GITHUB | Project Description | Notion | OneDrive

New Commonwealth Fund: LinkedIn Leadership

Boston University

Faculty of Computing and Data Science

Team Members

Natalie Malave Koprowska	CDS'27	nmakop@bu.edu
Aryan Singh	CDS'25	aryan3@bu.edu
Samritha Aadhi Ravikumar	MSDS'25	samritha@bu.edu
Hongzun (Aimee) Zhang	MSDS'25	hongzun@bu.edu
Atul Aravind Das	MSDS'25	atuladas@bu.edu

March 2025

Introduction

Project Overview and Goals:

This project aims to explore racial demographics within nonprofit leadership by analyzing approximately 100,000 public LinkedIn profiles of nonprofit leaders, primarily based in Massachusetts. The primary objective is to identify trends in leadership representation, with a focus on people of color (POC) in executive and officer roles. Currently, nonprofits are not required to self-report leadership demographics, which creates a gap in understanding the diversity of leadership within the sector.

The analysis leverages keyword-based inference techniques to identify leadership titles and group affiliations that may suggest racial demographics. This approach ensures transparency and ethical integrity by relying solely on publicly available information. The project seeks to provide actionable insights into diversity trends, supporting broader efforts to increase equity and transparency across the nonprofit sector.

Key Goals:

- Analyze leadership demographics to understand the representation of POC in nonprofit leadership roles.
- Identify trends and disparities in executive and officer roles within nonprofit organizations.
- Provide data-driven insights to inform more equitable leadership practices in the nonprofit sector.
- Maintain data integrity and ethical responsibility by using only publicly shared information.

Big Impact:

The current political administration has significantly impacted cultural, diversity, equity, and inclusion (DEI) initiatives, which has posed challenges for organizations and hiring employers who are still committed to advancing equity among their clients and applicants. Identifying clients who align with these values has become more complex, but platforms like LinkedIn provide a space for people using online profiles to self-describe their activities, experiences, and identities. The ability to freely express one's background is crucial for people to highlight their commitment to the DEI effort. Even as many informal programs face setbacks, keyword inferencing can prove as an effective way of identifying candidates who align with these values.

This project addresses a critical gap in understanding diversity within nonprofit leadership. Currently, Black and Latino nonprofit leaders receive only 4% of philanthropic funding in the U.S., despite making up roughly 10% of nonprofit leadership. By uncovering trends in leadership representation, this project empowers funders, policymakers, and nonprofit boards to make more informed decisions that promote equity and inclusion.

The insights generated from this project will help:

- **Inform funding decisions** to better support POC-led nonprofits.
- **Promote systemic change** by increasing transparency in leadership demographics.
- **Support equitable leadership development** through targeted strategies and resources.
- **Enhance accountability** within the nonprofit sector by providing measurable data on representation.

Client Information:

Matt Taylor, the Tech Lead at the Data Nutrition Project (DNP), is spearheading the partnership between DNP and the New Commonwealth Racial Equity and Social Justice Fund (NCF) for the *Equity in Funding: Data Project*. DNP works alongside NCF to tackle inequities in nonprofit funding and improve transparency around racial representation in nonprofit leadership. Through this collaboration, Matt and the DNP team are developing tools and methodologies that will help amplify the voices of leaders of color, ensuring that they receive equitable recognition and support within the nonprofit sector. Matt's leadership is pivotal in using data-driven approaches to foster systemic change and advance racial equity across the nonprofit sector.

Data Description

The dataset, sourced from BrightData, includes approximately 100,000 public LinkedIn profiles of *nonprofit leaders*, primarily in Massachusetts. It contains structured information such as positions, current companies, locations and professional affiliations. The analysis focuses on education, languages and organizations to infer demographic trends using keyword-based methods. This approach ensures transparency and accuracy by relying on publicly shared information rather than potentially biased name-based models.

