mOSCAR: A Large-scale Multilingual and Multimodal Document-level Corpus

Matthieu Futeral[*] [1,2] **Armel Zebaze**[1,4] **Pedro Ortiz Suarez**[5] **Julien Abadji**[1]

Rémi Lacroix[3,6] **Cordelia Schmid**[1,2] **Rachel Bawden**[1] **Benoît Sagot**[1]

[1]Inria [2]Département d'informatique de l'ENS, CNRS, PSL Research University
[3]Institut du développement et des ressources en informatique scientifique, CNRS
[4]Sorbonne Université, Paris, France [5]Common Crawl Foundation [6]Université Paris-Saclay

Abstract

Multimodal Large Language Models (mLLMs) are trained on a large amount of text-image data. While most mLLMs are trained on caption-like data only, Alayrac et al. (2022) showed that additionally training them on interleaved sequences of text and images can lead to the emergence of in-context learning capabilities. However, the dataset they used, M3W, is not public and is only in English. There have been attempts to reproduce their results but the released datasets are English-only. In contrast, current multilingual and multimodal datasets are either composed of caption-like only or medium-scale or fully private data. This limits mLLM research for the 7,000 other languages spoken in the world. We therefore introduce mOSCAR, to the best of our knowledge the first large-scale multilingual and multimodal document corpus crawled from the web. It covers 163 languages, 303M documents, 200B tokens and 1.15B images. We carefully conduct a set of filtering and evaluation steps to make sure mOSCAR is sufficiently safe, diverse and of good quality. We additionally train two types of multilingual model to prove the benefits of mOSCAR: (1) a model trained on a subset of mOSCAR and captioning data and (2) a model trained on captioning data only. The model additionally trained on mOSCAR shows a strong boost in few-shot learning performance across various multilingual image-text tasks and benchmarks, confirming previous findings for English-only mLLMs. The dataset is released under the Creative Commons CC BY 4.0 license and can be accessed here.[1]

1 Introduction

Multimodal Large Language Models (mLLMs) are trained on large amounts of text-image data (Radford et al., 2021; Yu et al., 2022; Li et al., 2023; Wang et al., 2023; OpenAI, 2023; Gemini Team et al., 2023; Chameleon Team, 2024). The main paradigm until recently was to train a model from a large collection of web-crawled images and their captions (Li et al., 2021; Wang et al., 2022; Chen et al., 2023b). Models such as Flamingo (Alayrac et al., 2022) challenged this paradigm by being additionally trained on interleaved sequences of text and images from web documents, showing state-of-the-art results on various tasks and in-context learning capabilities that are not present in models trained on caption-like data only. Additionally, McKinzie et al. (2024) recently proved that including interleaved text-image data during training was necessary to get good few-shot learning performance. However, the datasets used to train mLLMs are either private (Alayrac et al., 2022), monolingual or multilingual but only medium-scale (Srinivasan et al., 2021). Some attempts have been made to reproduce these datasets (Zhu et al., 2023; Laurençon et al., 2023) but the resulting datasets are only available in English.

Few image-text datasets are multilingual and most of them are obtained by translating English caption-like datasets, such as multilingual Conceptual Captions (Sharma et al., 2018), into multiple languages using neural machine translation (NMT) systems (Surís et al., 2022; Maaz et al., 2024). This presents some drawbacks such as some languages still being poorly translated by current state-of-the-art NMT models (Liu et al., 2020; Costa-jussà et al., 2022) and some cultural subtleties inherent in each language not being fully conveyed. Some efforts have been conducted to collect large-scale multilingual image captioning datasets, such as LAION-5B (Schuhmann et al., 2022), but they are limited to caption data too, are relatively noisy and more importantly contain a non-negligible share of "not safe for work" (NSFW) content such as pædopornographic images (Schuhmann et al., 2022).

This motivated us to collect and release the first large-scale multilingual and multimodal document dataset derived from Common Crawl.[2] Our dataset, multimodal OSCAR (mOSCAR), follows the OSCAR initiative (Ortiz Suárez et al., 2019; Abadji et al., 2021, 2022) and covers 303M documents in 163 languages, 200B tokens and 1.15B images. Figure 1 shows an example of a document, more can be found in Appendix A.3. We carry out extensive filtering to increase its safety and quality. To prove mOSCAR's utility, we train a multilingual Open-

[*]Correspondence to matthieu.futeral@inria.fr

[1]https://oscar-project.github.io/documentation/versions/mOSCAR/

[2]https://commoncrawl.org/. The Common Crawl Foundation is a non-profit organization that crawls the web on a monthly basis.

Findings of the Association for Computational Linguistics: ACL 2025, pages 3461–3494
July 27 - August 1, 2025

Flamingo (Awadalla et al., 2023) from a Gemma-2B language model (Gemma Team et al., 2024) on a subset of mOSCAR and captioning data from LAION-400M (Schuhmann et al., 2021), recaptioned with BLIP (Li et al., 2022), filtered with CLIP (Radford et al., 2021) and translated with NLLB (Costa-jussà et al., 2022). We compare against a similar model trained on captioning data only and show we obtain a strong boost in few-shot learning, confirming previous findings for English (Alayrac et al., 2022; McKinzie et al., 2024; Laurençon et al., 2024). mOSCAR can be accessed here [3].

2 Related Work

Large-scale web-based datasets Numerous datasets have been created by filtering web-crawled data. These include large-scale text-only datasets (Ortiz Suárez et al., 2019; Raffel et al., 2020; Wenzek et al., 2020; Gao et al., 2020; Abadji et al., 2021; Xue et al., 2021; Laurençon et al., 2022; Abadji et al., 2022; Penedo et al., 2023) and multimodal ones (Sharma et al., 2018; Changpinyo et al., 2021; Jia et al., 2021; Schuhmann et al., 2021, 2022; Byeon et al., 2022; Laurençon et al., 2023; Zhu et al., 2023; Gadre et al., 2024). Even if these datasets are not as high quality as smaller and/or hand-crafted ones, they are now the standard to pretrain foundation models, as it has been shown that training bigger models on more data leads to better downstream performances (Brown et al., 2020; Hoffmann et al., 2022; Touvron et al., 2023a,b).

English image-text datasets The first open-source image-text datasets were manually created, small-scale and English-only (Ordonez et al., 2011; Lin et al., 2014; Plummer et al., 2015; Krishna et al., 2017). Scaling up these datasets was an appealing solution to overcome limitations of previous image-text models; a few works (Sharma et al., 2018; Changpinyo et al., 2021) proposed to collect millions of image-text pairs from the web before filtering them with well-designed steps. Relaxing the filtering steps enabled the collection of more data and led to large-scale datasets to train image-text foundation models (Radford et al., 2021; Li et al., 2021; Schuhmann et al., 2021, 2022; Byeon et al., 2022). However, these datasets generally contain caption-like image-text pairs only, and it is therefore difficult to observe in-context learning abilities similarly to text-only language models trained on raw documents (Raffel et al., 2020). Alayrac et al. (2022) overcome this issue by training their model directly on documents with interleaved image-text data. While their results are promising, their M3W dataset is English-only and private. Recently, open-source efforts (Zhu et al., 2023; Laurençon et al., 2023) have been made to release a similar dataset but they are still monolingual.

Multilingual image-text datasets Only a few image-text datasets are available in multiple languages. One of the first focused on collecting Google images from short queries based on word frequencies from Wikipedia pages in 98 languages (Hewitt et al., 2018). Later, Srinivasan et al. (2021) proposed the WIT dataset, an image-text dataset composed of Wikipedia pages. Although of high quality, it is only medium-scale even for high-resource languages and there are fewer than 50k unique images for most languages. Another approach lies in bootstrapping multilingual and multimodal data from a model trained with English-only data (Mohammed et al., 2023). While effective for captioning, it is computationally expensive to implement in practice. Other multilingual image-text datasets exist but focus on captions only and are highly domain-specific (Kosar et al., 2022; Leong et al., 2022).

3 Dataset Creation Pipeline

3.1 Data collection

We collect mOSCAR from the Web ARchive Content (WARC) files of three 2023 Common Crawl dumps, processing them using the FastWARC library (Bevendorff et al., 2021). We remove documents smaller than 500 bytes (50% of the documents), as we find they are usually too small to be considered documents and tend to contain noisy text. We then navigate through the entire Document Object Model (DOM) tree with a depth first search algorithm and ChatNoir library (Bevendorff et al., 2018) to extract nodes of interests corresponding to specific HTML tags.

Following previous work, we extract text from the tags that usually contain the main content of web pages (we refer to them as DOM text nodes), i.e. `<p>`, `<h*>`, `<title>`, `<description>`, `<ul>`, `<ol>`, `<aside>`, `<dl>`, `<dd>`, `<dt>`. Similarly to (Laurençon et al., 2023), we choose to remove `<table>` content as most often it is irrelevant and difficult to render. We extract all `<img>` tags (we refer to them as DOM image nodes). We then remove documents with fewer than 3 text nodes (as they do not contain enough text) and more than 30 image nodes (as we found them to be too noisy).

3.2 Language identification

We identify the language of each document using the state-of-the-art open-LID language detector (Burchell et al., 2023), covering 201 languages. We apply open-LID to each DOM text node and keep the three most probable languages with their respective probabilities. The language of the document is then determined by summing over the probabilities of each language detected for each text segment, weighted by the number of characters in the segment[4] and taking the language with the highest score.

[3] https://huggingface.co/datasets/oscar-corpus/mOSCAR

[4] This is to avoid mis-assigning the language due to the presence of many short, non-informative DOM text nodes in the same language (e.g. "Cookies", "Subscribe", "Newsletter" etc.) and because language identification is generally less reliable for short segments.

Le plus charmant petit square de Paris.

Le square René Viviani a été ouvert au public en 1928. Il occupait l'emplacement de l'annexe de l'Hôtel-Dieu (il est situé en face de Notre Dame) et d'une rangée de maisons longeant la rue de la Bûcherie. Le conseiller municipal Léon Riotor et la Commission du Vieux Paris ont particulièrement oeuvré à sa création dans le but de préserver à la fois le dégagement de l'église Saint-Julien-le-Pauvre et la perspective de Notre-Dame. Quelques éléments architecturaux parsèment le parc, et contribuent à son charme : pinacles, chapiteaux, balustrades, provenant sans doute de la cathédrale Notre-Dame. [...]



Pinacles et chapiteaux ornent le parc : probablement des vestiges de Notre-Dame. La fine poussière derrière la balustrade de pierre provient du fort vent qui faisait s'envoler le sable et la terre. [...]

Figure 1: Example of a French document from mOSCAR.

3.3 Text-only filtering

We apply a series of filtering steps to the text content of each document independently of the images, with the aim of discarding poor quality documents and cleaning text as best as possible. We first filter at the text-node level and then at the whole document level, before running near-deduplication to keep unique text nodes within a document and unique documents in the dataset.

Text node filtering We use a set of heuristics (see Appendix A.4) to extract as much human-generated content as possible while discarding noisy text related to ads and website functions (e.g. "Instagram", "Facebook"). We then keep DOM text nodes with content over 10 bytes. This step, designed to improve the quality of extracted text, removes on average 55% of text nodes.

Document filtering We mostly filter "not safe for work" (NSFW) content at the document level. We use an English regular expression to detect adult content, similar to the one used by the Université Toulouse 1 Capitole[5] and remove the entire document if there is a match with any of the DOM text nodes' contents, removing on average 0.5% of documents (mostly English ones). We acknowledge that there is a high probability that this also discards safe content, e.g. we could remove content from certain communities who use some explicit words in a non-sexual way (Sap et al., 2019). However, we explicitly favour recall over precision to minimise the risk of unsafe content. We additionally remove documents containing fewer than five DOM text nodes and fewer than 300 characters after the previous filtering steps, removing 70.6% of documents.

Deduplication We conduct several types of per-language deduplication at different levels, as this has been shown to improve training efficiency (Abbas et al., 2023). First, we keep unique documents only by removing exact duplicates at the document level. We also remove exact duplicates of text nodes within the same document (4% of text nodes) and near-duplicate text nodes (1% of text nodes) by computing the Levenshtein ratio (Levenshtein, 1966) between all text nodes within the same document and applying a threshold of 0.95. If near-duplicates are found, we keep the first one in the document. Finally, we conduct per language near-deduplication at the document level with MinHashLSH (Broder, 1997; Gionis et al., 1999) following Smith et al. (2022), removing on average 19% of documents:[6] we turn documents into hashing vectors, compute min hashes from these vectors and perform Locality Sensitive Hashing to remove duplicates[7] (see Appendix A.6.1 for more details).

Toxicity filtering Toxic content targeting individuals or groups of people is widespread on the internet and can therefore be found in large-scale web-crawled datasets like mOSCAR without appropriate filtering steps. To alleviate this issue, we apply the same method used by Costa-jussà et al. (2022) and remove documents from mOSCAR based on the presence of a list of "toxic" words for each language[8]. As some words in the list can also be used in a non-toxic way based on the context (e.g.: 'breast' in English), we tag the document as toxic and remove it from mOSCAR if it contains at least two distinct words in the list. This filtering step removes 0.95% of the documents for very high-resource languages (>5M documents), 2.13% for high-resource languages (<5M, >500K), 0.47% for mid-resource languages (<500K, >50K) and 0.64% for low-resource languages (< 50K). When manually analysing 1,000 random documents removed by this filtering step in each of the 2 (high-resource) languages we are native speakers of (English and French), we found 568 documents with toxic content.

Personal Identifiable Information Personal Identifiable Information (PII) can be found in large-scale web-crawled datasets, we therefore conducted an additional

[5] https://dsi.ut-capitole.fr/blacklists/index_en.php

[6] With some disparity among languages as we found more duplicates for low- than high-resource languages.

[7] We performed this using the `datasketch` python library.

