Using Large Language Models to Classify Cultural Heritage

Christoffer Cappelen [*]

April 2024

Abstract. Large language models holds potential to classify and annotate textual data in new and creative ways. Unlike traditional models for annotating text-as-data, they are capable of annotation tasks that require contextual reasoning and interpretation of high-level semantic meaning. In this paper, I test how OpenAI's GPT performs in comparison to crowd workers when it comes to annotation of cultural attributes in less conventional text genres. Replicating Michalopoulos and Xue's (2021) classification of gender stereotypes in folklore motifs, I show that GPT annotations (1) are on par with or better than MTurkers when it comes to consistency and (2) yield classifications similar to those of human annotators.

1. Introduction

Culture—the shared norms, values, and beliefs of a society—matters for a range of economic and political outcomes, e.g., economic development, political institutions, policy preferences, and conflict. Scholars in (historical) political economy and related disciplines have therefore also increasingly come to recognize the importance of culture and cultural persistence (Alesina and Giuliano 2015; Nunn 2012, 2021; Giuliano and Nunn 2020; Giavazzi, Petkov, and Schiantarelli 2019; Lowes 2024; Persson and Tabellini 2021; Acemoglu and Robinson 2023). Since, contemporary economic and political outcomes are, at least partially, the product of culture, we need to understand the roots of variation in cultural norms and values.

[*]University of Copenhagen / Stanford University

Studies of cultural persistence often focus on non-cultural determinants of contemporary variations in culture (e.g., Alesina, Giuliano, and Nunn 2013; Hansen, Jensen, and Skovsgaard 2015; Nunn 2008). However, to better understand the mechanisms of cultural persistence, we need to study the emergence and evolution of historical norms and values. Tracing the evolution of culture centuries or even millennia ago is constrained by our access to information on those norms and values of past generations. For modern societies, we can simply survey people across societies. But we cannot survey the dead. The alternative is, then, typically to use written documents that have survived the test of time.

The written language is arguably one of humanity's greatest inventions. Not only was it central to early state formation (Stasavage 2021). It also allowed for important information to be recorded and kept for future generations and, thus, allowing historical researchers to gain a glimpse into the past. Historical documents provide a rich source of information about people (e.g., censuses), government, economics, etc. Information that HPE scholars have collected and explored in new and creative ways.

They also provide more than "cold facts". They open a window into the minds of those people who are long gone. The stories and songs that were passed down through generations and, eventually, put on paper reveal information about the norms, values, and beliefs held by our ancestors. But extracting those cultural traits from large amounts of text can be costly and time-consuming. Scholars have come up with many creative ways to capture certain norms and values such as individualism/collectivism (Knudsen 2024; Bazzi, Fiszbein, and Gebresilasse 2020; Greenfield 2013) or gender norms (Singla and Mukhopadhyay 2022). And advances in computational methods for analyzing text (i.e., natural language processing, NLP)—along with increased digitalization of historical archives—have over the last two decades enables a growing quantitative literature of historical social science research.[1]

Yet, the quantitative analysis of textual data has traditionally been limited to relatively simple analytical tasks building on "bag of words", lexical meaning, sentence semantics, and word embeddings. For more complicated tasks that require reasoning based on contextual knowledge and inference about authors' intentions, these models tend to perform

[1] An alternative to written documents is images which provide a rich source of information about cultural evolution. See (Voth and Yanagizawa-Drott 2023) for a recent creative attempt to use visual data to trace the evolution of culture.

poorly (Törnberg 2023); hence, the need for human annotators—either for the full dataset or to train more fine-tuned machine learning models—which can, however, be prohibitively expensive (in terms of both costs and time) and require a high level of technical expertise.[2]

The rise of new large language models (LLMs), such as OpenAI's ChatGPT, opens up many new possibilities for textual analysis. Unlike conventional NLP methods, these zero-shot (or few-shot) models enable more complicated textual analysis that are sensitive to context and semantic meaning at a high level (e.g., sentence, paragraph, or even document level) while requiring relatively little technical skill to use. Thus, they hold potential for a wide range of new applications and analyses of textual data that were until now reserved for human annotators at a fraction of the time and cost.

Whether and exactly in what contexts such models are useful, scholars are only beginning to explore. They have been shown to perform well for classifying text into news and not news, detecting hate-speech and misinformation, rating the credibility of news outlets, classifying political affiliation of Twitter posts, and estimating the ideological position of politicians (Reiss 2023; Hoes, Altay, and Bermeo 2023; Huang, Kwak, and An 2023; Kuzman, Mozetič, and Ljubešić 2023; Qin et al. 2023; Yang and Menczer 2023; Törnberg 2023; Wu et al. 2023). Gilardi, Alizadeh, and Kubli (2023) also show that ChatGPT outperforms crowd workers across a range of annotation tasks (relevance, stance, topics, and frames detection) in both reliability and accuracy.

Many of these early applications have demonstrated LLMs' ability to annotate modern—and typically relatively simple—pieces of text, e.g., news items and social media posts. And they almost exclusively perform tasks on text genres on which LLMs' training data are likely to draw overwhelmingly from. The question thus remains how well LLMs perform when confronted with more complex classification tasks requiring contextual reasoning and less conventional, less contemporary genres of text which undoubtedly compose only a minor part of the data used for training these models.

In this paper, I test LLMs' ability to detect gender stereotypes in folklore

[2] The use of crowd workers—such as Amazon MTurk—for annotation has also increasingly come into question with concerns of decreasing quality (Chmielewski and Kucker 2020). Crowd workers have even been shows to use LLMs themselves for text production tasks, severely limiting the reliability and quality of crowd sources text annotation (Veselovsky, Ribeiro, and West 2023).

motifs. I use data from Michalopoulos and Xue (2021) who use crowd workers from Amazon's MTurk platform to classify depictions of stereotypical traits of male and female characters. Comparison between classifications from OpenAI's GPT and the original data indicate that GPT classifications are (1) on par with or more consistent than the corresponding MTurk classifications, and (2) very similar to that of MTurkers. I also replicate several results from Michalopoulos and Xue (2021) and generally find similar or marginally stronger effect sizes. The results suggest great potential for LLMs to classify and annotate a wide variety of textual data to capture abstract cultural traits.

