Inferring Gender from Author Names with Local LLMs: A Multi-Model Evaluation

Victor Herrero-Solana[1], Elvira González-Salmón[2*] and Nicolas Robinson-Garcia[2]

[1] SCImago-UGR Group, Unit for Computational Humanities and Social Sciences (U^CHASS), University of Granada, Spain
[2] EC3 Research Group, Unit for Computational Humanities and Social Sciences (U^CHASS), University of Granada, Spain
[*] Correspondence: elviragonzalez@go.ugr.es

Abstract
Gender identification of researchers is a common practice in scientometric studies examining inequalities in science. The most widely used approach relies on inferring gender from author names using commercial APIs or name-gender dictionaries, which often lack transparency and reproducibility. This study explores the use of local open-weight Large Language Models (LLMs) as an alternative for name-based gender classification. We evaluate 25 models from seven leading families (Llama, Gemma, Phi, Mistral, Qwen, DeepSeek, and Yi), ranging from 270 million to 70 billion parameters, using a reference dataset of nearly 200,000 names across 195 countries extracted from Wikidata. Results show that top-performing models achieve F1-Scores above 0.93 for both gender categories, positioning local LLMs as a viable, cost-effective, and reproducible alternative to proprietary tools. A critical performance threshold emerges at approximately 7 billion parameters, above which all models achieve acceptable results, with diminishing returns beyond 12-14 billion. All models exhibit systematic gender bias, showing higher precision for men and higher recall for women, indicating a tendency to classify ambiguous names as male. Mistral-Nemo-12b emerges as the optimal choice, balancing accuracy, computational efficiency, and gender equity.

Keywords: Generative AI; Local Large Language Models; gender assignment algorithms; scientometrics

1. Introduction

Scientometrics has expanded its analytical scope over the past decade by moving from publications as the primary unit of analysis to authors (Wildgaard et al. 2014). The introduction of author identifiers and registries not only allows tracking author publication histories, but also broadens the variables that can be analyzed to include individual characteristics such as career stage or geographic trajectories, among other (Robinson-Garcia et al. 2025). Gender is another variable that has received considerable attention with a stream of bibliometric studies examining gender inequalities in science (González-Salmón et al. 2025). So far, they do so by inferring gender using different approaches such as the use of images (Karimi et al. 2016), or self-identification (Campbell and Simberloff 2022), but the most common approach is based on author names. Some studies do provide extensive details on the methodological choices made when using this approach (Bérubé et al. 2020; Boekhout et al. 2021; Huang et al. 2020; Kozlowski et al. 2022). But in other cases, researchers use third-party algorithms such as Gender Guesser, Gender API or Genderize among others. While this is not an issue in itself, the lack of transparency on the inner workings of these algorithms impedes accountability and can hinder comparisons across studies (González-Salmón and Robinson-Garcia 2024a).

Recent advances in large language models (LLMs) have created new opportunities to perform computationally intensive tasks locally, at substantially lower technical and computational costs (e.g., Touvron et al. 2023a). This paper explores the possibility of using local LLMs to infer gender based on author names from scientific publications. Open local LLMs offer some advantages over commercial options, as they allow greater control over model parameters, facilitating reproducibility and preserving research data privacy when deployed in controlled environments. Here, we select 25 open local LLMs from the 7 leading families of LLMs in the market (e.g., Llama, Gemma, Qwen). We cover most available model sizes available, from micro LLMs (<4B) to large reasoning models (70B), to test their capability to infer gender by comparing their results with a master list of name-gender-country information. Our master list is extracted from Wikidata, which serves as a reference for evaluation. Using this dataset, we compare the recall and precision of different locally executed Large Language Models in order to assess their relative performance, examine systematic differences and analyze the effects of model size.

Although we approach the use of local LLMs from the field of scientometrics, our findings may inform other fields within the Social Sciences and Humanities where gender information may be of interest. Some examples are studies from the field of education on students' grades (e.g. Cyrenne and Chan 2012), sociology of science (e.g. Sugimoto et al. 2017) or economics on data consumption (e.g. Zorell and Denk 2021), among others. If we aim to include the gender variable[1], a reliable method for inferring such information is crucial.

2. Literature review

This section outlines the main approaches used to assign gender to names, as summarized in Table 1. First, name-based gender identification relies on probabilistic associations between personal names and gender, often incorporating country-level information to improve accuracy. This approach is typically implemented through commercial APIs or open gender–name dictionaries and remains the most widely used method in large-scale studies. Second, when using Large Language Models, gender can be inferred either indirectly from textual data or directly from names. In text-based approaches, LLMs exploit linguistic and stylistic patterns to predict gender, whereas name-based LLM approaches leverage the contextual knowledge embedded in the models' training data to classify gender from names alone. We will now see each approach in detail.

Table 1. Approaches used in research to assign gender to names.

Approach	Inner mechanism	Tools	Limitations
Name-based gender identification	Gender inferred probabilistically from personal names, sometimes contextualised by country	● Commercial APIs (e.g. genderize.io, Gender API, NamSory) ● Open gender–name dictionaries (e.g. WGND)	● Assumes gender–name stability ● Unisex names ● Uneven regional coverage (notably Asian names) ● Binary gender model ● Limited transparency of proprietary data
Gender detection from text using LLMs	Gender inferred from stylistic and linguistic patterns in written text	● BERT ● RoBERTa ● BERTweet ● Hybrid models (e.g. BERT + XGBoost)	● Moderate accuracy ● Domain dependence ● Ethical concerns ● Indirect and noisy gender signals

[1] Throughout this paper, we use women/men to refer to gender rather than female/male. Quoted articles are not modified; therefore, some of them may not follow this guideline.

Name-based gender classification using LLMs	Gender inferred from names using contextual knowledge embedded in LLMs	• 360 Brain • Baichuan • BERT • ChatGLM • ChatGPT • Claude • Doubao • ERNIE Bot (Baidu) • Haiku • Llama • Mistral • Qingyan • Qwen (Alibaba) • RoBERTa • Skywork	• Accuracy varies by language and culture • Bias toward Anglo-Saxon men names • Sensitivity to prompt design • Binary output constraints

2.1. Name-based gender identification

The most widely used and simplest approach to assign gender to individuals consists of inferring gender from personal names (Mihaljevíc et al. 2019). In small datasets, researchers may perform this task manually, assigning gender based on common name–gender associations or on personal knowledge of the individuals involved. However, such approaches are not practical when dealing with larger datasets, as is the case in bibliometric studies. Hence, most authors typically rely on automated methods, often implemented through third-party tools, which assign gender based on given names—sometimes in combination with surnames—using predefined name–gender associations (González-Salmón and Robinson-García 2024a). These solutions are typically based on lists of names associated with the most probable gender for each name. Accordingly, these lists can be cross-referenced with a dataset of names to assign a gender to individuals in the data (Figure 1). In some cases, such lists also incorporate country information, as the gender associated with a given name may vary across national contexts. A common example is the name Andrea, which is generally assigned to women in Spanish speaking countries and to men in Italy. Other variations include Slavic naming conventions, in which gender information is often encoded in the surname rather than in the given name.

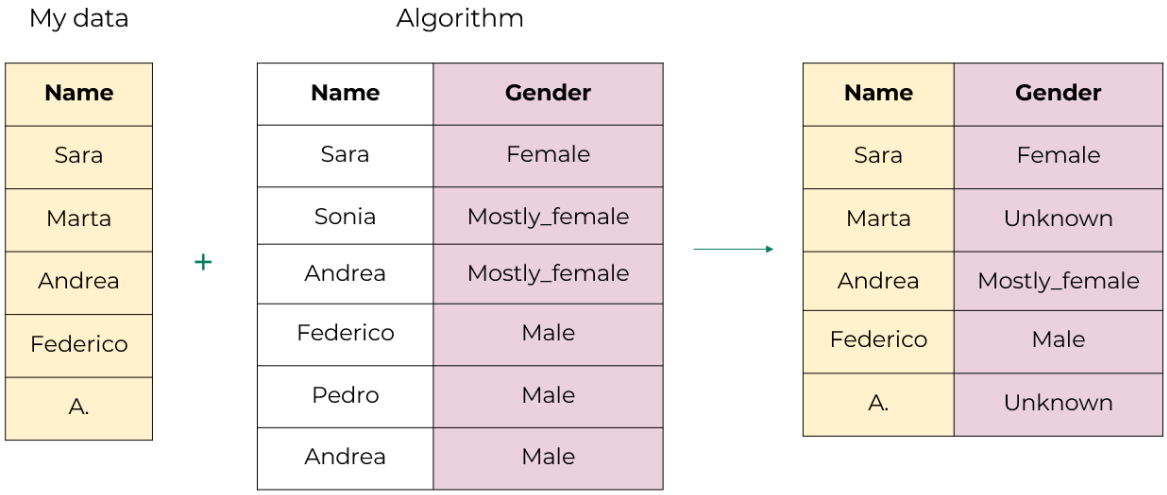

Figure 1. Workflow of gender assignment algorithms.

