

For the Misgendered Chinese in Gender Bias Research: Multi-Task Learning with Knowledge Distillation for Pinyin Name-Gender Prediction

Xiaocong Du, Haipeng Zhang*

ShanghaiTech University

{duxc2023, zhanghp}@shanghaitech.edu.cn

Abstract

Achieving gender equality is a pivotal factor in realizing the UN’s Global Goals for Sustainable Development. Gender bias studies work towards this and rely on name-based gender inference tools to assign individual gender labels when gender information is unavailable. However, these tools often inaccurately predict gender for Chinese Pinyin names, leading to potential bias in such studies. With the growing participation of Chinese in international activities, this situation is becoming more severe. Specifically, current tools focus on pronunciation (Pinyin) information, neglecting the fact that the latent connections between Pinyin and Chinese characters (Hanzi) behind convey critical information. As a first effort, we formulate the Pinyin name-gender guessing problem and design a Multi-Task Learning Network assisted by Knowledge Distillation that enables the Pinyin embeddings in the model to possess semantic features of Chinese characters and to learn gender information from Chinese character names. Our open-sourced method surpasses commercial name-gender guessing tools by 9.70% to 20.08% relatively, and also outperforms the state-of-the-art algorithms.

1 Introduction

Gender equality is an important pillar of the UN’s Global Goals for Sustainable Development and numerous gender bias studies work towards this goal. For these studies, not all times is gender information of the target population available, and employing a name-based gender inference tool to assign gender labels based on individual names is a popular solution. Such tools have been widely applied in various contexts, such as understanding the impact of gender composition on teams [Yang *et al.*, 2022a] and revealing gender inequalities among scientists [Larivière *et al.*, 2013; Huang *et al.*, 2020], academic publications [Dworkin *et al.*, 2020; Squazzoni *et al.*, 2021], and in faculty positions [Spoon *et al.*, 2023], as shown in Tabel 1.

*Corresponding author.

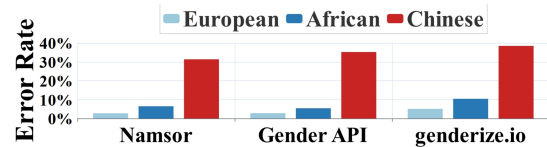


Figure 1: The gender classification error rates of the commercial gender detection tools for names of different origins.

However, as evidenced in Figure 1, popular name-gender guessing tools, namely Namsor¹, Gender API², and genderize.io³, perform poorly for Chinese names in Pinyin, which represents the romanized pronunciations of corresponding Chinese characters (Hanzi). The error rate in gender classification for Pinyin names is 7 to 12 times higher than that for European names and for the web service genderize.io, the error rate is as high as 38.39%. Some studies [Dworkin *et al.*, 2020; Squazzoni *et al.*, 2021; Yang *et al.*, 2022a] use these results for the Chinese in their datasets, while some [Huang *et al.*, 2020] decide to exclude these samples representing Chinese people from the analysis. Both practices seem problematic: if Chinese Pinyin results are kept, the low accuracy would shake the reliability of the analysis; if they are discarded, there is inevitably a bias since Chinese people are underrepresented. This problem can be more severe, as Chinese people take up considerably large portions of the datasets and there is a growing trend – for instance, in 2002, Chinese authors write 4.72% of international papers and this ratio grows to 19.70% in 2022⁴. Manual verification by searching for scientists’ homepages may produce accurate results [Larivière *et al.*, 2013], while it is not practical for datasets containing millions of individuals without homepages.

In most Western languages, gender information is encapsulated within the names that consist of letters. In languages such as English or German, there exist distinct lists of male and female names, and the gender tendency of a name can be easily computed as the ratios of males/females under that name. In languages such as Italian, gender can be marked on a specific part of a name (e.g., suffixes like ‘a’ for females

¹<https://namsor.app/>

²<https://gender-api.com/>

³<https://genderize.io/>

⁴<https://www.scimagojr.com/countryrank.php>

Study	Venue	Data Source	Gender Inferring Tool	Pinyin names excluded?
Huang <i>et al.</i> [2020]	PNAS	Web of Science	genderize.io	Yes
Dworkin <i>et al.</i> [2020]	Nature Neuroscience	Web of Science	genderize.io	No
Squazzoni <i>et al.</i> [2021]	Science Advances	145 journals	Gender API	No
Yang <i>et al.</i> [2022a]	PNAS	MAG	Namsor	No
Larivière <i>et al.</i> [2013]	Nature	Web of Science	manually annotated	No

Table 1: Instances of gender bias research that infer genders from romanized names. MAG stands for Microsoft Academic Graph.