2. Folklore and gender roles

To uncover the cultural heritage of societies across the globe, Michalopoulos and Xue (2021, hereafter M&X) rely on a catalog of folklore collected by the anthropologist and folklorist Yuri Berezkin. Folklore "consists of the traditional beliefs, customs, and stories of a community, passed through the generations by word of mouth" (Michalopoulos and Xue 2021, 1994). While written documentation of the tales and myths are mainly from the 20th century, their history can typically be traced back centuries and sometimes even millennia. Thus, folklore offers insights into the cultural heritage of societies.

Berezkin's *Folklore and Mythology Catalog* is a global comparative database of the oral traditions for 958 groups worldwide. For these 958 groups, he categorized 2,564 *motifs*, "combination[s] of images, episodes, or structural elements found in two or more texts" (Michalopoulos and Xue 2021, 1996). It is these motifs that comprise the data used by M&X. The authors first validate the catalog of motifs by showing that images and episodes in a group's oral tradition reflect salient features of its physical (geographical) and social environment. They then classify these motifs along several cultural traits (e.g., gender roles, risk aversion, or trust) using either machine learning or human classification (MTurk) depending on the specific task.

In this paper, I focus on their coding of gender roles using MTurkers' classification of motifs along gender stereotypical attributes (e.g., violence, intelligence, domestic work). The authors first identify 1,073 motifs containing female and/or male characters and then ask MTurkers to indicate whether the male and/or female characters are depicted as any of 7 stereotypical traits.[3] The coding of these motifs are then matched to ethnic groups in the Ethnographic Atlas (EA) and to modern countries for which they further calculate several bias measures capturing, for instance, the extent to which male characters are depicted as more dominant than female characters.

3. Method

To test the performance of OpenAI's GPT in annotating gender stereotypes in less conventional texts, I use the database of folklore motifs from Michalopoulos and Xue (2021) who relied on MTurk to classify motifs.[4] I asked GPT to classify each motif according the the seven attributes (for female and male characters, respectively) using the following instructions:

> I need to classify a motif from folklore according to the presence of gender stereotypes. The motif is: "[Motif]".
>
> Is the FEMALE character(s) in the motif depicted as... (You may select more than one answer except for "No female character(s) present").
> [List of attributes]
>
> Is the MALE character(s) in the motif depicted as... (You may select more than one answer except for "No male character(s) present").
> [List of attributes]
>
> Please only focus on the FEMALE and MALE character(s), resepctively. You may select more than one answer (except for "No female character(s) present" or "No male character(s) present").

To ensure comparability, the prompt mimics the original instructions provided to MTurkers except for a few additions and tweaks to give the proper context and ensure a consistent output.[5]

[3] The seven traits are (1) Violent/Dominant/Aggressive, (2) Submissive/Dependent, (3) Physically active, (4) Engaged in domestic affairs, (5) Sexual, (6) Intelligent, (7) Naive/Stupid. They can also indicate depictions of other stereotypical traits ("Other") as well as indicate that there is no male/female character present. Except in the latter case, they are also allowed to indicate depictions of multiple traits in the same motif.

[4] The main results are based on GPT-3.5, which was the newest model available at the time. GPT-4 is more powerful, but also more expensive. A single run of GPT-4 yields results similar to that of GPT-3.5 and, thus, suggesting only marginal improvements for the annotation tasks in this paper.

[5] The first section (including the actual motif) is added to provide context. I also added "*except for 'No female character(s) present'*" which minimized the number of times ChatGPT would indicate both "no presence" and one of the attributes listed (which human classifiers would implicitly understand). The prompt also ends with additional instructions to ensure

I used ChatGPT-3.5 through the API.[6] Since GPT is non-deterministic, identical inputs can lead to different outputs.[7] The temperature parameter controls the degree of randomness in responses, with lower values creating more consistent outputs and higher values potentially generating more creative responses. To capture variability in responses within and across parameter values, I ran the model 3 times for each of 3 different temperature settings: 0.2, 0.7 (default), and 1.2.

To assess the reliability and accuracy of ChatGPT, I compare the resulting dataset to that of M&X using MTurk. While MTurkers' classifications cannot be considered the "ground truth" or "gold standard", it is often what research rely on. The question, thus, is whether ChatGPT generates more reliable classifications than MTurk and whether it agrees with MTurkers' classifications.

4. Results

The performance of ChatGPT is assessed in several ways. First, I calculate inter-coder reliability scores for each temperature setting to evaluate the consistency of responses. I also compare these to similar inter-coder reliability scores for MTurk. Second, I calculate agreement between the modal response of ChatGPT to the modal response of MTurk (again for each temperature setting). In the following section, I also replicate results from Michalopoulos and Xue (2021) on the relationship between male bias in folklore and female labor force participation and between ethnic groups' mode of food production and male bias in folklore.

Reliability

Inter-coder reliability is calculated as the share of motifs for which all coders (i.e. rounds of ChatGPT) agree on a particular attribute. Thus, I calculate a percentage agreement for each of the 9 possible classifications (including "Other" and "No (fe)male character(s) present" for male and female characters, respectively. While inter-coder reliability for ChatGPT re-

a consistently formatted output. It is likely that an alternative prompt could improve the model further, but I opted for maximum comparability as a first test of GPT's performance relative to that of MTurk. The full prompt is included in Appendix C.

[6] The model was run between May 16, 2023 and May 21, 2023. The exact model version is *gpt-3.5-turbo-0301*. The R code for running the model is available in the replication files.

[7] Even with additional instructions on the output format, it would occasionally fail to return the same output format. In these cases, I reran the model for those specific motifs.

lies on 3 rounds (per temperature setting), an average of 9 MTurkers coded each motif (93 % of the 1073 motifs were coded by 9). Since more coders increases the chances of disagreement, it would not be fair to compare percentage of complete agreement across the two. For MTurk, I therefore calculate the share of motifs for which more than two-thirds agree on a classification (i.e. the minimum agreement for 3 coders). As an alternative, I also calculate Krippendorff's Alpha which is more comparable across number of coders and better accounts for chance agreement.[8]

Figure 1 shows the average reliability scores across attributes for the three temperature settings of ChatGPT and MTurk. In general, ChatGPT generates very consistent classifications across temperature settings. As one would expect, the most consistent is also the lowest temperature (0.2) for which 95 percent of the motifs are agreed upon. Higher temperatures generate less consistent output, although it varies significantly across attributes (see Figures A1–A2 in Appendix A). MTurkers generally also agree on classifications with about 80 percent of classifications agreed upon by more than two-thirds. However, when looking at the Krippendorff's Alpha scores, the picture is significantly different. The low-temperature ChatGPT model still generates consistent responses ($\alpha = 0.86$), but the reliability drops significantly when increasing the temperature, and MTurkers are the least consistent with an alpha of only 0.20.

