kaggle

Kaggle Winner: WSDM Cup - Multilingual Chatbot Arena

Maksym Zhuravinskyi, Michael Pieler

Agenda

Agenda

1. Background
2. Summary
3. Feature selection & engineering
4. Training methods
5. Important findings
6. Simple model

Background

Background

- We worked together as Machine Learning Engineers at Stability.AI.
- Michael has a PhD from a Max Planck Institute and Maksym has a Bachelor's degree in CS.
- Maksym competed in the last LMSYS competition and placed ~30th.
- Michael competed in some older kaggle competitions.

Background

- “WSDM Cup - Multilingual Chatbot Arena: Predict human preference across multiple languages from real votes in the Chatbot Arena.”
- Predict which responses users will prefer in a head-to-head battle between chatbots powered by LLMs.
- Code competition → inference via kaggle notebooks (2x T4)!
- Competition was selected for the WSDM Cup 2025 @ 18th ACM International Conference on Web Search and Data Mining

Summary

- We used pretrained LLMs and fine-tuned them on a mixture of different datasets.
- For most of the work we used the HF Transformers library.
- It took 8 hours to train each teacher (5 teachers total) and 7 hours to train a student.
  Training gemma-2-9b-it took 5 hours.

Features Selection/ Engineering

Data

- For training we used external data released by LMSYS/lmarena, and other open sources synthetic datasets:
  - mlabonne/orpo-dpo-mix-40k (40k)
  - opencsg/UltraFeedback-chinese (50k)
  - lmarena-ai/arena-human-preference-55k (57k)
  - lmsys/chatbot_arena_conversations (33k)
  - lmarena-ai/PPE-Human-Preference-V1 (16k) (despite it being an evaluation dataset and we even initially suspected that this is the LB due to a very close correlation, we still included it last minute fearing others would)
  - lmarena-ai/Llama-3-70b-battles (1k)
  - lmarena-ai/gpt-4o-mini_battles (1k)
  - Datasets from @nbroad 🤗 (v1, v2 and v3) (25k)

Training Methods

High-level setup

- We use a default HF Transformers `ForSequenceClassification` model setup with two output classes for binary classification.

- Our solution is a merge of the distillation approach from @sayoulala and the inference method from @tascj0, the 1st and the 2nd solutions from the last competition, but with a much larger base model, i.e., Qwen2.5-14B-Instruct.

- Prompt template (details see code):

```
<BOS><start_of_turn>user
prompt<end_of_turn>
<start_of_turn>model
completion a<end_of_turn>
<start_of_turn>assistant
completion b<end_of_turn>
<EOS>
```

- Prompt template example:

```
<BOS><start_of_turn>user
Which is heavier? 1 kg of cotton or 1
pound of steel. Just write one word answer
without any explanation.<end_of_turn>
<start_of_turn>model
Steel<end_of_turn>
<start_of_turn>assistant
Heavier<end_of_turn>
<EOS>
```

Training

- Teachers training:
  - CE with two classes (excl. ties)

- Student training:
  - Distillation with CE against the original label if present, otherwise a hard label, and KL and Cosine loss on soft labels
  - CE loss was weighed down (0.25) while other losses had coefficient 1.0.

Merging and inference

- The final model is a linear merge of two:
  - One trained on the full data and one excluding the last two datasets.

- For inference we used two models for two passes:
  - First, using the merged model on all samples.
  - Second, using the original model (trained on the complete data) with the response order swapped on only 33% of the samples with the most uncertain predictions.

