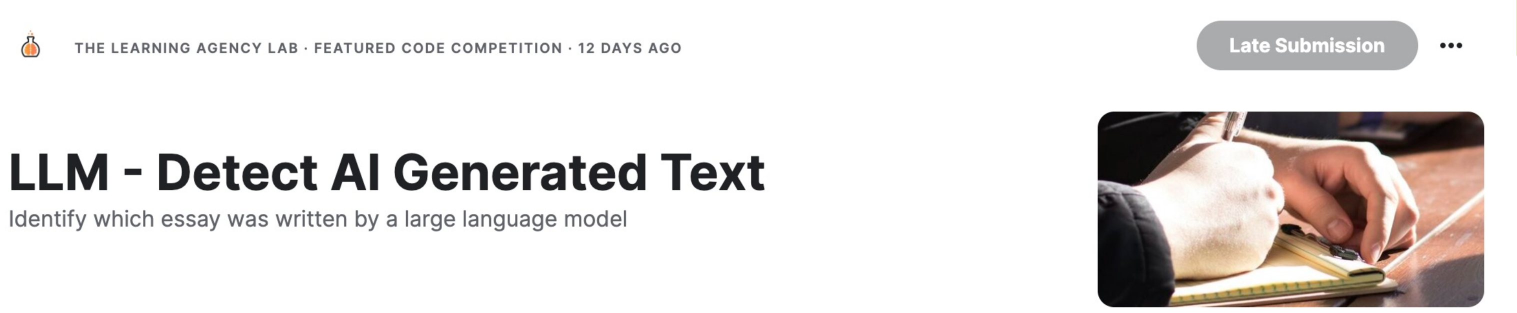
kaggle



THE LEARNING AGENCY LAB · FEATURED CODE COMPETITION · 12 DAYS AGO

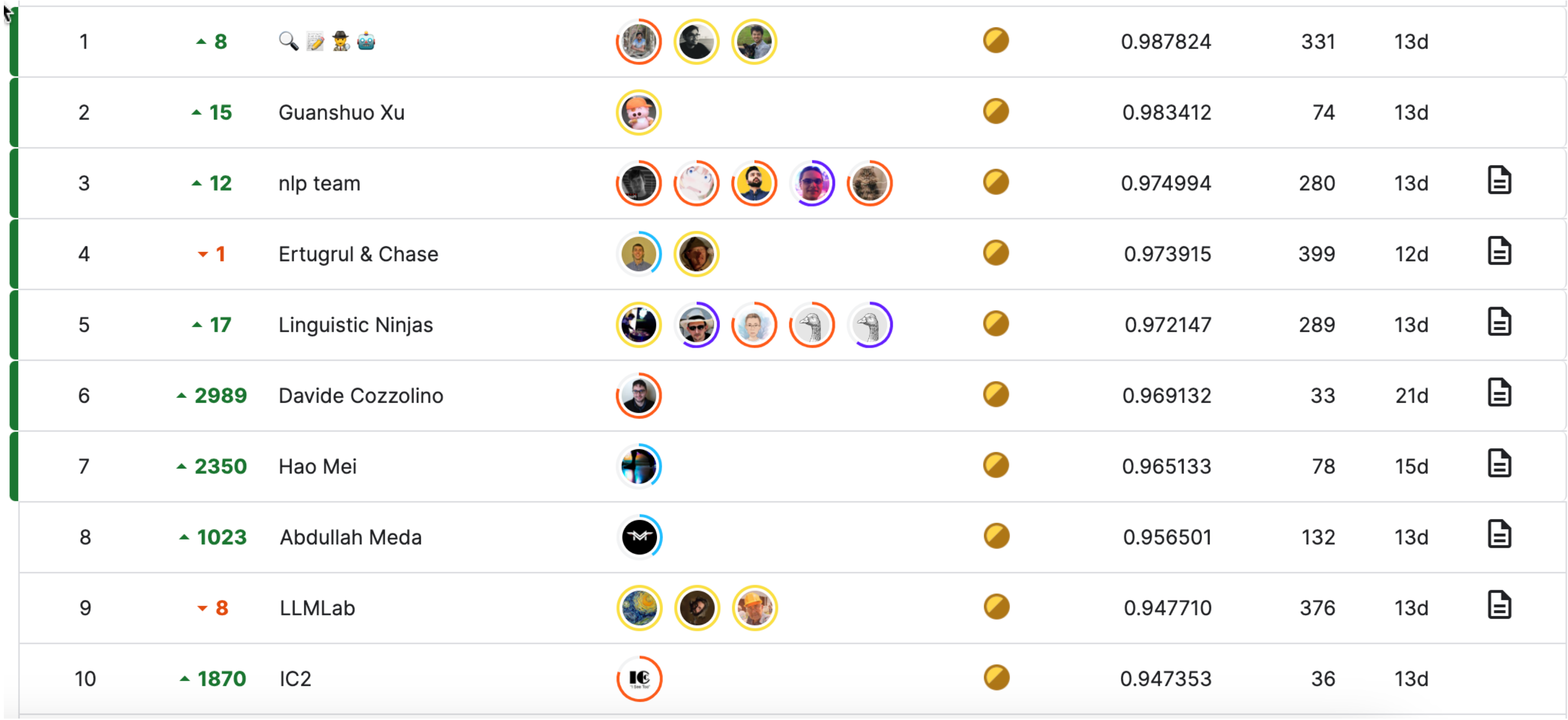
LLM - Detect AI Generated Text

Identify which essay was written by a large language model

2nd place solution

Guanshuo Xu
Data Scientist @ H2O.ai

Big Shakeup and Score Difference



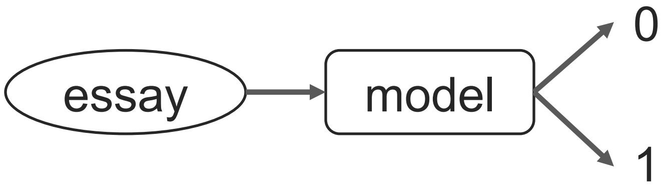
1	▲ 8	🔍📝🕵️🤖	●	0.987824	331	13d	
2	▲ 15	Guanshuo Xu	●	0.983412	74	13d	
3	▲ 12	nlp team	●	0.974994	280	13d	📄
4	▼ 1	Ertugrul & Chase	●	0.973915	399	12d	📄
5	▲ 17	Linguistic Ninjas	●	0.972147	289	13d	📄
6	▲ 2989	Davide Cozzolino	●	0.969132	33	21d	📄
7	▲ 2350	Hao Mei	●	0.965133	78	15d	📄
8	▲ 1023	Abdullah Meda	●	0.956501	132	13d	📄
