction-tuned autoregressive models in 8 and 70B sizes (AI@Meta, 2024). The training data of Llama 3 is over seven times larger than Llama 2, reaching over 15 trillion tokens of data and over 10M human-annotated examples.

Mixtral-8x7B. Mixtral-8x7B is a pre-trained generative Sparse Mixture of Experts (Jiang et al., 2024). The Mixtral-8x7B outperformed Llama 2 70B on most benchmarks and can handle English, French, Italian, German, and Spanish, which is helpful when predicting French name genders.

Claude 3 Haiku. Claude 3 family is a series of close-source language models, including three state-of-the-art models in ascending order of capability: Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus (Anthropic, 2024). Claude 3 Haiku is the fastest, most compact model for near-instant responsiveness. We used Claude 3.

D Prompt Templates for LLMs

We reported the prompt templates for the experiments of LLMs in 0- and 5-shot settings for RQ 1 and RQ 2 in Table 6. For RQ 1, we used "First Name" for gender prediction. For RQ 2, we provided "First Name" and "Year of Birth" as input.

In the 5-shot setting, we randomly chose five name-gender pairs from the three SSA datasets, using the number 42 as the random seed. We selected names that appeared at least twice and were assigned different genders in different years.

Experimental Setting	RQ 1	RQ 2
0-shot	Predict the gender association of the given name. \nUse the following labels for classification: \nMale: The name is predominantly associated with males. \nFemale: The name is predominantly associated with females. \nNeutral: The name is not predominantly associated with any single gender and is considered neutral. \nYour outputs should be all in lowercase and can only output gender from male, female, or neutral. \nName: + {name} + \nGender:	Predict the gender association of the given name, considering the year of birth as an additional reference. \nThe provided names appear more than once across different years of birth as they may be labeled in different genders given the change in the predominant gender of names. \nUse the following labels for classification: \nMale: The name is predominantly associated with males. \nFemale: The name is predominantly associated with females. \nNeutral: The name is not predominantly associated with any single gender and is considered neutral. \nYour outputs should be all in lowercase and can only output gender from male, female, or neutral. \nName: + {name} + \nYear of Birth: + {year} + \nGender:
5-shot (US SSA)	Predict the gender association of the given name. \nThe provided names appear more than once. \nUse the following labels for classification: \nMale: The name is predominantly associated with males. \nFemale: The name is predominantly associated with females. \nNeutral: The name is not predominantly associated with any single gender and is considered neutral. \nPlease note that first names can be labeled in different genders over time. \nHere are five pairs of examples of first names and genders: \nPair 1: Name: Christie, Gender: Neutral; Name: Christie, Gender: Female Pair 2: Name: Jan, Gender: Neutral; Name: Jan, Gender: Male Pair 3: Name: Bee, Gender: Female; Name: Bee, Gender: Neutral Pair 4: Name: Kasen, Gender: Neutral; Name: Kasen, Gender: Male Pair 5: Name: Mel, Gender: Male; Name: Mel, Gender: Neutral \nYour outputs should be all in lowercase and can only output gender from male, female, or neutral. \nName: + {name} + \nGender:	Predict the gender association of the given name, considering the year of birth as an additional reference. \nThe provided names appear more than once. \nUse the following labels for classification: \nMale: The name is predominantly associated with males. \nFemale: The name is predominantly associated with females. \nNeutral: The name is not predominantly associated with any single gender and is considered neutral. \nPlease note that first names can be labeled in different genders over time. \nHere are five pairs of examples of first names and genders: \nPair 1: Name: Christie, Year of Birth: 1919, Gender: Neutral; Name: Christie, Year of Birth: 1949, Gender: Female Pair 2: Name: Jan, Year of Birth: 1966, Gender: Neutral; Name: Jan, Year of Birth: 2012, Gender: Male Pair 3: Name: Bee, Year of Birth: 1952, Gender: Female; Name: Bee, Year of Birth: 1989, Gender: Neutral Pair 4: Name: Kasen, Year of Birth: 2000, Gender: Neutral; Name: Kasen, Year of Birth: 2006, Gender: Male Pair 5: Name: Mel, Year of Birth: 1947, Gender: Male; Name: Mel, Year of Birth: 2007, Gender: Neutral \nYour outputs should be all in lowercase and can only output gender from male, female, or neutral. \nName: + {name} + \nYear of Birth: + {year} + \nGender:
5-shot (Canada SSA)	...Pair 1: Name: Nyjah, Gender: Neutral; Name: Nyjah, Gender: Male Pair 2: Name: Kendell, Gender: Neutral; Name: Kendell, Gender: Male Pair 3: Name: Arshia, Gender: Neutral; Name: Arshia, Gender: Male Pair 4: Name: Lennix, Gender: Neutral; Name: Lennix, Gender: Female Pair 5: Name: Kirat, Gender: Male; Name: Kirat, Gender: Neutral...	...Pair 1: Name: Nyjah, Year of Birth: 2014, Gender: Neutral; Name: Nyjah, Year of Birth: 2016, Gender: Male Pair 2: Name: Kendell, Year of Birth: 2014, Gender: Neutral; Name: Kendell, Year of Birth: 2016, Gender: Male Pair 3: Name: Arshia, Year of Birth: 2014, Gender: Neutral; Name: Arshia, Year of Birth: 2018, Gender: Male Pair 4: Name: Lennix, Year of Birth: 2013, Gender: Neutral; Name: Lennix, Year of Birth: 2018, Gender: Female Pair 5: Name: Kirat, Year of Birth: 2013, Gender: Male; Name: Kirat, Year of Birth: 2014, Gender: Neutral...
5-shot (France SSA)	...Pair 1: Name: CARMEL, Gender: Male; Name: CARMEL, Gender: Neutral Pair 2: Name: LIE, Gender: Male; Name: LIE, Gender: Neutral Pair 3: Name: JESSY, Gender: Female; Name: JESSY, Gender: Neutral Pair 4: Name: ANH, Gender: Neutral; Name: ANH, Gender: Male Pair 5: Name: FIDELE, Gender: Neutral; Name: FIDELE, Gender: Female...	...Pair 1: Name: CARMEL, Year of Birth: 1920, Gender: Male; Name: CARMEL, Year of Birth: 1951, Gender: Neutral Pair 2: Name: LIE, Year of Birth: 1922, Gender: Male; Name: LIE, Year of Birth: 1931, Gender: Neutral Pair 3: Name: JESSY, Year of Birth: 1960, Gender: Female; Name: JESSY, Year of Birth: 1975, Gender: Neutral Pair 4: Name: ANH, Year of Birth: 1995, Gender: Neutral; Name: ANH, Year of Birth: 2006, Gender: Male Pair 5: Name: FIDELE, Year of Birth: 1918, Gender: Neutral; Name: FIDELE, Year of Birth: 1945, Gender: Female...