Column Name	Description	Our Analysis
position	Current job title	It is a string
about	Self-written bio	It is a string
groups	Organization they are part in	Many missing values; need to parse
activity	Post ID, Title, Links of the post	Lengthy; need to parse
current_company	Company Id, Name, Job title	Matching to experience, find current/past non-profit work; need to parse
experience	Full list of company name, id, job title. Some have duration, location	Could match with current_company; Need to parse
educations_details	College name	It is a string
education	Degree, description, college, field, start-end dates, institute logo	Need to parse
recommendations_count	# of recommendations	It is a string
courses	Courses they have done	It is a string
languages	Title/language name, subtitle/proficiency	Need to parse
certifications	meta/issued date, subtitle/course name, title/course company	Need to parse
volunteer_experience	Date, duration, their role, title, cause	Need to parse
followers	No. of followers	Low priority
connections	No. of connections	Low priority
current_company_company_id	Current company name	Same as experience
current_company_name	Company name, title,	It is a string
publications	Name, Title & where it is published	Many missing values/never published; Need to parse
patents	List of patents held	Plan to drop

projects	Description, title, date	Lengthy; need to parse
organizations	Title/organization name, membership_type, start_date / end_date	Lengthy; need to parse
location	City name. Could be where they work or where they live.	It is a string
activity	Post ID, Title of liked post, Live links to posts each person has liked	Lengthy, need to parse. Contains real names
linkedin_num_id	Unique ID value	Required if needed to merge dataset

Data Cleaning

The initial issue faced was that some columns, such as education, organization, and languages, were in the format of a JSON file, i.e., a list of dictionaries. The JSON-formatted data contained important parameters necessary for the analysis. The precise data was extracted from the JSON format by identifying and retaining valid strings while removing invalid ones. For example, 83,670 valid strings were extracted for education and 25,872 for organization. The extracted strings were loaded into a JSON parser to convert them into individual dictionaries, which were then transformed into structured data frames. To enhance the data frames further, the LinkedIn numeric IDs were concatenated with the values to associate them with personal details, ensuring consistency and completeness in the dataset.

Keyword Research

The selection of keywords for this task was guided by our commitment to equity and cultural awareness. We took a data-driven approach to ensure keyword accuracy and captured the span of LinkedIn users's racial and ethnic self-identification completely. combining both automated methods and manual scraping to identify relevant LinkedIn profiles. We manually researched for references supporting Massachusetts' demographic data, and found keywords in official reports, census data, and research studies that focus on marginalized communities. The chosen keywords aim to accurately capture racial and ethnic diversity within Massachusetts' nonprofit leadership sector while ensuring respect for privacy and ethical data usage.

To ensure a comprehensive and unbiased keyword selection, we utilized resources such as the Massachusetts diversity reports, census records, and academic publications. For example:

- Native American Representation: Information about the Mashpee Wampanoag and Aquinnah Wampanoag tribes was sourced from Brookline's Indigenous Peoples resource to ensure that Indigenous communities were identified accurately in our analysis.
- Middle Eastern and North African (MENA) Representation: Recent guidance on 2023 Census Bureau data on MENA demographics on each US state's population specifies the most prevalent countries of origin.
- Asian American and Pacific Islander (AAPI) Representation: The APIAVote Massachusetts Report includes their findings from statewide population and electorate numbers using the Census data, reflecting the most prevalent communities of origin.
- Latinx Representation: Research from UMass Boston's Gastón Institute for Latino Community Development and Public Policy Publications informed the inclusion of Latinx-specific terms and the most prevalent countries of origin in Massachusetts.
- Black Representation and HBCU Greek Organizations: Along with baseline demographic keywords, we also included historically Black Greek organizations (Alpha Phi Alpha, Delta Sigma Theta, etc.) qs keywords using prior knowledge of Howard University's Pan-Hellenic/Divine Nine umbrella Greek organization council, which signal values of high community engagement. As keywords in our project, this ensures that leadership connections in Black communities are recognized.

Along with focusing on the keywords for broader racial groups and ethnic identification in Massachusetts, we manually analyzed the organizations affiliated with the LinkedIn members in our dataset to identify additional keyword patterns and trends related to academic and leadership roles beyond identity. Below is a chart showing a select group of keywords and their accuracies in comparison with ground truth manual analysis. We included keyword examples that their queries produced. Keywords with low detection accuracy will need to be reconsidered or expanded for clarity.