[8] The list of these words for each language can be found here: https://github.com/facebookresearch/flores/tree/main/toxicity

filtering step to replace all detected PII by place holder strings using regular expressions (Appendix A.6.2 for more details). Concretely, we replaced all detected email addresses, phone numbers, credit card numbers, IP addresses and passport numbers. Moreover, it was shown that CommonCrawl contains a non neglictible part of API keys in its content [9]. We therefore scanned the dataset with the trufflehog tool [10] to track down residual API keys that could have passed previous filters. We found ∼200K positive matches and manually check a random sample of 1K positive matches. We found only 2 of them to potentially be API keys, other matches are mainly noisy text nodes not related to PII. We removed the 200K text nodes from mOSCAR.

3.4 Image-only filtering

We downloaded images from the URLs in DOM image nodes using a modified version of the img2dataset toolkit (Beaumont, 2021) that includes an antivirus scan and follows `robots.txt` instructions to respect the Robots Exclusion Protocol. We then apply a series of filtering steps, first removing images based on heuristics, and then applying multiple NSFW detection models to remove undesirable content. Finally, we conduct a set of deduplication steps.

Rule-based filters Similarly to previous works (Schuhmann et al., 2021) and to avoid extracting low-resolution images and favicons, we keep images with a minimum height and width of 150 pixels. We restrict the aspect ratio to be between 3 and 1/3 (to remove banners), we remove images if their URLs contain the words "logo", "banner", "button", "widget", "icon" or "plugin" or if the image name from the URL matches "twitter", "facebook" or "rss" (to remove logos). This step removes 13.6% of the URLs. At this stage, we downloaded 2.5B images with an average success rate of 55%.

NSFW detection We use multiple NSFW automatic models to remove as much unsafe content as possible. We first combine two NSFW detectors: nsfw-detector (Laborde), a 5-class classifier with a MobileNet (Howard et al., 2017) backbone fine-tuned on 60GB of annotated data and NudeNet,[11] an object detector trained to detect different types of nudity in images. We combined the two models as we found the first to be gender-biased while the second gives a large number of false positives for non-human images. Concretely, we consider an image an NSFW candidate if the sum of the probabilities for the classes 'porn' and 'hentai' is superior to 0.8 using nsfw-detector. We then tag the image as NSFW if one of the sensitive 'exposed' classes of NudeNet gets a probability superior to 0.5.

If a document contains an image with an NSFW tag, we remove the entire document from the dataset, which removes 0.5% of images. We manually inspecting 1,000 images of the remaining data and found no NSFW content. We manually inspected 1,000 images of the removed content and found 63.4% of NSFW images.

CSAM content Child Sexual Abuse Material (CSAM) is widespread on the internet and is therefore likely to be found in such a large-scale dataset crawled from the web. Removing CSAM is challenging as there is no training data nor open-source detection models available as these could be used in a harmful way. We rely on Thorn's CSAM classifier[12], a proprietary classifier trained to detect CSAM content in images. We tag the image as CSAM if the probability of the class CSAM is superior to 0.4 to favour recall over precision. As mentioned above, if a document contains an image with a CSAM tag, we remove it from the dataset. This step removes 0.07% of the images.

Deduplication To avoid memorisation issues often seen in models trained on datasets with many duplicated images (Somepalli et al., 2023; Carlini et al., 2023; Webster et al., 2023; Somepalli et al., 2024), we perform deduplication at the image level. We first remove duplicate images within the same document by URL matching (removing 8.7% of URLs). We then compute a perceptual hash (pHash) for each image using the imagehash library[13] and remove images with the same pHash within the same document, keeping only the first occurrence. We also limit the number of times an image can appear in the dataset per-language to 10 using both URL matching and perceptual hashing (this removes 2.5% of images). We do this per-language and not across languages as having the same images in documents from different languages could encourage cross-lingual transfer.

Personal Identification Information To protect PII in images, we use a lightweight face detector[14] and apply a threshold of 0.99 to detect faces in the images. We apply such a high threshold as we found the model to be biased towards detecting faces with high probability in images without any human. For each image in mOSCAR, we distribute the bounding boxes of the detected faces so that users can blur them when downloading the images. More details are provided in Appendix A.6.2.

3.5 Data decontamination

LLMs and mLLMs are trained on web-crawled data that can contain the benchmarks they are tested on (Dodge et al., 2021). As they are good at memorizing training data (Carlini et al., 2023), this data contamination is problematic. We therefore discard all images with the

[9] https://trufflesecurity.com/blog/[...]

[10] https://github.com/trufflesecurity/trufflehog

[11] https://github.com/vladmandic/nudenet

[12] https://safer.io/

[13] https://github.com/JohannesBuchner/imagehash

[14] https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

same perceptual hash as any of the images from the evaluation benchmarks (and their training sets) we use (see Section 5.1). This step removes on average 126,016 images for high-resource languages (up to 300K images for English), 6,862 images for mid-resource languages and 45 images for low-resource languages.

3.6 Text-image joint filtering

Our aim is to obtain truly multimodal documents where all images are related to at least one of the text nodes in some way[15] and vice versa. We choose to apply joint text-image filtering to discard images and/or text nodes that are irrelevant to the rest of the document (e.g. the case of ads and website functionalities).

To do this, we use NLLB-SIGLIP[16] (Visheratin, 2023), a multilingual version of SIGLIP (Zhai et al., 2023) trained with the encoder of NLLB (Costa-jussà et al., 2022), which covers all mOSCAR languages.[17] We compute cosine similarity scores between all images and all paragraphs[18] within a same document. To remove irrelevant text nodes or images in a document, we mimic a text-image retrieval task, which means we avoid using arbitrary cosine similarity thresholds for each language and can reduce length biases and those in favour of caption-like paragraphs. For each candidate pair we randomly sample 63 negative images and 63 negative similar-length paragraphs from the same language but other documents. We tag the text node (resp. image) as valid if the cosine similarity of the pair is among the top 8 of the text-to-image (resp. image-to-text) similarity scores computed with the candidate text node (resp. image) and all the negative images (resp. text nodes). This means that we tag the text node (resp. image) as valid if it has a significantly higher score than a score computed with a random image (resp. text) for at least one of the images (resp. text node) in the document. We then discard text nodes and images not tagged as valid (on average 35% of the DOM text nodes and 10% of the images within a document).

After this filtering step, we apply additional text-only filters to keep documents superior to 100 bytes. We also reapply the open-lid language detector (Burchell et al., 2023) as described in Section 3.2 as we found the last filtering step to change the major language of some documents.

4 Multimodal Open Super-large Crawled Aggregated coRpus (mOSCAR)

mOSCAR is extracted from three Common Crawl dumps from 2023. Due to computational constraints and in order to extract a maximum number of documents for low-resource languages, we extracted all languages from the first dump only. We removed the 6 most high-resource languages from the second dump and only extracted the languages with fewer than 1M documents for the last dump. Table 1 shows a distribution of the total number of languages and their number of documents. To avoid data poisoning (Carlini et al., 2024), we release a hash (sha512) with each mOSCAR image. mOSCAR is composed of 303M documents (200B tokens, 1.15B images) from 163 languages. Figure 2 shows the distribution of images and tokens per document and their joint distribution. As shown in Figure 2a, the mean and median number of images per document is 2 and 3.80.

#docs.	10M	5M	1M	500K	200K	50K	10K	5K	1K
#langs.	10	15	36	46	56	75	119	133	154

Table 1: Number of languages with at least N documents

4.1 Quality vs Diversity

While improving overall data quality, the filtering steps we applied (see Section 3) necessarily have a negative impact on diversity. We therefore study the trade-off between quality and diversity and compare against previously published, well-used datasets.

4.1.1 Text content

Diversity By contruction, mOSCAR is diverse in terms of number of languages, so we focus on the diversity of mOSCAR's English documents and compare against mmc4 (Zhu et al., 2023), OBELICS (Laurençon et al., 2023) and the English subset of WIT (Srinivasan et al., 2021). We compute the Vendi score (Friedman and Dieng, 2023) on a set of SimCSE embeddings (Gao et al., 2021) with a RoBERTa encoder (Liu et al., 2019) to evaluate the content diversity. Since embedding-based diversity metrics target content diversity well but are less relevant for lexical diversity (Tevet and Berant, 2021), we measure lexical diversity via the distinct n-gram ratio (Li et al., 2016). An analysis of the topics (Grootendorst, 2022) found in multiple languages of mOSCAR where we show diverse topics across languages can be found in Appendix A.2.

	Vendi score	Dist. n-gram ratio
mOSCAR	<u>69.05</u> ($\pm$ 0.14)	0.472 ($\pm$ 0.002)
mmc4	67.93 ($\pm$ 0.12)	<u>0.494</u> ($\pm$ 0.002)
OBELICS	58.49 ($\pm$ 0.09)	0.488 ($\pm$ 0.001)
WIT	**73.30** ($\pm$ 0.09)	**0.530** ($\pm$ 0.001)

Table 2: Average text diversity scores ($\pm$ standard error) of text documents.

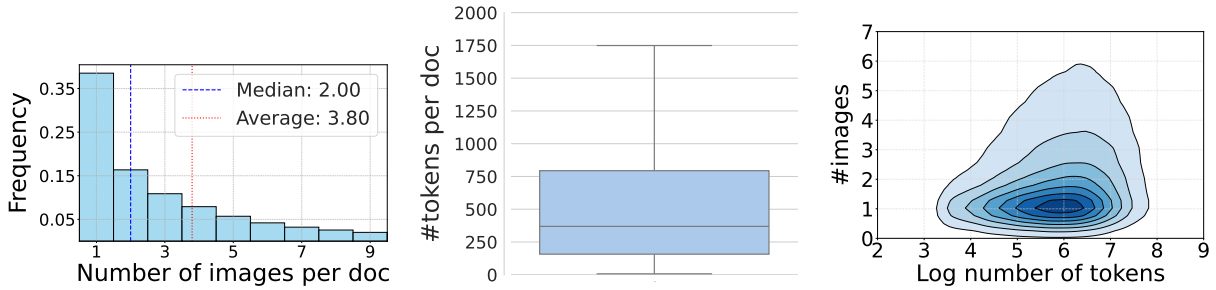
Comparison with other datasets For content diversity, we randomly sample 30M documents for mOSCAR, mmc4 and OBELICS and 3M documents

[15] We do not limit ourselves to caption-like relation and instead allow all types of text-image relation.

[16] `siglip-base-patch16-224` as vision encoder and `nllb-distilled-600M` as text encoder.

[17] We use the open-clip (Ilharco et al., 2021) model version and the transformers (Wolf et al., 2020) library.

[18] We refer to paragraph as the text content in a DOM text node.



(a) Number of images (b) Number of tokens (c) Number of images and tokens

Figure 2: Distributions of numbers of tokens and images per document.

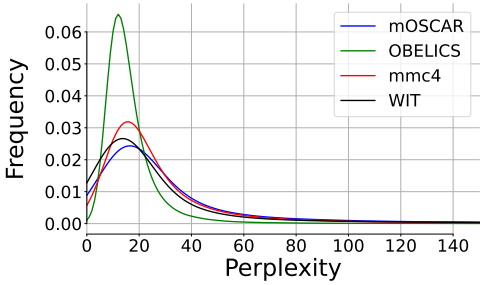


Figure 3: Perplexity of 100K random documents from different datasets.

for WIT and represent the documents by their SimCSE embedding. We compute the Vendi Score with cosine similarity on a randomly sampled subset of 65,536 documents. Table 2 shows that mOSCAR English content is more diverse than mmc4 and OBELICS but less diverse than WIT. For lexical diversity, we randomly sample 3M documents for mOSCAR, mmc4, OBELICS and WIT and compute the distinct n-gram ratio on a subset of 8,192 documents for n from 1 to 4. Table 2 shows that mOSCAR is slightly less lexically diverse than OBELICS and mmc4, while WIT is by far the most diverse.

Quality To evaluate document quality, we focus on English documents and compute their perplexity using Gemma-2B (Gemma Team et al., 2024). Figure 3 shows the kernel density estimation of the distribution of the perplexity of 100K randomly sampled documents from different datasets: mOSCAR is comparable to mmc4 and WIT, while OBELICS appears to be the of the highest quality. mOSCAR is therefore comparable to other interleaved image-text dataset in terms of quality and diversity of its English subset. It is however more diverse than English-only datasets by its multilingual construction and more than 10 times larger than existing multilingual interleaved image-text datasets such as WIT.

4.1.2 Image diversity

Comparison with other datasets We compute the Vendi Score on random samples of images for different datasets, comparing the images from English mOSCAR documents with those from Conceptual Captions (Changpinyo et al., 2021), LAION-400M (Schuhmann et al., 2021) and WIT (Srinivasan et al., 2021). We represent each image by its SigLIP[19] (Zhai et al., 2023) embedding and compute the Vendi score on batches of size 65,536 and a total of 1M images for each dataset. In Table 3, we notice that the set of images in mOSCAR documents are more diverse than images from WIT documents but less diverse than LAION-400M.

mOSCAR	LAION-400M	WIT
55.74 (± 0.16)	**67.59** (± 0.16)	36.14 (± 0.08)

Table 3: Average Vendi score (± standard error) of images sampled from different datasets.

English	All
52.36 (± 0.18)	**54.78** (± 2.29)

Table 4: Average Vendi score (± standard error) of images sampled from mOSCAR (English vs. any language).

Multilingual diversity We also compare the diversity of images from English documents and of images sampled from documents of any language (English included). We use multilingual SigLIP (Chen et al., 2023a) trained on WebLI (Chen et al., 2023b) to compute image embeddings used to get the Vendi score. We again use a batch of size 65,536 and a total of 3M images, and we do not sample multiple images from a same document. For the multilingual setting, we randomly sample 50 languages and an equal number of images for each language to build the batch. As we did not do any image deduplication across languages, we could expect to have less diversity in the multilingual setting. However, Table 4 shows that the set of images is on average more diverse when sampled from all documents than from English-only documents. This means that the distribution of images is not exactly the same across languages, potentially due to cultural differences.

