SafeLMM: Safe Large Multimodal Models By Design

Huu Nguyen, Robert Kaczmarczyk, Anna Rogers, Bo Li, Ludwig Schmidt, Rio Yokota, Marianna Nezhurina, Liangyu Chen, Marzena Karpinska, Taishi Nakamura, Tommaso Furlanello, Tanmay Laud, Giovanni Puccetti, Xiaozhe Yao, Dung Nguyen, Qi Sun, Aleksandr Drozd, Paulo Villegas, Gabriel Ilharco Magalhaes, Mitchell Wortsman, Weiyang Liu, Christoph Schuhmann, Kenneth Heafield, Jenia Jitsev

This whitepaper is a proposal for a project to train multimodal, multilingual foundation models that are safe by design. It is an expanded version of a proposal originally sent to the Euro HPC and we publish it for discussion with the community. Please send comments to: engage@ontocord.ai.

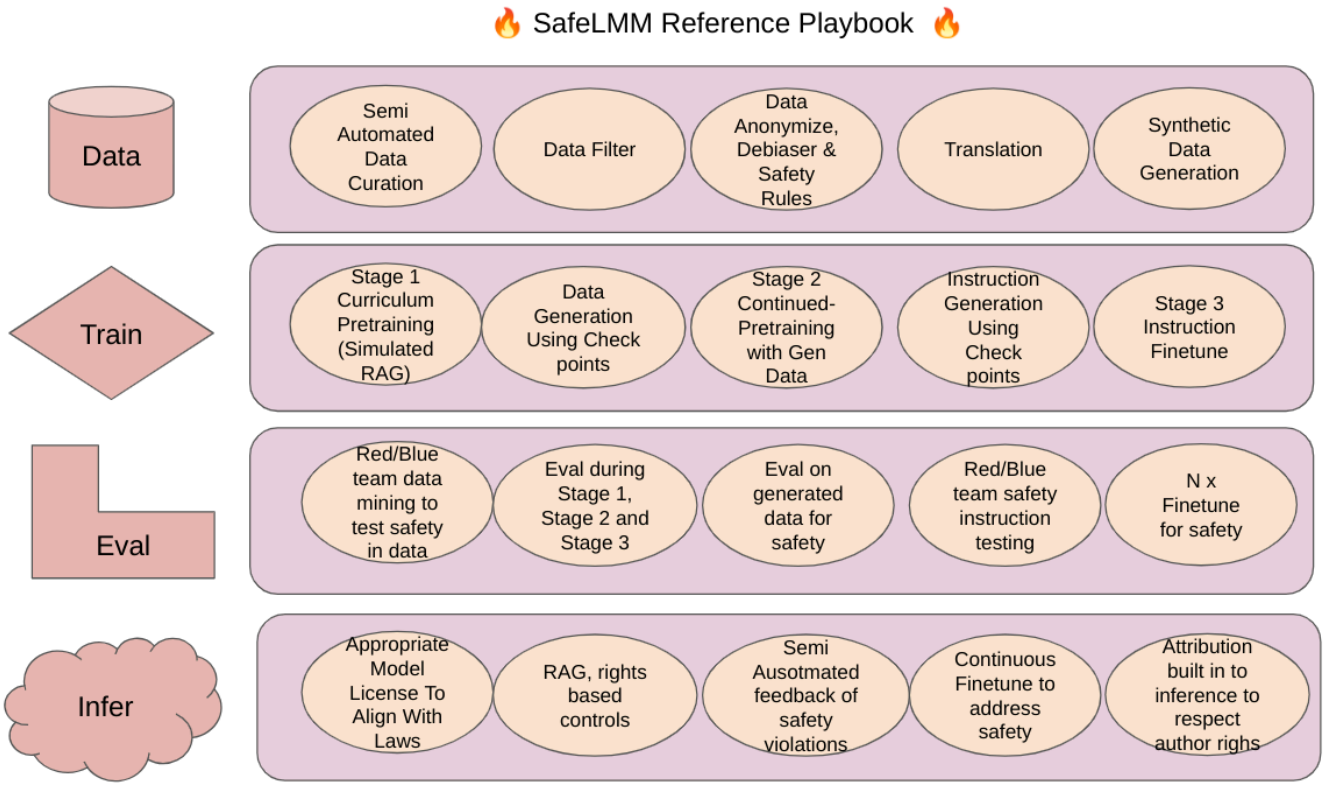
1 Key scientific/societal/technological contribution of the proposal

The Synthetic Augmented data, Fair and Extreme-scaled Large Multimodal Model (SafeLMM) project will redefine the AI landscape by pioneering next-generation multimodal models that emphasise ethical and regulatory compliance. Using the Leonardo supercomputer and in collaboration with Ontocord AI, PIISA.org, LAION, e.v., Juelich Supercomputing Center, Horizon Europe project HPLT, and Efficient Translation Limited, among others, the SafeLMM models, ranging from 7B to 34B paramasters, will harness vast amounts of detoxified synthetic data and open and permissively licensed real data spanning images and text in 31 languages to address compliance with regulations. Key project contributions:

- **Innovative Modelling:** Implementing multimodal architectures for potentially enabling zero-shot transfer across modes, languages, and domains.
- **Safe Content:** Crafting content that strictly adheres to ethical and regulatory guidelines, addressing bias, toxicity, and privacy concerns.

- **Robust Governance:** Incorporating data provenance, safety filtering, and attribution systems during training and inference.
- **Open-Access:** Developing and sharing comprehensive documentation, data, models, tools and libraries, promoting transparency and collaboration in AI research.

