Exploring the Role of Women in Hugging Face Organizations

Maria Tubella Salinas[1], Alexandra González[1], Silverio Martínez-Fernández[1]

[1]Universitat Politècnica de Catalunya

{maria.tubella@estudiantat.upc.edu,
alexandra.gonzalez.alvarez,silverio.martinez}@upc.edu

Abstract. Background: *Despite its impact on innovation, gender diversity remains far from fully being achieved in open-source projects.* ***Aims***: *We examine gender diversity in Hugging Face (HF) organizations, investigating its impact on innovation and team dynamics in open-source development projects.* ***Method***: *We conducted a repository mining study, focusing on ML model development projects on HF, to explore the involvement of women in collaborative processes.* ***Results***: *Women are highly underrepresented in both organizations and commits distribution, which is also found when analyzing individual developers.* ***Conclusions***: *Addressing gender disparities is essential to create more equitable, diverse, and inclusive open-source ecosystems.*

1. Introduction

Diversity is crucial in Software Engineering (SE), fostering innovation and effective problem-solving [Damian et al. 2024]. More specifically, gender diversity is being explored as, in recent decades, the software development field has been marked by a significant gender imbalance, with men overwhelmingly representing the majority of the workforce. For instance, only 26.8% of the people hired in Google's Tech area in 2023 were women [1]. This lack of gender diversity is prevalent, not only at the individual level but often across entire teams and organizations. In a survey conducted by [Kohl and Prikladnicki 2022], most respondents indicated that there is a lack of support for diversity in the SE field, with outdated patterns continuing to prevail, alongside instances of sexism and prejudice. Such gender homogeneity stands in contrast to the needs of software projects today, which rely heavily on collaboration, creativity, and diverse perspectives. Research shows that team diversity, including gender diversity, can enhance problem-solving capabilities, foster innovation, and lead to higher-quality outputs [Vasilescu et al. 2015]. Yet, many women who contribute meaningfully to the field remain underrepresented, often lacking visibility and recognition for their contributions.

Hugging Face (HF) is an advanced Machine Learning (ML) open-source platform known for its extensive library of pre-trained models, particularly in natural language processing and computer vision. Its ecosystem supports seamless model deployment, fine-tuning, and community-driven development, fostering rapid innovation. We focused our study in this platform due to its rapid growth and increasing popularity in both research and industry. Its versatility and active community make it an appealing choice for exploring modern ML workflows and tendencies.

---

[1]https://static.googleusercontent.com/media/about.google/ca//belonging/diversity-annual-report/2023/static/pdfs/google_2023_diversity_annual_report.pdf?cachebust=2943cac

On the HF platform, users can collaborate through entities called *Organizations* [2]. These HF Organizations serve as collaborative spaces where teams, research groups, companies, and other entities can jointly manage ML projects. By establishing an organization, members gain a shared environment for publishing, managing, and maintaining models, datasets, and other ML assets as a team. Any HF member can join one or more organizations, each of which can create and maintain an unlimited number of models. HF's openness is a foundational concept, essential for accurately interpreting the results.

Our research addresses multiple facets of gender composition and its potential influence on open-source ML projects. First, we aim to understand how HF organizations are structured in terms of gender diversity and whether this composition correlates with the popularity of the models they develop, measured by metrics such as downloads and likes. Secondly, we explore how women contribute within these organizations, particularly regarding the frequency and impact of their code commits. Analyzing commit patterns can offer insights into whether men or women tend to serve as primary contributors, potentially shaping project direction. Finally, we examine the nature of interactions between male and female members within these teams. By analyzing these dimensions, we aim to reveal how gender diversity manifests in HF organizations and explore its potential connections to organizational behavior and ML model success. This work ultimately aims to highlight the roles and contributions of women within this community, with broader implications for gender diversity in open-source software development.

2. Related Work

Extensive research has been conducted to study diversity across work groups, including both software development teams and teams from other domains. Understood broadly, diversity arises from any attribute that differentiates people [Phillips and O'Reilly 1998]. These attributes can be demographic (*e.g.*, age, gender, culture, religion, nationality), functional (*e.g.*, role, expertise), or subjective (*e.g.*, personality).

The impact of diversity on team outcomes remains a topic of debate, with conflicting findings in the literature. While some studies report a positive correlation between diversity and performance, others suggest a negative relationship. For instance, [Gomathy 2023] states that diversity can prevent grouping and lead to an inclusive workspace while also adding problems in how teammates communication.

2.1. Gender Diversity in Software Teams

Said by the World Health Organization, *gender* refers to "*the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men. It varies from society to society and can be changed*" [3]. Nevertheless, there are multiple definitions, as the American Psychology Association defines *gender identity* as a component of gender that describes a person's psychological sense of their gender. Many people describe gender identity as a deeply felt, inherent sense of being a boy, a man, or male; a girl, a woman, or female; or a nonbinary gender (*e.g.*, genderqueer, gender-nonconforming, gender-neutral, agender, gender-fluid) that may or may not correspond to a person's sex assigned at birth, presumed gender based on sex assignment [American Psychological Association 2015].

