

DATA REFLECTION:

Looking at the dashboard, a few results caught my eye. First, the share of international entries is about 48.57 percent, which feels high and makes me wonder if international applicants are more likely to post on GradCafe. Second, there was a high number of fall accepted applicants (6000+), specially when considering the ~20k dataset we had. This is probably due the entries in the GradCafe being organized from oldest to newest and fall 2025 is the most recent semester. This also indicates that there are thousands of entries per semester. Third, the average GPA for accepted Fall 2025 applicants is about 3.75, very close to the overall GPA card near 3.76, which suggests that GPA alone does not separate accepted from general posters in this sample. Finally, some school specific cards are thin. I only found one JHU MS in Computer Science entry under my filters, and zero for Georgetown CS PhD accepts in 2025. That says more about sparse reporting and labeling than about the programs.

The biggest struggle I faced was the language model sometimes hallucinating or mislabeling universities. I saw a clear Johns Hopkins row come back as a different school. To keep the analysis usable, I stored a raw university field in addition to the model generated one, then wrote the queries to match if either field indicated the school, checking the model first and falling back to raw when it looked wrong. The other issue is simple scarcity. Once I stack several filters together the sample collapses. Very specific slices like a particular school plus program plus degree plus year often have zero or one posts. Even at a large school like Johns Hopkins I only see a handful of entries per program, which means a single post can move the result a lot.

For questions nine and ten, my goal was curiosity. I wanted to see how the newer AI program at JHU is being reported compared with CS for master admits in the current year to date. I also wanted a fun comparison, so I looked at GPA buckets for Medicine versus Engineering. Obviously Engineering is better, just kidding, please be kind to our health professionals.

There are broader limits when analyzing anonymously submitted and self reported data. People with stronger outcomes are more likely to post, accepted students are more likely to share than rejected students, and fields are messy with abbreviations and missing values. Posting dates do not always match decision dates, and duplicates are hard to find without identifiers. This helps explain why GRE numbers in my cards run hot compared with public norms. CS and Engineering entries are over represented, higher scorers are more willing to share, and some people round up or skip weaker sections.

If I kept building this, I would harden normalization and grow the dataset. On the model side, I would try a larger hosted model or break the job into smaller prompts with guardrails. On the data side, I would ingest more pages, schedule regular refreshes, and keep the either model or raw logic in queries. With those changes and more data, the cards would feel sturdier while staying easy to run.

Issue with llm below:

School

Program

Added On

Decision

Sort

Johns Hopkins University

AI · Masters

September 15, 2025

Accepted on 2 Sep

...

Fall 2025

American

GPA 3.2

Online AI master

Lates

What

Grad

Local - gradcafe5432 | gradcafe_db

1 SELECT

2 url, date_added, program, llm_generated_program, llm_generated_university,

3 status, term, degree, gpa

4 FROM applicants

5 WHERE url = 'https://www.thegradcafe.com/result/986442';

6

PROBLEMS

PORTS

POSTGRESQL QUERY RESULTS

DEBUG CONSOLE

TERMINAL

Results

Messages

	url	date_added	program	llm_generated_program	llm_generated_university	status	term	degree	gpa
1	https://www.thegradcafe.com/result/986442	2025-09-15	AI	Artificial Intelligence	University of Pisa	Accepted	Fall 2025	Masters	3.2

Running site:

