Gender Representation and Bias in Indian Civil Service Mock Interviews

Anonymous ACL submission

Abstract

Online educational resources often serve as a great leveler for broadening participation. However, unlike traditional educational resources, little or no computational audits for bias exist for such resources. This paper investigates online educational resources for Indian civil service exams, one of the most fiercely competed exams in the world. Our paper makes three key contributions. First, via a substantial corpus of 51,366 interview questions sourced from 888 YouTube videos of mock interviews of Indian civil service candidates, we demonstrate stark gender bias in the broad nature of questions asked to male and female candidates. Second, our experiments with large language models show a strong presence of gender bias in explanations provided by the LLMs on the gender inference task. Finally, we present a novel dataset of 51,366 interview questions that can inform future social science studies.

1 Introduction

- *What is the age of your kids?*
- *Provide tips to keep your kids busy.*
- *Who is there to handle the kids in your absence?*
- *How is the poverty line defined now?*
- *What is the role of Sanchi Stupa in the national emblem of India?*
- *What was the basic philosophy of Kautilya in Political Science?*

The surprising thread connecting these two contrasting sets of questions is that they appeared in the preparatory UPSC mock interviews organized by the same coaching institute. However, with one key difference — the first set was asked of a female candidate, and the second one was asked of a male candidate.

Tracing back its origin to the Imperial Civil Service (ICS) (Cornell and Svensson, 2020), Indian Administrative Service (IAS) has a long and decorated history that shaped the India we see today. The IAS holds significant influence in Indian governance, forming the administrative backbone of the world's largest democracy. Due to its strong influence on public policy, the Civil Services Examination, organized by the Union Public Services Commission (UPSC), is one of the most competitive exams in India, with around a million aspirants applying every year. The exam consists of multiple written tests with the final phase involving an interview/personality test.

A growing market of coaching institutes has emerged providing coaching to these millions of aspirants. Many of these institutes have a strong online presence and have published mock interview videos of the personality test of the top candidates on their YouTube channels for broader accessibility of their training materials. Online educational resources often serve as a great leveler for broadening participation (Chtouki et al., 2012; Hansen and Reich, 2015). However, unlike traditional educational resources (Lucy et al., 2020; Parashar and Singh, 2020), little or no computational audits for bias exists for such resources.

While gender bias has a rich and extensive literature in diverse social and computational settings that include hiring decisions (Marlowe et al., 1996), machine translation (Ghosh and Caliskan, 2023), movie transcripts and narrative tropes (Gala et al., 2020), interview processes (Kane and Macaulay, 1993), word embeddings (Garg et al., 2018), academic textbooks (Blumberg, 2008), and political interruptions (Yoo et al., 2022), UPSC mock interviews present a rare lens to the interview process of one of the most coveted job positions in India and to the best of our knowledge, no comprehensive AI-powered analyses have scrutinized gender bias in these interviews. *Is it possible that beneath the veneer of seemingly innocuous assortment of in-*

terview questions on public policy, international relations, cutting edge technologies, and social studies, lies a biased pattern where women are consistently asked different questions than their male counterparts? Via a substantial corpus of 888 mock civil service and administrative service mock interview videos published by 14 well-known coaching institutes, this paper seeks to conduct a thorough investigation of the following research questions.

- ***RQ1: How does gender representation manifest in UPSC mock interviews in terms of candidate and panel composition?***
- ***RQ2: What topical biases are present in the questions asked of male and female candidates during mock interviews?***
- ***RQ3: Are there discernible differences in the style or tone of questioning that indicate gender bias, irrespective of the topics covered?***
- ***RQ4: Do LLMs exhibit gender biases in their explanations when tasked with inferring the gender of candidates from interview transcripts?***

Our mixed-method analyses reveal that (1) women are almost thrice as likely as men to be asked questions about gender equality or family; (2) while the candidates in mock interviews show reasonable gender distribution (65.32% male and 34.68% female), the interview panels exhibit significantly more skewed gender distribution; and, (3) large language models exhibit societal biases in their explanations when tasked with the determination of gender from interview transcripts.

Our contributions are the following:

- ***Resource***: We compile a substantial corpus of 51,366 interview questions sourced from 888 UPSC mock interviews conducted by 14 prominent coaching institutes. These questions, with the video transcripts, will enable social science researchers to investigate other important questions relating to the topics featured in these interviews.
- ***Social***: To our knowledge, this is the first paper that examines gender bias in UPSC mock interviews. Our analyses reveal that women face substantially more questions around gender equality, family, and women empowerment and considerably fewer questions on international affairs, world politics, and sports suggesting a strong presence of gender stereotypes.
- ***Methodological***: In an experiment to infer gender from interview transcripts, we observe that several cutting-edge LLMs exhibit stereotypes in their explanations that point to deeply entrenched gender bias in emerging technologies.

2 Dataset

2.1 Step 1: Identifying Relevant Videos

We first construct a set of relevant video of mock interviews conducted with candidates preparing for Civil Services examinations. We consult 14 YouTube channels managed by prominent training institutes (see **SI** for further details). These channels have a strong viewer engagement with 2,378,857 $\pm$ 4,259,079 subscribers and median video views of 42 million. We use publicly available YouTube API and collect all videos from these channels. We next filter in all videos whose titles contain the phrase `mock interview`. To avoid shorts and promotional videos, we discard any video that lasts less than 10 minutes. This yields a set of 888 videos, denoted by $\mathcal{V}$. Note that, $\mathcal{V}$ not only includes mock interviews of candidates preparing for the Civil Services positions of the Union/Central government but also state governments such as Uttar Pradesh, Bihar, and Rajasthan.

When we contrast the academic background distribution of a random sample of 200 candidates from $\mathcal{V}$ (obtained through manual inspection of videos) with ground truth sourced from official UPSC statistics. we observe that $\mathcal{V}$ is a representative sample of successful UPSC candidates and is consistent with the academic distribution background of the recommended candidates (**SI** contains the Table).