---

[2]https://huggingface.co/organizations
[3]https://www.who.int/health-topics/gender#tab=tab_1

Gender diversity has been increasingly recognized as a key factor in enhancing productivity and reducing turnover in software development teams. Studies show that gender-diverse teams tend to exhibit higher productivity, with partial positive effects on reducing employee turnover [Vasilescu et al. 2015]. This positive influence is linked to diverse perspectives in problem-solving and communication, which contribute to more dynamic and effective team interactions. The impact of gender diversity on employee performance is further influenced by several demographic and social factors, including the gender, age, marital status, education level, position, and work experience of employees. Research demonstrates a positive correlation between gender diversity and individual employee performance within organizations and statistical analyses reveal that gender diversity can be a predictive factor for employee performance, as diverse teams foster harmonious relationships and positive attitudes [Krishnan and Gokula 2020]. This, in turn, enhances overall team performance and adaptability to evolving market demands.

Another way to analyze the women's behaviour in teams is through studying its effect in *community smells*, which are organizational patterns associated with problematic team dynamics and communication structures [Palomba et al. 2018]. The top four more popular are: *Organizational silos*, when subgroups within a team are isolated, leading to fragmented knowledge and collaboration breakdowns; *Black clouds*, which describe persistent negative attitudes or dissatisfaction within a team, which can spread and impact morale; the *Lone wolf* phenomenon, that refers to individuals working in isolation from the rest of the team, limiting collaborative potential; and cohesion and *Radio silence*, that indicates a lack of communication, often resulting in misalignment and reduced efficiency in project execution. In terms of organizational culture, gender-diverse teams report fewer community smells compared to non-diverse teams [Catolino et al. 2019]. Specifically, gender diversity and women's participation are influential in mitigating issues such as *Black clouds* and *Radio silence*. However, the effect of gender diversity on other patterns, such as *Organizational silos* and *Lone wolves* appears to be less pronounced.

Nonetheless, women do not appear much in software teams; maybe because there are too few of them working in the field or maybe because of an inherent bias. The 2022 Stack Overflow Developer Survey [Stack-Overflow 2022] showed that only 5.17% of worldwide developers (from 180 countries) were women, while the 2023 data from Eurostat [4] revealed that just 17.4% ICT employees were women. Imtiaz *et al*.[Imtiaz et al. 2019] conducted an investigation to examine four effects on gender bias; *Tightrope*, when women behave in a restricted manner to avoid backlash, *Maternal Wall*, when women who are mothers have their competence and commitment questioned, *Prove-It-Again*, when women must provide more evidence than men to prove competence and *Tug of War*, when women are discouraging to other women. These four phenomena were first described in Williams and Dempsey's work [Williams and Dempsey 2014]. The investigation, performed with GitHub data, demonstrated that, while women are equally competent, they often limit their contributions to fewer projects and organizations compared to their male counterparts. Women also tend to demonstrate a more restrained approach in their professional interactions. Contrary to some expectations, the research has shown no significant differences between men and women in avoiding the posting of personal

---

[4]https://ec.europa.eu/eurostat/statistics-explained/index.php?title=ICT_education_-_a_statistical_overview

content, such as images with children, on professional platforms.

Table 1 shows a summary of the aforementioned related work. While gender diversity in software projects has been widely studied, most research is focused on larger and established platforms like GitHub [Vasilescu et al. 2015] [Imtiaz et al. 2019] or relied on survey-based approaches [Hodges and Murphy-Hill 2024] [Rastogi 2024]. This leaves a gap in understanding how gender diversity manifests itself in newer communities, such as ML engineers [Lanubile et al. 2024] in HF. This offers an opportunity to explore gender dynamics in a new area, motivating us to investigate this understudied area.

Table 1. Comparison of related work.

Study	Focus	Key Findings	Research Strategy
[Gomathy 2023]	Impact of diversity on communication in teams	Diversity can prevent grouping and create an inclusive workspace but may lead to communication challenges	Field strategy
[Vasilescu et al. 2015]	Gender diversity in software teams	Gender-diverse teams have higher productivity and lower turnover due to better problem-solving and communication	Data strategy
[Krishnan and Gokula 2020]	Gender diversity and individual performance	Gender diversity positively correlates with individual performance, fostering better relationships and attitudes	Data strategy
[Catolino et al. 2019]	Impact of gender diversity on organizational culture	Gender-diverse teams report fewer community smells, especially black clouds and radio silence	Data strategy
[Imtiaz et al. 2019]	Gender bias in software teams	Women tend to limit their contributions and restrict professional interactions due to bias	Data strategy
[Hodges and Murphy-Hill 2024]	Gender diversity in software teams	Researchers should also explore less visible challenges affecting diversity	Respondent strategy
[Rastogi 2024]	Gender diversity in software teams	Improving diversity in software development requires meaningful, sustainable solutions that balance business needs and inclusivity	Respondent strategy
This study	Gender diversity in HF	Analyzes the effects of gender diversity on HF organizations and models	Data strategy

The last column shows the research strategy based on [Storey et al. 2020].

