OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents

Hugo Laurençon[*,1,2] **Lucile Saulnier**[*,1] **Léo Tronchon**[*,1]
Stas Bekman[*,1] **Amanpreet Singh**[*,1] **Anton Lozhkov**[1]
Thomas Wang[1] **Siddharth Karamcheti**[1,3] **Alexander M. Rush**[†,1]
Douwe Kiela[†,1,3] **Matthieu Cord**[†,2] **Victor Sanh**[*,†,1]
[*]Equal contributions, [†]Senior contributions
hugo@huggingface.co
[1]Hugging Face [2]Sorbonne Université [3]Stanford University

Abstract

Large multimodal models trained on natural documents, which interleave images and text, outperform models trained on image-text pairs on various multimodal benchmarks. However, the datasets used to train these models have not been released, and the collection process has not been fully specified. We introduce the OBELICS dataset, an open web-scale filtered dataset of interleaved image-text documents comprising 141 million web pages extracted from Common Crawl, 353 million associated images, and 115 billion text tokens. We describe the dataset creation process, present comprehensive filtering rules, and provide an analysis of the dataset's content. To show the viability of OBELICS, we train vision and language models of 9 and 80 billion parameters named IDEFICS, and obtain competitive performance on different multimodal benchmarks. We release our dataset, models and code.[1]

1 Introduction

Recent systems demonstrate the effectiveness of training large multimodal models such as Flamingo on naturally occurring multimodal documents (Alayrac et al., 2022; Aghajanyan et al., 2022; Huang et al., 2023). A multimodal document is a succession of text paragraphs interleaved by images, such as web pages that contain images. Models trained on these web documents outperform vision and language models trained solely on image-text pairs on various benchmarks (Alayrac et al., 2022). They can also generate long and coherent text about a set of multiple images.

While these results are compelling, they have not been replicable. The datasets used in these works are not publicly available, and relatively little information is known about their creation process and composition. This state motivates the creation of large-scale collections of high-quality multimodal web documents to support the creation of the next generation of models.

We take inspiration from existing large open image-text datasets such as LAION (Schuhmann et al., 2022) and COYO (Byeon et al., 2022), comprised of hundreds of millions of image-text

OBELICS: https://huggingface.co/datasets/HuggingFaceM4/OBELICS
[1]OBELICS reproduction code: https://github.com/huggingface/OBELICS
IDEFICS models: https://huggingface.co/HuggingFaceM4/idefics-80b

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

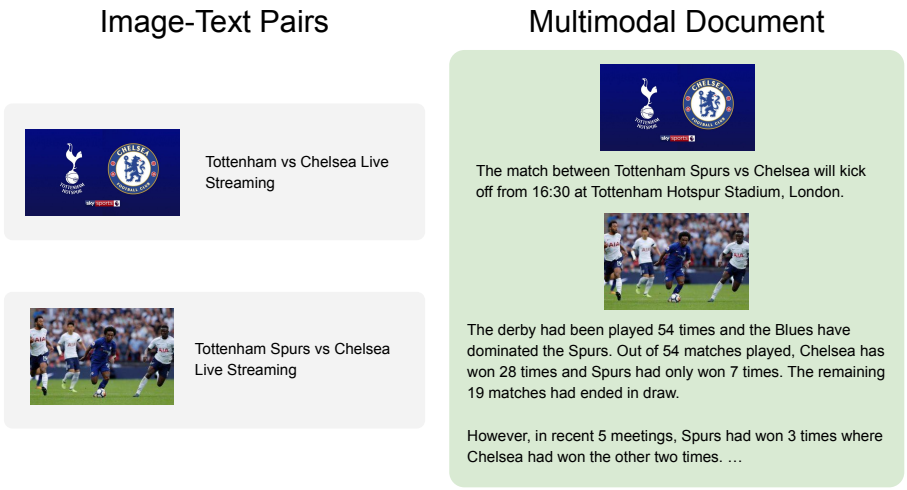

Figure 1: A comparison of extraction from the same web document. For image-text pairs, the alt-text of images is often short or non-grammatical. For OBELICS, the extracted multimodal web document interleaves long-form text with the images on the page.

pairs obtained through web crawling. These datasets have been critical to developing and replicating numerous recent multimodal models (Radford et al., 2021; Wang et al., 2022; Yu et al., 2022; Wang et al., 2022; Liu et al., 2023). While this approach allows for building extremely large and diverse training datasets, we note several limitations to using only image-text pairs. From a language perspective, these datasets rely primarily on alt-text, meaning the text given is brief, captures an approximate snapshot of the image's content, and often lacks grammatical correctness. From a document perspective, image-text pairs remove an image from its natural context on a page and its relationship with other documents.

In this work, we introduce OBELICS[2], an openly-accessible curated web-scale dataset consisting of 141 million multimodal English web documents which contain 353 million associated images and 115 billion tokens. OBELICS collects full multimodal documents interleaving text and images as shown in Figure 1. We describe the dataset creation process, outline the filtering and curation steps and shed light on the dataset's content and limitations. To demonstrate the viability of OBELICS, we train IDEFICS, an 80 billion parameter multimodal model and show competitive performance against large-scale multimodal models such as Flamingo (Alayrac et al., 2022).