9	▼ 8	LLMLab	●	0.947710	376	13d	📄
10	▲ 1870	IC2	●	0.947353	36	13d	

Content

- Problem description
- Data
- Solution
- Results

Problem description

- Determine whether a given essay was student written (negative: 0) or LLM generated (positive: 1)


essay

model

0

1

Data

- 15 prompts (essay topics) from persuade corpus
  - Only student written provided
  - 5 prompts in test data
  - 9000 data points in test
  - Private and public test data split by prompt

- Need LLM generated data for modeling

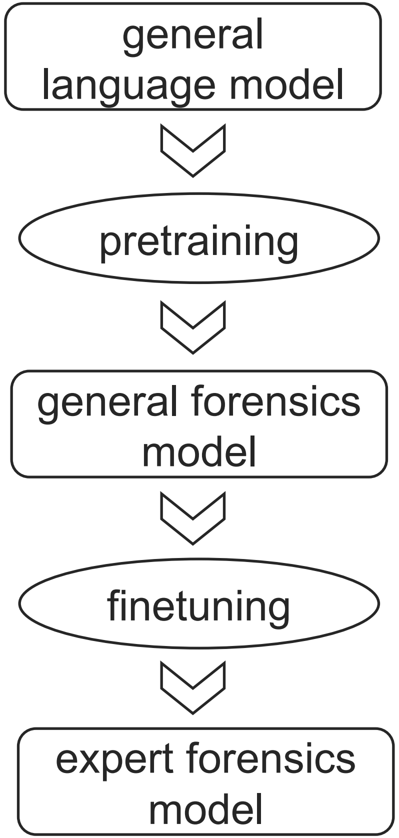
Difficulties in Modeling

i. Distribution difference between our generated and organizer generated. (major issue)
  - Prompts, LLMs, generation hyperparameters, postprocessing