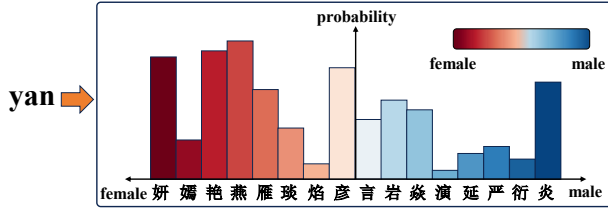


Figure 2: The possible correspondences between Pinyin and Chinese characters, as well as the gender information embedded in Chinese characters names, take ‘yan’ as an example. All the Chinese characters above have the same Pinyin ‘yan’.

or ‘o’ for males). The former is categorized as the conventional type of gender marking languages, while the latter is classified as the formal type [Ackermann and Zimmer, 2021].

Pinyin only represents the pronunciations of the Chinese characters or in other words, Chinese characters are converted into Pinyin. Pronunciations, convey some helpful information for gender guessing [Pan *et al.*, 2023; Jia and Zhao, 2019] but we have to note that the Chinese characters behind, are more informative [Pan *et al.*, 2023]. If it were one-to-one mapping from Pinyin to Chinese characters, it would be a (almost) solved problem since we can infer genders from names in Chinese characters with state-of-the-art tools such as CHGAT [Pan *et al.*, 2023]. Unfortunately, Pinyin can be associated with different Chinese characters based on varying probabilities, as shown in Figure 2.

Chinese represents the third type separated from the two aforementioned language types, namely, the semantic type. In Chinese names, gender is indicated by the semantics of the name. For example, at the left end of the horizontal axis in Figure 2, the character ‘妍’ connotes ‘beautiful’, aligning with the female gender stereotype, hence it is commonly used in female names. Conversely, at the other end, ‘炎’ signifies ‘flame of fire’, which relates to the male gender stereotype, making it prevalent in male names. Characters ‘妍’ and ‘炎’ exhibit distinct gender tendencies. However, when they are converted to their Pinyin forms ‘yan’, they become indistinguishable. Pinyin only represents the pronunciation of the characters, resulting in the loss of the tonal, symbolic, and semantic information. This transition leads to the loss of gender information embedded in the Chinese names. Given that characters are pivotal in inferring gender in Chinese names, previous methods [Hu *et al.*, 2021; Van Buskirk *et al.*, 2023] overlook the potential link between Pinyin and Chinese characters.

We therefore want to model the uncertainty in the con-

nections between Pinyin and Chinese characters and utilize the gender information in Pinyin and more importantly, Chinese characters. A straightforward approach to establish the Pinyin-character connection is converting Pinyin into the most possible Chinese character. However, this results in the loss of information from other possible characters. Instead, we introduce a Pinyin-to-character prediction task as an intermediate auxiliary task to learn the associations between Chinese characters and Pinyin, as well as the relationships among different characters, particularly those sharing the same Pinyin. Since our main task is Pinyin name-gender prediction task, we employ Multi-Task Learning (MTL) to integrate two tasks into a cohesive framework. By allowing two tasks to share encoder parameters, the model can learn external knowledge from the supervision signals of characters and generate characters enhanced embeddings. Consequently, the Pinyin embeddings learned from multi-task learning essentially integrate the information of all Chinese characters associated with that Pinyin, while preserving the one-to-many relationship between Pinyin and Chinese characters.

The MTL module mainly captures the Pinyin gender tendency and the association between Pinyin and Chinese characters, represented by the vertical dimension in Figure 2. However, the gender-related information embedded within the Chinese characters, i.e., the horizontal dimension in Figure 2, remains underexplored. Though we can assign binary gender labels to the embeddings learned from the MTL step, this only provides gender information at sample level and does not encompass the overall gender tendency of the Chinese characters behind each Pinyin. The gender tendency of each character can be learned by training a Chinese character name-gender prediction model. We then want to seamlessly transfer this learned information to the embeddings that model the Pinyin-character connections from MTL. This can be naturally done by Knowledge Distillation that encapsulates the knowledge learned from a teacher model (i.e., Chinese character name-gender prediction model here) in the form of soft targets, which are used to supervise the training of a student model (i.e., Pinyin name-gender prediction model here). By performing feature and response based Knowledge Distillation between two models, the gender tendency in Pinyin embeddings becomes clearer under the guidance of gender-specific characteristics of Chinese character names.

We summarize our contributions as follows:

1. We bring the problem of misgendering or excluding Chinese individuals in current gender bias studies to attention, which is attributed to the poor performance of commercial name-gender inference tools designed for West-

ern names. Accordingly, we define the new and important task of Chinese Pinyin name-gender prediction.

2. We develop a knowledge distillation assisted multi-task learning network for this task to model the connection between the Pinyin and Chinese character forms of Chinese names. Our method outperforms commercial gender inferring services by 9.70% to 20.08% relatively and it is open-sourced⁵.
3. We determine the potential upper bound of accuracy in the task of Pinyin name-gender guessing. For names in other languages (e.g., Japanese) that are romanized in a ‘many-to-one’ fashion, our model is a possible solution.