2 Related Works

Image-text pairs datasets The largest multimodal datasets, such as LAION (Schuhmann et al., 2021, 2022), Conceptual Captions (Sharma et al., 2018; Changpinyo et al., 2021), ALIGN (Jia et al., 2021), COYO (Byeon et al., 2022), and DataComp (Gadre et al., 2023), contain billions of image-text pairs and are usually obtained through web-crawling and alt-text extraction. A variety of multimodal models have been trained on this type of dataset: multimodal encoder models which use a contrastive objective (Radford et al., 2021; Wang et al., 2022), image generation based on Transformers or diffusion processes (Nichol et al., 2022; Ramesh et al., 2022; Rombach et al., 2021; Saharia et al., 2022). While the scale of these datasets makes them attractive candidates for training, our work focuses on extracting images and the textual context in which they appear instead of extracting the associated alternative text.

Web document datasets Insights from scaling language models (Kaplan et al., 2020; Hoffmann et al., 2022) emphasize the need for increasingly bigger datasets. For instance,

[2]Open Bimodal Examples from Large fIltered Commoncrawl Snapshots

LLaMA (Touvron et al., 2023) was trained on a dataset of 1.4T tokens created exclusively from openly accessible English web content. The authors noticed that an even bigger dataset would have benefited the model. To address that need, multiple web-scale datasets have been introduced and made available: c4 (Raffel et al., 2019), ROOTS (Laurençon et al., 2022), Pile (Gao et al., 2020), OSCAR (Ortiz Suárez et al., 2020). Although `OBELICS` falls in the same category of making accessible large collections of curated web documents, the additional extraction of images changes the nature of the resulting dataset. It allows training models with additional vision capabilities.

Multimodal web document datasets The recent most performant vision and language models are trained on large sets of multimodal web documents. For instance, Flamingo (Alayrac et al., 2022), an 80 billion multimodal model, was trained on a mix of 2.1 billion image-text pairs, 27 million video-text pairs, and 43 million multimodal web documents. The latter, called `M3W`, includes 185 million images. Similarly, KOSMOS-1 (Huang et al., 2023) was trained on a mixture containing 71 million multimodal web documents. However, in both cases, the dataset is not publicly available, and little information is accessible as to the dataset's content, the strategies employed to create that dataset (including filtering strategies), and the quality of the resulting web documents, which ultimately hinders further research.

Concurrently to our work, the Multimodal C4 (`mmc4`) dataset (Zhu et al., 2023) was recently made accessible. It consists of 103 million multimodal web documents that include 585 million images. Although there are similarities between our datasets, it is important to highlight particular distinctions. First, our dataset is based on more recent documents from February 2020 to February 2023, whereas `mmc4` uses documents from April 2019. Additionally, our filtering heuristics appear to be more comprehensive: we leverage the HTML DOM trees to filter out undesirable texts and images, whereas `mmc4` uses the HTML to find images in order to merge them with the original C4 dataset by solving a bipartite assignment problem based on CLIP model similarities. Last, we implement additional deduplication steps at the image, document, and paragraph levels.

3 Creation of the Multimodal Web Document Dataset

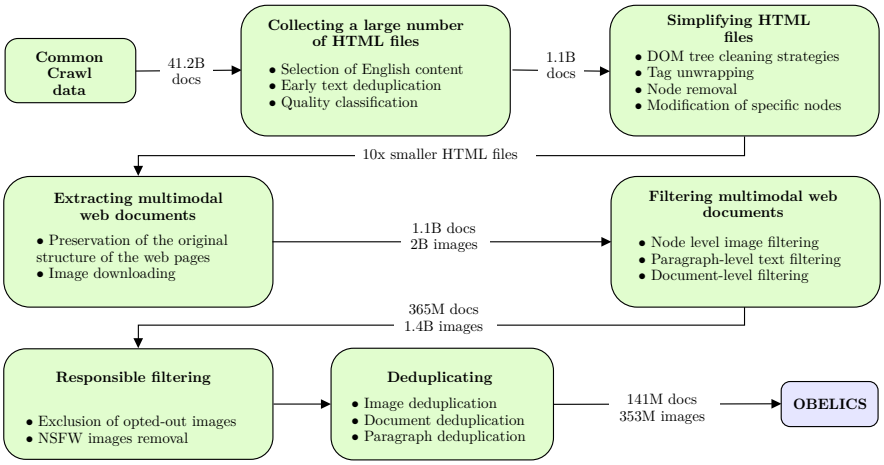

Figure 2: Overview of the steps involved in creating `OBELICS`.

This section provides an overview of the critical choices of the creation and filtering process. Figure 2 gives a high-level summary of the main steps involved. Many details are omitted from this section, and we invite the reader to refer to the appendix A.1 for completeness.

3.1 Collecting a Large Number of HTML Files

First, we collect a vast amount of raw web documents by considering the 25 most recent Common Crawl dumps at the time of the creation, spanning from February 2020 to January/February 2023[3]. We extract the main text from the documents while discarding documents with text of insufficient quality. This process results in 41.2 billion documents.

To filter out non-English content, we apply the FastText classifier (Joulin et al., 2017) to the extracted text, which removes 63.6% of the documents. We perform a MinHash (Broder, 1997) deduplication to remove duplicate content. Additionally, we filter out documents with significant proportions of repeated paragraphs and n-grams, following the methodology used in MassiveText (Rae et al., 2022). Previous studies (Lee et al., 2022; Abbas et al., 2023) have demonstrated the prevalence of duplication in crawled data and the benefits of training on deduplicated data.

Similar to Brown et al. (2020), we employ a logistic regression classifier with hashed token frequencies to ensure high-quality text. This classifier, trained using curated datasets like Wikipedia or OpenWebText (Gokaslan and Cohen, 2019) as positive examples and documents sampled from Common Crawl as negative ones, is fast and effective at detecting human-written text. After these steps, we are left with 1.1 billion documents and their HTML sources from the associated Common Crawl WARC files.