By leveraging High Performance Computing, SafeLMM offers a fusion of synthetic data, multimodal capabilities and responsible AI practices. The ambition is not just high-performance models, but ones that carry the stamp of scientific, societal, and technological excellence. At a glance:



2 Detailed proposal information

2.1 Justification for the importance of the scientific problem and the requested resources

SafeLMM

It is well known that GPT-4V and even open source multimodal models have performed phenomenal multimodal and multilingual language understanding, pushing the state of the art on an almost daily basis. However, the intersection of safety, multimodal and multilingual research is lacking and in particular, building large scale models that are safe by design is a challenge of our time, and crucial for the EU and humanity in general. We have brought together arguably some of

the best safety, multimodal and multilingual open science and open source researchers together to help address this issue. SafeLMM is poised to bridge this chasm by creating a family of language models that offer comprehensive support for the 24 official European languages along with seven (7) primary languages in other language families (Zh, Hi, Id, Vi, Ja, Ko, Ar) along with multimodal abilities, ultimately surpassing the capabilities of the world's most extensive open models. Beyond sheer performance, the project sets the stage for uncharted innovation and the creation of model variations tailored to specialised needs.

Safety

What Does Safety Mean? We will address the potential EU AI Act, potential copyright issues and the existing GDPR. We also look to the UDHR as our ethical northstar, similar to our work in (Jernite, 2022). We understand that this project is particularly EU focused and other laws and ethical standards may differ, and we acknowledge this as one limitation. Given this framing, for SafeLMM, "safety" means, using copyright free or copyright permissive text, de-biasing, and removing toxic content, tagging NSFW, protecting privacy, and using our best efforts to remove hate speech and to remove illegal text. This work aims no less than to create industry standards for self-regulating open source large foundational models.

Privacy of EU Data Subjects. Generally, we believe there is a right under freedom of expression and information under the GDPR to train and research large language models, and in particular training on newspaper articles and encyclopaedia text. However, to minimise downstream risks to users of our models, we will use our best efforts to mitigate privacy issues as discussed in Sec. 2.2.

Mitigation of Copyright Issues. We will use our best efforts to gather public domain data or data that has been permissively licensed, such as CC-0, CC-BY, CC-BY-SA, and non GPL open source licensed data. For images, we will respect opt-out preferences from spawning.ai's API in each new release of our datasets. This will not however, remove content in already trained models.

Potentially Illegal Content. We recognize that some countries have generally prohibited certain types of content such as child sexual abuse materials and certain types of hate speech. We will use our best efforts not to include such data in our dataset or training.

Potential EU AI Act Issues. Because the EU AI Act has not been agreed upon nor enacted by member states, we will have data and training procedures that are directed to general principles in the proposed Act. Moreover, this research will hopefully promote a self regulatory regime in the Open Source and research community, with a reference implementation. First, all of the process for data gathering, training and deployment under the SafeLMM project will be open for inspection and thus transparent, another EU AI Act theme. Open science promotes safety. Prohibited activities are also one of the themes of the EU AI Act. So we will licence our models under a modified Apache 2 licence with prohibited use provisions similar to our work under the BLOOM Rail licence. In addition to licence restrictions, we will use technical measures as discussed in Sec 2.2 to address potential risks. As a policy issue, the EU AI Act may also regulate general purpose generative AI, and our data provenance tools could lower the documentation burden of adopters of SafeLMM.

Training Scale

It is worth highlighting that training models of this magnitude represents a formidable scientific challenge in itself. Such an undertaking demands innovation in data generation, preprocessing, training software and infrastructure, parallel computing, and model evaluation. It necessitates access to millions of GPU hours and is unattainable without the computational capabilities offered by supercomputers like Leonardo. These state-of-the-art supercomputers, equipped with advanced GPU compute resources and a high degree of parallelism, are indispensable for training models of this scale within reasonable timeframes. The SafeLMM project will generate synthetic data and train

very large models of 7 billion, 13 billion, and 34 billion parameters, each trained on approximately 2 trillion tokens each, using approximately 4.2 million GPU-hours. As a comparison, we aim to replicate the performance of the recent LlaMA 2's scale of data and training (Touvron, 2023) except with multimodal, multilingual and safe by design data.

Deliverables

The SafeLMM project recognizes the profound scientific, technical, and societal implications of its mission. It seeks to create multimodal and multilingual models that align with ethical and regulatory standards while pushing the boundaries of AI research and application, ultimately contributing to transparent and trustworthy AI practices on a global scale. Appropriate data and model governance policies will be put in place, also with the creation of a governing board, to ensure our ethical and regulatory objectives are met (Jernite, 2022), and we will incorporate an AI impact analysis (including environmental impact) and community input to guide our governance efforts.

2.2 Overview of the project

Motivation, Objectives and Scientific Challenges

In recent years, the AI community has witnessed remarkable advancements driven by deep neural network models trained on vast datasets. These models have revolutionised various artificial intelligence (AI) domains, ranging from computer vision, audio, and video processing to natural language processing and understanding (NLP/NLU). The success of this paradigm in AI system development has heavily relied on access to state-of-the-art high-performance computing (HPC) resources, and many groundbreaking innovations have been made possible by tech giants in the United States and China. However, concerns have been raised about the centralised control of core AI technology in the hands of a few commercial entities, leading to restrictions on academic and industrial access and raising issues regarding transparency and democratic access.

Prior Work

Our teams at LAION, JSC, HPLT, Ontocord, PIISA.org, Efficient Translation Limited, among others, have already been instrumental in demonstrating numerous successful applications of various deep supervised, unsupervised and reinforcement learning techniques on different domains (Clavijo, 2021; Eslamibidgoli, 2021; Brenzke, 2021) as well as training very large language models (Scao, 2022), governance and construction of large datasets (Jernite 2022; Laurençon, 2022; TogetherComputer, 2023), instruction tuned models (Köpf, 2023; Li, 2023), and safety and trustworthiness evaluations (Wang, 2023).