Keyword	Match Counts	Dataset Ground Truth Count	Detection Accuracy (Comparing with Excel)	Select Keyword Examples
equity	82	83	98.8%	"Actors' Equity Association", "Engaged Donors for Global Equity - EDGE Funders Alliance"
diversity	44	46	95.7%	
minority / resource group / leadership initiative	31	31	100%	"Minority Women in Medicine"
black	152	159	95.6%	"Black Professionals Association Charitable Foundation", "Black Student Forum"
alpha kappa alpha / delta sigma theta / alpha phi alpha	66	66	100%	
nigerian / african american	44	44	100%	"Nigerian Institution of Mechanical Engineering"
national association of black / naacp / black student union / national black mba association	64	64	100%	"National Association of Black Journalists", "National Association of Black Professional Organizers", "NAACP Boston Branch"
vietnamese	9	10	90%	
chinese	22	22	100%	"Association of Chinese Americans"
indian	32	49	65.3%	"Narragansett Indian Tribe"
guatemalan	1	3	33.3%	"Guatemalan Cultural Group"

Large Language Model Implementation

In our research, we used the Mistral-Nemo-Instruct-2407, an instruct fine-tuned Large Language Model (LLM) version of the Mistral-Nemo-Base-2407.Trained jointly by Mistral AI and NVIDIA, it significantly outperforms existing models that are smaller or similar in size. We set the max token as 1000 to fully reason the likelihood of POC-indication in terms of language and education background factors. We set the temperature as 0.0 to ensure its output is fixed every time we run the model.

Originally, we chose Meta Llama as our language model, because it's the most popular and common model now. After doing some further research, we decided to implement the Mistral-Nemo model to get a more precise answer.

	Context Window	HellaSwag (0-shot)	Winogrande (0-shot)	NaturalQ (5-shot)	TriviaQA (5-shot)	MMLU (5-shot)	OpenBookQA (0-shot)	CommonSense QA (0-shot)	TruthfulQA (0-shot)
Mistral NeMo 12B	**128k**	**83.5%**	**76.8%**	**31.2%**	**73.8%**	68.0%	**60.6%**	**70.4%**	**50.3%**
Gemma 2 9B	8k	80.1%	74.0%	29.8%	71.3%	**71.5%**	50.8%	60.8%	46.6%
Llama 3 8B	8k	80.6%	73.5%	28.2%	61.0%	62.3%	56.4%	66.7%	43.0%

Table x: Mistral NeMo base model performance compared to Gemma 2 9B and Llama 3 8B.

As with working with any LLM, the prompt is the most essential part that must be carefully designed to get the most precise answer. Below is our prompt sentence and the sample answer, which includes strong reasoning and precise calculation process. We asked the model to give a likelihood of POC that falls into four categories: less than 25%, 25%-50%, 50%-75% and above 75%. Respectively group (A, B, C, D). We took into account that each user_id might have multiple rows that implied different degrees they acquired, so we chose the category with the highest percentage of POC-indication as the final result.

```
for college in colleges:
    question = (
        f'Please determine the likelihood of students being people of color at {college}. '
        'First, based on the information of the college, describe demographic information of students'
'Then, based on the above information, determine the percentage of students being non-hispanic white (NHW)'
'Finally, describe the likelihood in a percentage way in four categories: "A: less than 25%", "B: 25%-50%", "C: 50%-75%", "D: more than 75% .'
'and provide your final answer in the end. '
'In conclusion, your final answer (!letter only! !no other text!):'
    )
```

```
Prompt: 'Please determine the likelihood of students being people of color at UMass Boston. First, based on the information of the college, d
escribe demographic information of studentsThen, based on the above information, determine the percentage of students being non-hispanic whit
e (NHW)Finally, describe the likelihood in a percentage way in four categories: "A: less than 25%", "B: 25%-50%", "C: 50%-75%", "D: more than
75% .and provide your final answer in the end. In conclusion, your final answer (!letter only! !no other text!):', Generated text: " A, B, C,
or D\n\nUMass Boston, a public research university, has a diverse student body. According to the university's data, here's the demographic br
eakdown of students as of Fall 2021:\n\n- Undergraduate Students:\n  - Total: 16,572\n  - Non-Hispanic White: 3,744 (22.6%)\n  - Hispanic/Lat
ino: 4,212 (25.4%)\n  - Black or African American: 3,024 (18.2%)\n  - Asian: 2,204 (13.3%)\n  - Two or More Races: 1,310 (7.9%)\n  - Non-Resi
dent Alien: 1,028 (6.2%)\n  - Race and Ethnicity Unknown: 1,050 (6.4%)\n\n- Graduate Students:\n  - Total: 4,265\n  - Non-Hispanic White: 1,1
24 (26.3%)\n  - Hispanic/Latino: 843 (19.7%)\n  - Black or African American: 644 (15.1%)\n  - Asian: 704 (16.5%)\n  - Two or More Races: 304
(7.1%)\n  - Non-Resident Alien: 301 (7.1%)\n  - Race and Ethnicity Unknown: 209 (4.9%)\n\nBased on the above information, the percentage of s
tudents being non-hispanic white (NHW) at UMass Boston is approximately 24.4% (22.6% undergraduate + 26.3% graduate).\n\nTherefore, the likel
ihood of students being non-hispanic white at UMass Boston is: B (25%-50%)."
```

