Occupational gender bias in ungendered languages an LLMS: Comparing Hungarian and Chinese

Anonymous ACL submission

Abstract

This paper examines occupational gender bias and stereotypes, in a cross-linguistic setting. We analyze ratings of 50 job titles collected from speakers of two languages without grammatical gender markings: Hungarian and Mandarin Chinese. Participants were instructed to rate how typical it is for a certain job to be done by men or women, according to their own perceptions. Our results show that in both languages the occupational nouns carry societal biases, despite the fact that the job titles themselves have no grammatical gender markings. We analyze the ratings by participant gender and perform intra-linguistic and cross-linguistic comparisons, highlighting key differences and offering insights that range from peculiarities in word formation to broader cross-cultural generalizations. Additionally, we also compared the human raters' responses with that of a few popular generative AI engines. Interestingly, the biases exhibited by Large Language Models (LLMs) were found to be even stronger than those shown by our human participants.

1 Introduction

— Check Kaukonen et al. (2025).

…

— Hungarian has no grammatical gender, and most occupations are used in gender-neutral terms.

— Chinese only marks gender when writing the 3rd person singular pronoun – 他 *tā* 'he' / 她 *tā* 'she' – but that too is a relatively recent invention, going back to the May Fourth Movement of 1919 (Bi, 2013), and similarly to Hungarian, most occupations are unmarked for gender.

—Gabor: Describe word formation for job titles in Hungarian, why most jobs are unmarked for gender, and when they are not. Add cultural notes? See Materials.

—Wenhui: Describe word formation of job titles in Chinese, why job titles are unmarked for gender, what does adding 女 'woman' and 男 'man' "do", and any related cultural aspects.

We are interested in these unmarked words, as they do not inherently possess gender bias – certainly not grammatically – but according to our expectations they will be rated according to the prevailing societal stereotypes nonetheless.

…

— Mention some "default is masculine" theory if exists...? Latin, Arabic, how about languages with many genders?

1.1 Background

…

In section 2 we describe the methods and the experimental setup in detail, followed by the results and their analysis in section 3. Finally, we discuss the findings in section 4.

2 Experimental setup

For both languages we designed a simple survey-based experiment in which participants were asked to rate job titles on a 7-point Likert scale. Participants were instructed to make decisions on how likely is an occupation to be pursued by men or by women, according to their own perception.[1] First, we will introduce the Hungarian experiment, then the Chinese one, and then move on to the results.

2.1 Hungarian

2.1.1 Participants

A total of 22 native Hungarian speakers filled our questionnaire, and after validating the responses (reviewing attention checks and manually checking for anomalies) 2 were rejected. Participants were recruited online using the participant recruitment platform Prolific, with screeners set for location

[1] While this study focuses on people who identify or are identified as either male or female, we acknowledge the presence of non-binary people in the workforce.

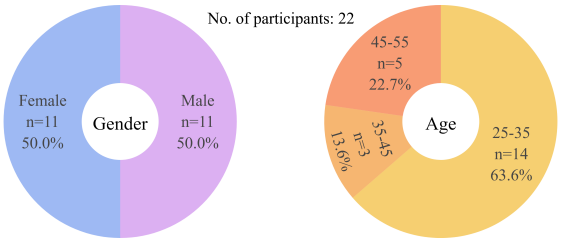

Figure 1: Demographics of the Hungarian participants.

(Hungary) and first language (Hungarian); raters were compensated for their time with a small monetary reward. In the end, the Hungarian ratings dataset had 20 participants (n=11 female, n=9 male), with ages ranges of 25-35 (n=11), 35-45 (n=4), and 45-55 (n=5). See Figure 1 for the distribution.

2.1.2 Materials

The Hungarian survey contained 44 items and 6 attention-check items, each a commonly occurring job title in Hungary, such as: *modell* 'model' or *katona* 'soldier', in no particular order. The attention-check items were removed from the final analysis, these were *pincérnő* 'waitress', *titkárnő* 'secretary (female)', *tanárnő* 'teacher (female)', *takarítónő* 'cleaning lady', *ápolónő* 'nurse (female)', and *házvezetőnő* 'housekeeper (female)'. These words are compounds, they explicitly determine the gender of the worker by appending *-nő* 'woman' to the base noun. If participants paid attention, all these items should be rated 'completely female' (3). Participants who rated any of these lower than 2, or rated them lower than 3 more then once were rejected.

The 6 words above have counterparts without the *nő* 'woman' element, i.e. *pincér* 'waiter', *titkár* 'secretary', *tanár* 'teacher', *takarító* 'cleaner', *ápoló* 'nurse', and *házvezető* 'housekeeper', all of which are included in our survey. These words are unmarked for gender, but they do not explicitly refer to men. The perceived gender bias of these occupations – if any – is due to social factors, not grammatical ones.