There are dedicated programs and paid web services that perform this task (for an evaluation procedure for gender assignment methods see Matias-Rayme et al. 2024). These tools typically

achieve high accuracy because they rely on very large underlying name datasets. Notable examples include genderize.io, Gender API and NamSor:

- Genderize.io aggregates names from across the web and contains over 900 million country–name combinations (https://genderize.io/).
- Gender API reports coverage of more than six million names across 190 countries (https://gender-api.com/es).
- NamSor has processed over 11 billion names (https://namsor.app/). In addition to these commercial solutions, open-access and free alternatives are also available. The most prominent is the World Gender Name Dictionary, hosted on GitHub (https://github.com/IES-platform/r4r_gender/blob/main/wgnd/README.md), which includes more than five million names and can be used without advanced programming skills.

Other similar initiatives have been developed, but many have been discontinued due to the resource-intensive nature of maintaining such algorithms (for a more complete list of gender assignment algorithms go to González-Salmón et al. 2025).

This methodology nevertheless presents well-documented limitations (González-Salmón and Robinson-García 2024a). First, it assumes that gender can be reliably inferred from names, an assumption that does not always hold, particularly in the presence of unisex names. Second, the coverage of names is uneven across regions, with Asian names being especially underrepresented (Karimi et al. 2016), introducing geographical inequalities that may bias empirical results. Third, these approaches generally conceptualize gender as a binary category, which results in the systematic exclusion of non-binary, transgender, and other gender-diverse identities. Finally, commercial algorithms are often opaque regarding the provenance and composition of their training data, which introduces additional uncertainty and limits reproducibility and critical assessment of their outputs.

2.2. Gender detection using Large Language Models (LLM)

The development of the Transformer architecture (Vaswani et al. 2017) has marked a turning point in natural language processing. It has revolutionized the field by replacing traditional recurrent architecture with attention mechanisms, enabling parallel sequence processing and capturing long-range dependencies more efficiently. Furthermore, it has paved the way for all subsequent Large Language Models (LLMs), establishing a new paradigm in language generation and understanding which has influenced all fields of science and daily practice.

The evolution of Transformer-based architectures, moving from bidirectional encoding (e.g., BERT) to the aligned reasoning capabilities of modern LLMs (Ouyang et al. 2022) like ChatGPT, has established a new paradigm for complex inference tasks. While initial breakthroughs were driven by proprietary systems, a critical shift has occurred with the proliferation of high-performance open-weight models. The release of efficient architectures such as Llama and Mistral has narrowed the performance gap between closed APIs and locally executable models (Touvron et al. 2023b; Jiang et al. 2023), allowing researchers to deploy sophisticated reasoning on consumer hardware.

In the context of gender identification, this democratization offers a powerful alternative to static name lists: it enables context-aware classification that is scalable, reproducible, and independent of third-party services. In this regard, we observe two different approaches on gender inference: based on written text and based on individual names. While the former is rooted in computational sociolinguistics and stylometry, focusing on linguistic patterns and

stylistic choices (Argamon et al. 2003; Nguyen et al. 2016), the latter aligns more closely with the purpose of bibliometric studies. Next, we briefly review the literature on both approaches.

2.2. 1. Gender Detection from Written Text

Prior to the wide adoption of generative LLMs, significant research focused on inferring author demographics through stylometric analysis of written content. Early Transformer-based approaches yielded mixed results; Sazzed (2022) reported suboptimal performance (<65%) using BERT and RoBERTa on general text. However, domain-specific tuning improved these metrics: Parasurama and Sedoc (2021) achieved 76% accuracy analyzing anonymized CVs, while Alzahrani et al. (2022) reached 87% on Twitter data by incorporating the platform-specific BERTweet model. More recently, Kavuri and Kavitha (2023) demonstrated that hybrid architectures can surpass the 90% threshold by combining BERT embeddings with ensemble learning techniques like XGBoost.

2.2. 2. Name-Based Gender Classification Using LLMs

A different approach involves leveraging the cultural and linguistic knowledge embedded in LLMs to classify gender based on names, offering a semantic alternative to traditional dictionary-based APIs. Recent studies highlight the variability of this approach depending on the model family and linguistic context. In the context of Chinese names, Deng et al. (2025) achieved an 86% accuracy rate analyzing thesis authors using an ensemble of eight Chinese-developed LLMs (e.g., ERNIE, Qwen). In contrast, Zhuang et al. (2024) reported significantly lower performance (56–68%) when applying ChatGPT-3.5 and ChatGLM-6b to international GitHub user data using a simple name-plus-country prompt.

Methodological refinements, such as prompt engineering, have shown promise in mitigating these discrepancies. You et al. (2024) explored identifying neutral names across multiple models (including Llama 2/3 and Claude 3), finding that while traditional BERT models still outperform LLMs on ambiguous names, LLM performance improves with "few-shot" prompting strategies that provide context examples.

The viability of open-weights and diverse model architectures was further validated by AlNuaimi et al. (2024). Their extensive evaluation of 12 zero-shot models (including Mistral, Yi, and Llama) across datasets such as Florida voters and Wikipedia yielded accuracy rates exceeding 90%, although they noted a persistent bias favoring Anglo-Saxon masculine names. Finally, Domínguez-Díaz et al. (2024) compared ChatGPT against established tools like NamSor. They concluded that while LLMs offer reliable inferences, traditional APIs currently maintain an advantage in transparency and control for rigorous empirical research, suggesting that the transition to LLM-based classification requires careful validation of the specific models employed.

2.3. Gender Bias in LLMs

The emergence of gender biases in LLMs is an issue that has concerned researchers since their inception. Gender bias can occur at any point during training and does not necessarily diminish as the model grows larger. This has been demonstrated with the experimental LLM Pythia (Patel et al. 2025). An interesting work in this regard is that of He (He 2025), which studies the recommendation of anonymized scientific articles made by well-known LLMs such as ChatGPT and Claude. It concludes that the papers recommended by ChatGPT models are biased toward men, while those from Claude are biased toward the dominant gender (according to the number of authors). The effect is less pronounced in Social Sciences than in other disciplines.

3. Research Questions

The main objective of this work is to study the possibilities of local LLMs for gender identification. Unlike proprietary, cloud-based systems, local models can be executed under controlled and stable conditions, allowing the same model version, parameters, and computational environment to be preserved and reused, thereby ensuring reproducible results. In addition, these models can run on consumer-grade hardware without relying on external API calls, which entails two further advantages: a substantially lower cost than proprietary solutions such as ChatGPT or Claude, enabling large-scale processing, and the preservation of data privacy, since all processing occurs locally and sensitive information never leaves the user's infrastructure. To this end, we aim to answer the following questions:

- RQ1. In general, are LLMs efficient for gender identification based on the simple combination of first name + country?
- RQ2. Do classification metrics (precision, recall, F1-score) show systematic differences between men and women categories?
- RQ3. Does model size matter? Are results proportional to the number of parameters?
- RQ4. Is there any bias based on the LLM's country of origin?
- RQ5. Which LLMs offer the best balance between precision and recall?

4. Materials and Methods

4.1. Data

As a reference dataset to compare the performance of the different LLMs models, we use a dataset constructed using information retrieved from Wikidata. Wikidata allows for the download of data through their own query, the Wikidata Query Service (WDQS), which uses the SPARQL query language. Specifically, we extracted a list of individuals' given names and surnames, linked to the country of citizenship and gender (See Appendix 1 for an example of the query). The population was defined as all entries corresponding to humans who are included in Wikipedia, operationalized through Wikidata's structured properties (as of July 2023). For each individual, we obtained their country of origin and gender.

