Prize recipient solution documentation guide

**Congratulations!** You’ve gone up against dataheads from around the globe and emerged victorious! Laugh, dance, brush your shoulders off. You demonstrated serious skills, and helped make this world a better place in the process. Awesome job. Now you’ve finished in one of the top spots of the private leaderboard, which makes you eligible to receive a monetary prize. You’re almost there.

In accordance with the official competition rules, the DrivenData terms of use, and applicable State and Federal law, we both have some due diligence to take care of before we can announce winners and disburse prizes.

You will receive a separate document to submit your legal documentation so that we can verify your legal identity. We will use this to verify your eligibility to participate and then review the specific laws and rules about giving out prizes based upon your nationality and our tax reporting obligations.

This document details the steps required to submit your solution materials.

1. **Code submission and result reproducibility**  
   You package up and send us your code, documentation of dependencies, and any other assets you used. We review your package and make sure that it works and that we can fully reproduce the workflow from raw data to a submission comparable to your best submission. See the [Competition prize winner submission template](https://github.com/drivendataorg/prize-winner-template) to get started.
2. **Basic information for winner announcement**  
   Provide your preferred information for use in announcing the winners of the competition.
3. **Model documentation and write-up**  
   You write up answers to our questionnaire, providing important context and documentation so that the beneficiary and the community get the most out of your work.

Please read this document carefully. Each section details exactly what is needed from you—the faster we can check all the boxes for our mutual responsibilities, the faster we can disburse your prize!

Thanks for your hard work, and congratulations for making it this far.

Best,

The DrivenData Team

# I. Code submission and result reproducibility

You will need to submit a compressed archive of your code. You don’t need to include the raw data we provided, but everything else should be included and organized. If the files are too large to be e-mailed, a Google Drive or Dropbox share (or other comparable method of transferring data) works.

|  |
| --- |
| **Note: *please follow these instructions carefully*.** The spirit and purpose of the competition (and the reason for offering prizes) is to give our beneficiary organizations the best possible solution *along with working code they can actually use*. In accordance with the competition rules, if we can’t get your code working and reproduce your results with a reasonable effort, or if your entry is too disorganized to be practically usable, then your entry may be disqualified! |

The overall concept is to **set up this archive as if it were a finished open source project**, with clear instructions, dependencies and requirements identified, and code structured logically with an obvious point of entry. Check out the competition prize winner [README template](https://github.com/drivendataorg/prize-winner-template) to get started. We also have a [data science project template which may be helpful](http://drivendata.github.io/cookiecutter-data-science/).

At a minimum, your solution must include **an extremely clear README** that details all of the steps necessary to get to your submission from a fresh system with no dependencies (e.g., a brand new Linux, Mac OS X, or Windows installation depending on what environment you choose to develop under) and no other data aside from the raw data you downloaded from us.

This will probably entail the following:

* Necessary tools and requirements (e.g. “You must install Word2Vec 0.1c” or “Install the required Python packages in requirements.txt”).
  + **All requirements** **should be clearly documented**, for instance in either a requirements.txt file with versions specified or environment.yml file.
* The series of commands, in order, that would get a reasonably experienced and savvy user from your code to a finished submission.
  + **Ideally, you will have a main point of entry to your code** such as an executable script that runs all steps of the pipeline in a deterministic fashion. A well-constructed Jupyter notebook or R script meets this standard.
  + **The next best thing is a list of specific, manual steps** outlining what to do. For example, “First, open Word2Vec and set these parameters. [...] Take the output file and run the script src/make\_preds.R with the following parameters [...]”. *(The limitations of this approach should be clear to any experienced data scientist!)*
* **Make sure to provide access to all trained model weights necessary to generate predictions from new data samples** without needing to retrain your model from scratch. Note that model weights can be contained in your archive or shared via a cloud storage service. The solution should provide clear instructions to perform inference on a new data point, whether or not it is included in the test set.
* Any other instructions necessary to end up with your winning submission file (or comparable — we understand that certain parts of model fitting are stochastic and won’t result in exactly the same parameters every time).