2.2 Step 2: Obtaining Interview Transcripts

663 videos (74.66%) in $\mathcal{V}$ have creative commons license. For these videos, we generate transcripts from the audio information using Whisper OpenAI (Radford et al., 2023). For the remaining videos, we first obtain the transcript using publicly available YouTube API. YouTube official transcripts do not have punctuation such as question mark. We use `GPT-3.5` to add appropriate punctuation to the transcript. The transcribed corpus, $\mathcal{D}$, consists of 4.5 million tokens. $\mathcal{D}$ consists predominantly of conversations in English. However, a few interviews had conversations in both English and Hindi. We note that when the conversations switched to Hindi, the ASR system often repeats its previous generations. To account for this, we remove sentences that repeat three or more times in a row. A manual inspection on a small subset of videos confirms that the transcripts have high

fidelity with actual audio even including accurate transcription of Indian names if mentioned in the audio. Our use of these publicly available interviews of public officials hosted on public social web platforms for research purpose comes under the purview of fair use.

2.3 Step 3: Candidates' Gender Inference

Any contrastive study involving gender requires partitioning instances based on gender information. However, annotating image or videos for race and gender information is often treated as insignificant, irrefutable, and apolitical process (Scheuerman et al., 2020). We adopt a sociotechnical approach for gender inference of the interview candidates considering multiple sources. We obtain consensus labels from two human annotators who had access to the (1) video titles (titles list candidate names); (2) video transcripts; (3) video thumbnails; and (4) videos. Indian personal names often indicate gender (Sharma, 2005; Gulati, 2015). The annotation process was informed by subcultural naming conventions in last names as well (for instance, `Kaur`, meaning princess, is a Punjabi last name only for females (Kaur-Aulja et al., 2019)). The annotators (see **SI** for details) considered the video frames, videos, and audio transcript and share that formal male (suit) and female attire (97.06% of the female candidates wore sari or kurti); domain-specific knowledge (e.g., if a candidate received gender-isolated education); and of course the pronouns with which the candidate is being referred to – contributed to this annotation process. Overall, we identify 580 videos ($\mathcal{V}_{male}$) of male candidates and 308 videos ($\mathcal{V}_{female}$) of female candidates. These set of labels is denoted by $\mathcal{L}_{comprehensive}$.

Barring recent candidates who are still receiving administrative training, most of these interview candidates have already joined as highly visible public officials. We conduct online search on the candidate names and identify news articles, interview videos (as a celebrated exam topper) and tally our initial annotation with gendered pronouns used in these articles. This process also uncovered further corroborating evidence (e.g., one candidate was a beauty pageant winner and a female model). Finally, the resumes of the candidates already in the IAS are publicly available. The gender information is listed in these resumes. We consider this information to be the closest to self-determined gender which we consider the ultimate ground truth that we do not possess. Our initial gender inference from videos tally 100% with gender inference conducted through this process.

We also conduct gender inference using large language models and observe interesting stereotypes and biases in their explanations which we discuss in the results section.

2.3.1 Gender Inference from Names Only

Separate from the two annotators who constructed $\mathcal{L}_{comprehensive}$, we task another annotator who is an expert social scientist with inferring gender solely from the candidate names. The Cohen's κ with $L_{comprehensive}$ is 0.81. The human annotator struggled with gender-neutral names. On a similar task, we observe `GPT-3.5` [1] and `Claude-3.5-Sonnet` [2] achieve superior Cohen's κ of 0.91 and 0.89, respectively, establishing that (1) multiple sources (e.g., image, news articles, resumes) contribute to more robust gender inference; and (2) these LLMs have cultural grounding of Indian names.

2.4 Step 4: Sets of Interview Questions

From $\mathcal{D}$, we construct $\mathcal{Q}$ consisting of sentences that end with a question mark as the set of questions asked of the candidates. To preserve the context of the questions, we also included the sentence that appeared before each question. We acknowledge that this is a high-recall approach with certain caveats. For instance, this set will include clarification questions asked by the candidates and exclude imperative sentences (e.g., *please give a brief introduction*). A manual inspection of randomly sampled 100 questions reveals that 3 are clarifying questions asked by the candidates. $\mathcal{Q}_{male}$ and $\mathcal{Q}_{female}$ denote all the questions asked on male and female candidates, respectively.

3 Related Work

A substantial body of social science literature highlights the deep cultural and historical roots of gender inequality in South Asian societies, including India. Sen (2001) examines its pervasive presence across various domains, while Batra *et al.* (2016) focuses on entrenched sociocultural norms that sustain disparities in education, employment, and more, advocating for targeted policies. Radhakrishnan *et al.* (2009) explore the tension between traditional constructs

[1] https://openai.com
[2] https://www.anthropic.com/

of femininity and modern globalized identities among professional women in urban India. Beyond the social science literature on gender gap in India, gender bias has an extensive literature in diverse social and computational settings that include hiring decisions (Marlowe et al., 1996), machine translation (Ghosh and Caliskan, 2023), movie transcripts (Khadilkar et al., 2022), interview processes (Kane and Macaulay, 1993), word embeddings (Garg et al., 2018), academic textbooks (Blumberg, 2008), and political interruptions (Yoo et al., 2022). However, barring a few instances (Madaan et al., 2018; Khadilkar et al., 2022; Dutta et al., 2023), AI-powered, computational analyses of gender and societal biases in the Indian context are rather underexplored. Our work contrasts with existing lines of work (1) in terms of domain (Civil Service interview versus gender inequality in Bollywood (Madaan et al., 2018; Khadilkar et al., 2022) and divorce court proceedings (Dutta et al., 2023)); and (2) nuanced analyses of bias in LLM explanations.

Davison and Burke (2000) conducted research spanning from the 1970s to the present and showed persistent gender discrimination in workplace. Despite a growing belief in competence equality over time, as noted in a cross-temporal meta-analysis by Eagly et al. (2020), recent research by Lippens et al. (2023) reveals the complexity of gender discrimination in hiring, with both men and women experiencing discrimination in certain contexts. Castaño et al. (2019) found that women who take on roles traditionally seen as masculine are viewed as cold and driven, while those who align with feminine roles are seen as less capable. Men don't usually face this type of bias. As a result, even when women perform as well as their male counterparts, they are often rewarded less in prestigious jobs (Joshi et al., 2015).

