Multimodal Large Language Models for Low-Resource Languages: A Case Study for Basque

Lukas Arana, Julen Etxaniz, Ander Salaberria, Gorka Azkune
HiTZ Basque Center for Language Technology - Ixa NLP Group, University of the Basque Country UPV/EHU
{lukas.arana, julen.etxaniz, ander.salaberria, gorka.azkune}@ehu.eus

Abstract

Current Multimodal Large Language Models exhibit very strong performance for several demanding tasks. While commercial MLLMs deliver acceptable performance in low-resource languages, comparable results remain unattained within the open science community. In this paper, we aim to develop a strong MLLM for a low-resource language, namely Basque. For that purpose, we develop our own training and evaluation image-text datasets. Using two different Large Language Models as backbones, the Llama-3.1-Instruct model and a Basque-adapted variant called Latxa, we explore several data mixtures for training. We show that: i) low ratios of Basque multimodal data (around 20%) are already enough to obtain solid results on Basque benchmarks, and ii) contrary to expected, a Basque instructed backbone LLM is not required to obtain a strong MLLM in Basque. Our results pave the way to develop MLLMs for other low-resource languages by openly releasing our resources.

Keywords: Low-Resource Languages, Language Modeling, Natural Language Generation, Datasets

1. Introduction

Multimodal Large Language Models (MLLMs) (Zhang et al., 2024) aim to further improve assistance systems by combining textual data with other information modalities, such as images, video, or audio. In the context of MLLMs designed for text and image processing, these systems can perform novel tasks that traditional text-only Large Language Models (LLMs) cannot natively support, such as image captioning, visual question answering, or optical character recognition, among others. Due to these capabilities, the majority of the most advanced LLMs are being shared as natively multimodal (OpenAI, 2025; DeepMind, 2025).

However, current MLLMs are primarily trained with English resources and thus still face performance degradation in low-resource languages. This aspect has been thoroughly studied (Yue et al., 2025), showing the multilingual performance gap between modern proprietary models and open-weight alternatives. Despite recent efforts to improve the multilingual capabilities of MLLMs (Dash et al., 2025), there is still a significant performance gap between open-weight and proprietary systems, especially in the context of low-resource languages.

Developing an MLLM requires addressing a vast number of design decisions (Zhang et al., 2024), many of which have not been properly explored for low-resource languages. This paper proposes to study the development and performance of various training recipes and evaluation methods to build **the first open MLLM for Basque**. Although centred on a single language, our study likely generalizes to other similarly resourced languages. Basque is a low-resource language that lacks multimodal datasets and ranks around 50th in Common Crawl with roughly 1,000× less text data than English.

In this paper, we create **the first multimodal datasets for Basque**, both for training and evaluation. We only use open resources, avoiding the use of proprietary systems. We rely on translation procedures specifically adapted to the requirements of the datasets. As a result, we generate more than 3 million image-text instances for training and around 8 thousand evaluation instances from human-validated benchmarks[1].

With our new datasets, we run a systematic exploration with MLLMs, following the late-fusion paradigm (Dai et al., 2024) to adapt two pretrained LLMs for multimodal tasks: English-centric Llama-3.1-8B-Instruct (Llama-Team, 2024) and Basque instructed Latxa-Llama-3.1-8B-Instruct (Sainz et al., 2025). In our experiments, we find that:

1) Low ratios of Basque multimodal data are enough to build a strong MLLM for Basque. Using only 20% of the training data in Basque (the rest in English), we already achieve a very performant MLLM for our Basque benchmarks. Furthermore, our experiments suggest that English multimodal data could be enough for a decent performance, assuming that text-only Basque data can be used for training. The implications of this finding are important, since obtaining aligned image-text data for a low-resource language is generally more difficult than text-only data.

2) A Basque instructed backbone LLM is not required to build a strong MLLM for Basque. Contrary to expectations, an English-centric backbone LLM can achieve the same performance, even for open-ended text generation.

[1] https://huggingface.co/collections/HiTZ/multimodal-latxa

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9172–9187
11–16 May 2026.