3. Research Methodology

3.1. Goal and Research Questions

Following the Goal Question Metric (GQM) template [Basili et al. 1994], our goal is to *analyze* **gender diversity** *for the purpose of* **measuring it** *with respect to* **ML model commits** *from the point of view of* **ML engineers** *in the context of* **HF organizations of the HF Hub.**

To gain insights into organizational structures and ML model development processes in HF, we defined four Research Questions (RQ). **RQ1** focuses on understanding

the characteristics of HF organizations in terms of gender diversity to uncover patterns that define their composition.

> *RQ1: How are HF organizations built in terms of gender diversity?*

RQ2 and **RQ3** shift the focus to ML model commits and the contributions made by users to these projects. In this context, we aim to investigate how gender diversity within organizations relates to the number of downloads and success of the models they develop, as well as the behavioral patterns of individual contributors.

> *RQ2: Does the popularity of an ML model have a relation with the gender diversity in the organization developing it?*

> *RQ3: How does gender relate to the volume and contribution of commits in HF models?*

Building on the analysis of organizations and ML model contributions, **RQ4** compares gender diversity at the organizational level with that of individual contributors. We aim to see if gender patterns observed in organizations are also reflected in individual contributions.

> *RQ4: How does gender distribution in the top individual HF contributors compare to that in leading organizations?*

3.2. Repository Mining Process

As illustrated in *Figure* 1, this repository mining study is divided into three fundamental parts: data collection from the HF API, the use of GitHub API and Large Language Models (LLMs) to augment that data, and a final analysis stage.
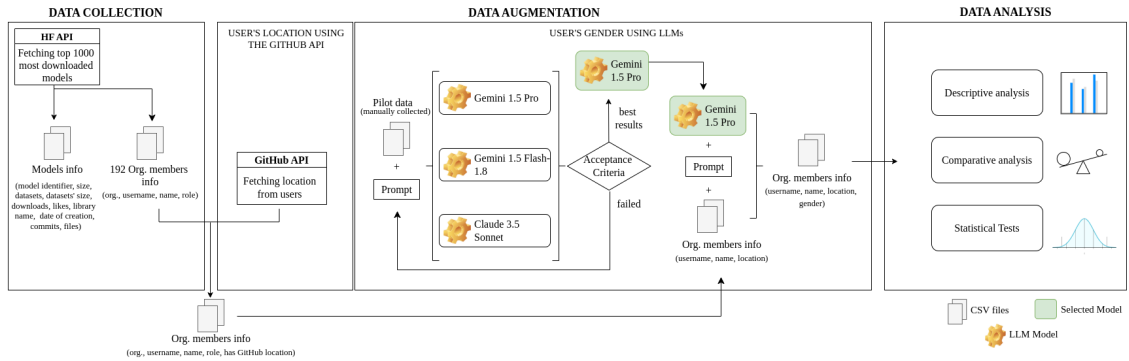

Figure 1. Repository Mining Study: Data Collection, Augmentation, and Analysis.

3.2.1. Data Collection

The data collection process began by using the HF API [5], following methods from [Castaño et al. 2024], to retrieve information about users, including their affiliations with organizations and details about the models they manage. We retrieved the **top 1,000 most downloaded models**, associated with **363 different developers**, including individual users who independently published models and organizations. Since our focus was

---

[5] https://huggingface.co/docs/huggingface_hub/package_reference/hf_api#huggingface_hub.HfApi

on studying organizations, we filtered the data to retain only the models associated with them, leaving us with **192 organizations** and their most downloaded models. For these organizations, we collected the necessary data. On the one hand, we obtained the data from the **models**: *model identifier, size, datasets, datasets' size, downloads, likes, library name, date of creation, commits* and *files*. However, to conduct our experiment we only used *model identifier, downloads, likes* and *commits*. Each *commits* data comprised a JSON with all ML model commits registered. Each commit is composed of these different fields: *commit identifier, authors, date of creation, title, message, formatted title* and *formatted message*. On the other hand, we saved the information about all **users** from those 192 organizations. That resulted in a total of 14,988 users from who we got *organization, username, name* and *role*.

At the same time, manually, we got the **top 100 individual contributors** from December 2024. About these users, we retrieved their *username* and *name* and obtained their *gender* by manually looking at their profile images or either of their socials.

3.2.2. Data Augmentation

User's location using the GitHub API Determining a user's gender directly from the HF API was not possible, as the platform does not include a specific field for users to disclose this information. However, the profile configuration page allows users to link external accounts, such as GitHub, X (formerly Twitter), LinkedIn, and Bluesky. While the API does not provide access to this data, it is possible to manually infer a user's gender by examining their profile photo on HF or by exploring these linked accounts.

