Vo Declaration
Exhibit D

Exhibit 8

Filed Under Seal

Message

From: Ahmad Al-Dahle [redacted]@meta.com]
Sent: 10/12/2023 2:44:39 PM
To: Hugo Touvron [redacted]@meta.com]; Ahmad Al-Dahle [redacted]@meta.com]
Subject: Message summary [{"otherUserFbId":100035214585313,"threadFbId":null}]
Attachments: 380442431_2581322495364059_6205606330417476910_n.png

Ahmad Al-Dahle (10/12/2023 00:02:35 PDT):
>Did you change the data recipe from llama 2?

Ahmad Al-Dahle (10/12/2023 00:02:55 PDT):
>Very very exciting

Ahmad Al-Dahle (10/12/2023 00:03:02 PDT):
>i'm so glad we're betting on this direction seriously

Ahmad Al-Dahle (10/12/2023 00:03:30 PDT):
>By the way, I pushed on the team to build better data visualization tools for the underlying pretrained dataset. Do you have ideas for what we should be building there?

Ahmad Al-Dahle (10/12/2023 00:03:44 PDT):
>(do you see any improvement on rigor and urgency on your side)

Ahmad Al-Dahle (10/12/2023 00:07:18 PDT):
>the point on C is super key for your approach on SPM. What do you think are the drawbacks here?

Hugo Touvron (10/12/2023 00:10:02 PDT):
>Yes I keep only the data that seem good and I use more code & science doc

Hugo Touvron (10/12/2023 00:10:40 PDT):
>Good question, I think they're working on it right now.

Ahmad Al-Dahle (10/12/2023 00:10:48 PDT):
>I'm still digging deeper on data (and I can tell people are scrambling a bit on it)

Ahmad Al-Dahle (10/12/2023 00:11:06 PDT):
>but curious if you're seeing improvements on rigor or not yet

Ahmad Al-Dahle (10/12/2023 00:11:11 PDT):
>be very very direct

Hugo Touvron (10/12/2023 00:13:12 PDT):
>Slightly for data I think we have to rebuild the preprocessing pipeline. For instanec there is a bug with the parsing of the code for common crawl

Hugo Touvron (10/12/2023 00:13:24 PDT):

shared: 380442431_2581322495364059_6205606330417476910_n.png

Hugo Touvron (10/12/2023 00:13:56 PDT):
>They are still running experiments with this bug I don't know if the impact on perf is important but it add some noise

Ahmad Al-Dahle (10/12/2023 00:14:04 PDT):
>ooffff

Ahmad Al-Dahle (10/12/2023 00:15:22 PDT):
>how does this happen though

Ahmad Al-Dahle (10/12/2023 00:15:29 PDT):
>unusual

Hugo Touvron (10/12/2023 00:15:47 PDT):
>Here the feedback I gave to the data team yesterday on that:
>
>Hey
>It seem the issue is still present in the new code ablation.
>```
>231009_punitkoura_cc_code_replace_github_v2_run000
>231009_punitkoura_cc_code_replace_cc_v2_run000
>```
>I just run

>```
>cd [redacted]/fair_llm/data/ablations[redacted]
>cat * | grep '1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
34'
>```
>
>It seem this is the parsing of line number as mentioned previously by Todor.
>Here some example of doc with the issue:
>```
>https://graphics.rwth-aachen.de:9000/OpenFlipper-Free/Plugin-Datacontrol/[redacted]
>
>https://source.puri.sm/Librem5/uboot-imx/[redacted]
>
>
>Is this expected? Or did we think we'd removed the documents with this issue?
>
>Maybe the easiest way is to fix this when we pre-process common crawl.

Hugo Touvron (10/12/2023 00:15:57 PDT):
>you have an exemple here
>https://graphics.rwth-aachen.de:9000/OpenFlipper-Free/Plugin-Datacontrol[redacted]

Ahmad Al-Dahle (10/12/2023 00:15:59 PDT):
>it's clearly happening around dates

Hugo Touvron (10/12/2023 00:16:24 PDT):
>there is line count but the parsing change that to a list of number

Ahmad Al-Dahle (10/12/2023 00:18:36 PDT):
>so how would data visualizations count this?