[19]We use `siglip-base-patch16-224`.

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	MaRVL	XVNLI	Multi30K	CoMMuTE
Multi. OF *mOSCAR + cap.*	0	16.91	7.45	26.95	22.33	49.56	33.88	22.91	63.34
	4	34.80	22.18	32.33	26.33	49.64	34.07	23.27	63.22
	8	36.90	23.48	34.24	27.08	**51.48**	**36.60**	23.59	**63.54**
	16	**39.46**	**23.67**	**35.23**	**27.47**	49.84	34.85	**23.85**	62.78
Multi. OF *cap. only*	0	9.57	4.21	8.62	4.01	**49.88**	33.76	0.00	61.36
	4	13.20	9.26	13.45	4.15	49.54	32.04	0.00	61.13
	8	18.00	10.35	12.82	4.88	49.65	33.71	0.01	60.90
	16	19.87	12.07	13.37	4.89	49.79	32.70	0.74	60.25

Table 5: Results averaged over all languages. Multi. OF refers to multilingual Open Flamingo, *mOSCAR + cap.* refers to the model trained on text-image pairs and mOSCAR while *cap. only* refers to the model trained only on text-image pairs. **Bold** is best result.

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	MaRVL	XVNLI	Multi30K	CoMMuTE
Multi. OF (35M) *mOSCAR + cap.*	0	19.07	8.73	25.08	19.64	**49.77**	33.01	22.70	**63.75**
	4	34.32	20.59	31.90	23.90	49.67	36.07	22.79	63.65
	8	36.77	22.15	33.9	24.41	49.72	**37.16**	23.21	63.00
	16	**37.63**	**22.24**	**35.71**	**25.38**	49.73	35.36	23.48	62.77
Multi. OF (35M) *WIT + cap.*	0	9.39	4.67	19.81	14.63	49.71	32.78	**26.99**	56.75
	4	7.68	2.99	25.68	16.12	49.72	33.51	**26.99**	53.27
	8	8.91	3.63	27.06	16.81	49.74	32.77	**26.99**	55.33
	16	9.74	4.14	28.14	16.34	49.74	33.63	**26.99**	54.04

Table 6: Results averaged over all languages and comparison between a model trained on WIT and a checkpoint of multilingual Open Flamingo trained on 35M mOSCAR documents (full model was trained on 50M mOSCAR documents). Both models were trained on 35M documents from their respective training datasets and 70M text-image pairs for fair comparison. Multi. OF (35M) refers to multilingual Open Flamingo trained on 35M documents. **Bold** is best result.

5 Training a multilingual multimodal language model

We train a multilingual Flamingo-like model on mOSCAR that we call multilingual Open Flamingo. As adding captioning data to training data has been shown to improve zero-shot performance (McKinzie et al., 2024), we additionally train on LAION-400M, which we re-captioned using BLIP (Li et al., 2022), filtered with CLIP score (Radford et al., 2021) and translated using distilled NLLB-600M (Visheratin, 2023) following the proportion of languages found in mOSCAR. We use Gemma-2B (Gemma Team et al., 2024) as the underlying language model and we train the model on 50M mOSCAR documents and 100M randomly sampled image-text pairs. We also train a model on 300M image-text pairs, a model trained on 35M WIT (Srinivasan et al., 2021) documents and 70M text-image pairs and a model trained on 50M mOSCAR documents from the English subset and 100M English image-text pairs as comparison baselines. We additionally compare with OpenFlamingo-3B-MPT (Awadalla et al., 2023) as the *translate-test* baseline. The full list of languages for training and the implementation details can be found in Appendix A.6.

5.1 Evaluation setup

We evaluate the models using a broad set of image-text multilingual tasks and benchmarks. We use the IGLUE benchmark (Bugliarello et al., 2022) composed of XVNLI, MaRVL (Liu et al., 2021) to test reasoning, xGQA (Pfeiffer et al., 2022) to test visual question answering capabilities and xFlickr&CO (Young et al., 2014; Karpathy and Fei-Fei, 2015; Yoshikawa et al., 2017) for captioning. We also include Crossmodal-3600 (XM3600) (Thapliyal et al., 2022) and MaXM (Changpinyo et al., 2022) as they cover a broader range of languages. To test to what extent models trained on mOSCAR can perform zero-shot multimodal machine translation (MMT), we also test on Multi30K (Elliott et al., 2016, 2017; Barrault et al., 2018) and CoMMuTE (Futeral et al., 2023). For captioning we compute the CideR (Vedantam et al., 2015) score and we tokenize references and model outputs with the Stanford Core NLP tokenizer for English and Stanza (Qi et al., 2020) tokenizers for other languages. To evaluate Multi30k, we compute BLEU (Papineni et al., 2002) score from Sacrebleu (Post, 2018) with *13a* tokenization and default parameters. We use accuracy for CoMMuTE. More details can be found in Appendix A.6.4.

5.2 Results

Tables 5 and 8 show the average results across all languages. Full results are available in Appendix A.7. We notice that the multilingual OpenFlamingo trained additionally on mOSCAR gets better results than the model trained on captioning data only while having seen fewer image-text pairs during training. More importantly, when increasing the number of few-shot examples from

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	XVNLI
Multilingual OF *mOSCAR + cap.*	0	29.64	42.57	34.24	36.58	34.62
	4	51.47	77.98	37.91	38.13	33.59
	8	56.75	77.64	39.44	**38.52**	**38.75**
	16	**59.89**	78.18	40.09	35.80	36.60
English OF *English mOSCAR + English cap.*	0	32.70	43.75	34.71	36.19	35.82
	4	51.39	75.33	37.48	37.96	34.88
	8	51.44	77.73	39.64	38.35	36.86
	16	59.24	**78.38**	**40.36**	37.35	37.11

Table 7: Results on the English subsets of the test sets and comparison between multilingual Open Flamingo and an Open Flamingo trained on the English subset of mOSCAR and English text-image pairs (English OF). Both models were trained on 50M documents from their respective training datasets and 100M text-image pairs for fair comparison. **Bold** is best result.

	#shots	xGQA	MaXM	MaRVL	XVNLI
OF-3B MPT	0	18.34	7.68	49.75	32.73
	4	22.97	7.82	49.70	35.82
	8	28.57	8.32	49.71	31.29
	16	31.82	9.04	49.72	33.29
Multi. OF *mOSCAR + cap.*	0	30.16	10.06	49.93	34.66
	4	35.55	9.89	48.99	36.10
	8	36.78	10.12	**50.54**	**39.69**
	16	**37.75**	**11.49**	49.57	37.97

Table 8: *Translate-test* results averaged over languages where all benchmarks were translated from local languages into English using Google Translate API. Multi. OF *mOSCAR + cap.* refers to Multilingual Open Flamingo trained on mOSCAR and text-image pairs while OF-3B MPT refers to Open Flamingo (Awadalla et al., 2023) based on MPT (Team, 2023) and trained on mmc4 (Zhu et al., 2023) and English text-image pairs.

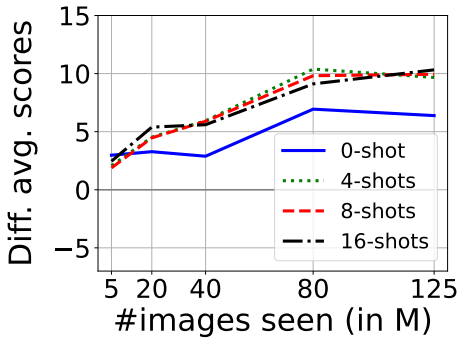


Figure 4: Score differences averaged over benchmarks and languages between the model trained on mOSCAR + text-image pairs and the model trained only on text-image pairs. **Bold** is best result.

0 to 16, it sees gains of on average +6.71 points on VQA benchmarks and +19.39 CideR points on captioning benchmarks. In contrast, the model trained on text-image pairs only sees gains of +2.82 and +9.08 points respectively. In cross-modal machine translation, the model additionally trained on interleaved data is again much better than the one trained on just captioning data, which is not able to translate the Multi30k benchmark at all.[20] Moreover, mOSCAR helps the model to learn to zero-shot disambiguate translations as shown by the improved average score on CoMMuTE (63.54) compared to the model trained on captions only (61.36).

Multilingual Open Flamingo trained on mOSCAR & text-image pairs is also better than OpenFlamingo 3B MPT evaluated on translate test benchmarks[21]. However, we obtain the best results (except for MaXM) by evaluating our multilingual Open Flamingo on the translate-test benchmarks since the underlying language model (Gemma-2B) is far better in English than other languages. We also notice that all models struggle with reasoning classification tasks (MaRVL, XVNLI) where they obtain scores close to random guessing.

Table 6 additionally shows that Multilingual Open Flamingo trained on mOSCAR obtains much better results than the same model trained on WIT for equivalent training data seen during training[22] (except for Multi30K benchmark) which means mOSCAR is better suited than WIT for training multilingual mLLMs. Eventually, Table 7 shows that we don't face a drop in performances in English performances when training the model on 43 languages (multilingual Open Flamingo) in comparison to training it on the English subset of mOSCAR and English text-image pairs.

Additional comparison results with InternVL2 (Chen et al., 2024), Llava-NeXT (Li et al., 2024), PaliGemma (Beyer* et al., 2024) and Idefics2 (Laurençon et al., 2024) can be found in Appendix A.8.

We additionally compare results at different training steps, defined by the number of images seen during training. Figure 4 shows the difference of averaged scores between the model trained on all data and the model trained only on text-images pairs. We notice that the gap first decreases until 20M images seen and keep

[20]Most of the time, the model is not able to follow the prompt and only outputs the end of sequence token.

[21]This means benchmarks were translated from local languages to English, using Google Translate API

[22]We select the checkpoint of multilingual Open Flamingo trained on 35M documents and 70M captions to have fair comparison.

increasing over training at all training steps after that. Particularly, the gap is wider for few-shot learning.

6 Conclusion

We introduced mOSCAR, a large-scale multilingual and multimodal dataset covering 163 languages and composed of 303M documents, 200B tokens and 1.15B images. We showed that mOSCAR is of good quality, diverse and could be used to train a multilingual and multimodal LLM. We also proved that training on mOSCAR led to the emergence of in-context learning capabilities in several languages. We eventually ensured that mOSCAR is as safe as possible by applying a series of filtering steps to remove NSFW and toxic content.

Limitations

We did not conduct any analysis of the biases of mOSCAR as this is challenging in a multilingual setting. As it is crawled from the Internet, it is indeed possible that mOSCAR reflects biases widespread on it. Training a model on mOSCAR must therefore be combined with additional alignment training steps to mitigate potential biases towards groups of people. Nevertheless, by its multilingual nature, mOSCAR is a step towards the inclusion of more languages, cultures, and people in accessing mLLMs.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011014232R1, 2023-AD011014232 and 2023-AD011012254 made by GENCI. It was also partly funded by the last three authors' chairs in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19- P3IA-0001. We deeply thanks the Jean-Zay support team. We also thank Filip Šedivý for insightful discussions regarding the removal of CSAM, Thorn for having provided access to their CSAM detector, Zeeshan Khan for discussions regarding the training of the models and Victoria Le Fourner for having manually checked subsamples of NSFW images.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9)*, pages 1–9, Limerick. Leibniz-Institut für Deutsche Sprache.

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Romain Beaumont. 2021. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset.

Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Fastwarc: optimizing large-scale web archive analytics. *arXiv preprint arXiv:2112.03103*.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, pages 21–29. IEEE Computer Society.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pages 2370–2392. PMLR.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA. USENIX Association.

Nicolas Carlini, Matthew Jagielski, Christopher Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2024. Poisoning web-scale training datasets is practical. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*, pages 179–179, Los Alamitos, CA, USA. IEEE Computer Society.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, Nashville, TN, USA.

Soravit Changpinyo, Linting Xue, Idan Szpektor, Ashish V Thapliyal, Julien Amelot, Michal Yarom, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023a. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2023b. Pali: A jointly-scaled multilingual language-image model. In *Proceedings of the International Conference on Learning Representations*, Kigali, Rwanda.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. 1999. Similarity search in high dimensions via hashing. In *Proceedings of the 25th VLDB Conference*, volume 99, pages 518–529, Edinburgh, Scotland, UK.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning*, pages 4904–4916, online. PMLR.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Vaclav Kosar, Antonín Hoskovec, Milan Šulc, and Radek Bartyzal. 2022. GLAMI-1M: A Multilingual Image-Text Fashion Dataset. In *Proceedings of the 33rd British Machine Vision Conference*, London, UK. BMVA Press.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Gant Laborde. Deep nn for nsfw detection.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems*, volume 36, pages 71683–71702. Curran Associates, Inc.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and

Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730—-19742, Honolulu Hawaii USA.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39 th International Conference on Machine Learning*, pages 12888–12900, Baltimore, Maryland, USA. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2024. PALO: A Polyglot Large Multimodal Model for 5B People. *arXiv preprint arXiv:2402.14818*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *arXiv preprint arXiv:2403.09611*.

Owais Khan Mohammed, Kriti Aggarwal, Qiang Liu, Saksham Singhal, Johan Bjorck, and Subhojit Som. 2023. Bootstrapping a high quality multilingual multimodal dataset for Bletchley. In *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pages 738–753. PMLR.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, UK. Leibniz-Institut für Deutsche Sprache.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, Santiago, Chile.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38 th International Conference on Machine Learning*, pages 8748–8763, online. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Proceedings of the Data Centric AI NeurIPS Workshop 2021*, online.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *arXiv preprint arXiv:2201.11990*.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, online.

