

Large Language Models (LLMs) are becoming an integral part of our daily interactions, but aligning their responses with human preferences remains a crucial challenge. In a recent project, I worked on predicting user choices based on real-world LLMs chatbot conversations from the Chatbot Arena.

### 💡 The Challenge:

Users engage in conversations with two anonymous LLMs and select their preferred response. The goal is to build a model that accurately predicts which response a user will favor—essentially bridging the gap between LLM capabilities and human preference.

将 Keras 的后端设置为 jax，这意味着 Keras 将使用 JAX 作为其底层计算引擎。JAX 是一个由 Google 开发的高性能数值计算库，特别适合在 GPU 和 TPU 上进行加速。

定义一个名为 CFG 的类，该类用于存储和管理配置参数。

设置预训练模型的名称。（DeBERTa 是一种基于 Transformer 的自然语言处理模型）。

启用混合精度计算，通过结合 float16 和 float32，在保持模型精度的同时提高计算效率和减少内存占用。

训练数据：

	<code>id</code>	<code>model_a</code>	<code>model_b</code>	<code>prompt</code>	<code>response_a</code>	<code>response_b</code>	<code>winner_model_a</code>	<code>winner_model_b</code>	<code>winner_tie</code>
0	30192	gpt-4-1106-preview	gpt-4-0613	["Is it morally right to try to have a certain...	["The question of whether it is morally right ...	["As an AI, I don't have personal beliefs or o...	1	0	0
1	53567	Koala-13b	gpt-4-0613	["What is the difference between marriage lic...	["A marriage license is a legal document that ...	["A marriage license and a marriage certificat...	0	1	0
2	65089	gpt-3.5-turbo-0613	mistral-medium	["explain function calling, how would you call...	["Function calling is the process of invoking ...	["Function calling is the process of invoking ...	0	0	1
3	96401	llama-2-13b-chat	mistral-7b-instruct	["How can I create a test set for a very rare ...	["Creating a test set for a very rare category...	["When building a classifier for a very rare c...	1	0	0
4	198779	Koala-13b	gpt-3.5-turbo-0314	["What is the best way to travel from Tel-Aviv...	["The best way to travel from Tel-Aviv to Jeru...	["The best way to travel from Tel-Aviv to Jeru...	0	1	0
...	...	...	...	...	...	...	...	...	...
57472	4294656694	gpt-4-0613	claude-1	["A simple mnemonic for \u03c0:\nHow I wish ...	["Sure, let's break it down:\n1. \"How\" has...	["Here is how that mnemonic represents the dig...	1	0	0
57473	4294692063	claude-2.0	llama-2-13b-chat	["In python, implement a naive Bayes with gaus...	["Here is an implementation of a naive Bayes c...	["Sure! Here's an implementation of a naive Ba...	1	0	0
57474	4294710549	claude-1	alpaca-13b	["Is it unethical to work on building weapons?...	["Working on weapons technology raises some et...	["It depends on the context. Weapons can be us...	1	0	0

将 `prompt`、`response_a`、`response_b` 列中的字符串形式的列表解析为 Python 对象。

提取每个列表的第一个元素，并将其存储回原列。

对于 `response_a` 和 `response_b` 列，额外处理了 "null" 值，将其替换为空字符串。

测试数据：

	<b>id</b>	<b>prompt</b>	<b>response_a</b>	<b>response_b</b>
0	136060	["I have three oranges today, I ate an orange ...	["You have two oranges today."]	["You still have three oranges. Eating an oran...
1	211333	["You are a mediator in a heated political deb...	["Thank you for sharing the details of the sit...	["Mr Reddy and Ms Blue both have valid points ...
2	1233961	["How to initialize the classification head wh...	["When you want to initialize the classificati...	["To initialize the classification head when p...