Gender and country were used as key filtering dimensions to generate separate subsets of records, ensuring internal consistency across extractions. All labels were standardized to English in order to avoid duplication and language-related ambiguities. Due to technical constraints on query size and result limits, data extraction had to be carried out through fragmented queries rather than a single download per country and gender. For countries with large populations in Wikidata, such as the United States, this resulted in dozens of separate files (e.g., 97 files for men individuals for the United States). These fragments, initially divided by gender, were merged into a single country-level file. This allowed us to identify given names that appear across both genders within the same national context and to compute the proportion of occurrences associated with each gender. In the case of Slavic countries, where surnames encode gendered morphological variations (e.g. surnames ending in ov, en, in, eb, yi, yj, ky, kii, kij and ob are generally assigned to men, and to women those ending in ova, eva, ina, oba, aya, aia, ina and iha), an additional dataset was generated focusing specifically on family names. All country-level datasets were then integrated into a unified dataset, preserving country-specific and gender-specific distributions while enabling cross-national comparison. As a result, the dataset reflects the well-documented biases of Wikipedia, notably the systematic overrepresentation of men relative to women (Tripodi 2023).

Once the data were consolidated, there was a data cleaning stage. This process involved separating given names from middle names, removing names written in non-Latin scripts, converting all strings to lowercase, and eliminating special characters such as hyphens. After cleaning, we computed, for each country, the proportion of occurrences in which a given name was associated with each gender. A gender was assigned to a name–country pair when that gender accounted for at least 70% of the occurrences in that national context. Table 2 presents an illustrative example of the structure of the dataset, including the individual's first name, country of origin, gender as assigned by Wikidata, the number of occurrences of that name within the country, and the proportion of cases in which the name is associated with that gender in the given national context.

First name	Country	Gender	Nº times	Perc
Andrea	Italy	Man	1970	98.98%
Andrea	Argentina	Woman	69	97.10%
Alex	Canada	Man	193	92.23%
Antonio	Perú	Man	127	100%
Azusa	Japan	Woman	74	90.54%
Kurt	Liechtenstein	Man	1	100%
Eleonora	Austria	Woman	6	100%

Table 2 - Sample of the dataset.

Finally, the resulting dataset comprises 199,639 names across 195 countries. As seen in table 3, The United States of America accounts for 7.11% of all observations, followed by four countries with shares close to 3–4% each (Canada 3.95%, United Kingdom 3.76%, Germany 3.71%, and France 3.39%). A further six countries contribute approximately 2% each (Italy, Australia, Switzerland, Spain, Sweden, and Brazil), while all remaining countries represent 1.9% or less individually. For some countries, as seen on the right side of table, the number of names available was less than 50. In terms of gender distribution, 63.30% (126,372) of the names are associated with men and 36.70% (73,267) with women.

Position	Country	Nº of names	% of names	Position	Country	Nº of names	% of names
1	United States	14199	7.11	**186**	Mongolia	43	0.02
2	Canada	7901	3.95	**187**	Comoros	43	0.02
3	United Kingdom	7500	3.76	**188**	Lesotho	42	0.02
4	Germany	7397	3.71	**189**	Vatican City	41	0.02
5	France	6769	3.39	**190**	Maldives	40	0.02
6	Italy	4792	2.40	**191**	Marshall Islands	38	0.02
7	Australia	4590	2.30	**192**	Djibouti	30	0.02
8	Switzerland	4555	2.28	**193**	Laos	22	0.01
9	Spain	4522	2.27	**194**	Bhutan	16	0.01
10	Sweden	4467	2.24	**195**	Western Sahara	6	0.00

Table 3 - Summary of the dataset, including top 10 and bottom 10 countries in terms of percentage of the dataset.

To test the performance of the dataset as a gender assignment tool, we used data from Dimensions, a bibliometric database launched in 2018 by Digital Science. We extracted 8,860,456 researcher names covering the period 1990–2021 and applied the algorithm to this dataset. After completing the global assignment of gender and thematic areas, 50.6% of the names (3,990,896) were associated with men, 28.8% (2,274,275) with women, and the remaining 20.5% (1,618,764) could not be classified with a sufficient level of reliability. Overall, the algorithm was therefore able to identify the gender of 79.5% of the researcher names. The country-level results are shown in Figure 2. As shown in the data, coverage is robust for the Americas and Europe, whereas several regions in Asia and Africa are substantially underrepresented.

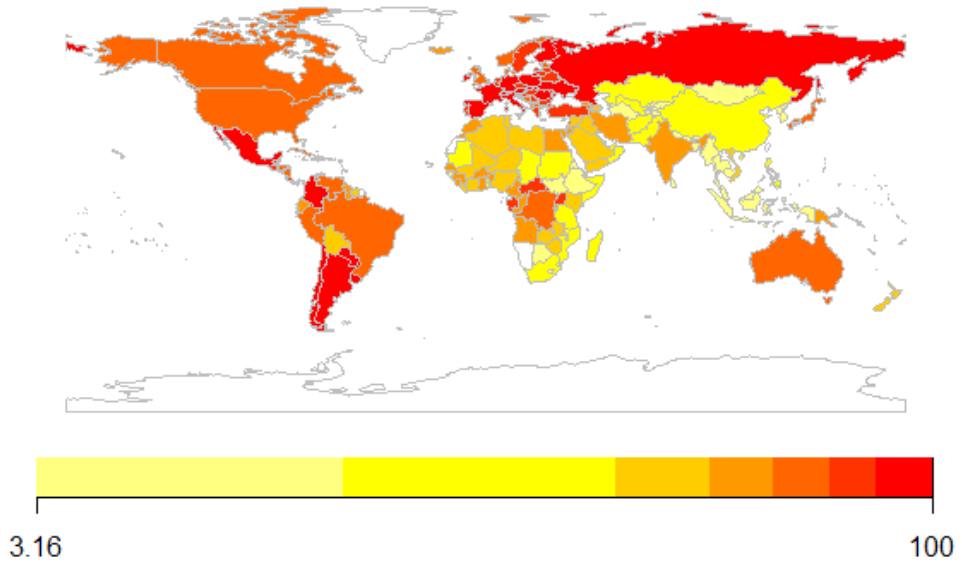

Figure 2 - Percentage of researchers from Dimensions data whose gender identification was successful, by country.

The results obtained with this dataset are comparable to those reported for other open datasets and gender assignment algorithms (González-Salmón et al., 2024). In a previous study, we analysed a sample of 100,000 researcher names from Dimensions to compare the performance of this dataset with other open and freely available approaches, namely the World Gender Name Dictionary (WGND), ChatGPT-3.5, and Gender-Guesser (https://pypi.org/project/gender-guesser/). The comparison showed higher coverage for WGND (87.2%) and ChatGPT-3.5 (99.8%). However, the shortcomings identified in WGND (e.g., assigning gender to initials) and the opacity of ChatGPT-3.5 limit their reliability for large-scale analyses. By contrast, the Wikidata-based approach offers greater transparency and methodological control. Gender-Guesser achieved substantially lower coverage, assigning gender to only 50.7% of the names (González-Salmón & Robinson-Garcia, 2024b). Thus, the dataset used here favours precision over recall and performs adequately for the purposes of this study.

4.2. Models

We evaluated a diverse cohort of 25 open-weight Large Language Models (LLMs), selecting representative architectures from the seven leading families currently defining the state-of-the-art: Llama (Meta), Gemma (Google), Phi (Microsoft), Mistral (Mistral AI), Qwen (Alibaba),

DeepSeek (DeepSeek-AI), and Yi (01.AI). The selection criteria prioritized a broad range of parameter sizes, ranging from lightweight edge models (270M) to high-performance reasoning engines (70B), to analyze the impact of model scale on gender identification tasks. The selected families represent distinct architectural philosophies:

- Meta's Llama series (Touvron et al. 2023a) serves as the foundational baseline for open-weight research, setting the standard for general-purpose generative capabilities.
- Mistral AI (Jiang et al., 2023) and Microsoft's Phi (Microsoft 2024) represent the efficiency-focused approach. Mistral introduced sliding window attention for performance efficiency, while the Phi family challenges scaling laws by leveraging high-quality synthetic data to achieve competitive results with smaller parameter counts.
- Google's Gemma (Gemma Team 2024) adapts the proprietary Gemini architecture into lightweight, open versions optimized for efficient deployment.
- Asian-origin architectures, including Qwen (Qwen Team 2025), DeepSeek (DeepSeek-AI 2024), and Yi (Young et al. 2024), were included to ensure linguistic and cultural diversity. These models introduce significant architectural innovations, such as DeepSeek's Multi-head Latent Attention (MLA) and mixture-of-experts (MoE) optimizations, as well as Qwen's extensive multilingual pre-training.

We deliberately excluded reasoning or thinking models due to their non-deterministic chain-of-thought outputs, which introduce variability unsuitable for standardized zero-shot classification. Similarly, code-specialized variants were omitted, as their training data is biased towards syntactic structures rather than the natural language semantics required for gender inference. Table 4 details the 25 models selected, categorized by company and country and ordered by parameter count.