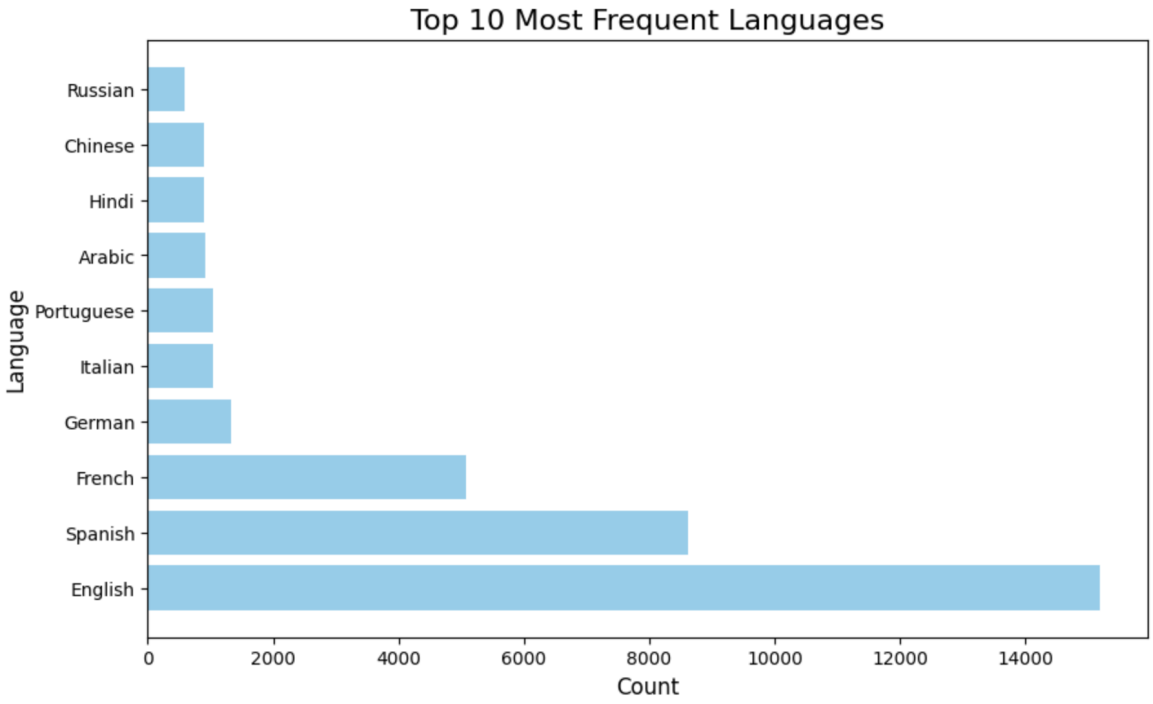
The same process was implemented for language, where LinkedIn users indicate proficiency.