2 Related Work

2.1 Name-based Gender Prediction

Many efforts have been dedicated to enhancing name-based gender prediction tools for better service in scientific research. Santamaría and Mihaljević [2018] evaluate the performance of five frequently used paid web services for gender inference. Hu *et al.* [2021] develop character-level machine learning and deep learning methods for gender prediction tasks based on English names. Van Buskirk *et al.* [2023] propose a Cultural Consensus Theory (CCT) model that can unify data from multiple sources and offer a taxonomy of names to assess dataset composition. According to their analysis, the performance of current name-gender methods is nearing its ceiling. They believe that potential improvements can only come from increasing data volume and enhancing methodological transparency. However, this conclusion is somewhat one-sided, as the authors focus primarily on languages like English and other Latin-based languages, overlooking the nuances of Chinese names.

Chinese names differ from English names in that they possess richer lexical meanings, often gender-related. Additionally, Chinese characters provide more gender clues compared to English letters. In Chinese, leveraging the inherent information in names to improve gender prediction performance is feasible. Ngender⁶ determines the gender probability of a first name by calculating the product of the gender probabilities of each letter. PBERT [Jia and Zhao, 2019] learns gender characteristics of Chinese names through character embeddings and pronunciation embeddings of Chinese names. CHGAT [Pan *et al.*, 2023] represents Chinese character names by constructing a heterogeneous graph attention network of Chinese character radicals and pronunciations. While progress has been made in gender prediction for Chinese names in character form, in more international contexts, Chinese names are often presented in Pinyin form. To our knowledge, there has yet to be research on the task of gender classification for Pinyin names. Multilingual language models may handle different languages and their transliterated versions. However, this transfer relies on language structure [Pires *et al.*, 2019], which does not exist in extremely short texts such as names.

⁵https://github.com/PinyinNameGenderPrediction/pinyin_name_gender_prediction

⁶<https://github.com/observerss/ngender>

2.2 Deep Multi-task Learning

The basic idea of multi-task learning is that knowledge acquired from one task may prove beneficial when performing another related task [Crawshaw, 2020]. Traditional multi-task learning architectures are quite simplistic: All tasks share a global feature encoder through either hard or soft parameter sharing, with each task having its separate output branch. Decoder-based architecture [Xu *et al.*, 2018] introduce intermediate auxiliary prediction tasks in a multi-task network to extract features that have been disentangled according to the output tasks, making it easier for other tasks to extract relevant information [Vandenhende *et al.*, 2020]. Recently, Yang *et al.* [2022b] employ a multi-task network in automatic speech recognition to jointly learn the features of Pinyin and Chinese characters, demonstrating that Pinyin and Chinese characters can mutually enhance each other. As they treat Pinyin and Chinese characters equally, in our scenario, we only concern with the performance of Pinyin-related task. We initially facilitate knowledge transfer between Pinyin gender prediction task and Pinyin-to-Character prediction task through hard parameter sharing. Here Pinyin-to-Character prediction task serves as an auxiliary task to enhance our main task. It is known that a suitable auxiliary task for the target task can reap the benefits of multi-task learning [Ruder, 2017]. We then distill the gender-specific features from ground truth Chinese character names instead of from intermediate predictions as in [Xu *et al.*, 2018].

2.3 Knowledge Distillation

Knowledge distillation transfers knowledge from a large teacher model to a smaller student model, enabling the smaller model to achieve performance close to that of the larger model. Hinton *et al.* [2015] first propose response-based knowledge distillation. It views the output of the teacher network’s logits layer as soft labels containing additional information, which are used to supervise the learning of the student model. Feature-based knowledge distillation works on the principle of feature alignment. It is first introduced in FitNet [Romero *et al.*, 2014]. Though knowledge distillation is traditionally used for model compression, it can also be viewed as a way of transferring knowledge across domains [Gou *et al.*, 2021]. In our scenario, we set the Chinese character names gender prediction network as the teacher model and the Pinyin name-gender prediction network as the student model. The transfer of knowledge from Chinese characters to Pinyin is achieved through knowledge distillation between the two models.

3 Problem Statement

We denote the Pinyin form of Chinese names as X_{pinyin} , whereas the logographic (Chinese character) representation is indicated as X_{hanzi} . Given a training dataset $D = \{X_{\text{pinyin}}, X_{\text{hanzi}}, y\}$, our main goal is to train a model f_1 with trainable parameters θ_1 to classify X_{pinyin} into binary category $y \in \{0, 1\}$:

$$f_1(X_{\text{pinyin}}, \theta_1) \rightarrow y \quad (1)$$

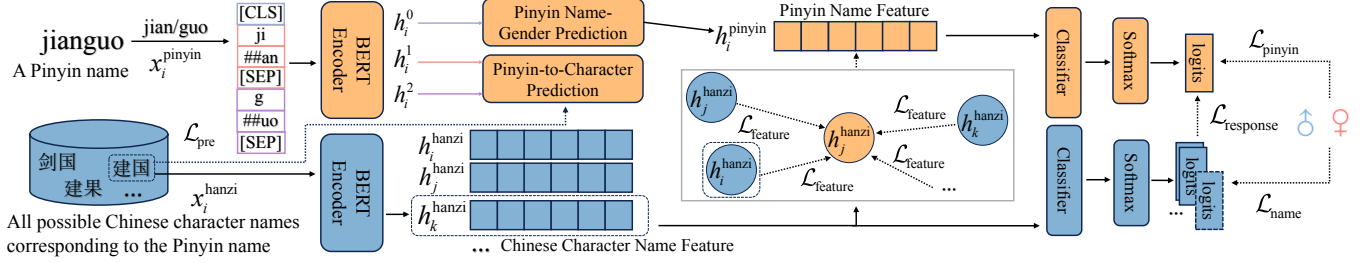


Figure 3: System structure. Orange represents the Pinyin name-gender prediction model (i.e., the student model) that incorporates the Multi-Task Learning module, while blue denotes the Chinese character name-gender prediction model (i.e., the teacher model).

Without loss of generality, we let $y = 1$ denote female and $y = 0$ signify male. To capitalize on supplementary information contained within the Chinese names in character form, we develop two additional models f_2 and f_3 to serve as auxiliaries to the main model f_1 :

$$f_2(X_{\text{hanzi}}, \theta_2) \rightarrow y \quad (2)$$

$$f_3(X_{\text{pinyin}}, \theta_3) \rightarrow X_{\text{hanzi}} \quad (3)$$

Specifically, f_2 is a Chinese character name-gender prediction model, whereas f_3 focuses on predicting Chinese character names based on their Pinyin form.

4 Method

The overall structure of our model is shown in Figure 3. First, we preprocess Pinyin names and feed them into the BERT encoder to generate vector representations. Second, a Pinyin-to-Character prediction task is co-trained with the Pinyin name-gender prediction task, constituting the Multi-Task Learning module (i.e., the orange part) that captures the Pinyin gender tendency and the association between Pinyin and Chinese characters. Third, we design a Knowledge Distillation module and apply a Chinese character name-gender prediction model (i.e., the blue part) as the teacher model. The output of the last layer and intermediate layers of the teacher model are used to supervise the training of the Pinyin name-gender prediction model, capturing the gender-related information embedded within the Chinese characters. Finally, the distilled information is used for gender classification. The input of our model is Pinyin names and Chinese character names during training. While during the testing phase, the model solely utilizes Pinyin names for gender prediction. We detail the design of our model in the following sections.

4.1 Pinyin Preprocessing

Since we want to use BERT as our encoder, the default tokenizer will erroneously separate letters that do not belong to a Pinyin syllable. Taking the two-character name ‘建国’ (jianguo) as an example, the default tokenizer separates ‘jianguo’ into ‘ji’, ‘ang’ and ‘uo’, while ‘建国’ actually corresponds to the Pinyin syllables ‘jian’ and ‘guo’. To avoid this, we use `pinyin-split`⁷ to segment Pinyin names into syllables that correspond to individual Chinese characters.

⁷<https://github.com/llb0536/PinyinSplit>

4.2 Pinyin-to-Character Prediction

To obtain a Pinyin representation enriched with information from Chinese characters, we introduce a Pinyin-to-Character Prediction task as an auxiliary task for Multi-Task Learning. For an input Pinyin name $x_i^{\text{pinyin}} \in X_{\text{pinyin}}$, taking a two-character name as an example, we first segment the letter string x_i^{pinyin} into Pinyin syllables corresponding to Chinese characters $P_i = \{p_i^1, p_i^2\}$. Then, we feed the split Pinyin name into BERT encoder and get a Pinyin representation $H_i^{\text{pinyin}} = \{h_i^0, h_i^1, h_i^2\}$. Here, h_i^0 represents the feature of the entire Pinyin name, which is used for subsequent gender classification. While h_i^1 and h_i^2 represent the features of individual syllables, used for Chinese character name prediction. The target of the Pinyin-to-Character prediction task is $x_i^{\text{hanzi}} = \{c_i^1, c_i^2\} \in X_{\text{hanzi}}$. The prediction loss of this task is:

$$\mathcal{L}_{\text{pre}} = \frac{1}{N} \sum_i -[\log P(c_i^1 | \{h_i^1, h_i^2\}) + \log P(c_i^2 | \{h_i^1, h_i^2\})] \quad (4)$$

4.3 Response-Based and Feature-Based Knowledge Distillation

The Pinyin representation learns the connection between Pinyin and Chinese characters from the Pinyin-to-Character prediction task, but the gender-related features in Chinese characters should be given more attention. To acquire the gender features of Chinese character names, we establish a Chinese character name-gender prediction task. Another BERT encoder and a single layer fully connected network extract gender information from the Chinese characters. For an input Chinese character name $x_i^{\text{hanzi}} = \{c_i^1, c_i^2\} \in X_{\text{hanzi}}$ and the gender label y_i , the prediction loss of this task is:

$$\mathcal{L}_{\text{name}} = \frac{1}{N} \sum_i -\log P(y_i | \{c_i^1, c_i^2\}) \quad (5)$$