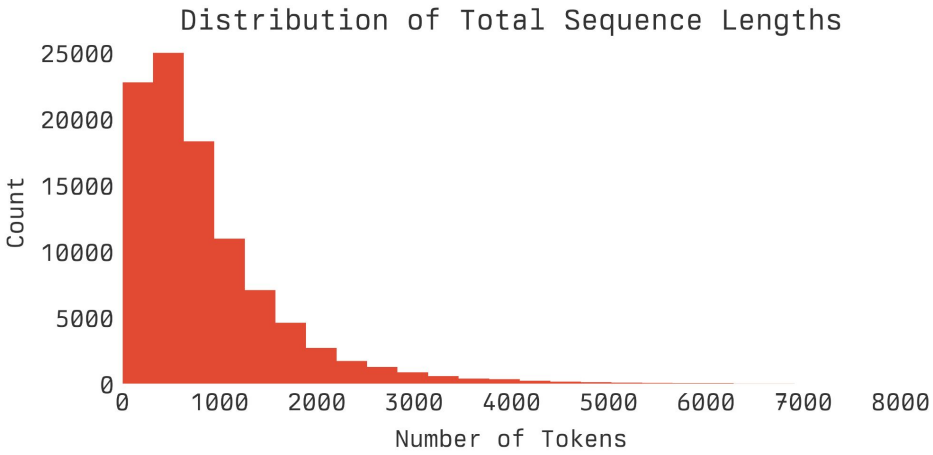
Important and Interesting Findings

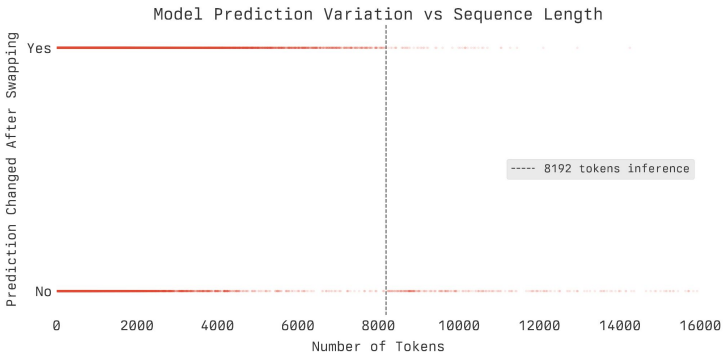
Important and Interesting Findings

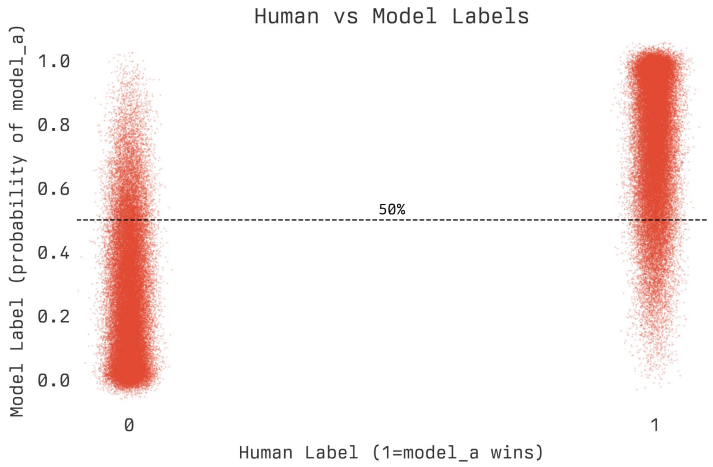
- The most important part is distillation from a larger teacher (as was done in the first place last competition).

- We added some improvements:
  - Adding model names before the responses in the input format.
  - Manually relabeled ~100 of samples.
  - Model merging.
  - Adaptive selection of samples for test-time augmentation (TTA).

Important and Interesting Findings







Simple Model

Simple Model

- Training gemma-2-9b-it on the competition data and previous lmsys dataset should get approximately 98% of our LB score.

Solution Overview

In 3 min or less, can you provide an overview of your participation in this competition?

- Key insights: Execute well on previously outlined recipes, with a focus on distillation and inference!
- Biggest advantage for us was combining data curation, model training, and optimized inference!
- Most fun was learning and trying out new setups! 😀

Leaderboard

#	Team	Members	Score	Entries
1	**whitefebruary**		0.716092	2
2	zhudong1949		0.716047	2
3	PLaMo 1000000B		0.714967	2
4	HKUST-GZ DSA KIMI Lab		0.714067	2
5	Just doing		0.712627	2
6	Oh Your Mercy Never Fails		0.712582	2
7	Team Turing		0.712222	2
8	sayoulala		0.712042	2
9	quenn		0.711817	2
10	tascj		0.711817	2

You won a gold medal!
Your team placed 1st out of 950 teams.

kaggle

Question and Answer

kaggle

Links

- Competition: https://www.kaggle.com/competitions/wsdm-cup-multilingual-chatbot-arena/

- Solution summary: https://www.kaggle.com/competitions/wsdm-cup-multilingual-chatbot-arena/discussion/569902

- Code: https://github.com/maxreciprocate/kaggle-lmarena-1st-place