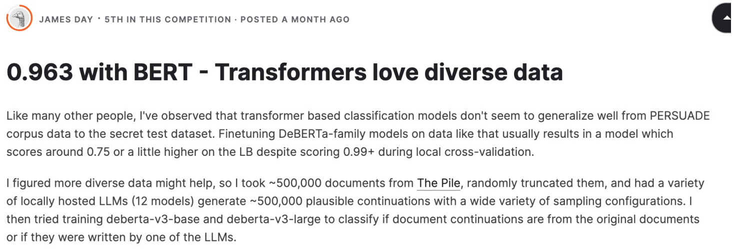
ii. Distribution difference of public and private test data due to splitting by prompts. (minor issue)

iii. 9000 test data big enough? No, because of strong correlation between data. (high variance, minor issue)

- Pretraining
  - Move the general language model closer to the competition problem domain.
  - Alleviate the issue caused by distribution difference because the transfer learning relies less on the in-domain data.
- Finetuning
  - Move the model closer to the five essay prompts.



Pretraining



JAMES DAY · 5TH IN THIS COMPETITION · POSTED A MONTH AGO

0.963 with BERT - Transformers love diverse data

Like many other people, I've observed that transformer based classification models don't seem to generalize well from PERSUADE corpus data to the secret test dataset. Finetuning DeBERTa-family models on data like that usually results in a model which scores around 0.75 or a little higher on the LB despite scoring 0.99+ during local cross-validation.

I figured more diverse data might help, so I took ~500,000 documents from The Pile, randomly truncated them, and had a variety of locally hosted LLMs (12 models) generate ~500,000 plausible continuations with a wide variety of sampling configurations. I then tried training deberta-v3-base and deberta-v3-large to classify if document continuations are from the original documents or if they were written by one of the LLMs.

➢ Steps to reproduce the James Day's post

1) Download the SlimPajama Dataset.
2) Collect around 500,000 long docs from C4 subset.
3) Randomly set a truncation point from each doc and use the 1024 tokens before the truncation point as prompt, and ask an LLM to generate the next 1024 tokens (by setting min and max token both to 1024). The LLM-generated 1024 tokens is the positive text, the original 1024 tokens after the truncation point is the negative text.
4) Finetune deberta models for one epoch on this dataset with max length 768.

➢ What is good about this pretraining approach

- Simulate the process of essay writing but with diverse prompts.
- Endless number of training data.

Pretraining examples: talk about some NBA players

prompt

Stephen Curry. Young is currently a free agent, with no teams taking the chance on him. Swaggy P is hoping to find an NBA home during the season.

Another forgotten bench piece from the Golden State Warriors run last season was David West. The toughness and gritty play of West made him a success in the NBA for many seasons, with his best runs coming for the New Orleans Hornets and Indiana Pacers.

West reached a point where he just wanted to win NBA Championships, so he signed with the San Antonio Spurs and

AI

won the 2018 NBA Championship there. West will be a free agent at the end of the season. The Spurs considered signing him, but he was only able to score around 4.2 points per game last season and was largely a backup. Still, if he can find a team that will take a chance on him, West can find another way to finish his NBA career on top of the podium.

Former big player Tim Duncan will be entering the league as a player development coach. The San Antonio Spurs hired Duncan as a player development coach in

human

later the Warriors. As a Warrior, West won two NBA Championship rings to fulfill his dream. West made the decision to retire this offseason, citing the difficulties of keeping up with the younger players today as his reason.