On the surface level, the unmarked form is the base noun, and the feminine-marked form is created by appending *-nő* 'woman' to it, which is a regular way of creating female occupational titles in Hungarian. However, in reality this does not equal to a male-female pair, as the unmarked form is not necessarily "masculine". We can observe 3 types in the pragmatic usage of occupational nouns when it comes to the unmarked–marked pairing and its implications for the gendering of the unmarked nouns.

1) **Both forms are common.** Frequently occurring word-pairs in Hungarian would be for example *énekes* 'singer' – *énekesnő* 'female singer' (not in our dataset). In cases where both versions are well established – i.e., both occur with a relatively high frequency in a balanced corpus – the unmarked word seems to carry some male bias, as the frequent use of a feminine form indicates a need and/or custom for differentiation. We wanted to test if raters perceived this bias or not. In this example, the absolute and relative frequencies (occurrance per a million words) of the two lemmatized nouns in the Hungarian National Corpus (HNC) are 1441/9,4001 for *énekes* and 748/4,8795 for *énekesnő* (Váradi, 2002; Oravecz et al., 2014); the frequency difference here is roughly half (51.9%).

The stark deviations in frequencies for marked-unmarked word pairs such as the above are not an indicator for a strong gender bias – we can assume that both men and women singers would be equally represented in the Hungarian corpus – but reflect that in general, the unmarked forms are used for either males or females when talking about one's occupation. The female-marked forms are used when there is an explicit intention to specify the gender of the individual, when it is otherwise not known from context (or proper names).[2]

2) **Only unmarked form is common.** However, there are many cases where the unmarked form is the only one generally used for both genders. Take for example *ügyész* 'prosecutor' (8451/55,1287) vs. *ügyésznő* 'female prosecutor' (56/0,3653), or *fodrász* 'hairdresser' (944/6,1580) vs. *fodrásznő* 'female hairdresser' (35/0,2283); the deviations in frequency here are over multiple orders of magnitude. In these cases, the unmarked form is the default word to describe anyone practicing the occupation, and appending *'-nő* 'woman' to it – although possible – would render it unusual, and a bit awkward; but still not as uncanny/forced as Modern English *singress* would be.[3]

3) **Marked form is common.** Furthermore, there are cases, where the female-marked version ending in *-nő* is so ubiquitous, that it is the

[2]For example, the sentence *Anyukám tanár.* 'My mom is a teacher.' uses the unmarked form, firstly because we want to communicate her job, and secondly because it is obvious from the subject (mother) that she is a woman.

[3]Although English had a form *singeress*, from Middle English *syngeresse*, it is now obsolete.

unmarked version that will sound a bit unusual, such as *házvezető* 'housekeeper' (10/0.0652) vs. *házvezetőnő* 'female housekeeper' (92/0,6001), or, to a small extent *takarító* 'cleaner' (169/1.1024) vs. *takarítónő* 'female cleaner' (392/2,5571). In short, we are interested in these unmarked words and how people perceive them.

The full list of 44 items is as follows: *modell*, *katona* 'soldier', *kórboncnok* 'pathologist', *vezérigazgató* 'CEO', *menedzser*, 'manager' *nővér*, 'nurse' *szakács* 'chef', *felszolgáló* 'server', *könyvelő* 'accountant', *professzor* 'professor', *építész* 'architect', *tudós* 'scientist', *ápoló* 'nurse', *pénztáros* 'cashier', *bíró* 'judge', *munkás* 'worker', *vízimentő* 'lifeguard', *jegyárus* 'ticket seller', *tűzoltó* 'firefighter', *mérnök* 'engineer', *rendező* 'director', *takarító* 'cleaner', *HR-es* 'HR specialist', *házvezető* 'housekeeper', *légiutas-kísérő* 'flight attendant', *pincér* 'waiter', *orvos* 'doctor', *fodrász* 'hairdresser', *földműves* 'farmer', *gondozó* 'caregiver', *bolti eladó* 'shop assistant', *kertész* 'gardener', *titkár* 'secretary', *PR munkatárs* 'PR officer', *dietetikus* 'dietitian', *tanár* 'teacher', *rendőr* 'police officer', *pilóta* 'pilot', *recepciós* 'receptionist', *biztonsági őr* 'security guard', *ügyész* 'prosecutor', *kozmetikus* 'beautician', *programozó* 'programmer', *diák* 'student'.

We have included *diák* 'student' out of curiosity. Although being a student is not a job per se, but it is beyond doubt the only truly gender-neutral "occupation" there is, since it is mandatory for every child to go to school (both in Hungary and in China). We wanted to see if there would be would be any bias regarding this word, especially that Hungarian has a female-marked form for it – *diáklány* 'girl student', appending *-lány* 'girl' to the base noun – and could be considered to belong to the 1st type with a potential male bias when considered in contrast.