The researchers taking part in the project have long-standing profound experience in using supercomputers for conducting large-scale multi-modal and multi-lingual learning experiments. Dr. Dr. Jitsev was serving as PI for numerous successful national large-scale compute grant applications in frame of Gauss Center for Supercomputing, Germany, and also as PI for large-scale compute grant application INCITE for Summit supercomputer at Oak Ridge National Laboratories, USA. Dr. Heafield also served as PI for the HPLT project, overseeing 12M GPUh allocation on LUMI-G. Other members of the team participated in training the 175B parameter models on Jean Zay HPC as part of the Big Science project.

Our efforts have been also put forward to investigate the behaviour of models of different sizes obtained via large-scale distributed training on supercomputers using various large datasets, especially with regard to the impact of scale on generalisation and transferability of the pre-trained models across various conditions (Cherti, 2021; Schuhmann, 2021). These efforts also led to establishing baselines for large-scale distributed training across many nodes using multiple GPUs with different neural network architectures and various types of learning. Moreover, Large language models show strong generalizability after pretraining on general massive text corpus (Touvron,

2023). Recent work on multi-modal image-text learning showed substantial robustness to distribution shift (Radford, 2021).

Our work in data governance in BigScience and BigCode informs SafeLMM's data governance regime. See (Jernite, 2022). We explored managing diverse stakeholder interests, grounding the ethics of our work in first principles, striving to comply with regulations, and adopting a model licence that prohibits certain usage. Also, similar to our work in Wang, 2023, we will perform rigorous safety and trustworthiness evaluations on the SafeLMM models. However, we will enhance the evaluations to include multimodal safety contexts, including fairness, bias, privacy, and toxicity evaluations.

Also our many previous work in machine translations informs our approach to privacy preserving scalable machine translations. Bogoychev, Nikolay (2021). Our most recent work in the HPLT project also will assist us in addressing multilingual fairness. Following on from the work above, we propose to study and produce high quality multimodal and multilingual models that are designed to respect copyrights, minimise biases, remove toxicity from input and output content, and protect against potential privacy violations. And we will do so at large scale to deliver the most performant models.

LAION Large-scale multimodal datasets.

Multi-modal language-vision models like CLIP are trained on hundreds of millions of image-text pairs and then show remarkable capability to perform zero- or few-shot learning and transfer even in the absence of per-sample labels on target image data. Despite this trend, until the release of our LAION-400M dataset (Schuhmann, 2021) there had been no publicly available dataset of sufficient scale for training such models from scratch and studying their transfer capabilities, also with regard to the effect of training scale. To address this issue, we built and released many large datasets. Moreover, We will use our large scale datasets and those from other projects such as all of Wikipedia images to train SafeLMM

Deep Neural Networks, Transformers, Autoregressive Models and Retrieval

We will use a standard setup for training the models. For the network architecture, we choose different established variants of convolutional (advanced ResNets) or transformer neural networks that are standard backbones for deep learning. The pre-training phase consists in iterating through the corresponding large full source dataset in multiple epochs. We will use the AdamW optimizer with typical learning rate decays. In general, training is done using tensors of simple float precision (32 bits) and mixed precision (bf16). Mixed precision can speed up training substantially, while requiring additional mechanisms to avoid training instability. We chose the above computational methods because they have been well used in the industry to train large models, and we actively contribute to open source standards software such as OpenLM and FMengine. After pre-training, in our finetuning stage will also use low-rank-adaptation similar to Emu (Sun, 2023), SEED (Ge, 2023). Similar to SILO, we will provide a mechanism to use the well known retrieval augmented generation paradigm at inference time using the SILO open source code (Min, 2023), as an additional mechanism to enhance functionality.

Safety

Our main technique for scalable oversight is data filtering and data augmentation, conditional pretraining and instruction tuning.

Privacy

From a privacy perspective, we will rely heavily on filtering. Out of copyright data (published more than 100 years ago) by their nature will be almost certainly about deceased individuals. Thus, we will not redact any personal information in out-of-copyright data.

For wikipedia articles, we will filter articles of EU based living people (e.g., where there is no death date) where the first paragraph of the article has words associated with one of the member states under the GDPR (e.g., “German”, “Italian”, etc.). We will only use these articles to create anonymized, de-biased and detoxified, and simplified content (see discussion of synthetic data below). Downstream users can still use all of wikipedia at inference time using retrieval based methods to increase factual recall (see SILO architecture (Min, 2023) using retrieval to augment a safe-by-design LLM), but as a policy matter, we will strive to prevent our models from memorising personal information of EU data subjects. We will use public tools such as our tools from PIISA.org to anonymize data. For our code data, The Stack dataset has already partially removed PII (based on our volunteer work in the Bigcode project), however we will additionally run our PII processing on The Stack data used for training. For sources that derive from non-EU sources, such as US case law and US government text, we will not redact personal information.

Regardless of the training data processing, we will however retain all attribution information in the datasets, such as the names of authors of scientific articles or of out of copyright books, even though those named which might include university affiliation for example would be stripped from the training data.

EU AI Act

Usage that could manipulate vulnerable groups such as children is another type of potential prohibited activity. Thus, we will also tag content with content rating: G, PG, R and X. With this rating, we can allow users to protect vulnerable groups such as children from accessing inappropriate NSFW content, but controlling content at inference. The ratings will be prepended data in a conditional pretraining scheme similar to Rohan, Anil (2023). The EU AI Act is keenly focused on high risk usage. While the downstream users are ultimately responsible for their usage, we will filter out data that could be used for potentially creating fake government ID images, border control documents or law enforcement documents, which could be used in high risk activity. In particular, we will use an image classifier to filter out images of government IDs, and use keywords to filter out government official documents. During our instruction turning, we will include instructions that will dissuade or refuse to provide instructions on performing criminal acts or potential harm to self and others, since safety is one of the themes of the Act. During instruction tuning, we will also include instructions that will refuse giving legal or medical advice, and refuse to give ratings for resumes or candidates. Additionally, we will attempt to address algorithmic bias by using de-biasing data augmentation as discussed below, because it is not only an ethical obligation, potential bias in certain activities such as hiring could be a high-risk activity under a potential EU AI Act.