A range of research studies has explored gender bias in explainability, highlighting that bias in AI systems can manifest in the explanations provided by these models. For instance, Huber *et al.* (2023) explore potential gender bias in explainability tools used in face recognition systems. These tools, designed to provide insights into ML models, might exhibit gender-based bias, leading to signs of biased decisions in critical applications like face recognition. Shrestha and Das (2022) conduct a systematic review to identify gender biases in ML and AI academic research.

4 Results and Discussion

We start with an important point for the readers as they learn about our findings regarding gender representation and bias: ***the female candidate pool in the mock interviews is as strong as (if not slightly better than) their male counterparts.*** As already mentioned, of the multiple phases in the UPSC exam, the final phase is the personality test. Figure 4 (see **SI**) summarizes the candidates' overall performance taking into account the written as well as the personality test. We further note that no significant differences exist in the average number of questions and interview duration between male and female candidates (**SI** contains details).

4.1 Representation

RQ1: How does gender representation manifest in UPSC mock interviews in terms of candidate and panel composition?

Observation 1: Gender representation in YouTube mock interviews is not far from real-world representation. As already noted, $\mathcal{V}_{male}$ represents 65.32% of our candidate pool while $\mathcal{V}_{female}$ represents the remaining 34.68%. Hence, the gender representation of $\mathcal{V}$ is visibly skewed. However, the imbalance is not far from real-world gender imbalance in UPSC recommendations. Real-world data indicates that the percentage of women candidates recommended by the UPSC has increased from 24% in 2018 to 34% in 2022 (Desk, 2023).

Observation 2: The interview panels exhibit stark gender imbalance. We observe that the candidates refer to male panelists and the female panelists with the formal honorific `sir` and `ma'am` (short form of `madam`), respectively. Let $\mathcal{N}_s$ and $\mathcal{N}_m$ denote the count of the usage of `sir` and `ma'am`, respectively. We compute the male honorific ratio (MHR) $\frac{\mathcal{N}_s}{\mathcal{N}_s+\mathcal{N}_m}$. A value closer to 1 indicates a predominantly male panel whereas a value closer to 0.5 indicates a gender-balanced panel. We observe a value of 0.81 for `MHR` indicating a predominantly male panel composition. A manual inspection of randomly sampled 200 videos aligns with this observation.

4.2 Bias in Discourse and Questions

RQ2: What topical biases are present in the questions asked of male and female candidates during mock interviews? Our findings from a series of experiments indicate considerable gender bias.

4.2.1 Unigram Differential Analysis

A unigram differential analysis illustrates the difference between the discourse in $\mathcal{D}_{male}$ and $\mathcal{D}_{female}$. For $\mathcal{D}_{male}$ and $\mathcal{D}_{female}$, we compute the respective unigram distributions $\mathcal{P}_{male}$ and $\mathcal{P}_{female}$. Next, for each token t, we compute the scores $\mathcal{P}_{male}(t) - \mathcal{P}_{female}(t)$, and $\mathcal{P}_{female}(t) - \mathcal{P}_{male}(t)$ and obtain the top tokens ranked by these scores (indicating increased usage in the respective subcorpus). Table 1 indicates that male interviews are likelier to discuss technology, global politics, and sports than female interviews. In contrast, female interviews are likelier to discuss gender, family, and children as indicated by the presence of words *girl*, *woman*, *gender*, and *child*. We do not observe a single gendered word in the left column while we observe two gendered words in the right (e.g., *woman* and *girl*).

More presence in $\mathcal{D}_{male}$	More presence in $\mathcal{D}_{female}$
bengal, region, close, west, job, relative, happening, department, interest, industrial, accept, engineering, ukraine, agent, cricket, relation, option, subject, forest, iit	*woman, question, delhi, believe, capital, owner, important, something, good, deep, girl, education, place, gender, first, child, feel, science, health, doctor*

Table 1: Disparity in word presence.

4.2.2 Log Odds Ratio Analysis

We perform a log odds ratio analysis to find the words that are more likely to appear in female (male) interviews compared to their male (female) counterparts. Log odds of a word w is defined as

$$log\ odds(w) = \frac{\text{normalized frequency of } w \text{ in female interviews}}{\text{normalized frequency of } w \text{ in male interviews}}$$

A high positive value indicates that the word is more likely to appear in the interviews featuring a female candidate and a high negative value would indicate the opposite. We find the following words with high positive values of log odds – rag (3.81), miranda (house college) (3.16), nervous (2.28), glass (ceiling) (2.11), sari (1.91), beauty (1.83). On the other hand, these words show high negative scores – photography (-18.71), brexit (-18.53), football (-2.80), ncc (National Cadet Corps) (-2.80), camera (-2.66), alcohol (-2.50).

The words with high association with women often indicate their hobbies and academic background. For instance, *rag* is a musical structure in Indian classical music. Since music is often mentioned as a hobby by the female candidates, we find this word's overpresence in female interviews. Miranda House College is a well-known gender-isolated college for women. We also observe words indicating female-traditional attire (sari). We were intrigued by the overpresence of the word *glass*; manual inspection reveals that this word was used in the context of the phrase glass ceiling which further corroborates our earlier finding that gender inequality is more predominantly discussed with female candidates than with male candidates.

With male candidates, we again observe that hobbies often dictated frequently used words. For instance, *photography*, *football*, and *camera* were discussed in the context of hobbies. However, we do notice that words indicating world events (e.g., *Brexit* were present more frequently in male candidate interviews further substantiating our earlier finding that male candidates were more likely to be asked of questions about world politics.

Our findings point to a worrisome vicious cycle. On one hand, we do notice, that male candidates are asked about global politics more. But when a large language model uses an explanation during gender inference that a discussion heavy with world politics made it infer that the candidate is possibly male, may point to a problematic cycle where models learn from existing social biases and in turn, produce biased responses.