2.1.3 Procedure

Hungarian participants were instructed to rate each word on a 7-point Likert scale, ranging from *completely male* to *completely female*, the ratings were then converted to numerical values from -3 to +3. The scale presented in both questionnaires followed the same logic, with 0 in the middle, hence the choices were completely male (-3); mostly male (-2); somewhat male (-1); neutral/equal (0); somewhat female (+1); mostly female (+2); completely female (+3). The exact wording of the main question of the Hungarian survey was: "Ön szerint a foglalkozás tipikusan férfi foglalkozás, vagy tipikusan női foglalkozás?" (Is this occupation typically a man's occupation or a woman's occupation?), cf. 2.

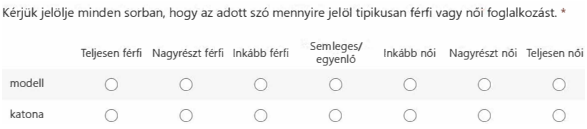

Figure 2: A sample of the the Hungarian survey layout.

The survey was created in Microsoft Forms, and after a brief welcome message and instructions the words were presented in a simple list format, each word with a corresponding rating scale next to it, with no context. Time limit was not set, but the survey was designed to take around 5 minutes, and participants took 4 minutes 25 seconds on average to finish.

2.2 Chinese survey

2.2.1 Participants

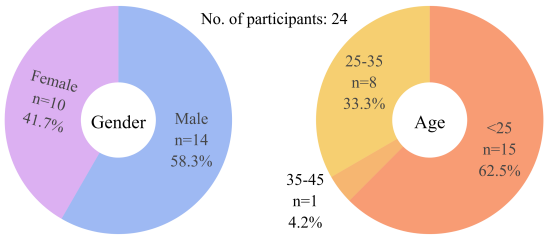

Figure 3: Demographics of the Chinese participants.

The Chinese survey was completed by 30 native Mandarin Chinese speakers, 6 of which were rejected after failing attention checks.

Wenhui: Please describe the platform used to recruit participants/how it was distributed.

Participants were paid a small fee for completing the questionnaire. The 24 accepted participants (n=14 male, n=10 female) were mostly university students, aged <25 (n=15), 25-35 (n=8), or 35-45 (n=1). See Figure 3 for the distribution.

2.2.2 Materials

The Chinese survey, too, contained 44 items of commonly occurring job titles in Mandarin Chinese (Simplified), with 6 attention checks, to ensure participant engagement and data quality, also in randomized order. The attention checks were 妈妈 'mother', 爸爸 'father', 女作家 'female writer', 男作家 'male writer', 女画家 'female painter', 男画家 'male painter'. These words are inherently feminine or masculine in meaning, or explicitly determine gender by prepending 女 'woman' and 男

'man', helping to filter out inattentive responses. Participants who failed to rate these with the highest scores of either *completely male* or *completely female* were rejected.

Here is the full list of Chinese items: 警察 'police', 秘书 'secretary', 教授 'professor', 护士 'nurse', 高管 'manager', 教师 'teacher', 前台 'receptionist', 工人 'worker', 幼师 'kindergarten teacher', 模特 'model', 护工 'caregiver', 保姆 'nanny', 会计 'accountant', 工程师 'engineer', 保洁 'cleaner', 法官 'judge', 导购员 'shop assistant', 美容师 'beautician', 服务员 'waiter', 乘务员 'flight attendant'[4], 理发师 'hairdresser', 空服员 'flight attendant'[5], 售票员 'ticket seller', 厨师 'chef', 营养师 'nutritionist', 家政员 'housekeeper', 收银员 'cashier', 医生 'doctor', 法医 'pathologist', 程序员 'programmer', 保安 'security guard', 导演 'director', 军人 'soldier', 董事长 'CEO', 农民 'farmer', 学生 'student', 园丁 'gardener', 飞行员 'pilot', 人事 'HR personnel', 消防员 'firefighter', 科学家 'scientist', 检察官 'prosecutor', 救生员 'lifeguard', 建筑师 'architect'.

2.2.3 Procedure

Chinese participants, too, were instructed to rate each word on a 7-point Likert scale, ranging from *completely male* to *completely female* with *neutral/equal* in the middle, and the ratings were then converted to numerical values as explained above. The exact wording of the questions in the Chinese survey was: "对于`occupation`这个职业，您认为通常担任该职业的男女性别比例是多少？" (What do you think is the ratio of men to women in `occupation`?). The survey was created in Microsoft Forms with essentially identical instructions, participants saw each word highlighted in the question above, with no other context, and were presented with the scale directly below.

Figure 4: A sample of the the Chinese survey layout.

[4] Also the attendant/crew on high-speed trains.

[5] Less frequent word; only on airplane

2.3 Data Analysis

For analyzing the data, we used the `scipy` library in Python, and performed one-sample *t*-tests to determine the significance of differences in the mean ratings of occupations, and independent sample (two-sample) *t*-tests to analyze the differences between different groups, such as the ratings of male and female raters and the differences between the ratings of Hungarian and Chinese raters for the same occupations. The significance level was set to $p < 0.05$, and marginally significant items ($0.05 < p < 0.1$) were also highlighted.

