# Cornell Course Classification and South Asian Representation

**Katie Huntley**
(kah294)

**Jupyter Notebook**: https://drive.google.com/file/d/11F0hHomCntNcM4GsjaBhLDSClSTb-ZcV/view?usp=sharing

## 1 Introduction

“I would found an institution where any person can find instruction in any study.” On October 7, 1868, at the official dedication of Cornell University, Ezra Cornell expressed that Cornell University would welcome people from all walks of life to pursue their personal academic interests. Present day, Cornell University strives to uphold its founding principle, offering nearly 80 majors and more than 120 minors across 9 colleges and schools. I write this in November 2024, as many students eagerly await to pre-enroll in their spring 2025 semester classes, but scrolling through the course roster, some students have begun to question whether the university truly sustains “any person, any study.”

In late October 2024, the Cornell Daily Sun (the university newspaper) published an opinion piece from the South Asian Council at Cornell titled “‘Any Person, Any Study’ Cannot Stop with South Asia" (1). The council expresses concern over the lack of South Asian courses, stating that “in the past five years, many faculty members teaching topics relevant to South Asia have either taken extended sabbaticals or have left Cornell due to fellowship opportunities and retirement.” More specifically, the authors cite that “Of the 49 faculty affiliated with the [South Asian] program, 14 are emeritus or retired. Another 18 are associate professors whose main focus lies outside the subcontinent.”

The South Asian Council at Cornell backs their claims with faculty statistics, but lacks data related to course offerings directly. Using the Cornell University course roster as my dataset and binary, no-shot classification, I intend to more thoroughly test the student testimonial that South Asian course offerings at Cornell University have decreased over time.

## 2 Data and Methodology

### 2.1 Data

I restrict my dataset to the Cornell University humanities classes offered during fall and spring semesters within the past 4 years (fall 2021 to spring 2025). I choose this timeframe given that, according to the Daily Sun article, the number of faculty teaching South Asian courses has decreased “in the past five years.” I shorten my timeframe to 4, rather than 5 years, because of potential impacts of the covid-19 pandemic in the 2020/2021 school year.

I classify humanities courses as any course that falls under one of the 31 departments listed in Appendix A, for example music, philosophy, and history. I created the list of humanities departments with the help of ChatGPT; I prompted the model with the complete list of Cornell University departments (187 in total) and the statement “I have this list of departments. Which would you consider humanities departments?”

For each course, I extract the semester in which it was offered, the course title, and the course description. From the fall 2021 semester through the spring 2025 semester, I gathered 8239 total humanities courses.

## 2.2 Pre-processing

After skimming through the dataset, I chose to pre-process my data in the following four steps:

First, I chose to remove "FWS: " from the beginning of freshman writing seminar course titles, since the classification model may overfit to this feature. For example, a course initially titled “FWS: Culture, Society, and Power” would be renamed to “Culture, Society, and Power.”

Second, I eliminated any courses with a description length of less than 50 characters. This narrowed my dataset from an initial size of 8239 to 8001 (about 200 samples discarded).

Third, I treat cross listed courses as the same course. The term, "cross listed courses", refers to courses that have the same course title listed across different departments. For example, in the fall 2021 semester, Modern China falls under both the Asian Studies and History departments. In my modified dataset I keep just one copy of courses that share the same title in the same semester. This further narrows my dataset from 8001 courses to 4804.

Lastly, I chose to remove courses related to independent study or honors research. Although self-guided work offers plenty of merit, for the purpose of this study I want to focus on what type of directed curricula Cornell University offers its students. I filter out these courses by removing any course with the keywords “Independent,” “Honors,” or “Thesis” in its title.

After all pre-processing steps, 4451 (roughly 54%) of the original 8239 courses remain in my dataset.

## 2.3 Classification

I opt to use zero-shot learning with BART-large-mnli to perform binary classification (“South Asian” versus “not South Asian”) on my dataset. I choose zero-shot learning because my dataset lacks gold standard labels and manually annotating a sufficient number of documents to train, validate, and test a model would be time intensive. As input for each model I combine the course title and full description, separated by a newline character. Note that the input does not include the course’s department. I define a South Asian course as any course relating to one or more of the following countries: Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. I gather this list of countries from a Cornell University library website, "Statistics Sources for Asia: An Introduction: South Asia".

As compared to other large language models like FLAN-T5, BART-large-mnli takes a list of categories as input without the need for a prompt. Category lists can be of arbitrary lengths.