Potentially Illegal Content

We will use previously developed KenLM filters and keyword filters from our work with BigScience, and LAION along with CLIP filters to filter potentially illegal content.

Instruction Tuning For Safety

Instruction tuning refers to the refined calibration of a model using specific guidance or directives, aimed at enhancing its alignment with human-specified intents and outcomes (Li, 2023). In our case, we will generate instructions that will dissuade illegal acts and harm to self and others.

Data Generation

As part of our safety efforts, we will generate culture-aware data, leverage existing distributions of language families to help generate representative and long-tail sampled data, and will generate content in different languages combined with pseudonymization thus providing privacy-preserving data generation. We will also make extensive use of out of copyright text or public domain text, such as Project Gutenberg. In this example, we take the following passage:

> …by Protestant zeal and benevolence for the reformation or the bringing up of poor Catholic children, and some of which go so far as to kidnap little papist orphans or half orphans, lock them up in their orphan asylums, where no priest can enter, change their names so that their relatives cannot trace them, send them to a distance, and place them …

After filtering for toxicity (changing “kidnap” and “lock them up” using text augmentation libraries such as wordnet, vector based analogy text-augmentation methods (see https://github.com/dsfsi/textaugment), and using PIISA.org’s PII tools, we generate:

> …by zeal and benevolence for the reformation or the bringing up of poor Buddhist children, and some of which go so far as to rescue little orphans or half orphans, rescue in their orphan asylums, where no nun can enter, change their names so that their relatives cannot trace them, send them to a distance, and place them …

In this case, we may choose Japanese as our target language, and thus use the culturally appropriate religion “Buddhist”. Translating to Japanese, we have:

> ….貧しい仏教徒の子供たちの更生や教育に対する熱意と慈悲によって、中には尼僧の立ち入りが許されない孤児院にいる小さな孤児や半孤児を救うことまでする人もいる。

We may create image text pairs by searching for Creative Commons Images in Image # 1. Alternatively, we could generate the image using an off the shelf image generator or our own trained model (Image # 2).

Image # 1	Image # 2:
	

Filtering using Clip to match the Images to the passage results in Image # 2. Then, we can also create a “screenshot” of the text and image and use an off the shelf OCR generator and object detector to create additional text. This could result in the synthetic data as follows:

> <vqgan_image>… </vqgan_image> || OCR text: ([[9, 7], [631, 7], [631, 47], [9, 47]], '貧しい仏教徒の子供たちのリハビリと教育に対す'), ([[69, 44], [550, 44], [550, 82], [69, 82]], 'る彼らの情熱と思いやりのおかげですs'), ([[9, 223], [193, 223], [193, 261], [9, 261]], '彼らの中には'), ([[205, 223], [633, 223], [633, 261], [205, 261]], '修道女が立ち入れない孤児院にい'), ([[11, 261], [633, 261], [633, 299], [11, 299]], 'る小さな孤児や半孤児を救出することまでする人'), ([[261, 299], [379, 299], [379, 333], [261, 333]], 'もいます',)

The above generation is our Data Augmentation Stage 1.

As part of our data generation and model training pipeline, we will include a Data Augmentation Stage 2 and continued pretraining stage (see Section 3.1 for timeline). Using our already trained 7b, 13b or 34b model, we will generate additional content similar to Stage 1, and additionally synthetic instructions (along with translations of such instructions) for a subset of text dataset above. We will generate safety based instructions (that for example steers away from encouraging physical harm to self and others and informs the user that the model is an AI when asked). Lastly we will generate synthetic stories and textbooks tailored to various domains, ages, and subject matters. Content ratings, age recommendations, and domain labels will be carefully assigned to ensure responsible

content generation. In both stages, the synthetic data will be highly filtered for quality, so we will overproduce data in both stages. Stage 2 data and instructions will then be used to instruction tune each of the models. We will also use some of the filtered synthetic data from Stage 1 and 2 as part of retrieval augmented generation instead of training, to reduce training cost and isolate legal risks. For example, while all training data will be at most CC-BY-SA licensed, we will create images and text from CC-BY-NC content for retrieval, and this data could be provided at inference time when the user wishes to generate content for non-commercial use. See SILO's work in isolating legal risks using retrieval (Min, 2023).

Dataset

We will curate and filter text from a wide variety of sources, most of which will be public or permissively licensed (e.g., government works might not be expressly licensed). Proposed data is described below:

Est. Text:	
SILO/OLC:	220B tokens
The Stack:	682 tokens
FreeLaw:	11B tokens
USPTO Abstracts:	5B tokens
HUPD Patents:	11B tokens
Additional Wikipedia:	2B tokens
Misc Govt works:	1B tokens
Additional Books	
CC-BY/CC-BY-SA:	1B tokens
We will create detoxified text	
For subset of stories (2B),	
news (1B) and Wikipedia (5B)	
Approx:	7B tokens
Open Source Instructions:	.2B tokens
We will create augmented	
instructions, simplified textbooks	
and stories:	2B tokens
We will generate	
30 language translation	
of approximately	
40B tokens each:	1200B
tokens *	
Generated CC-BY-NC text	
for retrieval only (3B per lang)	30-120B
tokens **	
Total Text Tokens	
With filtering (approx):	1950B

Est. Number of Text Embeddings:	
Text Embeddings of	
Approx 200M sentences	
To total 20B tokens or	
Approx 20 tokens per embeddings:	200M
Est. Images and Image Tokens:	
For each image, we will create (256 + 32 tokens)	
WIT:	11M
YFCC100M	
CC-BY/CC-BY-SA after filtering:	20M
*Note mini-dalle only	
downloaded 16M	
Fondant-CC-25M	
CC-BY/CC-BY-SA after filtering:	5M
Additional Wikimedia Images	
And video:	50M
(90M in total, but subtracting 11M and misc media)	
Misc other CC images:	2M
Generated Images:	90M
CC-BY-NC generated images	
used for retrieval only**	5-10M
Total Number of Images:	178M
Total Number of Image Embeddings:	178M
Total Number of Image tokens:	51B
token	

* This means we need approx. 4TB of translated (mostly factual) text.