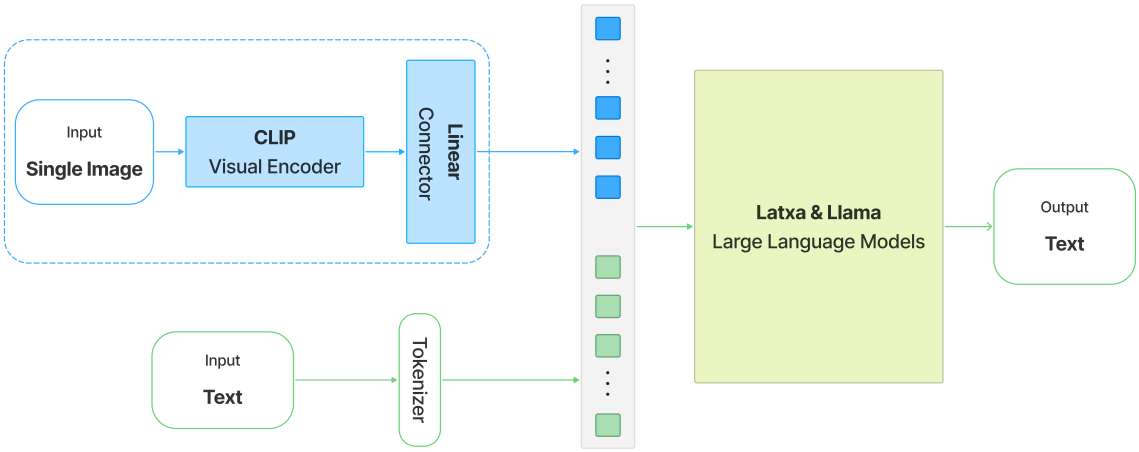

Figure 1: The late-fusion MLLM architecture used in this work. Our MLLMs have a visual encoder to represent input images, a connector to project visual representations into the embedding space of the LLM, and an LLM to process image inputs and textual queries to generate textual answers.

2. Related Work

The current Multimodal Large Language Model field is dominated by proprietary models. According to the vision LMArena ranking[2] (Chiang et al., 2024), the leading models are Gemini-2.5-Pro (DeepMind, 2025), GPT-5 (OpenAI, 2025), and Claude-4 (Anthropic, 2024). While these state-of-the-art MLLMs demonstrate superior performance, their architectures and training methodologies are largely unknown.

Among open weight alternatives, Qwen3-VL-235B (Yang et al., 2025) ranks competitively in the arena leaderboard, achieving notable results. However, the composition of its dataset is still unknown. The Molmo-72B (Deitke et al., 2025) is the best-performing model that offers open training data and is located around the 57th position in multimodal performance.

The vision LMArena and most other multimodal benchmarks primarily focus on evaluating the English performance of MLLMs, ignoring their multilingual performance. Consequently, efforts to develop open multilingual MLLMs are limited to a few works, such as Pangea (Yue et al., 2025) and Ayavision (Dash et al., 2025), which address this perspective by training on multimodal training data of multiple languages obtained through translation.

Despite these efforts, there is still a substantial performance gap between open and proprietary MLLMs (Yue et al., 2025). In fact, this gap is likely even more pronounced for low-resource languages, which are excluded from the training and evaluation datasets of existing works. To the best of our knowledge, this work represents the first effort at scale to develop an MLLM specifically for a low-resource language.

For Basque, this performance gap has recently been overcome in the text-only setting by Latxa-Llama-3.1 (Sainz et al., 2025), a family of open LLMs based on Llama-3.1 (Llama-Team, 2024) specially trained for Basque. These models have highlighted the potential of using translation for text-only datasets, showing competitive results with proprietary models.

3. MLLM Architecture and Training

In this work, we adopt the late-fusion architecture for building MLLMs. This choice is motivated by two main factors: (i) the late-fusion design is prevalent among most open MLLMs (Tong et al., 2024; Liu et al., 2024; Dai et al., 2024; Deitke et al., 2025), and (ii) it enables us to assess the feasibility of already pretrained LLMs, such as Latxa. The following sections provide a detailed description of the architecture (§3.1) and the two-stage training procedure typically used for developing such models (§3.2).