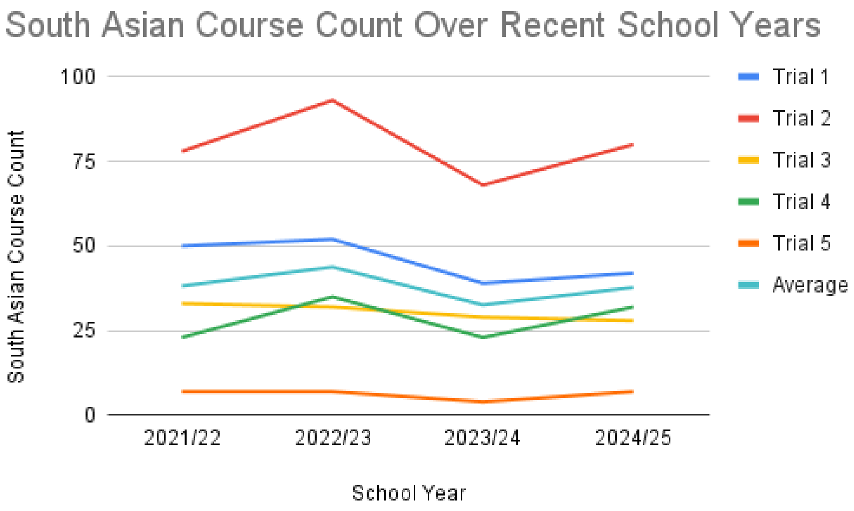
As such, rather than prompt engineering, I shift my attention to category engineering. In five separate trials, I use each of the following category lists as input to BART-large-mnli for Cornell course classification:

| Trial | Category List |
|---|---|
| Trial 1 | ["South Asian", "not South Asian"] |
| Trial 2 | ["South Asian topics", "miscellaneous topics"] |
| Trial 3 | ["South Asian countries", "miscellaneous"] |
| Trial 4 | ["Afghanistan", "Bangladesh", "Bhutan", "India", "Maldives", "Nepal", "Pakistan", "Sri Lanka", "other"] |
| Trial 5 | ["Afghan", "Bangladeshi", "Bhutanese", "Indian", "Maldivian", "Nepalese", "Pakistani", "Sri Lankan", "other"] |

In each list, all labels except the last serve as positive labels (indicative of a course with South Asian content).

# 3 Results

The graph below displays predicted South Asian course counts from fall and spring semesters in each academic year from 2021/22 to 2024/25 given the category lists described above in section 2.3.



Furthermore, the following chart presents agreement between trials expressed as a percentage rounded to 3 decimal places. Two trials are said to agree if they both assign a given course a positive label (South Asian) or they both assign a given course a negative label (not South Asian). Agreement scores are calculated with the Cohen Kappa formula, which accounts for agreement that could occur by chance. For brevity, repeated values are excluded from the chart.

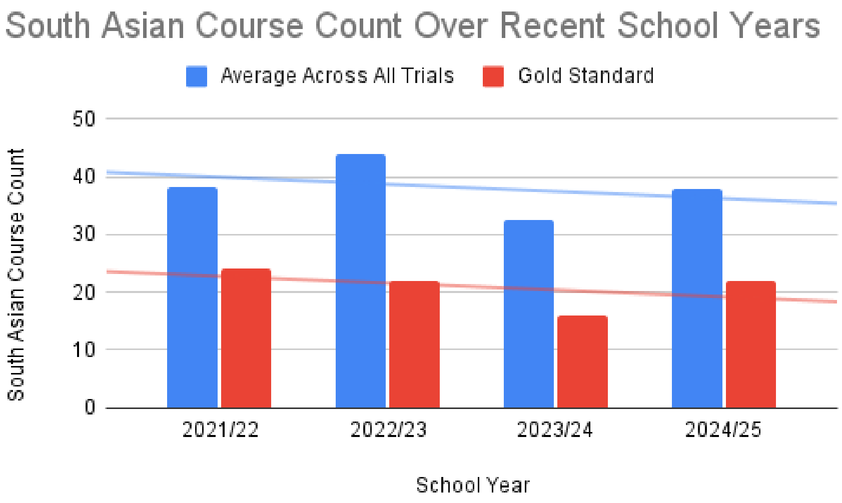
| | **Trial 1** | **Trial 2** | **Trial 3** | **Trial 4** | **Trial 5** |
|---|---|---|---|---|---|
| **Trial 1** | 1.000 | 0.378 | 0.393 | 0.058 | 0.068 |
| **Trial 2** | - | 1.000 | 0.464 | 0.139 | 0.095 |
| **Trial 3** | - | - | 1.000 | 0.156 | 0.183 |
| **Trial 4** | - | - | - | 1.000 | 0.225 |
| **Trial 5** | - | - | - | - | 1.000 |

Trials two and three hold the highest agreement score at 0.464, and trials one and four hold the lowest agreement score at 0.058. All trial combinations score above 0, indicating that agreement exceeds random chance.

However, given the vast differences among trials, I chose to manually annotate any course assigned at least one positive label across the five trials. This subset included 496 of the original 4451 courses (roughly 11%). Courses assigned negative labels across all five trials were automatically and equally assigned a negative label in my manual annotation. Given these gold standard labels, I calculate precision, recall, and F1 scores for each trial as shown below.

| | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|
| **Trial 1** | 0.322 | 0.702 | 0.442 |
| **Trial 2** | 0.260 | 0.988 | 0.412 |
| **Trial 3** | 0.418 | 0.607 | 0.495 |
| **Trial 4** | 0.204 | 0.273 | 0.234 |
| **Trial 5** | 0.640 | 0.190 | 0.294 |

Lastly, I aggregate trial averages and gold standard data to again graph South Asian course counts from fall and spring semesters in each academic year from 2021/22 to 2024/25. The overlaid lines indicate the line of best fit.



