EXHIBIT THIRTEEN

From: Mary Williamson[/O=THEFACEBOOK/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=7B844EC3D774479A94F24F3C679C955E-[redacted]

Sent: Mon 10/10/2022 7:05:46 PM (UTC)

To: Mary Williamson[redacted]@meta.com]; Stephen Roller[redacted]@meta.com]

Subject: Message summary [{"otherUserFbId":100020258253907,"threadFbId":null}]

Attachment: 310360090_442897674498228_1598685202944442996_n.png

Attachment: 310236819_1172234356709771_7648913428283707991_n.png

Mary Williamson (10/10/2022 11:13:32 PDT):
>Does EMEA have some corpus total of 5T tokens? Is there some miscommunication of counting tokens that is going on?

Stephen Roller (10/10/2022 11:20:16 PDT):
>they have 4.5TB which was about 1T tokens. After I ran the deduplication it was about 3.2TB and 778T tokens

Stephen Roller (10/10/2022 11:20:24 PDT):
>probably ppl confusing TB vs Trillion tokens

Mary Williamson (10/10/2022 11:21:51 PDT):
>And that's about what you have right

Stephen Roller (10/10/2022 11:22:06 PDT):
>https://docs.google.com/spreadsheets/d/[redacted]PrdA/edit#gid=494549991

Stephen Roller (10/10/2022 11:22:11 PDT):
>everything is extensively broken down and accurately

Mary Williamson (10/10/2022 11:25:14 PDT):
>Thanks much - I'll have a look

Stephen Roller (10/10/2022 11:25:32 PDT):
>TP = theorem proving

Stephen Roller (10/10/2022 11:25:59 PDT):
>there's an EMEA tokens at the top

Mary Williamson (10/10/2022 11:26:25 PDT):
>Yeah I see it but then I have to go check the Chinchilla paper to figure out this all compares ha

Stephen Roller (10/10/2022 11:27:22 PDT):

shared: 310360090_442897674498228_1598685202944442996_n.png

Mary Williamson (10/10/2022 11:27:35 PDT):
>Have you talked to look at this "10 Trillion Tokens" effort? I'm trying basically articulate what is the north star goal for text (the next step after whatever we train on now) --> 10X ish bigger for example. How would you articulate that and how far off are we?

Mary Williamson (10/10/2022 11:27:46 PDT):

>look at -> Luke about

Mary Williamson (10/10/2022 11:29:42 PDT):
>Given what you just sent me, I have a more educated opinion after I look ☺ I just keep having EMEA say we have this many tokens already

Stephen Roller (10/10/2022 11:35:22 PDT):
>CC-XL was already 12T tokens of english but i bet deduplication and quality filtering you would pull that down to ~3-4T tokens.

Stephen Roller (10/10/2022 11:37:39 PDT):
>maybe even less

Stephen Roller (10/10/2022 11:38:36 PDT):
>tbh i don't think there's much more than that available from publicly available crawls of any sort

Stephen Roller (10/10/2022 11:38:55 PDT):
>you'd have to start either going to fb data (which, prolly we should), or start being much more creative

Stephen Roller (10/10/2022 11:39:30 PDT):
>you know why i think openai created whispernet? i bet they ran it on every video and every podcast and more in order just mop up all the data locked into spoken dialogue

Stephen Roller (10/10/2022 11:40:17 PDT):
>and gopher/chinchilla reports some 500B tokens of a mysterious "books" dataset that christ, i have no idea where that is from

Mary Williamson (10/10/2022 11:40:21 PDT):
>What is CC-XL? Was that used for one of the LLMs?

Stephen Roller (10/10/2022 11:40:36 PDT):
>xlmg created it. it's _every_ CC snapshot concattenated together

Mary Williamson (10/10/2022 11:40:49 PDT):
>Ah OK but we can't use it?

Stephen Roller (10/10/2022 11:40:50 PDT):
>i think there's about 80 of them? EMEA is currently using ~4

Stephen Roller (10/10/2022 11:41:07 PDT):
>it's not been cleaned or deduplicated. enormous effort required there

Stephen Roller (10/10/2022 11:41:34 PDT):
>we're comparing things at the _start_ of pipelines with things at the _end_ of pipelines

Stephen Roller (10/10/2022 11:43:17 PDT):
>we're already reaching into big troves of pirated books on bittorrent and still only coming back with 20% of that...

Mary Williamson (10/10/2022 11:44:53 PDT):
>Legal Redacted - Privileged

Stephen Roller (10/10/2022 11:45:25 PDT):
Redacted - Privileged

Stephen Roller (10/10/2022 11:45:29 PDT):

>that's the books3 dataset

Stephen Roller (10/10/2022 11:45:43 PDT):
>will be iffy, idk

Stephen Roller (10/10/2022 11:45:53 PDT):
>emea went and grabbed it lol

Mary Williamson (10/10/2022 11:46:21 PDT):
>FFS

Stephen Roller (10/10/2022 11:46:22 PDT):
>might be okay in that someone else did the torrenting and putting it into corpus form and oh hey, look what just fell off the back of a truck

Stephen Roller (10/10/2022 11:46:38 PDT):
>sorta like reddit is okay because the pushshift people do all the curation and we just download their shit

