Validation and Cleaning Report of Synthetic Dataset for Viking LLM 7B Fine-Tuning

This document summarizes the validation and cleaning procedures applied to multilingual (Finnish and Swedish) synthetic instruction–response dataset intended for fine-tuning the Viking LLM 7B model. The dataset consists of 13 batches, each containing 250 Finnish and 250 Swedish samples, for a total of 3,250 examples per language.

1. Schema Validation

- Each sample was checked for valid JSON formatting.
- Verified presence of mandatory fields:
  - "instruction": a user prompt.
  - "response": a relevant and coherent model-generated answer.
- Invalid or malformed entries were removed.

2. Duplicate Removal

- Samples were deduplicated using SHA-256 hashing of the combined instruction and response.
- Only the first occurrence of each unique pair was retained.
- Removed exact duplicates across and within batches.

3. Content Quality Filtering

- Applied rules to remove:
  - Instructions shorter than 5 words.
  - Responses shorter than 15 words or longer than 300 words.
- Ensured proper formatting and coherent text.

4. Language Verification (Deferred)

- Language detection was initially planned to use `langdetect`, but the module was not available in the environment.
- Manual review or offline language validation is recommended for final release.

5. Output and Results

- Successfully processed Batches due to availability.
- Cleaned and validated samples were split back into their original batches.
- Each validated batch now contains exactly 250 unique, quality-filtered Finnish/Swedish samples.