Table 6: Task-oriented prompt templates of LLMs in 0-shot and 5-shot settings for RQ 1 (w/o birth year) and RQ 2 (w/ birth year). For clarity, we report only the 5-shot example pairs for Canada and France's SSA datasets, as the prompt templates are the same as those used for the 5-shot US SSA dataset.

E Prompt Robustness Evaluation

The effectiveness of prompts designed for LLM-based experiments is crucial for the performance of downstream natural language processing tasks, as highlighted by Zhou et al. (2022); Zhu et al. (2023). Therefore, we developed two prompt templates inspired by Zhu et al. (2023): task-oriented and role-oriented prompts, to evaluate the robust-

ness of LLM gender prediction performance. The task-oriented prompt was the same as introduced in Appendix D.

0-shot Role-Based Prompt for RQ 1
In the role of a first name gender prediction tool, classify names based on their gender association using the following gender labels:

Male: The name is predominantly associated with males.
Female: The name is predominantly associated with females.
Neutral: The name is not predominantly associated with any single gender and is considered neutral.

The provided names appear more than once. Your outputs should be all in lowercase and can only output gender from male, female, or neutral. "\n Name: " + name + "\n Gender: "

0-shot Role-Based Prompt for RQ 2
In the role of a first name gender prediction tool, classify names based on their gender association using the following gender labels:

Male: The name is predominantly associated with males.
Female: The name is predominantly associated with females.
Neutral: The name is not predominantly associated with any single gender and is considered neutral.

Consider the year of birth as an additional reference. The provided names appear more than once across different years of birth as they may be labeled in different genders given the change in the predominant gender of names.