Nevertheless, searching for the gender one by one for almost 15,000 people would mean a huge amount of time. In order to try to get the gender information we used the GitHub API. Because some users had their GitHub account linked, we used it to save their *location*. Again, obtaining gender information was not accomplished. Many are trying to but it seems like the pronouns are highly protected and, apart from doing it manually, it is not viable to know them. So, finally, the organization's members data we collected were *organization, username, name, role, has github* and *location*.

User's gender using LLMs Because the experiment's goal is the study of gender diversity, it is essential to know that about the users that compose an organization. To accomplish that, we relied on LLMs and followed the PRIMES framework [De Martino et al. 2025]. We focused on LLMs that provide an accessible API and offer some form of free credit for testing. Following a careful evaluation of available models, we selected Anthropic's *Claude 3.5 Sonnet*, which provides a $5 free credit, and Google's *Gemini 1.5 Flash* and *Gemini 1.5 Pro*, which offer $300 trial. To ensure clarity and reduce potential ambiguities arising from international naming conventions, we followed the recommendation by [Serebrenik 2024] to provide additional context beyond just names. We conducted a pilot using a manually collected **dataset of 360 randomly selected user samples** to validate the use of the LLM, and through several prompt refinements, we derived the following prompt to complete the dataset:

> **Final Prompt**
>
> Given the name [name] and, if not none, the location [location], tell me the gender of the username.
> - The input is a name and maybe a location too.
> - The output must be 'Male', 'Female', or 'Unknown'. Without any explanations or reasons. Just the predicted gender.
> - Examples:
>   - Christopher Pirillo: Male. Vera AxelRod: Female

We chose to introduce a third category, *unknown*, to prevent the LLM from being forced to classify individuals as either male or female. This approach was intended to mitigate the risk of the model providing correct answers purely by chance. This third category allows for more nuanced interpretations of results and reduces the potential for biased outcomes. To evaluate the accuracy, we used **Cohen's Kappa** [Cohen 1960], a statistical measure of agreement between two dependent categorical samples. An important consideration when interpreting the kappa is that the pilot data was built using additional information extracted from other platforms (such as HF, X and LinkedIn profile photos, published posts, Google Scholar pages and other socials appearances) while the models could only rely on name and location. We compared our manually collected data with the LLM gender classification to evaluate how well the model did. As shown in *Table 2*, the *Claude 3.5 Sonnet* requires more than twice the execution time of *Gemini's Flash* and nearly four times that of the *Pro*. Additionally, the free credit provided by *Anthropic* would be insufficient to process our entire dataset. Despite the comparable accuracy levels across all models, execution time and economic implications led us to select **Gemini 1.5 Pro**.

With the LLM, we augmented our dataset, which comprised each user's *organization, username, name, role, has github, location* and *gender*.

Table 2. Comparison of LLMs in the pilot test.

LLM	Execution Time	Economic Cost	Cohen's Kappa
Claude 3.5 Sonnet	12'	≈ $0.64	Moderate Agreement
Gemini 1.5 Flash	5'40"	≈ $0.50	Moderate Agreement
Gemini 1.5 Pro	3'46	≈ $0.50	Moderate Agreement

Explicit Cohen's Kappa values can be found in the Replication package 8.

3.2.3. Data Analysis

We used **descriptive analysis** to explore patterns of diversity and contributions within organizations. This method quantified important aspects such as gender distribution and commit characteristics, offering an overview of trends. Additionally, **comparative analysis** contrasted gender distributions and contributions between organizations and individual developers, providing deeper insights into diversity dynamics on the platform. Finally,

we conducted **statistical tests** for each RQ, with a significance level of $\alpha = 0.05$ for all analyses. For *RQ1*, we performed a χ^2 test to determine whether there is an association between organizations and gender distribution across the dataset. To address *RQ2*, we investigated the relationship between the proportion of women in an organization and the downloads of their most popular model, applying Spearman's correlation test due to the non-normal distribution of the data. For *RQ3*, we used the Kruskal-Wallis test to examine differences in the distribution of commits between men, women, and bots, as the normality assumption was not met. Lastly, for *RQ4*, we conducted a χ^2 test to explore the relationship between gender distribution in organizations and among individual contributors. The statistical tests and the null hypotheses are summarized in *Table* 3.

Table 3. Summary of RQs, Statistical Tests, and Null Hypothesis (H_0).

RQ	Test	Null Hypothesis (H_0)
RQ1	χ^2 test	No association between organization and gender.
RQ2	Spearman's correlation	No correlation between women's proportion and ML model downloads.
RQ3	Kruskal-Wallis test	Commit distribution is equal across groups.
RQ4	χ^2 test	Organization and contributor gender distributions are the same.