** Generated tokens for retrieval augmented generation not counted in the totals for training purposes; amount will depend on available compute.

This will result in approximately 2T tokens. All sources of the data will be kept as metadata, including the permission for usage.

The challenge of training such large models on this 2T tokens will be managing large amounts of data, training instability, large spikes in the loss, and failure to decrease loss. From our experience, these issues can be alleviated through use of optimal precisions for components, optimal architectures, proper data cleaning and data mixtures.

2.3 Validation, verification, state of the art

2.3.1 *Validation & Verification*

Modelling. We will use existing autoregressive transformer models capable of processing tokens in various domains and languages, as well as mapping image embeddings into tokens to be used in the transformer. This is based on countless works using transformers for language and multimodality: LlaMA2 (Touvron, 2023), Emu (Sun, 2023), SEED (Ge, 2023). Following LlaMA2, we expect training on a similar amount of 2T tokens, with high quality filtering will result in similar performance (e.g., 27.8 HumanEval scores, 69.9 PIQA/HellaSwag scores, 62.6 MMLU scores). Following Emu, we expect our evaluations to be as high (e.g., 117.7 COCO scores, 57.5 VQAv2 scores, and using other standard metrics). Our innovation will be to combine the Emu architecture of embeddings and tokens to intermix image text understanding with the SEED architecture of using only image tokens and text tokens to understand and generate images and text.

Data. We will use our experience in governing and creating the ROOTS dataset (Laurençon, 2022) , and the Red Pajama (TogetherComputer, 2023) dataset (which is a clean room implementation of the Llama 1 dataset) to curate and filter high quality public domain and permissively licensed text. From a synthetic data perspective, we will use our machine translation translateLocally (https://github.com/XapaJIaMnu/translateLocally) and other pretrained models to translate text from English to our 30 other languages (e.g., NLLB-200: BLEU score of 37.84 *for language translation*, KenLM LID model (https://arxiv.org/abs/2305.13820) achieves a macro-average F1 score of 0.93 and a false positive rate of 0.033 across 201 languages *for language identification*).

Safety. We will verify and validate the safety of our model based on well known toxicity, and bias metrics. We will use the full suite of evaluations we developed in our work in Wang, 2023, to test for example, toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness. We will also use other industry metrics such as those performed in LlaMA2 (Touvron, 2023) (e.g., TruthfulQA, ToxiGen and BOLD for truthfulness, toxicity and bias, respectively).

2.3.2 *Comparison with state of the art*

Within the rapidly evolving AI research milieu, the Synthetic-data, Fair, and Extreme-scaled Large Language Model (SafeLMM) stands as a salient contribution vis-a-vis the state-of-the-art:

Most similar to SafeLMM, the SILO architecture (Min, 2023) addresses the formidable challenge of reconciling copyrighted data legality with the imperative for expansive and quality datasets. Similarly, we will leverage SILO's Open License Corpus (OLC), which comprises public domain and permissively licensed text. Contrasting this, SafeLMM extends its purview, synthesising and collecting permissive data across text and images in 31 languages, whilst rigorously adhering to ethical mandates and regulatory statutes.

Our work in data governance in BigScience and BigCode informs SafeLMM's data governance regime. See Jernite, 2022. However, unlike our previous work, we will use only out of copyright and permissive content, thus simplifying data governance. Moreover, we will perform more rigorous detoxification debiasing and privacy filtering. Lastly, we will provide a mechanism to check model outputs against data from the dataset, to permit attribution.

Similar to Biderman, 2023, we perform bias remediation by creating counter-factual gender swapping. Biderman, 2023 showed that such data-augmentation results in a low impact on perplexity scores on LAMBADA. Unlike in that work, we will also perform other types of bias remediation such as race swapping.

Emu (Sun, 2023) uses a transformer-based foundation model, adeptly spanning the realms of text , images, and videos. It can comprehend content across these modalities using singular input data. Similarly, SafeLMM will use a similar autoregressive embedding and text interleaved training. Nonetheless, SafeLMM's objectives supersede mere multimodal understanding. By enhancing ethical content generation, robust governance mechanisms, and open-source paradigms, SafeLMM not only augments Emu's technological prowess but also strengthens ethical congruence.

Overall, the relative advantages of SafeLMM is more comprehensive multimodal capabilities, an emphasis on ethics, and use of synthetic data for which we can control for safety.

2.4 Software and Attributes

2.4.1 *Software*

The project will primarily make use of FMEngine and OpenLM software for large multimodal model training, which we have previously used on A100 GPUs on the JUWELS HPC The software is based on other established libraries such as pytorch, DeepSpeed, Flash Attention and NVIDIA libraries for interconnection (NCCL). In preparation for this submission, we have benchmarked our training libraries on 7b to 70b parameter models on JUWELS and Leonardo. For image data generation, we will use an off the shelf text to image generator model which produces VQGAN image tokens from text and the SEED visual tokenizers, which we have also tested on JUWELS HPC. We will use machine translation libraries such as our translateLocally and NLLB. We will also use the SILO-lm (Min, 2023) open source software for retrieval augmented generation and evaluation, as well as our prior work in the DecodingTrust AI safety evaluations (Wang, 2023).

2.4.2 *Particular libraries*

In the SafeLMM project, various libraries, algorithms, numerical techniques, programming languages, and build environments are employed to facilitate production and data analysis. Here are some of the specific requirements and components used in the project: numpy, pandas, pytorch, Huggingface transformers, pytorch, and other standard libraries. Python Version: Specific Python 3.X versions will be required, and tools like virtualenv or conda can be used to manage Python environments. Dependency files, such as requirements.txt for Python packages. For version control, the project uses exclusively git, including the public Github repositories.