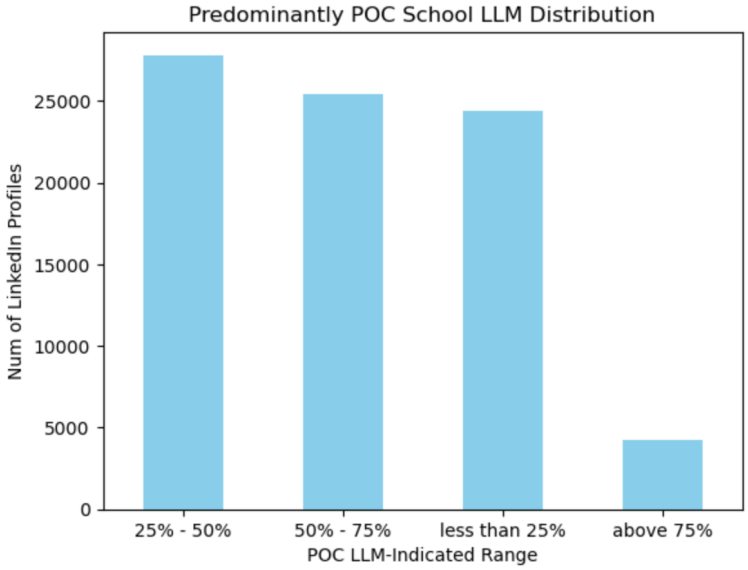
```
for l in language:
    question = (
        f'Please determine the likelihood of person being people of color at {l}. '
        'You should know that all of people are now working in the United States'
        'First, from the LinkedIn information about the language they speak and the relative proficiency level, describe possible geographic information for their upbring'
'Then, from the above information, infer and describe possible race and ethnicity of this person'
'Finally, based on the informataion above, if this person is more likely to be non-hispanic white, output 0, else output 1 '
'your final answer (0 for non-hispanic white; 1 for otherwise): '
    )
    prompts.append(question)
```

In short, the prompt was set in three steps: first, to clarify the background that all the users are now working within the United States, and the data was collected from web-scraping LinkedIn profile

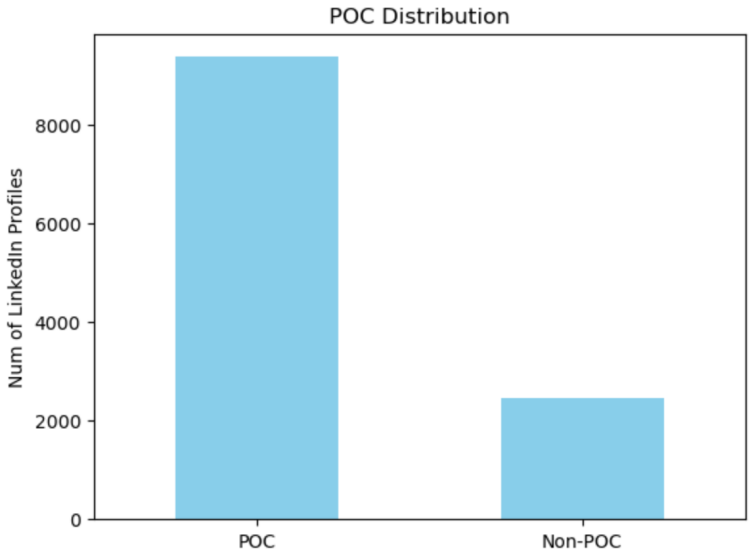
pages. Then, we asked what is the most likely place the user may have been born in. Finally, to make the model more efficient, we solely let it decide the likelihood of the user showing indications of being a POC if the probability is above 50%.

