Environment / Software

Setup:
We used Anaconda on a Windows machine with VS Code (Python 3.12) and GPU4EDU for faster processing.

Libraries:
Key libraries included PyTorch, TensorFlow, NumPy, Pandas, and Hugging Face's Transformers, allowing efficient data handling and model training.

Data

Initial data was sourced from Kaggle's PetFinder Adoption Prediction dataset (2.48 GB, 14,000 rows). We filtered the dataset to retain descriptions associated with adoption rates of 0, 1, and 2, thereby increasing the quality of retained descriptions (~7,000 rows). Chinese descriptions and non-essential columns were removed for consistency. From this refined dataset, two additional datasets were created:

- **English to Dutch Translation Dataset**: Generated by translating descriptions to Dutch using the T5 model.
- **Enhance Dataset**: Created by translating descriptions from English to French and then back to English to refine text quality.

Preprocessing: We split the data into 90% training and 10% validation.

Models

1. **Form to Description**: Converts simplified inputs into detailed descriptions. Trained for 30 epochs, batch size 256, learning rate 5e-4, completing in 3 hours.
2. **English to Dutch Translation**: Translates descriptions to Dutch for dataset variety. Trained on the refined PetFinder dataset (filtered for high-quality, English-only text), using 30 epochs, batch size 256, learning rate 5e-4, finishing in 3.5 hours.
3. **Description Enhancement**: Enhances quality by translating English to French and back to English. Trained for 36 epochs, batch size 256, learning rate 5e-4, requiring 9 hours.

Each model used the T5 Transformer, trained with consistent configurations, and saved based on validation loss to ensure optimized performance across tasks.