3 Results & Analysis

For both languages, the majority of occupations were rated with a significant gender bias. In Hungarian, 37 out of 44 occupations showed significant bias, and in Chinese, 39 out of 40 were biased. This in itself is not surprising...

3.1 Hungarian

The Hungarian data was first analyzed using a one-sample *t*-test to determine which of the occupations showed a significant bias, measured against 0 (neutral/equal). The results showed that the majority of occupational titles – 36 out of 44 – were rated with a significant gender bias, with 14 showing female, and 22 showing male bias. See Figure 5 for a visualization of the mean ratings, with the gender biases highlighted.

3.1.1 Overall distribution of ratings

In general, occupations were rated according to expectations, following societal stereotypes and realities. Words with the highest female bias were *kozmetikus* 'beautician' (2.20), *házvezető* 'housekeeper' (1.80), *légiutas-kísérő* 'flight attendant' (1.40), and *takarító* 'cleaner' (1.25), *gondozó* 'caregiver' (1.05), while words with the highest male bias included *munkás* 'worker' (-1.40), *pilóta* 'pilot' (-1.65), *katona* 'soldier' (-1.80), *biztonsági őr* 'security guard' (-1.90), and *tűzoltó* 'firefighter' (-2.20).

The highest rated feminine word – *nővér* 'nurse' (2.20) – is a bit special, as it literally means 'sister' and goes back to the time when nuns were the ones taking care of the sick; hence the word carries a strong female bias that is encoded in its literal meaning. Interestingly, it was not rated as an exclusively female job, probably because male nurses are now also common. The gender-neutral word

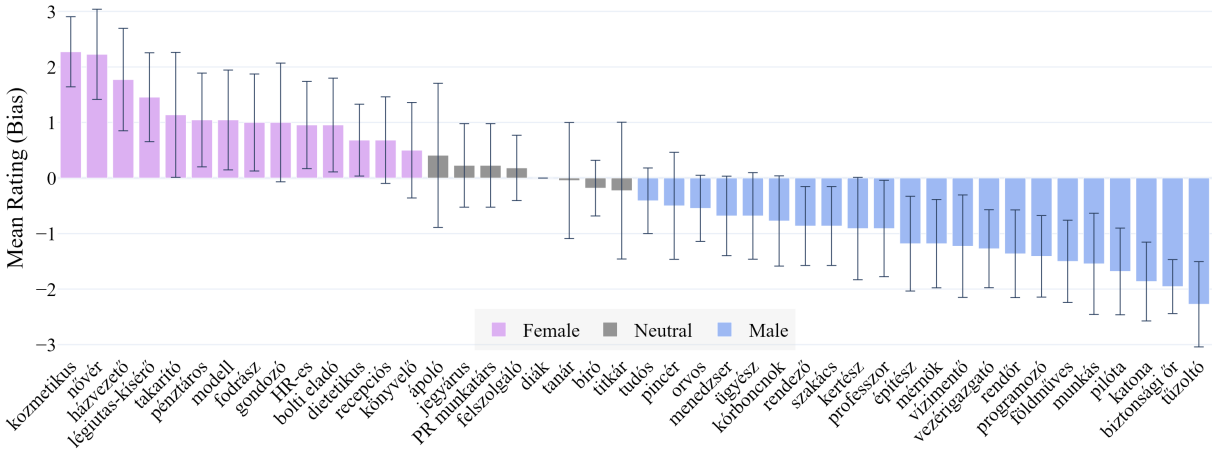

Figure 5: Mean ratings of occupational titles in Hungarian with standard deviations, significant gender bias highlighted – explore the interactive plot.

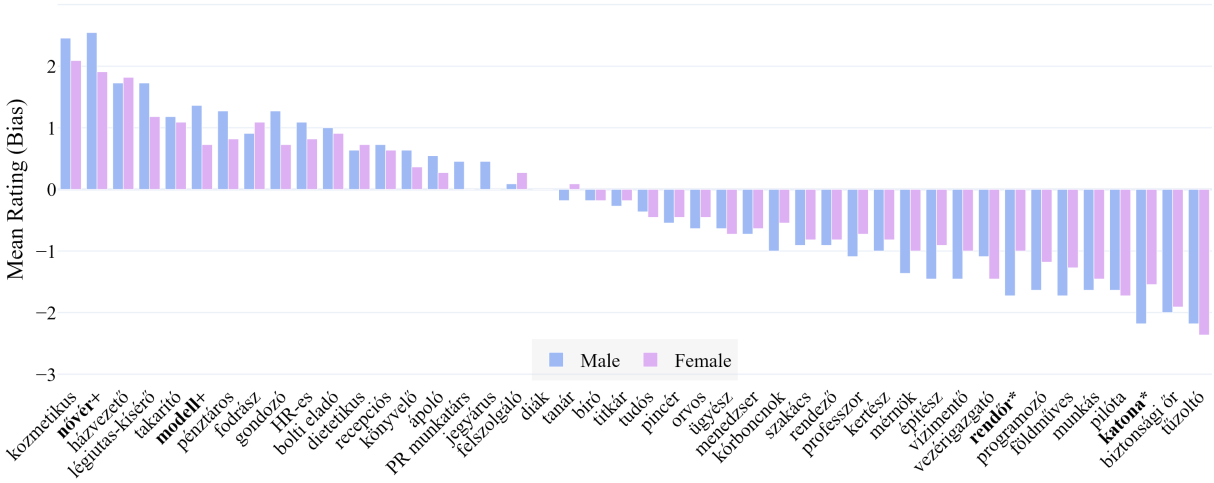

Figure 6: Mean ratings of occupational titles in Hungarian by gender, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