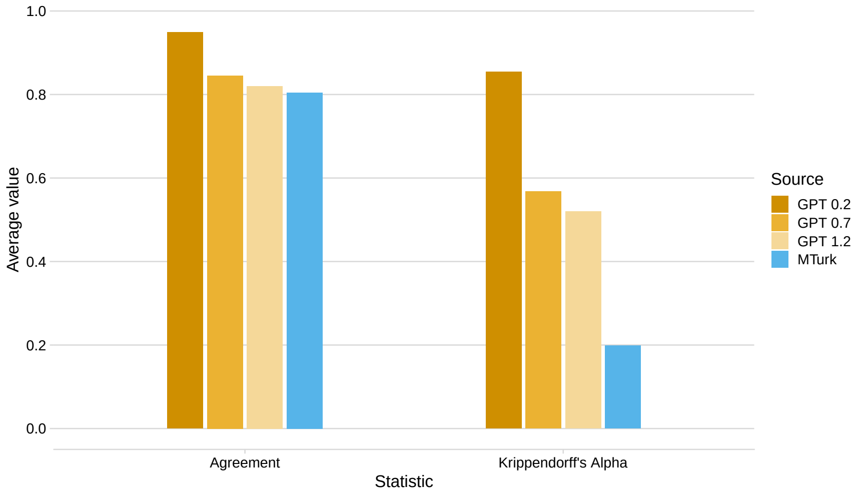
Accuracy

The second important question is whether, or to what extent, GPT generates responses similar to those of MTurkers (i.e., the benchmark). To assess this, I follow M&X's procedure of first calculating the modal response for each motif, that is, choosing only the attributes that most "coders" agree on (in case of ties, more than one attribute is allowed). To assess the accuracy, I calculate for each attribute the share of motifs for which GPT (for each temperature setting) and MTurk agree.

Figure 2 shows the average agreement across attributes for male and female characters, respectively. The figure shows a remarkable degree of similarity between the two. ChatGPT and MTurk agree on the presence of certain character traits in more than 75 percent of motifs. Interestingly, there is no discernible difference between temperature settings. If anything,

[8] Since Krippendorff's Alpha requires the same number of coders for each motif, I randomly select 9 for those with more than 9 coders and exclude the 12 motifs coded by less than 9.

FIGURE 1. Average intercoder reliability for GPT and MTurk



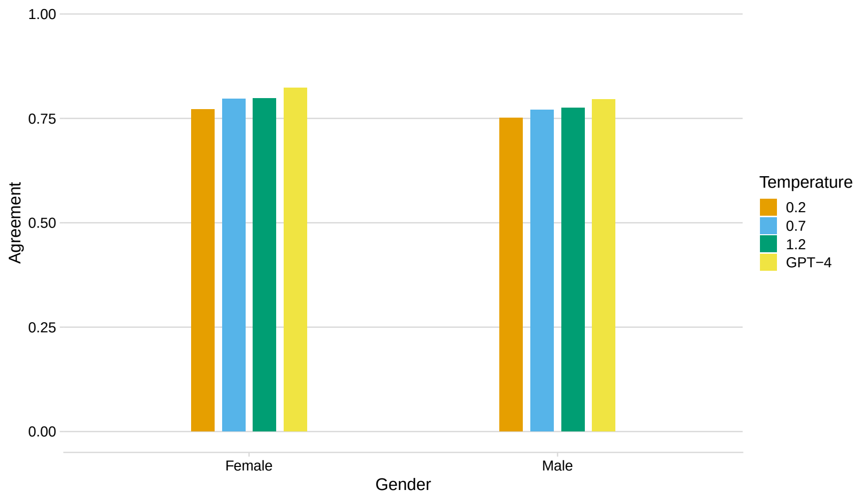
Note: Agreement is the share of motifs for which (1) all GPT round agree on coding or (2) more than two thirds of MTurk coders agree on coding.

the more consistent temperature of 0.2 appears marginally less accurate, but the difference is negligible. Given ChatGPT's non-deterministic nature, it is generally recommended to run the mode multiple times, not only to evaluate its consistency but also to aggregate responses. Sometimes, depending on the research task at hand, more creative (i.e. random) responses may be preferable, and even under those circumstances, aggregating multiple responses may still produce unbiased outputs.

The high average agreement, however, also masks significant variation across character traits (see Figure A5 in Appendix A). Most attributes fluctuate around 75 percent, and some traits such as ”Violent / Aggressive / Dominant” among female characters show very high agreement (more than 90 %). One notable exception is the trait ”Physically active” for male characters, for which ChatGPT and MTurk only agree on about 20–25 percent (depending on temperature setting). This may be due to ChatGPT for some reason having trouble identifying that character trait. On average, however, ChatGPT performs well when it comes to annotating character traits in folklore motifs.

The figure also shows accuracy for classifications based on GPT-4, the newer and more powerful model. Since the cost of GPT-4 is still markedly higher than that of GPT-3.5, I ran this version only once using a tempera-

FIGURE 2. Average accuracy of GPT coding



Note: Average share of motifs for which GPT and MTurk agree on the modal coding.

ture setting of 0.2.[9] Overall, the accuracy of GPT-4 is very similar to that of GPT-3.5, although marginally higher on average. Of the 16 individual traits, GPT-4 performs better than GPT-3.5 on 13 (81%). The most notable difference is for physically active male characters for which GPT-3.5 performed poorly. The accuracy more than doubles when using GPT-4. Overall, there is suggestive evidence that there is improved accuracy to be gained by GPT-4, although this must we weighed against the added cost (simpler annotation tasks may not gain much precision by using GPT-4).

