WOODHOUSE EXHIBIT 4

REDACTED

EXHIBIT C

From: Guillaume Lample[/O=THEFACEBOOK/OU=EXTERNAL (FYDIBOHF25SPDLT)/CN=RECIPIENTS/CN=D5DFE144828C4CF7980638CE3E199ADF]

Sent: Tue 2/28/2023 11:01:12 PM (UTC)

To: Melanie Kambadur[redacted]@meta.com]; Xavier Martinet[redacted]@meta.com]; Nikolay Bashlykov[redacted]@meta.com]; Peter Albert[redacted]@meta.com]; Moya Chen[redacted]@meta.com]; Marie-Anne Lachaux[redacted]@meta.com]; Guillaume Lample[redacted]@meta.com]

Subject: Message summary [{"otherUserFbId":null,"threadFbId":8865995773473436}]

Attachment: 332846787_9020492504659760_2751155617077653862_n.png

Attachment: 332530197_533621921983081_4735798491804002938_n.png

Attachment: 332024563_230609189352290_3833951687220188647_n.png

Moya Chen (2/28/2023 10:14:04 PST):
>Starting a thread with folks that might have opinion on the [redacted]data acquisition.
>
>Context here is that @Nikolay did the initial ingestion + analysis of data, and while the data that we've got is high quality, there's only going to be maybe about 1b usable tokens.

Moya Chen (2/28/2023 10:14:06 PST):

shared: 332846787_9020492504659760_2751155617077653862_n.png

Moya Chen (2/28/2023 10:14:34 PST):

shared: 332530197_533621921983081_4735798491804002938_n.png

Melanie Kambadur (2/28/2023 10:16:13 PST):
>Books data acquisition perhaps? Since we are also talking to scribd and others?

Guillaume Lample (2/28/2023 10:20:03 PST):
>1B usable tokens seems very small

Guillaume Lample (2/28/2023 10:20:27 PST):
>books3 + gutenberg is 30B and it's quite small already

Peter Albert (2/28/2023 10:20:34 PST):
>I think the books seemed like good quality, so I don't think there will be issues on that front. So I don't think we need another sample dataset.
>
> The following things would be interested to know:
>- how many books in total can they offer us (if it is way to small it might not make sense for us. For comparison the bookcorpus has about 11k books, but mostly fiction)
>- how is the distribution between books and articles? For articles we can already get Pubmed and arxiv for free, so I think mostly the long form books are interesting
>- how is the distribution of topics

Guillaume Lample (2/28/2023 10:20:41 PST):
>are we talking about 1B for [redacted]or just what they sent us ?

Peter Albert (2/28/2023 10:20:50 PST):
>just what they sent us i think

Moya Chen (2/28/2023 10:21:11 PST):
>cynical thought, but how much do we think they would do something like only send us the *good* quality books that they have

Nikolay Bashlykov (2/28/2023 10:21:28 PST):
>the dataset has both ebooks and journals (scientific articles): 380k journals, 170k books chapters. Both are well structured and have high quality texts.

Guillaume Lample (2/28/2023 10:21:49 PST):
>i was afraid of this, but given that they sent us many books with just a few chapters here and there, it reduces this probability a bit

Nikolay Bashlykov (2/28/2023 10:22:05 PST):
>* it's the sample that we have

Nikolay Bashlykov (2/28/2023 10:22:44 PST):
>just what they sent us

Guillaume Lample (2/28/2023 10:22:46 PST):
>what was the fraction again? of this sample among the full size ?

Nikolay Bashlykov (2/28/2023 10:23:19 PST):
>I don't have data on that ☹

Melanie Kambadur (2/28/2023 10:24:28 PST):
>I wasn't helping with early [redacted]convos, just have this doc:
https://docs.google.com/document/d/[redacted]

Moya Chen (2/28/2023 10:24:40 PST):
>access requested

Melanie Kambadur (2/28/2023 10:24:43 PST):
>But we can look into this with BD

Guillaume Lample (2/28/2023 10:24:49 PST):
>no access

Melanie Kambadur (2/28/2023 10:25:12 PST):
>Ah don't have sharing. It's Faisal's doc let me copy

Peter Albert (2/28/2023 10:25:23 PST):
>Books3 is 197k books for comparison

Melanie Kambadur (2/28/2023 10:25:47 PST):
>https://docs.google.com/document/[redacted]

Peter Albert (2/28/2023 10:28:55 PST):
>Mhm, the document doesn't say how many books they have in total. Also for pricing its 200 books for 60k dollars?

Moya Chen (2/28/2023 10:30:16 PST):
>actually as a thought - do we have any [redacted]enforced deadlines to make a decision here? Given how long it took them to get back to us, I'm guessing no, right?

Guillaume Lample (2/28/2023 10:30:35 PST):
>seems unreasonably expensive

Moya Chen (2/28/2023 10:30:38 PST):
>Kind of half wondering if we can just decide to... hold off and wait and see if bizdev can make contact with other books sources

Melanie Kambadur (2/28/2023 10:30:48 PST):
>No not at all but they are incredibly slow to respond to each email and make meetings and they take like 4+ weeks to deliver data

Moya Chen (2/28/2023 10:30:49 PST):
>cause this lol

Melanie Kambadur (2/28/2023 10:31:14 PST):
>Yes and this is the not commercial friendly price and we pay per year

Guillaume Lample (2/28/2023 10:31:42 PST):
>how much would it be for commercial friendly?

