EXHIBIT FOURTEEN

Message

From: Wojciech Galuba [@meta.com]
Sent: 1/16/2023 3:54:44 PM
To: Wojciech Galuba [@meta.com]; Marie-Anne Lachaux [@meta.com]; Guillaume Lample @meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":5948813541836446}]

Wojciech Galuba (1/16/2023 02:15:57 PST):
>Hi there, @Guillaume, for Genesis you which CC text were you using? From Edouard's pipelines? Or something else branched off of CCNet?

Wojciech Galuba (1/16/2023 02:16:49 PST):
>And coming back to this - Legal Redacted - Privilege

Guillaume Lample (1/16/2023 02:18:30 PST):
>hi, I use the CCNet of Edouard yes

Guillaume Lample (1/16/2023 02:18:47 PST):
>some people from legal pinged me directly me on Workchat to Redacted - Privilege

Guillaume Lample (1/16/2023 02:18:51 PST):
>and they also pinged Aurelien

Guillaume Lample (1/16/2023 02:18:58 PST):
>i have no idea who told them Redacted - Privilege

Guillaume Lample (1/16/2023 02:19:16 PST):
>I got approval from Antoine & Joelle to work on this so i just didn't reply to any of them

Wojciech Galuba (1/16/2023 02:20:13 PST):
>but they explicitly told you Redacted - Privilege
>(just trying to understand the constraints of working with LLMs at FAIR)

Guillaume Lample (1/16/2023 02:20:42 PST):
>let me check

Wojciech Galuba (1/16/2023 02:21:25 PST):
>So the only code that goes into WARCs that you know of would be Marie's right?
>(because CCNet uses the CC text directly)

Guillaume Lample (1/16/2023 02:22:30 PST):
>yes

Guillaume Lample (1/16/2023 02:23:01 PST):
>going from WARC should be much better that the CC text

Guillaume Lample (1/16/2023 02:23:58 PST):
>cant retrieve conversation, forgot the name of the person (i can search if it's necessary) but yeah I was told Redacted - Privilege Antoine also said this, Redacted - Privilege Redacted - Privilege

Guillaume Lample (1/16/2023 02:24:11 PST):
>which is an issue because RSC does not have internet access so you cant download lib-gen from tehre

Guillaume Lample (1/16/2023 02:24:34 PST):
>so I worked mostly on the AWS server we have for theorem proving, which conveniently has a huge amount of CPUs and where no one else has access

Wojciech Galuba (1/16/2023 02:24:36 PST):
>I'm putting together a project to do just that this week.
>Figuring out how we could to this on Spark.

Wojciech Galuba (1/16/2023 02:25:55 PST):
>Genesis code should only ever be on RSC? - sorry, not sure what you mean by Redacted - Privilege?

Guillaume Lample (1/16/2023 02:26:43 PST):
>they were referring to Redacted - Privilege

Wojciech Galuba (1/16/2023 02:27:23 PST):
>But code != data?
>Or they were Redacted - Privilege

Guillaume Lample (1/16/2023 02:27:38 PST):
>i dont know :D i think they had no idea

Wojciech Galuba (1/16/2023 02:27:44 PST):
>😀

Guillaume Lample (1/16/2023 02:27:45 PST):
Redacted - Privilege

Guillaume Lample (1/16/2023 02:27:50 PST):
Redacted - Privilege

Guillaume Lample (1/16/2023 02:28:09 PST):
Redacted - Privilege

Guillaume Lample (1/16/2023 02:28:28 PST):
>point was precisely not to make a SRT task on this, because the idea was to discretely investigate how much books help

Guillaume Lample (1/16/2023 02:28:44 PST):
>and if they help, then proceed to buying large collection of books with springer or something

Guillaume Lample (1/16/2023 02:29:37 PST):
>the springer book dataset looks quite amazing, million of books for only $2M, which is extremely cheap if you think about it

Wojciech Galuba (1/16/2023 02:29:44 PST):
>OK - let's rewind here a bit ☺ - the way we started going down this thread is I asked you if the Genesis data pipleines are anywhere in some repo - so the pipelines are not anywhere because part of the pipelines touch the book torrents and we want that under wraps?