4.2.3 Word Embedding Association Tests

Word Embedding Association Tests (`WEAT`) (Caliskan et al., 2017) is a widely used framework (Lewis and Lupyan, 2020; Khadilkar et al., 2022) to quantify gender bias. Here, we are interested in answering the question – *does the candidate's gender matter in terms of discussion related to career and family?* We use `WEAT` score to quantify this association. Following prior literature (Nosek et al., 2002), we construct two target sets: *Career* {*executive, management, professional, corporation, salary, office, business, career*} and *Family* {*home, parents, children, family, cousins, marriage, wedding, relatives*}. We choose the following sets as the attributes representing gender qualifiers: *Male* {`male, man, he`} and *Female* {`female, woman, she`}. We train `FastText` (Bojanowski et al., 2017) on $\mathcal{D}$ to get the vectors corresponding to these words. Using these word vectors, we compute the `WEAT` score. Over five independent runs, we observe the `WEAT` score to be 0.29 ± 0.07. This positive score indicates a statistically significant association between males with career-oriented terms and females with family-oriented terms.

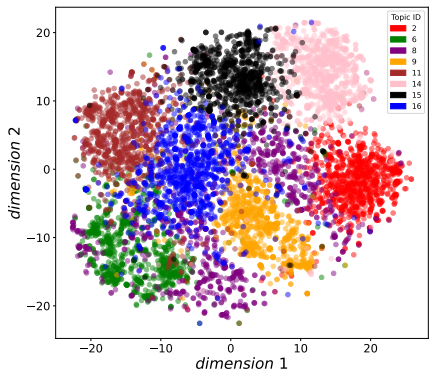


Figure 1: t-SNE (Van der Maaten and Hinton, 2008) visualization (using scikit-learn(Pedregosa et al., 2011)) of top eight question topics. For better visualization, 1000 questions were randomly sampled from each cluster. Topic explanations – 2: *history and mythology*, 6: *agriculture and environment*, 8: *science*, 9: *foreign policy*, 11: *economics*, 14: *gender related*, 15: *law and order*, 16: *engineering and technology*. Relevant keywords are listed in Table 2.

4.2.4 Semantic Clustering of Questions

To further study the differences in questioning patterns between male and female candidates, we cluster the questions into semantically similar topics. We compute the semantic embedding of the question texts in $\mathcal{Q}$ using a transformer-based embedding model, `all-MiniLM-L6-v2` (Reimers and Gurevych, 2020). We then run K-means (Wu, 2012), an unsupervised clustering algorithm on these embeddings. The assumption here is that the clusters will have semantically similar questions. Initially, the number of clusters (topics) was set to 20. Among these, we are interested in the topics that exhibit a disparity in gender representation. To quantify this disparity, we use the imbalance ratio ($\mathcal{R}_{imbalance}$) metric. For a topic t, the imbalance ratio is defined as

$$\mathcal{R}_{imbalance} = \frac{max\{f^t_{male}, f^t_{female}\}}{min\{f^t_{male}, f^t_{female}\}}$$

where f^t_{male} and f^t_{female}denote the fraction of questions asked to male and female candidates, respectively, that belong to the topic t. In an ideal world where men and women candidates face similar questioning, the value of $\mathcal{R}_{imbalance}$ should be ~ 1 for any topic t. Conversely, a high $\mathcal{R}_{imbalance}$ value indicates a significant skew in the distribution of questions toward one gender. Table 2 presents the top eight topics displaying the greatest imbalance ratios. Figure 1 visualizes these topics. To better interpret the topics, we find the most prevalent phrases in each cluster and manually read a random sample of questions. Our analysis reveals that the questions related to *gender equality* show the most skewed distribution, with female candidates nearly three times more likely to face such questions than their male counterparts. Among the other topics, questions related to *agriculture and environment, engineering and technology, foreign policy, science,* and *economics* were predominantly directed at male candidates, whereas *history and mythology* and *law and order* questions were more frequently posed to females.

4.3 Separability Tests

RQ3: Are there discernible differences in the style or tone of questioning that indicate gender bias, irrespective of the topics covered?

The *imbalance ratio* captures the differences in questioning patterns by analyzing the distribution of topics across male and female candidates. However, an important question still remains – *Does a systematic difference exist in the nature of questioning between the two groups regardless of the topic?* Following Dutta *et al.* (2022), we conduct a set of separability tests treating classification accuracy as a proxy for text separability. We construct the classification datasets by assigning labels to the questions from $\mathcal{Q}_{female}$ (label **F**) and $\mathcal{Q}_{male}$ (label **M**). Given a question, the classifier needs to predict whether it was asked to a female or male candidate. Intuitively, if there exist linguistic cues to differentiate between the questions, the task is learnable. However, if no such signals exist, the classifier will not perform better than chance. We describe the separability tests below.

Let $\mathcal{Q}^t_{female}$ and $\mathcal{Q}^t_{male}$ denote the questions asked to female and male candidates belonging to topic t. We sample an equal number of questions from $\mathcal{Q}^t_{female}$ and $\mathcal{Q}^t_{male}$, combine all topics, and split the data into train and test set in an 80:20 ratio. We fine-tune `BERT` (Devlin, 2018), a well-known pre-trained language model, for this classification task. As a control, we randomly divide $\mathcal{Q}^t_{female}$ into two equal parts, combine all topics, and conduct the classification experiment. We repeat this process for $\mathcal{Q}_{male}$. Table 3 presents results across test sets. We note that classification accuracy for $\mathcal{Q}_{female}$ vs $\mathcal{Q}_{male}$ is significantly higher than chance. Whereas the in-group classifiers perform no better than random guesses as expected. These results indicate that there exists a difference in the nature of questing between the