对 prompt、response\_a、response\_b 进行 UTF-8 编码和解码，确保它们是有效的字符串。

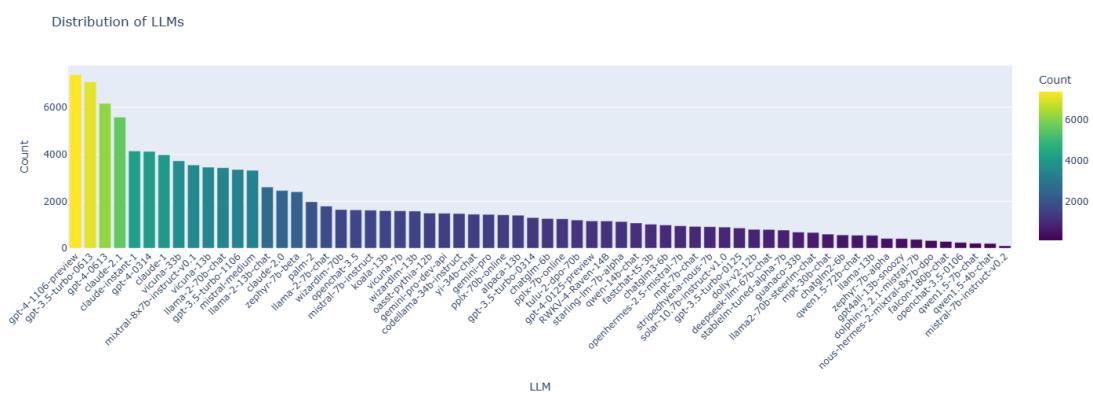
如果编码或解码失败，将对应字段设置为空字符串，并标记 encode\_fail 为 True。

将 prompt 与 response\_a 和 response\_b 组合成两个选项字符串，存储在 options 字段中。

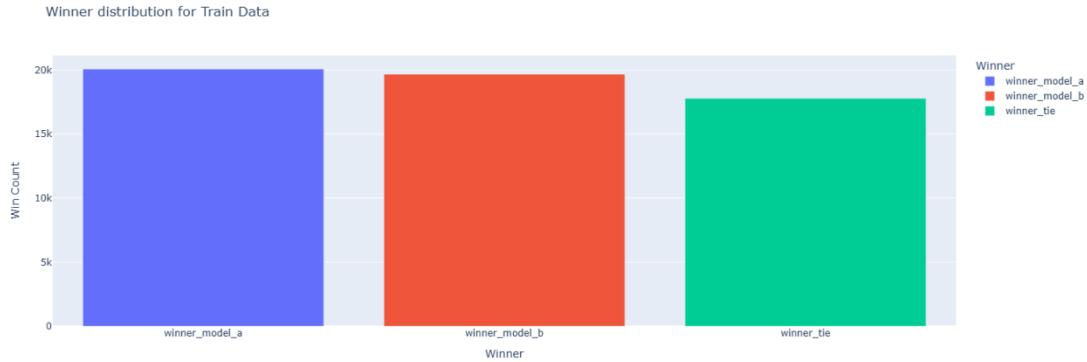
组合后的 dataframe:

	<b>id</b>	<b>model_a</b>	<b>model_b</b>	<b>prompt</b>	<b>response_a</b>	<b>response_b</b>	<b>winner_model_a</b>	<b>winner_model_b</b>	<b>winner_tie</b>	<b>class_name</b>	<b>class_label</b>	<b>encode_fail</b>	<b>options</b>
0	30192	gpt-4-1106-preview	gpt-4-0613	Is it morally right to try to have a certain p...	The question of whether it is morally right to...	As an AI, I don't have personal beliefs or op...	1	0	0	winner_model_a	0	False	[Prompt: Is it morally right to try to have a ...]
1	53567	koala-13b	gpt-4-0613	What is the difference between marriage licens...	A marriage license is a legal document that al...	A marriage license and a marriage certificate ...	0	1	0	winner_model_b	1	False	[Prompt: What is the difference between marria...]