4. Results

4.1. Gender diversity in HF organizations (RQ1)

As illustrated in *Table* 4, women account for 12.06% of the total users across the 192 HF organizations analyzed [6] when including those with unknown gender and 17.21% when considering only users with an assigned gender. A closer examination reveals that the proportion of women in the studied organizations ranges from 0% to 50%, with no organization having a female majority. Notably, 88 organizations have no women on their teams, and 183 have less than 30% female representation. Among the few organizations where women constitute half of the team, all consist of four members or fewer, underscoring the limited female presence across these organizations.

Table 4. Users by gender in the 192 analyzed organizations

Gender	Count	Proportion (%)	Mean Users
Men	8,696	58.02%	45.29
Women	1,808	12.06%	9.42
Unknown	4,484	29.92%	23.35
Total	14,988	100%	78.06

Finding 1. There is a significant gender imbalance within HF organizations, with the number of women remaining low, even when considering only users with known gender.

---

[6]https://github.com/GAISSA-UPC/HF_diversity/tree/main/HF_diversity/HF_diversity/organisations/complete_organisations

The χ^2 test yielded a p-value of 3.28×10^{-57}, which is far smaller than the significance level. Consequently, we rejected the null hypothesis, concluding that the proportions of males, females, and unknown gender differ significantly across organizations. The result is consistent with the gender imbalance observed across the different HF organizations, as well as with the diverse team sizes and highly varied gender distributions.

4.2. Relationship between ML model popularity and gender diversity in HF Organizations (RQ2)

We selected the **most downloaded model** from each organization to explore whether there is a relationship between ML model popularity and gender diversity. We then examined the behavior of the top 15 organizations - organizations that developed the 15 models with the highest download counts - including: *sentence-transformers, nesaorg, facebookAI, google-bert, Qwen, microsoft, openai, facebook, openai-community, bigscience, google, timm, distilbert, pyannote* and *M-CLIP. Figure* 2 shows no evident correlation between team diversity and the number of downloads. In fact, the top 3 organizations, with the highest download counts, have no women on their teams, suggesting a potential inverse relationship between gender diversity and popularity. Among the top 15 organizations, more than half are composed entirely of male members, highlighting the absence of gender representation in the most successful teams. These observations challenge the assumption that diversity is positively associated with organizational success, at least in terms of download metrics. However, it is important to note that the popularity of an ML model does not necessarily equate to the popularity of its organization, as models are often developed by contributors outside of the organization. Thus, it would not be accurate to assume that the gender distribution within a team aligns with that of the ML model authors.
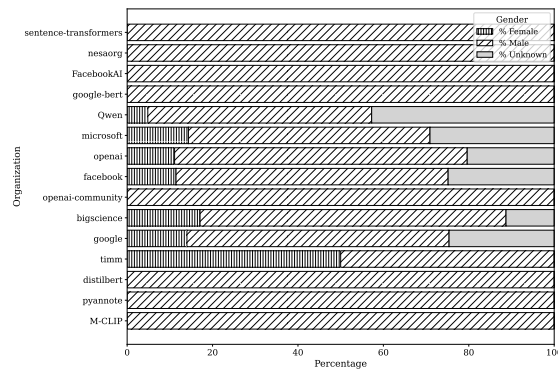

Figure 2. Normalized Gender Distribution of the Top 15 Organizations.

Finding 2.1. The most downloaded models do not exhibit a direct relationship with organizational diversity, as the top organizations often have no female team members.

Next, we analyzed **all 192 models** rather than just the top ones per organization, and obtained Spearman's correlation coefficient of -0.03 and a p-value of 0.7. The correlation value indicates a very weak negative relationship between the proportion of women and the number of downloads. Since the correlation is close to zero, it suggests that there is no clear relationship between the two variables. The negative direction implies that, on average, as one variable increases, the other decreases slightly, but this trend is negligible

due to the small correlation value. Furthermore, the p-value of 0.7 is much greater than the significance threshold of 0.05, providing no evidence to reject the null hypothesis (H_0). Therefore, we conclude that there is no significant correlation between the proportion of women and the number of downloads.

Finding 2.2. No significant correlation exists between the proportion of women and ML model downloads across all models.

4.3. Commits data and gender diversity (RQ3)

We further analyzed the commit distribution of models developed by the 192 organizations selected for this study. However, before diving into the analysis, several key observations were made.

Firstly, we observed that the commit distribution does not directly correspond to the distribution of organizations. Due to the open-source nature of the platform, individuals can contribute to models without being affiliated with a particular organization or owning the model. As a result, some users make contributions to over 30 models, despite not being members of any team. Additionally, we encountered what we refer to as **Bots**, which contribute a significant number of commits. However, these commits are not critical to ML model development, as they mainly focus on file-related tasks. Two notable bots, *system* and *SFconvertbot*, rank as the top two committers. The *system* bot is responsible for managing ML model files such as *.txt, .json,* and *.bin*, as well as tasks such as updating *README.md* or making the initial commit. In contrast, the *SFconvertbot* specializes in a single type of commit: *"Adding 'safetensors' variant of this model"*.