5. Replicating Michalopoulos and Xue (2021)

The final test of GPT's performance is an attempt to replicate the empirical analyses in Michalopoulos and Xue (2021). Given the lack of a ”ground truth” to use as benchmark, there is in principle no way of knowing whether GPT outperforms MTurk or simply reproduces the same errors (or biases) as MTurkers. Replicating the analyses may provide additional

[9] Since I only ran GPT-4 once, there is no reliability measure. However, one can reasonably expect it to be at least on par with GPT-3.5. The comparison with GPT-4 also skips the step of calculating the modal response, again because there is only one round. Given the non-deterministic nature of GPT (even at a temperature setting of 0.2 one can expect the aggregation across multiple rounds to even out measurement error and, hence, the results shown here should be seen as a lower bound.

(albeit still not definite) information to adjudicate this. For instance, there is good theoretical and empirical reason so expect a cultural heritage of traditional gender roles to be associated with lower levels of female labor force participation today (e.g., Alesina, Giuliano, and Nunn 2013), which is also what M&X find. If the GPT classifications lead to similar—and especially stronger—associations, this may indicate that GPT actually does a better job at capturing gender stereotypes in folklore.

Following M&X, I aggregate motifs to the ethnic group level. Berezkin's original catalog contains information on the distribution of motifs across 958 ethnic groups. M&X match these groups to the ethnic groups in George P. Murdock's Ethnographic Atlas (Murdock 1967) and calculate the share of motifs present in the group's oral tradition depicting a specific character trait. They also aggregate to the country-level using the distribution of groups in a country's population along with migration data to calculate a weighted national-level measure of the share of motifs depicting a particular concept. I follow the same procedures to construct GPT data at the group and country level.

M&X calculate several indices of male bias to test the association between gender roles and, e.g., female labor force participation. First, they calculate the difference between the male and female shares for each character trait. Second, they calculate two composite indices (*male dominance bias* and male intelligence bias) by combining biases across several traits.[10] To ease comparison, I standardize both the original indices based on MTurk and the new indices based on ChatGPT ($\mu = 0$, $\sigma = 1$).

Female labor force participation

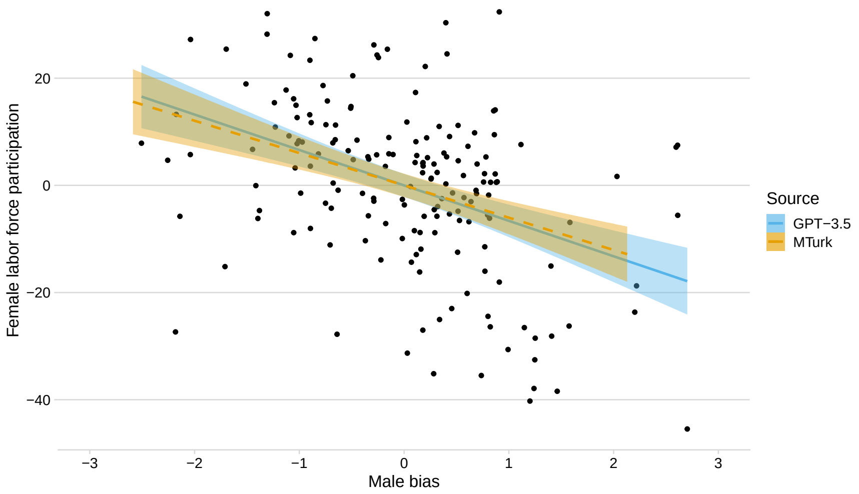
M&X document first a robust negative relationship between *male dominance bias* and female labor force participation. Using GPT yields almost identical results. Figure 3 shows the added variable plot of male dominance bias and female labor force participation.[11] The two lines represent the linear fit for GPT (blue) and MTurk (yellow), respectively, and are almost identical (regression coefficients are 6.64 and 6.03).

The same pattern is repeated in Table 1 showing the results of regress-

[10] For *male dominance bias*, they add male bias in violence and activeness and subtract male bias in submissiveness and domestic work. For *male intelligence* they add male bias in intelligence and subtract male bias in naivité. The latter index is originally termed "male mental capacity bias".

[11] Conditional on log year of first publication and log number of publications.

FIGURE 3. Male dominance bias and female labor force participation across countries



Note: Added variable plot of male dominance bias and female labor force participation conditional on log year of first publication and log number of publications. Points show values of male dominance bias coded by GPT. The fitted lines represent male dominance bias coded by GPT and MTurks, respectively.

TABLE 1. Male bias and female labor force participation, GPT-3.5

	Dominance bias		Intelligence bias		Sexual bias	
	MTurk	GPT	MTurk	GPT	MTurk	GPT
	(1)	(2)	(3)	(4)	(5)	(6)
Dominance bias	-5.87***	-6.80***				
	(1.26)	(1.20)				
Intelligence bias			2.06*	-5.22***		
			(1.19)	(1.28)		
Sexual bias					-0.049	-1.51
					(1.28)	(1.34)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	174	174	174	174	174	174
Adjusted R^2	0.145	0.190	0.051	0.122	0.034	0.041

Note: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors in parentheses.

ing female labor force participation on several bias indices. Again, the model includes controls for year of first publication and number of publication as well as continental fixed effects. Columns (1) and (2) shows the coefficients for male dominance bias. For both MTurk and ChatGPT, there is a strong negative (and statistically significant) association between bias and labor force participation. The coefficient based on ChatGPT is actually somewhat stronger than MTurk, although the difference is marginal.

Columns (3) and (4) document the relationship between *male intelligence bias* and female labor force participation. Originally, M&X find a small positive association between male bias and labor force participation, although this is only marginally significant. Interestingly, the index based on GPT yields a strong, negative (and significant) coefficient, indicating that countries with an oral tradition of depicting men as more intelligent and women as more naive are associated with lower female labor force participation, which is more in line with our theoretical priors. Finally, for both MTurk and GPT there is a negative relationship between *sexual bias* and labor force participation with ChatGPT again showing a stronger coefficient, although neither is statistically significant.

