**大規模言語モデル講座 最終課題 LLMコンペティション**

## 

## **概要:**

LLM講座 コンペティション

## **修了:**

修了要件の1つとして含まれる

## **上位入賞者:**

優秀者の表彰を行う可能性あり

## **開催時期**

2023/09/25 - 2023/10/10 23:59（予定）

## **計算リソースについて**

## **コンペ概要**

下記の３種類のベンチマーク性能を高めよ。ただし、事前学習、FinetuningやRLHF等、プロンプティングの工夫など、授業で学んだことを自由に利用して構わない。モデルや学習に利用できるデータについてはルールを参照のこと。

* Type１：日本語QA
  + 日本に関する知識を答える選択肢問題
  + 出力は選択肢番号(1~5)を出力してください。(int型)
  + 評価は正解率で行います。
* Type２：文章要約
  + 出力は要約文を出力してください。(str型)
  + 評価はROUGE-2 score(F1 score)で行います。
* Type３：Instruction Following
  + 与えられた指示に対して適切な出力を返してください。
  + 出力は生成文を出力してください。(str型)
  + 評価は人手 or 外部の大規模言語モデルを用いて行います。
  + ※本評価に関しては、Type1・Type2ベンチマークでの成績上位者を対象に行います。
    - Type1・Type2成績上位者のうち、Type3のコンペに参加希望の方には人手による評価に協力いただきます

締切は2023/10/10 23:59までです。提出の際は以下のファイルをOmnicampusへ提出してもらいます。

* 提出物１：学習・推論コード(train\_and\_predict.ipynb) Omnicampusの宿題competition\_codeへ提出してください
  + 学習データを作成、読み込んで学習させるコード、評価データを読み込み推論結果のjsonファイルを出力するコード
  + 結果が再現されるか確認に使用します。
  + 成績優秀者の提出コードはコンペ終了後、他の受講生に対して公開させていただきます。
  + starter\_code.ipynbを参考にしてください。
* 提出物２：推論結果(submission.json) Omnicampusの宿題competitionへ提出してください
  + id, task\_type, text, answerのkeyがあることを確認してください(正しく入っていない場合、スコア付けが行われません)
* 提出物3: コンペの工夫についてのレポート(submission\_report.json, 想定所要時間15-30分) Omnicampusの宿題compretition\_reportへ提出してください
  + コンペへの最終submitが終わったらご自身が取り組まれた工夫や、苦労した点について簡単にレポートを提出していただきます
  + 最終課題で選んだモデルについてや工夫点、苦労した点について簡単にレポートを提出していただきたいと思います  
    コンペへ最終的な提出が終わった方は以下の例を参考に項目を埋めていただき提出していただきます  
    <https://colab.research.google.com/drive/1bvzg6mDZI_aMEbOt41mdiwcwRqIUgvIh?usp=drive_link>
  + 提出先 <https://edu.omnicamp.us/courses/34/assignments/266/>
  + 締切はコンペの締切の1日後10/11 23:59に設定しております

なお、課題の最終的な提出はOmnicumus上で行います。なお、課題に取り組むに当たり簡単なベンチマークコードを提供する予定です。

## **ルール**

### ■ 計算リソース等について

* 配布したGPUリソース(50GPU時間以内)で学習・推論を終えてください
  + つまり1GPUのノードであれば最大50時間、2GPUのノードであれば25時間、といったように利用するGPU数によってインスタンスを利用可能な時間が異なります。
  + 1GPU50時間、2GPU25時間、4GPU12.5時間、8GPU6.25時間など、どのようにGPUリソースを使っても構いません。
  + なお、GPU時間には最終的な提出物の推論時間を含みます。推論のためのGPU時間を確保しておいてください。  
    ※軽いモデルや軽いデータで学習を終えて、1度提出の流れまでは確認することを推奨します。
* ご自身で用意したGPUリソースで学習・試行錯誤をすることは可能ですが、提出する結果の元となるモデルは、配布したGPUリソースで学習を終えてください。
  + 簡単な動作確認などはgoogle colabなどを活用し、大規模な学習をomnicampus上で行うことをオススメします。
* workspace/assetsというディレクトリ以下は、インスタンスを閉じた後、再度立ち上げても作業中のファイルやダウンロードや学習させたモデルなどが残るようになっております