Visualizations

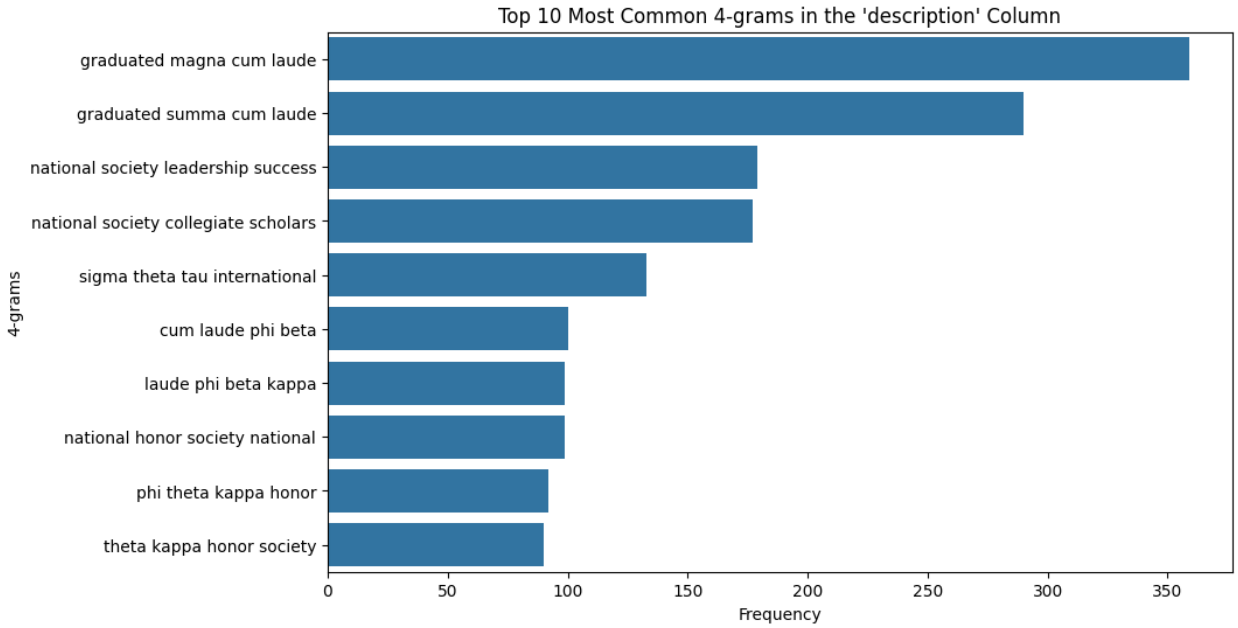


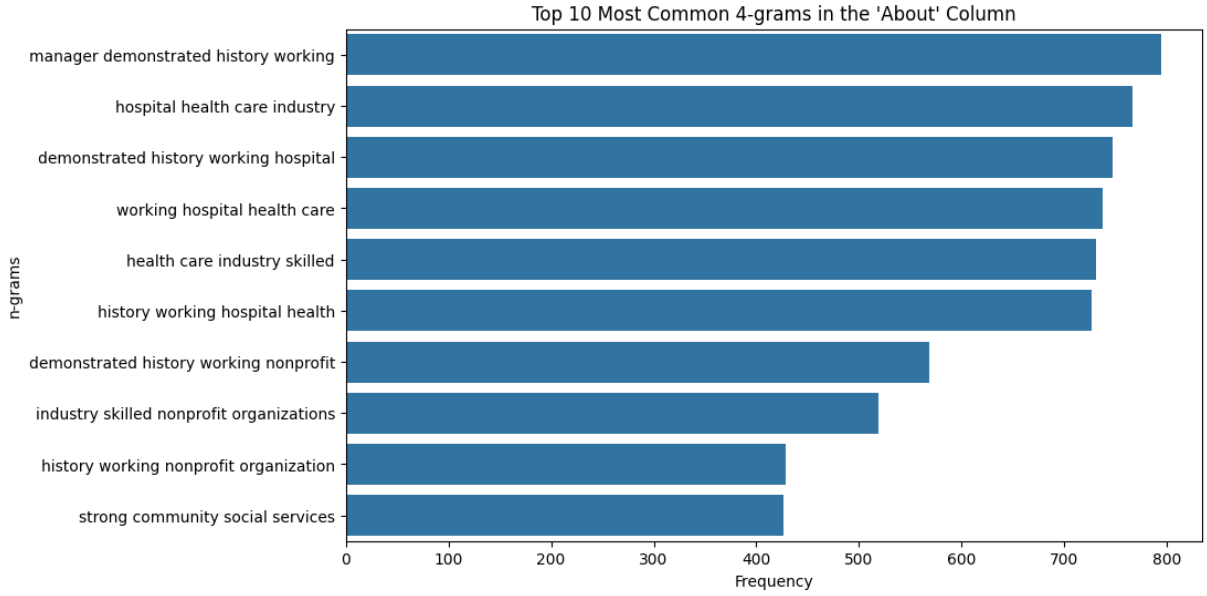


The chart above shows the result of our LLM algorithm based on the education background of each LinkedIn user. Since some of the users may have acquired multiple degrees, instead of combining all of the factors, we chose to select the college