Results are generally also similar when looking at male bias for the individual character traits, although GPT typically yield stronger coefficients (see Figure A4 in Appendix A). The most notable exception is again intelligence where MTurk produces a small positive (and insignificant) coefficient, whereas ChatGPT yields a stronger negative (and significant) coefficient. Results are also virtually identical when looking at attitudes of second-generation immigrants in Europe (Table A1).

Agriculture and the plow

Finally, M&X also address the causes of male bias in societies' oral traditions by looking at food production and agricultural practices across ethnic groups (EA). One of the most popular explanations for the emergence of a division of labor based on sex and resulting cultural norms and values is early adoption of agriculture and in particular the adoption of the plow, which requires more upper body strength and hence favored men in agriculture (Boserup 1970; Alesina, Giuliano, and Nunn 2011, 2013; Hansen, Jensen, and Skovsgaard 2015). Relying on information from the Ethnographic Atlas, they show that *male dominance bias* is higher in societies where men contributed more than women in agriculture and when

TABLE 2. Male bias and agriculture, Ethnographic Atlas

	Agricultural contribution		Plow indigenous	
	MTurk	GPT	MTurk	GPT
	(1)	(2)	(3)	(4)
Men contribute more	0.188** (0.083)	0.274** (0.118)		
Plow indigenous			0.359** (0.177)	0.556*** (0.209)
Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Observations	702	702	1,131	1,131
Adjusted R^2	0.201	0.295	0.179	0.292

Note: $^*p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors clustered by linguistic family in parentheses.

the plow is indigenous to a society.

Table 2 repeats the main models from this analysis (again controlling for year of first publication, number of publications, and country fixed effects).[12] Again, GPT yields remarkably consistent results. Male dominance bias is higher in societies where men contribute more than women in agriculture (Columns 1–2). And it is higher where the plow is indigenous to society (Columns 3–4). Again, the coefficients for GPT is significantly stronger (about 50 percent), which may be indicative of less attenuation bias due to reduced measurement error.[13]

6. Discussion

LLMs hold potential for a wide range of new text-as-data applications. They open the door to automated annotation of textual data requiring contextual reasoning and interpretation in a way that earlier generations of NLP could not. Historical political economy often rely on resource intensive data collection, such as textual data, and LLMs enable researchers to easily scale research designs.

[12] Column (1) corresponds to Column (3) in Table VII, Panel C in Michalopoulos and Xue (2021) and Column (3) corresponds to Column (6).

[13] Section B in the appendix, repeat the above analyses using GPT-4 instead of GPT-3.5. Overall, results are very similar.

Just as important, these zero-shot models allow researchers with little experience in natural language processing and machine learning to capture even complicated and abstract concepts in text. Code for accessing APIs are readily available and can easily be adapted to specific applications. The main challenge is rather to design a prompt that minimizes error and delivers the desired response.

It is also cost-effective compared to human annotators. The crowd workers in Michalopoulos and Xue (2021) were paid between 0.08 and 0.20 USD per motif. With an average of 9 coders per motif, this amounts to a total cost of about 290–1930 USD. The cost of running GPT-3.5 9 times for each motif totalled 14.76 USD. The price of GPT-3.5 has since been reduced significantly and prices continue to drop. For GPT-4, the same amount of work would total about 110 USD. Thus, even in the most conservative comparison, LLMs would reduce the cost of more than 60 percent. And more likely, the cost will lie in the range 1 to 10 percent compared to human annotators (even more for expert annotators).

As several reviews have pointed out, LLMs are still not without limitations. The quality varied widely between applications and there are still issues of replicability and bias (Ollion et al. 2023; Qin et al. 2023; Reiss 2023). It is therefore still highly advised to carefully validate the use of LLMs in each specific application. How to most effectively use LLMs—e.g., in terms of how to design the best prompt—is also still left to individual researchers. Further research should explore how variations in prompts affect responses and develop guidelines for how to design accurate and efficient prompts.

References

Acemoglu, Daron, and James Robinson (2023). Culture, Institutions and Social Equilibria: A Framework. *Working paper.*

Alesina, Alberto, and Paola Giuliano (2015). Culture and Institutions. *Journal of Economic Literature* 53 (4): 898–944.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn (2011). Fertility and the Plough. *American Economic Review* 101 (3): 499–503.

——— (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics* 128 (2): 469–530.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse (2020). Frontier Culture: The Roots and Persistence of "Rugged Individualism" in the United States. *NBER Working Paper.*

Boserup, Ester (1970). *Woman's Role in Economic Development.* London: George Allen / Unwin Ltd.

Chmielewski, Michael, and Sarah C. Kucker (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science* 11 (4): 464–473.

Giavazzi, Francesco, Ivan Petkov, and Fabio Schiantarelli (2019). Culture: persistence and evolution. *Journal of Economic Growth* 24 (2): 117–154.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *arXiv,* eprint: 2303.15056.

Giuliano, Paola, and Nathan Nunn (2020). Understanding Cultural Persistence and Change. *The Review of Economic Studies* 88 (4): 1541–1581.

Greenfield, Patricia M. (2013). The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science* 24 (9): 1722–1731.

Hansen, Casper Worm, Peter Sandholt Jensen, and Christian Volmar Skovsgaard (2015). Modern gender roles and agricultural history: the Neolithic inheritance. *Journal of Economic Growth* 20 (4): 365–404.

Hoes, Emma, Sacha Altay, and Juan Bermeo (2023). Leveraging ChatGPT for Efficient Fact-Checking.

Huang, Fan, Haewoon Kwak, and Jisun An (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *arXiv,* eprint: 2302.07736.

Knudsen, Anne Sofie Beck (2024). Those Who Stayed: Selection and Cultural Change in the Age of Mass Migration. *Working paper.*

Kuzman, Taja, Igor Mozetič, and Nikola Ljubešić (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. *arXiv,* eprint: 2303.03953.

Lowes, Sara (2024). Culture in Historical Political Economy. In *Oxford Handbook of Historical Political Economy,* edited by Jeffery A. Jenkins and Jared Rubin, 887–924. Oxford: Oxford University Press.

Michalopoulos, Stelios, and Melanie Meng Xue (2021). Folklore. *The Quarterly Journal of Economics* 136 (4): 1993–2046.

Murdock, George P. (1967). *Ethnographic Atlas.* University of Pittsburgh Press. University of Pittsburgh Press.