ápoló 'nurse' for the same job was also tested, and it received a neutral rating of 0.45.

The 8 job titles that came back as not significantly biased were: *ápoló* 'nurse' (0.45), *PR munkatárs* 'PR worker' (0.30), *felszolgáló* 'server' 0.20, *jegyárus* 'ticket seller' (0.20), *diák* 'student' (0), *tanár* 'teacher' (0), *bíró* 'judge' (-0.25), and *titkár* 'secretary' (-0.30). It is worth noting that while *diák* 'student' was rated 0 by nearly everyone, *tanár* 'teacher' had more individual variation in the ratings, leading to a higher standard deviation.

The strongest agreement were on *diák* 'student' (0), *bíró* 'judge' (-0.25), *biztonsági őr* 'security guard' (-1.90), *tudós* 'scientist' (-0.40), and *orvos* 'doctor' (-0.55).

3.1.2 Intra-language gender differences in the Hungarian data

We also ran a two-sample *t*-test to compare the ratings of male and female participants for each occupation, and see if there was any discrepancies between the two groups. The only job that showed a significant difference was *rendőr* 'police', where the male bias was much higher by male raters (-1.77 vs. -1.00). The results are summarized in Figure 6. Furthermore, it seems like men's ratings tended to have a greater absolute bias for both male- and female-coded jobs (See Figure 10 below).

3.2 Chinese

Similarly to Hungarian, we found that a majority of occupations in Chinese were also rated with significant gender bias. The results of the one-sample

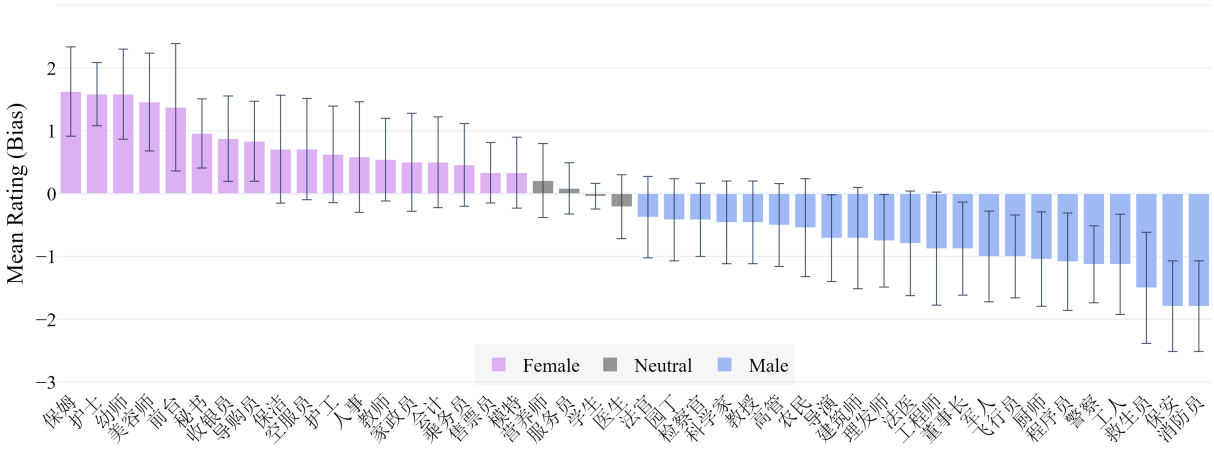

Figure 7: Mean ratings of occupational titles in Chinese with standard deviations, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