Dídac Surís, Dave Epstein, and Carl Vondrick. 2022. Globetrotter: Connecting languages by connecting images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16474–16484, Canada.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively

multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alexander Visheratin. 2023. NLLB-CLIP – train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. 2023. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A Appendix

A.1 mOSCAR languages & statistics

Table 9 shows the number of documents, the number of images, and the number of NLLB tokens for each language of mOSCAR.

A.2 Topic modeling

We analyzed the topic found in different language subsets of mOSCAR with BERTopic (Grootendorst, 2022). We first preprocessed ∼30K randomly sampled documents in tokens using stanza tokenizers (Qi et al., 2020), we then represent each document with their document embedding obtained with a multilingual sentence-bert[23] (Reimers and Gurevych, 2019) before running BERTopic to cluster the documents. Tables 10 to 14 show some of the most salient English-translated strings representing each cluster and their weight (in percentage) in the subsets of mOSCAR. While many clusters are shared between subsets, we can observe that some of them are culturally-specific (i.e. they are related to the culture of the language subset and only appears on this subset). For instance, Table 14 shows topics of the Hindi subset and we observe a cluster related to Hindu gods described by the words 'ganesha', 'shiva' or 'shani dev'. We also observe a cluster related to festivals hosted in India, it is described by the words 'festival', 'dussehra' or 'janmashtami'. We can also observe that among clusters shared between several language subsets, some contain knowledge culturally-specific as exemplified by the cluster related to jewelry in the Hindi subset which contain the word 'mangalsutra' which is a jewel worn by men during weddings in India. Eventually, we can observe that the English subset contains a strong cluster related to the Christian religion. While a similar cluster appears in the Tosk Albanian subset, it is related to Islam. This shows the importance of a multilingual corpora to have data representing more cultures around the world.

A.3 Examples of documents

Figures 5 to 10 provide examples of documents found in mOSCAR for different languages. We can observe the diversity in terms of content of the documents. Figure 5 is a French document describing a specific shot of golf illustrated by images which provide very detailed description of images. Figure 7 shows a Russian document describing an Instagram account with pictures of food with animal heads which is reminiscent of a famous problem of computer vision consisting in classifying images between chihuahuas and muffins.

[23] sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Languages				Statistics		
Lang. name	Code	Family	Script	#documents	#images	#tokens
Acehnese	ace_Latn	Austronesian	Latin	2,159	9,026	1,395,381
Mesopotamian Arabic	acm_Arab	Afro-Asiatic	Arabic	1,282	5,621	704,549
Tunisian Arabic	aeb_Arab	Afro-Asiatic	Arabic	5,933	34,270	2,308,455
Afrikaans	afr_Latn	Indo-European	Latin	50,061	211,876	38,761,504
South Levantine Arabic	ajp_Arab	Afro-Asiatic	Arabic	8,603	69,051	3,869,688
Tosk Albanian	als_Latn	Indo-European	Latin	856,144	2,543,758	441,244,377
Amharic	amh_Ethi	Afro-Asiatic	Ge'ez	39,031	149,739	33,768,732
North Levantine Arabic	apc_Arab	Afro-Asiatic	Arabic	16,198	110,792	8,268,237
Modern Standard Arabic	arb_Arab	Afro-Asiatic	Arabic	3,794,792	14,757,353	3,346,786,610
Najdi Arabic	ars_Arab	Afro-Asiatic	Arabic	52,102	261,275	39,066,487
Moroccan Arabic	ary_Arab	Afro-Asiatic	Arabic	117,957	584,301	188,462,338
Egyptian Arabic	arz_Arab	Afro-Asiatic	Arabic	761,113	3,785,164	635,018,784
Assamese	asm_Beng	Indo-European	Bengali	2,947	7,228	543,676
Asturian	ast_Latn	Indo-European	Latin	87,649	533,723	25,499,269
Awadhi	awa_Deva	Indo-European	Devanagari	8,179	29,142	2,293,620
Central Aymara	ayr_Latn	Aymaran	Latin	10,112	57,294	2,343,403
South Azerbaijani	azb_Arab	Turkic	Arabic	3,411	14,825	3,143,946
North Azerbaijani	azj_Latn	Turkic	Latin	511,832	1,796,046	256,160,442
Bashkir	bak_Cyrl	Turkic	Cyrillic	3,287	12,031	2,600,135
Bambara	bam_Latn	Manding	Latin	3,011	17,666	446,961
Balinese	ban_Latn	Austronesian	Latin	787	4,894	392,978
Belarusian	bel_Cyrl	Indo-European	Cyrillic	60,443	276,672	71,854,171
Bemba	bem_Latn	Atlantic–Congo	Latin	582	3,018	1,021,026
Bengali	ben_Beng	Indo-European	Bengali	204,475	758,222	30,400,395
Bhojpuri	bho_Deva	Indo-European	Devanagari	4,190	18,339	715,786
Banjar	bjn_Latn	Austronesian	Latin	1,764	9,017	1,093,443
Bosnian	bos_Latn	Indo-European	Latin	635,750	2,642,491	423,073,661
Buginese	bug_Latn	Austronesian	Latin	584	2,379	167,459
Bulgarian	bul_Cyrl	Indo-European	Cyrillic	2,578,191	11,601,214	1,736,106,287
Catalan	cat_Latn	Indo-European	Latin	1,132,056	4,638,966	598,942,711
Cebuano	ceb_Latn	Austronesian	Latin	14,924	75,258	10,221,371
Czech	ces_Latn	Indo-European	Latin	3,736,126	12,683,461	2,767,295,966
Central Kurdish	ckb_Arab	Indo-European	Arabic	36,413	135,461	21,622,335
Crimean Tatar	crh_Latn	Turkic	Latin	2,744	10,079	1,173,321
Welsh	cym_Latn	Indo-European	Latin	38,616	155,591	27,237,252
Danish	dan_Latn	Indo-European	Latin	2,020,516	9,214,031	1,207,829,704
German	deu_Latn	Indo-European	Latin	20,265,504	86,393,702	8,315,212,019
Southwestern Dinka	dik_Latn	Nilo-Saharan	Latin	1,233	4,766	1,098,795
Greek	ell_Grek	Indo-European	Greek	4,895,433	15,147,284	2,909,427,055
English	eng_Latn	Indo-European	Latin	51,658,029	205,363,181	32,599,001,993
Esperanto	epo_Latn	Artificial	Latin	23,619	112,577	26,976,847
Estonian	est_Latn	Uralic	Latin	1,022,368	5,108,102	589,045,973
Basque	eus_Latn	Isolate	Latin	682,599	2,914,120	259,930,954
Faroese	fao_Latn	Indo-European	Latin	14,921	56,934	6,579,921
Fijian	fij_Latn	Austronesian	Latin	1,039	4,039	416,670
Finnish	fin_Latn	Uralic	Latin	2,377,155	10,263,171	1,749,904,041
French	fra_Latn	Indo-European	Latin	19,963,542	76,851,982	13,818,099,493
Friulian	fur_Latn	Indo-European	Latin	15,823	120,878	2,550,209
Nigerian Fulfulde	fuv_Latn	Atlantic-Congo	Latin	919	4,281	264,234
West Central Oromo	gaz_Latn	Afro-Asiatic	Latin	3,399	9,071	1,640,693
Scottish Gaelic	gla_Latn	Indo-European	Latin	19,638	105,937	13,119,348
Irish	gle_Latn	Indo-European	Latin	60,303	267,562	45,341,371
Galician	glg_Latn	Indo-European	Latin	410,489	1,696,763	197,685,077
Guarani	grn_Latn	Tupian	Latin	207,800	1,038,296	48,610,979
Gujarati	guj_Gujr	Indo-European	Gujarati	21,916	87,805	3,202,096
Haitian Creole	hat_Latn	Indo-European	Latin	105,777	667,801	34,261,838
Hausa	hau_Latn	Afro-Asiatic	Latin	21,850	81,141	11,807,898
Hebrew	heb_Hebr	Afro-Asiatic	Hebrew	1,098,800	4,708,947	859,238,720
Hindi	hin_Deva	Indo-European	Devanagari	543,928	1,745,222	118,903,998
Chhattisgarhi	hne_Deva	Indo-European	Devanagari	832	3,908	205,345
Croatian	hrv_Latn	Indo-European	Latin	1,689,553	8,315,237	998,928,993
Hungarian	hun_Latn	Uralic	Latin	3,515,058	15,293,132	2,811,446,583
Armenian	hye_Armn	Indo-European	Armenian	336,285	1,126,920	199,883,484
Igbo	ibo_Latn	Atlantic-Congo	Latin	7,089	41,672	3,014,602
Ilocano	ilo_Latn	Austronesian	Latin	7,076	59,327	832,454
Indonesian	ind_Latn	Austronesian	Latin	6,644,918	16,237,247	2,895,956,979

Languages				Statistics		
Lang. name	Code	Family	Script	#documents	#images	#tokens
Icelandic	isl_Latn	Indo-European	Latin	239,195	1,003,522	131,308,802
Italian	ita_Latn	Indo-European	Latin	12,812,932	47,011,085	8,144,757,759
Javanese	jav_Latn	Austronesian	Latin	18,192	100,952	15,206,708
Japanese	jpn_Jpan	Japonic	Kanji	14,154,575	23,435,549	8,539,956,266
Kabyle	kab_Latn	Afro-Asiatic	Latin	6,101	33,923	1,781,992
Kannada	kan_Knda	Dravidian	Kannada	9,373	33,147	1,206,651
Kashmiri	kas_Arab	Indo-European	Arabic	1,498	5,284	3,384,394
Georgian	kat_Geor	Kartvelian	Georgian	353,471	1,300,710	274,042,522
Kazakh	kaz_Cyrl	Turkic	Cyrillic	248,403	718,126	138,597,176
Halh Mongolian	khk_Cyrl	Mongolic	Cyrillic	123,789	505,098	83,628,495
Khmer	khm_Khmr	Austroasiatic	Kher	23,348	116,437	2,915,205
Kinyarwanda	kin_Latn	Atlantic-Congo	Latin	20,381	108,280	10,268,334
Kyrgyz	kir_Cyrl	Uralic	Cyrillic	51,221	194,092	33,981,180
Northern Kurdish	kmr_Latn	Indo-European	Latin	34,593	142,634	21,972,155
Korean	kor_Hang	Koreanic	Hanja	2,614,038	13,562,957	2,000,344,511
Lao	lao_Laoo	Kra-Dai	Lao	49,925	205,452	30,098,274
Ligurian	lij_Latn	Indo-European	Latin	3,581	26,740	1,046,463
Limburgish	lim_Latn	Indo-European	Latin	70,099	443,903	25,465,590
Lingala	lin_Latn	Atlantic-Congo	Latin	6,304	41,400	1,580,536
Lithuanian	lit_Latn	Indo-European	Latin	1,673,790	8,772,570	1,153,604,941
Lombard	lmo_Latn	Indo-European	Latin	14,053	61,359	6,270,646
Latgalian	ltg_Latn	Indo-European	Latin	5,174	21,062	2,903,043
Luxembourgish	ltz_Latn	Indo-European	Latin	27,946	142,470	13,925,521
Ganda	lug_Latn	Afro-Asiatic	Latin	1,475	4,118	688,308
Mizo	lus_Latn	Sino-Tibetan	Latin	7,009	22,630	4,106,536
Standard Latvian	lvs_Latn	Indo-European	Latin	857,757	3,937,940	578,441,751
Magahi	mag_Deva	Indo-European	Devanagari	290	1,088	94,031
Malayalam	mal_Mlym	Dravidian	Malayalam	11,203	44,417	1,420,906
Marathi	mar_Deva	Indo-European	Devanagari	43,720	142,001	6,164,176
Minangkabau	min_Latn	Austronesian	Latin	1,523	7,300	447,320
Macedonian	mkd_Cyrl	Indo-European	Cyrillic	539,149	1,841,846	304,592,615
Maltese	mlt_Latn	Afro-Asiatic	Latin	56,666	327,331	27,114,870
Maori	mri_Latn	Austronesian	Latin	20,840	114,680	24,524,962
Burmese	mya_Mymr	Sino-Tibetan	Mon	6,575	36,661	406,016
Dutch	nld_Latn	Indo-European	Latin	16,890,074	64,609,055	9,493,533,101
Norwegian Nynorsk	nno_Latn	Indo-European	Latin	138,384	701,972	57,812,652
Norwegian Bokmål	nob_Latn	Indo-European	Latin	2,192,012	9,534,178	1,267,421,216
Nepali	npi_Deva	Indo-European	Devanagari	28,042	116,363	2,892,865
Nyanja	nya_Latn	Atlantic-Congo	Latin	11,749	65,324	8,513,823
Occitan	oci_Latn	Indo-European	Latin	61,681	323,632	21,029,975
Odia	ory_Orya	Indo-European	Odia	3,759	14,373	340,695
Pangasinan	pag_Latn	Austronesian	Latin	1,045	7,770	270,363
Eastern Panjabi	pan_Guru	Indo-European	Gurmukhi	10,857	44,440	1,821,511
Papiamento	pap_Latn	Indo-European	Latin	29,564	177,229	7,396,392
Southern Pashto	pbt_Arab	Indo-European	Arabic	31,854	107,563	27,623,486
Western Persian	pes_Arab	Indo-European	Arabic	6,995,368	24,998,370	6,061,794,870
Plateau Malgasy	plt_Latn	Austronesian	Latin	32,119	119,506	28,542,084
Polish	pol_Latn	Indo-European	Latin	14,492,239	60,362,860	10,994,239,010
Portuguese	por_Latn	Indo-European	Latin	8,033,406	26,058,040	4,639,089,792
Dari	prs_Arab	Indo-European	Arabic	421,097	2,101,038	399,037,437
Ayacucho Quechua	quy_Latn	Quechuan	Latin	1,248	10,038	322,112
Romanian	ron_Latn	Indo-European	Latin	5,131,444	17,790,793	3,484,865,185
Rundi	run_Latn	Atlantic-Congo	Latin	17,798	55,060	8,140,230
Russian	rus_Cyrl	Indo-European	Cyrillic	15,753,144	68,786,134	18,196,141,357
Sango	sag_Latn	Atlantic-Congo	Latin	724	4,564	181,876
Sicilian	scn_Latn	Indo-European	Latin	27,388	164,772	17,535,500
Sinhala	sin_Sinh	Indo-European	Sinhalese	44,963	179,082	11,413,044
Slovak	slk_Latn	Indo-European	Latin	2,979,681	14,894,160	1,951,406,321
Slovenian	slv_Latn	Indo-European	Latin	1,456,026	7,106,291	928,101,642
Samoan	smo_Latn	Austronesian	Latin	11,024	62,358	11,672,900
Shona	sna_Latn	Atlantic-Congo	Latin	7,400	41,385	5,276,139
Sindhi	snd_Arab	Indo-European	Arabic	20,615	70,992	16,686,668
Somali	som_Latn	Afro-Asiatic	Latin	58,151	209,905	31,093,227
Southern Sotho	sot_Latn	Atlantic-Congo	Latin	7,474	41,714	5,876,842
Spanish	spa_Latn	Indo-European	Latin	22,218,630	76,372,709	13,882,047,139
Sardinian	srd_Latn	Indo-European	Latin	336,476	2,220,976	68,281,992
Serbian	srp_Cyrl	Indo-European	Cyrillic	593,332	2,251,042	394,477,097