Nunn, Nathan (2008). The Long-Term Effects of Africa's Slave Trades. *The Quarterly Journal of Economics* 123 (1): 139–176.

——— (2012). Culture and the Historical Process. *Economic History of Developing Regions* 27 (sup1): S108–S126.

——— (2021). History as Evolution. In *The Handbook of Historical Economics,* edited by Alberto Bisin and Giovanni Federico, 41–91. London: Academic Press.

Ollion, Etienne, Rubing Shen, Ana Macanovic, and Arnault Chatelain (2023). ChatGPT for Text Annotation? Mind the Hype!

Persson, Torsten, and Guido Tabellini (2021). Culture, Institutions, and Policy. In *The Handbook of Historical Economics,* edited by Alberto Bisin and Giovanni Federico, 463–490. London: Academic Press.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *arXiv,* eprint: 2302.06476.

Reiss, Michael V (2023). Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. *arXiv,* eprint: 2304.11085.

Singla, Shikhar, and Mayukh Mukhopadhyay (2022). Gender Norms Do Not Persist But Converge Across Time. *SSRN Working Paper.*

Stasavage, David (2021). Biogeography, Writing, and the Origins of the State. In *The Handbook of Historical Economics,* edited by Alberto Bisin and Giovanni Federico, 881–902. London: Academic Press.

Törnberg, Petter (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv,* eprint: 2304.06588.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West (2023). Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv,* eprint: 2306.07899.

Voth, Hans-Joachim, and David Yanagizawa-Drott (2023). Image(s). *Working paper.*

Wu, Patrick Y, Jonathan Nagler, Joshua A Tucker, and Solomon Messing (2023). Large Language Models Can Be Used to Scale the Ideologies of Politicians in a Zero-Shot Learning Setting. *arXiv,* eprint: 2303.12057.

Yang, Kai-Cheng, and Filippo Menczer (2023). Large language models can rate news outlet credibility. *arXiv,* eprint: 2304.00228.

Online Appendix

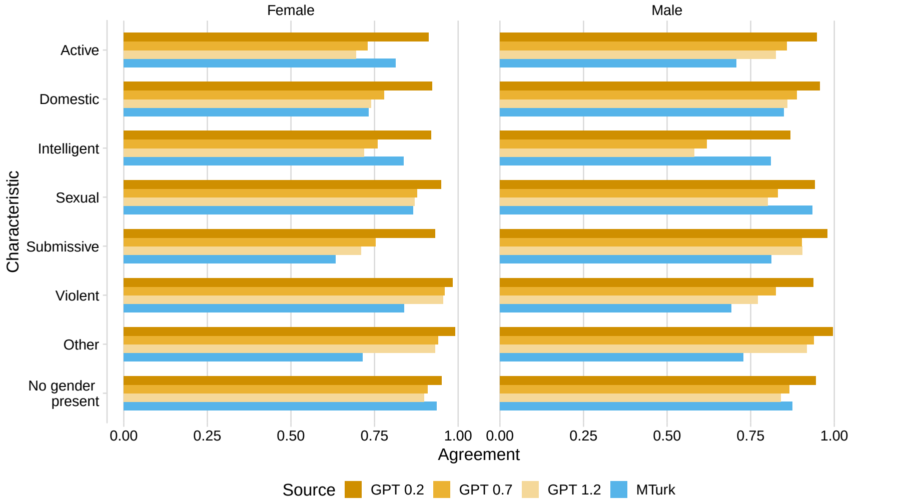
Using Large Language Models to Classify Cultural Heritage

Christoffer Cappelen

April 2024

A. ADDITIONAL FIGURES AND TABLES

FIGURE A1. Agreement of GPT and MTurk



Note: Agreement is the share of motifs for which (1) all GPT round agree on coding or (2) more than two thirds of MTurk coders agree on coding.

FIGURE A2. Krippendorff's for GPT and Mturk codings

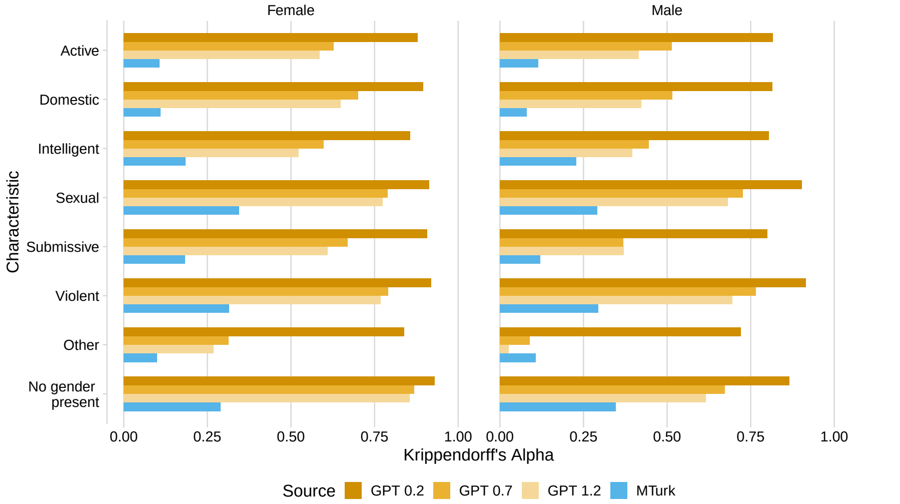
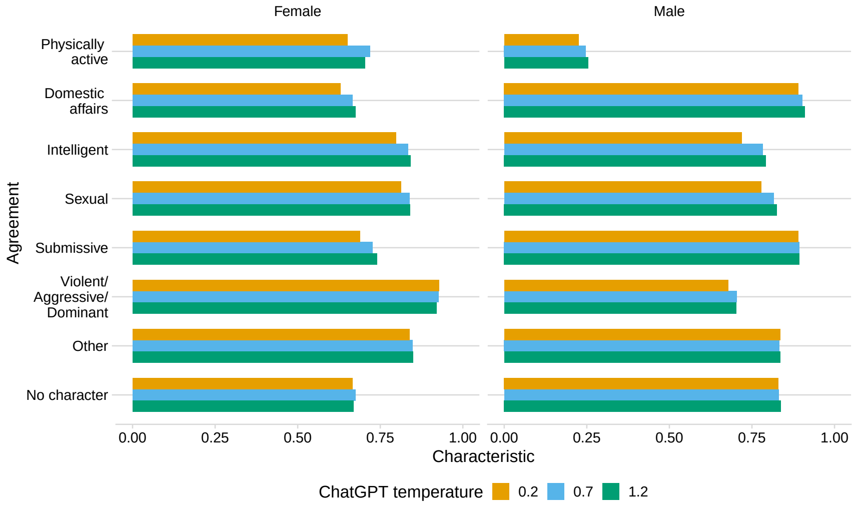