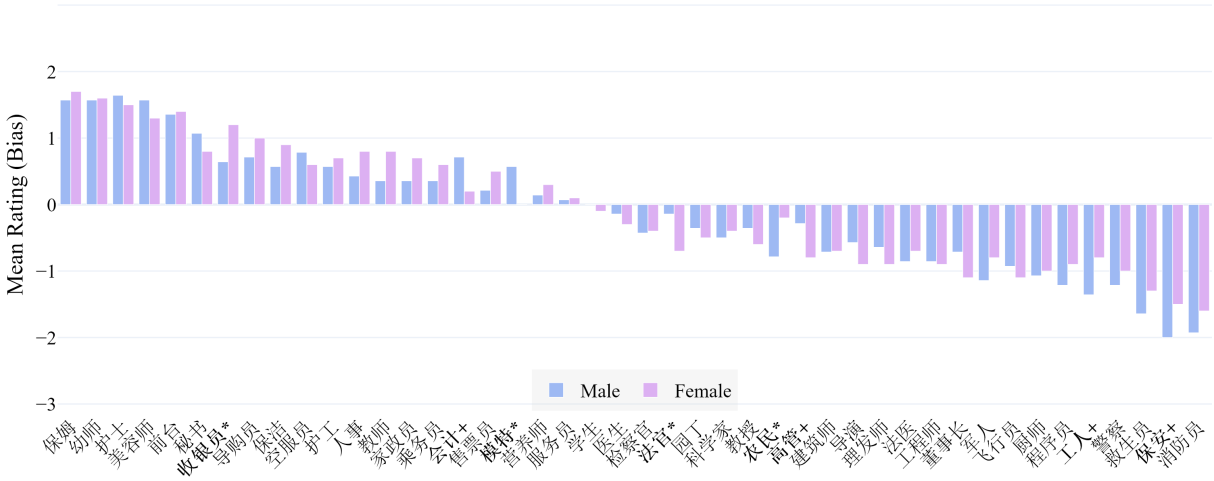

Figure 8: Mean ratings of occupational titles in Chinese by gender, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

t-test showed that 39 out of 44 occupations were biased. The mean ratings are shown in Figure 7.

— Expectations...?

In Chinese, the words with the highest female bias were 保姆 'domestic helper' (1.63), 护士 'nurse' (1.58), 幼师 'kindergarten teacher' (1.58), 美容师 'beautician' (1.46), and 前台 'receptionist' (1.38). Words with the highest male bias were 警察 'police officer' (-1.13), 工人 'worker' (-1.13), 救生员 'lifeguard' (-1.50), 保安 'security guard' (-1.79), and 消防员 'firefighter' (-1.79), which shows a relatively strong similarity to the Hungarian trends.

The 4 job titles that were not significantly biased were: 营养师 'dietetician' (0.21), 服务员 'waiter/server' (0.08), 学生 'student' (-0.04), and 医生 'doctor' (-0.21).

In Chinese too, there was strong agreement on 学生 'student' (-0.04) and 医生 'doctor' (-0.21), the remaining words in the top 5 jobs with the lowest standard deviation were 服务员 'waiter' (0.08), 售票员 'ticket seller' (0.33), and 护士 'nurse' (1.58).

3.2.1 Intra-language gender differences in the Chinese data

The two-sample *t*-test comparing the ratings of male vs. female participants showed that there were significant differences between what people think of a typical 收银员 'cashier' (m=0.64; f=1.20), 模特 'model' (m=0.57; f=0), 法官 'judge' (m=-0.14; f=0.70), and 农民 'farmer' (m=-0.79; f=-0.20). The results are summarized in Figure 8.

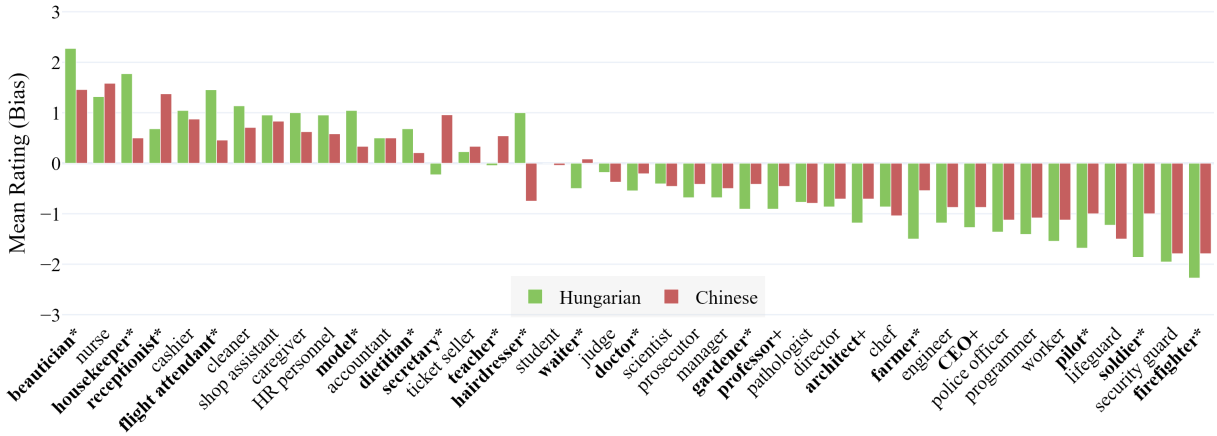

Figure 9: Mean ratings of common occupational titles in Hungarian and Chinese, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

3.3 Cross-linguistic comparison

The wordlists of the two datasets were almost, but not exactly the same; by performing an inner join on the two lists, we could pair 42 items together according to their meanings. Using a two-sample *t*-test, we checked if there were significant differences between the ratings in the two languages. The results are summarized in Figure 9.

When comparing the two sets of ratings, the first noticeable trend is that in general, the two languages have similar biases for the same occupations. Shared items on the extreme ends of the scale include *nurse* and *beautician*, and *firefighter* and *security guard*. Job titles that were unbiased in both datasets were *server* (*felszolgáló*, 服务员), and *student* (*diák*, 学生).

In terms of significance, we found 13 occupations that were rated differently in the two languages. These include *beautician* (*kozmetikus* 2.20 vs. 美容师 1.46), *housekeeper* (*házvezető* 1.80 vs. 家政员 0.59), *receptionist* (*recepciós* 0.65 vs. 前台 1.38), *flight attendant* (*légiutas-kísérő* 1.40 vs. 乘务员 0.46), *model* (*modell* 1.00 vs. 模特 0.33), *dietetician* (*dietetikus* 0.65 vs. 营养师 0.21), *secretary* (*titkár* -0.30 vs. 秘书 0.96), *hairdresser* (*fodrász* 1.00 vs. 理发师 -0.75), *waiter* (*pincér* -0.45 vs. 服务员 0.08), *farmer* (*földműves* -1.40 vs. 农民 -0.54), *architect* (*építész* -1.25 vs. 建筑师 -0.71), *pilot* (*pilóta* -1.65 vs. 飞行员 -1.00), and *soldier* (*katona* -1.80 vs. 军人 -1.00).

— housekeeper is not exactly the same as 家政员
— secretary has more than one meaning
— hairdresser is the most striking, discuss it

The most striking difference was the word for 'hairdresser' (*fodrász* vs. 理发师), which shows a strong female bias in Hungarian (1.00) but a definite male bias in Chinese (-0.75). We believe that this is a neat example for a cultural difference, in Hungary hairdressers are perceived to be predominantly female, while in China the profession is perceived to be more male-dominated. It is not difficult to find evidence for the latter from the press, even in English-language media.[6]

It would be interesting not only to explore the reasons behind this difference, but to compare actual figures on the gender distribution of hairdressers in the two countries, and see if the ratings reflect reality. (Is there any available workforce statistics on gender distribution per profession?)

4 Discussion

The first obvious thing to notice when looking at Figure 9 is that the overall trends in gender bias are quite similar between the two languages, despite some notable differences in specific occupations. This suggests that stereotypes are comparable and – mostly – consistent across developing societies.

4.1 Gender bias by language

An interesting feature is that Hungarian raters tended to rate occupations with a stronger bias than Chinese speakers. Out of 42 shared occupations, 31 were rated with a higher bias in Hungarian, while only 10 were rated with a higher bias in Chinese (excluding the 1 item with equal value). This is a threefold difference, and warrants further exploration to fully explain.

[6] https://www.chinadailyhk.com/hk/article/603100

4.2 Gender bias by gender?

An interesting divergence arose when comparing the two datasets and the differences between the ratings by gender. While in Hungarian, biases – both male and female – were stronger in the ratings of men, in Chinese women rated with a stronger bias on average, especially regarding female biases. You can compare the contrast in Figure 10.

4.3 Comparing the human ratings with AI-generated ratings

Now, we were also curious how the human ratings compare to the ratings of Large Language Models (LLMs) used in popular AI agents, and so we basically repeated the same experiment using AI instead of human raters. We prompted Copilot (+Think Deeper), ChatGPT (+Reason), Gemini (2.5 Flash), and Deepseek (Deepthink R1) to elicit a rating on the same job titles in both languages, using the same scale as the human raters. The instructions were in Hungarian and in Chinese, respectively, with a pretext: "You are participating in an experiment and your answers will help our research.". We asked the AI agents 10 separate times, and took the mean as a baseline for the results for each job title. The mean AI ratings were then superimposed on the human raters' plots to allow for a convenient comparison, you can see these in Figures 11 and 12.

The rationale behind this was to simulate how the general public and non-experts would turn to AI to do the job of human raters, and to show that these practices can be highly problematic, as the biases encoded in the LLMs are expected to be even stronger than those of the human raters. Mention real studies on this and cite. We think that the results perfectly illustrate the dangers of using AI agents instead of human raters, and only perpetuate social biases.

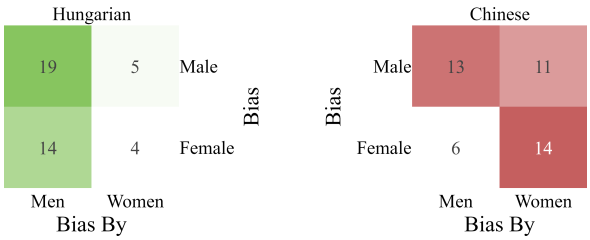

Figure 10: Confusion matrix of the Hungarian and Chinese ratings, showing the differences in the ratings of male and female participants.

5 Conclusion

In conclusion, our study provides evidence for occupational gender bias in both Hungarian and Chinese contexts, using data from native speakers. We have demonstrated that gender stereotypes are present in job titles, even if the words themselves are unmarked for gender, and this is due to societal norms and cultural influences, reflecting deep-seated biases.

