EXHIBIT THREE

From: Praveen Krishnan[/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=2CC6EBFAAC61420798D315B345B9A90E[redacted]

Sent: Wed 3/27/2024 4:50:06 AM (UTC)

To: Praveen Krishna[redacted]meta.com]; Ankit Ramchanda[redacted]meta.com]; Todor Mihaylo[redacted]@meta.com]; Simon Vandenhende[redacted]meta.com]; Viktor Kerke[redacted]@meta.com]

Subject: Message summary [{"otherUserFbId":null,"threadFbId":7587376214648091}]

Attachment: 433266281_1156194791991204_9160559402101587450_n.png

Todor Mihaylov (3/26/2024 12:46:52 PDT):
>Hi Ankit and Praveen,
>
>Me and Simon were discussing the multimodal performance, especially MMMU which we are doing worse than GPT4.
>Simon mentioned that if they have 1000 high quality examples of interleaved problems per subtask, this should be good be very good for fine-tuning.

Todor Mihaylov (3/26/2024 12:47:05 PDT):
>There is a low hanging fruit here

Todor Mihaylov (3/26/2024 12:47:51 PDT):
>We have scitech potion of Libgen which has mostly high school textbooks which are very easy to associate with given task.

Todor Mihaylov (3/26/2024 12:48:09 PDT):
>16% of the data is ebook which are easy to parse since they are usually html + images.

Ankit Ramchandani (3/26/2024 12:48:24 PDT):
>This sounds doable. If we source relevant images, is it possible to have the examples be annotated to get high quality Q&A pairs?

Todor Mihaylov (3/26/2024 12:48:36 PDT):
>https://docs.google.com/document/d[redacted]

Todor Mihaylov (3/26/2024 12:49:20 PDT):
>Can you look this up.
>basically search for “geometry” in the title of the book. And pull some examples to google docs.

Todor Mihaylov (3/26/2024 12:49:35 PDT):
>A few people can take a look in each category and determine if it is worth exploring

Simon Vandenhende (3/26/2024 12:50:06 PDT):
>I think we can do either: (i) parse the data and merge with pre-training data towards end. (ii) Curate question-answer pairs for fine-tuning. But, for the latter there might be easier sources where can directly have annotators get the question-answer pairs into some format we desire.

Todor Mihaylov (3/26/2024 12:50:07 PDT):
>Regardless, it is worth parsing the epubs

Todor Mihaylov (3/26/2024 12:50:13 PDT):
>Since this will have interleaved data

Todor Mihaylov (3/26/2024 12:51:16 PDT):
>I am adding Viktor, who currently processes the ebooks from commoncrawl.

Todor Mihaylov (3/26/2024 12:51:31 PDT):
>@Viktor what do you use for processing epubs

Ankit Ramchandani (3/26/2024 12:51:44 PDT):
>yes sounds good! We also have pdf parsing code in our pipeline now so we can extract images and text out of textbooks which could be very helpful

Todor Mihaylov (3/26/2024 12:52:09 PDT):
>We were discussing that we need to know the location of images.

Todor Mihaylov (3/26/2024 12:52:49 PDT):
>If you have this for pdfs, I would suggest to reprocess the whole scitech.

Todor Mihaylov (3/26/2024 12:53:45 PDT):
>Simon do you have some pointers on MMMU

Ankit Ramchandani (3/26/2024 12:53:45 PDT):
>we can get locations of images and texts. let me look more into libgen later today. I was trying to access the data earlier today actually but I dont know how to set up aws s3 properly

Simon Vandenhende (3/26/2024 12:54:15 PDT):
>https://mmmu-benchmark.github.io/ This is the dataset

Ankit Ramchandani (3/26/2024 12:54:48 PDT):
>btw is libgen safe for use legally? (from googling around, seems no)

Simon Vandenhende (3/26/2024 12:54:50 PDT):
>We have not run a very deep analysis on what domains see the largest failure cases. As said, we only started focusing on it today, as we've been fixing the other tasks first.

Todor Mihaylov (3/26/2024 12:56:04 PDT):
>We currently use it for training.