### ■ 利用データおよびモデルについて

* 使うデータ・モデルは基本的に自由ですが、以下の条件を満たしてください
* データについて
  + 学習で使うデータの中に、日本語ベンチマーク(JGLUEなど研究機関が作成したデータセット)のval/testのデータが含まれないようにしてください。trainデータの利用は認めます。
    - 最低限JGLUE(<https://github.com/yahoojapan/JGLUE> )のval/test、xl\_sum(<https://huggingface.co/datasets/csebuetnlp/xlsum> )のval/test、<https://huggingface.co/datasets/elyza/ELYZA-tasks-100> のtestが含まれていないかの確認はお願いします。
    - [日本語ベンチマークの例](https://drive.google.com/file/d/1kbXIIFp4dgqO0frXoWu9B2L05Y99PIKR/view)(p4, 評価用データ)
  + ご自身で新しくデータセットを作成して学習に利用しても問題ありません。
  + 配布するsample.jsonのデータを学習に使用しても問題ありません。
* モデルについて
  + 学習・推論に使用するモデルは、コンペ開始時点(2023/09/25)で公開されているモデルのみ使用を認めます。
    - ただし2023/09/28にPFN社から公開されたPlamo-13B <https://huggingface.co/pfnet/plamo-13b> の使用は例外として認めます
  + 他者が学習させたモデルをベースに学習を行う場合には、そのモデルにベンチマークのデータセットが含まれていないかも確認してください。
  + GPT-4やPaLMなど、モデルの重みを公開していないモデルは推論には使えません。データ作成においては使っても構いません。(利用規約はご自身でご確認ください)
  + 利用可能なモデルの例(あくまでも一例なのでここに記載がないモデルも使用可能です、ベンチマークのデータセットを学習データに含んでいないかの確認をお願いします)
    - Falcon
      * <https://huggingface.co/tiiuae>
    - LLaMA or Llama 2
      * <https://huggingface.co/meta-llama>
      * <https://github.com/facebookresearch/llama>
    - OPT
      * <https://github.com/facebookresearch/metaseq/tree/main/projects/OPT>
    - Bloom
      * <https://huggingface.co/bigscience/bloom>
    - GPT Neo, J, NeoX, Pythia
      * <https://huggingface.co/collections/EleutherAI/pythia-64fb5dfa8c21ebb3db7ad2e1>
    - T5
      * <https://huggingface.co/collections/google/flan-t5-release-65005c39e3201fff885e22fb>
    - rinna社のモデル
      * <https://huggingface.co/rinna/bilingual-gpt-neox-4b> など
    - cyberagent社のモデル
      * <https://huggingface.co/cyberagent/open-calm-7b> など
    - Stability AI Japan社のモデル
      * <https://huggingface.co/stabilityai/japanese-stablelm-instruct-alpha-7b> など
    - ELYZA社のモデル
      * <https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b> など
    - Weblab-10B
      * <https://huggingface.co/matsuo-lab>
    - 有名なモデルとその性能について
      * <https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard>
      * <https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable>
      * <https://note.com/elyza/n/na405acaca130>

### ■ テストデータとLeaderboardについて

* テストデータはtest.jsonというデータをコンペ開始とともに配布します、そのデータにType1, Type2, Type3のデータが全て含まれています
  + 注意: テストデータはコンペ期間中に追加される可能性があります、追加した際はslackにてアナウンスを行うのでご確認ください。追加する場合は10/01までにアナウンスします。推論用のGPUリソースを残しておいてください。
* Leaderboardのスコアは、テストデータ全件についての評価値ではありません、コンペ終了時に全件の評価を行いType1, Type2の成績上位者を決定します
  + スコアの計算式は以下のように計算されています
    - (Type1のスコア / ベースラインのType1のスコア) + (Type2のスコア / ベースラインのType2のスコア)