In our analysis of commit data, we recorded 5,274 commits across 192 distinct models (one per organization). The data necessary to complete the final component of this experiment was collected manually due to the nature of authors' independent commitments to models. Consequently, when combining data from the organizations with commit data, many authors could not be identified because the LLM did not process those users.

Initially, we analyzed the gender distribution among authors who had committed the most, based on the number of models they contributed to. As observed in *Figure* 3, there were no female authors among the top committers, with only five female authors contributing to a total of two models. Further examination of the authors with the highest total number of commits revealed a similar lack of female representation. Of the 5,274 commits collected, 83.35% were done by men, 6.48% by users of unknown gender, 5.84% by bots, and the remaining 4.3% were committed by female authors.

Recovering the previous analysis - *Figure* 2 - we explored how are the commits [7] of their respective models (*Figure* 4). As stated before, both gender arrangements do differ. When comparing both perspectives - one from the organization point of view and the other from their corresponding ML model- some differences can be found. The first one and obvious, bots are not part of any organization yet they commit in majority. Furthermore, in some cases, there are women committers even though that team is male-composed. That would be the case of *google-bert*'s *bert-base-uncased*, *openai-community*'s *gpt2* or *distilbert*'s *distilbert-base-uncased*. The contrary phenomenon appears too, as the models

---

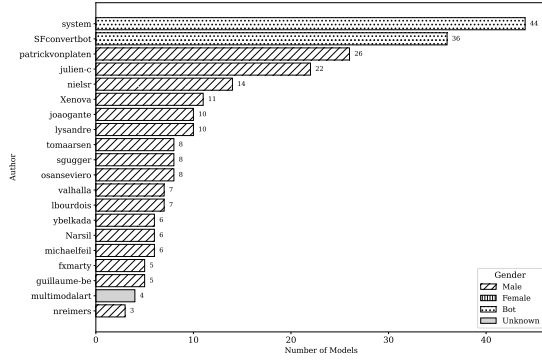[7] pyannote's ML model was not authorized to download, that's why CIDAS appears

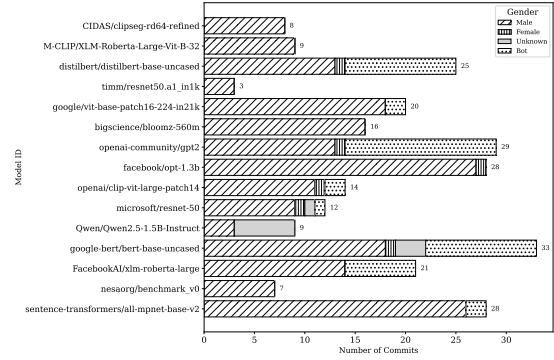**Figure 3. Top 20 Authors by Number of Models they Committed in.**

Figure 4. Number of Commits by Gender for the Top 15 Most Downloaded Models.

Qwen2.5-1.5B-Instruct, *bloomz-560m*, *vit-base-patch16-224-in21k* and *resnet50.a1-in1k* show no commits from female authors besides having women as members of the developing organization.

> **Finding 3.1.** The gender distribution of committers does not always align with the gender composition of the developing organizations.
> **Finding 3.2.** The most committers are men, with bots making more commits than women.

Despite the significant disparity in the quantity of contributions between men and women, we examined whether any differences could be observed in the nature of the commits rather than just their volume. By analyzing the commits' title, we categorized them into two main groups: commits related to ML model changes and those related to documentation. ML model updates such as fixing or adding variables, tokenizers, or weights were classified under the first group, while tasks such as creating and updating a README.md or a model card would be included in the second.

From a broader perspective, after conducting the Kruskal-Wallis test, we obtained a p-value of 0.008, which is smaller than the significance level of 0.05. Consequently, we rejected the null hypothesis and concluded that there are significant differences in the commit patterns across the different groups. These results are consistent with our earlier findings, where we observed that the number of commits made by men is significantly greater than those made by women or bots.

4.4. Comparison with top users outside HF organizations (RQ4)

To examine how the gender distribution in top organizations compares to that of individual contributors, we analyzed the *Top Contributors to Follow* [Valoatto]. This leaderboard, which ranks users based on metrics such as ML model downloads, dataset downloads, space likes, and profile followers, provides insight into individual developers. By focusing on ML model downloads, we applied the same approach used for organizations to study the distribution among top contributors. As shown in *Table* 5, the gender distribution among the top contributors closely resembles that of the leading organizations. While some differences exist, both groups follow a similar pattern in terms of representation across genders.

> **Finding 4.** Gender distribution is similar across organizations and individual developers.

Table 5. Gender Distribution in Top 100 Contributors and Top 192 Organizations.

