Vo Declaration
Exhibit K

EXHIBIT A

Email Title: [A/C Priv] Request for decision: LibGen Escalation OneLLM

To: Joelle Pineau
From: Sony Theakanath
CC: [HYPERLINK "mailto[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]l@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h],[HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h]

[A/C PRIV]

Hi Joelle,

The team is writing for a decision on proceeding with using libgen for training OneLLM after guidance from Legal and Policy. After a prior escalation to MZ, GenAI has been approved to use LibGen for Llama 3 (with VP sponsor requiring to accept full risk) with a number of agreed upon mitigations. Upon discussion with Legal, the OneLLM team was asked to escalate to you for use of the Libgen dataset including implementation of said mitigations (see below).

Libgen is essential to meet SOTA numbers across all categories, and it is known that OpenAI and Mistral are using the library for their models (through word of mouth). Without LibGen, research scientists within OneLLM believe it would not be able to achieve the SOTA numbers the industry shows. We're beating Llama V2 7B across most benchmarks and are close to their 30B benchmarks, but are not able to reach Mistral. Results from experiments in GenAI suggest that adding Libgen would significantly close this gap, and possibly get us to SOTA.

Model	**MMLU**	**GSM8k**	**MATH**	**BoolQ**	**PiQA**	**SiQA**	**Winogrande**	**O**
Llama V2 7B *Text only*	45.30	14.6	2.50	77.40	78.80	48.30	69.20	58
Mistral 7B **(using Libgen)** *Text only*	60.10	52.10	13.10	83.00	-	75.30	-	8
OneLLM 7B (using Llama 2 data) *Image and text*	46.83	32.98	6.58	82.11	79.43	57.52	69.77	50

The team presents four options. In **no case** would we disclose publicly that we had trained on libgen, however there is practical risk external parties could deduce our use of this dataset, especially if we release models trained on it (e.g. option #3).

Option 1: Use libgen internally only, without external publication of numbers or external model release.
The team would use libgen for internal evaluation, and the team would not release the libgen-trained model externally or publish results in a paper, discuss externally, or allow usage of the model internally, e.g. for production use cases. Purpose is to show internally that we can reach SOTA numbers for MMLU for 7B, provided we have the right datasets, such as LibGen.
Risk: Medium
Requires no acceptance of risk and just an FYI.

Option 2: Use libgen, and publish benchmarked numbers on blog post.
The team would use libgen for internal evaluation and publish results externally. For extreme clarity, this would mean we would publish benchmarks similar to the table above. Additionally, we would not disclose use of LibGen datasets used to train, but there is a possibility that external parties may deduce that we are using libgen. Please note that we will not open source a model with Libgen.
Risk: Medium-High
[Recommended by Research Scientists]
Requires Joelle to accept medium-high legal risks (with mitigations) and high policy risk.

Option 3: Use libgen, publish benchmarked numbers on blog post, and open source a model trained on LibGen.
The team would use libgen for internal evaluation and publish results externally. We would also either research license / commercial license a model. We would not disclose use of Libgen datasets used to train, but there is a possibility that external parties may deduce that we are using libgen.
Risk: Medium-High
Requires Joelle to accept Medium-high legal risk (with mitigations) and high policy risk .

Option 4: Do not proceed with libgen.
Risk: Low

Legal and policy ([HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h]) have outlined the legal and policy risks and mitigations:

<u>**Legal Risk**</u>:

Redacted

Redacted

Mitigations:

1. Remove data clearly marked as pirated/stolen
2. Do not externally cite the use of any training data including LibGen
3. Perform evaluations and appropriate mitigations as necessary to reduce memorization metrics (e.g. propensity to generate outputs containing copyrighted material). The memorization rates should meet the same thresholds achieved by Llama 2 team for which leadership has already accepted the risk.

Policy Risks:

1. Legislative/lobbying:
    - Risks:
        - Copyright and IP is top of mind for legislators around the world, including in the US and the EU.
        - US legislators expressed concern in a recent hearing about AI developers using pirated websites for training. It's unclear what their legislative actions would be if the concern spreads, but it reflects some of the negative lobbying rights holders have been doing, related to our litigation on this topic (along the lines that this is "stolen" content that then taints the output of the model).

        - If there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues.
        - The recently agreed EU AI Act will include measures requiring disclosure of training data, and the right for rights holders to opt out of their material being used for training AI models. The latter is expected to extend beyond the EU's borders, meaning rights holders outside of Europe (many of whose works are in LibGen) could leverage this clause. **The penalty for non-compliance related to requirements on data is up to €35m or 7% of the total worldwide annual turnover of the preceding financial year (whichever is higher).**

2. Misuse:
    a. Risks:
        - A secondary concern is about the potential malicious applications of a model trained on large quantities of scientific data. A 2022 [HYPERLINK "https://www.nature.com/articles/s42256-022-00465-9.epdf?sharing_token=I84MQc16g0O-kOA6bzPRzdRgN0jAjWeI9jnR3ZoTv0M6VuGuVWKcBJFL5U5ocXOA5zcnGmZOUPQzouuai7vI0XuOG1hxcfSUpHakkMxyD1NjtXRFBgFxUa9ZQI7okPtQc-7YkJa4BSKUXZqV75Cr1BQONFfkK_B6nn67L7Rh7c_gK5wvkMkmpSea3sj2flJNMKSj4uWwVONITDnES4qbG2V5jcdfJi6QkMWbQIgc0YE%3D&tracking_referrer=www.theverge.com" \h] on the risk of misuse of AI drug discovery models was widely covered in the [HYPERLINK "https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx" \h], and subsequent [HYPERLINK "https://arxiv.org/pdf/2306.03809.pdf" \h] have highlighted that training on large corpuses of medical data can elevate risks relating to bioweapons and CBRNE. One [HYPERLINK "https://arxiv.org/pdf/2306.03809.pdf" \h] is to remove some of the data that would enable these capabilities (e.g. toxicity datasets), which represented a fraction of the overall data in the paper in question ([HYPERLINK "https://arxiv.org/pdf/2306.03809.pdf" \h]).
        - If the model is to be used internally for evaluation only, then this risk may not warrant further action. However, if a model trained on LibGen is to be used for anything other than internal evaluation then it would need to be red teamed for these risks.
    - Potential mitigations:
        - In its [HYPERLINK "https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf" \h], Anthropic proposes containment measures for models that could present a risk of 'catastrophic misuse' (see ASL 3) which include internal controls on who can access information about the model. OpenAI [HYPERLINK

"https://cdn.openai.com/openai-preparedness-framework-beta.pdf" \h] in its Preparedness policy. This might be something to consider even for a version that is for internal evaluation only.

- If we were to use models trained on LibGen for a purpose other than internal evaluation, we would need to red team those models for bioweapons and CBRNE risks to ensure we understand and have mitigated risks that may arise from the scientific literature in LibGen.
    - As well as providing assurances, this would be a way to demonstrate that we have upheld commitments to developing and deploying AI responsibly, in the event that we are forced by future legislation to disclose that we used LibGen.
- We might also consider filtering the dataset to reduce risks relating to both bioweapons and CBRNE, and also based on the year that a publication was added to reduce the risk that it is trained on papers that include harmful stereotypes. It is also possible that a model trained on LibGen may produce inaccurate scientific sounding information, and this is again something we should test for.

Decision needed: Which option should the team go forward with?

Thanks,
Sony