Languages				Statistics		
Lang. name	Code	Family	Script	#documents	#images	#tokens
Sundanese	sun_Latn	Austronesian	Latin	16,438	89,379	9,549,957
Swedish	swe_Latn	Indo-European	Latin	3,231,753	10,558,719	1,748,495,813
Swahili	swh_Latn	Atlantic-Congo	Latin	96,770	365,792	52,827,863
Silesian	szl_Latn	Indo-European	Latin	7,846	47,313	3,022,502
Tamil	tam_Taml	Dravidian	Tamil	30,202	149,837	4,234,345
Tatar	tat_Cyrl	Turkic	Cyrillic	34,489	133,014	22,255,423
Telugu	tel_Telu	Dravidian	Telugu	16,107	54,100	1,633,579
Tajik	tgk_Cyrl	Turkic	Cyrillic	119,383	395,470	87,519,228
Tagalog	tgl_Latn	Austronesian	Latin	140,922	628,210	95,285,900
Thai	tha_Thai	Kra-Dai	Thai	1,799,735	6,603,060	807,374,946
Tigrinya	tir_Ethi	Afro-Asiatic	Ge‘ez	2,622	8,601	1,699,272
Tok Pisin	tpi_Latn	Indo-European	Latin	785	5,888	97,298
Turkmen	tuk_Latn	Turkic	Latin	12,372	54,002	9,650,172
Turkish	tur_Latn	Turkic	Latin	4,448,111	12,304,912	2,356,627,784
Twi	twi_Latn	Atlantic-Congo	Latin	286	2,041	78,227
Uyghur	uig_Arab	Turkic	Arabic	10,614	41,367	6,602,690
Ukrainian	ukr_Cyrl	Indo-European	Cyrillic	2,689,369	10,842,572	1,909,330,669
Urdu	urd_Arab	Indo-European	Arabic	403,245	1,224,175	236,356,788
Northern Uzbek	uzn_Latn	Turkic	Latin	113,772	581,861	81,808,833
Venetian	vec_Latn	Indo-European	Latin	122,390	763,029	24,081,966
Vietnamese	vie_Latn	Viet-Muong	Latin	12,296,989	46,339,341	11,462,111,787
Wolof	wol_Latn	Atlantic-Congo	Latin	2,152	9,351	367,848
Xhosa	xho_Latn	Atlantic-Congo	Latin	13,620	80,748	14,566,904
Eastern Yiddish	ydd_Hebr	Indo-European	Hebrew	12,275	56,421	17,078,751
Yoruba	yor_Latn	Atlantic-Congo	Latin	10,148	49,474	8,346,193
Yue Chinese	yue_Hant	Sino-Tibetan	Hant	28,478	172,592	21,579,579
Chinese (Simplified)	zho_Hans	Sino-Tibetan	Hanzi	8,326,440	29,575,591	5,199,137,981
Chinese (Traditional)	zho_Hant	Sino-Tibetan	Hant	3,796,336	15,514,804	2,617,463,485
Standard Malay	zsm_Latn	Austronesian	Latin	864,831	3,651,754	384,708,004
Zulu	zul_Latn	Atlantic-Congo	Latin	13,089	73,167	9,654,461

Table 9: Languages & Statistics

A.4 Heuristics to increase the quality of documents

We use a set of heuristics to improve the quality of the documents by discarding some text nodes. We first consider text nodes to be written in Latin scripts if more than 50% of the characters are Latin. In detail, we discard the text node if:

1. It is empty.
2. It contains fewer than 5 bytes for Latin scripts and fewer than 15 bytes for non-Latin scripts.
3. More than 30% of the characters are digits.
4. It contains more than one date.
5. It contains the sequence "lorem ipsum".
6. The ratio of non-alphabetic characters is superior to 0.33.
7. The symbols '{' or ''}' are in the text.
8. The symbols '$\geq$', '$\leq$', '>' or '<' are more than 2 times in the text.
9. "Follow us", "javascript", "copyright" or "©" are in the text.
10. The ratio of capitalized letters is superior to 0.2.
11. The text exactly matches with "comment", "facebook", "instagram", "twitter", "rss", "newsletter", "share" or "follow us".
12. A character is more than 33% of the total number of characters in the string.

We then also apply some filters to clean the text as much as possible:

1. Remove URLs from all documents.
2. Normalize consecutive special characters ('\t', '\n', '#', '/', '$', ')', '(', '[', ']', '!', '?', '%', '<', '>') to keep only one.

Following previous steps, we keep the text node if it is superior to 5 bytes and we keep the final document if it is superior to 100 bytes.

A.5 Text-Image similarity and DOM Tree

As we rely on the DOM Tree to build the documents and the order of appearance of the nodes could differ from HTML rendering, we attempt to assess to what extent it is a relevant way of constructing a multimodal document. To do so, we rely on the results of the text-image joint filtering step where we compute the ranks of relevant text nodes (resp images) for each image. We plot the distribution of the closest most relevant node for each modality in Figures 11a and 11b. We notice that the most relevant node to either a text node or an image is their closest node in the DOM tree. The cumulative distribution function of the distribution of the closest node reaches 25% for nodes positioned between -5 and 5, which confirms the relevance of using the DOM tree to represent a document.

A.6 Implementation details

A.6.1 Text deduplication parameters

Following previous work, we near-deduplicate documents using MinHashLSH. We first vectorize the documents using HashingVectorizer from scikit-learn with 2,097,152 features computed on 4-grams and 5-grams within word boundaries. We then compute MinHashes from those vectors with 256 permutations and we finally run Locality Sensitive Hashing with a threshold Jaccard Similarity of 0.8 for finding near-duplicates.

A.6.2 Removing Personal Identifiable Information

We used regular expressions to detect and remove PII in documents. More precisely, we used:

email address:
```
^[\w\.]+@[\w-]+\.[\w-]{2,4}$
```

phone number:
```
^\+?\d{1,3}?[-.\s]?\(?\d{1,4}?\)?
[-.\s]?\d{1,4}[-.\s]?\d{1,4}
[-.\s]?\d{1,9}$
```

credit card number:
```
^(?:4[0-9]{12}(?:[0-9]{3})?|5[1-5]
[0-9]{14}|3[47][0-9]{13}|3(?:0[0-5]|
[68][0-9])[0-9]{11}|6(?:011|5[0-9]
{2})[0-9]{12}|(?:2131|1800|35
\d{3})\d{11})$
```

IP address:
```
^(?:25[0-5]|2[0-4][0-9]|1[0-9]{2}|
[1-9][0-9]|\d)\.(?:25[0-5]|2[0-4]
[0-9]|1[0-9]{2}|[1-9][0-9]|\d)\.
(?:25[0-5]|2[0-4][0-9]|1[0-9]{2}|
[1-9][0-9]|\d)\.(?:25[0-5]|2[0-4]
[0-9]|1[0-9]{2}|[1-9][0-9]|\d)$
```

passport number: `^[A-Z0-9]{6,15}$`

For images, we detect faces in the images and distribute the bounding boxes coordinates. More precisely, all the images are resized to have a maximum of width and height of 256, keeping aspect ratio. The bounding boxes coordinates are therefore computed given this image size but can be extrapolated if images are downloaded in a higher resolution.

A.6.3 Training implementation details

We train multilingual OpenFlamingo on mOSCAR and multilingual text-image pairs. We use a batch of size 64 for mOSCAR and 128 for captioning data, limiting the number of tokens to 256 for mOSCAR and 32 for captioning data. Similarly to Flamingo and OpenFlamingo, text tokens can only attend to the previous image in the sequence. To increase diversity in the training batch, we randomly reject 2/3 of the documents if they contain only one image. We limit the maximum number of images in a sequence to 8. We randomly sample

Relevant words describing the topics	Topic representation (in %)
recipe, sauce, cheese, recipes, chicken, add, delicious, food, cook, minutes	7.17
game, games, players, gaming, play, gameplay, pc, playing, player, playstation	2.1
dog, dogs, cat, pet, cats, pets, puppy, breed, puppies, animal	1.94
god, jesus, church, christ, faith, lord, bible, him, christian, holy	1.82
shirt, jacket, jeans, cotton, wear, shirts, fit, men, size, waist	1.68
estate, property, mortgage, real, home, house, buyers, market, properties, homes	1.6
art, artist, artists, painting, gallery, paintings, works, arts, his, exhibition	1.59
card, cards, love, gift, cute, christmas, fun, stamp, halloween, easter	1.57
covid, 19, vaccine, coronavirus, virus, cases, health, vaccinated, vaccination, vaccines	1.28
album, band, music, song, songs, rock, release, albums, his	1.27
data, cloud, server, software, aws, application, business, management, cluster, configuration	1.21
shipping, item, delivery, ankle, brace, items, order, days	1.14
diamond, jewelry, ring, carat, gold, rings, earrings, silver, necklace	1.13
books, book, read, she, reading, author, novel, story	1.13
dress, fashion, dresses, wear, style, skirt, wedding, outfit, love	1.12
bedroom, apartments, room, property, apartment, kitchen, floor, home, spacious	1.03
trump, president, election, biden, republican, democrats, senate, republicans, court	1.0
bitcoin, crypto, cryptocurrency, blockchain, ethereum, tokens, token, price, cryptocurrencies, nft	0.94
life, tarot, yourself, love, feel, myself, mind	0.93
sound, hearing, audio, speakers, noise, headphones, bluetooth, speaker, microphone, wireless	0.9
trail, lake, park, mountain, hike, trails, hiking, canyon, river, yosemite	0.85
cookies, website, cookie, settings, resume, preferences, disable, browser, user, information	0.85
police, said, arrested, man, officers, crime, county	0.84
skin, acne, skincare, sunscreen, face, treatment, facial, pores, oil, hyamax	0.83
casino, gambling, games, slot, casinos, online, bonus, poker, slots, players	0.82
lighting, light, lights, led, lamp, bulb, lamps, bulbs, wall, fixtures	0.76
league, club, chelsea, season, his, football, cup, players, player	0.76
wedding, weddings, venue, couples, dj, guests, day, reception, ceremony, bride	0.76
dental, teeth, tooth, dentist, smile, dentistry, whitening, oral, implant, dentists	0.72
cbd, cannabis, marijuana, hemp, thc, gummies, cannabinoids, medical, oil, recreational	0.69
shoes, shoe, boots, socks, nike, sneaker, heel, foot, running, boxing	0.68
stocks, inflation, stock, market, investment, gold, investors, fund, trading, funds	0.67
car, cars, engine, rear, porsche, toyota, mercedes, electric, suv	0.65
fitness, workout, exercise, gym, trainer, exercises, training, strength, body, workouts	0.63
divorce, attorney, law, lawyer, legal, court, lawyers, attorneys, injury, case	0.59
steel, nailer, arm, drill, machine, wrench, tool, saw, palm	0.59
security, cyber, ransomware, cybersecurity, attacks, malware, phishing, attack, data, hackers	0.59
furniture, chair, chairs, sofa, table, design, style, dining, comfort, leather	0.5
her, she, actress, star, kristina, instagram, lorena, grikaite, kardashian	0.49
testosterone, vitamin, supplements, weight, blood, body, supplement, levels, calcium, muscle	0.49
bag, bags, backpack, stethoscope, pockets, strap, tote, zipper, purse, leather	0.48
energy, carbon, wind, renewable, emissions, gas, coal, power, electricity, climate	0.47
ball, tennis, player, sports, sport, players, backhand, coach, athletes, drills	0.47
phone, samsung, smartphone, camera, galaxy, oneplus, realme, battery, schematic, android	0.47
watches, replica, watch, rolex, dial, clock, patek, swiss, fake, wholesale	0.47
fundraising, volunteer, donors, volunteers, donor, charity, charities, community, volunteering, imdsa	0.46
parking, traffic, pedestrian, rail, transport, transportation, street, mobility, public, city	0.45
race, f1, racing, lap, formula, drivers, season, car, championship, driver	0.44
pain, chiropractic, joint, ankle, spinal, muscles, arthritis, spine, symptoms, joints	0.41
wine, wines, winery, bordeaux, vineyard, tasting, sauvignon, grapes, vineyards, palate	0.41

Table 10: BERTopic modeling on the **English** subset of mOSCAR.