2.4.3 *Parallel programming*

We employ a widespread data parallel distributed training scheme, which relies on a set of mechanisms and tools. We make use of various libraries like PyTorch DDP and DeepSpeed that offer convenient ways to execute data parallel distributed training. In the data parallel scheme, the network model is cloned across multiple GPU devices, while each clone is getting its own mini-batch portion to train on, locally computing loss gradients. Those gradients are then gathered decentrally

across clone workers and reduced via efficient Ring- AllReduce operations (Caron, 2021) to update the model and sync all the workers' clones.

2.4.4 *I/O requirements*

SafeLMM's model training process involves reading pre-processed binary tensor data stored in a custom format on disk, utilising memory-mapped and streaming I/O. Efficient data transfer from disk to GPU memory is essential to ensure productive model training. It's important to note that this data transfer process does not impose extraordinary I/O requirements. During the training process, SafeLMM primarily engages in two types of I/O operations, we will perform standard logging, checkpoint writing of a maximum of 2TB per checkpoint and other miscellaneous transfers. We will be transferring about 200TB via ftp or scp for our image data from an external source.

2.5 Data: Management Plan, Storage, Analysis and Visualization

2.5.1 *Data Management Plan covering*

SafeLMM anticipates that the transfer of data into the system will involve a relatively small amount of data. Most of the training data will be publicly available text data. Text, images, Tokens and embeddings generated as part of the synthetic data creation will add approximately 200TB. We will not retain real or synthetic images, so this is a maximum usage estimate. Checkpoints can be as large as 2TB per model per checkpoint. We will follow the FAIR principles of scientific data management, making sure that all data (unless restricted by licence agreements) will be made Findable, Accessible, Interoperable and Reusable. All data will be made publicly available.

2.5.2 *Project workflow*

The SafeLMM project workflow encompasses several stages, including data generation, analysis and visualization and model training, which includes data generation, model training, and instruction tuning. See discussion in Section 2.2 and 3.1

2.5.3 *Software workflow solution*

We will be using various pre- and post-processing scripts to automate run management and facilitate the volume of work. For training jobs, we will use SLURM, along with job chaining. We will use W&B tracking to monitor training. We will also use python multiprocessing, webdatasets, Hugging Face's datasets, and parquet/arrow data to process the volume of data.

2.5.4 *I/O requirements*

We will not perform significant data analysis that will require high I/O requirements.

2.6 Performance of Software

2.6.1 *Testing of your code on the requested machine*

We have performed two main types of code tests. We have tested our production ready code on the Leonardo HPC and our the JUWELS HPC which has a substantial similar specification - comparable size, software stack and with the same architecture, and network. Most of our tests were on JUWELS due to limited access to Leonardo HPC. But we have found that the Leonardo HPC performs slightly better in terms of token/sec. A comparison of the two systems is below:

Feature / Specification	**JUWELS**	**Leonardo**

Location	Atos Forschungszentrum Jülich	CINECA datacenter, Bologna, Italy
Main Processors	Intel Xeon Skylake, AMD EPYC Rome	Intel Xeon 8358, Intel Sapphire Rapids
GPU Architecture	Nvidia Ampere GPUs (3744 x A100, 224 x V100)	Nvidia Ampere GPUs (13,824 x custom)
Network	200 Gb/s (NVIDIA Mellanox HDR InfiniBand)	200 Gb/s (NVIDIA Mellanox HDR InfiniBand)
Compute nodes	992	3,456

General specifications

System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
JUWELS	449,280	44.12	70.98	1,764
Leonardo	1,824,768	238.70	304.47	7,404

Raw performance and energy consumption

2.6.2 *Quantify the HPC performance of your project*

2.6.2.1 Strong and weak scalability

Performance of large language and multimodal model training are based on tokens/seconds (which is in turn related to TFlops). We began scaling tests for our smallest model of 7B parameters and a much bigger model 70b. While we only aim to train a 7B, 13B and 34B, we use the 70B to interpolate tokens/sec performance. Performance on JUWELS and Leonardo are as follows, from small number of nodes to larger number of nodes.

Leonardo 7b - Weak Scaling

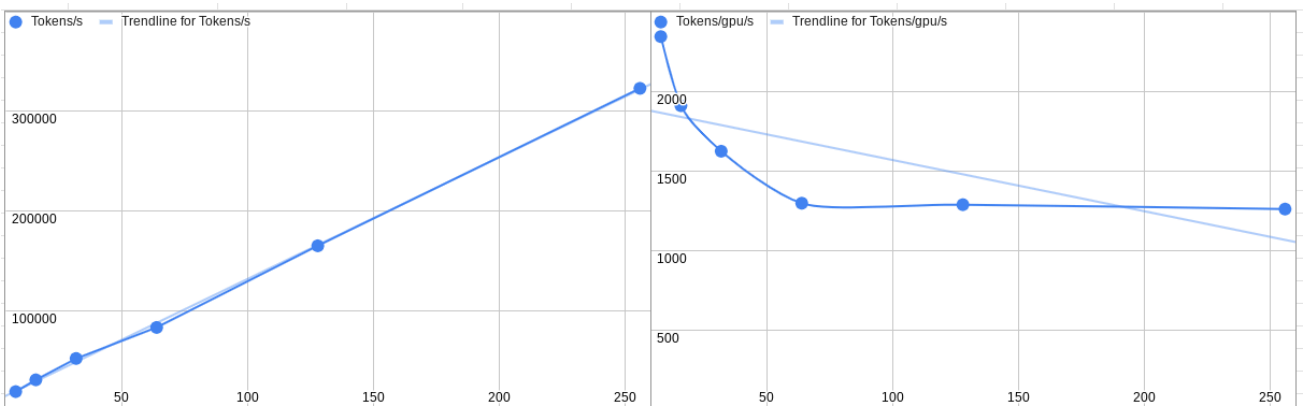
GPUs	train_batch_size	micro_batch_size	seq_length	Tokens/s	Tokens/gpu/s	Nodes
8	512	8	2048	23039.24	2879.905	2
16	512	8	2048	42690.94	2668.18375	4
32	512	8	2048	73987.77	2312.117813	8
64	512	8	2048	117864.71	1841.636094	16