To transfer the gender information of Chinese character names to Pinyin names, we perform the feature-based and response-based distillation between the Chinese character model and the Pinyin model. Specifically, we use the features and the logits of Chinese character names to supervise the training of the Pinyin names. In Chinese character model, we denote the output of the BERT encoder as h_i^{hanzi} and the output of the last fully connected layer as logits vector z_i^{hanzi} . In

Pinyin model, a linear layer projects Pinyin name representation h_i^0 to h_i^{pinyin} , while another generates the logits of Pinyin name z_i^{pinyin} from h_i^{pinyin} . The distillation loss for feature-based knowledge transfer is defined as:

$$\mathcal{L}_{\text{feature}} = \frac{1}{N} \sum_i \|h_i^{\text{hanzi}}, h_i^{\text{pinyin}}\|_2 \quad (6)$$

The distillation loss for response-based knowledge transfer is defined as:

$$p(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \quad (7)$$

$$\mathcal{L}_{\text{response}} = \frac{1}{N} \sum_i D_{\text{KL}}(p(z_i^{\text{pinyin}}) \| p(z_i^{\text{hanzi}})) \quad (8)$$

Then, the loss function of Pinyin name-gender prediction with response-level and feature-level knowledge distillation can be defined as:

$$\mathcal{L}_{\text{pinyin}} = \frac{1}{N} \sum_i -\log P(y_i | z_i^{\text{pinyin}}) \quad (9)$$

Finally, we combine all the loss functions to formulate our ultimate objective function:

$$\mathcal{L} = \mathcal{L}_{\text{pre}} + \mathcal{L}_{\text{name}} + \mathcal{L}_{\text{feature}} + \mathcal{L}_{\text{response}} + \mathcal{L}_{\text{pinyin}} \quad (10)$$

5 Experiments

5.1 Experimental Setup

Datasets

We evaluate our model with three distinct datasets. The first is a dataset of 9,800 names compiled from the 2005 China’s 1% Population Census [Bao *et al.*, 2021]. The second encompasses a collection of 20,000 names from several public sources in China [Sebo, 2022]. To facilitate a fair comparison with paid web services that rely on large-scale data, we utilized a third dataset comprising 58,393,173 records collected from Guangdong province government [Pan *et al.*, 2023], named as 58M Names dataset. The detail information of the datasets is shown in Table 2.

Train/Test Split

The 9,800 Names dataset is split into 8:1:1 for training, validation, and test. For a fair comparison with paid web services, we train our model using a significantly larger 58M Names dataset. Subsequently, we conduct testing across the entire scope of the 9,800 Names dataset and 20,000 Names dataset.

	Records	UH Name	UP Name
9,800 Names	9,800	6,972	4,203
20,000 Names	20,000	N/A	17,948
58M Names	58,393,173	560,706	65,971

Table 2: Statistics of the datasets, where UH Name means Unique Hanzi (Chinese Character) first name and UP Name means Unique Pinyin first name. For the 20,000 Names dataset, it lacks the first name in Chinese character form.

confusion matrix		predicted class		
		male	female	unknown
true class	male	m_m	m_f	m_u
	female	f_m	f_f	f_u

Table 3: Confusion matrix of six possible classification outcomes.

Evaluation Metrics

The traditional metrics are Accuracy (Acc), Precision (P), Recall (R), and F1-score (F1). While in our scenario, we are not assessing the outcomes of a strict binary classification problem. Besides ‘male’ and ‘female’, an ‘unknown’ category can also be returned by paid web services. To comprehensively evaluate the performance of different methods, we adopt a evaluation metric [Wais, 2016] that incorporates the ‘unknown’ category. This metric has already been utilized in evaluating the performance of paid web services [Santamaría and Mihaljević, 2018; Sebo, 2022]. The new confusion matrix is defined in Table 3.

Elements m_m and f_f are the correct classifications, while elements m_f and f_m are misclassifications. The sum of both can be simply referred to as classifications, and elements m_u and f_u represent non-classifications. Both misclassifications and non-classifications are included under the term inaccuracies. Based on the confusion matrix, four performance metrics are introduced:

$$\text{errorCoded} = \frac{f_m + m_f + m_u + f_u}{m_m + f_f + m_f + f_m + m_u + f_u} \quad (11a)$$

$$\text{errorCodedWithoutNA} = \frac{f_m + m_f}{m_m + f_f + m_f + f_m} \quad (11b)$$

$$\text{naCoded} = \frac{m_u + f_u}{m_m + f_f + m_f + f_m + m_u + f_u} \quad (11c)$$

$$\text{errorGenderBias} = \frac{m_f - f_m}{m_m + f_f + m_f + f_m} \quad (11d)$$

The first metric, `errorCoded`, treats both misclassifications and non-classifications as errors, thereby measuring the overall performance of the method. The second metric, `errorCodedWithoutNA`, assesses the errors excluding the ‘unknown’ category, proving useful when high predictive accuracy is demanded and a significant number of omissions is acceptable. The third metric, `naCoded`, quantifies the rate of missed outcomes. The last metric, `errorGenderBias`, estimates the direction of bias in gender prediction and is used to assess whether errors are more frequent with male or female names. A positive value indicates that males are misclassified more frequently than females, and vice versa.