## 4 Discussion

From these experiments we can see that BART-large-mnli tends to overestimate the number of South Asian courses per year. For example, on average, the model predicted 43 South Asian courses for the 2022/23 school year when in actuality, there were only half that figure (22). We examine one instance of a false-positive below.

Given the ["South Asian", "not South Asian"] label list, the model incorrectly marks the following course as South Asian. For briefness, I truncate the course description to just the first couple of sentences.

> Phonetics I
> This course provides advanced instruction in phonetic analysis and experimental methodology. Students learn about various theories of speech perception, production, and cognitive representation...

Of the five trials, trial two (labels "South Asian topics" and "miscellaneous topics") counted highest number of South Asian courses of 79.75 on average across the four school years. By comparison, the gold standard annotation counted an average of 21. Despite various false-positives, trail two held nearly no false-negatives, achieving a nearly perfect recall score of 0.988.

Trials one, two, and three outperformed trials four and five in terms of F1 scores. Although trial five achieved the highest recall score of any trials (0.640), both trials four and five demonstrated low recall compared to former trials. Trials one, two, and three held F1 scores between 0.4 and 0.5 with trial three (labels "South Asian countries" and "miscellaneous") producing the highest F1 score at 0.495. On the other hand, trials four and five did not surpass an F1 score of 0.3.

Even with scattered prediction inaccuracies, the model accurately captures the downward trend exhibited in the gold standard annotation.

## 5 Conclusion

Given the data shown above, the South Asian Council at Cornell is right to feel that the number of South Asian courses has decreased in the last few school years. Both in BART-large-mnli zero-shot predictions and manual annotations, we see that the number of South Asian courses per school year has steadily declined from 2021/22. Optimistically, we see a rise in South Asian courses from the 2023/24 school year to the 2024/25 school year.

One minor limitation of this study is the quality of input data. Upon closer inspection, it appears that some formatting was lost in downloading course names and descriptions from the Cornell class roster. For example, certain punctuation like "–" failed to transfer and was instead replaced by odd characters. For example, in one description, we see the following mismatch between the original and downloaded text.

| **Original Text** | **Downloaded Text** |
|---|---|
| ...the end of the world – at least as we know it – is looming... | ...the end of the world â€“Â at least as we know it â€“Â is looming... |

These symbols could potentially confuse the model, leading to misinformed predictions. Improvements to this study could detect and correct or detect and discard poorly formatted course descriptions.

Future studies can experiment with different inputs to the BART-large-mnli model or with different model architectures entirely. For example, BART-large-mnli may make different predictions given only the course titles or course descriptions but not both.

On a broader scale, we can apply this methodology to other universities to answer questions like "Do Cornell University's peer institutions offer a similar representation of South Asian courses?" and "Is diminishing South Asian representation a widespread trend?"

This study provides insight into how we can utilize zero-shot learning large language models to evaluate quality and diversity of educational curricula. Although language models are far from perfect, we see in the results above that BART-large-mnli is successful in capturing the general trend of South Asian course representation displayed in gold standard annotations. With the help of large language models, we can work to more quantitatively identify under-representation in curricula and make adjustments accordingly.

# References

[1] Kamani, Ameya. "‘Any Person, Any Study’ Cannot Stop with South Asia." *The Cornell Daily Sun*, Cornell University, 24 Oct. 2024, cornellsun.com/2024/10/24/guest-room-any-person-any-study-cannot-stop-with-south-asia.

# A Humanities Departments

The following chart outlines selected humanities departments from the Cornell course roster. Humanities courses include any courses that fall under these departments and serve as the foundation of the dataset.

| Department Name | Department Abbreviation |
|---|---|
| Asian American Studies | AAS |
| American Indian and Indigenous Studies | AIIS |
| American Studies | AMST |
| Anthropology | ANTHR |
| Arabic | ARAB |
| Archaeology | ARKEO |
| Art | ART |
| History of Art | ARTH |
| Asian Studies | ASIAN |
| American Sign Language | ASL |
| Africana Studies and Research Center | ASRC |
| Classics | CLASS |
| Comparative Literature | COML |
| Communication | COMM |
| English | ENGL |
| Feminist, Gender and Sexuality Studies | FGSS |
| German Studies | GERST |
| History | HIST |
| Jewish Studies | JWST |
| Latin American Studies | LATA |
| Linguistics | LING |
| Medieval Studies | MEDVL |
| Music | MUSIC |
| Philosophy | PHIL |
| Religious Studies | RELST |
| Romance Studies | ROMS |
| Sociology | SOC |
| Spanish | SPAN |
| Science and Technology Studies | STS |
| Visual Studies | VISST |
| Writing Program | WRIT |

For a full list of all Cornell departments, please refer to the Cornell course roster.