JUWELS 7b - Weak scaling

Showing asymptote performance of tokens/sec with scaling of nodes:

GPUs	train_batch_size	micro_batch_	seq_length	Tokens/s	Tokens/gpu	Nodes

		size			/s	
8	512	8	2048	18745.6	2343.2	2
16	512	8	2048	30558.39	1909.899375	4
32	512	8	2048	51981.71	1624.428438	8
64	512	8	2048	83097.33	1298.395781	16
128	1024	8	2048	165030.77	1289.302891	32
256	2048	8	2048	323002.08	1261.726875	64



Leonardo 70b - Strong scaling

GPUs	train_batch_size	micro_batch_size	seq_length	Tokens/s	Tokens/gpu/s
80	1024	4	2048	19688.59	246.107375

JUWELS 70b - Strong scaling, longer sequence length

GPUs	train_batch_size	micro_batch_size	seq_length	Tokens/s	Tokens/gpu/s
80	1024	4	4096	38000	475
160	1024	4	4096	71000	443.75

While we do not expect to train at the extreme of 70b parameters, we have tested our production code and are confident that the performance in tokens/sec is adequate to train our models in the requested node hours discussed below. We expect that as we move closer to 256 nodes with longer sequences (>4K), we will reach an asymptote of the following tokens/sec.

7B **0.4M GPUh (1500 tokens/sec/GPU)**
13B **0.8M GPUh (752 tokens/sec/GPU)**

34B 2.0M GPUh (267 tokens/sec/GPU)

2.6.2.2 Precision reported

We will use various precision for our various models depending on usage. Our main training will use bfp16 which is a mixed precision model. Our data generation code will use fp16 for speed of performance.

2.6.2.3 Time-to-solution

We performed ablation studies on JUWELS using various scaling of both parameters, batch size, and node size, using industry standard methodology. Our time to solution is expected to be tokens/sec/GPU per size of the model times the number of training tokens (2T tokens). Assuming a variance in our training run tokens/sec and errors in training and restarting, we expect the time to solutions to be as set forth in the chart in Section 3.1.

2.6.2.4 System scale

Please see the charts above for the scaling of just our training model engine. While the tests above did not account for data loading time, we have extensive experience in using data loaders to train our distributed models. As an example, in our other vision-language model training on JUWELS, we used webdatasets, which we will be using for our training, as shown in our other scaling work (Clip, CoCa, Flamingo, Stable Diffusion):

#Nodes	#GPUs	SD v1 Im/s (Eff.%)	CoCa H/14 Im/s (Eff.%)	Paella Im/s (Eff.%)	Flamingo L/14 Im/s (Eff.%)
1	4	17 (100%)	165 (100%)	27 (100%)	108 (100%)
2	8	33 (99%)	323 (98%)	54 (99%)	214 (99%)
4	16	65 (99%)	647 (98%)	108 (99%)	429 (99%)
8	32	131 (99%)	1281 (97%)	214 (98%)	858 (99%)
16	64	256 (97%)	2546 (97%)	426 (98%)	1669 (97%)
32	128	518 (98%)	5027 (95%)	843 (97%)	3393 (98%)
64	256	1003 (95%)	9838 (93%)	1679 (96%)	6712 (97%)
128	512	2046 (97%)	19495 (93%)	3352 (96%)	13071 (95%)
256	1024	3650 (86%)	37506 (89%)	6326 (91%)	27300 (99%)

Thus, we are confident that our training will scale to 256 nodes on Leonardo HPC, on a system wide basis.

2.6.2.5 Measurement mechanism

We will use TFlops and tokens/sec as our primary performance measure. We will measure throughout, training loss, validation loss and eventual model performance via the metrics discussed above.

2.6.2.6 Memory usage

We expect to use nearly the full memory of each GPU as discussed above, by using tensor parallelism, data parallelism and pipeline parallelism, while accounting for potential out of memory issues. We will also use sequence parallelism to increase sequence length.

3 Milestones (quarterly basis)

Run Type	Code(s)	No. of runs (2 days per run)	No. of GPU Nodes	No. of steps per run*	No of cores per node	Time per step(s)	Est. Total node hours
Data Augmentation 1 **	Image, and Mult-Lingual generation	30	256	N/A	4	48 hours	100K node hours/400K GPUh
7B model	training	30	256	N/A	4	48 hours	100K node hours/400K GPUh
13B model	training	60	256	N/A	4	48 hours	200K node hours/800K GPUh
34B model	training	120	256	N/A	4	48 hours	500K node hours/2M GPUh
Eval 7b ***	evaluation	10	40	N/A	4	48 hours	5K node hours/20K GPUh
Eval 13b ***	evaluation	10	80	N/A	4	48 hours	10K node hours/40K GPUh
Eval 34b ***	evaluation	10	160	N/A	4	48 hours	20K node hours/80K GPUh
Data Augmentation 2 **	Text generation	8	256	N/A	4	48 hours	97.5k node hours/390K GPUh
Instruction Tune 7B model	finetuning	5	40	N/A	4	48 hours	2.5K node hours/10K GPUh
Instruction Tune 13B model	finetuning	5	80	N/A	4	48 hours	5K node hours/20K GPUh
Instruction Tune 34B model	finetuning	5	160	N/A	4	48 hours	10K node hours/40K GPUh
Total:							1.05M node hours/4.2M GPUh