Baselines

We compare our method with nine baselines. In the first baseline, we directly convert Pinyin names to Chinese characters. The second one, `char-BERT`, is a character-based name-gender prediction model tailored for alphabetic writing systems. The subsequent two (`Ngender` and `PBERT`) are representative methods for predicting gender from Chinese character names. We modify these models to work on Pinyin names.

Methods	P	R	F1	Acc
CCT (NA=M)	0.6431	0.6133	0.5880	0.6071
CCT (NA=F)	0.6764	0.6604	0.6550	0.6643
GPT3.5	0.6948	0.6191	0.5781	0.6191
Conversion	0.6824	0.6825	0.6824	0.6827
char-BERT	0.7142	0.7126	0.7117	0.7122
Pinyin-Ngender	0.7158	0.7157	0.7154	0.7155
PBERT (Pinyin)	0.7180	0.7175	0.7169	0.7171
Ours	0.7349	0.7341	0.7334	0.7337

Table 4: Performance comparison for models trained and tested on the 9,800 Names dataset with 5-fold cross validation. GPT3.5 is given the prompt and the 9,800 names.

We also compare our method with three popular name-based gender inferring tools provided by paid web services, as well as the current most popular large language model, GPT-3.5⁸. The last CCT is a recently proposed dictionary-based method for name-based gender classification.

- **Conversion:** It simply converts Pinyin names into most possible Chinese character names and employ a SOTA Chinese character names gender classification tool CHGAT [Pan *et al.*, 2023] to predict genders from these Chinese character names.
- **char-BERT [Hu *et al.*, 2021]:** This model segments names into individual letters (i.e., characters) and generates a vector representation for each letter. It captures letter-level information within names and performs well with English names.
- **Pinyin-Ngender:** Ngender is a commonly used Chinese character name-gender prediction tool based on the Naive Bayes. We implement a Pinyin version of Ngender, called Pinyin-Ngender, by replacing the Chinese characters with their corresponding Pinyin.
- **PBERT [Jia and Zhao, 2019]:** Pinyin BERT (PBERT) employs a pre-trained BERT to generate vector representations for names in both their Chinese character and Pinyin forms, using their concatenation for gender classification. Here, we implement its Pinyin-only version.
- **Paid Web Services:** Namsor, Gender API and genderize.io are mainstream paid web services which perform well on European names with wide applications in scientific research [Santamaría and Mihaljević, 2018; Van Buskirk *et al.*, 2023].
- **GPT-3.5:** GPT-3.5 has demonstrated remarkable capabilities in name-gender tasks [Michelle *et al.*, 2023]. Here we use this prompt: “I will give you the first names of some people, please infer their gender, use 0 to represent male and 1 to represent female.”.
- **CCT [Van Buskirk *et al.*, 2023]:** The CCT model is based on Cultural Consensus Theory. It forms a gender consensus for names from different data sources by iteratively calculating the capability of each data source and the collective estimation of all sources on the name.

⁸<https://platform.openai.com/docs/models/gpt-3-5-turbo>

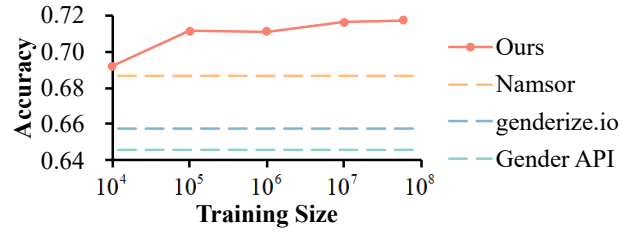


Figure 4: Model performance under different training set sizes. (The x-axis is in log scale.)

Since the original CCT model is based on datasets from 36 sources collected by its authors, and our data sources are far fewer, we only implement a simplified version of this model. Due to CCT’s inability to assign genders to names not present in the training data, we are compelled to uniformly designate all instances of missing gender information (NA) as either ‘male’ or ‘female’, denoted as CCT (NA=M) and CCT (NA=F).

5.2 Experimental Results

Table 4 shows the average results of 5-fold cross-validation on the 9,800 Names dataset beyond paid services since they cannot be evaluated using traditional metrics. Among all these baselines, our method achieves the best overall performance and exhibits a relative improvement of 2.31% in accuracy compared to the best baseline model, which suggests the effectiveness of incorporating Chinese character information into the Pinyin name-gender prediction task. The observation that directly converting Pinyin into Chinese characters performs even worse than methods only using Pinyin underscores the complexity of the one-to-many relationship between Pinyin and Chinese characters. We also notice that the syllable-based method, PBERT (Pinyin only) and Pinyin-Ngender, perform better than character-based method, char-BERT, which indicates that Pinyin syllables convey gender information. GPT-3.5 does not perform very well on this task, compared with specifically designed models.