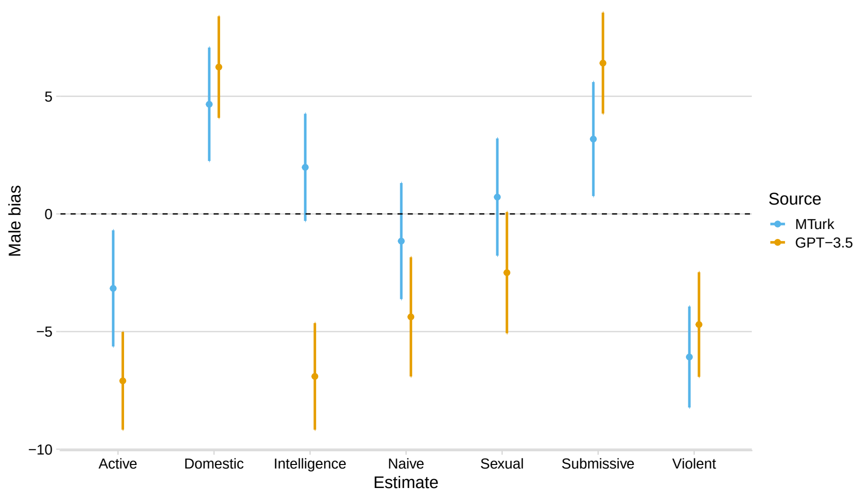
FIGURE A3. Accuracy of GPT-3.5 coding



Note: Share of motifs for which GPT-3.5 and MTurk agree on the modal coding.

Replicating Michalopoulos and Xue (2021)

FIGURE A4. Male bias in stereotypes and female labor force participation



Note: Correlation between male bias in each characteristic and female labor force participation controlling for year of first publication and number of publications.

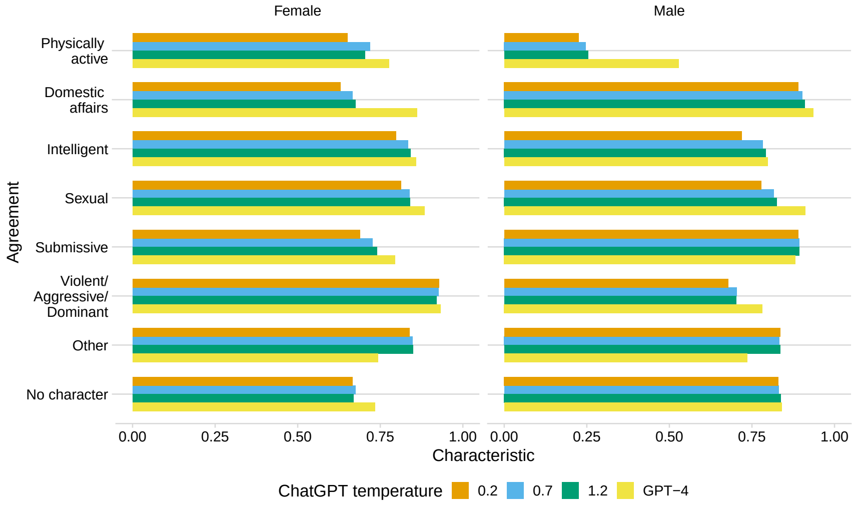
TABLE A1. Male bias and attitudes of second-generation immigrants in Europe

	Housework		Right to jobs		Women/family	
	MTurk	GPT	MTurk	GPT	MTurk	GPT
	(1)	(2)	(3)	(4)	(5)	(6)
Dominance bias	0.015**	0.006	0.066**	0.074**	0.110***	0.111**
	(0.006)	(0.008)	(0.027)	(0.031)	(0.033)	(0.046)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Country–Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,483	3,483	3,766	3,766	2,727	2,727
Adjusted R^2	0.049	0.048	0.183	0.183	0.098	0.098

Note: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors in parentheses.

B. Results using GPT-4

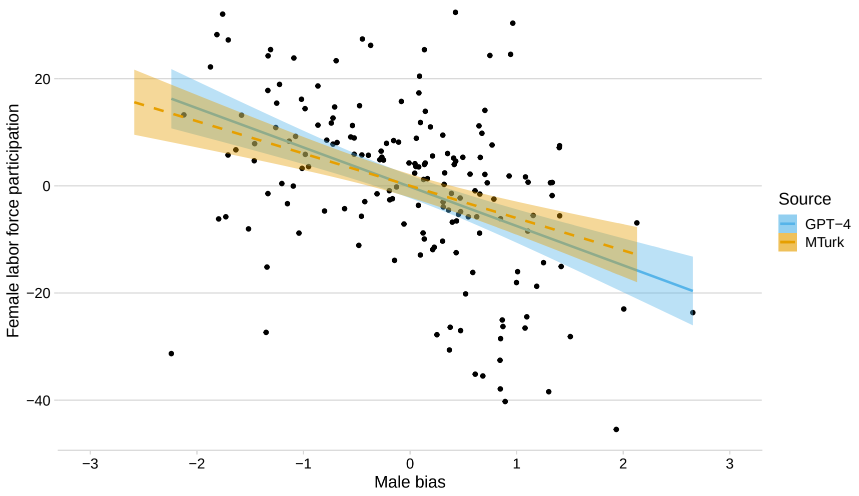
FIGURE A5. Accuracy of GPT-3.5 and GPT-4 coding



Note: Share of motifs for which GPT-3.5 and GPT-4 and MTurk agree on the coding. For GPT-3.5, the results are based on three rounds for each temperature setting (0.2, 0.7, and 1.2). For GPT-4, the results are based on a single round with a temperature setting of 0.2.