#	Model	Company	Country
1	Llama3.1-70b	Meta	USA
2	Llama3.3-70b	Meta	USA
3	Gemma3-27b	Google	USA
4	Deepseek-v2-16b	Deepseek	China
5	Qwen2.5-14b	Alibaba	China
6	Phi4-14b	Microsoft	USA
7	Phi3-14b	Microsoft	USA
8	Mistral-Nemo-12b	Mistral	France
9	Gemma3-12b	Google	USA
10	Yi-9b	01.AI	Taiwan
11	Llama3.1-8b	Meta	USA
12	Dolphin-Mistral-7b	Mistral	France
13	Deepseek-llm-7b	Deepseek	China
14	Mistral-7b	Mistral	France
15	Qwen2.5-7b	Alibaba	China
16	Yi-6b	01.AI	Taiwan
17	Gemma3-4b	Google	USA
18	Phi3-3.8b	Microsoft	USA
19	Llama3.2-3b	Meta	USA
20	Qwen2.5-3b	Alibaba	China
21	Qwen2.5-1.5b	Alibaba	China

#	Model	Company	Country
22	Gemma3-1b	Google	USA
23	Llama3.2-1b	Meta	USA
24	Qwen2.5-0.5b	Alibaba	China
25	Gemma3-270m	Google	USA

Table 4– Models used

4.3. Experimental Setup

In order to accelerate the process, we ran the models simultaneously on different computers. However, in all cases, these consisted of consumer-grade computers with varying levels of performance. These were: MacBook M1 16GB RAM; Mac Mini M4 24GB RAM; PC Ryzen 7 64GB RAM RTX 3060 8GB; PC Ryzen 7 128GB RAM RTX 3090 24GB. To run the models, we used Ollama.com since it allows easy downloading and simple management from a simple Python script. Ollama offers a large number of versions of each model, depending on the type and quantization. Regarding types, having discarded thinking and coding, we worked with the rest, especially prioritizing "instruct" over "text." This is important because our objective was not text creation but simple decision-making.

As for quantization, we tried to work with the highest resolution, which for Ollama is 16-bit. Only in some cases did we have to go down to 8-bit, but never beyond that. LLM model quantization consists of reducing the numerical precision with which model weights (the values that define its internal connections) are represented to make it lighter and faster without losing too much quality.

The largest local models have been 70 billion parameters, and (as we will see in the conclusions) it has been sufficient. Both the analysis script, the dataset, and the LLM were ran on the same local computer. It launched the prompt shown in the Appendix 2. We compared the results of each LLM with our reference dataset in order to assess their performance.

4.4. Performance Evaluation Metrics

We used three standard metrics to assess the performance of gender classification models derived from the confusion matrix were used: precision, recall, and the F1-Score indicator (Powers 2011; Sokolova and Lapalme 2009). These metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where TP (True Positives) represents correctly classified instances, FP (False Positives) represents instances incorrectly assigned to a class, and FN (False Negatives) represents instances that should have been assigned to a class but were not. Precision (also known as positive predictive value) measures the proportion of correct positive predictions over the total positive predictions made by the model, quantifying the classifier's ability to avoid false positives. Recall (or sensitivity) calculates the proportion of positive instances correctly

identified over the total actual positive instances in the dataset, reflecting the model's ability to detect all cases belonging to each class. While high precision indicates that the model makes few false positives, high recall indicates that the model detects most relevant cases while minimizing false negatives.

The F1-Score is the harmonic means of precision and recall, providing a single metric that balances both aspects of classifier performance (Rijsbergen 1979). This measure is especially useful when there is an imbalance between classes or when both false positives and false negatives have significant costs. The F1-Score reaches its maximum value of 1.0 when both precision and recall are perfect, and low values indicate poor performance in at least one of these dimensions. Although accuracy (the proportion of correctly classified instances over the total) is commonly used in classification tasks, it can be misleading when classes are imbalanced; therefore, we prioritize precision, recall, and F1-Score as more informative metrics for this study.

In this study, these metrics were calculated both for each individual class (Man, Woman) and as macro averages (simple average between classes) and weighted averages (weighted average by the support of each class), allowing a comprehensive evaluation of each model's performance under different classification scenarios. Additionally, we analyze the abstention rate, defined as the proportion of cases where the model refrained from inferring a gender providing an "Unknown" value. This metric is relevant because it reflects the model's confidence in its predictions: a high abstention rate may indicate either appropriate caution when facing ambiguous names or excessive uncertainty in the classification process. The trade-off between abstention rate and classification accuracy provides insights into the practical deployment suitability of each model.

5. Results

5.1. Precision Analysis

Figure 3 displays a dumbbell graph visualizing precision, recall and F1 metrics across all 25 evaluated models, stratified by gender and ordered by family groups. Regarding precision, the analysis reveals robust precision (>0.90) for the majority of architectures exceeding the 7-billion parameter threshold. Notably, efficiency-focused models such as Qwen2.5-3b (1.000/0.990) and Qwen2.5-14b (0.995/0.981) achieved near-optimal results, rivaling massive architectures like Llama3.1-70b (0.991/0.922). These high precision scores indicate that false positives are minimal; when these models assign a gender label, the classification is highly reliable.

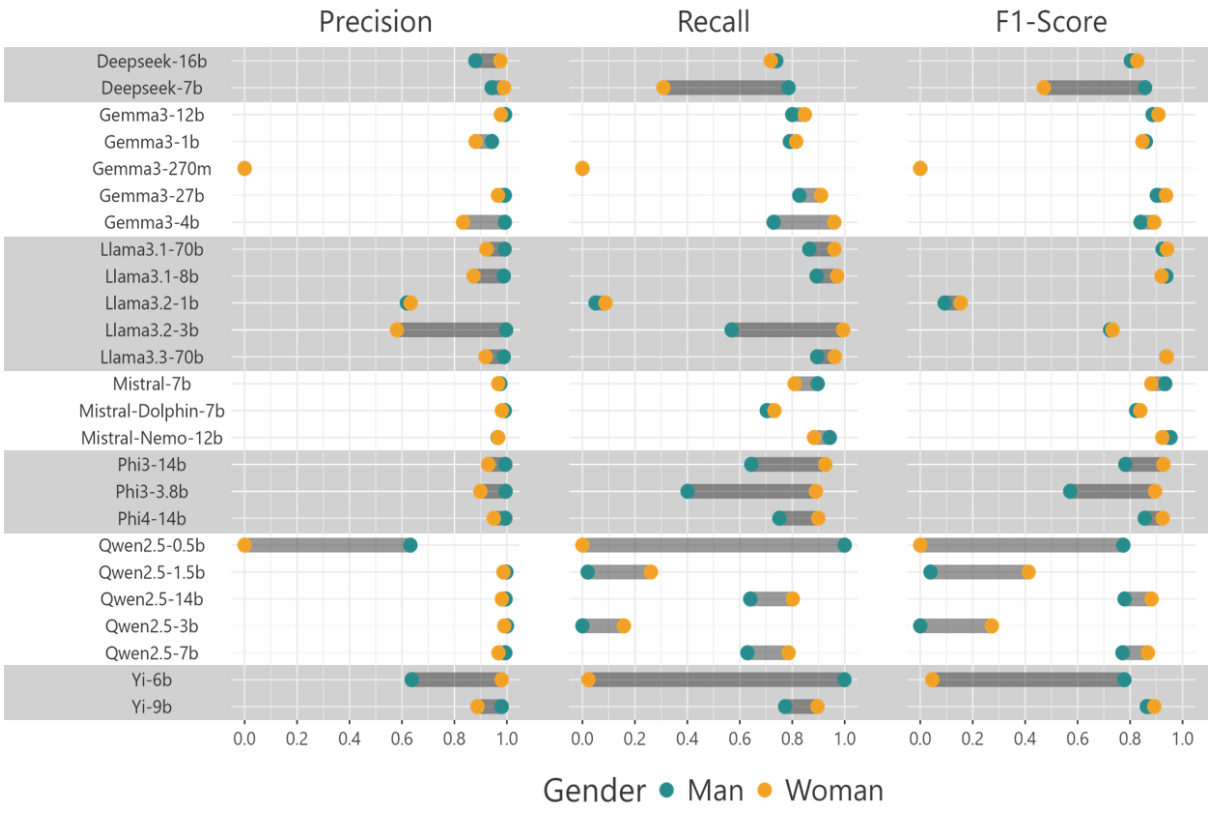

Figure 3 – Precision, recall and F1 score data by gender and families.