Paid web services typically rely on large-scale datasets. Given that we start with a dataset of 9,800 names, which is significantly smaller than the datasets used by paid web services, we want to see what happens as the size of the training set continues to increase. In each iteration, we randomly select a subset of names from the 58M Names dataset for training and evaluate the performance of our model on the entire 9,800 Names dataset. We report the experimental results in Figure 4. When the training size is 10,000, which is approximately equal to that of the test set, our method has surpassed the best commercial gender detection tools. As the size of the training set continues to increase, our method performs increasingly better compared to paid web services.

Finally, we train our model utilizing the entire 58M Names dataset and evaluate on two distinct datasets representing scenarios with and without Chinese character first names. The experimental results are presented in Table 5. On the 9,800 Names dataset, our method outperforms the leading mainstream name-based gender inferring tools, achieving relative improvements of 9.70% to 20.08% (i.e., 3.04% to 7.11% ab-

	Methods	errorCoded	errorCodedWithoutNA	naCoded	errorGenderBias
9,800 Names	Gender API	0.3540	0.3030	0.0732	-0.1588
	genderize.io	0.3425	0.3156	0.0393	-0.1130
	Namsor	0.3133	0.3133	0.0000	-0.0671
	Ours	0.2829	0.2829	0.0000	-0.1196
20,000 Names	Gender API	0.6496	0.4183	0.3976	-0.0906
	genderize.io	0.6124	0.4306	0.3193	-0.0438
	Namsor	0.5253	0.4065	0.2001	-0.2021
	Ours	0.3847	0.3847	0.0000	-0.0862

Table 5: Performance comparison with paid web services. Models are trained on the 58M dataset.

Methods	P	R	F1	Acc
w/o KD&MTL	0.7141	0.7142	0.7142	0.7143
w/o logits&feat	0.7253	0.7253	0.7253	0.7255
w/o logits	0.7287	0.7288	0.7285	0.7286
Ours	0.7349	0.7341	0.7334	0.7337

Table 6: Performance of different variants under the 9,800 dataset.

solutely) in errorCoded, the proportion of misclassifications and nonclassifications. On the 20,000 Names dataset, the improvements brought by our method were even more significant, achieving relative gains ranging from 26.77% to 40.78% (i.e., 14.06% to 26.49% absolutely) in errorCoded.

Additionally, we observe that in almost all results of various methods, errorGenderBias is consistently negative. This means that females are more prone to being misclassified compared to males, which could likely be caused by an imbalance in male/female ratio within the training set.

5.3 Ablation Study

To validate the effectiveness of our proposed multi-task learning module and knowledge distillation module, we incrementally remove the response-based knowledge distillation (w/o logits), feature-based knowledge distillation (w/o logits&feat), the entire teacher network and the task of predicting Chinese character names (w/o KD&MTL) from our model. We test various variants of our model on the 9,800 Names dataset, and the experimental results are presented in Table 6. Unsurprisingly, our method outperformed all other variants. Furthermore, as response-based knowledge, feature-based knowledge, and the multi-task learning module are removed, the model’s accuracy progressively decline.

6 Discussion

Performance Upper Bound. It is noticed that, as training set size grows, the performance increase of our method slows down. This is consistent with observations in other name-gender studies [Van Buskirk *et al.*, 2023; Lockhart *et al.*, 2023]. It appears that, for name-gender guessing tasks, there exists an information-theoretical limit to accuracy. Beyond this threshold, additional reference data or more advanced modeling techniques may not enhance the performance.

As mentioned in Introduction, if we know the exact Chinese characters behind each Pinyin name, the problem would be almost solved. In the previous work for Chinese character names gender prediction, similarly trained on the 58M dataset and tested on the 9,800 dataset, the accuracy of current SOTA model [Pan *et al.*, 2023] reaches 81.47%. This may be the most ideal scenario for Pinyin name-gender classification, which is (almost) impossible to achieve since it is impossible to determine the exact Pinyin-to-characters mappings. Our proposed method achieves 71.71% in accuracy (i.e., 28.29% in errorCoded), which means the potential upper bound for this task may lie between these two extremes and further enhancements towards the upper bound would be increasingly difficult.

Future Directions. Though this work focuses on gender detection of romanized Chinese names, a similar approach could be applied to other languages that are romanized in English records, such as Japanese in which many-to-one mappings exist. If the conversion of names from different languages into alphabetic script is inevitable, our method can to some extent mitigate the loss of gender information resulting from this transliteration process.

Naming preferences in China change over time and across regions, according to the Ministry of Public Security of China⁹. It is unclear how these specifics of name-gender dictionaries affect the model performance on test sets from different time and regions and we leave it to future explorations.

7 Conclusion

We identify the problem of misgendering Chinese Pinyin names and the exclusion of Chinese individuals from analyses in previous gender bias studies, which is resulted from the low accuracy of name-gender guessing tools on Pinyin names. We define a new Pinyin name-gender guessing task accordingly and design a Multi-Task Learning Network assisted by Knowledge Distillation to transfer semantic and gender information from the rich Chinese characters to Pinyin. This open-sourced method outperforms commercial gender classification tools and other baselines. Aligned with the UN’s SDGs to promote gender equality, we hope that it can help consolidate gender bias studies such that Chinese individuals’ genders are inferred with higher accuracy and they would not be excluded from these studies.