Guillaume Lample (2/28/2023 10:31:49 PST):
>that's absurd. books do not cost $300

Moya Chen (2/28/2023 10:32:32 PST):
>...when you're a publisher that makes its income off of charging large institutions for access to scientific data... (like I think companies like [redacted] is part of the reason why arxiv exists, yeah?)

Moya Chen (2/28/2023 10:33:20 PST):
>(anyway that snark is prolly note useful)

Melanie Kambadur (2/28/2023 10:33:56 PST):
>Ah forgot @Xavier was involved with original [redacted] deal

Peter Albert (2/28/2023 10:34:10 PST):
>I think if we get their whole dataset (and its like 30x larger than the sample, which was only maths books) for a few million it would make sense as these are high quality long books, that we propably won't get somewhere else.

Melanie Kambadur (2/28/2023 10:34:29 PST):

shared: 332024563_230609189352290_3833951687220188647_n.png

Melanie Kambadur (2/28/2023 10:34:30 PST):
>Also dig up this email (again research prices I believe)

Peter Albert (2/28/2023 10:41:25 PST):
>So 200k dollars per year for 3.84 million books/articles (even if only some percentage are books and the rest are articles) seems good, even if the commercial license is more expensive.

Xavier Martinet (2/28/2023 10:50:03 PST):
>My 2 (x10^8) cents: with the current focus, are the [redacted]books really necessary? shouldn't we take $1M and buy as much as we can from Barnes & Nobles instead ?

Moya Chen (2/28/2023 10:50:36 PST):
>...copywrite?

Nikolay Bashlykov (2/28/2023 10:50:47 PST):
>in the sample that we have we have there are ~8k books and ~70k journals, so books are ~10%. I the

whole dataset (3.8M according to the screenshot above) has the same ratio, than we would get ~380k books, which is quite good

Moya Chen (2/28/2023 10:50:48 PST):
>*right

Xavier Martinet (2/28/2023 10:50:48 PST):
>fair use, no?

Moya Chen (2/28/2023 10:50:56 PST):
>legal gray area

Xavier Martinet (2/28/2023 10:51:25 PST):
>this is why they set up this gen ai org for: so we can be less risk averse

Xavier Martinet (2/28/2023 10:52:20 PST):
>my opinion would be (in the line of "ask for forgiveness, not for permission"): we try to acquire the books and escalate it to execs so they make the call

Peter Albert (2/28/2023 10:53:17 PST):
>So what is the idea here? To buy ebooks from barnes and nobles?

Moya Chen (2/28/2023 10:53:21 PST):
>mmm i think it's one thing if it were a new area where we're not sure yet... I think it's another thing when there's already been lawsuites in the pipeline

Moya Chen (2/28/2023 10:53:39 PST):
>notably

Moya Chen (2/28/2023 10:53:40 PST):
>https://www.theverge.com/2023/1/28/23575919/microsoft-openai-github-dismiss-copilot-ai-copyright-lawsuit

Moya Chen (2/28/2023 10:53:59 PST):
>(it's about code data rather than books, but I'm guessing that's what the lawyers are watching)

Xavier Martinet (2/28/2023 10:54:03 PST):
>sure, but have the courts ruled that the tech company were guilty?

Xavier Martinet (2/28/2023 10:54:40 PST):
>I mean, worst case: we found out it is finally ok, while a gazillion start up just pirated tons of books on bittorrent

Moya Chen (2/28/2023 10:55:08 PST):
>i'd maybe agree with you, but given how FB also threw idk $2 mil or something like that at shutterstock to not deal with potentially getting sued

Moya Chen (2/28/2023 10:55:28 PST):
>regardless.. I think @Melanie Kambadur's been the one in contact with the lawyers

Moya Chen (2/28/2023 10:55:45 PST):
>and would have best context for what sort of appetite the org has for asking for forgiveness, not permission

Moya Chen (2/28/2023 10:56:18 PST):
>but for better or for worse -- yeah, it is a trend that startups tend to have more leeway on this

Xavier Martinet (2/28/2023 10:56:20 PST):
>my 2 cents again: trying to have deals with publishers direcly takes a long time, as the [redacted] shows (we started in September if I remember correctly). So better buy books at retailers directly

Moya Chen (2/28/2023 10:56:38 PST):
>cause the established companies are worth more $$ to patent trolls

Peter Albert (2/28/2023 10:59:06 PST):
>but isn't that more expensive than making a deal? 10$ * 3 million books is more than what we are paying [redacted]

Melanie Kambadur (2/28/2023 11:00:05 PST):
>Yeah we definitely need to get licenses or get approvals on publicly available data still

Melanie Kambadur (2/28/2023 11:00:34 PST):
>difference now is we have more money, more lawyers, more bizdev help, ability to fast track/escalate for speed, and lawyers are being a bit less conservative on approvals.

Xavier Martinet (2/28/2023 11:02:52 PST):
>There is a trade off between speed and money efficiency I suppose

Melanie Kambadur (2/28/2023 15:01:12 PST):
>Know it's not quite books but posted a question on PDF docs here:
https://fb.workplace.com[redacted]