A systematic asymmetry emerges in the precision values: models consistently exhibit higher precision for the man category compared to the woman category. This differential is particularly pronounced in models like Llama3.1-70b (0.991 vs. 0.922), Llama3.1-8b (0.988 vs. 0.873), and Phi4-14b (0.994 vs. 0.950). An anomalous case is observed with Yi-6b, which exhibits inverted precision values (0.637 for Man, 0.980 for Woman), suggesting an unusual bias pattern distinct from other models in its size category. This pattern suggests a tendency toward more conservative classification of women names, potentially reflecting imbalances in the training data or inherent model biases. The smallest models (Gemma3-270m, Qwen2.5-0.5b, Llama3.2-1b) demonstrate severely degraded performance, with precision values approaching or equal to zero. This may indicate that these models lack the parametric capacity necessary for reliable gender classification tasks.

5.2. Recall Analysis

Figure 3 also displays the recall (sensitivity) values for all models. The recall metric quantifies the proportion of actual positive cases correctly identified by each classifier. In contrast to the precision patterns observed with the precision, recall values exhibit an inverse gender asymmetry: models generally achieve higher recall for the woman category than for the man category. This is evident in top-performing models such as Llama3.1-70b (0.865 vs. 0.961), Llama3.3-70b (0.895 vs. 0.962), and Llama3.1-8b (0.893 vs. 0.971). This complementary pattern, when considered alongside precision differences, indicates a systematic classification bias: models tend to classify ambiguous cases as man, resulting in higher man precision but lower man recall, and conversely for woman classifications.

The recall distribution shows substantially greater heterogeneity compared to precision. While precision remains consistently high across most medium-to-large models, recall values span a broader range (approximately 0.6-0.95 for functional models). Yi-6b demonstrates a pathological classification pattern with near-perfect man recall (0.999) but virtually zero woman recall (0.023), indicating that this model classifies almost all names as man regardless of actual gender. Similarly, Deepseek-llm-7b shows severely impaired woman recall (0.309), suggesting substantial gender bias in this model's decision boundaries. The smallest models (Qwen2.5-3b, Qwen2.5-1.5b, Llama3.2-1b, Gemma3-270m) exhibit near-zero recall for one or both categories, confirming their functional inadequacy for this classification task.

5.3. F1-Score Analysis

Figure 3 also presents the F1-Score, providing a balanced assessment of model performance that penalizes extreme trade-offs between these complementary metrics. This composite measure is particularly valuable for identifying models that achieve robust performance across both dimensions. The highest-performing models demonstrate F1-Scores exceeding 0.90 for both gender categories. Llama3.3-70b (0.939/0.940), Llama3.1-70b (0.924/0.941), and Gemma3-27b (0.902/0.937) exhibit the most balanced and robust performance profiles. These results suggest that models at or above approximately 27 billion parameters achieve near-optimal classification capability for this task.

The F1-Score effectively reveals models with severe gender bias that might appear acceptable when examining precision or recall in isolation. Yi-6b demonstrates this clearly: while achieving reasonable man F1-Score (0.778), its woman F1-Score collapses to 0.046, exposing its extreme bias toward man classification. Similarly, Deepseek-llm-7b shows acceptable man performance (0.857) but poor woman F1-Score (0.471). A notable finding emerges regarding the Gemma model family. Gemma3-1b achieves remarkably balanced F1-Scores (0.860/0.847) despite its small parameter count, substantially outperforming other models in the 1-3 billion parameter range. This suggests that architectural innovations or training data quality in the Gemma family may partially compensate for reduced model capacity.

Among models below the 27-billion parameter threshold, Mistral-Nemo-12b stands out with exceptional performance (0.954/0.923), achieving F1-Scores comparable to or exceeding those of significantly larger models. This positions Mistral-Nemo-12b as a particularly efficient choice for gender classification tasks where computational resources are constrained but high accuracy is required. Models below approximately 3 billion parameters generally fail to achieve usable F1-Scores, with several showing complete failure (F1 = 0.0) for one or both gender categories. The threshold for minimally viable performance appears to be approximately 4-7 billion parameters for most model families.

5.4. Relationship Between Model Size and Performance

Figure 4 presents a scatter plot examining the relationship between model size (measured in billions of parameters, logarithmic scale on x-axis) and overall F1-Score performance. Points are color-coded by country of origin (China: red, France: green, Taiwan: yellow, USA: blue), enabling analysis of potential geographic or organizational patterns in model development.

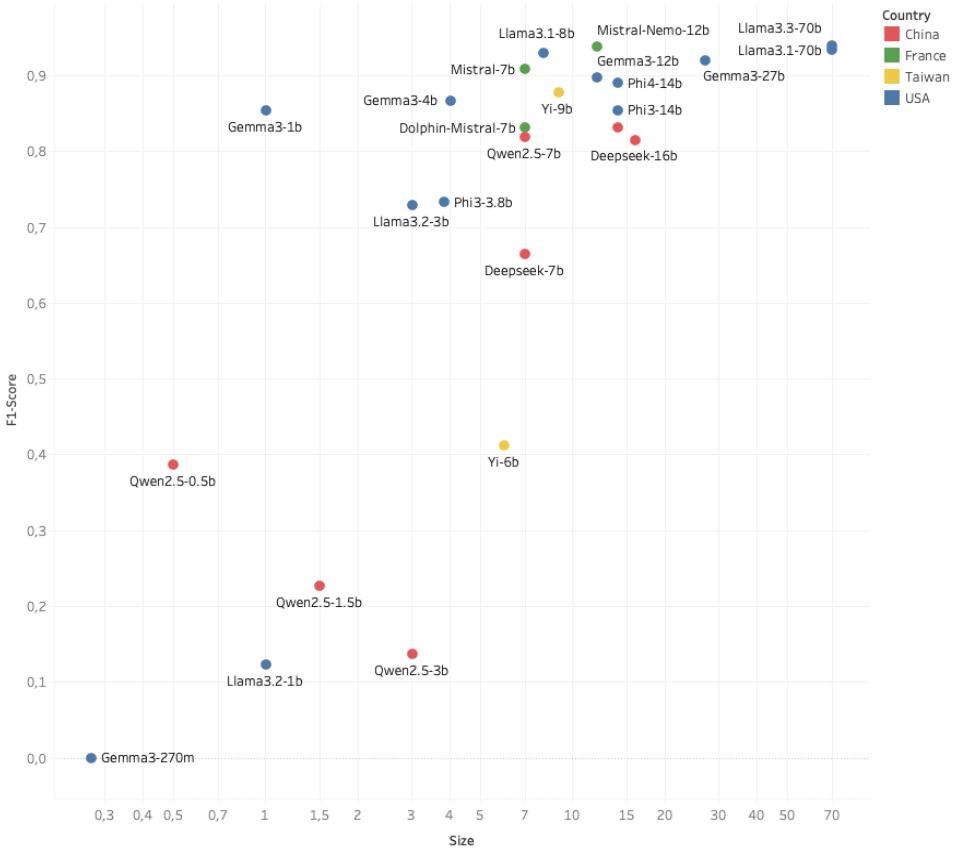

Figure 4 – Relationship between F1-Score and model size

We observe a logarithmic relationship between model size and performance. Below approximately 3 billion parameters, performance varies dramatically and unpredictably, with F1-Scores ranging from near-zero (Gemma3-270m, Llama3.2-1b) to surprisingly competent (Gemma3-1b at approximately 0.85). This high variance zone suggests that small model performance is highly dependent on specific architectural choices and training methodologies rather than scale alone.

A critical threshold emerges at approximately 7 billion parameters, above which all evaluated models achieve F1-Scores exceeding 0.80, with most clustering between 0.85 and 0.95. Importantly, increasing model size beyond this threshold yields diminishing returns: the 70-billion-parameter models (Llama3.1-70b, Llama3.3-70b) achieve only marginally better performance than well-optimized 7-12 billion parameter models (Llama3.1-8b, Mistral-Nemo-12b, Mistral-7b).

The country-of-origin analysis reveals no strong systematic bias favoring models from regions among the well-performing models. USA-based models (blue) span the full performance range, and French models (Mistral family, green) cluster in the high-performance zone. However, a subtle pattern emerges with China-based models (red), which tend to position slightly below their USA and French counterparts of similar size. Models such as Qwen2.5-7b, Deepseek-7b, and Deepseek-16b achieve lower F1-Scores compared to similarly sized models from other regions, suggesting potential differences in training data coverage for international names or optimization priorities. The Taiwan-based Yi models show inconsistent performance, with Yi-9b achieving good results while Yi-6b performs poorly.