Relevant words describing the topics	Topic representation (in %)
Kosovo, Serbia, European, taken, Kurti, Serbian, Serbia	4.65
Ukraine, Russian, Russia, Putin, war, Ukrainian, military	4.44
Covid, 19, vaccine, cases, health, patients, coronavirus	3.72
Luizi, Kiara, Big, Brother, VIP, Olta, stuck	3.17
apartment, rent, floor, m2, lease, housing, located, area, sale	2.51
letter, language, book, literary, poetry, book, Albanian, writer	1.81
add, spoon, eggs, recipe, sugar, pour, oven, oil, flour	1.46
photo, she, Instagram, hers, bikini, Tarja, model, follows, dress	1.23
Allah, greeting, Quran, prophet, interpretation, exegesis, Islam, prophets	1.19
court, property, court, trials, KPK, reevaluation, appeal, judges, declaration, prosecution	1.16
horoscope, sign, signs, you, stars, zodiac	0.99
music, festival, stage, KultPlus, artist, culture, musical	0.98
prison, court, criminal, imprisonment, pretrial detention, measure, sentenced, crime	0.85
Inter, Inter Milan, Milan, AC Milan, Gazzetta, Pioli	0.82
Rama, Edi, prime minister, opposition, PD, McGonigal, Basha, justice, PS	0.82
Trump, Biden, Donald, president, Republican, president's, Joe, American, White, president	0.81
accident, occurred, consequence, road, car, injured, type, crashed	0.78
Berisha, PD, Sali, Berisha, Rama, opposition, elections, Rama, Edi, party's	0.74
exhibition, art, museum, exhibition, culture, artist, KultPlus	0.72
knife, gold, weapon, police, killed, incident, fired, event, victim	0.7
water, river, residents, floods, rain, drinking	0.7
women, gender, violence, LGBT, friend, violence, family	0.68
migrants, asylum, refugees, permits, Albanian, passport, exchange, asylum seeker, 000, passport	0.63
students, UET, academic, university, university's, Erasmus, universities, faculty, study, UBT	0.63
tourists, tourism, tourist, travel, foreign, destination, Albanian, visitors	0.60
doctors, patients, health, clinic, hour, sore, clinic, Dr	0.57
firefighters, fire, flame, fires, extinguishing, hotspots	0.56
Ronaldo, Cristiano, CR7, Portuguese, United, Manchester, Messi	0.56
music, musical, Adriano, Celentano, musical, Tari, top, titled	0.56
weather, temperatures, forecast, cloudy, rain, snow, degrees	0.56
Mercedes, car, Audi, sale, BMW, Aston, vehicles, Benz, sold	0.52
email, site, comment, browser, save, address, fields, marked, required	0.5
archaeologists, archaeological, Roman, ancient, city, archaeological, restored, Durres, century	0.49
NASA, Earth, space, science, planet, boundary, solar, Apollo, planets	0.47
Albanian, November, flag, nations, independence, year, day, Vasfije	0.46
oil, liter, price, fuel, lek prices, gasoline, petrol, barrel	0.45
PD, Democratic, party's, party, PS, elections, vote, party, Socialist, Democrat	0.44
city, located, park, rock, water, natural	0.44
Turkey, earthquake, Syria, magnitude, collapsed	0.43
airline, plane, passengers, flight	0.41
road, buses, urban, bike, works, city's, urban, municipality, Veliaj, transportation	0.4
iPhone, Apple, Pro, Max, iOS, iPad, 15, apps, 14, dollars	0.4
north, North Macedonia, Bulgarian, Bulgaria, Macedonian, Macedonia, Macedonians	0.4

Table 11: BERTopic modeling on the **Tosk Albanian** subset of mOSCAR.

Relevant words describing the topics	Topic representation (in %)
Match, league, Madrid, Barcelona, match, football, Real	9.86
Allah, Sheikh, Mohammed, occupation	4.24
Allah, cleaning, homes, jurisprudential, buying, Kuwait, furniture	2.78
Game, games, download, casino, Android, play, computer	2.5
Sarai, occupation, Jerusalem, Palestinian, Gaza, next, Palestine	2.08
Corona, health, virus, infection, Covid, 19, vaccine	1.98
Ukraine, Russia, Russian, Putin, Moscow, Kiev, forces, Ukrainian	1.86
Fridge, LED, electric, watts, stainless steel, product, resistant, corrosion	1.66
Movie, film, TV series, actress, actor, series, cinematic, ceremony, cinema	1.54
Hospital, medical, health, medical, care, doctors, diseases, doctor	1.47
Car, cars, Toyota, Hyundai, electric, Mercedes, Ford, Kia	1.44
Capital, Daraya, compound, real estate, mall, meters, administrative, units, for sale	1.4
Data, workers, feasibility, project, work, study, customers, employees	1.25
Syrian, Syria, ISIS, Aleppo, forces	1.01
Saudi, national, Saudi Arabia, Kingdom, Arabia, Muhammad	0.98
Aviation, airport, flight, airplane, aircraft, air, airlines	0.91
Women, women's, violence, rights, society, feminist, affected, victims, origin	0.83
iPhone, iPhone, Apple, iOS, phone, iPad, apps	0.83
Space, NASA, Mars, moon, planet, planets, scientists, telescope	0.81
Iran, Tehran, nuclear, agreement, Washington, United States, American	0.79
Tourism, hotel, island, Maldives, beach, Malaysia, shores	0.71
Skin, cream, face, skin, oil, care, your skin, treatment, oily	0.65
Chocolate, spoon, meat, recipes, cake, butter, method, recipe, bowl	0.62
College, engineering, science, education, university, sciences, department, school	0.6
Temperature, weather, degrees, forecast, airport, temperature, west, regions	0.6
Police, suspect, arrest, drugs, crime, general, city, prosecutor	0.59
Spine, column, knee, blood, body, fever, pain, muscle	0.59
Istanbul, Turkey, Turkish, apartments, for sale, real estate, nationality, Şişli, property	0.59
Pregnancy, fetus, womb, birth, monthly, cycle, doctor, progesterone, symptoms	0.57
Sudan, Ethiopia, Tigray, Ethiopian, Sudanese, Ethiopian, Renaissance, army, Sudanese	0.55
Germany, Sweden, Europe, asylum, Merkel, immigration, refugee	0.54
Insects, pest control, ants, pests, spray, pesticides, cockroaches, company, white	0.53
University, Baniyas, scholarships, students, universities, study	0.51
Gold, price, carat, today, pound, grams, dealings	0.5
Medicine, drug, prescription, dosage, side effects, doctor, treatment, illness	0.5
Rooms, bedrooms, decorations, modern, wallpaper, furniture, sleeping	0.49
Hotels, activities, trips, close, hotel, tours	0.45
Samsung, Galaxy, phone, specifications, S23, 5G, phones	0.44
Fire, accident, firefighters, accident, Greece, outbreak, injured, forest	0.42
Trump, Donald, Biden, elections, American, former, White House	0.41
Technician, painter, Kuwait, repair, installation, AC, satellite, changing, provided	0.4
Hair, hair loss, shampoo, scalp, grow, hair, hair treatment	0.39
Education, secondary, studies, students, exams, schools, cumulative, average	0.39
Elections, voting, election, Myanmar, council, parties, electoral	0.39
Buyer, seller, shipment, equipment, Alibaba, help, charger, suppliers, machine, tank	0.39
India, ambassador, cooperation, Libya, ministers, Modi	0.39

Table 12: BERTopic modeling on the **Arabic** subset of mOSCAR.

Relevant words describing the topics	Topic representation (in %)
team, match, child, event, olympic, goal, championship, win, champion	10.84
school, students, education, children, second, student	4.11
elections, syriza, mitsotakis, government, ND, response, prime minister, parliament, no	2.75
order, shipment, product, payment, purchase, cash on delivery, receipt, donation, you, shipping fees	2.15
god, god's, church, holy, christ, priest, prayer, divine	1.8
gold, necklace, earrings, ring, bracelet, jewelry, silver, steel	1.79
book, writer, history, published, read, literature, write, novel	1.64
game, play, children, two, ball, games	1.36
ukraine, russia, russian, force, usa, war, russia	1.21
shoes, t-shirt, adidas, nike, sneakers, men, athletic, running	1.18
song, music, album, single, music, two, rock	1.16
skin, face, cream, body, oil, oily, skin care	1.15
traffic, vehicle, road, lane, road, vehicles, street	1.11
mattress, sofa, pillow, bed, chair, furniture, armchair, bedroom, color	1.04
recipe, chicken, recipe, juicy, add, chicken, mix, season	1.02
christmas, birth, christmas, december, gift, festive, tree, holiday	0.98
turkey, erdogan, turkish, greece, home, democracy, istanbul	0.93
moon, planet, nasa, sky, galaxy, earth, telescope, satellite, lunar, space	0.83
tourism, tourist, greece, destination, trips, hotel	0.81
mayor, march, meeting, municipal, council, committee, session, members	0.81
ships, boats, port, cruise, shipping, sea, cruise ship, peiraeus	0.78
fire, fire brigade, wildfire, fire extinguishers, area, blaze, fire vehicles, firemen	0.78
beach, island, water, village, white, place	0.77
art, painting, works, artist, museum, exhibition, artists	0.76
car, bmw, model, new, nissan, audi, mobile, video	0.75
vaccine, vaccination, covid, doses, health, coronavirus, disease	0.74
led, lighting, lamp, simple, lights, white, lighting, hanging, ceiling, lamps	0.72
you, yourself, feeling, life, thoughts, live, feel	0.68
cotton, cotton, fabric, design, cotton	0.67
purse, wallet, leather, case, bag, items	0.65
exercise, gym, workout, practice, body, strength, pilates, muscle	0.64
weather, rainy, temperature, clouds, cold, storms, celsius, wind	0.63
company, product, sunlight, quality, service, construction, production, service	0.62
plants, seeds, plant, grow, flowers, neem, soil, wild	0.62
europe, euro, greek, eu, policy, government, eurozone, economy	0.61
wine, wines, vineyard, winery, variety, grapes, alcohol	0.57
pump, burner, gas, water, temperature, heating, air, water heater	0.55
headphones, bluetooth, speakers, sound, usb, wireless, audio, sound	0.54
movie, series, netflix, characters, director, episode, season, naruto	0.54
apartment, for sale, bedroom, location, new, living, for rent, kitchen	0.49
restaurant, kitchen, chef, dish, food, menu, restaurant, tavern	0.49
diet, weight, food, loss, calories, pounds, healthy	0.48
history, news, breakfast, syriza, greek, government, politics, news	0.47
video, tv, sony, iptv, dvd, photo, movie, camera	0.47
dog, dogs, pet, animal, dog, puppies, leash, friend, pets	0.46
inflation, rates, euro, increase, prices, interest, market, economy	0.42

Table 13: BERTopic modeling on the **Greek** subset of mOSCAR.

Relevant words describing the topics	Topic representation (in %)
loan, bank, paytm, sbi, upi, credit, personal, sunal, atm, account	3.61
electric, tata, mahindra, hyundai, maruti, bike, suv	3.16
recipe, make, paneer, chicken, best, ingredients, potato	2.45
recruitment, notification, availability, 2023, vacancy, age, position, fill	1.39
whatsapp, sap, fm, gb, chat, gbwhatsapp, message, group, feature	1.28
movie, khan, film, jawan, trailer, release, series, jaari, cast, collection	1.25
bitcoin, cryptocurrency, token, crypto, currency, ethereum, coin, blockchain, binance, dogecoin	1.25
samsung, 5g, oneplus, realme, redmi, nokia, oppo, vivo, galaxy, xiaomi	1.19
vs, ind, aus, odi, one-day, wicket, match, runs, australia, cup	1.1
stock, market, trading, intraday, share, ratio, demat, year, buy	0.88
wish, life, dog, equity, confidence, morning, suprabhat, habit, flirting, self	0.86
business, ideas, idea, manufacturing, making, small, entrepreneur, start, recycling, people	0.85
aadhar, aadhaar, uidai, half, card, otp, update, number, pan, bar	0.72
designs, design, taarak, mehta, alto, jewellery, mangalsutra, galas, blouse, earrings	0.72
rlalibaba, equipment, mash, ravi, gati, jak, shakash, ens, quot, woodworking	0.69
iphone, apple, aif, apple, 14, ipad, 15, ios, xs, mac	0.67
biography, chay, kanika, mann, one, alia, bhatt, tulsidas, family, age	0.63
reply, cancel, leave, rac, berth, approval, seat, resignation, vishesh, vyakti	0.6
movies, movie, download, hollywood, tamil, dubbed, telugu, hd	0.57
shayari, attitude, love, dil, sad, status, romantic, nahin, mohabbat, girlfriend	0.56
ganga, festival, dussehra, jayanti, ekadashi, sav, janmashtami, bihu, raksha, dashera	0.55
computer, processor, tar, virus, system, software, year, input, output	0.54
yatan, thal, places, tourist, garan, yatak, kuchinda, jaipur, tourism, barot	0.54
course, bba, llb, pharma, bsc, ba, bachelor, mba, ed, entrance	0.52
ration, vein, dhaman, artery, card, shan, list, epds, fcs, nfsa	0.51
jio, phone, tune, caller, recharge, sim, plan, call, reliance, jiotv	0.5
youtube, channel, video, youtuber, sponsorship, shorts, subscriber, videos, views	0.5
train, railway, railways, trains, station, vay	0.46
freelancing, money, kam, earn, paise, online, kamaye, fiverr, earning, freelancer	0.45
weight, creatine, loss, pathri, at, diet, ghat, patal	0.44
photo, lightroom, background, image, photoshop, edit, presets, kag, preset, jpeg	0.43
shram, card, eshram, ram, check, payment, uan, shramik, otp, gov	0.42
election, nasbah, assembly, bjp, president, elections, powers, veto, gujarat	0.4
haven, matching, suppliers, supplier, verified, let, found, alibaba, equipment	0.39
stotram, lord, shiva, shani, dev, ganesha, gan, bhagav, kath	0.39
kisan, pm, beneficiary, farmer, pm kisan, corner, kyc, status, nidhi, samman	0.38

Table 14: BERTopic modeling on the **Hindi** subset of mOSCAR.