with the highest percentage of POC as their unique education background.

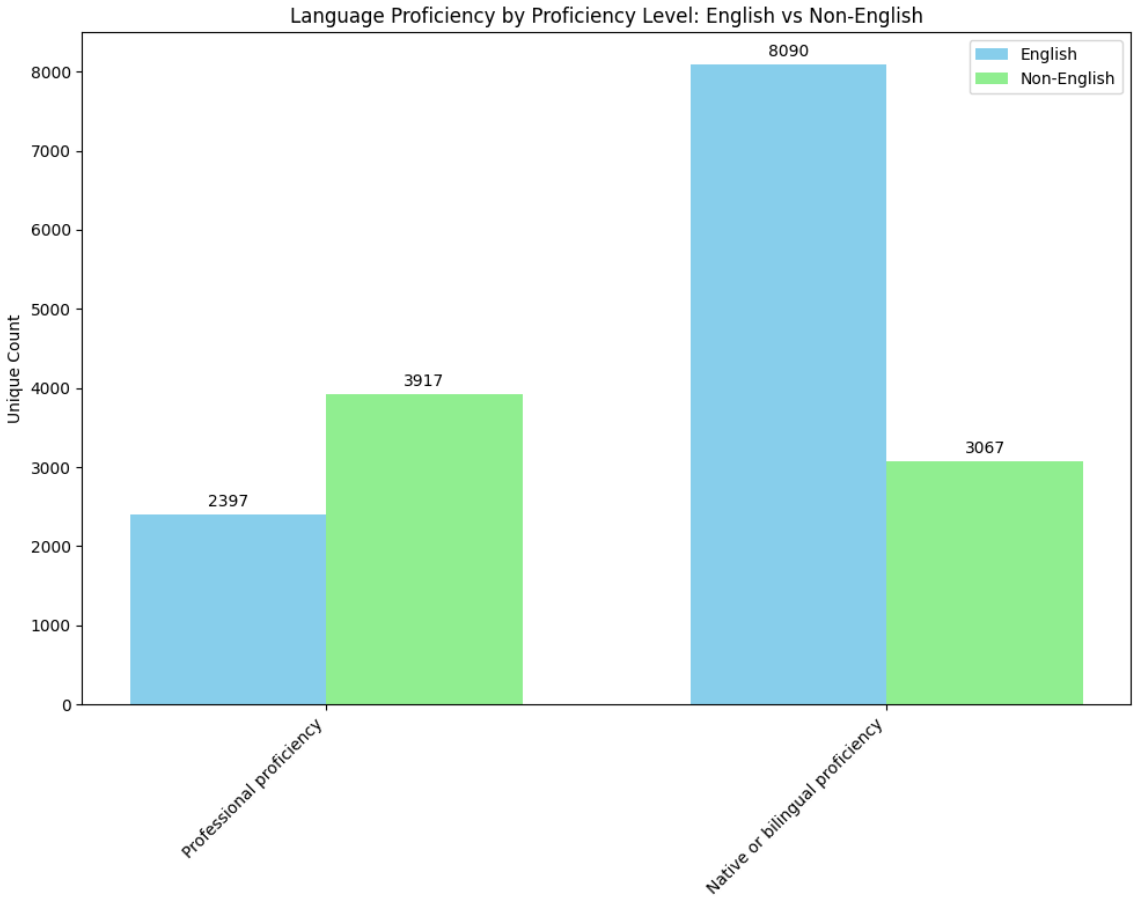


Also, The chart above indicates the result based on the proficiency of the language that each user speaks. According to the chart, there are more than 8000 users who are probably POC, while only around 2000 users are Non-POC.