3.2 Simplifying HTML Files

The original HTML content of a document contains a wealth of valuable information that proves highly beneficial in the process of filtering out undesirable text and images. Therefore, we prioritize pre-processing the raw HTML into simplified HTML, making the subsequent extraction of textual and visual elements more efficient.

To this aim, we devise multiple pre-processing strategies for an HTML DOM tree. By manually inspecting instances of all HTML nodes, we differentiate nodes likely to contain relevant texts or images from those that should be discarded, and we formulate specific rules for each type of node. After these pre-processing steps, the resulting simplified HTML files are more than ten times smaller and have been stripped of a large proportion of generic text (spam, ads, boilerplate template, etc.) and generic images, such as logos, while retaining the relevant content.

3.3 Extracting Multimodal Web Documents

In this step, we transform the simplified HTML files previously obtained into a structured web multimodal web document format. This format consists of interleaved texts and images.

We meticulously preserve the original structure of the web pages from the simplified HTML files by extracting the texts and image links while maintaining their rendering defined by the DOM tree. Given that each HTML tag denotes a distinct separation between the preceding and subsequent nodes, we leverage that information to retain line breaks and line feeds on the original page, preserving the formatting and visual rendering of the content.

We obtain 3.6 billion image links and successfully download 55% of them (approximately 2 billion images).

3.4 Filtering Multimodal Web Documents

The filtering process comprises two distinct steps operating at different granularity levels. In the first step, filtering occurs at the node level for images and the paragraph level for text. This step guarantees that only high-quality and relevant images and paragraphs are retained. Each paragraph or image is evaluated based on specific criteria and may undergo modifications or be eliminated if necessary. The second step, conducted at the document level, involves deciding whether to retain or discard the output documents obtained from the

[3] https://commoncrawl.org/

first step. Most text filters used in both steps are primarily derived from Laurençon et al. (2022).

Node-level image filtering We discard images that are too small, excessively large or have disproportionate dimensions. We observe that these images are often indicative of low-quality or irrelevant content. To eliminate some logos and generic images, we remove images whose URLs contain one of the banned sub-strings, like *logo*.

Paragraph-level text filtering We apply multiple filters to text paragraphs to remove undesirable content. Specifically, paragraphs that contain an insufficient number of words are discarded. Additionally, we filter out paragraphs with high repetition ratios, excessive ratios of special characters, low ratios of stop words, low punctuation ratios, high proportions of flagged words associated with adult or inappropriate content, or excessively high perplexity scores (as measured by an n-gram language model trained on Wikipedia (Heafield, 2011)). To identify boilerplate sentences or invitations to share articles on social networks, we create a list of frequently used words associated with these paragraphs and remove paragraphs containing an excessive proportion of words from this list. To further identify machine-generated content, we extract words from web-crawled documents to form a list of common words and discard documents with a low ratio of common words.

Document-level filtering At the document level, we remove all documents with no or excessively high number of images. For text filters, the same filters used at the paragraph level are applied, with sometimes stricter cutoff values.

After these filtering steps, we are left with 365 million web documents and 1.4 billion images. At this step, images can be duplicated across documents.

3.5 Responsible Filtering and Deduplication

We take measures to minimize the amount of inappropriate content in the dataset. In particular, based on manual inspections and tool availability, we implement filters to respect data consent and remove images with pornographic content. Additionally, we also heavily deduplicate content.

Exclusion of opted-out images To respect the preferences of content creators, we remove all images for which creators explicitly opted out of AI model training. We used the Spawning API[4] to verify that the images in the dataset respect the original copyright owners' choices.

Image deduplication based on URL Some images could be present across different documents. We observe that it is particularly true for browser-specific icons or common advertisements encountered during the crawling process. To address this issue, we remove all images that appear more than ten times across the entire dataset. We intentionally do not perform strict deduplication, as we notice that when an image is duplicated only a few times across different documents, the surrounding text and contextual information tend to be different. We also deduplicate images within the same document.

NSFW image filtering To reduce explicit adult content, we use an open-source NSFW classifier to remove entire documents containing pornographically classified images. We also filter out images with URLs containing banned sub-strings.

Document deduplication based on URL and set of images We complete the initial deduplication step by forming clusters of documents with the same URLs and retaining the most recent document within each cluster. We repeat this operation by forming clusters of documents containing identical sets of images.

Paragraph deduplication across documents of the same domain names To remove generic spam phrases commonly found at the end of documents, we perform paragraph-level

[4] `https://api.spawning.ai/spawning-api`

exact deduplication within documents sharing the same domain name, resulting in the elimination of approximately 15% of the text.

Following these filtering and deduplication steps, the final dataset contains 141 million documents and 353 million images, of which 298 million are unique. We observe that using stricter values for the filtering steps yields fewer multimodal documents, although not of higher quality. As such, we invite users who are interested in manipulating a smaller subset of OBELICS to start with a random subset.

4 Analysis of OBELICS

Figure 1 provides an example showcasing an original webpage alongside the resulting multimodal web document. Extracting and filtering the multimodal document is non-trivial as it requires carefully removing undesirable information on the left, top, and bottom of the page, such as menus and navigation bars. We provide other examples at https://huggingface.co/spaces/HuggingFaceM4/obelics_visualization and in Figures 7, 8 and 9.