Jamal Crawford has been one of the best bench players in NBA history. His ability to score in bunches and the reputation of being a great teammate made Crawford stand out for seven different teams in his playing career. Crawford had his best run with the Los Angeles Clippers, as he won

Pretraining examples: talk about some energy saving program

prompt

of the Home Energy Savings Program and the Energy Efficiency and Fuel Switching in Public Buildings Program.

Originally announced as part of provincial Budget 2017, the Home Energy Savings Program (HESP) is a provincial initiative designed to assist low-income households in making energy efficiency upgrades to their home. Delivered by the Newfoundland and Labrador Housing Corporation (NLHC), this program provides non-repayable grants of up to $5,000 for cost-effective energy efficiency upgrades, including

AI

the installation of energy-efficient heating systems and appliances, and major renovation works that reduce carbon emissions associated with home heating. The Province plans to implement three new HESP regions to better service communities in the southern and east-central part of the province and to support retrofits of schools and community facilities.

To date, the program has provided assistance to more than 3,200 low-income households by delivering approximately $10 million in home energy efficiency

human

insulation and air sealing, to homes heated by electricity, propane or wood biomass.

Through the LCELF, the Province is expanding program eligibility to include households reliant on fuel-oil for spacing heating. This will reduce greenhouse gas emissions and energy bills. This expansion is supported by an investment of $8.57 million over four years (2018-2022), with $6.43 million from the Province and $2.14 million from the Federal Government. This funding is in addition to the $5.0 million provided for

Pretraining examples: references

prompt

modes of adapted receptor admittance to the somatodendritic claret membrane. Nature Neurosci. 9, 622–627 (2006). Yang, Y., Wang, X. B., Frerking, M. & Zhou, Q. Commitment of AMPA receptors to perisynaptic sites precedes the abounding announcement of abiding potentiation. Proc. Natl Acad. Sci. USA 105, 11388–11393 (2008). Holtmaat, A. & Svoboda, K. Experience-dependent structural synaptic bendability in the beastly brain. Nature Rev. Neurosci. 10, 647–658 (2009). Takahashi, T., Svoboda, K., Malinow, R.

AI

J. & Suzuki, M. The postsynaptic morphological plasticity of the nematode brain. Nature Rev. Neurosci. 6, 627–636 (2005). Liu, F., Shen, N., Zou, S. & Xu, S. Mechanisms responsible for synapse assembly and plasticity in the invertebrate central nervous system. Physiol. Rev. 90, 2341–2373 (2010). Kumar, V., Zajic, N., Reinsch, E., Lattimer, A. R. & Bekinschtein, M. Synaptic plasticity with dendritic actin remodeling. Science 315, 1101–1103 (2007). Kaneko, T., Miyagawa, S., Kato, K., Murata, K. & Tsuda, H. I

human

Experience deepening manual by active AMPA receptors into synapses. Science 299, 1585–1588 (2003). Whitlock, J. R., Heynen, A. J., Shuler, M. G. & Bear, M. F. Acquirements induces abiding potentiation in the hippocampus. Science 313, 1093–1097 (2006). Matsuo, N., Reijmers, L. & Mayford, M. Spine-type-specific application of anew actinic AMPA receptors with learning. Science 319, 1104–1107 (2008). Zamanillo, D. et al. Importance of AMPA receptors for hippocampal synaptic bendability but not for spatial learn

Pretraining future improvement

➢ Don’t fix the length of generated text

- Low generation quality when forcing LLM to generate long text. Potentially lowered the pretraining quality.
- Remove min_tokens to generate let the generation to stop more naturally.

