Limitations GradCafe Analysis

The primary limitation of using GradCafe as a data source is that it is entirely self-reported and anonymous. Applicants voluntarily submit their own admissions outcomes, often without verification, which introduces significant bias into the dataset. People are more likely to report extreme outcomes such as acceptances, rejections, or unusually high test scores, while average or uneventful applications may go unreported. Additionally, fields such as GPA, GRE scores, and decision status are often embedded within free-form comments rather than structured input forms. This makes automated extraction difficult and increases the risk of missing or misinterpreting important information. The dataset therefore cannot be considered a representative or statistically rigorous sample of the overall applicant population.

A second major limitation is inconsistency and ambiguity in how users describe their programs, universities, and terms. Applicants frequently use abbreviations, nicknames, or informal language (e.g., “JHU” instead of “Johns Hopkins University”), which requires significant cleaning and normalization. Even with local language-model processing, some program names remain ambiguous or incorrect. Furthermore, scraping only the most recent pages creates a strong selection bias toward the newest submissions, making year-to-year comparisons unreliable. These issues highlight that while GradCafe data can be useful for exploratory analysis, any conclusions drawn from it should be treated as approximate and descriptive rather than definitive.