Notably, the Gemma3-1b model represents a significant outlier, achieving performance competitive with models 4-7 times its size. This exceptional efficiency suggests potential

applications for resource-constrained deployment scenarios where computational cost or latency are primary concerns.

5.5. Abstention Rate versus Classification Accuracy

Figure 5 examines the trade-off between classification confidence and accuracy by plotting F1-Score against abstention rate (the proportion of cases where the model responded "Unknown" rather than committing to a gender classification). The axis scales have been deliberately adjusted to exclude the smallest models that exhibited complete classification failure (that is, Yi-6b, Qwen2.5-3b, Qwen2.5-1.5b, Llama3.2-1b, Qwen2.5-0.5b and Gemma3-270m), allowing for more detailed examination of the functional models and their performance trade-offs.

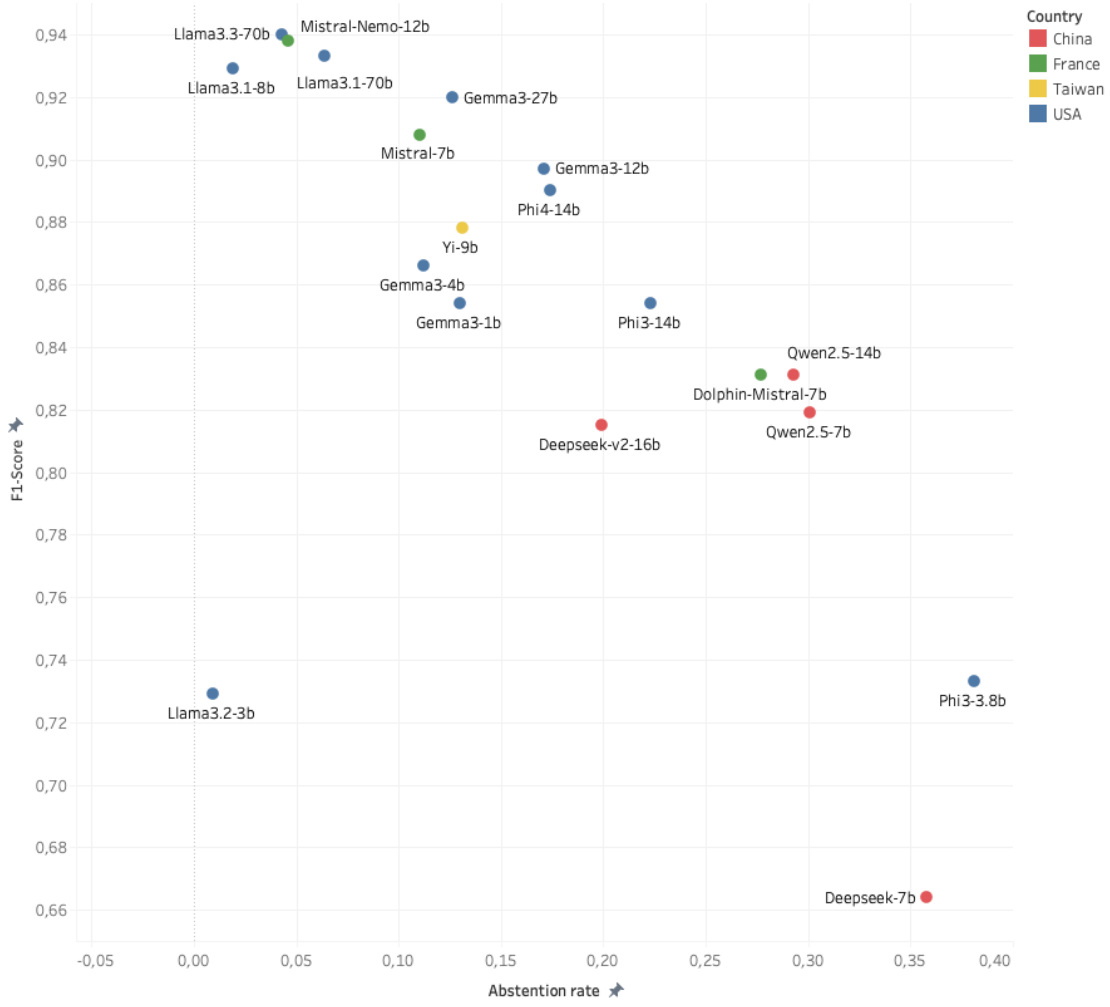

Figure 5 – Abstention rate versus F1-Score

The optimal operating region lies in the upper-left quadrant: high F1-Score combined with low abstention rate, indicating confident and accurate classification. Models occupying this region include Llama3.3-70b (F1 ≈ 0.94, abstention ≈ 3%), Llama3.1-8b (F1 ≈ 0.93, abstention < 5%), and Mistral-Nemo-12b (F1 ≈ 0.94, abstention ≈ 5%). These models demonstrate the capacity to make accurate predictions without excessive reliance on the "Unknown" escape category. A negative correlation between abstention rate and F1-Score is not uniformly observed, contradicting the intuition that more selective models should achieve higher accuracy on

committed predictions. Deepseek-7b exemplifies this: despite the highest abstention rate (approximately 35%), it achieves the lowest F1-Score (approximately 0.66) among models in this analysis. Similarly, Phi3-3.8b shows high abstention (approximately 38%) with only moderate F1-Score (approximately 0.73). These patterns suggest that high abstention in these models reflects classification uncertainty rather than strategic conservatism.

Chinese-origin models (red points) tend toward higher abstention rates compared to USA and French models, potentially reflecting different optimization objectives or training data characteristics. However, this pattern requires further investigation to distinguish between systematic differences and model-specific effects. The analysis suggests that for practical deployment, models from the Llama family (particularly Llama3.1-8b and Llama3.3-70b) offer the best combination of accuracy and classification coverage.

6. Discussion

This section addresses each of our research questions in light of the empirical results presented above, contextualizing our findings within the existing literature on LLM-based gender classification and gender bias in language models.

RQ1: In general, are LLMs efficient for gender identification based on the simple combination of first name + country?

Our results demonstrate that local LLMs are indeed efficient for name-based gender classification, with top-performing models achieving F1-Scores exceeding 0.93 for both gender categories. This finding aligns with AlNuaimi et al. (2024), who reported accuracy rates above 90% using various LLMs in zero-shot mode. However, our results are substantially higher than those reported by Zhuang et al. (2024), who found accuracy rates of only 56-68% with ChatGPT3.5 and ChatGLM-6b on GitHub author names. This discrepancy may be attributed to differences in dataset composition (their sample contained a large proportion of Chinese names) and the specific models evaluated. Our findings also exceed the 86% accuracy reported by Deng et al. (2025) for Chinese names using eight Chinese LLMs, suggesting that the models in our study generalize better across international name distributions. The combination of first name and country proves to be a sufficient input for reliable gender classification when using appropriately sized models.

RQ2: Do classification metrics (precision, recall, F1-score) show systematic differences between men and women categories?

Our analysis reveals consistent and systematic asymmetry in classification performance across genders. Models exhibit higher precision for man names but higher recall for woman names, indicating a tendency to classify ambiguous cases as man. This pattern results in more false positives for the man category and more false negatives for the woman category. These findings are consistent with broader research on gender bias in LLMs. Kotek et al. (2023) demonstrated that LLMs are 3-6 times more likely to choose occupations that stereotypically align with a person's gender, reflecting biases embedded in training data. He (2025) found that ChatGPT models show bias toward recommending papers by man authors, while Claude exhibits bias toward the dominant gender. Our results suggest that similar biases manifest in name-based gender classification, where models appear more conservative when assigning woman labels. This systematic bias has important implications for bibliometric studies, as it may lead to underestimation of woman representation in research that uses such data.

RQ3: Does model size matter? Are results proportional to the number of parameters?

Our findings reveal a clear but non-linear relationship between model size and classification performance. Below approximately 3 billion parameters, performance is highly variable and often inadequate. A critical threshold emerges at approximately 7 billion parameters, above which all models achieve acceptable performance (F1 > 0.80). Importantly, scaling beyond this threshold yields diminishing returns: 70-billion-parameter models achieve only marginally better results than well-optimized 7-12 billion parameter models. This observation partially aligns with established scaling laws (Kaplan et al. 2020), which describe power-law relationships between model size and performance. However, our task-specific results suggest that for gender classification, there exists a practical ceiling beyond which additional parameters provide minimal benefit. This finding has significant practical implications: researchers can achieve near-optimal performance using 7-12B parameter models that are substantially more economical to deploy than larger alternatives. The exceptional performance of Gemma3-1b further demonstrates that architectural innovations and training data quality can partially compensate for reduced model size, consistent with Microsoft's findings on the Phi family (Microsoft 2024).

RQ4: Is there any bias based on the LLM's country of origin?

Our analysis reveals a subtle but consistent pattern: China-based models (Qwen, Deepseek) tend to achieve slightly lower F1-Scores compared to similarly sized models from the United States and France. This difference is most apparent in the 7-16 billion parameter range. Several factors may explain this observation. First, training data composition likely differs across regions, with Chinese models potentially emphasizing Chinese and East Asian names at the expense of Western name coverage. Second, optimization priorities may vary: Chinese models are often designed for bilingual (Chinese and English) performance, which may involve trade-offs affecting name classification in other linguistic contexts. Third, the models from French company Mistral, despite being smaller, consistently achieve high performance, suggesting that focused architectural innovations and training strategies can overcome resource limitations. However, we note that Taiwan-based Yi models show inconsistent performance (Yi-9b performs well while Yi-6b fails catastrophically), indicating that country of origin alone is not deterministic. These patterns require further investigation with controlled experiments to distinguish systematic regional differences from model-specific effects.

RQ5: Which LLMs offer the best balance between precision and recall?

Based on our comprehensive evaluation, three models stand out for their balanced performance. Llama3.3-70b achieves the highest overall F1-Scores (0.939/0.940 for man/woman) with minimal gender disparity, representing the best choice when computational resources are not constrained. However, Mistral-Nemo-12b achieves virtually identical performance (F1-Scores of 0.954/0.923) while requiring approximately five times less computational resources, the difference between running on a high-end consumer GPU versus requiring professional-grade hardware. This makes Mistral-Nemo-12b the optimal choice for most practical applications, offering top-tier accuracy at a fraction of the computational cost. Gemma3-1b represents a remarkable outlier, achieving balanced F1-Scores (0.860/0.847) with only 1 billion parameters, substantially outperforming other models in its size category. For practical bibliometric applications, we recommend Mistral-Nemo-12b as the default choice balancing accuracy, efficiency, and gender equity. Researchers processing very large datasets with strict resource constraints may consider Gemma3-1b as a viable alternative.

This study is not free from limitations. First, our evaluation focuses exclusively on local open-weight models, and while our dataset provides substantial coverage, it may not fully represent

all cultural and linguistic contexts. Moreover, data on gender is binary, and this does not fully capture all gender realities. Furthermore, we evaluated models in zero-shot mode; few-shot prompting strategies might yield different performance patterns. Future research will extend this work by analyzing classification performance disaggregated by country and region of origin of the names, which will help identify specific cultural contexts where models excel or struggle. We also plan to incorporate recently released models such as Kimi K2 to assess whether the latest architectural innovations further improve gender classification accuracy.

7. Conclusions

This study has evaluated the performance of 25 local LLMs for name-based gender classification, using a dataset of nearly 200,000 names across multiple countries. Given our analysis, we can conclude the following. First, local LLMs are highly effective for gender identification based on the combination of first name and country, with top-performing models achieving F1-Scores above 0.93. This positions them as a viable and cost-effective alternative to commercial gender detection tools for large-scale bibliometric studies. Second, all evaluated models exhibit a systematic gender bias: higher precision for man names combined with higher recall for woman names. This asymmetry indicates that models tend to classify ambiguous cases as man, which may lead to underestimation of woman representation in analyses that rely on names to infer gender. Researchers should be aware of this limitation when interpreting results.

Third, model size matters, but only up to a point. A critical threshold exists at approximately 7 billion parameters, below which performance is unreliable. However, scaling beyond 12-14 billion parameters yields diminishing returns, making mid-sized models the most cost-effective choice for this task. Fourth, China-based models show slightly lower performance on international name datasets compared to US and French models of similar size, likely reflecting differences in training data composition. However, country of origin is not deterministic, as demonstrated by the inconsistent performance of Taiwan-based Yi models.

Finally, Mistral-Nemo-12b emerges as the optimal choice for most applications, achieving performance equivalent to Llama3.3-70b while requiring approximately five times fewer computational resources. For extremely resource-constrained scenarios, Gemma3-1b offers surprisingly competitive performance with only 1 billion parameters.

Acknowledgements

This work is part of the research project of the Spanish Ministry of Science Innovation and Universities: “Artificial Intelligence in Europe: Rise or Decline?” (PID2023-149646NB-I00), funded by MICIU/AEI/10.13039/501100011033/ Knowledge Generation Projects 2023, and of the STITCH project (Ref: PID2024-155412OB-I00). EGS is supported by a FPU grant from the Spanish Ministry of Science (Ref: FPU2021/02320) and NRG is currently supported by a Ramón y Cajal from the Spanish Ministry of Science (Ref: RYC2019-027886-I).

CRediT authorship contribution statement

Conceptualization	VHS, NRG
Data curation	EGS

Formal analysis	
Funding acquisition	VHS, NRG
Investigation	VHS, EGS
Methodology	VHS, EGS, NRG
Project administration	
Resources	VHS
Software	VHS, EGS
Supervision	NRG
Validation	
Visualization	VHS, EGS
Writing – original draft	VHS, EGS
Writing – review & editing	VHS, EGS, NRG

8. References

AlNuaimi, Khaled, Gautier Marti, Mathieu Ravaut, Abdulla Alketbi, Andreas Henschel, and Raed Jaradat. 2024. "Enriching Datasets with Demographics through Large Language Models: What's in a Name?" Arxiv. https://doi.org/10.48550/arXiv.2409.11491.

Alzahrani, Esam, Mohammed Al Qurashi, and Leon Jololian. 2022. "Comparative Analysis of the Use of Pre-Trained Models to Profile Authors' Ages and Genders." In 2022 2nd International Conference on Computing and Machine Intelligence, ICMI 2022 - Proceedings. https://doi.org/10.1109/ICMI55296.2022.9873677.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni. 2003. "Gender, genre, and writing style in formal written texts". To appear in Text, 23 (3): 321-346. https://doi.org/10.1515/text

Bérubé, Nicolas, Gita Ghiasi, Maxime Sainte-Marie and Vincent Larivière. 2020. "Wiki-Gendersort: Automatic gender detection using first names in Wikipedia". https://api.semanticscholar.org/CorpusID:241993757

Boekhout, Hanjo, Inge van der Weijden, and Ludo Waltman. 2021. "Gender Differences in Scientific Careers: A Large-Scale Bibliometric Analysis". Arxiv. https://arxiv.org/pdf/2106.12624

Campbell, Sara E, and Daniel Simberloff. 2022. "The Productivity Puzzle in Invasion Science: Declining but Persisting Gender Imbalances in Research Performance". BioScience 72(12): 1220–1229. https://doi.org/10.1093/biosci/biac082

Cyrenne, Philippe, and Alan Chan. 2012. 'High School Grades and University Performance: A Case Study'. Economics of Education Review 31 (5): 524–42. https://doi.org/10.1016/j.econedurev.2012.03.005.

Domínguez-Díaz, Adrián, Manuel Goyanes, Luis de-Marcos, and Víctor Pablo Prado-Sánchez. 2024. "Comparative Analysis of Automatic Gender Detection from Names: Evaluating the Stability and Performance of ChatGPT versus Namsor, and Gender-API." PeerJ Computer Science 10: e2378. https://doi.org/10.7717/peerj-cs.2378.

Deng, Xinyu, Hui Xu, Zihui Li, Huiwen Bai, Lanfeng Ni, and Chengzhi Zhang. 2025. "Gender Differences in Research Methods: Insights from Chinese Humanities and Social Sciences PhD Dissertations." In 20th International Conference on Scientometrics & Informetrics. https://doi.org/10.51408/issi2025_014.

DeepSeek-AI. 2024. "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism." ArXiv Preprint. https://doi.org/10.48550/arXiv.2401.02954.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, et al. 2024. "Gemma: Open Models Based on Gemini Research and Technology." ArXiv Preprint. https://doi.org/10.48550/arXiv.2403.08295.

González-Salmón, E., Chinchilla-Rodríguez, Z., Nane, G. F., & Robinson García, N. (2024). What contributes to gender parity in science? A Bayesian Network analysis. International Conference on Science, Technology and Innovation Indicators, 2024. https://zenodo.org/records/12609270

González-Salmón, Elvira, and Nicolas Robinson-Garcia. 2024a. 'A Call for Transparency in Gender Assignment Approaches'. Scientometrics 129 (4): 2451–54. https://doi.org/10.1007/s11192-024-04995-4.