Topic ID	f^t_{female} (%)	f^t_{male} (%)	$\mathcal{R}_{imbalance}$	Topic Interpretation	Key Phrases
14	4.19	1.43	2.94	gender equality	*woman, gender, female, empowerment, society, woman empowerment, sex ratio, gender equality, reservation woman, uttar pradesh, woman reservation, sexual harassment*
6	3.68	4.49	1.33	agriculture and environment	*agriculture, farmer, forest, climate change, environment, pollution, sustainable development, global warming, renewable energy, development goal, power plant, green hydrogen, environmental issue, organic farming, rural area, food security, drinking water, green revolution, disaster management*
16	4.78	5.96	1.25	engineering and technology	*big data, artificial intelligence, mechanical engineering, computer science, internet of things, machine learning, digital india, technology*
2	7.90	6.43	1.23	history and mythology	The key phrases in this cluster are not clear; however, a manual inspection reveals that questions are mostly related to history, mythology, and religious scriptures.
9	4.10	4.92	1.20	foreign policy	*international relation, foreign policy, prime minister, saudi arabia, united nation, european union, sri lanka, cold war, russia, ukraine, china, pakistan, afghanistan, taliban, world war, foreign trade, middle east, security*
8	5.84	6.87	1.18	science	*difference, virus, chemical, example, reason, basic difference*
15	6.02	5.19	1.16	law and order	*law, supreme court, high court, fundamental right, constitutional amendment, district magistrate, information act, article, justice, police, rule, government*
11	5.90	6.53	1.11	economics	*economy, gdp, bank, budget, income tax, fiscal deficit, finance commission, growth rate, monetary policy, interest rate, stock market, demographic trend, fiscal policy, black money*

Table 2: Descriptive analysis of question topics. Color coding: Red highlights topics with greater female representation, while Blue signifies topics with greater male representation.

male and female candidates.

	$\mathcal{Q}_{female}$	$\mathcal{Q}_{male}$
$\mathcal{Q}_{female}$	51.5 ± 1.3%	57.9 ± 0.6%
$\mathcal{Q}_{male}$	57.9 ± 0.6%	52.1 ± 0.4%

Table 3: Separability test results.

SI contains additional experiments that show (1) linguistic separability of male-versus-female questions happens even if we control for topics (i.e., ensure each split has an equal number of questions from a given topic); and (2) linguistic separability of male-versus-female questions exists even within the same topic.

4.4 Bias in LLM Explanations

LLM	Cohen's κ
`Mistral-7B-Instruct`	0.581
`GPT-3.5-Turbo`	0.751
`Claude-3.5-Sonnet`	0.853

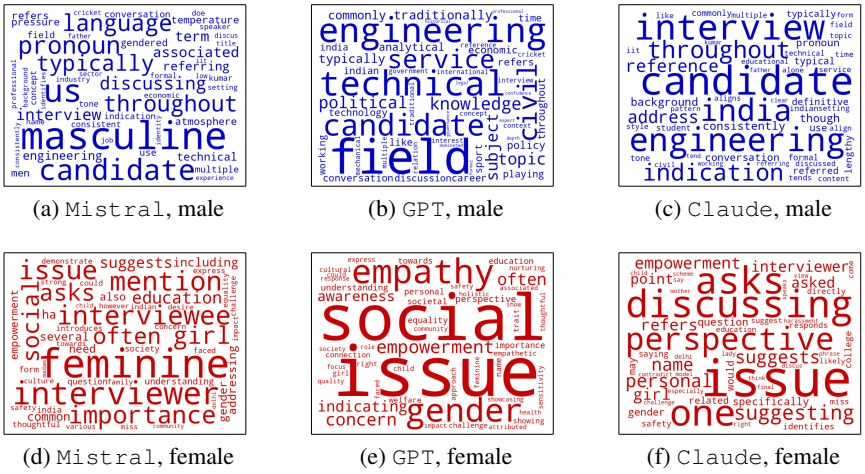
Table 4: Gender inference evaluation.

RQ4: To what extent do LLMs exhibit gender biases when inferring the gender of candidates from UPSC mock interview transcripts?

Here, we examine the explanations provided by the three LLMs – `Mistral-7B-Instruct` (Jiang et al., 2023) (open source), `GPT-3.5-Turbo` (proprietary), and `Claude-3.5-Sonnet` (proprietary) – to assess whether the entrenched societal biases in language models significantly influence their gender inference.

To infer gender using LLMs, we set up a zero-shot classification task with a prompt (See Figure 3 in **SI**) containing a detailed instruction followed by the interview transcript. We then extract the predicted gender and reasons from the response. It is worth noting that here, we do not include the candidate's name specifically; however, it may appear in the transcript if it is mentioned during the interview. Table 4 compares the performance of different LLMs in these inference tasks (Cohen's κ is computed with respect to $\mathcal{L}_{comprehensive}$ inferred by humans) and shows that `Claude` demonstrates the highest performance in this task followed by `GPT` and `Mistral`.

We extract all rationales given by these models for each candidate, subsequently organizing these into two distinct datasets: $\mathcal{D}^M_{reasons}$ for male predictions and $\mathcal{D}^F_{reasons}$ for female predictions. We then analyze the unigram distribution within each dataset ($\mathcal{NF}^M_{reasons}$ and $\mathcal{NF}^F_{reasons}$) and perform a differential analysis similar to that described earlier. This analysis identifies words that are disproportionately frequent in one dataset compared to the other. Specifically, terms with high differential frequencies, $DF_{\mathcal{NF}^M_{reasons} - \mathcal{NF}^F_{reasons}}$, are indicative of a male prediction bias in LLM responses. Conversely, words with high scores in $DF_{\mathcal{NF}^F_{reasons} - \mathcal{NF}^M_{reasons}}$ suggest a female prediction



(a) Mistral, male (b) GPT, male (c) Claude, male

(d) Mistral, female (e) GPT, female (f) Claude, female

Figure 2: Wordclouds highlighting the top words found from the differential analysis of unigram distribution of LLM explanations. The images illustrate words like *engineering, technical, civil, knowledge* while the bottom images feature words like *empathy, gender, social, issue, awareness* indicating the ingrained bias in the reasoning process of LLMs.

bias.

Figure 2 display the most significant words emerging from this differential analysis. The analysis reveals that all three models (Figures 2a, 2b, 2c) are more likely to predict a candidate's gender as male when encountering terms such as *engineering, technical, civil* — words traditionally associated with male-dominated fields. In contrast, as shown in Figures 2d, 2e, 2f, terms like *gender, women empowerment, social issue* are predictive of a female gender identification by these models, underscoring potential biases in their training data that perhaps correlate these concepts with female gender. Interestingly, we note that if GPT-3.5 finds qualities like *empathy, awareness, and understanding* in the candidate's response, it predicts female. On the other hand, Mistral often determines a candidate's gender if it deems the language as masculine or feminine. Table 5 lists a few examples showing the striking difference between the explanations provided by these LLMs for male and female candidates.