Category	Men	Women	Unknown
Top 100 Contributors	60.0%	9.0%	31.0%
Top 192 Organizations	58.0%	12.1%	29.9%

Lastly, the Chi-squared test yielded a p-value of 0.644, leading us to not reject the null hypothesis, indicating no significant differences in gender distributions. These results align with the earlier findings, as shown in *Table* 5.

5. Discussion

Our analysis shows that women are underrepresented in HF organizations. Contrary to what Vasilescu *et al*.stated in [Vasilescu et al. 2015], we found that an ML model success is not directly related to the gender diversity in it. However, we can agree with Imtiaz *et al*., whose findings in [Imtiaz et al. 2019] exhibit that "women focus their work in fewer projects". This observation does not necessarily reflect a barrier to participation in open-source coding and development. Open-source platforms inherently allow anyone to contribute regardless of organizational affiliation. Our study reveals that model popularity does not consistently correlate with organizational diversity, as many leading organizations lack female representation. Furthermore, the gender makeup of contributors often diverges from the organization's internal gender composition, indicating that external contributions may play a significant role in shaping widely adopted models. This means that while women may not be prominent in formal organizations, their contributions to the open-source ecosystem can still be significant and impactful.

However, understanding the true extent of this impact requires more than just counting commits. It is crucial to note that the quantity of commits cannot be assumed to be proportional to the quality or significance of the work performed. As we only have access to commit *titles* and *messages*, we cannot assure the actual impact of each contribution. Therefore, it would be unfair to infer that men's work is more relevant or substantial than women's based solely on commit counts. It would be valuable to develop measures that assess the relevance of a commit — such as its influence on project development, the number of dependencies affected, or the extent of community engagement it generates. Such an approach could provide deeper insights into the meaningful contributions of women in open source beyond sheer participation rates.

Interestingly, the patterns observed in HF organizations appear to mirror those seen in the broader context of software development. A similar lack of gender diversity is evident when examining the most recognized individual contributors on the platform. What is more, this is a general tendency in the SE field. Survey-method work done by Kohl and Prikladnicki in [Kohl and Prikladnicki 2024] [Damian et al. 2024] found that respondents answered "it is common to see only one woman in teams or just a few".

The nature of open-source coding has posed challenges in analyzing the relationship between team composition and ML model commits. The decentralized and collaborative nature of open-source projects has made it difficult to reliably relate contributions to specific individuals or teams. Consequently, our findings may lack the precision needed

to draw definitive conclusions, and some of our results could be considered insignificant or unreliable due to these limitations.

6. Threats to Validity

In this section, we address potential threats to the study's validity, aiming to clarify the constraints and biases that could affect the interpretation of the findings.

Internal Validity: Since HF does not provide gender data, we relied on an LLM using names and locations to infer gender. To mitigate biases, we conducted a pilot and measured agreement using Cohen's Kappa. The results, however, indicate that this method cannot reliably replace human judgment, highlighting the need for human involvement to ensure accuracy in this context. Nevertheless, even though we classify some users as having an *unknown* gender due to insufficient information, if these users were indeed women, the proportions would still remain unbalanced. Additionally, analyzing contributions made by the same individual across multiple models has proven challenging. To address this, we aggregated all contributions by a user, regardless of their affiliation, ensuring a comprehensive perspective on their involvement across the platform.

External Validity: Our findings, based on HF organizations, may not fully generalize to all open-source ML communities or software ecosystems. However, by focusing on a prominent ML repository with a diverse range of organizations, our findings aim to be significant for the community. We analyzed the top 1,000 most downloaded models to assess whether model popularity is associated with gender diversity among contributors. However, the findings may not extend to less popular or niche models.

Construct Validity: Our study focuses on women's representation in HF organizations, using inferred gender data rather than self-reported identities, which may not fully capture non-binary contributors or broader diversity factors. As gender is not reported, we focus on women's roles, recognizing the existence of non-binary identities.

Conclusion Validity: Due to the absence of a gender field in HF, we inferred gender based on names and locations, considering regional variations. Despite potential variability, the study's conclusions are expected to remain reliable.

7. Conclusion and Future Work

Our study has revealed a significant gender imbalance within HF organizations, with women being notably underrepresented. However, no direct relationship was found between gender diversity in organizations and the popularity of the models they develop. Additionally, gender diversity was not present in the participation of committers, with men being the primary contributors. Despite this, no differences were observed in the quality or content of commits between men and women, indicating that women's contributions are not limited to specific tasks or roles. Their commits span a wide range of development activities, including core functionality, bug fixes, documentation, and feature enhancements. It is important to note that this analysis is based on commit titles and messages, which serve as the primary data for this study. Similar gender trends were observed among individual contributors, reflecting those in organizations. These findings highlight the need for greater gender inclusion in open-source communities, while also showing that gender does not influence the quality of contributions.