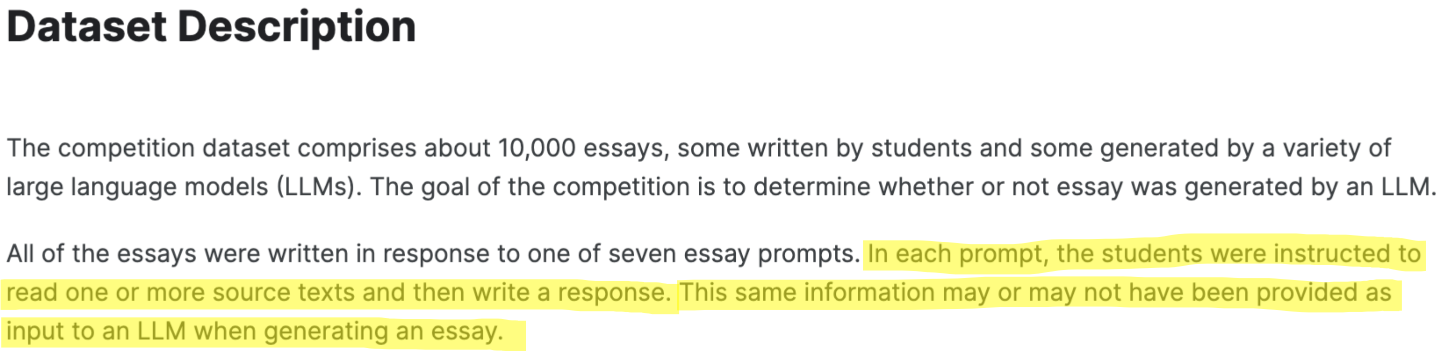
➢ Randomize hyperparameters during generation

- For example, randomize temperature.

➢ Use larger LLMs for generation

- In this competition, I mostly limited the LLM to 7B or lower to save time.
- Using larger and stronger LLMs can potentially improve the generation quality.

Finetuning



Dataset Description

The competition dataset comprises about 10,000 essays, some written by students and some generated by a variety of large language models (LLMs). The goal of the competition is to determine whether or not essay was generated by an LLM.

All of the essays were written in response to one of seven essay prompts. In each prompt, the students were instructed to read one or more source texts and then write a response. This same information may or may not have been provided as input to an LLM when generating an essay.

- We don’t have the source texts
- LLM generated without the source texts are easier to identify. There will be strong distribution difference on positive data.
- We need the source texts for generation to reduce this gap.

Finetuning: example prompts

prompt_name: The Face on Mars

assignment: You have read the article 'Unmasking the Face on Mars.' Imagine you are a scientist at NASA discussing the Face with someone who thinks it was created by aliens. Using information in the article, write an argumentative essay to convince someone that the Face is just a natural landform.Be sure to include: claims to support your argument that the Face is a natural landform; evidence from the article to support your claims; an introduction, a body, and a conclusion to your argumentative essay.

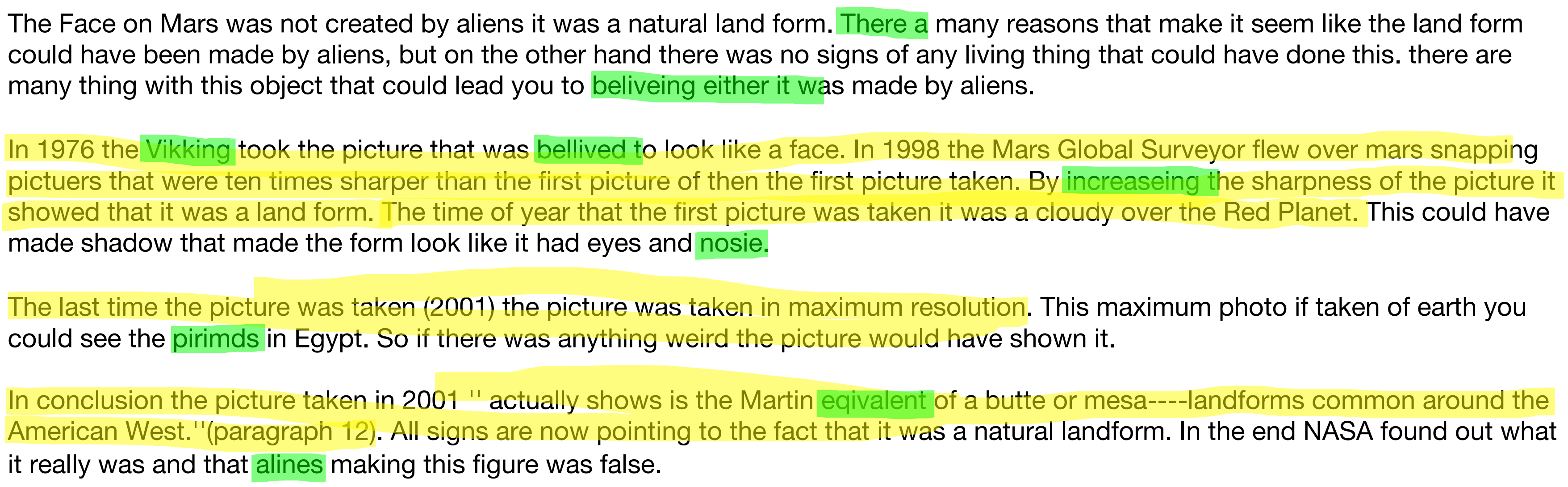
source_text: "Unmasking the Face on Mars"