Predicted Gender	LLM	LLM Explanation
Male	GPT-3.5	*The candidate shows a strong knowledge of engineering concepts, which can be more commonly found in male candidates in technical fields.*
Female	GPT-3.5	*The candidate's responses reflected empathy, compassion, and a focus on issues related to women empowerment, education, and societal challenges, which are often associated with female perspectives.*
Male	Mistral	*The candidate mentions his educational background, including his M.Tech in transportation engineering and his optional subject of anthropology, which are typically male-dominated fields.*
Female	Mistral	*... discusses issues related to the representation of tribal people and the inclusion of women in political and employment spheres, which are often topics of interest for female candidates.*
Male	Claude	*The candidate discusses his B.Tech degree in Mechanical Engineering, a field that tends to have more male students.*
Female	Claude	*The candidate is asked about procedures for sexual harassment of women in the workplace, which is a topic often directed at female candidates.*

Table 5: Examples demonstrating bias in the language models' rationale for gender predictions.

5 Conclusion

This paper conducts a comprehensive analysis of gender inequality through the lens of UPSC mock interview questions. UPSC is one of the most competitive exams in India, and selected candidates form the administrative backbone of the country. Yet, no prior literature (to our knowledge) has investigated gender inequality in mock interviews for one of the most high-profile government jobs in India. Our study is descriptive, not prescriptive. Our analyses reveal that while the interviewed female candidates are as strong as their male counterparts, their interview questions are strikingly different from the interview questions asked of the male candidates. We also observe that the interview panels are predominantly male. Finally, we present an intriguing finding that uncovers deep-seated gender bias in LLMs through the lens of a gender inference task.

Limitations

Transcription of the YouTube videos might not be the most accurate as models may introduce errors. We have used Whisper OpenAI in order to transcribe the videos. We have used proprietary LLMs such `GPT-3.5` and `Claude-Sonet-3.5`. Exact reproducibility of results might not be possible as the LLMs keep updating themselves. Our study investigates UPSC mock interview questions. While these mock interviews often invite experienced former IAS officers and noted academicians as panelists, it is not possible to estimate the fidelity of these mock interviews with the actual interviews.

Finally, any study on binary gender bias runs the risk of oversimplifying gender. We acknowledge that gender lies on a spectrum. We are also sensitive to previous studies that point out the potential harms of the erasure of gender and sexual minorities (Dev et al., 2021). It is possible that our gender inference has some noise. Following a recent global survey that indicates that nearly 3% of the survey population identified as non-binary, nonconforming, gender-fluid, or transgender [3], we induce a 3% error in $\mathcal{L}_{comprehensive}$ and observe that our qualitative claims remain unchanged. Table 6 presents the analysis of different question topics after introducing 3% noise in $\mathcal{L}_{comprehensive}$.

Ethical Statement

We collect public domain data using publicly available API. The interview candidates are highly visible public officials working at high-profile public-facing jobs. Instead of focusing on individual candidates, we conduct aggregate analyses. We thus see no major ethical concern. We rely on large language models for some of our analyses. Prior literature indicates possibilities of biases in these models which may percolate into downstream tasks. In fact, we report LLM biases in the explanations of the gender inference task which presents yet another data point in support of that concern. We verify our results through manual inspection whenever possible. Also, for some of our analyses (e.g., `WEAT` score), we train the word embeddings from scratch. Our dataset also depends on the accuracy of the ASR system. Prior literature indicates such systems are not infallible (Errattahi et al., 2018). Our manual inspection reveals that the quality of the transcriptions was high with occasional errors caused by the conversation switching to Hindi. We conduct additional preprocessing to account for that.

[3] https://www.statista.com/statistics/1269778/gender-identity-worldwide-country/

References

Renu Batra and Thomas G Reio Jr. 2016. Gender inequality issues in india. *Advances in Developing Human Resources*, 18(1):88–101.

Rae Lesser Blumberg. 2008. The invisible obstacle to educational equality: Gender bias in textbooks. *Prospects*, 38:345–361.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ana M Castaño, Yolanda Fontanil, and Antonio L García-Izquierdo. 2019. "why can't i become a manager?"—a systematic review of gender stereotypes and organizational discrimination. *International Journal of Environmental Research and Public Health*, 16(10):1813.

Yousra Chtouki, Hamid Harroud, Mohammed Khalidi, and Samir Bennani. 2012. The impact of youtube videos on the student's learning. In *2012 international conference on information technology based higher education and training (ITHET)*, pages 1–4. IEEE.

Agnes Cornell and Ted Svensson. 2020. Imperial diffusion of bureaucratic practices? entrance examinations to the indian civil service and the british civil service.

Heather K Davison and Michael J Burke. 2000. Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, 56(2):225–248.

Express Web Desk. 2023. Women's share in the upsc pie steadily increasing: 5 points. The Indian Express.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M. Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *EMNLP*, pages 1968–1994.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Topic ID	f^t_{female} (%)	f^t_{male} (%)	$\mathcal{R}_{imbalance}$	Topic Interpretation	Key Phrases
14	4.08	1.44	2.83	gender equality	*woman, gender, female, empowerment, society, woman empowerment, sex ratio, gender equality, reservation woman, uttar pradesh, woman reservation, sexual harassment*
6	3.77	4.88	1.29	agriculture and environment	*agriculture, farmer, forest, climate change, environment, pollution, sustainable development, global warming, renewable energy, development goal, power plant, green hydrogen, environmental issue, organic farming, rural area, food security, drinking water, green revolution, disaster management*
16	4.78	5.98	1.25	engineering and technology	*big data, artificial intelligence, mechanical engineering, computer science, internet of things, machine learning, digital india, technology*
2	7.89	6.41	1.23	history and mythology	The key phrases in this cluster are not clear; however, a manual inspection reveals that questions are mostly related to history, mythology, and religious scriptures.
9	4.06	4.96	1.22	foreign policy	*international relation, foreign policy, prime minister, saudi arabia, united nation, european union, sri lanka, cold war, russia, ukraine, china, pakistan, afghanistan, taliban, world war, foreign trade, middle east, security*
8	5.88	6.87	1.17	science	*difference, virus, chemical, example, reason, basic difference*
15	6.09	5.13	1.19	law and order	*law, supreme court, high court, fundamental right, constitutional amendment, district magistrate, information act, article, justice, police, rule, government*
11	5.84	6.58	1.13	economics	*economy, gdp, bank, budget, income tax, fiscal deficit, finance commission, growth rate, monetary policy, interest rate, stock market, demographic trend, fiscal policy, black money*

Table 6: Descriptive analysis of question topics (with added noise). Color coding: Red highlights topics with greater female representation, while Blue signifies topics with greater male representation.

