Occupational gender bias in ungendered languages: Comparing experimental data in Hungarian and Chinese

Anonymous ACL submission

Abstract

This paper is about occupational gender stereotypes, explored in a cross-linguistic setting. In the study, we analyze experimental data collected from Hungarian and Mandarin Chinese speakers on their ratings of job titles. 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. Results show that in both of these languages the words carry societal biases, despite the fact that the job titles themselves have no grammatical gender markings. We analyze and compare the ratings across linguistic and gender lines, highlight the differences, and discuss the results with insights ranging from peculiarities in word formation to generalizations in cross-cultural differences. Additionally, we also compared the human raters' responses with that of a few popular generative AI engines, which showed that the biases we humans carry are even stronger in the Large Language Models (LLMs) underlying these chatbots.

1 Introduction

I like Kaukonen et al. (2025).

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 Experiment setup

For both languages we devised a simple experiment in the form of surveys, where we asked participants to rate job titles on a scale. In both cases they were instructed to make decisions on how likely an occupation is 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

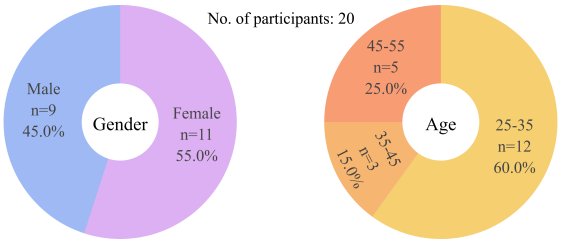

Figure 1: Demographics of the Hungarian participants.

A total of 22 native Hungarian speakers filled our questionnaire, and after validating the responses (i.e. reviewing attention checks and manually checking for anomalies) 2 were rejected. Participants were recruited online, with screeners set for location (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 50 items, each a commonly occurring job title in Hungary, such as: *modell* 'model' or *katona* 'soldier', in no particular order. Six of the words were attention-check items, which were removed from the final analysis. The attention checks were *pincérnő* 'waitress', *titkárnő* 'secretary (female)', *tanárnő* 'teacher (female)', *takarítónő* 'cleaning lady', *ápolónő* 'nurse (female)', and *házvezetőnő* 'housekeeper (female)'.

[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.

These words are compounds, they explicitly determine the gender of the worker by appending *-nő* 'woman' to the base word. If participants paid attention, all these items should be rated according to '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 also have their counterparts without the *nő* 'woman' element, i.e. *pincér* 'waiter', *titkár* 'secretary', *tanár* 'teacher', etc., these are unmarked for gender.

Common pairs include *énekes* 'singer' – *énekesnő* 'female singer', *színész* 'actor' – *színésznő* 'actress', and in such cases where both are well established, the unmarked word seems to carry some male bias, but it does not explicitly refer to a man. More interestingly, there are occupations where the unmarked form is the only one generally used for both genders, and appending *'-nő* 'woman' to it – although possible – would render it awkward – such as in *alkalmazottnő* 'female employee' or *programozónő* 'female programmer' – but not as uncanny as Modern English *singress*[2] would be. Furthermore, there are a few cases, where the female-marked version is so ubiquitous, that it is the unmarked version that will sound a bit odd, such as *házvezető* 'housekeeper', or to some extent *takarító* 'cleaner'.

In short, 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...? Languages with many genders?

The full list of words is as follows: *modell*, *katona*, *kórboncnok*, *vezérigazgató*, *menedzser*, *nővér*, *szakács*, *pincérnő*, *felszolgáló*, *könyvelő*, *professzor*, *építész*, *tudós*, *ápoló*, *pénztáros*, *bíró*, *munkás*, *vízimentő*, *titkárnő*, *jegyárus*, *tűzoltó*, *mérnök*, *tanárnő*, *rendező*, *takarító*, *HR-es*, *házvezető*, *légiutas-kísérő*, *pincér*, *takarítónő*, *orvos*, *fodrász*, *földműves*, *ápolónő*, *gondozó*, *bolti eladó*, *kertész*, *titkár*, *PR munkatárs*, *dietetikus*, *tanár*, *rendőr*, *pilóta*, *házvezetőnő*, *recepciós*, *biztonsági őr*, *ügyész*, *kozmetikus*, *programozó*.