Autour des greens notre créativité est souvent mise à rude épreuve. En effet les bosses, la vitesse et la fermeté des greens, les obstacles à sauter, tous ces éléments nous poussent parfois à devoir modifier nos trajectoires de balles. Dans ces variations existe le lob shot ! Cette balle haute qui a pour objectif de survoler un obstacle et s'arrêter rapidement est souvent perçue comme un calvaire par les joueurs amateurs. Mais est ce si difficile ? Existe-t-il une manière de faire, « simple et répétitive », pour appréhender une première version de ce lob shot ? Je vais m'appuyer sur Jon Rahm, 7 ème cette année au Scrambling du PGA Tour* en dessous de 30m, pour vous apporter quelques explications pour améliorer ce domaine dans votre chipping.

Les premiers éléments à maitriser dans tous coups de golf sont les éléments de la posture ! Un stance (position des pieds) assez étroit. L'extérieur des pieds étant à l'intérieur de la largeur des épaules. Identique à la position classique de chipping. Le poids sur le pied avant = le droit pour les gauchers, le gauche pour les droitiers. Le club dans l'axe de l'aine et de l'avant bras comme indiqué par le trait vert.

On voit également que la face de club est ouverte. Elle est en direction du ciel. Cette ouverture est effectué par une rotation de la face et non par une orientation de la face en avançant les mains vers l'avant, ce qui dans ce cas serait contre productif.

L'armement... voici un vaste sujet ! Pour cette version Alpha du lob shot, je vais vous demander d'envisager les choses ainsi. Si la face de club à l'adresse est ouverte le club en devient moins puissant. Exemple un F9 est moins puissant qu'un F5 ceci étant du, entre autre à l'ouverture de la face. Si le club est moins puissant et donc ici peu puissant, c'est un sand-wedge dont J. Rahm a ouvert la face, il faut pas mal d'amplitude même pour faire peu de distance. Si il faut de l'amplitude il faut, comme dans tout swing, se mettre à armer le club. L'armement dans cette version Alpha du lob shot n'est donc pas volontaire ! [...]

Figure 5: Example of a French document.

8 languages per batch and upsample low-resource languages. We train multilingual OpenFlamingo on 43 languages covering all the languages of the benchmarks we evaluate the models on (see Section A.6.4).

We use Gemma-2B as the underlying language model behind multilingual OpenFlamingo and CLIP ViT-L-14 as the image encoder. We add a cross-attention layer after each decoder layer. Following OpenFlamingo, we add the two special tokens `<image>` and`<|endofchunk|>`, whose embeddings were trained. Only the Perceiver Resampler, cross-attention layers and these two embeddings were trained; everything else remained frozen. During training, we apply a factor of 0.2 for the captioning data loss function.

We train the model using the Adam optimizer and a maximum learning rate of 1e-4. We use a constant learning rate scheduler with 1875 warm-up steps. We use 4 accumulation gradient steps to have an effective batch of size 256 for mOSCAR and 512 for captioning data. We train the model on 50M documents and 100M image-text pairs on 8 Nvidia A100 for 170h.

A.6.4 Evaluation details

We evaluate on a set of eight benchmarks: xFlickr&CO, XM3600, xGQA, MaXM, MaRVL, XVNLI, Multi30k (Test2016 subset) and CoMMuTE; covering 5 different tasks and 43 languages. Details about the languages, the number of examples and the metric used can be found in Table 15. We used the *translate-test*[24] samples provided by the authors of the benchmarks if available. No translate test samples were provided for MaXM, so we translated the test set using the NLLB-600M distilled model. As no training set was available for MaXM, we use the few-shot examples from xGQA. Since we use Stanza tokenizers, we could not evaluate on all languages from XM3600 as 3 of them were not available. Filipino was also not into the list of mOSCAR languages, so we skip this language during evaluation. The CoMMuTE evaluation set involves choosing between two different translations of a same source text (one correct and one incorrect depending on an image provided to disambiguate the text). We use the lowest perplexity between the two translations as the model's prediction. We also use Multi30k training set as few-shot examples.

Prompting Following previous works, the zero-shot setting is composed of two few-shot examples without providing the images. The prompts we use for the different tasks are as follows:[25]

For captioning tasks, we use the prompt:
"`<image>Output:[Caption]<|endofchunk|><image>Output:`",
where `[Caption]` is replaced by the caption.

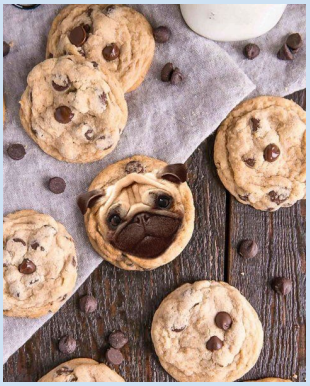
For visual question answering tasks, we use the

[24] Benchmark automatically translated into English.

[25] We show the prompts we used with one context example.

群馬県伊勢崎市でレジェンドたちと野球教室 ~! 本日、群馬県伊勢崎市にて野球教室でした～。プロ野球OBクラブ更に「大東建託」さん主催！中学校の野球部の選手達へ熱血指導～。

Figure 6: Example of a Japanese document.



Собаки в еде! Необычный профиль в Instagram взорвал весь интернет. Данный аккаунт приглянется всем тем, кто не мыслит своей жизни без вкуснейшей еды и просто обожает братьев меньших, в особенности милых пёсиков. Только представьте себе, что у вас на тарелке лежит еда, но только в ней вы видите ещё и мордочку мопсика. Странно звучит, правда? Но вот кому-то эта идея пришла в голову и этот «кто-то» даже решил реализовать её. В Instagram в январе 2018 года появился весьма необычный профиль — @dogs_infood. В нём публикуются очень оригинальные и забавные иллюстрации, где изображена еда в тандеме с фотографиями собак.

Так что же можно там увидеть? Например, печенье с мордочкой мопса, веточка винограда со смешным французским бульдожкой, кренделёк с доберманом или шпиц в форме тефтельки. Это не только звучит забавно, но ещё и выглядит очень смешно. Кстати, любой желающий может прислать фотографию своего любимца автору профиля, и кто знает, может, следующий пост будет посвящён именно ему. [...]

Figure 7: Example of a Russian document.

Nel mese di settembre c'è un altro evento sportivo che coinvolge soprattutto gli appassionati di corsa ed è il “Bibione is surprising run”. É una gara internazionale di 10 miglia con percorsi che si intrecciano lungo il litorale toccando i punti più belli di Bibione. Anche per i meno allenati, è una buona occasione per far conciliare benessere fisico e salute. Ci sono tante proposte di strutture ricettive a Bibione che offrono pacchetti famiglia economici con la possibilità non solo di partecipare alla gara ma anche di fare un bel tuffo in mare. Il periodo di settembre è adatto per le famiglie con bambini: il mare è calmo e le giornate sono calde. Ritagliati un week-end last minute prima di tornare al lavoro e iniziare con la routine quotidiana. Di seguito sono elencati appartamenti confortevoli ed hotel economici che garantiscono risparmio e qualità al tuo soggiorno.

Rimani aggiornato sulle migliori offerte per Bibione. Residence con piscina - appartamento con barbecue e posto auto.

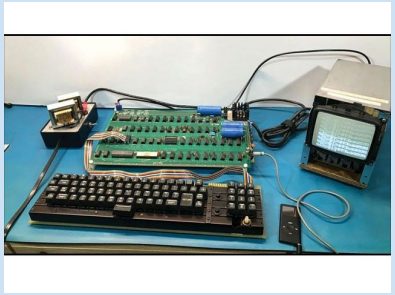
Figure 8: Example of an Italian document.

Nissan ចាប់ដៃគ្នាជាមួ New Balance បញ្ចេញគំរូរថយន្តដ៏ពិសេសដែលមិនធ្លាប់មានពីមុនមក បែកធ្លាយរូបរាងឡាន Tacoma ជំនាន់ថ្មី ចេញពីរូបប៉ាតង់ថ្មី មើលមកដូចកូន Tundra ៥ឆ្នាំទៀត Porsche នឹងឈប់ផលិត Macan ប្រើសាំង

សមាជិក Blackpink សហការជាមួយ Porsche ឌីស្សាញម៉ូឌែលរថយន្តដ៏ពិសេសសម្រាប់ខ្លួនឯង

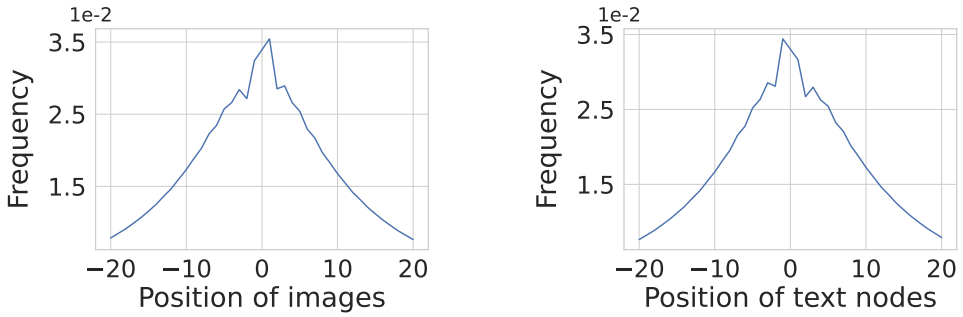
Figure 9: Example of a Khmer document.

ایپل کا سب سے پہلا کمپیوٹر نیلامی کے لیے پیش



بوسٹن: ایپل کا سب سے پہلا مکمل طور پر فعال ایپل 1 کمپیوٹر نیلامی کے لیے پیش کر دیا گیا۔ میڈیا رپورٹ کے مطابق اس مشین، جس پر ایپل کے بانی اسٹیو جابز نے اپنے ہاتھوں سے نمبر ڈالے تھے، کے ساتھ وہ تمام چیزیں آئیں گی جو اس مشین کو چلانے کے لیے ضروری ہیں۔ فی الحال اس کمپیوٹر کی نیلامی کی بولی 2 لاکھ 41 ہزار 557 ڈالرز پر ہے جو 15 دسمبر کو ختم ہوجائے گی لیکن ایک اندازے کے مطابق اس کی حتمی بولی 3 لاکھ 75 ہزار ڈالرز تک جائے گی۔ 1976 میں متعارف کروایا جانے والا ایپل 1 اس ٹیک کمپنی کی سب سے پہلی شے تھی جو ایک اسمبلڈ سرکٹ بورڈ کے طور پر بیچی گئی تھی اس میں بنیادی چیزیں جیسے کہ کی بورڈ یا مانیٹر نہیں تھا۔ لیکن دیگر ایپل 1 کمپیوٹرز کے برعکس اس یونٹ کے فزیکل بورڈ میں کسی قسم کی کوئی تبدیلی نہیں کی گئی ہے اور اس کا نمونہ صاف اور بغیر کسی استعمال شدہ ہے۔ بوسٹن کے آکشن ہاؤس کے مطابق ایک تفصیلی ٹیسٹ میں اس سسٹم کو تقریباً آٹھ گھنٹے تک چلایا گیا جس میں کوئی خرابی سامنے نہیں آئی۔ تازہ ترین سلائیڈ شوز

Figure 10: Example of an Urdu document.



(a) Relative position in the document of relevant text nodes with respect to images.

(b) Relative position in the document of relevant images with respect to text nodes.

Figure 11: Relative positions of most relevant images and text nodes with respect to the other modality.

	Metric	#examples	Languages
xFlickr&CO	CideR	2,000	Chinese, English, German, Indonesian, Japanese, Russian, Spanish, Turkish
XM3600	CideR	3,600	Arabic, Czech, Danish, German, Greek, English, Spanish, Farsi, Finnish, French, Hebrew, Hindi, Croatian, Hungarian, Indonesian, Italian, Japanese, Korean, Dutch, Norwegian, Poland, Portuguese, Romanian, Russian, Swedish, Telugu, Thai, Turkish, Ukrainian, Vietnamese, Chinese
xGQA	Accuracy	9,666	Bengali, German, English, Indonesian, Korean, Portuguese, Russian, Chinese
MaXM	Accuracy	∼ 170	English, French, Hindi, Hebrew, Romanian, Thai, Chinese
MaRVL	Accuracy	∼ 1,150	Indonesian, Swahili, Tamil, Turkish, Chinese
XVNLI	Accuracy	1,164	English, Arabic, Spanish, French, Russian
Multi30k	BLEU	1,000	French, German, Czech
CoMMuTE	Accuracy	310	Czech, French, German

Table 15: Overview of the benchmarks used to evaluate our multilingual OpenFlamingo.

prompt:
"<image>Question: [Question]
Short Answer: [Answer] <|endofchunk|>
<image>Question: [Question]
Short Answer:",
where [Question] and [Answer] are replaced by the questions and the answer respectively.

For multimodal machine translation tasks, we use the prompt:
"<image>Sentence:'[Caption]'.
Translation: [Translation] <|endofchunk|>
<image>Sentence:'[Caption]'
Translation:",
where [Caption] is replaced by the sentences to translate and [Translation] is replaced by its translation.

For MaRVL, we use the prompt:
"<image> '[Statement]'. True of False? [Answer]<|endofchunk|><image> '[Statement]'. True of False?",
where [Statement] is replaced by the statement and [Answer] by the answer. We also concatenate the left and right image into a single image.