II. Basic information for winner announcement  
Please provide your preferred information for use in announcing the winners of the competition.

* Name (first and last name or first name and last initial): Yang Xu
* Hometown: LiShui, ZheJiang, China
* A recent picture of yourself or digital avatar (feel free to attach separately):
* Social handle or URL (optional): https://www.kaggle.com/steamedsheep

III. Model documentation and write-up

Information included in this section may be shared publicly with challenge results. You can respond to these questions in an e-mail or as an attached file. Please number your responses.

1. Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.

I am an experienced data engineer and work in bioinformatic and computer vision field. I join many data science competition at both DrivenData and Kaggle, with a Kaggle GrandMaster tie.

1. What motivated you to compete in this challenge?

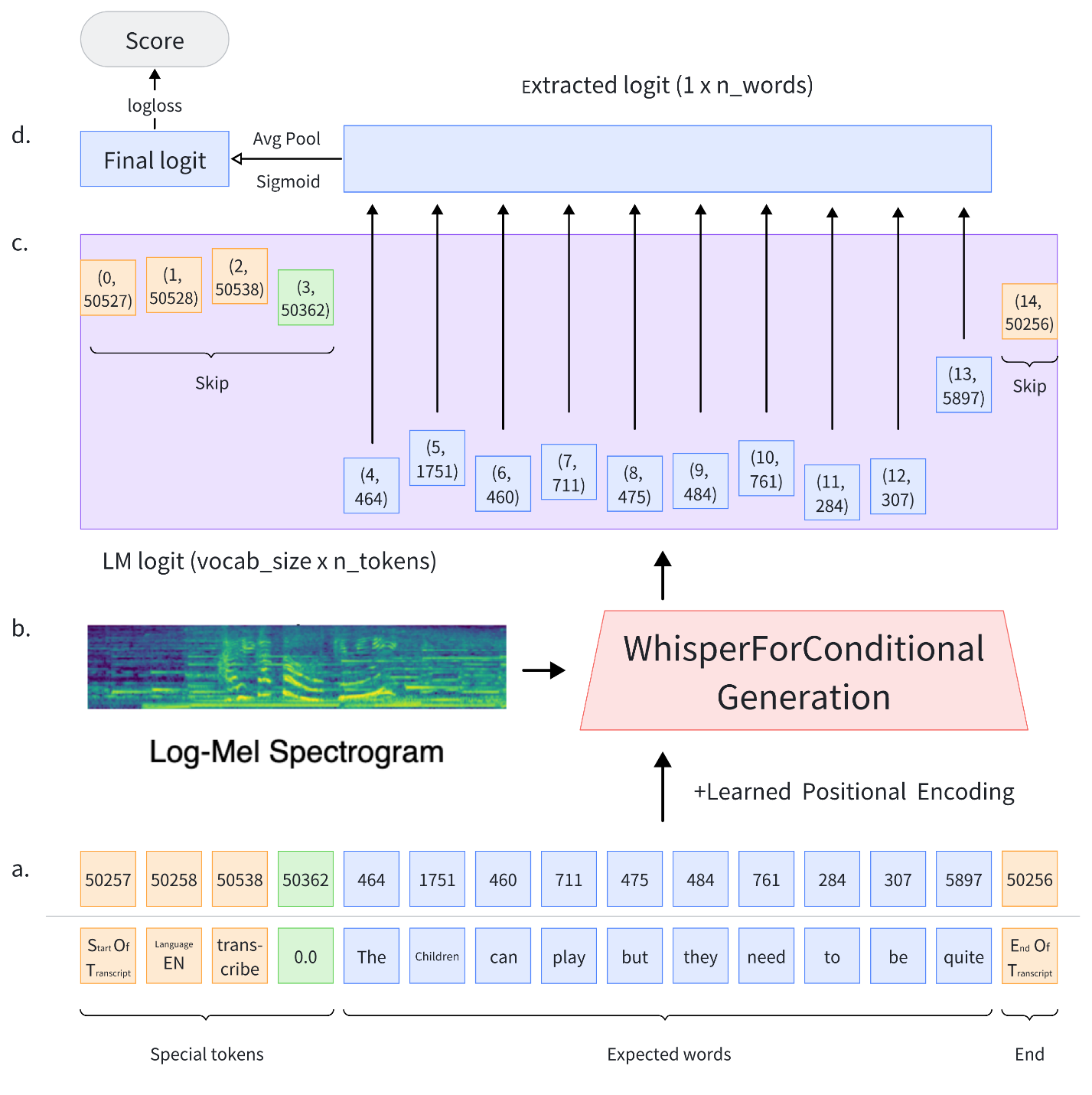
First I am interested in both audio and multi-model problem, the ermengecy of openai’s whisper provide a good opptunity to solve this kind of problem. Then the dataset of this challenge is will annotated and clean, only few has wrong label. This make this challenge a good benchmark for audio-text matching.

