# EXHIBIT SIX

Message

**From:** Armand Joulin [[redacted]@meta.com]
**Sent:** 3/22/2023 5:11:08 PM
**To:** Sergey Edunov [[redacted]@meta.com]; Angela Fan [[redacted]@meta.com]; Hugo Touvron [[redacted]@meta.com]; Armand Joulin [[redacted]@meta.com]; Guillaume Lample [[redacted]@meta.com]; Sharan Narang [[redacted]@meta.com]
**Subject:** Message summary [{"otherUserFbId":null,"threadFbId":6404277759582213}]
**Attachments:** 334895747_734815558092396_5910972950723858311_n.png; 334880939_229569982970193_1619066219322410334_n.png; 334708285_749736926545718_7298159112984152035_n.png

Armand Joulin (3/22/2023 01:39:32 PDT):
>Hahaha amateurs

Armand Joulin (3/22/2023 01:40:12 PDT):
>by the way there are 4k instances of "i m sorry" I think it may be worth removing the dialogues that contains them. It is a bit to obvious when x2 is saying that

Guillaume Lample (3/22/2023 01:55:07 PDT):
>ah i thought we were removing those already

Armand Joulin (3/22/2023 02:20:31 PDT):
>not the "i'm sorry"

Guillaume Lample (3/22/2023 02:21:59 PDT):
>we should probably go through a bit more of these generations and detect what is typical of ChatGPT responses

Armand Joulin (3/22/2023 02:22:13 PDT):
>yes

Armand Joulin (3/22/2023 02:22:17 PDT):
>I m on it

Armand Joulin (3/22/2023 02:22:58 PDT):
>and paraphrasing seems more and more like a good idea

Armand Joulin (3/22/2023 02:31:02 PDT):

shared: 334895747_734815558092396_5910972950723858311_n.png

Armand Joulin (3/22/2023 02:31:26 PDT):
>x2 was trained without the filtering on sharegpt?

Hugo Touvron (3/22/2023 02:32:46 PDT):
>Yes the new (with filtering) model will be release in the demo this week

Hugo Touvron (3/22/2023 04:56:29 PDT):
>Today or tomorow (with openai filtering) and 3-4 days for the full filtering

Armand Joulin (3/22/2023 05:34:34 PDT):
>We are very agressive in our filtering but we are able to keep a significant fraction of the dialogues. It would be great if we could red team the model once it is trained to test if it still makes this type of mistakes.

Sergey Edunov (3/22/2023 08:49:51 PDT):
>I was thinking that just filtering might not be enough. Some of that data can leak into model during pretraining stage. And the prompts we filter out, are real and perhaps important ones. So we should probably re-label them. Would that be helpful? wdyt?

Guillaume Lample (3/22/2023 08:51:40 PDT):
>our pre-training data is not ultra recent so unless we have the latest latest dumps i dont think there is any risk of leak during pre-training. we can relabel though, to reduce risk even further

Guillaume Lample (3/22/2023 08:52:00 PDT):
>we have a few samples already in our fine-tuning dataset with "I'm a model trained by Meta AI" etc that we manually wrote

Guillaume Lample (3/22/2023 08:52:38 PDT):
>independently of that, it seems the Bard model can easily be prompted to return "I'm trained by OpenAI"

Sergey Edunov (3/22/2023 08:55:14 PDT):
>yup, this prompt that Sharan shared was pretty insane...

Guillaume Lample (3/22/2023 08:56:26 PDT):
>i believe if we red prompt the model and manually annotate the good answers on everything it failed, we are good to go !