The two graphs above gave us insights into the major recurring word combinations in the description column as well as the “about” column. This helped us in the Pre-analysis stage by giving us guided keyword research.



Another category that was very helpful in coming up with our column weightings and as indicators of POC was the language column. As you see in the graph above there was a good amount of LinkedIn profiles with indications of being POC, as highlighted by the Professional Proficiency in another language.

```
Keyword Match Counts:
equity: 696
diversity: 205
black: 70
racial equity: 26
indian: 20
latino: 17
latina: 16
```

```
Keyword Match Counts:
black: 152
equity: 82
diversity: 44
hispanic: 44
latino: 44
indian: 32
alpha kappa alpha: 29
```

These two screenshots above are the counts of some of our keywords listed in the “about” and “organizations” columns, providing some proof that our research is accurate.

Conclusion

This project has successfully addressed a critical gap in understanding the diversity of nonprofit leadership by analyzing approximately 100,000 public LinkedIn profiles from nonprofit leaders in Massachusetts. Our keyword-based inference approach—bolstered by rigorous data cleaning, JSON parsing, and the innovative use of a fine-tuned Large Language Model—enabled us to identify meaningful trends in leadership representation, particularly for POC in executive and officer roles.

By leveraging publicly available data, our methodology ensures ethical integrity and transparency, offering actionable insights for those who may use our insights. These insights not only support more equitable funding decisions but also contribute to systemic change in advancing diversity, equity, and inclusion within the nonprofit sector. Ultimately, our findings pave the way for continued efforts to enhance accountability and foster a more inclusive leadership landscape.

Practical Recommendations/Future Improvements

We could explore columns that were preliminarily dropped for further analysis, create an additional keyword list that helps identify Non-POC, and merge individual data files into one large dataset to accomplish a holistic analysis. In the future, we would like to merge datasets together based on a unique linkedin ID and use an algorithm to check which columns support the person being POC. It would also help to have a third party proof check if POC inference was correct based on real profile.
In the future scope of this project, we would like to implement a weighing model: Highest level of proficiency in that language could be weighed with whether they are a part of relevant organizations, or whether the university was indicated to have high demographic diversity rates.

Below details our recommendation for a weighing model's percentage rates. The higher weights on the About and Organization sections are due to the personal and self-descriptive nature on LinkedIn, and a lower weight was assigned to the University section to account for the fact that our LLM model analysis may not give an accurate reading of the POC-indication of someone who went to a University outside of the US.

- About section: ~ 35%
- Languages: ~20%
- Organizations: ~ 30%
- University: ~ 15%

References

1. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

2. https://datausa.io/
3. https://www.brooklinema.gov/DocumentCenter/View/20453/
4. https://www.census.gov/library/stories/2023/09/2020-census-dhc-a-mena-population.html
5. https://apiavote.org/wp-content/uploads/Massachusetts-2020.pdf
6. https://www.bestcolleges.com/resources/hbcu/greek-life/
7. https://www.nphchq.com/about
8. https://scholarworks.umb.edu/cgi/viewcontent.cgi?article=1261&context=gaston_pubs

Team Contribution

- **Natalie Malave Koprowska:** Team facilitator, Keyword Research, Processing and Analysis
- **Aryan Singh:** Scrum Master, Data Pre-analysis, Data Processing, Analysis, and ideation.
- **Samritha Aadhi Ravikumar:** Design Lead, Data Processing and Analysis, Documentation and Presentation
- **Hongzun (Aimee) Zhang:** Clients Liaison, Data Processing and Analysis, Large Language Model Implementation and Documentation
- **Atul Aravind Das:** Tech Lead, Data Cleaning, GitHub handling