Given the scale of OBELICS, it would be prohibitive to describe its content exhaustively. Instead, we provide high-level statistics and analyses that shed light on the dataset's properties.

4.1 General Statistics

Dataset	Images	% unique images	Docs	Tokens	Open
KOSMOS-1	-	-	71M	-	✗
M3W	185M	-	43M	-	✗
mmc4-ff	385M	60.6%	79M	34B	✓
mmc4	**585M**	-	103M	43B	✓
OBELICS	353M	**84.3%**	**141M**	**115B**	✓

Table 1: General statistics of OBELICS and the current largest alternatives.

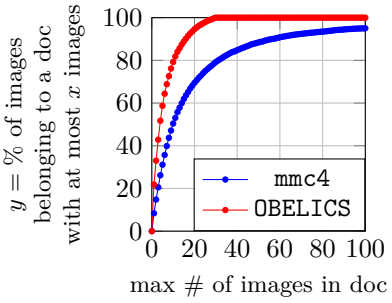

Figure 3: Distribution of images.

Table 1 compares OBELICS against the largest existing alternatives. mmc4-ff is the mmc4 dataset with fewer faces. Our dataset has the highest number of unique documents and total tokens while containing a huge number of images.

It is worth mentioning that we have fewer images than mmc4 (Zhu et al., 2023). This discrepancy can be attributed to two reasons. First, our analysis reveals that mmc4 contains many duplicated images, with only 60.6% being unique compared to 84.3% for OBELICS. We found that images duplicated multiple times often indicate spam or unrelated generic content. Second, mmc4 does not limit the number of images within a document. As a result, the distribution of images across documents is highly uneven, with a substantial portion of them concentrated in documents with excessive image counts (see Figure 3). The images in these documents are often unrelated to each other and exhibit spam or advertisement content. Moreover, these documents often have little text, making them unsuitable for learning the alignment between text and images (see an example in Figure 10).

Figure 4 shows the joint distribution of a number of tokens and a number of images in OBELICS. Although we limit the number of images in a document to 30, we cut the plot at 6 images for clarity. The documents of OBELICS contain a median number of images of 1 and a median number of tokens of 677.

Perplexity analysis To assess the quality of our text in comparison to reference datasets used for training large language models, we leverage an n-gram language model trained on Wikipedia (Heafield, 2011; Laurençon et al., 2022). This allows us to compute perplexity

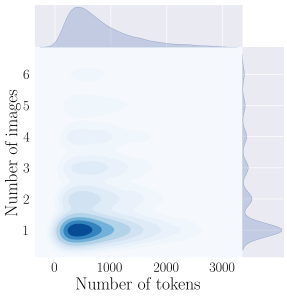

Figure 4: Heatmap displaying the joint distribution of the number of tokens and the number of images in OBELICS documents, accompanied by their respective marginal distributions.

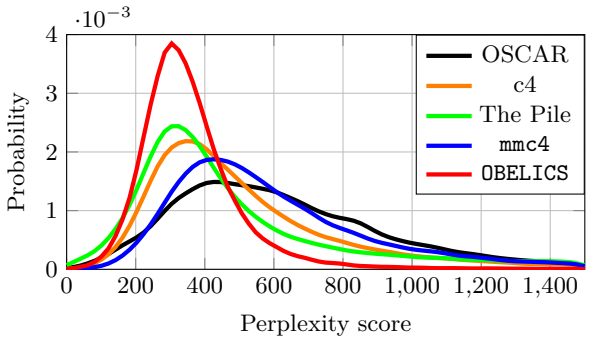

Figure 5: Kernel density estimations representing the distribution of perplexity scores for OBELICS compared to reference datasets. The lower the perplexity for a document, the more it resembles a Wikipedia article.

scores for 100,000 documents from each dataset. Lower perplexity scores indicate a higher resemblance to Wikipedia documents. Figure 5 displays the distributions of these scores. Our results demonstrate that the texts in OBELICS have a significantly lower average perplexity compared to the texts in c4 (Raffel et al., 2019), mmc4 (Zhu et al., 2023), and OSCAR (Ortiz Suárez et al., 2020). Furthermore, our distribution aligns closely with the one from The Pile (Gao et al., 2020), which was thoughtfully curated from diverse, high-quality sources.

4.2 Topic Modeling

Similar to Zhu et al. (2023), we employ a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to understand the diversity of the dataset. The LDA gives us insights into the distribution of topics in the dataset, along with estimated proportions and frequently associated words. Table 5 and 6 present the results of the LDA with respectively 20 and 200 topics, offering both a high-level and a more granular analysis of the dataset's content. We observe that the dataset covers topics ranging from Politics to Health by way of Music. Additionally, we compute the most frequent domains and show that news sites are systematically the most represented (Table 4).

4.3 Qualitative Assessment of Dataset Samples

We manually inspect 250 documents from OBELICS to verify the dataset's quality and asses the risks contained in the dataset. We focus on the images' content in relation to the text since it's the core addition compared to a language modeling dataset.

80% of documents have photo images, while 29% have graphic images (drawings, cartoons, etc.). 90% of the documents have all images clearly related to the text content. 30% of documents have images containing at least one written word, and 5% of documents have images that are structured text (slides, tables, scanned documents, etc.), which can help models learn OCR capabilities. 7% of documents have content (images or text) that hasn't been captured by cleaning filters (non-English text, spam or advertisement, etc.). 46% of documents contain images with faces (portraits or group photos). No obvious Personally Identifiable Information (PII) texts were found, except for public personalities and people mentioned in news articles. No NSFW images were found. Only 3% of documents contain images with watermarks, and 2% have images with logos.