Ahmad Al-Dahle (10/12/2023 00:18:42 PDT):
>would we look for n grams?

Ahmad Al-Dahle (10/12/2023 00:18:50 PDT):
>like we need to audit manually

Ahmad Al-Dahle (10/12/2023 00:18:58 PDT):
>and sample different parts of the data distribution

Ahmad Al-Dahle (10/12/2023 00:19:01 PDT):
>especially around the pile

Hugo Touvron (10/12/2023 00:19:30 PDT):
>We can display the original document and the preprocess version of the data

Ahmad Al-Dahle (10/12/2023 00:19:32 PDT):
>but visualizations may not catch mistakes like this

Hugo Touvron (10/12/2023 00:20:46 PDT):
>Yes we can do some clustering based on fasttext embedding and data quality this can help to identify some issue. We do that 2 weeks ago and we identified some issue. But we have to go deeper in this analysis

Ahmad Al-Dahle (10/12/2023 00:21:01 PDT):
>totally

Ahmad Al-Dahle (10/12/2023 00:21:09 PDT):
>do we have a person dedicated to doing this audit

Ahmad Al-Dahle (10/12/2023 00:21:27 PDT):
>I think we should consider asking someone to build this stuff full time and do nothing but dig deep and audit for mistakes

Ahmad Al-Dahle (10/12/2023 00:21:36 PDT):
>maybe multiple people honestly

Hugo Touvron (10/12/2023 00:22:19 PDT):
>Not sure Laurens looks at the clusters, me too I but I don't know if anyone is actively working on it

Ahmad Al-Dahle (10/12/2023 00:22:40 PDT):
>Well, we have 1 person ;) ... me

Ahmad Al-Dahle (10/12/2023 00:22:51 PDT):
>i'm going to ask to start seeing clusters

Ahmad Al-Dahle (10/12/2023 00:23:04 PDT):
>and I also want to know whose we have auditing this every day

Ahmad Al-Dahle (10/12/2023 00:23:16 PDT):
>(I already pushed based on this conversation)

Hugo Touvron (10/12/2023 00:23:17 PDT):
>Clearly it's important to have more advanced analysis for data quality

Ahmad Al-Dahle (10/12/2023 00:23:28 PDT):
>super key

Hugo Touvron (10/12/2023 00:23:29 PDT):
>We can also do that for eval imo

Hugo Touvron (10/12/2023 00:23:53 PDT):
>to better understand models' limitations

Ahmad Al-Dahle (10/12/2023 00:23:54 PDT):
>also, I pushed on the team to show me a 7B run with high quality tokens only that can beat llama 2 and measure where it is relative to Mistral

Ahmad Al-Dahle (10/12/2023 00:24:04 PDT):
>(I want them to try to prove that our data quality is good)

Ahmad Al-Dahle (10/12/2023 00:24:23 PDT):
>without your recipe

Ahmad Al-Dahle (10/12/2023 00:24:26 PDT):
>with your recipe it will be 1000x better ;)

Ahmad Al-Dahle (10/12/2023 00:24:53 PDT):
>wdyt?

Ahmad Al-Dahle (10/12/2023 00:25:16 PDT):
>you suggested this in our call 2 days ago

Hugo Touvron (10/12/2023 00:25:40 PDT):
>For mistral I have some info:
>- with the same data mix and their new common crawl they reach llama2 performance with 1T tokens
>- their final run is on new data mix and on more token than llama2
>- They have a bigger model (smaller than 150B) that reach 77-78 on mmlu

Hugo Touvron (10/12/2023 00:26:00 PDT):
>Yes 100% agree with you

Ahmad Al-Dahle (10/12/2023 00:26:24 PDT):
>this is good signal

Ahmad Al-Dahle (10/12/2023 00:26:29 PDT):
>but honestly ... our goal needs to be GPT4

Ahmad Al-Dahle (10/12/2023 00:26:36 PDT):
>Mistral is peanuts for us

Ahmad Al-Dahle (10/12/2023 00:26:41 PDT):
>we have 64k GPUs coming!

