Datasets Used in Our R&D (Product Based)

As part of our **product-based project**, we conducted early stage **R&D** to explore various publicly available datasets for essay evaluation. This research helped benchmark different models and techniques, but **none of these datasets were used in our final product**. The deployed version is entirely based on **LLaMA3-70B** and **prompt engineering**, with no dataset dependency.

1. HP Essays Dataset (Kaggle)

- **Link**: https://www.kaggle.com/c/asap-aes/data
- Used as a standard dataset in the AES (Automated Essay Scoring) domain.
- Consists of human-scored essays on multiple prompts with rubric-based scoring.
- Enabled LSTM and ML model evaluation in early testing.

2. Custom English Essay Dataset

- **Link**: https://drive.google.com/file/d/1E2PVYltjXNKxYwr7IrVgjh4FL4CaIWZy/view
- Internally curated to train and evaluate various models (BERT, GPT-2, LSTM, CNN, etc.).
- Designed for feedback generation, scoring accuracy, and rubric alignment experiments.

3. Web-Scraped Educational Dataset (Essay-Relevant)

Edulix University Data (Essay Mentions in Profiles)

- **Link**: https://github.com/joemanley201/universityRecommendationSystem/tree/master/scraper/univJSON
- Included references to student essays as part of academic profiles (indirectly related).
- Used to understand contextual use of essays in academic systems.