5 Validating the Viability of OBELICS

To confirm the viability of our dataset, we first show that vision and language models trained on our multimodal web documents outperform the same models trained on image-text pairs on various multimodal benchmarks. Following that, we demonstrate the effectiveness of

OBELICS as an alternative to closed datasets by training models of different sizes on par with closed-source models.

Model details We follow the Flamingo (Alayrac et al., 2022) architecture closely: we combine two frozen unimodal backbones - LLaMA (Touvron et al., 2023) for the language model, and OpenClip [5] for the vision encoder - add learnable cross-attention Transformer blocks to connect the language and vision blocks. For multimodal web documents, we feed the model sequences corresponding to the succession of text paragraphs and images. For image-text pairs, we form the training sequences by packing images with their captions. The images are encoded with the vision encoder and vision hidden states are pooled with Transformer Perceiver blocks and then fused into the text sequence through the cross-attention blocks. The training objective is the standard next token prediction. For more details, we refer to the original paper.
Following Alayrac et al. (2022), we evaluate our models on a series of multimodal benchmarks spanning visual question answering (VQAv2 (Antol et al., 2015), OKVQA (Marino et al., 2019), TextVQA (Singh et al., 2019), VizWiz (Gurari et al., 2018)), visual dialogs (VisDial (Das et al., 2017)), hateful speech detection (HatefulMeme (Kiela et al., 2020)), image captioning (COCO (Lin et al., 2014), Flickr30k (Young et al., 2014)), and OCR (IIIT5k (Mishra et al., 2012)).
Additional details about the architecture, the training, the compute and the evaluation are present in Appendix A.4.

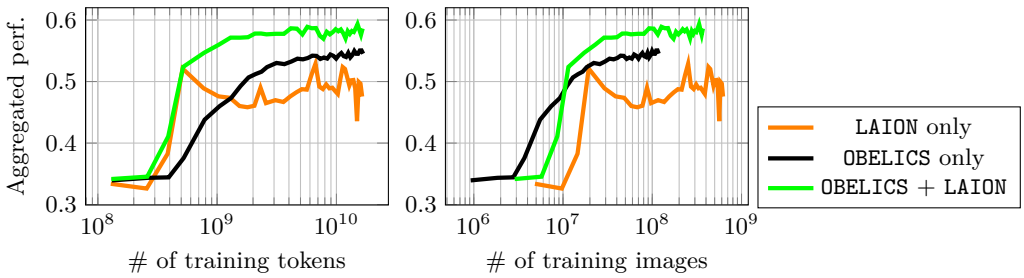

Figure 6: Aggregated 4-shot performance through the training using LAION only, OBELICS only and a mixture of both. The training sequences from multimodal documents and the packed sequences obtained from image-text pairs have different numbers of images but the same number of tokens. Thus, we plot the performance over two log x-axes. The initial uptick of the model trained on image-text pairs is attributed to the fact the performance on VQA tasks starts by increasing and then slowly degrades.

Training on different mixture of data Figure 6 shows the result of the first experiment, which consists in training 9B-parameter models on different mixture of data. Training on multimodal web documents allows reaching the same performance using an order of magnitude fewer images than training on image-text pairs, even though the images from the two datasets come from Common Crawl. This underlines the benefit of having longer text contexts for training multimodal models. Moreover, the model trained on multimodal web documents performs better on average. This is particularly striking on visual question-answering benchmarks on which the model trained on image-text pairs slowly degrades through the training. We note, however, that the model trained on image-text pairs has a slight advantage performance-wise in captioning, classification, and OCR tasks (see more details in Appendix A.4.5). We hypothesize that this is due to the nature of image-text pairs: captions can be seen as fuzzy class labels. Last, similarly to Alayrac et al. (2022), we observe that combining the two types of datasets leads to increased performance for a given number of images, tokens, or training compute.

Models trained on OBELICS achieve competitive performance at different scales Following these insights, we show that OBELICS is a viable open alternative to other datasets.

[5] https://laion.ai/blog/large-openclip/

	Shot	COCO	Flickr30k	VQAv2	OKVQA	TextVQA	VizWiz	VisDial	HatefulMemes
Flamingo-9B	0	79.4	61.5	51.8	44.7	31.8	22.8	48.0	57.0
OpenFlamingo-9B		79.5	59.5	52.7	37.8	24.2	27.5	-	51.6
IDEFICS-9B		46.0	27.3	50.9	38.4	25.9	35.5	48.7	51.8
Flamingo-9B	4	93.1	72.6	56.3	49.3	**33.6**	34.9	50.4	62.7
OpenFlamingo-9B		89.0	65.8	54.8	40.1	28.2	34.1	-	54.0
IDEFICS-9B		93.0	59.7	55.4	45.4	27.6	36.9	47.9	50.7
Flamingo-9B	8	99.0	**73.4**	58.0	50.0	**33.6**	39.4	51.2	63.9
OpenFlamingo-9B		96.3	62.9	54.8	41.1	29.1	38.5	-	54.7
IDEFICS-9B		97.0	61.9	56.4	47.7	27.5	40.4	47.6	51.1
Flamingo-9B	16	102.2	72.7	59.4	50.8	33.5	43.0	**51.3**	**64.5**
OpenFlamingo-9B		98.8	62.8	54.3	42.7	27.3	42.5	-	53.9
IDEFICS-9B		99.7	64.5	57.0	48.4	27.9	42.6	-	50.1
Flamingo-9B	32	**106.3**	72.8	**60.4**	**51.0**	32.6	**44.0**	50.4	63.5
OpenFlamingo-9B		99.5	61.3	53.3	42.4	23.8	**44.0**	-	53.8
IDEFICS-9B		98.0	64.3	57.9	49.6	28.3	43.7	-	49.8
Flamingo	0	84.3	67.2	56.3	50.6	35.0	31.6	52.0	46.4
IDEFICS		91.8	53.7	60.0	45.2	30.9	36.0	48.9	60.6
Flamingo	4	103.2	75.1	63.1	57.4	36.5	39.6	55.6	68.6
IDEFICS		110.3	73.7	63.6	52.4	34.4	40.4	48.4	57.8
Flamingo	8	108.8	78.2	65.6	57.5	37.3	44.8	56.4	**70.0**
IDEFICS		114.3	76.6	64.8	55.1	35.7	46.1	47.9	58.2
Flamingo	16	110.5	78.9	66.8	**57.8**	37.6	48.4	**56.8**	**70.0**
IDEFICS		**116.6**	80.1	65.4	56.8	36.3	48.3	-	57.8
Flamingo	32	113.8	75.4	**67.6**	**57.8**	**37.9**	49.8	55.6	**70.0**
IDEFICS		**116.6**	**81.1**	65.9	**57.8**	36.7	**50.0**	-	52.5

Table 2: Performance of `IDEFICS` against OpenFlamingo and Flamingo. The evaluations were done with random in-context examples, and in an open-ended setting for VQA tasks. (Task, Metric, Query split): (COCO, CIDEr, test), (Flickr30k, CIDEr, test (Karpathy)), (VQAv2, VQA acc., testdev), (OKVQA, VQA acc., val), (TextVQA, VQA acc., val), (VizWiz, VQA acc., testdev), (VisDial, NDCG, val), (HatefulMemes, ROC-AUC, test seen).

We train IDEFICS, an 80 billion parameters Flamingo-like model on a mixture of image-text pairs from LAION (Schuhmann et al., 2022), openly accessible captioning datasets (Singh et al., 2022), OBELICS and multimodal web documents obtained from Wikipedia using a similar extraction strategy. We also train a smaller version of 9 billion parameters, IDEFICS-9B. We compare these models against OpenFlamingo v2 (Awadalla et al., 2023) and Flamingo of the same sizes and trained on a similar mixture of multimodal web documents and image-text pairs. We report the results in Table 2.

IDEFICS is often on par with Flamingo on various multimodal benchmarks. Out of the 8 evaluation tasks, with 32 in-context examples, it either performs better or obtain the same result as Flamingo on 4 of them. At the 9 billion parameter scale, we are still behind Flamingo-9B. However, it is important to highlight that we outperform OpenFlamingo-9B, which was trained on mmc4, in terms of aggregated performance. We achieved a score of 56.5, compared to their score of 55.8, by selecting the best performance across all numbers of in-context examples for each task. This highlights the advantages of OBELICS as an open alternative to a multimodal web document dataset.

6 Conclusion

With the goal of supporting open-source large multimodal models, we introduce OBELICS, an open web-scale collection of filtered interleaved multimodal web documents based on Common Crawl snapshots. We document a collection and filtering process that balances the scale and removal of undesirable texts and images while addressing some of the well-documented ethical concerns of large-scale multimodal datasets, notably data consent and pornographic content. To demonstrate the usefulness of models trained on multimodal documents, we train IDEFICS on OBELICS and show that it is a viable alternative to closed datasets. Open datasets of multimodal documents with scale, quality, and diversity of sources can help support the ability to train competitive open models.

Acknowledgments and Disclosure of Funding

The authors were granted access to the HPC resources of the Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS) under the allocation 2022-A0121013450 made by Grand équipement national de calcul intensif (GENCI). The initial development of the dataset was done on Jean-Zay cluster of IDRIS, and we thank the IDRIS team for their responsive support throughout the project, in particular Rémi Lacroix. We thank Guillaume Salou for setting up the virtual machines used to download the images of our dataset, and Sebastian Nagel for his valuable assistance in providing insights on Common Crawl. We thank Yacine Jernite and Daniel van Strien for conducting a bias analysis of the models trained on `OBELICS`.

References

Abbas, A., K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos (2023). Semdedup: Data-efficient learning at web-scale through semantic deduplication.

Aghajanyan, A., B. Huang, C. Ross, V. Karpukhin, H. Xu, N. Goyal, D. Okhonko, M. Joshi, G. Ghosh, M. Lewis, and L. Zettlemoyer (2022). Cm3: A causal masked multimodal model of the internet. *ArXiv abs/2201.07520*.

Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan (2022). Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 23716–23736. Curran Associates, Inc.

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Awadalla, A., I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models.

Bai, Y., A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback.

Beaumont, R. (2021). img2dataset: Easily turn large sets of image urls to an image dataset. `https://github.com/rom1504/img2dataset`.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Biderman, S. and W. J. Scheirer (2020, 12 Dec). Pitfalls in machine learning research: Reexamining the development cycle. In J. Zosa Forde, F. Ruiz, M. F. Pradier, and A. Schein (Eds.), *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, Volume 137 of *Proceedings of Machine Learning Research*, pp. 106–117. PMLR.

Biderman, S., H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal (2023). Pythia: A suite for analyzing large language models across training and scaling.

Birhane, A., V. U. Prabhu, and E. Kahembwe (2021). Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv abs/2110.01963*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, mar). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*(null), 993–1022.