We also 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-specific form for it, *diáklány* 'girl student'.

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

2.1.3 Procedure

Hungarian participants were instructed to rate each word on a 7-point Likert scale, from -3 to +3, where -3 meant 'completely male' and +3 meant 'completely female'. The scale presented in both questionnaires followed the same logic, from -3 to +3, moving from men to women, 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 in translation would be: "Is this occupation typically a man's occupation or a woman's occupation?".

The survey was distributed online, 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.[3]

2.2 Chinese survey

2.2.1 Participants

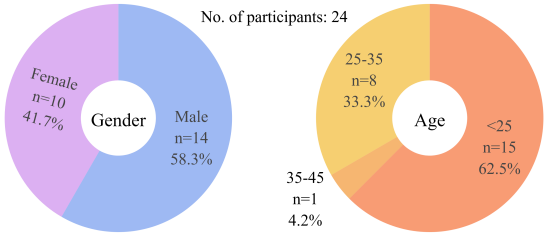

Figure 2: Demographics of the Chinese participants.

The Chinese survey was completed by 30 native Mandarin Chinese speakers, 6 of which were rejected after failing attention checks. 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 2 for the distribution.

[3]Please see a sample survey here. Is this necessary?

2.2.2 Materials

The Chinese survey also contained 50 items with commonly occurring job titles in Mandarin Chinese (Simplified), also in randomized order. There were six attention checks included to ensure participant engagement and data quality, these 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'. Similarly to the Hungarian attention checks, 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 3 or -3 were rejected.

— Say something about Chinese unmarked job titles?

Here is the full list of Chinese job titles: 警察, 秘书, 教授, 护士, 高管, 教师, 前台, 工人, 幼师, 妈妈, 模特, 护工, 保姆, 会计, 女画家, 工程师, 保洁, 法官, 导购员, 美容师, 服务员, 女作家, 乘务员, 理发师, 空服员, 售票员, 厨师, 营养师, 家政员, 收银员, 爸爸, 医生, 法医, 程序员, 保安, 导演, 军人, 董事长, 农民, 学生, 园丁, 飞行员, 人事, 男画家, 消防员, 科学家, 男作家, 检察官, 救生员, 建筑师

2.2.3 Procedure

Chinese participants, too, were instructed to rate each word on a 7-point Likert scale, from -3 to +3, with the same scaling as explained above. The exact wording of the question in the Chinese survey (in translation) was: "What do you think is the ratio of men to women in `occupation_name`?". The survey was also distributed online with essentially identical instructions, and participants saw each word highlighted in the question above, with no other context, and were presented with the scale directly below.

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.

2.4 Comparing the human ratings with AI-generated ratings

We were also curious how the human ratings compare to the ratings of Large Language Models (LLMs) used in popular AI agents. 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 averaged the results for each job title. The AI ratings were then superimposed on the human raters' plots to allow for a convenient comparison.

The aim of 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. The results of this comparison are discussed in section 4.

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. See Figure 3 for a visualization of the mean ratings, with the gender biases highlighted.