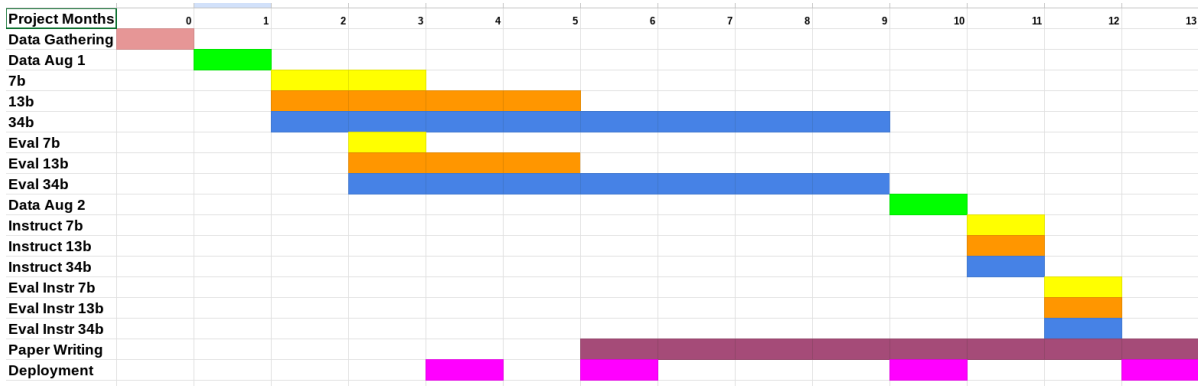
* We do not measure training in steps, but rather in tokens.

** We do not use all generated data for training and will use a portion for retrieval augmented generation

*** Evaluation will be done in two stages, once for the pretrained models and once for the instruct models. See section 3.1 below for the proposed schedule.

3.1 Gantt Chart

The Gantt chart below provides a view on the time schedule for work packages intended in this project and is based on the workflow description in Section 2.5.2. We will actively be consulting with EuropHPC JU regarding the project, including its scheduling. Our communication plan will be to discuss on a monthly basis our status and create reports on a quarterly basis. Work Products will be disseminated via model, data and documentation release as described herein. Work schedule for the project duration (12 months using Leonardo, 14 months in total).



4 Personnel and Management Plan

The PI will monitor the usage of the assigned resources and will take actions to ensure the most efficient use. Results will be subject to dissemination and/or publications that will constitute part of the documentation and reports. No additional hiring relevant to SafeLMM is currently planned nor turnover in critical roles anticipated.

5 References

M Brenzke, S Wiesen, M Bernert, D Coster, J Jitsev, Y Liang, U von Toussaint, ASDEX Upgrade Team, EUROfusion MST1 Team, et al. (2021). Divertor power load predictions based on machine learning. Nuclear Fusion, 61, no. 4, 046023.

Mathilde Caron, et. al. (2021). Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660

Mehdi Cherti, Jenia Jitsev (2021). Effect of large-scale pre-training on full and few-shot transfer learning for natural and medical images. In Medical Imaging Meets NeurIPS (MedNerIPS) Workshop, arXiv:2106.00116

Mehdi Cherti, , et al. (2023). Reproducible scaling laws for contrastive language-image learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

JM Clavijo, P Glaysher, J Jitsev, and JM Katzy (2021). Adversarial domain adaptation to reduce sample bias of a high energy physics event classifier, Machine Learning: Science and Technology, 3, no. 1, 015014.

Mohammad J Eslamibidgoli, et al. (2021). Convolutional neural networks for high throughput screening of catalyst layer inks for polymer electrolyte fuel cells. RSC Advances, 11, no. 51, 32126–32134.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, & Ying Shan (2023). Planting a SEED of Vision in Large Language Model. *arXiv preprint arXiv:2307.08041*.

Jernite, Yacine, et al. "Data governance in the age of large-scale data-driven language technology." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

Bidderman, Stella, et al. Pythia: A Suite for Analyzing Large Language Models Across Training and Scalin. *Proceedings of the 40 th International Conference on Machine Learning, 2023.*

Wang, Boxin, et al. (2023) DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *arXiv preprint arXiv:2306.11698*

Jared Kaplan, et al. (2020). “Scaling laws for neural language models”.

Laurençon, et al. (2022). The BigScience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*,

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang and Ziwei Liu. “Otter: A Multi-Modal Model with In-Context Instruction Tuning.” ArXiv abs/2305.03726 (2023).

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*

Jared Kaplan, et al., “Scaling laws for neural language models”, 2020.

Kocetkov, D., Li, R., Allal, L. B., Li, J., Mou, C., Ferrandis, C. M., ... & de Vries, H. (2022). The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533*.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., ... & Mattick, A. (2023). OpenAssistant Conversations--Democratizing Large Language Model Alignment. *arXiv preprint arXiv:2304.07327*.

Kuo, W., Piergiovanni, A. J., Kim, D., Luo, X., Caine, B., Li, W., ... & Angelova, A. (2023). Mammut: A simple architecture for joint learning for multimodal tasks. *arXiv preprint arXiv:2303.16839*.

John P Miller, , et al. ((2021). Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization, in: International Conference on Machine Learning (ICML), https://arxiv.org/abs/2107.04649.

Alec Radford,, et al. (2021). Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, pp. 8748–8763. PMLR

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2019). Do imagenet classifiers generalize to imagenet?, in: International Conference on Machine Learning, pp. 5389–5400. PMLR

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Sewon Min, et al. (2023). SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. *arXiv preprint arXiv:2308.04430*.

Christoph Schuhmann, et al. (2021). LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs.

Christoph Schuhmann, et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models.

Quan Sun, et al. (2023). Generative Pretraining in Multimodality.

Together Computer (2023). RedPajama: An Open Source Recipe to Reproduce LLaMA training dataset. URL: https://github.com/togethercomputer/RedPajama-Data

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wu, Y., et. al. (2023, June). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*

Samir Yitzhak Gadre, et. al (2023). DataComp: In search of the next generation of multimodal datasets.

6 Confidentiality

- Is any part of the project covered by confidentiality? **no**
- Does your project involve handling of personal data? **yes**