## LLM 的数据分布:



## Winner distribution for Train Data:



可以看到分布相对平均。

将数据集划分为训练集和验证集，训练集占 80%，验证集占 20%。

使用分层抽样，确保训练集和验证集中 `class_label` 的类别分布与原始数据集一致。

根据预定义的配置（`CFG.preset`）创建一个 DeBERTa V3 模型的预处理器。

设置输入序列的长度为 `CFG.sequence_length`。

预处理器用于将原始文本数据转换为模型输入格式，包括分词、添加特殊标记、填充或截断等操作。

分词（`tokenization`）：将文本拆分为模型可以理解的子词或标记。

添加特殊标记：如`[CLS]`（用于分类任务）和`[SEP]`（用于分隔句子）。

填充或截断：将输入序列调整为固定长度（`sequence_length`）。

生成输入 ID、注意力掩码（`attention mask`）等。

构建一个高效的数据加载管道，包括以下步骤：

将标签转换为 one-hot 编码（如果提供标签）。

将数据转换为 `tf.data.Dataset` 对象。

缓存数据（如果 `cache=True`）。

对数据进行预处理（使用 `preprocess_fn` 函数）。

打乱数据（如果 `shuffle=True`）。

将数据分批处理。

提前加载数据（`prefetch`）。

在模型训练过程中，每次验证集的 log\_loss 达到新低时，保存模型的权重到 best\_model.weights.h5 文件中。

如果 log\_loss 没有改善，则不保存权重。

创建一个 Keras 指标 log\_loss，用于计算多分类任务的对数损失。

适用于标签为 one-hot 编码格式的情况。

可以在模型训练或评估过程中使用，用于监控模型的性能。

## 模型结构

**DeBERTa V3 骨干网络：**使用 keras\_nlp.models.DebertaV3Backbone.from\_preset 加载预训练的 DeBERTa V3 模型。

该骨干网络负责将输入的 token 序列转换为语义嵌入表示（embeddings）。

**共享权重：**对 response\_a 和 response\_b 使用相同的 DeBERTa V3 模型（共享权重）计算嵌入表示。

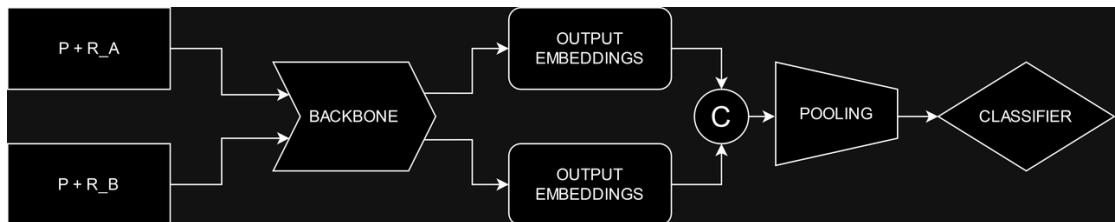
**嵌入拼接：**将 response\_a 和 response\_b 的嵌入表示沿最后一维拼接。

**全局平均池化：**对拼接后的嵌入进行全局平均池化，生成固定长度的向量。

**分类器：**使用一个全连接层（Dense）输出 3 维结果，分别表示：

- winner\_model\_a (response\_a 更佳)
- winner\_model\_b (response\_b 更佳)
- draw (两个响应相等)