Guillaume Lample (3/22/2023 08:57:07 PDT):
>the good thing about all these Bard failures is that it lowers the entry bar quite a bit ^^

Hugo Touvron (3/22/2023 08:57:44 PDT):
>Maybe we can plan chatGPT and Bard comparison with our model^^

Sergey Edunov (3/22/2023 08:58:31 PDT):
>maybe let's not use google's fuck up as a new bar :-P

Guillaume Lample (3/22/2023 09:00:17 PDT):
>sure, but at least this will show leadership and everyone that Meta stock will not drop by 10% because the model makes mistakes once in a while :p

Angela Fan (3/22/2023 09:05:34 PDT):
>The [redacted] deal is now approved on our side for budget, BD is finalizing the contract. This includes regular monthly dumps of web data ... I hope they can move fast and finish signing

Armand Joulin (3/22/2023 09:05:46 PDT):
>does google ToS allows us to use Bard data to pretrain? It is a way to get chatgpt data...

Armand Joulin (3/22/2023 09:07:43 PDT):
>Hugo is training a model on the filtered data and we should red team it to see if it is safer. We should expect some drop of performance in some areas, and we need to identify them (and replace data with paid annotations or be a bit smarter about our filtering -- paraphrasing for instance)

Hugo Touvron (3/22/2023 09:08:34 PDT):
>The new model with just OAI filtering tomorrow

Hugo Touvron (3/22/2023 09:08:49 PDT):
>The model with the full filtering will be ready on friday

Guillaume Lample (3/22/2023 09:13:29 PDT):
>I believe I have still never seen the [redacted] data. Is it good ?

Guillaume Lample (3/22/2023 09:18:34 PDT):

shared: 334880939_229569982970193_1619066219322410334_n.png

Guillaume Lample (3/22/2023 09:18:40 PDT):
>https://docs.google.com/spreadsheets/d/[redacted]

Guillaume Lample (3/22/2023 09:19:12 PDT):
>still dont understand why we are not making any progress on [redacted] (I ask the question to the Data meeting every week) and why we are going ahead with this

Sergey Edunov (3/22/2023 09:22:48 PDT):
>I believe [redacted] is in the process of signing the deal. I'll double check on that

Guillaume Lample (3/22/2023 09:23:14 PDT):
>okay. unless this is too late we should cancel the [redacted] deal i think

Guillaume Lample (3/22/2023 09:23:24 PDT):
>at least until we see a proper sample of it

Guillaume Lample (3/22/2023 09:23:43 PDT):
>makes no sense to ask random startups to preprocess common crawl for us

Armand Joulin (3/22/2023 09:23:54 PDT):
>these samples ar eunderwhelming. I also read that they have only 1.2T tokens.

Guillaume Lample (3/22/2023 09:23:58 PDT):
>never understood why we are dealing with them

Armand Joulin (3/22/2023 09:24:02 PDT):
>Is this accurate?

Armand Joulin (3/22/2023 09:27:32 PDT):
>I really have a hard time understanding how [redacted] aligns with our priority? Is this to replace CC data? because if it is the case, we will need 10x more data than what they can provide.

Guillaume Lample (3/22/2023 09:27:45 PDT):
>lol

Guillaume Lample (3/22/2023 09:27:46 PDT):

shared: 334708285_749736926545718_7298159112984152035_n.png

Armand Joulin (3/22/2023 09:28:03 PDT):
>if it is to benefit from their search ability, I really want for us to have access to it and evaluate it before we move forward

Armand Joulin (3/22/2023 09:28:29 PDT):
>just need clarity here

Angela Fan (3/22/2023 09:28:44 PDT):
>There are three parts of the [redacted] deal: [redacted]. I believe Robert's team evaluated (3) and Moya / Peter / Nikolay evaluated (1).

Armand Joulin (3/22/2023 09:29:01 PDT):
>how big is the web archieve?

Guillaume Lample (3/22/2023 09:29:22 PDT):
>all 3 are trivial given crawl dumps no ? and it seems they are only using that

Armand Joulin (3/22/2023 09:29:37 PDT):
>how does it compare with CC? And why is CC a problem (licencing? legal? just plain crap? :D )

Angela Fan (3/22/2023 09:30:21 PDT):
>Hm did Moya not check w/ this group when it was being evaluated?