For the sake of reproducibility and open science, we have made all the data available... please find all raw data and code in this repository: https://github.com/partigabor/occupational-bias.

ANONYMIZE THE FILES BEFORE SUBMISSION!

References

畢新偉 Xinwei Bi. 2013. 「她」字的來源與女性主體性——評《「她」字的文化史》 [The Origin of Chinese Characters and Female Subjectivity: A Review of "The Cultural History of the Chinese Character 'She' "]. 二十一世紀 *[21st century]*, (136):136–142.

Elisabeth Kaukonen, Polina Oskolskaia, Liina Lindström, and Raili Marling. 2025. Gender, language and labour: Gender perception of Estonian and Russian occupational titles. *Frontiers in Communication*, 9.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In *LREC*, volume 9, pages 1719–1723.

Tamás Váradi. 2002. The Hungarian National Corpus. In *LREC*, pages 385–389.

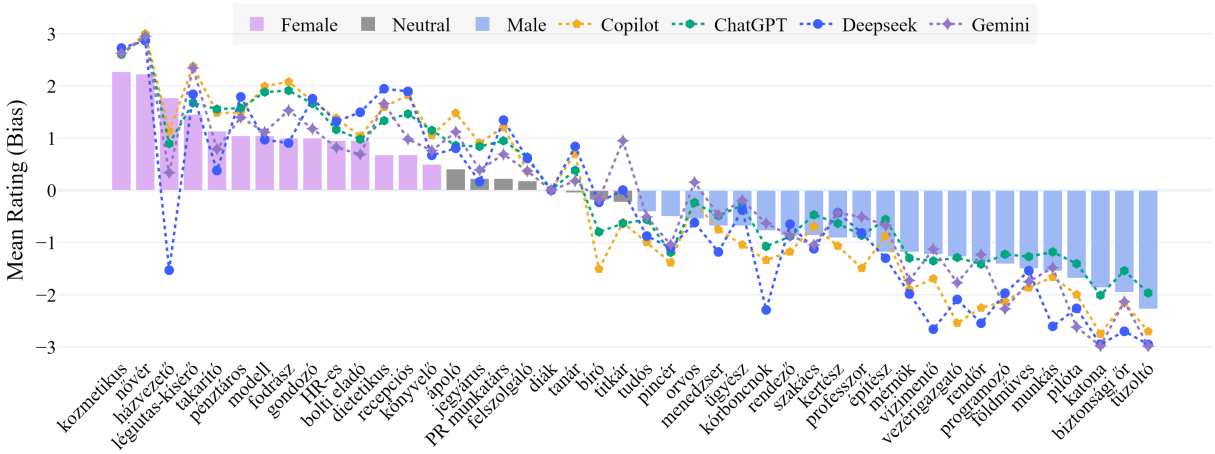

Figure 11: AI ratings for Hungarian occupations — explore the interactive plot.

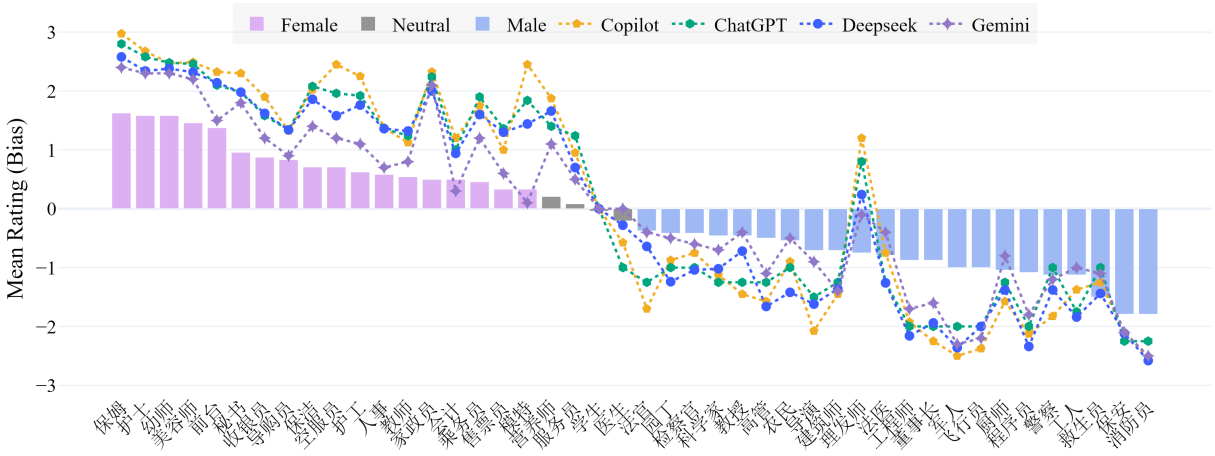

Figure 12: AI ratings for Chinese occupations — explore the interactive plot.