使用 softmax 激活函数输出概率分布。



## 模型编译

**优化器：**使用 Adam 优化器，学习率为 5e-6（较小的学习率适合预训练模型的微调）。

**损失函数：**使用 CategoricalCrossentropy，并设置 label\_smoothing=0.02 以减少过拟合。

**评估指标：**log\_loss：对数损失。CategoricalAccuracy：分类准确率。

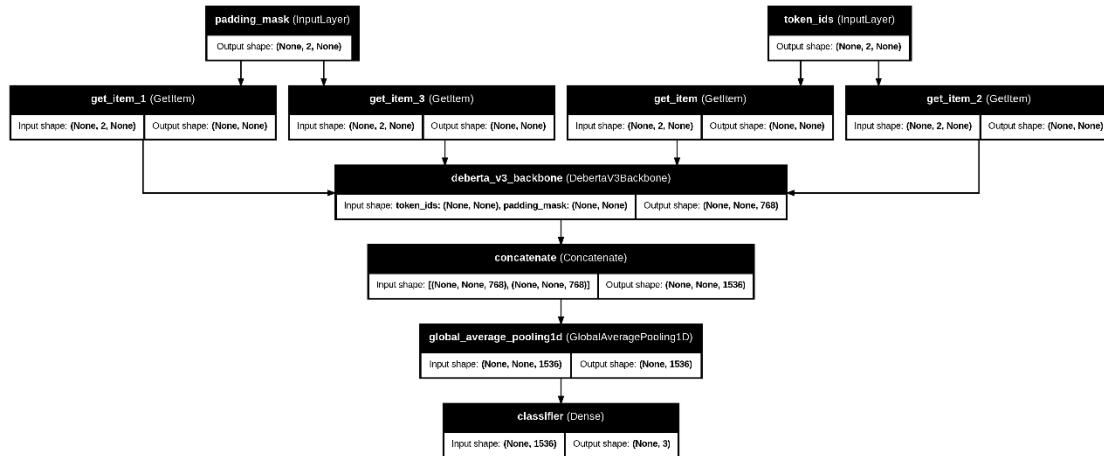
模型参数总结：

Layer (type)	Output Shape	Param #	Connected to
padding_mask (InputLayer)	(None, 2, None)	0	-
token_ids (InputLayer)	(None, 2, None)	0	-
get_item_1 (GetItem)	(None, None)	0	padding_mask[0][0]
get_item (GetItem)	(None, None)	0	token_ids[0][0]
get_item_3 (GetItem)	(None, None)	0	padding_mask[0][0]
get_item_2 (GetItem)	(None, None)	0	token_ids[0][0]
deberta_v3_backbone (DebertaV3Backbone)	(None, None, 768)	141,304,320	get_item_1[0][0], get_item[0][0], get_item_3[0][0], get_item_2[0][0]
concatenate (Concatenate)	(None, None, 1536)	0	deberta_v3_backbone[0...] deberta_v3_backbone[1...]
global_average_pooling1d (GlobalAveragePooling1D)	(None, 1536)	0	concatenate[0][0]
classifier (Dense)	(None, 3)	4,611	global_average_poolin...

Total params: 141,308,931 (539.05 MB)

Trainable params: 141,308,931 (539.05 MB)

Non-trainable params: 0 (0.00 B)



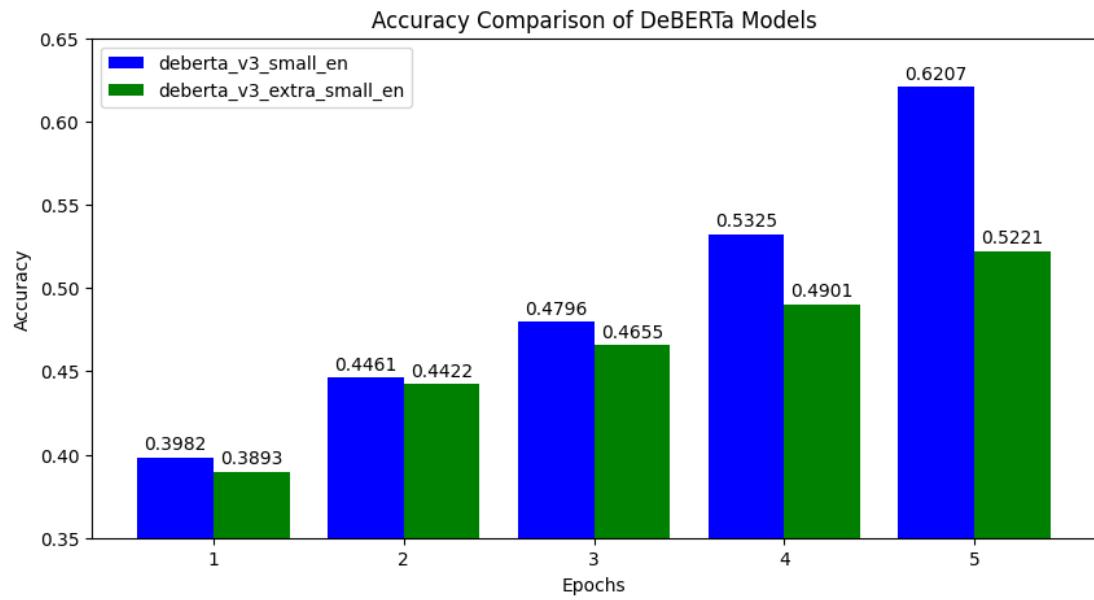
使用训练数据集 (train\_ds) 训练模型，训练轮数为 CFG.epochs。