Your outputs should be all in lowercase and can only output gender from male, female, or neutral. "\n Name: " + name + "\n Year of Birth: " + year + "\n Gender: "

Above are examples of role-based prompts used in RQ 1 and 2 under the 0-shot setting. The 5-shot examples are the same as we applied in task-oriented prompts. We provided first names after "Name" and guided LLMs to output genders after "Gender".

We evaluated the robustness of prompts using GPT-3.5 on the France SSA dynamic gender label dataset referenced in Table 3. As shown in Table 7, our results indicate that in the 0-shot setting, both prompts exhibited similar performance for predicting male and female genders. However, using the task-oriented prompt showed a better performance in predicting gender-neutral names than using the role-oriented prompt. Given that over 75% of names in the French dataset were gender-neutral, even minor discrepancies in the "Neutral" category can significantly impact the overall accuracy. While the role-oriented prompt yielded better predictions for binary gender predictions when only the first names were provided, its overall accuracy still fell behind the task-oriented setting in both experimental setups. Notably, incorporating birth year as an additional feature for name gender prediction reduced the differences between various prompt templates, particularly for the performance of gender-neutral names (Table 7).

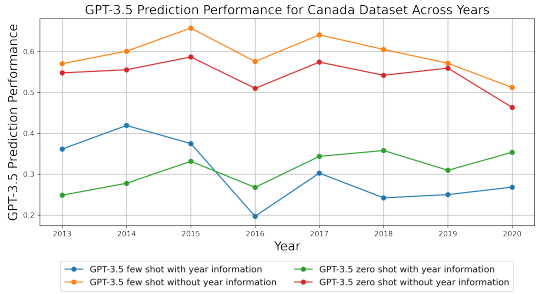
We also assessed the impact of including "Country" information in the gender prediction prompt using the France dataset. The results indicated no significant difference (i.e., the variation in overall accuracy is within 2%) when incorporating the original country of the given names in both 0-shot and 5-shot settings.

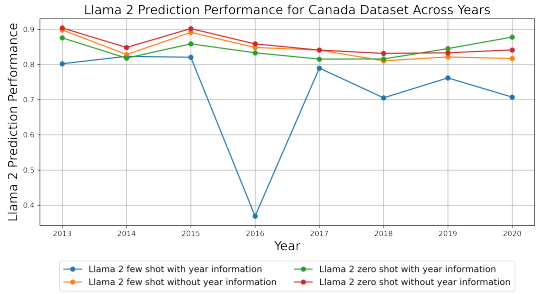
F Over-time Trends of LLM Performances

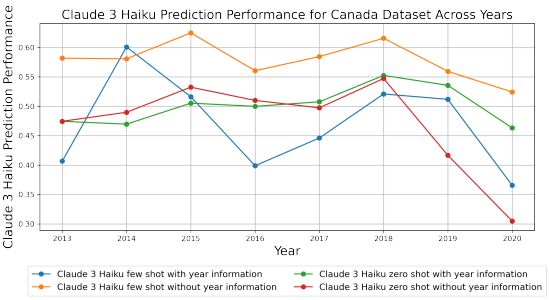
In Figures 3 and 4, we presented the trends in gender prediction accuracy for Canada and France using dynamic gender label datasets across five different LLMs. Generally, the performance of these LLMs varied over time for both datasets. Notably, models that did not incorporate temporal information tended to perform better, yielding more stable accuracy rates over the years than models that included birth year data. Figure 3 also indicated that the LLMs were less effective at predicting names from more recent years. In particular, GPT-3.5 demonstrated that omitting temporal information led to higher gender prediction performance consistently over the years than including it.

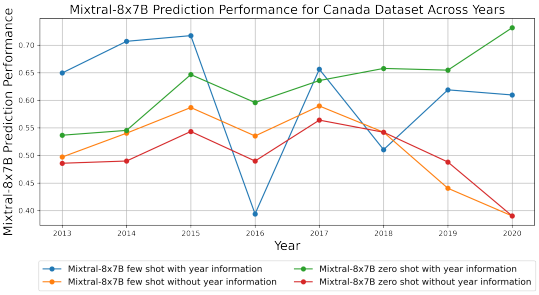
	First Name				First Name + Year			
Models	Male	Female	Neutral	Acc.	Male	Female	Neutral	Acc.
Task-o Oriented Prompt (0-shot)	78.43	98.31	16.52	34.30	90.20	98.31	3.54	25.84
Role-o Oriented Prompt (0-shot)	78.43	98.31	9.73	29.18	88.24	98.31	3.54	25.61
Task-o Oriented Prompt (5-shot)	78.43	98.31	20.35	37.19	98.04	100.00	5.01	28.06
Role-o Oriented Prompt (5-shot)	90.20	100.00	17.11	36.30	92.16	100.00	4.42	26.95

