EXHIBIT NINE

Message

From: David Esiobu [@meta.com]
Sent: 9/26/2023 6:51:39 AM
To: Melanie Kambadur [@meta.com]; David Esiobu [@meta.com]; Dieuwke Hupkes [@meta.com]; Sharan Narang [@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":6373379592791139}]

Dieuwke Hupkes (9/25/2023 06:04:54 PDT):
>Hi all,
>
>As promised, I drafted up a few potential follow-up questions for memorisation. As is written in the doc, it is currently mostly a list of potential questions that we could address with some arguments for why they may be interesting. My goal is to get this into a working doc + planning hopefully by the end of the week, and I am hoping for your help with prioritisation and timeline.
>
>https://docs.google.com/document/d/

Dieuwke Hupkes (9/25/2023 06:05:07 PDT):
>#shareadoc edit

Dieuwke Hupkes (9/25/2023 06:05:43 PDT):
>Some points that I'd specifically like your input on are:

Dieuwke Hupkes (9/25/2023 06:06:48 PDT):
>Could any of these questions be relevant to inform decisions during model training, or be helpful for model releases in other ways? If so: *when should they be delivered? For our previous results I think we were a bit too late, so it would be good to align our planning with the pretraining/release plannings

Dieuwke Hupkes (9/25/2023 06:08:24 PDT):
>Are there any other folks that I / we should reach out to to ask for their input? For instance, I chat sometimes with the public policy director of Benelux (Belgium Netherlands Luxembourgh), and he is usually very interested in these results and even uses them to seed arguments in government discussions. Could it make sense to reach out to them as well to ask if there are things that are particularly interesting for them? Or legal folks?

Dieuwke Hupkes (9/25/2023 06:08:34 PDT):
>Are there any categories / types of questions missing from this doc?

David Esiobu (9/25/2023 11:51:21 PDT):
>Thanks Dieuwke! The current cutoff for data mix changes for LLaMA-3 is Oct-9. For the short term, there are a couple of targeted analyses we can aim to have done before then:
>* What are the memorization rates in the ongoing LLaMA-3 test runs (https://docs.google.com/spreadsheets/d/)
>* What is the relative risk of epoching on high quality datasets (libgen, math, code), building on Nikolay's training experiments (https://docs.google.com/document/d/1)
>* What is the relative risk of including non-permissive github? The data and training runs for this will likely not be ready in time for Oct-9, but we can try do some code-specific testing to help inform that
>
>i'll write out the above in more detail and schedule some time for us to discuss this week

Melanie Kambadur (9/25/2023 11:54:01 PDT):
>yes, added some comments to your docs, thanks both for working on this area. I am particularly interested in focusing in on Redacted

Redacted

Melanie Kambadur (9/25/2023 11:54:40 PDT):
>technically we may have a little time than october 9th as we could always make last minute changes to the data mix before the run, but if we can figure out by october 9th that would be better.

Sharan Narang (9/25/2023 12:04:42 PDT):
>One somewhat tangential question - have we deduped across datasets for Llama 3?

David Esiobu (9/25/2023 12:15:33 PDT):
>i'll have to double-check with Jacob, but i don't think we've done more than URL dedup across datasets

Sharan Narang (9/25/2023 12:18:29 PDT):
>IMO, that might be high priority. Even some form of near dedup (so not exact match) but Melanie should chime in since she has more visibility of the data workstream.

Dieuwke Hupkes (9/25/2023 23:50:14 PDT):
>Great thank you! Let's schedule a meeting soon to work out a plan together.

Dieuwke Hupkes (9/25/2023 23:51:05 PDT):
>We can try to hit this, but I'm visiting Paris next week and will be pretty busy with meetings etc, Aurelien has his team offsite then. This is also why I am trying to reach out early this time so that we have a bit more time for the next round.

Dieuwke Hupkes (9/25/2023 23:51:39 PDT):
>I think Nikolay has been working on some experiments concerning this too? I asked him for some input on the MME doc, he also said he was planning on using the MME for something but I don't recall exactly what