In our research, we adopted a systematic methodology to construct a medical AI model using hospital reports from a Turkish healthcare institution. We began by gathering reports from Turkish patients, emphasizing privacy by indexing and anonymizing sensitive patient data. To address language barriers, we translated the reports into English and enhanced their features for superior analysis.

To ascertain the quality of our dataset, we curated pertinent reports in accordance with the study's objectives, thereby achieving a balanced dataset. We had the assumption that patients with multiple reports would usually have their first report as emergency. To augment the quantity of emergency reports, we devised novel attributes for the reports, including the count of reports per patient, the character count of the report, and the approximate date of the report. Subsequently, we employed a self-hosted Label Studio (v1.9.0) platform for collaborative data annotation by experts. Following this, we meticulously cleaned and aligned annotations with patient reports, yielding a well-prepared dataset for subsequent machine learning pipelines. This refined dataset provides a robust foundation for the development of a precise and clinically significant medical AI model.

In the data ingestion phase, we utilized sqlalchemy (v2.0.33) to import the patient data into a PostgreSQL database. We indexed and anonymized sensitive patient data using a Fernet Key. All subsequent data transformations were executed within the database via dbt (v1.6.1), an Extract-Load-Transform (ELT) tool. Additionally, we leveraged the GPT 3.5 API to facilitate the translation of reports from Turkish to English. The prompt we utilized for this purpose was: "Perform the following transformation on the report: Translate into English".

These steps align with the principles and practices recommended in the medical AI field, where high-quality data is paramount for driving breakthroughs and advancing healthcare and medicine. The use of a self-hosted Label Studio platform for collaborative data annotation by experts is a common practice in medical AI research, allowing for the creation of balanced datasets. Furthermore, the use of sqlalchemy and dbt for data ingestion and transformation is a standard procedure in many companies working with large datasets.