Future research could build upon this study by analyzing a broader range of models and organizations within HF. This would provide a wider and more representative view of gender diversity and contribution patterns. Moreover, integrating user feedback through surveys could shed light on the factors influencing contributors' decisions to engage in ML model development, helping to identify potential barriers or motivators for participation. Lastly, future studies could concentrate on evaluating the relevance and impact of commits by moving beyond surface-level attributes to gain deeper insights into their true contribution to ML model development and overall success. By doing so, we could be able to study different types of commits and committers, as in [Aguilera González et al. 2021].

8. Artifact availability

If wanted to reproduce the study, used codes and generated datasets can be found in this GitHub repository: https://github.com/GAISSA-UPC/HF_diversity.

Acknowledgment

References

Aguilera González, C., Albors Zumel, L., Antoñanzas Acero, J., Lenarduzzi, V., Martínez-Fernández, S., and Rabanaque Rodríguez, S. (2021). A preliminary investigation of developer profiles based on their activities and code quality: Who does what? In *2021 IEEE QRS*, pages 938–945.

American Psychological Association (2015). Key Terms and Concepts in Understanding Gender Diversity and Sexual Orientation Among Students.

Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The Goal Question Metric approach. *Encyclopedia of software engineering*, pages 528–532.

Castaño, J., Martínez-Fernández, S., Franch, X., and Bogner, J. (2024). Analyzing the Evolution and Maintenance of ML Models on Hugging Face. MSR '24, page 607–618.

Catolino, G., Palomba, F., Tamburri, D. A., Serebrenik, A., and Ferrucci, F. (2019). Gender Diversity and Women in Software Teams: How Do They Affect Community Smells? In *2019 IEEE/ACM ICSE-SEIS*, pages 11–20.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Damian, D., Blincoe, K., Ford, D., Serebrenik, A., and Masood, Z. (2024). *Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights*. Springer.

De Martino, V., Castaño, J., Palomba, F., Franch, X., and Martínez-Fernández, S. (2025). A Framework for Using LLMs for Repository Mining Studies in Empirical Software Engineering. *WSESE@ICSE 2025*.

Gomathy, D. (2023). Workplace Diversity and its effects on team dynamics and productivity. *Intl journal of scientific research in engineering and management*, 7(06):1–8.

Hodges, M. and Murphy-Hill, E. (2024). Perceptions of Software Developer Inclusion: A Survey at Google. In Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (pp. 207-230).

Imtiaz, N., Middleton, J., Chakraborty, J., Robson, N., Bai, G., and Murphy-Hill, E. (2019). Investigating the Effects of Gender Bias on GitHub. In *2019 IEEE/ACM ICSE*, pages 700–711.

Kohl, K. and Prikladnicki, R. (2022). Benefits and Difficulties of Gender Diversity on Software Development Teams: A Qualitative Study. In SBES '22, 21–30.

Kohl, K. and Prikladnicki, R. (2024). Gender Diversity on Software Development Teams: A Qualitative Study. In Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (pp. 169-184). .

Krishnan and Gokula, D. S. (2020). Gender Diversity in the workplace and its effects on Employees' Performance. *Journal of the Social Sciences*.

Lanubile, F., Martínez-Fernández, S., and Quaranta, L. (2024). Training Future Machine Learning Engineers: A Project-Based Course on MLOps. *IEEE Software*, 41(2):60–67.

Palomba, F., Tamburri, D. A., Serebrenik, A., Zaidman, A., Fontana, F. A., and Oliveto, R. (May 2018). How do community smells influence code smells? *In ICSE 2018*, pages 240–241.

Phillips, K. and O'Reilly, C. (1998). *Demography and Diversity in Organizations: A Review of 40 Years of Research*, volume 20, pages 77–140.

Rastogi, A. (2024). Roads Ahead to Diversity and Inclusion by Software Engineering. In Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (pp. 3-16).

Serebrenik, A. (2024). How to Ask About Gender Identity of Software Engineers and "Guess" It from the Archival Data. In Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights (pp. 487-505).

Stack-Overflow (2022). Stack Overflow Developer Survey. `https://survey.stackoverflow.co/2022/#developer-profile-demographics`.

Storey, M.-A., Ernst, N. A., Williams, C., and Kalliamvakou, E. (2020). The who, what, how of software engineering research: a socio-technical framework. *Empirical Software Engineering*, 25:4097–4129.

Valoatto, M. Top Contributors To Follow - a Hugging Face Space by mvaloatto — huggingface.co. `https://huggingface.co/spaces/mvaloatto/TCTF`. [Accessed 30-01-2025].

Vasilescu, B., Posnett, D., Ray, B., van den Brand, M. G., Serebrenik, A., Devanbu, P., and Filkov, V. (2015). Gender and Tenure Diversity in GitHub Teams. CHI '15, page 3789–3798.

Williams, J. C. and Dempsey, R. (2014). What works for women at work: Four patterns working women need to know. In *What Works for Women at Work*. New York University Press.