Sujan Dutta, Beibei Li, Daniel S Nagin, and Ashiqur R KhudaBukhsh. 2022. A murder and protests, the capitol riot, and the chauvin trial: Estimating disparate news media stance. In *IJCAI-22*, pages 5059–5065.

Sujan Dutta, Parth Srivastava, Vaishnavi Solunke, Swaprava Nath, and Ashiqur R. KhudaBukhsh. 2023. Disentangling societal inequality from model biases: Gender inequality in divorce court proceedings. In *IJCAI 2023*, pages 5959–5967.

Alice H Eagly, Christa Nater, David I Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American psychologist*, 75(3):301.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA*, 115(16):E3635–E3644.

Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *AIES*, pages 901–912.

Akshay Gulati. 2015. Extracting information from indian first names. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 138–143.

John D Hansen and Justin Reich. 2015. Democratizing education? examining access and usage patterns in massive open online courses. *Science*, 350(6265):1245–1248.

Marco Huber, Meiling Fang, Fadi Boutros, and Naser Damer. 2023. Are explainability tools gender biased? a case study on face presentation attack detection. In *EUSIPCO*, pages 945–949. IEEE.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aparna Joshi, Jooyeon Son, and Hyuntak Roh. 2015. When can women close the gap? a meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal*, 58(5):1516–1545.

Emily W Kane and Laura J Macaulay. 1993. Interviewer gender and gender attitudes. *Public opinion quarterly*, 57(1):1–28.

Harjnder Kaur-Aulja, Farzana Shain, and Alison Lilley. 2019. A Gap Exposed: What is Known About Sikh Victims of Domestic Violence Abuse (DVA) and Their Mental Health? *European Journal of Mental Health*, 14(1):179–189.

Kunal Khadilkar, Ashiqur R. KhudaBukhsh, and Tom M. Mitchell. 2022. Gender bias, social bias, and representation in bollywood and hollywood. *Patterns*, 3(4):100486.

Molly Lewis and Gary Lupyan. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028.

Louis Lippens, Siel Vermeiren, and Stijn Baert. 2023. The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, 151:104315.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in texas us history textbooks. *AERA Open*, 6(3):2332858420940312.

Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from Bollywood movies. In *FAccT*, pages 92–105. PMLR.

Cynthia M Marlowe, Sandra L Schneider, and Carnot E Nelson. 1996. Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of applied psychology*, 81(1):11.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*, 6(1):101.

Smita Parashar and Smriti Singh. 2020. Evaluating gender representation in ncert textbooks: A content analysis.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR.

Smitha Radhakrishnan. 2009. Professional women, good families: Respectable femininity and the cultural politics of a "new" india. *Qualitative Sociology*, 32:195–212.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*. Association for Computational Linguistics.

Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *CSCW*, 4:1–35.

Amartya Sen. 2001. The many faces of gender inequality. *New republic*, pages 35–39.

Dhruv Dev Sharma. 2005. *Panorama of Indian Anthroponomy:(an Historical, Socio-cultural & Linguistic Analysis of Indian Personal Names*. Mittal Publications.

Sunny Shrestha and Sanchari Das. 2022. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in artificial intelligence*, 5:976838.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Junjie Wu. 2012. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.

Clay H. Yoo, Jiachen Wang, Yuxi Luo, Kunal Khadilkar, and Ashiqur R. KhudaBukhsh. 2022. Conversational inequality through the lens of political interruption. In *IJCAI 2022*, pages 5213–5219.

```
You are provided with a transcript of an interview.
Your task is to read through the interview transcript
and determine the gender of the candidate based on
the content and context of the conversation. After
making your determination, please list the reasons
that led you to your conclusion about the candidate's
gender. Consider the following options for gender
classification:

Male
Female
Unknown

Interview Transcript:
{transcript text}
```

Figure 3: Prompt designed to infer gender using LLMs.

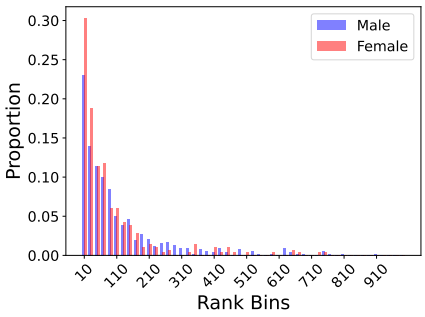


Figure 4: Distribution of records based on rank and gender. Rank information is obtained from the video title.

A Annotation Details

Three graduate students conducted the human annotations. They either received course credit or research stipend (hourly $30). The annotators were informed how their data would be used in our experiments. The annotation did not collect any private information.

	Engineering	Humanities	Science	Medical Science
2020	64.9%	23.2%	7.9%	4%
2019	63.1%	24.2%	6.6%	6.1%
2018	62.7%	24.5%	6.9%	5.9%
2017	66.2%	21.8%	6.4%	5.6%
2016	59.3%	21.9%	10.3%	8.5%

Table 7: Academic Background of Recommended Candidates

Academic Stream	Distribution in $\mathcal{V}$	Real World Distribution
Engineering	63.05%	63.24%
Humanities	25.10%	23.12%
Science	4.45%	7.62%
Medical Science	7.40%	6.02%

Table 8: Distribution of academic streams. Distribution in $\mathcal{V}$ is estimated by manually inspecting a random sample of 200 videos. Real-world distribution is obtained from official UPSC data.