prompt_name: Exploring Venus

assignment: In "The Challenge of Exploring Venus," the author suggests studying Venus is a worthy pursuit despite the dangers it presents. Using details from the article, write an essay evaluating how well the author supports this idea. Be sure to include: a claim that evaluates how well the author supports the idea that studying Venus is a worthy pursuit despite the dangers; an explanation of the evidence from the article that supports your claim; an introduction, a body, and a conclusion to your essay.

source_text: "The Challenge of Exploring Venus"

Finetuning: a typical student essay



The Face on Mars was not created by aliens it was a natural land form. There a many reasons that make it seem like the land form could have been made by aliens, but on the other hand there was no signs of any living thing that could have done this. there are many thing with this object that could lead you to beliveing either it was made by aliens.

In 1976 the Vikking took the picture that was bellived to look like a face. In 1998 the Mars Global Surveyor flew over mars snapping pictuers that were ten times sharper than the first picture of then the first picture taken. By increaseing the sharpness of the picture it showed that it was a land form. The time of year that the first picture was taken it was a cloudy over the Red Planet. This could have made shadow that made the form look like it had eyes and nosie.

The last time the picture was taken (2001) the picture was taken in maximum resolution. This maximum photo if taken of earth you could see the pirimds in Egypt. So if there was anything weird the picture would have shown it.

In conclusion the picture taken in 2001 '' actually shows is the Martin eqivalent of a butte or mesa----landforms common around the American West.''(paragraph 12). All signs are now pointing to the fact that it was a natural landform. In the end NASA found out what it really was and that alines making this figure was false.

- Essays are written based on the source text (article). See the yellow marks.
- Spell and grammar errors. See the green marks.

Step1: Create the csv file for language modeling

- Two columns in the csv file. One column contains the "assignment" as prompts, one column contains student essays.
- 5 prompts only.

Step2: Download and install h2o-llmstudio

- https://github.com/h2oai/h2o-llmstudio

Step3: Start llmstudio and import the csv file as dataset

Step4: Create experiments with the following settings

- dataset: the csv file we imported
- Problem type: causal language modeling
- LLM backbone: pick the LLM
- Max Length Prompt: 256 (the assignments)
- Max length answer: 1024 (the student essays)
- Max length: 1280
- Batch size: depends on gpu and LLM
- Epochs: 1
- The rest parameters use default values

Step5: Download the LM finetuned LLM

Step6: Generate essays with assignments as prompts

Finetuning: LLM generated examples

On April Fools day 2001 the Viking Mission for NASA snapped a picture of a "face" on Mars. " Face of Mars has since become a pop icon"(Unmask the face of Mars).

The first picture that people saw in 1976 reccomended an ancient civilization and the face was not real, Viking 2 only showed the "Martianequivalent of a butte or mesa - landform commong around the American West" (Unmasking the Face of Mars). Scientists also belive it is just a natural landform.

NASA went back with MOC and snapped the pictures all of the high-resolution imagery took time to obtain the data. The Face of Mars is like land formations on Earth because it is landform it is like a mesa common in the western states in the US and not just on Earth but on Mars because the MOC captured the image ten times sharper than the originale Vickings photo.