Wojciech Galuba (1/16/2023 02:29:59 PST):
>It IS amazing

Guillaume Lample (1/16/2023 02:30:05 PST):
>Antoine & Joelle said this would help to have some evidence that books are useful before going further

Guillaume Lample (1/16/2023 02:30:13 PST):
>so this is where lib-gen came

Guillaume Lample (1/16/2023 02:30:25 PST):
>ikr. ultra clean dataset, TB of data

Guillaume Lample (1/16/2023 02:30:37 PST):
>but legal department Redacted - Privilege

Guillaume Lample (1/16/2023 02:30:58 PST):
Redacted - Privilege

Wojciech Galuba (1/16/2023 02:31:04 PST):
>That's frustrating...

Guillaume Lample (1/16/2023 02:31:11 PST):
>even though we were asking to buy a small subset of math books (barely $50K)

Guillaume Lample (1/16/2023 02:31:26 PST):
>yeah

Wojciech Galuba (1/16/2023 02:32:27 PST):
>ok got it - wouldn't there be a way to isolate the libgen part of the code? and share in a repo everything else?

Guillaume Lample (1/16/2023 02:33:37 PST):
>i guess. but tbh there is not much in this code, it's essentially some python torrenting library which processes magnet links one by one :p

Guillaume Lample (1/16/2023 02:34:04 PST):
>and a couple of things to filter the types of downloaded files (eg epub / pdf only)

Guillaume Lample (1/16/2023 02:34:12 PST):
>(epubs are trivial to parse, it's essentially HTML)

Wojciech Galuba (1/16/2023 02:35:59 PST):
>Well, I'm not interested in the libgen part of the code - just anything having to do with web data (CC, reddit etc.) that is part of the Genesis pipelines.

Guillaume Lample (1/16/2023 02:36:26 PST):
>ah, then there is nothing interesting in this code I guess

Guillaume Lample (1/16/2023 02:36:38 PST):
>everything related to web stuff is in Marie-Anne PR

Guillaume Lample (1/16/2023 02:36:43 PST):
>on the Evariste repo

Wojciech Galuba (1/16/2023 02:37:30 PST):
>Would love to:
>1. Take a look at the code to grok it a bit
>2. Understand where data lives
>3. Sit down with someone for a quick walkthrough / q&a
>4. ???
>5. Profit

Wojciech Galuba (1/16/2023 02:37:52 PST):
>OK so if it's just Evariste I can take a look at that first

Guillaume Lample (1/16/2023 02:39:00 PST):
>yes that would be nice :) maybe we can make a meeting all 3 and @Marie-Anne can give a quick tour of the code and what's happening in details

Wojciech Galuba (1/16/2023 02:39:05 PST):
>Who'd be the best person to talk about the Genesis training data sets?
>(where they live, how they were prepped, esp. CC)

Guillaume Lample (1/16/2023 02:39:23 PST):
>me I guess

Guillaume Lample (1/16/2023 02:39:33 PST):
>Edouard processed CCNet, Thibaut Github, and I processed the rest

Wojciech Galuba (1/16/2023 02:40:35 PST):
>OK, everyone seems to be branching out of Edouard's processed CCNet (Zetta, Genesis etc.)

Guillaume Lample (1/16/2023 02:40:50 PST):
>yeah

Guillaume Lample (1/16/2023 02:41:05 PST):
>yes Zetta use the exact same corpus

Wojciech Galuba (1/16/2023 02:42:52 PST):
>OK, I'll dig into the code in Evariste.
>Can I figure out the data locations from the code? Or could you share a few main places where you store it?
>Thanks!

Marie-Anne Lachaux (1/16/2023 07:54:44 PST):
>hey it's holiday in the US, i'll read and answer tmr :)