### ■ 認められている工夫について

* 事前学習の工夫のみでも、プロンプティングの工夫のみでも、finetuningの工夫のみでも、また全ての工夫を組み合わせた工夫でも構いません。
* 量子化などを組み合わせても問題ありません。
* 他者が公開している学習用データセットを使用しても問題ありませんが、それらのデータに上記の日本語ベンチマークのval/testデータが含まれていないことを確認してください。結果を再現するために学習データも提出してもらいます、その際にval/testデータが含まれているなどの問題があれば失格とします。
  + 最低限JGLUE(<https://github.com/yahoojapan/JGLUE> )のval/test、xl\_sum(<https://huggingface.co/datasets/csebuetnlp/xlsum> )のval/test、<https://huggingface.co/datasets/elyza/ELYZA-tasks-100> のtestが含まれていないかの確認はお願いします。
* プロンプティングの工夫の際に、配布するsamples.jsonの例を使っても構いません。
  + ご自身で準備したデモンストレーションデータをfew-shot promptingの際に用いても問題ありませんが、日本語ベンチマークデータセットのval/testがpromptに含まれるのは禁止します。
* Retrievalを活用した手法はコンペで使用して良い手法として認めます、ただしデータベースに日本語ベンチマークのval/testが含まれる場合はNGとなります(学習データやモデルのルールと同じような方針です)

### ■ 受講生間での議論について

* 生成した出力結果自体を共有するのはやめてください。
* コンペに関しての受講生同士のディスカッションは#コンペ\_受講者ディスカッションで行ってください。
* コンペのルールに関してご不明な点があれば#コンペ\_運営へのルール確認で質問してください。

## **その他注意事項**

* 成績上位者は学習コードと、学習データを提出してもらい、50GPU時間以内で学習・推論が終わることを確認します。その際に著しく結果が再現されない・val/testデータが含まれているなどの問題があれば失格とします。
  + 再現性について
    - 提出いただくipynbファイルをこちら側で上から動かして、提出いただいたsubmission.jsonが概ね再現される、というのが再現性の確保でお願いしたいこととなります
    - from transformers import set\_seed
    - set\_seed(42)
    - このように設定すればdecoding手法でsamplingの手法を採用しても再現性が取れるかと思います、ご自身の手法でどこが確率的な振る舞いになるか、それをどのように再現性を保てるか、という観点でなるべく再現ができるようにしていただければと思います
    - 以下参考にしていただければと思います
    - https://huggingface.co/blog/how-to-generate#sampling
    - https://huggingface.co/docs/diffusers/using-diffusers/reproducibility
    - https://pytorch.org/docs/stable/notes/randomness.html

## QA

* 学習コードにてHugging Faceのプライベートレポジトリのデータセットを利用している場合、他の人はレポジトリにアクセスできないので学習が再現できないのですが、どのようにデータセットへのアクセスを確保すればいいでしょうか？
  + プライベートレポジトリの使用は想定していないです、パブリックにしていただくか、そのデータセットが再現されるスクリプトを提出するファイル内に含めて頂ければと思います
  + スクリプトと併せて再現確認の際に、作成していただいたデータセットをそのまま提出してもらうことになるかも知れません
* LlamaIndex等のEmbedding時間は学習時間に含まれますか？
  + ルールでこのように示している通り、50GPU時間以内で学習・推論が終わることを確認します。
  + 回答に必要なembeddingを取得、retrievalする時間も推論という処理に必要として、50GPU時間以内で終わるべき処理として考えております
  + 提出いただいたipynbを上から順に動かして50GPU時間以内で終わる、ということを意識していただければと思います
* Jupyterノートブックを使わないで学習や提出物の作成をしている場合は提出する ipynb 形式の提出ができないと思いますが、ノートブック上に学習から推論のコードを載せて実行することが前提でしょうか？ノートブックだと処理が煩雑になってしまってコードをいくつかのファイルに分けたりしたいのですが、その場合はどうすればいいでしょうか？
  + 提出するipynbには
  + !python scripts/train.py
  + !python scripts/eval.py
  + のようにterminal上で実行しているようなコマンドを整備して上からセルを動かしたら学習から推論を再現できるようにしておいてください
  + そしてipynbファイルの最後のセルに
  + scripts/train.pyの中身
  + scripts/eval.pyの中身
  + みたいなセルを準備していただき、そちらにコードをコピペして提出していただければと思います
  + その後成績上位者の場合は、提出いただいた実行に必要なコードを全て提出していただければと思います
  + なので結論
  + 上から実行したら結果が再現されるipynbファイルを提出してください、最後にコードのコピペも貼っておいてください
  + (コンペ終了時成績上位者の場合)↑の実行(コピペ部分以前のセル)に必要なpyファイルを別途送っていただきます
  + でお願いできればと思います
  + コンペ終了後こちらから連絡を差し上げますが、速やかに実行に必要なファイルを提出いただかなかったら失格と見做すかもしれないので、運営からの連絡にはすぐ対応いただければと思います