The Mars Global Surveyor (MGS) is like Earth Global Surveying System they use the MOC to catch the exact pictures on the Face and make all scientists happy because most all of the scientist believe that it is a normal natural landmark for Mars. But not all scientists are agreeig most scientist still believe that there is ancient civilization living on Mar's surface.

All of the new information that they took of the MOC showed that not one person or civilization on Mars could have possibly build this giant land mark. "So many people think its bona fide acient civilization"(Striving to meet the challenge 42).

- ➢ The generated text can use information from the article, without providing the actual article. See the yellow marks.
- ➢ The LLM more frequently creates spelling errors. See the green marks. So, the final classifier won't rely too much on spelling or grammar artifacts.

Finetuning

LLMs used for data generation

- h2ogpt-4096-llama2-7b, h2ogpt-4096-llama2-7b-chat, h2ogpt-4096-llama2-13b , h2ogpt-4096-llama2-13b-chat, Mistral-7B-v0.1.

Data generation

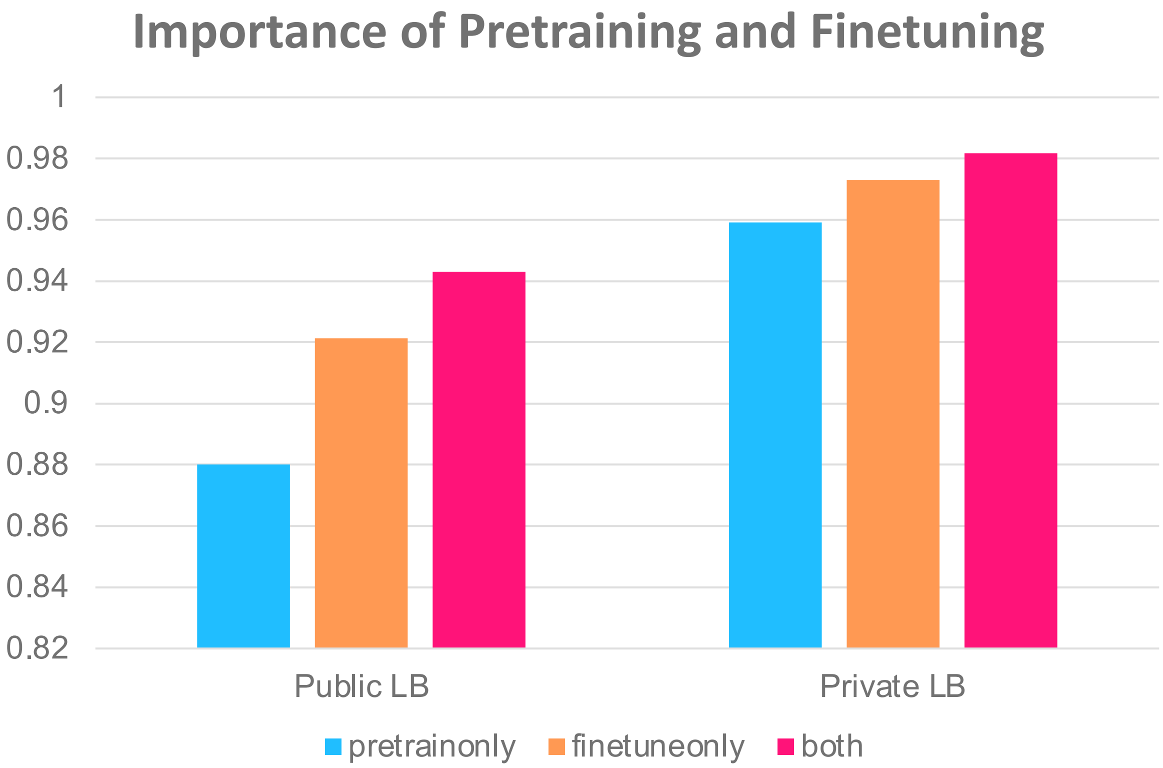
- On 5 prompts only, totally around 18,000 generated
- Generation temperature 1.0, 1.1, 1.2, 1.3

Finetuning: results

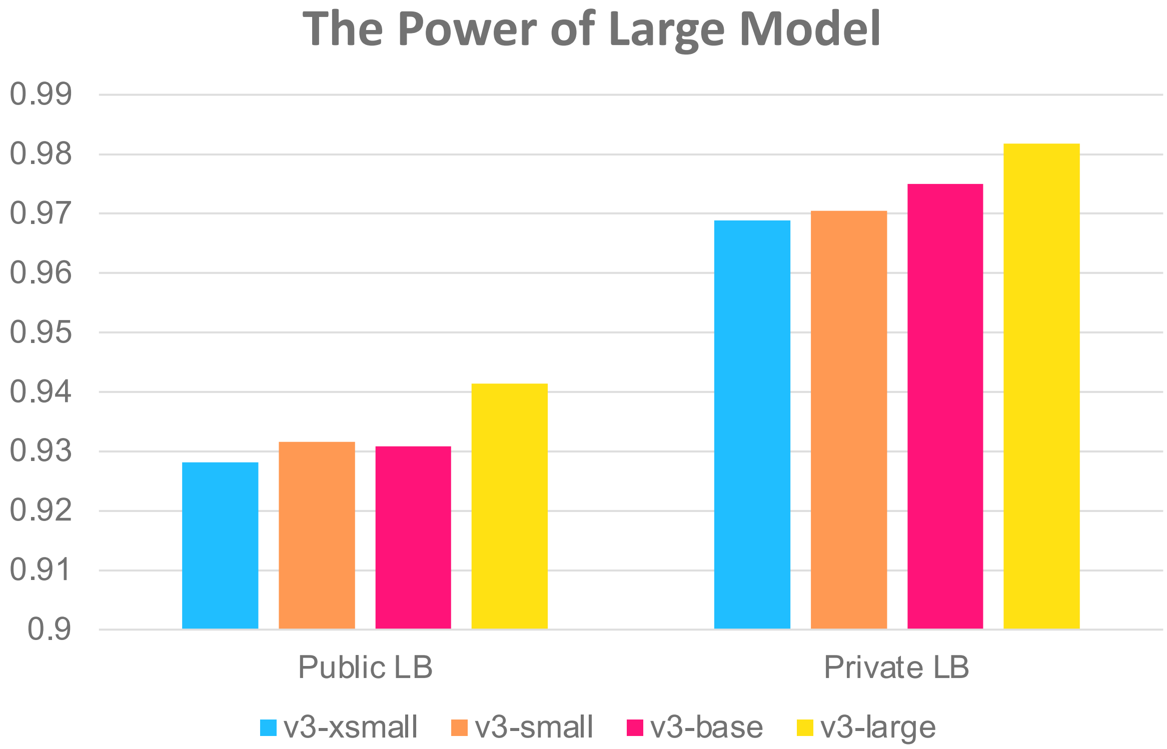
- Three pretrained models
  - deberta-v3-large1
  - deberta-v3-large2
  - deberta-large
- Two finetuning dataset
  - ftdata: the one we are talking about. 5 prompts only. Around 9,000 negatives VS 18,000 positives.
  - v4data: gathered by other competitors. All 15 prompts plus a little others. Around 28,000 negatives VS 45,000 positives.
- One epoch only
- Six model ensemble

model	Public LB	Private LB
(1) deberta-v3-large1_ftdata	0.9414	0.9819
(2) deberta-v3-large2_ftdata	0.9320	0.9804
(3) deberta-large_ftdata	0.9388	0.9786
(4) deberta-v3-large1_v4data	0.9441	0.9490
(5) deberta-v3-large2_v4data	0.9378	0.9372
(6) deberta-large_v4data	0.9052	0.9283
(1)+(2)+(3)	0.9431	0.9818
(1)+(2)+(3)+(4)+(5)+(6)	0.9672	0.9834

Importance of Pretraining and Finetuning



The Power of Large Model


The Power of Large Model
Public LB
Private LB
v3-xsmall
v3-small
v3-base
v3-large