Broder, A. (1997). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Byeon, M., B. Park, H. Kim, S. Lee, W. Baek, and S. Kim (2022). Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Caswell, I., T. Breiner, D. van Esch, and A. Bapna (2020). Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. *ArXiv abs/2010.14571*.

Changpinyo, S., P. Sharma, N. Ding, and R. Soricut (2021). Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel (2022). Palm: Scaling language modeling with pathways.

Das, A., S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra (2017, July). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dehghani, M., J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X. Wang, C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. van Steenkiste, G. F. Elsayed, A. Mahendran, F. Yu, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Mensink, A. Kolesnikov, F. Pavetić, D. Tran, T. Kipf, M. Lučić, X. Zhai, D. Keysers, J. Harmsen, and N. Houlsby (2023). Scaling vision transformers to 22 billion parameters.

Deng, X., P. Shiralkar, C. Lockard, B. Huang, and H. Sun (2022). Dom-lm: Learning generalizable representations for html documents. *ArXiv abs/2201.10608*.

Desai, K., G. Kaul, Z. Aysola, and J. Johnson (2021). Redcaps: Web-curated image-text data created by the people, for the people. In J. Vanschoren and S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Volume 1. Curran.

Dodge, J., A. Marasović, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Conference on Empirical Methods in Natural Language Processing*.

Gadre, S. Y., G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt (2023). Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Gao, L., S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy (2020). The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gokaslan, A. and V. Cohen (2019). Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Gu, J., X. Meng, G. Lu, L. Hou, N. Minzhe, X. Liang, L. Yao, R. Huang, W. Zhang, X. Jiang, C. XU, and H. Xu (2022). Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 26418–26431. Curran Associates, Inc.

Gurari, D., Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham (2018). Vizwiz grand challenge: Answering visual questions from blind people.

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, pp. 187–197. Association for Computational Linguistics.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre (2022). Training compute-optimal large language models.

Huang, S., L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei (2023). Language is not all you need: Aligning perception with language models.

Jaegle, A., F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira (2021). Perceiver: General perception with iterative attention.

Jia, C., Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Jiang, A. Q., S. Welleck, J. P. Zhou, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, and Y. Wu (2023). Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *The Eleventh International Conference on Learning Representations*.

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov (2017, April). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain, pp. 427–431. Association for Computational Linguistics.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models.

Kärkkäinen, K. and J. Joo (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1547–1557.

Kiela, D., H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 2611–2624. Curran Associates, Inc.

Koh, J. Y., R. Salakhutdinov, and D. Fried (2023). Grounding language models to images for multimodal generation.

Laborde, G. Deep nn for nsfw detection.

Laurençon, H., L. Saulnier, T. Wang, C. Akiki, A. Villanova del Moral, T. Le Scao, L. Von Werra, C. Mou, E. González Ponferrada, H. Nguyen, J. Frohberg, M. Šaško, Q. Lhoest, A. McMillan-Major, G. Dupont, S. Biderman, A. Rogers, L. Ben allal, F. De Toni, G. Pistilli, O. Nguyen, S. Nikpoor, M. Masoud, P. Colombo, J. de la Rosa, P. Villegas, T. Thrush, S. Longpre, S. Nagel, L. Weber, M. Muñoz, J. Zhu, D. Van Strien, Z. Alyafeai, K. Almubarak, M. C. Vu, I. Gonzalez-Dios, A. Soroa, K. Lo, M. Dey, P. Ortiz Suarez, A. Gokaslan, S. Bose, D. Adelani, L. Phan, H. Tran, I. Yu, S. Pai, J. Chim, V. Lepercq, S. Ilic, M. Mitchell, S. A. Luccioni, and Y. Jernite (2022). The bigscience roots corpus: A 1.6tb composite multilingual dataset. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 31809–31826. Curran Associates, Inc.

Lee, K., D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Li, J., D. Li, S. Savarese, and S. Hoi (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Li, J., D. Li, C. Xiong, and S. Hoi (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, R., L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim, Q. Liu, E. Zheltonozhskii, T. Y. Zhuo, T. Wang, O. Dehaene, M. Davaadorj, J. Lamy-Poirier, J. Monteiro, O. Shliazhko, N. Gontier, N. Meade, A. Zebaze, M.-H. Yee, L. K. Umapathi, J. Zhu, B. Lipkin, M. Oblokulov, Z. Wang, R. Murthy, J. Stillerman, S. S. Patel, D. Abulkhanov, M. Zocca, M. Dey, Z. Zhang, N. Fahmy, U. Bhattacharyya, W. Yu, S. Singh, S. Luccioni, P. Villegas, M. Kunakov, F. Zhdanov, M. Romero, T. Lee, N. Timor, J. Ding, C. Schlesinger, H. Schoelkopf, J. Ebert, T. Dao, M. Mishra, A. Gu, J. Robinson, C. J. Anderson, B. Dolan-Gavitt, D. Contractor, S. Reddy, D. Fried, D. Bahdanau, Y. Jernite, C. M. Ferrandis, S. Hughes, T. Wolf, A. Guha, L. von Werra, and H. de Vries (2023). Starcoder: may the source be with you!

Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár (2014). Microsoft coco: Common objects in context. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

Liu, S., L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar (2023). Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*.

Liu, Y., G. Zhu, B. Zhu, Q. Song, G. Ge, H. Chen, G. Qiao, R. Peng, L. Wu, and J. Wang (2022). Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 16705–16717. Curran Associates, Inc.