Guillaume Lample (3/22/2023 09:30:57 PDT):
>I think we have been pretty clear from the start that we do not understand the added value of this data compared to CC

Armand Joulin (3/22/2023 09:31:01 PDT):
>I repeatedly reduced this item to P2

Guillaume Lample (3/22/2023 09:31:12 PDT):
>it seems much worse than what we have for CC

Armand Joulin (3/22/2023 09:31:16 PDT):
>(or P1 to be more nice)

Armand Joulin (3/22/2023 09:31:19 PDT):
>but not P0

Guillaume Lample (3/22/2023 09:31:39 PDT):
>how was it evaluated ?

Armand Joulin (3/22/2023 09:32:15 PDT):
>I still don't understand what problems it is resolving. I don't say that it is necessarily bad, but just that replacing CC is not the highest priority

Angela Fan (3/22/2023 09:34:37 PDT):
>Alright, so let's disentangle these issues:
>- We believe that we did not prioritize our licensing accurately, and we should more heavily prioritize textbooks, fiction content, etc. I believe that is now the case, and BD is focusing on deals with [redacted]
>
>- We need to have more communication on the team, and the pre-training team needs to be more looped into data evaluation.
>
>- We believe that [redacted] does not provide substantially more data than Commoncrawl --- we should investigate this --- and we are unclear if the data workstream is actively downloading and processing Commoncrawl (which they absolutely absolutely should be).

Angela Fan (3/22/2023 09:34:53 PDT):
>What's the best way to address issue #2? That seems to be an important one

Guillaume Lample (3/22/2023 09:35:48 PDT):
>I think there is enough communication on 2) and that we gave extensive feedback on this dataset

Armand Joulin (3/22/2023 09:36:03 PDT):
>The pretraining team is in the data meeting and they were mentoring going on on some of the tasks (evaluating 2B tokens from [redacted] for instance) but maybe I didn't enforced it enough.

Armand Joulin (3/22/2023 09:36:29 PDT):
>We have never seen the [redacted] data to an extend where it can be properly evaluated

Guillaume Lample (3/22/2023 09:36:35 PDT):
>i believe separating the Data and Pretraining team was maybe not a good idea

Guillaume Lample (3/22/2023 09:36:49 PDT):
>data is an essential part of pretraining

Guillaume Lample (3/22/2023 09:37:06 PDT):
>there are so many critical things nobody is working on

Armand Joulin (3/22/2023 09:37:21 PDT):
>so far: we have been discussing the new CC pipeline (communication works very well there), the stack preprocessing (lost track of where this is.. need to circle back) and [redacted] where eval were unconclusive (but expected considering the sample size)

Guillaume Lample (3/22/2023 09:37:23 PDT):
>like, OCR for instance, to get clean content from PDF

Armand Joulin (3/22/2023 09:37:44 PDT):
>agree. I made a mistake with that split

Sergey Edunov (3/22/2023 09:41:39 PDT):
>> [redacted] where eval were unconclusive (but expected considering the sample size)
>
>The evals were positive, no? It was just expensive. Or did I misunderstand something?
>
>Btw, I checked on that, we are currently waiting for [redacted] to get back to us with a quote. Last update was yesterday.

Guillaume Lample (3/22/2023 09:42:18 PDT):
>eval didn't make a big difference because dataset was tiny

Guillaume Lample (3/22/2023 09:42:21 PDT):
>but data was really good

Armand Joulin (3/22/2023 09:43:35 PDT):
>It is very hard to get signal from 2B tokens + IC may not have follow the best experimental setup to ge tthat signal (but again probably impossible to star tweith)

Sharan Narang (3/22/2023 09:55:14 PDT):
>+1, it's really hard to get signal from 2B tokens. For data, we may have rely on inspection especially if we get such small samples.

Angela Fan (3/22/2023 10:02:02 PDT):
>Yeah. I agree. I talked to Armand about this this morning, but for these small samples, we should really move quickly and focus on eyeballing and rough measurements and just make a fast call.

Sharan Narang (3/22/2023 10:10:43 PDT):
>Also, we should really push legal to [redacted]

Redacted

Angela Fan (3/22/2023 10:11:08 PDT):
>i look into it