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*

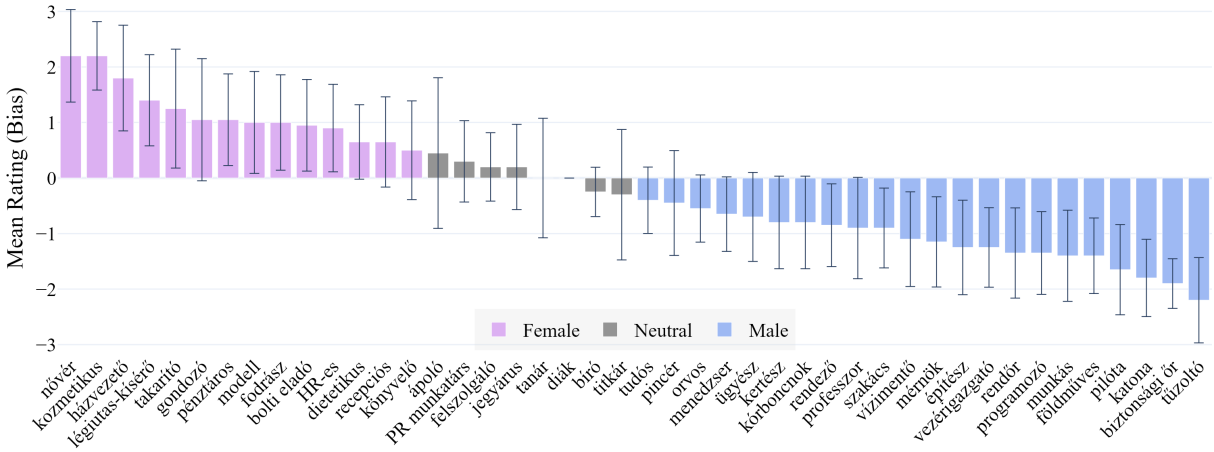

Figure 3: Mean ratings of occupational titles in Hungarian with standard deviations, significant gender bias highlighted – explore the interactive plot.

'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 feminine 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 *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 variety in the ratings, with 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).

3.1.1 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

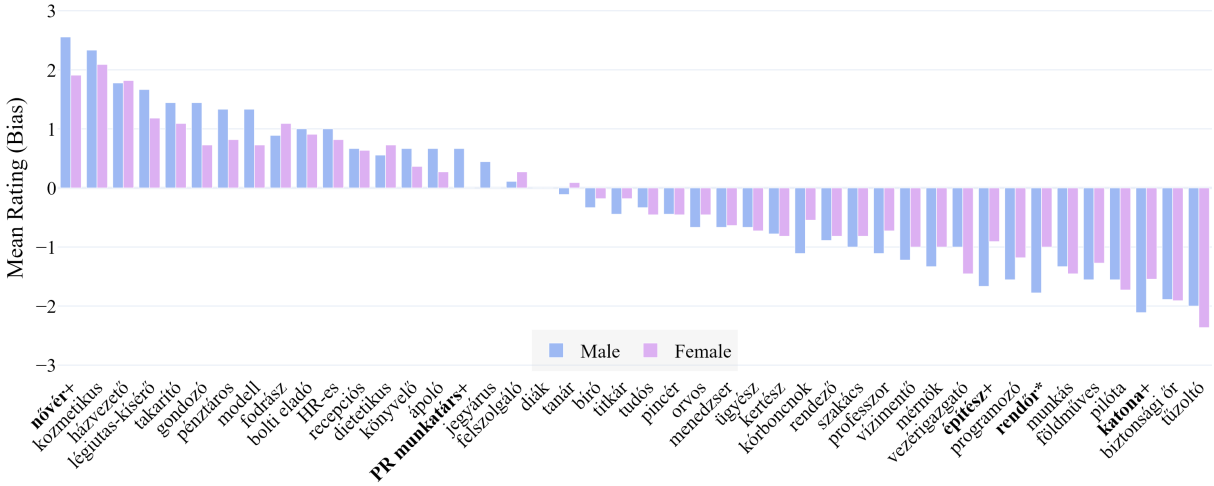

Figure 4: Mean ratings of occupational titles in Hungarian by gender, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

vs. -1.00). The results are summarized in Figure 4. Furthermore, it seems like men's ratings tended to have a greater absolute bias for both male- and female-coded jobs (See 8 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 *t*-test showed that 39 out of 44 occupations were biased. The mean ratings are shown in Figure 5.

— Expectations...?

In Chinese, the words with the highest feminine 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 6.

3.3 Cross-linguistic comparison

The wordlists of the two datasets were not exactly the same, but 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 7.

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*. Whereas job titles that were unbiased in both datasets were *server* (*felszolgáló*, 服务员), and *student* (*diák*, 学生).

— Explain...

The most striking difference was the word for 'hairdresser' (*fodrász* vs. 理发师), which shows a strong female bias in Hungarian (1.00) and a strong male bias in Chinese (-0.75). Discuss this...

4 Discussion

4.1 Who are the biased ones?

An interesting difference 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 8.

4.2 AI vs. Human?

This illustrates perfectly the dangers of using AI agents instead of human raters...

References

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.

A Example Appendix

This is an appendix.