Mary Williamson (10/10/2022 11:46:42 PDT):
>But this obviously needs approval before anybody should go grab it

Stephen Roller (10/10/2022 11:48:11 PDT):
>i've been blessed with offloading the SRT stuff to others

Stephen Roller (10/10/2022 11:48:26 PDT):
>and i've been focused on the hard technical numbers

Mary Williamson (10/10/2022 11:48:43 PDT):
>Ya that is good! ☺ I meant EMEA

Mary Williamson (10/10/2022 11:48:56 PDT):
>Should not have downloaded it prior to approval

Stephen Roller (10/10/2022 11:49:10 PDT):
>mary i'mma be honest with you

Stephen Roller (10/10/2022 11:49:16 PDT):
>given that some of these SRTs go on for 3 months

Stephen Roller (10/10/2022 11:49:28 PDT):
>that is really not practical

Mary Williamson (10/10/2022 11:49:43 PDT):
><sigh> ☺ I get it

Mary Williamson (10/10/2022 11:50:01 PDT):
>But especially for something you describe as "sketchy", we need approval first

Mary Williamson (10/10/2022 11:50:23 PDT):
>Anyway - another topic for another time

Stephen Roller (10/10/2022 11:50:29 PDT):
>i'm being a bit hyperbolic in about how sketchy it is Redacted - Privileged lawyers Redacted - Privileged Redacted - Privileged

Mary Williamson (10/10/2022 11:51:00 PDT):

>(Any time you guys have an SRT woe, someone should be escalating that to Joelle right away.)

Mary Williamson (10/10/2022 11:51:21 PDT):
>But so, do you have a view as to where to go that non-public?

Stephen Roller (10/10/2022 11:51:34 PDT):
>facebook.com ?

Stephen Roller (10/10/2022 11:51:42 PDT):
>i mean

Stephen Roller (10/10/2022 11:51:44 PDT):
>here's a hot take

Mary Williamson (10/10/2022 11:51:46 PDT):
>hehe

Stephen Roller (10/10/2022 11:51:48 PDT):
>our lawyers need to grow some balls

Stephen Roller (10/10/2022 11:51:57 PDT):
>openai and google are clearly violating IP left and right

Mary Williamson (10/10/2022 11:52:19 PDT):
>So, the way I'm thinking about it is that internal and external data need to be separate strategies. We may want to mix them (probably will want to), but we want to get the most we can from external as well as internal

Mary Williamson (10/10/2022 11:52:28 PDT):
>Ya it's a fair amount

Mary Williamson (10/10/2022 11:52:30 PDT):
>*fair point

Mary Williamson (10/10/2022 11:52:47 PDT):
>We'd have to articulate why it's worth the risk

Mary Williamson (10/10/2022 11:53:07 PDT):
>Do we have all the non-public scientific repositories?

Mary Williamson (10/10/2022 11:53:34 PDT):
>(Or better asked: could we get that via licensing assuming we do not beyond Arxiv and other public ones right now)

Stephen Roller (10/10/2022 11:53:59 PDT):
>we already have arxiv and the other public ones thanks to paperswithcode folks

Mary Williamson (10/10/2022 11:54:11 PDT):
>Yeah that I know

Stephen Roller (10/10/2022 11:54:13 PDT):
>going beyond that into paid journals is probably high cost and low yield

Mary Williamson (10/10/2022 11:54:20 PDT):
>K

Stephen Roller (10/10/2022 11:54:35 PDT):
>partnering with the library of congress would be more interesting lol

Stephen Roller (10/10/2022 11:55:00 PDT):
>but like, that's why i say i think openai is cracking into vaults of other places

Stephen Roller (10/10/2022 11:55:23 PDT):
>i think we're rapidly approaching literally at the limit of the internet unless we start pulling from behind our walled garden

Stephen Roller (10/10/2022 11:55:38 PDT):
>and that's using user data which lordy

Stephen Roller (10/10/2022 11:56:04 PDT):
>that's my 18 month prediction

Mary Williamson (10/10/2022 11:56:30 PDT):
>Right agreed (and IMHO we should *use that walled garden*. it's our competitive advantage), but what I'm not clear on yet: if I gave you a buunch of money, there's nothing you would *buy* outside of that walled garden that would get you 1T+ (or more) but like some large number

Stephen Roller (10/10/2022 11:57:46 PDT):
>i haven't put many thought cycles into what i would buy

Stephen Roller (10/10/2022 11:58:14 PDT):
>i'd almost suggest we'd be better off spending all that money on downstream annotators for human-in-the-loop

Stephen Roller (10/10/2022 11:58:45 PDT):
>this is gonna piss off all the scientists but the main finding from my baby zucchini ablations is that not even data matters if it seems to be sufficient in quantity and quality

Stephen Roller (10/10/2022 12:00:06 PDT):
>but yeah i guess i'd buy the rights to a _ton_ of books

Stephen Roller (10/10/2022 12:01:29 PDT):

shared: 310236819_1172234356709771_7648913428283707991_n.png

Stephen Roller (10/10/2022 12:05:46 PDT):
>that said i'm also explicitly proposing a strategy where we kick off training and use our 2 months of runway to go find more data, and set a goal of "never stop pretraining"