Simon Vandenhende (3/26/2024 12:56:33 PDT):
Redacted legal Redacted approved it in the end.

Praveen Krishnan (3/26/2024 12:57:46 PDT):
>This sounds very useful.

Viktor Kerkez (3/26/2024 13:20:52 PDT):
>I'm extracting pdfs at the moment, next ones on the list are epubs. having some problems with presto breaking with "Remote page too long" so if the fix I found works I'll have the ebooks until Thursday

Todor Mihaylov (3/26/2024 13:24:42 PDT):
>@Praveen or @Ankit would you be able to help with epubs - you probably have idea how to save interleaved data in a usable format.

Todor Mihaylov (3/26/2024 13:26:16 PDT):
[redacted]

[redacted]

>
>I got this setting from an infra person. It fixes the remote page too long.

Viktor Kerkez (3/26/2024 13:30:02 PDT):
>Yes that's the setting I'm also using.

Ankit Ramchandani (3/26/2024 13:31:12 PDT):
>> I'm extracting pdfs at the moment
>
>just for understanding, does this mean you are downloading pdfs or running ocr or other processing on it?

Ankit Ramchandani (3/26/2024 13:33:46 PDT):
>let's talk offline maybe so I understand this better and then I can summarize for the group here. are you at your desk?

Todor Mihaylov (3/26/2024 13:35:18 PDT):
>Me - yes.

Viktor Kerkez (3/26/2024 13:51:37 PDT):
>Yes, I'm creating a hive table with only pdfs, uhauling it to WS and processing

Ankit Ramchandani (3/26/2024 14:01:47 PDT):
>are the pdfs saved on manifold? If we have a hive table with paths to pdf files that would be great! That is what our pdf parsing pipeline expects

Todor Mihaylov (3/26/2024 14:07:20 PDT):
>https://www.amazon.com/Counting-Configurations-Combinatorics-Arithmetic-Mathematics/dp/0387955526/ref=sr_1_1?dib=eyJ2IjoiMSJ9.VXTOaNdS1P1658pBulnKyQ.A89RtUF2xbgloxzWiHrGqbk3VJEy4sGH7qQ0kvyOlEU&dib_tag=se&keywords=Counting+and+configurations%3A+problems+in+combinatorics%2C+arithmetic%2C+and+geometry&qid=1711487097&s=books&sr=1-1&asin=0387955526&revisionId=&format=4&depth=1

Todor Mihaylov (3/26/2024 14:07:41 PDT):

shared: 433266281_1156194791991204_9160559402101587450_n.png

Todor Mihaylov (3/26/2024 14:07:59 PDT):
>It seems very easy to map figure to problems and solutions.

Ankit Ramchandani (3/26/2024 14:57:18 PDT):
>it is tricky. The pdf does not know which figure is figure 1. So we have to first map image to its caption/name and then the name to its occurrence in regular text and then that text to the problem number

Ankit Ramchandani (3/26/2024 15:20:09 PDT):
>Can someone tell me how to access the s3 buckets in this doc? Running
>`aws s3 ls s3://fairspark-data/fair-use/scitech/` tells me to run aws configure, and I am not sure what to put for access key id and secret access key when running aws configure? maybe @Todor knows

Todor Mihaylov (3/26/2024 15:22:23 PDT):
>https://docs.google.com/document/d[redacted]tMHUWQoTFUGSBPXRwTUC1g/edit#bookmark=id.7ijfboawmuze you need to request access from here

Todor Mihaylov (3/26/2024 15:27:30 PDT):
>Yes. I suggest to look at ebooks where this is likely a table with 2 columns and Figure 1 is under the image of figure 1.
>It would not be super certain but you could do it with high precisions and lower recall.

Viktor Kerkez (3/26/2024 18:31:42 PDT):
>Now I'm getting:
>



>
>@Todor Mihaylov Are your queries finishing correctly?

Todor Mihaylov (3/26/2024 21:50:06 PDT):
>This seems like a different error. I probably did not have very big tables to generate "too many files"