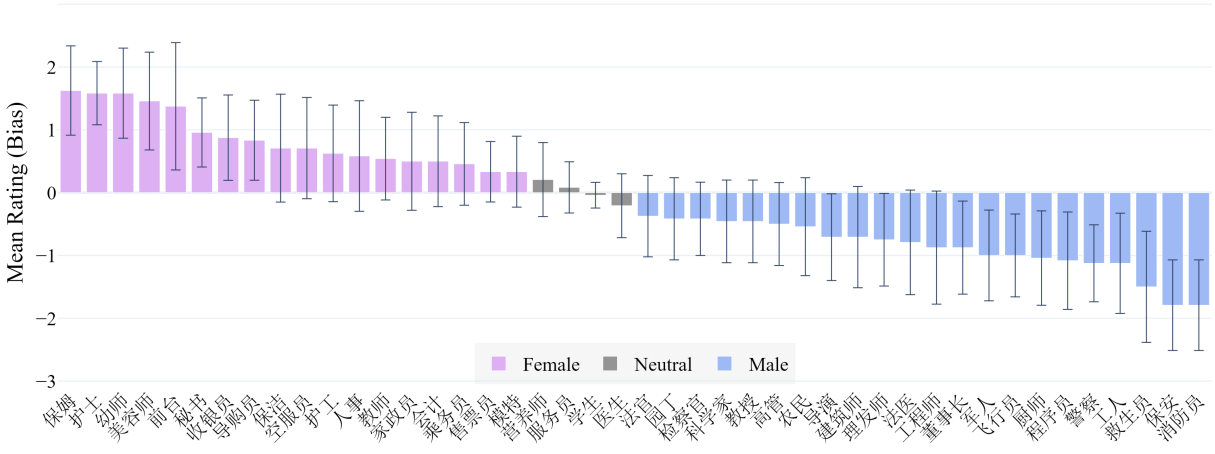

Figure 5: Mean ratings of occupational titles in Chinese with standard deviations, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

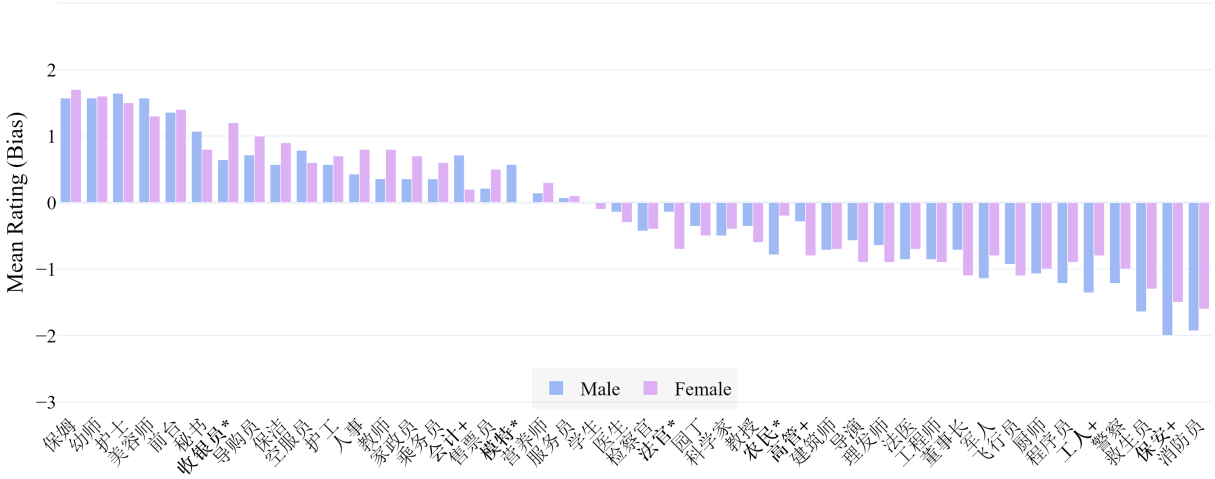

Figure 6: Mean ratings of occupational titles in Chinese by gender, significant differences highlighted (significant*, and marginally significant+ in **bold**) – explore the interactive plot.

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

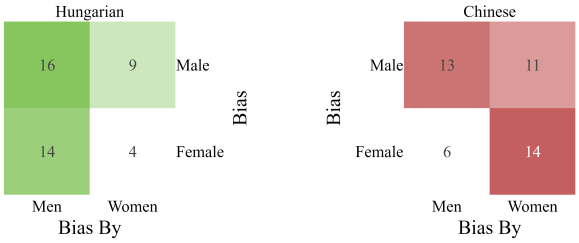

Figure 8: Confusion matrix of the Hungarian and Chinese ratings, showing the differences in the ratings of male and female participants.