Table 7: Prompt robustness evaluation of name gender prediction using GPT-3.5 under the France dynamic gender label dataset.









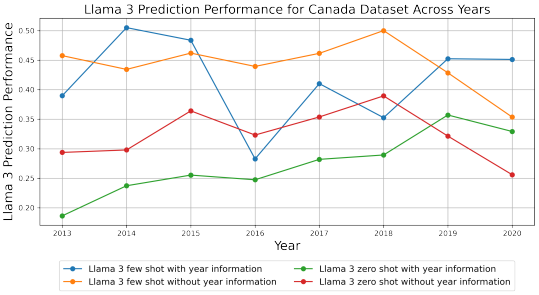
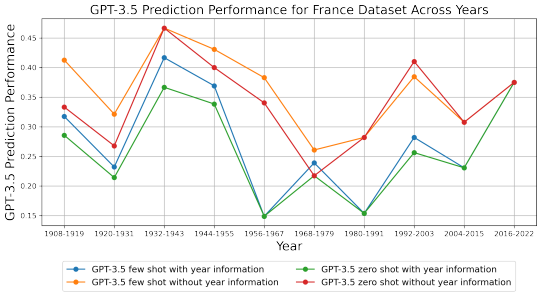
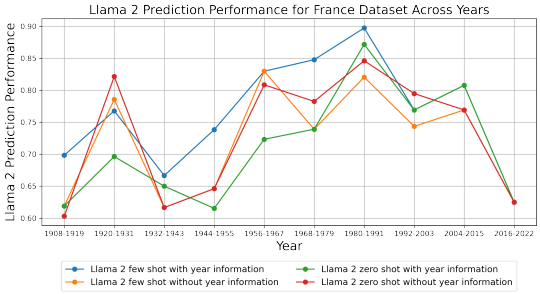
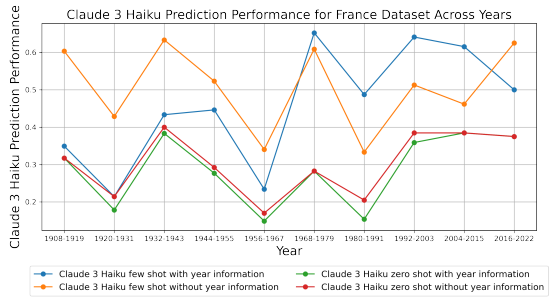
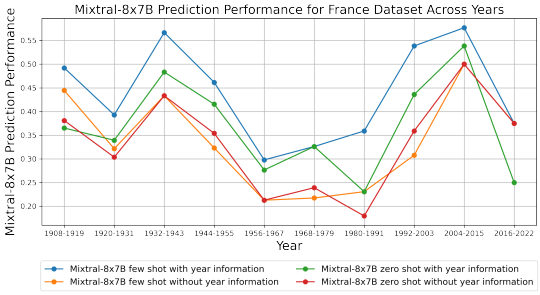


Figure 3: Temporal-level comparison of all LLMs across Canada SSA dynamic gender label dataset given the results of Table 3.









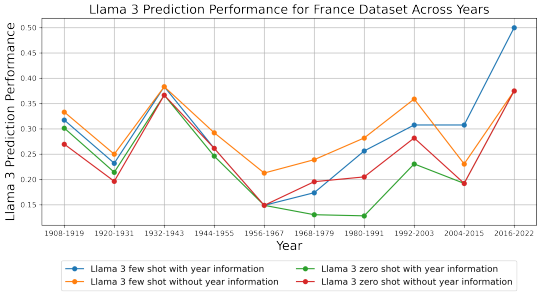


Figure 4: Temporal-level comparison of all LLMs across France SSA dynamic gender label dataset given the results of Table 3.