1. High level summary of your approach: what did you do and why?

We fine-tune OpenAI’s Whisper model with a custom loss function. For each word in expected\_text, we input its token into the decoder and compute the binary cross-entropy (BCE) loss between the token’s logit and its corresponding score. This approach enables us to evaluate the correctness of each word in a weakly supervised manner.

1. Do you have any useful charts, graphs, or visualizations from the process?

Yes, here is the figure of full workflow and the loss function design.



1. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.

The most impactful part is the loss function design, the code is listed below (WhisperForConditionalGenerationMask.forward):

|  |
| --- |
| loss\_fct = BCEWithLogitsLoss(reduction="none")  labels = labels.to(lm\_logits.device)  # extract the tokens (ignore first 4 special tokens)  lm\_logits = torch.gather(lm\_logits, dim=2, index=decoder\_input\_ids.unsqueeze(-1)).squeeze(-1)[:, 4:]  mask = decoder\_attention\_mask[:, 4:]  # only calculate those with words  lm\_logits = lm\_logits.masked\_fill(mask == 0, float('nan'))  lm\_logits = torch.nanmean(lm\_logits, dim=1)  loss = loss\_fct(lm\_logits, labels.float().reshape(-1)).mean() |

1. Please provide the machine specs and time you used to run your model.

* CPU (model): AMD ryzen 3960x
* GPU (model or N/A): RTX A6000 ada 48G
* Memory (GB): 128GB
* OS: Ubuntu.
* Train duration:~12h.
* Inference duration: ~4h (online submission)

1. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

You may take care of nonword performance, since this part is not pretrain by whisper, this part’s accuracy may be decayed.

1. Did you use any tools for data preparation or exploratory data analysis that aren’t listed in your code submission?

Nearly no, I only use potplayer to do error analysis.

1. How did you evaluate performance of the model other than the provided metric, if at all?

I only use log loss, since this is the both metric and loss function, no other metric is used.

1. What are some other things you tried that didn’t necessarily make it into the final workflow (quick overview)?

I tried many methods, and listed below:

* CLAP: nearly useless
* Two tower model, CNN or transformers to extract audio feature; bert to extract text feature, this best CV is about 0.36
* Hubert with CTC loss, very hard to make CTC loss penalty binary target

1. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?

I may try to pretrain whisper with common voice and this dataset, I don’t think any feature can help, the xgboost baseline only score 0.6 and our method score 0.2, this is a huge gap for logloss, I also tried groupkfold by expected\_text, our model also perform very good and hadn’t seen any decay on non-seen expected\_text.

1. What simplifications could be made to run your solution faster without sacrificing significant accuracy?

The inference speed is fast enough even for the large model. If we need more speed, the whisper small model is faster with abount 0.02 log loss decay.