⁹https://www.gov.cn/xinwen/2021-02/08/content_5585906.htm

Ethical Statement

Gender bias visibly or invisibly exists in many aspects of our lives and severely hinders humankind’s sustainable future. We design our model to help researchers and policymakers to confirm its systematic existence, to estimate its negative impact, to figure out its mechanism, to educate the public and to develop corresponding scientific and effective policies. However, not all names are strongly gendered. Although a name can serve as a gender characteristic, it does not necessarily reflect the actual sexuality of a person. Model predictions should still be treated with caution.

References

- [Ackermann and Zimmer, 2021] Tanja Ackermann and Christian Zimmer. The sound of gender–correlations of name phonology and gender across languages. *Linguistics*, 59(4):1143–1177, 2021.
- [Bao et al., 2021] Han-Wu-Shuang Bao, Huajian Cai, Yiming Jing, and Jianxiong Wang. Novel evidence for the increasing prevalence of unique names in china: A reply to oghihara. *Frontiers in Psychology*, 12:731244, 2021.
- [Crawshaw, 2020] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.
- [Dworkin et al., 2020] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature neuroscience*, 23(8):918–926, 2020.
- [Gou et al., 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.
- [Hinton et al., 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [Hu et al., 2021] Yifan Hu, Changwei Hu, Thanh Tran, Tejaswi Kasturi, Elizabeth Joseph, and Matt Gillingham. What’s in a name?—gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 35(4):1537–1563, 2021.
- [Huang et al., 2020] Junming Huang, Alexander J Gates, Roberta Sinatra, and Albert-László Barabási. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616, 2020.
- [Jia and Zhao, 2019] Jizheng Jia and Qiyang Zhao. 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, 2019.
- [Larivière et al., 2013] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.
- [Lockhart et al., 2023] Jeffrey W Lockhart, Molly M King, and Christin Munsch. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, pages 1–12, 2023.
- [Michelle et al., 2023] Alexopoulos Michelle, Lyons Kelly, Mahetaji Kaushar, Barnes Marcus Emmanuel, and Gutwillinger Rogan. 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, 2023.
- [Pan et al., 2023] Zihao Pan, Kai Peng, Shuai Ling, and Haipeng Zhang. For the underrepresented in gender bias research: Chinese name gender prediction with heterogeneous graph attention network. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’23/IAAI’23/EAAI’23*. AAAI Press, 2023.
- [Pires et al., 2019] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, 2019.
- [Romero et al., 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.
- [Ruder, 2017] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- [Santamaría and Mihaljević, 2018] Lucía Santamaría and Helena Mihaljević. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156, 2018.
- [Sebo, 2022] Paul Sebo. How accurate are gender detection tools in predicting the gender for chinese names? a study with 20,000 given names in pinyin format. *Journal of the Medical Library Association: JMLA*, 110(2):205, 2022.
- [Spoon et al., 2023] Katie Spoon, Nicholas LaBerge, K Hunter Wapman, Sam Zhang, Allison C Morgan, Mirta Galesic, Bailey K Fosdick, Daniel B Larremore, and Aaron Clauset. Gender and retention patterns among us faculty. *Science Advances*, 9(42):eadi2205, 2023.
- [Squazzoni et al., 2021] Flaminio Squazzoni, Giangiacomo Bravo, Mike Farjam, Ana Marusic, Bahar Mehmani, Michael Willis, Aliaksandr Birukou, Pierpaolo Dondio, and Francisco Grimaldo. Peer review and gender bias: A study on 145 scholarly journals. *Science advances*, 7(2):eabd0299, 2021.
- [Van Buskirk et al., 2023] Ian Van Buskirk, Aaron Clauset, and Daniel B Larremore. An open-source cultural consensus approach to name-based gender classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 866–877, 2023.

- [Vandenhende *et al.*, 2020] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Revisiting multi-task learning in the deep learning era. *arXiv preprint arXiv:2004.13379*, 2(3), 2020.
- [Wais, 2016] Kamil Wais. Gender prediction methods based on first names with genderizer. *R J.*, 8(1):17, 2016.
- [Xu *et al.*, 2018] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.
- [Yang *et al.*, 2022a] Yang Yang, Tanya Y Tian, Teresa K Woodruff, Benjamin F Jones, and Brian Uzzi. Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences*, 119(36):e2200841119, 2022.
- [Yang *et al.*, 2022b] Zhao Yang, Dianwen Ng, Xiao Fu, Liping Han, Wei Xi, Rui Wang, Rui Jiang, and Jizhong Zhao. On the effectiveness of pinyin-character dual-decoding for end-to-end mandarin chinese asr. *arXiv preprint arXiv:2201.10792*, 2022.