Loshchilov, I. and F. Hutter (2017). Fixing weight decay regularization in adam. *CoRR abs/1711.05101*.

Luccioni, A. S., C. Akiki, M. Mitchell, and Y. Jernite (2023). Stable bias: Analyzing societal representations in diffusion models.

Marino, K., M. Rastegari, A. Farhadi, and R. Mottaghi (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mishra, A., K. Alahari, and C. V. Jawahar (2012). Scene text recognition using higher order language priors. In *BMVC*.

Nichol, A., P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

Ortiz Suárez, P. J., L. Romary, and B. Sagot (2020, July). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 1703–1714. Association for Computational Linguistics.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 27730–27744. Curran Associates, Inc.

Piktus, A., C. Akiki, P. Villegas, H. Laurençon, G. Dupont, A. S. Luccioni, Y. Jernite, and A. Rogers (2023). The roots search tool: Data transparency for llms.

Radenovic, F., A. Dubey, A. Kadian, T. Mihaylov, S. Vandenhende, Y. Patel, Y. Wen, V. Ramanathan, and D. Mahajan (2023). Filtering, distillation, and hard negatives for vision-language pre-training.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Rae, J. W., S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P.-S. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J.-B. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d'Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. Johnson, B. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving (2022). Scaling language models: Methods, analysis & insights from training gopher.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu, and M. Chen (2022). Hierarchical text-conditional image generation with clip latents.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer (2021). High-resolution image synthesis with latent diffusion models.

Saharia, C., W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi (2022). Photorealistic text-to-image diffusion models with deep language understanding.

Schuhmann, C., R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave,

K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 25278–25294. Curran Associates, Inc.

Schuhmann, C., R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Sharma, P., N. Ding, S. Goodman, and R. Soricut (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Singh, A., R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela (2022). FLAVA: A foundational language and vision alignment model. In *CVPR*.

Singh, A., V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach (2019). Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326.

Sorscher, B., R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos (2022). Beyond neural scaling laws: beating power law scaling via data pruning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 19523–19536. Curran Associates, Inc.

Srinivasan, K., K. Raman, J. Chen, M. Bendersky, and M. 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*, SIGIR '21, New York, NY, USA, pp. 2443–2449. Association for Computing Machinery.

Team, M. N. (2023). Introducing mpt-7b: A new standard for open-source, commercially usable llms.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample (2023). Llama: Open and efficient foundation language models.

Wang, P., A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang (2022, 17–23 Jul). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 23318–23340. PMLR.

Wang, Q., Y. Fang, A. Ravula, F. Feng, X. Quan, and D. Liu (2022). Webformer: The web-page transformer for structure information extraction.

Wang, W., H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei (2022). Image as a foreign language: Beit pretraining for all vision and vision-language tasks.

Webster, R., J. Rabin, L. Simon, and F. Jurie (2023). On the de-duplication of laion-2b.

Workshop, B., :, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. D. Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. V. Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy,

M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. D. Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf (2023). Bloom: A 176b-parameter open-access multilingual language model.

Xie, S. M., H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu (2023). Doremi: Optimizing data mixtures speeds up language model pretraining.

Yang, Z., Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang (2022). An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 36, pp. 3081–3089.

Young, P., A. Lai, M. Hodosh, and J. 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.

Yu, J., Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu (2022). Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Yuan, S., S. Zhao, J. Leng, Z. Xue, H. Zhao, P. Liu, Z. Gong, W. X. Zhao, J. Li, and J. Tang (2022). Wudaomm: A large-scale multi-modal dataset for pre-training models.

Yuksekgonul, M., F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou (2023). When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Zhang, B. and R. Sennrich (2019). Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada.

Zhang, J., Y. Zhao, M. Saleh, and P. J. Liu (2019). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Zhang, R., J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao (2023). Llama-adapter: Efficient fine-tuning of language models with zero-init attention.

Zhang, S., S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer (2022). Opt: Open pre-trained transformer language models.

Zhou, Y., Y. Sheng, N. H. Vo, N. Edmonds, and S. Tata (2021). Simplified dom trees for transferable attribute extraction from the web. *ArXiv abs/2101.02415*.

Zhu, W., J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi (2023). Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We think that the release of such a dataset strikes a constructive trade-off between the risks associated with datasets built on top of crawled web pages (for instance, the presence of images with faces, the potential of PII in texts, offensive, insulting or threatening, etc.) with the future works that a dataset of such scale, quality and thoughtful filtering can enable. We further discuss these points in A.3.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We read the ethics review guidelines and tried our best to match the expectations. Our content is extracted from publicly available websites at the time of the web crawl. Given the size of our dataset, it would be prohibitive to get the explicit consent of the authors of these websites. Instead, we respect the choice of content creators by removing opted-out images. Such a strategy cannot be exhaustive and we remain available for content creators to opt-out of the dataset.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will release the code used for the creation of the model and its training, along with the model itself.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We mentioned the libraries we used.
   (b) Did you mention the license of the assets? [Yes] We only used open-source libraries.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See the ethics review guidelines part.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The dataset we are releasing is built from publicly accessible websites. As such, there is no content in our dataset that hasn't been publicly visible on the web at some point. Similarly, the dataset might contain texts or images that can be considered offensive, insulting, or threatening, as such data is prevalent on the web. We took measures to remove pornographic content and low-quality texts as much as possible. We did not take additional intentional measures to remove personal information. A manual inspection of 250 random samples reveals that there isn't obvious

personally identifiable information (excluding celebrities and people mentioned in news articles), although it is likely that the dataset contains some.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]