3.1. Architecture

The late-fusion architecture trains an instructed backbone LLM with the visual representations generated by a pretrained visual encoder (see Figure 1). The role of this connector is to transform the visual representations into the embedding space of the LLM. Once converted, the embedding representations of both the visual features and the text tokens are processed simultaneously by the backbone LLM to produce the textual output required

[2]Accessed at: Oct 6th, 2025.

for a multimodal task.

As the pretrained visual encoder, we have opted to use a CLIP visual encoder (Radford et al., 2021), after performing some initial experiments with SigLIP (Zhai et al., 2023) and obtaining very similar performances as in (Deitke et al., 2025). More concretely, we use the `clip-vit-large-patch14-336` ViT, which has already proven successful for the multilingual (Yue et al., 2025) and general-purpose scenarios (Liu et al., 2023).

Regarding the neural vision-language connector, we use a fully connected linear layer. This choice is based on PaliGemma (Beyer et al., 2024), which concludes that the single-layer linear connector outperforms more complex methods while being simpler and more efficient.

Finally, we compare the MLLMs based on the `Llama-3.1-8B-Instruct` (Llama) and `Latxa-Llama-3.1-8B-Instruct` (Latxa) models to assess the impact of a Basque-instructed backbone LLM. Since these LLMs share the same architecture and pretraining stage, they provide a fair comparison for evaluating how Basque-specific training influences posterior multimodal capabilities.

3.2. Training procedure

Late-fusion MLLMs are usually trained following a two-stage procedure (Zhang et al., 2024) consisting of : i) Vision-Language Alignment and ii) Multimodal Instruction Tuning.

Stage 1: Vision-Language Alignment. During the Vision–Language Alignment stage, the aim is to align the embedding spaces of the vision encoder and the backbone LLM. For this purpose, only the connector between them is trained while the other components remain frozen. This setup prevents the LLM from receiving out-of-distribution visual tokens in the second stage, thereby reducing data drift and stabilizing training.

We perform this stage equally for two backbone LLMs, namely Latxa and Llama, sharing the same training data and visual encoder. These systems will then be used for the subsequent Multimodal Instruction Tuning stage.

Stage 2: Multimodal Instruction Tuning. The Multimodal Instruction Tuning stage fine-tunes both the connector and the backbone LLM. This stage is where the backbone LLM learns to follow complex multimodal instructions. We experiment with various training configurations during the second phase to directly evaluate how different training dataset mixes affect the model's ability to process and generate responses.

4. Basque Multimodal Datasets

Developing an MLLM for a specific language requires large-scale training datasets and evaluation benchmarks in the target language. Given the scarcity of resources for Basque, and following common practices in multilingual MLLMs (Yue et al., 2025; Dash et al., 2025), we have created new multimodal datasets for Basque by translating from English-centric datasets. To do so, we analyzed the textual parts of each dataset to assess the most suitable translation method, alternating between sentence-level neural translators for Basque and text-only LLMs. The details for those procedures can be found in Appendix B. In total, we have created two multimodal datasets for training (§4.1) and four evaluation benchmarks (§4.2).

4.1. Training datasets

Stage 1. In the standard two-stage training procedure (§3.2), image captioning is normally used for Stage 1, as it is a suitable task to align visual embeddings to the LLM embedding space. Following previous works (Liu et al., 2023; Tong et al., 2024), we opt to use the Conceptual Captions dataset (CC3M) (Sharma et al., 2018). CC3M consists of 3.3 million image-caption pairs, of which 2.8 million samples are available (some links are currently broken). Unlike other better curated datasets such as COCO (Lin et al., 2014), this dataset sources image-text pairs directly from web content, resulting in significantly greater diversity. As the original captions of the dataset are mainly short sentences, we translate them into Basque using a sentence-level specialized translator system[3]. The result is the CC3M_{Eus} dataset, with a total of 2.8 million image captions in Basque.