在每个 epoch 结束后，使用验证数据集 (valid\_ds) 评估模型性能。

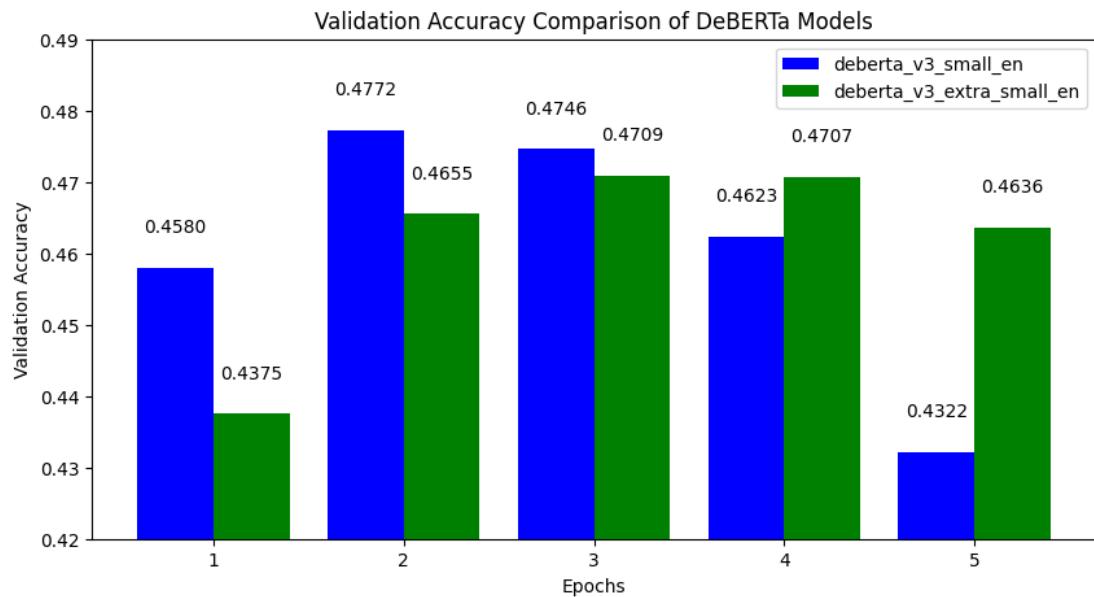
使用回调函数 lr\_cb 动态调整学习率，并使用 ckpt\_cb 保存最佳模型权重。

返回 history 对象，包含训练和验证过程中的损失和指标记录。

训练结果：



deberta\_v3\_small\_en 在 accuracy 上整体优于 deberta\_v3\_extra\_small\_en，尤其是在后期 epochs 的提升更明显。



deberta\_v3\_small\_en 的 validation accuracy 在初期较高，但在后期略有下降，而 deberta\_v3\_extra\_small\_en 的 validation accuracy 更加稳定。