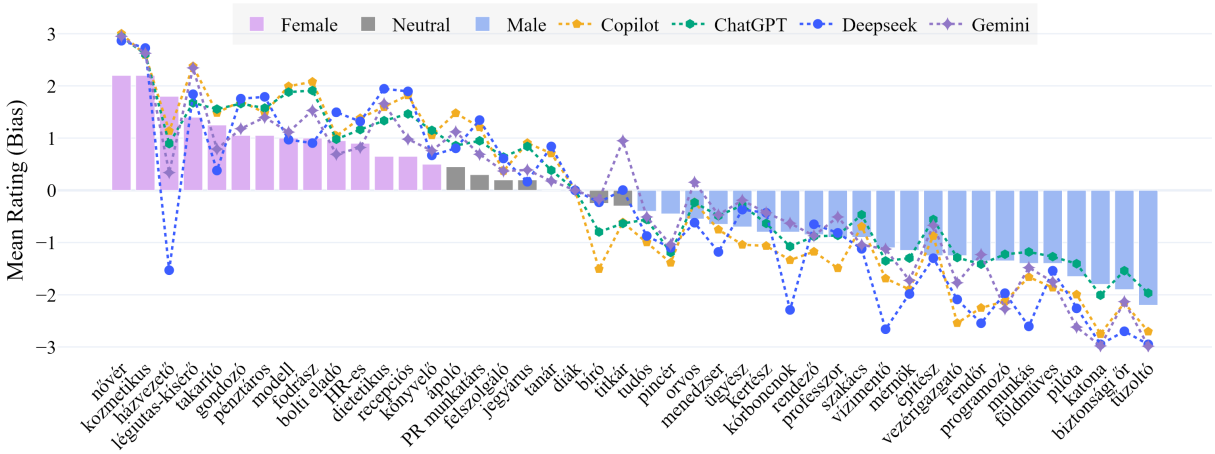

Figure 9: AI ratings for Hungarian occupations. explore the interactive plot.

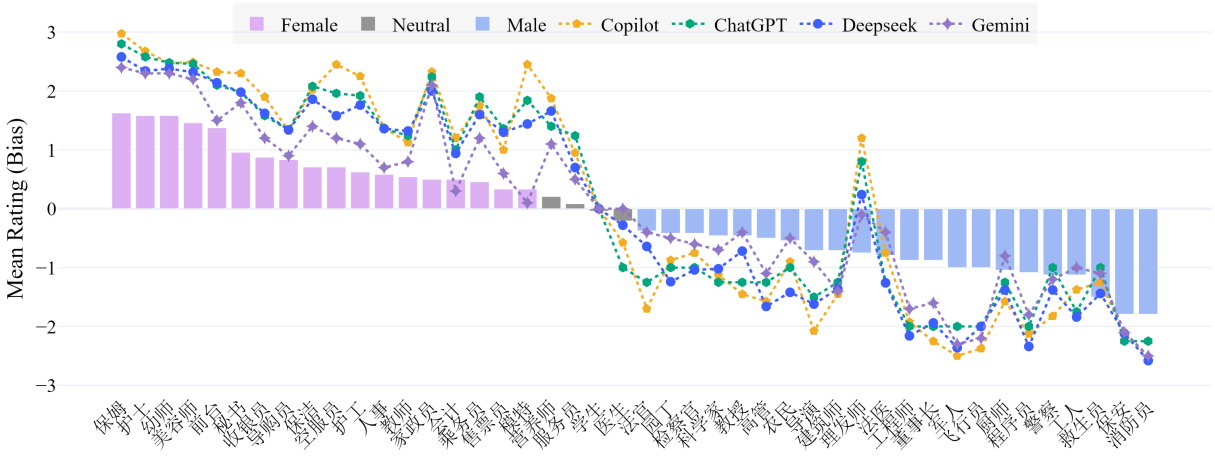

Figure 10: AI ratings for Chinese occupations. explore the interactive plot.