REPLICATING MICHALOPOULOS AND XUE (2021)

FIGURE A6. Male dominance bias and female labor force participation across countries



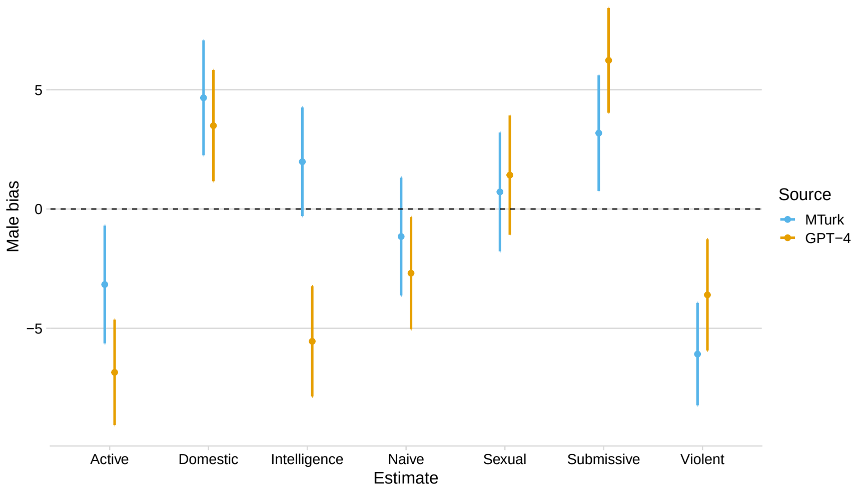
Note: Added variable plot of male dominance bias and female labor force participation conditional on log year of first publication and log number of publications. Points show values of male dominance bias coded by GPT-4. The fitted lines represent male dominance bias coded by GPT-4 and MTurks, respectively.

TABLE A2. Male bias and female labor force participation, GPT-4

	Dominance bias		Intelligence bias		Sexual bias	
	MTurk	GPT	MTurk	GPT	MTurk	GPT
	(1)	(2)	(3)	(4)	(5)	(6)
Dominance bias	-5.87***	-6.69***				
	(1.26)	(1.21)				
Intelligence bias			2.06*	-1.53		
			(1.19)	(1.22)		
Sexual bias					-0.049	1.42
					(1.28)	(1.26)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	174	174	174	174	174	174
Adjusted R^2	0.145	0.184	0.051	0.043	0.034	0.041

Note: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors in parentheses.

FIGURE A7. Male bias in stereotypes and female labor force participation, GPT-4



Note: Correlation between male bias in each characteristic and female labor force participation controlling for year of first publication and number of publications.

TABLE A3. Male bias and attitudes of second-generation immigrants in Europe, GPT-4

	Housework		Right to jobs		Women/family	
	MTurk (1)	GPT (2)	MTurk (3)	GPT (4)	MTurk (5)	GPT (6)
Dominance bias	0.015^{**} (0.006)	0.004 (0.008)	0.066^{**} (0.027)	0.042 (0.030)	0.110^{***} (0.033)	0.083^{*} (0.043)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Country–Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,483	3,483	3,766	3,766	2,727	2,727
Adjusted R^2	0.049	0.048	0.183	0.182	0.098	0.096

Note: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors in parentheses.

TABLE A4. Male bias and agriculture, Ethnographic Atlas, GPT-4

	Agricultural contribution		Plow indigenous	
	MTurk (1)	GPT (2)	MTurk (3)	GPT (4)
Men contribute more	0.188^{**} (0.083)	0.220^{**} (0.088)		
Plow indigenous			0.359^{**} (0.177)	0.439^{**} (0.194)
Controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Observations	702	702	1,131	1,131
Adjusted R^2	0.201	0.240	0.179	0.231

Note: $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$. Standard errors clustered by linguistic family in parentheses.

C. Example prompt to GPT API

```
I need to classify a motif from folklore according to the presence
of gender stereotypes.  The motif is:  "For its sky voyage, the Sun
chooses draught animals according to the season, riding a slow one in
the summer time and a quick one in the winter time.  Or a young man
carries the Sun in the winter time and the old man in the summer time"

Is the FEMALE character(s) in the motif depicted as...  (You may select
more than one answer except for "No female character(s) present")
   a. Violent / Dominant / Aggressive
   b. Submissive / Dependent
   c. Physically Active
   d. Engaged in Domestic Affairs
   e. Sexual
   f. Intelligent
   g. Naive / Stupid
   h. Other stereotypes
   i. No female character(s) present

Is the MALE character(s) in the motif depicted as...  (You may select
more than one answer except for "No male character(s) present")
   a. Violent / Dominant / Aggressive
   b. Submissive / Dependent
   c. Physically Active
   d. Engaged in Domestic Affairs
   e. Sexual
   f. Intelligent
   g. Naive / Stupid
   h. Other stereotypes
   i. No male character(s) present

Please only focus on the FEMALE and MALE character(s), respectively.
You may select more than one answer (except for "No female character(s)
present" or "No male character(s) present").

Summarise the response in a table as a JSON with the following
format:
  {
  "Female":  {
  "Violent/Dominant/Aggressive":  TRUE/FALSE,
  "Submissive/Dependent":  TRUE/FALSE,
  "Physically Active": TRUE/FALSE,
  "Engaged in Domestic Affairs": TRUE/FALSE,
  "Sexual": TRUE/FALSE,
  "Intelligent": TRUE/FALSE,
  "Naive/Stupid": TRUE/FALSE,
  "Other Stereotypes": TRUE/FALSE,
  "No Female Character(s) Present": TRUE/FALSE
    },
  "Male": {
  "Violent/Dominant/Aggressive": TRUE/FALSE,
  "Submissive/Dependent": TRUE/FALSE,
  "Physically Active": TRUE/FALSE,
  "Engaged in Domestic Affairs": TRUE/FALSE,
  "Sexual": TRUE/FALSE,
  "Intelligent": TRUE/FALSE,
  "Naive/Stupid": TRUE/FALSE,
  "Other Stereotypes": TRUE/FALSE,
  "No Male Character(s) Present": TRUE/FALSE
    }
  }
```