Module 3 Limitations Short Paper:

Anonymous, self-reported data is inherently noisy and incomplete. People do not follow a consistent format for program names, universities, dates, or scores, and many records are missing key fields (GPA, GRE, term). This creates bias in aggregates because the dataset is a non-random sample of applicants who chose to post. It also introduces ambiguity: the same school may be written in dozens of ways, and the same decision timeline can be expressed in different formats. Even with cleaning and LLM normalization, some errors remain, and the results should be interpreted as approximate trends rather than precise facts.

Some analytic responses can be surprising because small changes in cleaning rules or cohort definitions (for example, how "Fall 2026" is inferred) can materially change counts and averages. This is different from standard, institutional datasets where entries are validated and normalized at ingestion. Here, we must infer structure and meaning from semi-structured text, so the analysis reflects the quality of the input and the assumptions used to clean it. The takeaway is that anonymous data can still be useful, but it requires careful standardization and clear disclosure of its limitations.