Ahmad Al-Dahle (10/12/2023 00:26:48 PDT):
>we need to learn how to build frontier and win this race

Ahmad Al-Dahle (10/12/2023 00:27:26 PDT):
>do we know their underlying size of their dataset?

Ahmad Al-Dahle (10/12/2023 00:27:41 PDT):
>1T to reach llama 2 but how big did they train on

Hugo Touvron (10/12/2023 00:27:56 PDT):
>Yes but if they manage to reach 78 on mmlu it's a good signal it's close to claude and a step to reach GPT4 (86.4%)

Ahmad Al-Dahle (10/12/2023 00:28:17 PDT):
>We should be able to do better

Hugo Touvron (10/12/2023 00:28:22 PDT):
>8T I think

Ahmad Al-Dahle (10/12/2023 00:28:23 PDT):
>I will be VERY VERY sad if we don't

Ahmad Al-Dahle (10/12/2023 00:28:32 PDT):
>we just need our 13T to be good!

Hugo Touvron (10/12/2023 00:28:46 PDT):
>but we can do more epoch on the data if use more regularisation

Ahmad Al-Dahle (10/12/2023 00:28:47 PDT):
>Do you think our mix is at least good? Are we being aggressive enough?

Ahmad Al-Dahle (10/12/2023 00:29:13 PDT):
>if we clean up the stupidity in the data

Hugo Touvron (10/12/2023 00:29:16 PDT):
>Currently I m very agressive for the data mix
>llama2 data mix was bad

Ahmad Al-Dahle (10/12/2023 00:29:33 PDT):
>do we have the right datasets in there

Ahmad Al-Dahle (10/12/2023 00:29:46 PDT):
>is there anything you wanted to use but couldn't for some stupid reason?

Hugo Touvron (10/12/2023 00:30:30 PDT):
>Libgen is a good one :) I'm currently using it

Ahmad Al-Dahle (10/12/2023 00:30:44 PDT):
>yep, I cleared the path to use that one

Hugo Touvron (10/12/2023 00:30:46 PDT):
>For other dataset I don't know if we update wikipedia, arxiv etc.. since llama2 I ll ask the data team

Ahmad Al-Dahle (10/12/2023 00:30:50 PDT):
>I can clear the path for other things

Ahmad Al-Dahle (10/12/2023 00:31:04 PDT):
>we should!

Ahmad Al-Dahle (10/12/2023 00:31:55 PDT):
>this is a lot of fun! We're going to make this awesome together (me pushing from the top and you making it happen!)

Ahmad Al-Dahle (10/12/2023 00:32:25 PDT):
>when does your run finish? (so I know when to bug you ;))

Hugo Touvron (10/12/2023 00:32:46 PDT):
>Thank you very much for your help, it's really precious :)

Ahmad Al-Dahle (10/12/2023 00:32:58 PDT):
>We're a team! We both want the same thing

Hugo Touvron (10/12/2023 00:33:09 PDT):
>In 1 month it s a 16T token run but I'll have enough signal before that

Ahmad Al-Dahle (10/12/2023 00:33:13 PDT):
>llama 3 is literally all I care about

Hugo Touvron (10/12/2023 00:33:20 PDT):
>I'll do some lr annealing

Ahmad Al-Dahle (10/12/2023 00:33:26 PDT):
>I did products only because I want the compute and budget to build frontier models

Ahmad Al-Dahle (10/12/2023 07:32:38 PDT):
>Free for a call to catch up?

Hugo Touvron (10/12/2023 07:37:30 PDT):
>Yes :)

Ahmad Al-Dahle (10/12/2023 07:40:12 PDT):
>http://www.[redacted]net:8083/index_quality2.html

Ahmad Al-Dahle (10/12/2023 07:44:39 PDT):
>https://docs.google.com/document/[redacted]