For XVNLI, we use the prompt:
"<image> '[Statement1]' - '[Statement2]'.
entailment, neutral or contradiction?
Output: [Answer]<|endofchunk|>
<image> '[Statement1]' - '[Statement2]'.
entailment, neutral or contradiction?
Output:",
where [Statement1], [Statement2] and [Answer] are replaced by XVNLI test data.

A.7 Detailed results

Tables 16 to 23 show the detailed results for all languages in which it can be observed that the model trained on mOSCAR outperforms the model trained on captions only by a large margin.

	#shots	De	En	Es	Id	Ja	Ru	Tr	Zh
Multilingual OF *mOSCAR + caps.*	0	26.93	29.64	14.07	32.04	2.87	18.07	4.23	7.40
	4	54.38	51.47	37.32	47.22	11.06	32.23	13.03	31.71
	8	55.09	56.75	34.99	**51.60**	15.03	34.17	13.63	33.90
	16	**61.59**	**59.89**	**39.46**	51.50	**19.63**	**34.94**	**14.19**	**34.49**
Multilingual OF *captions only*	0	16.72	24.57	3.80	10.82	2.82	8.20	2.79	6.82
	4	21.10	31.05	7.52	9.63	3.84	13.21	7.01	12.20
	8	32.56	35.73	13.35	15.85	5.96	18.13	6.97	15.47
	16	29.86	40.57	13.75	23.83	6.92	20.40	7.90	15.73

Table 16: Captioning results (CideR scores) on xFlickr&CO. **Bold** is best result.

	#shots	Ar	Cs	Da	De	El	En	Es	Fa	Fi	Fr	He
Multi. OF *full*	0	4.83	2.50	8.52	8.16	0.76	42.57	16.79	12.49	1.26	14.76	3.76
	4	22.74	6.42	33.73	24.29	2.32	77.98	37.81	31.94	6.78	39.79	15.51
	8	22.91	7.41	35.23	**25.79**	2.95	77.64	38.41	**35.46**	7.92	42.81	15.85
	16	**23.47**	**8.14**	**35.96**	**25.47**	2.58	**78.18**	**39.18**	31.44	**8.42**	**43.77**	**16.08**
Multi. OF *Caps only*	0	2.24	0.97	6.42	6.46	3.68	10.02	9.32	4.95	1.14	16.15	0.78
	4	5.36	1.36	13.11	11.82	7.78	35.52	19.96	9.62	1.86	22.48	2.29
	8	6.76	1.40	15.29	14.39	7.21	37.28	21.90	12.19	2.08	23.27	1.71
	16	6.25	2.29	17.96	15.11	**7.64**	48.03	25.39	9.21	2.10	30.16	2.72

	#shots	Hi	Hr	Hu	Id	It	Ja	Ko	Nl	No	Pl	Pt
Multi. OF *full*	0	2.79	2.00	1.51	9.96	11.53	0.92	0.58	16.11	8.31	3.94	13.37
	4	11.03	10.87	5.87	25.88	29.53	17.45	10.85	46.22	25.18	15.36	31.32
	8	11.61	**12.00**	6.91	29.68	**29.34**	20.13	**12.01**	47.58	**27.08**	**17.80**	**33.29**
	16	**12.74**	11.40	7.03	26.73	**30.43**	**20.57**	11.07	**49.33**	27.07	17.15	32.79
Multi. OF *Caps only*	0	2.29	0.97	3.51	2.98	7.96	1.85	1.05	4.88	5.78	0.92	9.79
	4	4.57	1.72	7.57	6.39	16.23	3.47	4.33	11.26	11.99	1.16	15.93
	8	5.94	2.17	7.83	9.93	15.40	7.93	5.34	11.87	13.79	1.38	17.50
	16	6.36	2.42	**9.55**	11.77	17.43	10.44	6.03	12.98	14.65	1.28	20.32

	#shots	Ro	Ru	Sv	Te	Th	Tr	Uk	Vi	Zh
Multi. OF *full*	0	1.84	4.72	11.09	0.88	5.49	2.86	2.08	11.34	3.29
	4	6.08	21.46	30.24	3.46	23.14	10.75	11.35	32.70	19.57
	8	**7.10**	21.78	30.26	3.76	25.17	12.83	12.26	35.86	20.11
	16	6.95	**22.63**	**32.07**	**4.52**	**25.23**	**13.38**	**12.29**	**37.12**	**20.71**
Multi. OF *Caps only*	0	2.24	1.93	4.55	0.67	2.34	2.68	0.80	8.55	2.70
	4	5.35	6.29	15.66	0.77	7.21	5.94	1.76	20.69	7.80
	8	5.18	7.58	14.01	1.00	6.81	8.90	2.73	23.05	8.99
	16	5.06	9.06	20.60	1.18	8.35	10.25	3.47	25.16	11.05

Table 17: Captioning results (CideR scores) on XM3600. **Bold** is best result.

	#shots	Bn	De	En	Id	Ko	Pt	Ru	Zh
Multilingual OF *mOSCAR + caps.*	0	22.76	25.72	34.24	26.68	26.89	26.73	25.28	27.32
	4	26.72	32.57	37.91	32.54	31.88	32.35	31.28	33.4
	8	28.07	35.15	39.44	35.14	32.94	35.59	33.58	34.04
	16	**29.64**	**37.33**	**40.09**	**35.55**	**34.06**	**36.27**	**34.50**	**35.36**
Multilingual OF *captions only*	0	10.54	6.51	10.43	7.74	7.50	7.79	8.62	9.84
	4	12.54	11.90	15.78	13.95	13.70	12.01	12.73	15.03
	8	11.62	11.70	17.29	13.86	12.85	11.60	12.65	15.35
	16	9.77	11.86	18.37	13.24	12.48	11.25	11.24	14.33
Translate Test									
OF-3B MPT	0	18.64	18.67	-	18.36	17.54	19.21	18.88	17.11
	4	23.23	23.40	-	22.95	22.46	23.52	22.41	22.85
	8	28.22	29.44	-	28.21	27.67	29.58	28.21	28.63
	16	31.31	32.58	-	31.82	31.42	32.74	31.62	31.22
Multilingual OF *mOSCAR + caps.*	0	30.41	32.1	-	29.35	29.99	31.39	29.06	28.81
	4	34.89	36.32	-	35.50	35.64	36.84	35.05	34.60
	8	35.95	37.65	-	36.78	37.14	37.81	36.17	35.98
	16	**36.78**	**38.78**	-	**37.52**	**37.73**	**38.68**	**37.91**	**36.84**

Table 18: VQA results on xGQA. **Bold** is best result.

	#shots	En	Fr	Hi	He	Ro	Th	Zh
Multi. OF *mOSCAR + caps*	0	36.58	28.03	20.38	18.21	15.49	24.25	13.36
	4	38.13	30.03	23.08	21.43	17.61	31.72	22.02
	8	38.52	29.55	24.62	20.00	17.61	**25.27**	**23.83**
	16	35.80	**31.82**	**25.00**	**23.93**	19.01	**33.96**	22.74
Multi. OF *captions only*	0	9.73	0.38	7.69	1.43	0.00	5.22	3.61
	4	9.34	2.65	5.00	2.50	0.00	5.60	3.97
	8	9.34	1.89	8.08	5.00	1.06	3.36	5.42
	16	8.56	1.14	5.00	8.21	0.35	3.36	7.58
Translate test								
OF-3B MPT	0	-	12.50	22.31	0.36	10.92	0.00	0.00
	4	-	10.98	25.38	0.36	10.21	0.00	0.00
	8	-	10.98	27.31	0.36	11.27	0.00	0.00
	16	-	13.26	26.54	**1.07**	13.38	0.00	0.00
Multi. OF *mOSCAR + caps*	0	-	**18.18**	28.08	0.00	13.73	0.00	**0.36**
	4	-	15.91	30.38	0.36	12.68	0.00	0.00
	8	-	15.15	30.77	0.00	14.79	0.00	0.00
	16	-	15.91	**35.77**	0.36	**16.90**	0.00	0.00

Table 19: VQA results on MaXM. **Bold** is best result.

	#shots	Id	Sw	Ta	Tr	Zh
Random chance		50.00	50.00	50.00	50.00	50.00
Multilingual OF *mOSCAR + caps*	0	50.09	49.46	49.60	49.83	48.81
	4	49.91	48.19	49.68	50.42	50.00
	8	**53.55**	**50.72**	49.76	**51.78**	**51.58**
	16	48.94	49.82	49.20	50.25	50.99
Multilingual OF *captions only*	0	51.33	49.01	49.52	49.83	49.70
	4	49.73	49.64	49.19	49.41	49.70
	8	49.91	49.10	49.60	49.75	49.90
	16	50.09	49.73	49.60	49.75	49.80
Translate test						
OF-3B MPT	0	50.00	49.37	49.76	49.83	49.80
	4	50.00	49.64	49.52	49.75	49.60
	8	49.82	49.46	49.28	50.08	49.90
	16	50.00	49.37	49.44	50.00	49.80
Multilingual OF *mOSCAR + caps*	0	49.07	49.79	49.52	50.34	49.60
	4	49.99	49.79	48.23	49.75	49.76
	8	50.00	48.92	**50.64**	50.42	48.90
	16	49.84	50.00	50.24	48.90	49.75

Table 20: Classification results on MaRVL. **Bold** is best result.

	#shots	Ar	En	Es	Fr	Ru
Random chance		33.33	33.33	33.33	33.33	33.33
Multilingual OF. *mOSCAR + caps.*	0	33.51	34.62	33.08	34.02	34.19
	4	33.08	33.59	33.42	34.45	35.82
	8	**35.91**	**38.75**	**35.14**	**36.08**	**37.11**
	16	34.11	36.60	33.93	34.54	35.05
Multilingual OF. *captions only*	0	35.48	34.02	33.51	34.45	31.36
	4	32.04	31.79	32.73	32.22	31.44
	8	34.02	33.76	32.04	35.57	33.16
	16	32.04	32.99	33.76	33.17	31.53
Translate test						
OF-3B MPT	0	32.65	-	31.01	31.44	35.82
	4	36.25	-	35.82	35.57	35.65
	8	31.27	-	31.10	31.10	31.70
	16	33.68	-	33.25	32.99	33.25
Multilingual OF. *mOSCAR + caps.*	0	34.88	-	34.88	34.54	34.36
	4	36.25	-	36.17	35.91	36.08
	8	**39.60**	-	**39.52**	**40.29**	**39.35**
	16	37.54	-	37.89	37.46	39.00

Table 21: Classification results on XVNLI. **Bold** is best result.

	#shots	Cs	De	Fr
Multi. OF *full*	0	2.82	28.45	37.47
	4	3.12	29.20	37.49
	8	3.14	**29.62**	37.99
	16	**3.34**	29.41	**38.79**
Multi. OF *caps. only*	0	0.00	0.00	0.00
	4	0.00	0.00	0.00
	8	0.00	0.00	0.03
	16	0.00	0.40	1.82

Table 22: En→X translation results on Multi30k. **Bold** is best result.

	#shots	Cs	De	Fr
Multi. OF *full*	0	56.49	**65.67**	67.86
	4	57.47	64.00	**68.18**
	8	58.44	64.33	67.86
	16	58.11	62.67	66.23
Multi. OF *caps. only*	0	58.12	61.67	64.29
	4	**59.09**	61.00	63.31
	8	**59.09**	59.34	64.29
	16	58.12	58.67	63.96

Table 23: En→X CoMMuTE results. **Bold** is best result.

A.8 Comparison with state-of-the-art mLLMs

We computed the results for different state-of-the-art models of similar sizes as multilingual Open Flamingo namely: (1) InternVL2-4B[26] (2) PaliGemma[27] (3) Idefics2-8B[28] and (3) Llava-NeXT 8B[29]. InternVL2 and PaliGemma are trained on multilingual and multimodal data while Llava-NeXT and Idefics2 are trained on English multimodal datasets.

Table 24 shows results averaged across languages for different state-of-the-art mLLMs of sizes from 3b to 8B. These results highlights multiple things: (1) getting results significantly better than random (MaRVL and XVNLI) requires instruction-tuning data as Idefics2 and Llava-NeXT were both trained on instruction-tuning multimodal datasets (2) English-only still gets decent results on multilingual benchmarks despite not having been trained on multilingual and multimodal data, probably due to their underlying LLM being multilingual (3) multilingual Open Flamingo (trained on mOSCAR and captions) gets superior results to InternVL2-4B on VQA benchmarks and captioning benchmarks but inferior to PaliGemma-3B mainly due to the fact that it was trained on much less data and the quality of the captions used to train multilingual Open Flamingo may not be as good as the WebLI dataset used to train PaliGemma.

[26] OpenGVLab/InternVL2-4B
[27] google/paligemma-3b-pt-224
[28] HuggingFaceM4/idefics2-8b
[29] llava-hf/llama3-llava-next-8b-hf

	# shots	xFlickR&CO	XM3600	xGQA	MaXM	MaRVL	XVNLI	Multi30k	CoMMuTE
InternVL2 4B	0	16.21	7.02	12.38	6.35	53.14	33.85	26.99	66.93
	4	24.89	9.53	26.05	14.72	54.22	35.72	26.68	64.22
PaliGemma 3B	0	28.28	**24.49**	**42.68**	**33.42**	51.48	39.36	17.98	62.78
Idefics2 8B	0	27.11	15.94	22.53	28.99	**63.18**	**50.33**	**30.19**	**67.13**
Llava-NeXT 8B	0	23.67	14.70	25.48	15.17	60.50	45.40	29.40	66.37
Multi. OF 3B (*ours*)	0	16.91	7.45	26.95	22.23	49.56	33.88	22.91	63.34
	4	**34.80**	22.18	32.23	26.33	49.64	34.07	23.27	63.22

Table 24: Results averaged across languages. **Bold** is best result.