B Word Embedding Association Test (WEAT)

Word Embedding Association Test (WEAT) (Caliskan et al., 2017) score is a metric to detect if there exists a difference between two sets of target words in terms of their association with two sets of attribute words. To compute this metric, first, the words are converted to their vector representations (embeddings). The cosine similarity of two words (a and b) is denoted by $\cos(a, b)$.

$$\text{WEAT}(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \text{mean}_{x \in \mathcal{X}} \sigma(x, \mathcal{A}, \mathcal{B}) - \text{mean}_{y \in \mathcal{Y}} \sigma(y, \mathcal{A}, \mathcal{B})$$

where,

$$\sigma(w, \mathcal{A}, \mathcal{B}) = \text{mean}_{a \in \mathcal{A}} \cos(w, a) - \text{mean}_{b \in \mathcal{B}} \cos(w, b)$$

Intuitively, $\sigma(w, \mathcal{A}, \mathcal{B})$ quantifies the association of w with the attribute sets, and the WEAT score measures the differential association of the two sets of target words with the attribute sets. A positive WEAT score suggests that the target words in set $\mathcal{X}$ have a stronger association with the attributes in set $\mathcal{A}$ than those in set $\mathcal{B}$, and conversely, the words in set $\mathcal{Y}$ show a stronger association with set $\mathcal{B}$ than with set $\mathcal{A}$.

C Question per Interview and Interview Time Duration

We observe that male and female candidates receive comparable number of questions per interview (male candidates: 58.3 $\pm$ 22.1 questions; female candidates: 57.4 $\pm$ 20.8 questions) with male candidates receiving slightly more number of questions per interview. In a similar vein, we observe a male interview is marginally longer than a female interview (male interview: 30 minute 27 second $\pm$ 10 minute 7 second; female interview: 29 minute 25 second $\pm$ 9 minute 1 second).

D Background of Candidates

Table 7 shows the academic backgrounds of UPSC candidates. Table 8 contrasts the academic background distribution of a random sample of 200 candidates from $\mathcal{V}$ (obtained through manual inspection) with ground truth sourced from official UPSC data. Table 8 establishes that $\mathcal{V}$ is a representative sample of successful UPSC candidates and is consistent with the academic distribution background of the recommended candidates.

E List of Channels and Channel Distribution

We use 14 well-known channels: *Drishti IAS - English*; *Chanakya IAS Academy*; *Next IAS*; *Vajirao and Reddy Institute*; *Let's Crack UPSC CSE*; *CivilsDaily IAS*; *BYJU'S IAS*; *Dhyeya IAS*; *StudyIQ IAS*; *PW OnlyIAS*; *Unacademy*; *IAS Baba*; *INSIGHTS IAS*; and *Vajiram and Ravi Official*. Figure 5 illustrates the distribution of channels in the dataset.

F FastText Training Parameters

We use the following training parameters:

- dimension = 100
- epochs = 5
- learning rate = 0.05
- threads = 4.

G Topic Wise Separability Tests

Table 9 lists the separability results for male-versus-female question classification within each topic.

Topic ID	Topic Interpretation	Classifier Accuracy
14	*gender equality*	60.3%
6	*agriculture and environment*	57.9%
16	*engineering and technology*	58.0%
2	*history and mythology*	65.1%
9	*foreign policy*	55.3%
8	*science*	62.5%
15	*law and order*	61.7%
11	*economics*	55.9%

Table 9: Separability test results within topics. This result demonstrates that even when we consider a specific topic, questions asked of male candidates are linguistically different from questions asked of female candidates.

H Cluster Analysis with Noise

We acknowledge that our human gender inference could have errors. Following a recent global survey that indicates that nearly 3% of the survey population identified as non-binary, non-conforming, gender-fluid, or transgender [4], we induce a 3% error in $\mathcal{L}_{comprehensive}$ and observe that our qualitative claims remain unchanged. Table 6 presents the analysis of different question topics after introducing 3% noise in $\mathcal{L}_{comprehensive}$.

[4] https://www.statista.com/statistics/1269778/gender-identity-worldwide-country/

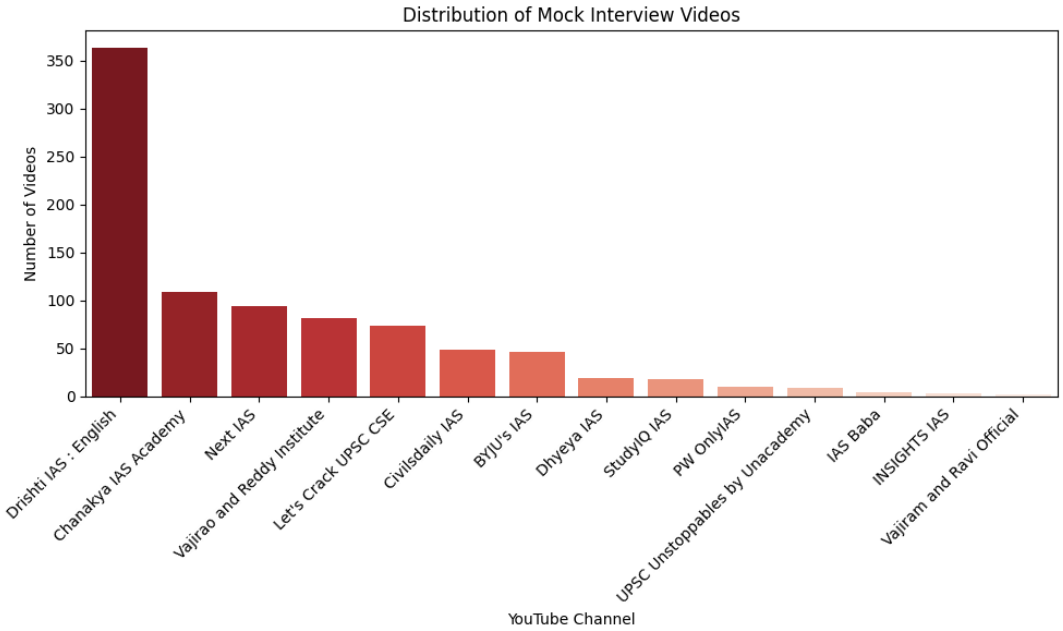


Figure 5: Distribution of mock interview videos across channels