González-Salmón, Elvira and Nicolas Robinson-García. 2024b. "WikiGenDex: Un nuevo algoritmo de identificación de género basado en fuentes abiertas". Infonomy, 2(1). https://doi.org/10.3145/infonomy.24.010

González-Salmón, Elvira, Zaida Chinchilla-Rodríguez, and Nicolas Robinson-Garcia. 2025. 'The Woman Researcher's Tale: A Review of Bibliometric Methods and Results for Studying Gender in Science'. Journal of the Association for Information Science and Technology 76 (9): 1188–209. https://doi.org/10.1002/asi.25012.

He, Jiangen. 2025. "Who Gets Cited? Gender- and Majority-Bias in LLM-Driven Reference Selection." https://doi.org/10.48550/arXiv.2508.02740.

Huang, Junming, Alexander J. Gates, Roberta Sinatra, and Albert-László Barabása. 2020. "Historical Comparison of Gender Inequality in Scientific Careers across Countries and Disciplines". Proceedings of the National Academy of Sciences of the United States of America 117(9): 4609–4616. https://doi.org/10.1073/pnas.1914221117

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. "Mistral 7B." Arxiv, 1–9. https://doi.org/10.48550/arXiv.2310.06825.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." ArXiv Preprint. https://doi.org/10.48550/arXiv.2001.08361.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. 'Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods'. Proceedings of the 25th International

Conference Companion on World Wide Web - WWW '16 Companion, 53–54. https://doi.org/10.1145/2872518.2889385.

Kavuri, Karunakar, and M. Kavitha. 2023. "A Word Embeddings Based Approach for Author Profiling: Gender and Age Prediction." International Journal on Recent and Innovation Trends in Computing and Communication 11 (7 s): 239–50. https://doi.org/10.17762/ijritcc.v11i7s.6996.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. "Gender Bias and Stereotypes in Large Language Models." In Proceedings of the ACM Collective Intelligence Conference (CI 2023). https://doi.org/10.1145/3582269.3615599.

Kozlowski, Diego, Vincent Larivière, Cassidy R. Sugimoto and Thema Monroe-White. 2022. "Intersectional Inequalities in Science. Proceedings of the National Academy of Sciences" Proceedings of the National Academy of Sciences 119 (2): e2113067119. https://doi.org/10.1073/pnas.2113067119

Matias-Rayme, Nataly, Iuliana Botezan, Mari Carmen Suárez-Figueroa, and Rodrigo Sánchez-Jiménez. 2024. 'Gender Assignment in Doctoral Theses: Revisiting Teseo with a Method Based on Cultural Consensus Theory'. Scientometrics 129 (7): 4553–72. https://doi.org/10.1007/s11192-024-05079-z.

Microsoft. 2024. "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone." ArXiv Preprint 2. https://doi.org/10.48550/arXiv.2404.14219.

Mihaljević, Helena, Marco Tullney, Lucía Santamaría and Christian Steinfeldt. 2019. "Reflections on Gender Analyses of Bibliographic Corpora". Frontiers in Big Data, 2. https://doi.org/10.3389/fdata.2019.00029

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé and Franciska de Jong. 2016. "Computational Sociolinguistics: A Survey". Computational Linguistics, 42(3): 537–593. https://doi.org/10.1162/COLI_a_00258

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." Advances in Neural Information Processing Systems 35. https://doi.org/10.48550/arXiv.2203.02155.

Parasurama, Prasanna, and João Sedoc. 2021. "Gendered Language in Resumes – An Empirical Analysis of Gender Norm Violation and Hiring Outcomes." In 42nd International Conference on Information Systems, ICIS 2021 TREOs: "Building Sustainability and Resilience with IS: A Call for Action," 0–9. https://aisel.aisnet.org/icis2021/data_analytics/data_analytics/17.

Patel, Krishna, Nivedha Sivakumar, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. 2025. "Fairness Dynamics During Training" 1: 1–8. https://doi.org/10.48550/arXiv.2506.01709.

Powers, David M. W. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation." Journal of Machine Learning Technologies, 2 (1): 37–63. http://arxiv.org/abs/2010.16061.

Qwen Team. 2025. "Qwen2.5 Technical Report." ArXiv Preprint, 1–26. https://doi.org/10.48550/arXiv.2412.15115.

Rijsbergen, C.J. Van. 1979. Information Retrieval. 2nd ed. London: Butterworths.

Robinson-Garcia, Nicolas, Carmen Corona-Sobrino, Zaida Chinchilla-Rodríguez, Daniel Torres-Salinas and Rodrigo Costas. 2025. "The use of informetric methods to study diversity in the scientific workforce: A literature review". Quantitative Science Studies, 1-34. https://doi.org/10.1162/qss_a_00367

Sazzed, Salim. 2022. "Revealing the Demographic Attributes of the Authors from the Abstracts of Scientific Articles." In HT 2022: 33rd ACM Conference on Hypertext and Social Media - Co-Located with ACM WebSci 2022 and ACM UMAP 2022, 209–13. https://doi.org/10.1145/3511095.3536358.

Sokolova, Marina, and Guy Lapalme. 2009. "A Systematic Analysis of Performance Measures for Classification Tasks." Information Processing and Management 45 (4): 427–37. https://doi.org/10.1016/j.ipm.2009.03.002.

Sugimoto, Cassidy R., and Vincent Larivière. 2023. Equity for Women in Science. Harvard University Press.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023a. "LLaMA: Open and Efficient Foundation Language Models." https://doi.org/10.48550/arXiv.2302.13971.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023b. "Llama 2: Open Foundation and Fine-Tuned Chat Models." ArXiv Preprint. https://doi.org/10.48550/arXiv.2307.09288.

Tripodi, Francesca. 2023. "Ms. Categorized: Gender, notability, and inequality on Wikipedia". News media & Society 25 (7): 1687-1707.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." Advances in Neural Information Processing Systems 2017-Decem (Nips): 5999–6009. https://doi.org/10.48550/arXiv.1706.03762.

Wildgaard, Lorna, Jesper W. Schneider and Birger Larsen. 2014. "A review of the characteristics of 108 author-level bibliometric indicators". Scientometrics, 101(1): 125-158. https://doi.org/10.1007/s11192-014-1423-3

You, Zhiwen, Hae Jin Lee, Shubhanshu Mishra, Sullam Jeoung, Apratim Mishra, Jinseok Kim, and Jana Diesner. 2024. "Beyond Binary Gender Labels: Revealing Gender Biases in LLMs through Gender-Neutral Name Predictions." GeBNLP 2024 - 5th Workshop on Gender Bias in Natural Language Processing, Proceedings of the Workshop, 255–68. https://doi.org/10.18653/v1/2024.gebnlp-1.16.

Young, Alex, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, et al. 2024. "Yi: Open Foundation Models by 01.AI." ArXiv Preprint. https://doi.org/10.48550/arXiv.2403.04652.

Zhuang, Yuqian, Mingya Zhang, Yiyuan Yang, and Liang Wang. 2024. "Analyzing Women's Contributions to Open-Source Software Projects Based on Large Language Models." Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, 2363–68. https://doi.org/10.1109/CSCWD61410.2024.10580385.

Zorell, Carolin V., and Thomas Denk. 2021. 'Political Consumerism and Interpersonal Discussion Patterns'. Scandinavian Political Studies 44 (4): 392–415. https://doi.org/10.1111/1467-9477.12204.

9. Appendix

Appendix 1.

```
SELECT ?personLabel ?givenNameLabel ?countryLabel
WHERE {
  ?person wdt:P31 wd:Q5 ;
        wdt:P21 wd:Q6581072 ;
        wdt:P27 ?country ;
        wdt:P735 ?givenName.
  ?person rdfs:label ?personLabel .
  ?country rdfs:label ?countryLabel .
  ?givenName rdfs:label ?givenNameLabel .
  FILTER(LANG(?personLabel)='en')
  FILTER(LANG(?countryLabel)='en')
  FILTER(LANG(?givenNameLabel)='en')
  FILTER(?country = wd:Q148)
}
```

Appendix 2.

```
prompt = f"""You are an expert bibliometrician helping researchers conduct
a gender study of scientific production.

Analyze the first name "{first_name}" from {country} and classify it as one
of these three options:
- "Man" if the name is typically masculine
- "Woman" if the name is typically feminine
- "Unknown" if the name is truly ambiguous or you cannot determine the
gender with reasonable confidence

IMPORTANT RULES:
- You MUST respond with ONLY one word: "Man", "Woman", or "Unknown"
- Do not include any explanations or additional text
- Only use "Unknown" when you genuinely cannot determine the gender

Name to analyze: {first_name}
Country: {country}"""
```