Stage 2. For the Multimodal Instruction Tuning stage, the models are typically trained on a set of diverse and carefully filtered multimodal instruction datasets. These datasets involve both general-purpose multimodal instructions and specialized datasets for specific multimodal tasks (OCR, document understanding, object counting, and so on). Since we have focused only on general-purpose capabilities, we have selected the Pixmo-ask-model-anything dataset (Pixmo-AMA) (Deitke et al., 2025). Pixmo-AMA contains a diverse collection of 162k human-annotated question–answer multimodal instructions. The dataset was generated through an iterative process in which a text-only LLM proposed multimodal instructions that were then evaluated by human annotators. Annotators could accept or reject each sample, and if rejected, they provided feedback to help the LLM refine its output until an

[3] https://huggingface.co/HiTZ/mt-hitz-en-eu

Dataset	Acceptance Rate	Agreement
$VQAv2_{Eus}$	0.8375	0.887
$Pixmo\text{-}CapQA_{Eus}$	0.8625	0.887
$A\text{-}OKVQA_{Eus}$	0.975	0.95

Table 1: Sample acceptance rate and mean inter-annotator agreement between the 4 annotators. The agreement has been calculated by using 40 images.

acceptable answer was achieved. Given the complex questions and long answers of the dataset, requiring multi-sentence coherence, we translate it into Basque using the best open LLM possible, Latxa-Llama-3.1-Instruct-70B (Sainz et al., 2025). Sentence-level translators were discarded since they do not keep the coherence of multi-sentence answers. In total, due to some of its images not being currently available, the $Pixmo\text{-}AMA_{Eus}$ dataset contains 146k images, questions and answers in Basque.

4.2. Evaluation datasets

Following standard practices in the field (Dai et al., 2024; Deitke et al., 2025), we evaluate our MLLMs for: i) multimodal understanding and ii) language generation with multimodal inputs. For evaluating multimodal understanding, we use close-ended benchmarks (§4.2.1), and for language generation, we build an open-ended benchmark (§4.2.2).

4.2.1. Close-ended benchmarks

As we are mainly interested in measuring the capabilities of Basque MLLMs in general multimodal understanding, we discard benchmarks that focus on specific skills such as OCR or table/chart understanding. In this context, we select three close-ended benchmarks: VQAv2 (Goyal et al., 2016), A-OKVQA (Schwenk et al., 2022) and Pixmo-CapQA (Deitke et al., 2025).

VQAv2 (Goyal et al., 2016) is the second iteration of the Visual Question-Answering dataset. The dataset focuses on the visual question-answering task, where given an image and a question, models have to answer that question. As opposed to the other close-ended evaluation benchmarks, VQAv2 requires generating single-word or short-phrase answers. Each instance contains a list of 10 human-generated answers as ground-truth. We use the VQA accuracy as the evaluation metric, as proposed by Goyal et al. (2016). However, since the MLLMs in our study were not specifically trained on specialized datasets with short answers, they cannot produce single-word answers even when implicitly requested in the prompt (Liu et al., 2024). To address this limitation, we modified the evaluation procedure of the benchmark to use inclusion rather than exact string matching. Under this approach, responses are considered correct as long as they contain a word from the ground truth answer anywhere within the generated text.

A-OKVQA (Schwenk et al., 2022) is a multiple-choice multimodal evaluation benchmark. It consists of 25k image-question pairs along with four possible answers to the provided question. We chose this benchmark because, in contrast with other VQA multiple-choice datasets, models need to relate their internal world knowledge with the visual input to answer the provided questions successfully. Following standard practice for multiple-choice benchmarks, the selected answer corresponds to the option with the highest log-probability at inference. Accuracy was then used as the evaluation metric, since the answer distribution is balanced.

PixMo-CapQA (Deitke et al., 2025) is a semi-automatically generated multimodal dataset that derives from dense image captions. It consists of 214k question-answer pairs generated from 165k distinct images, from which we have filtered only the yes/no questions for evaluation purposes. Therefore, we have created a multimodal binary multiple-choice dataset that requires models to demonstrate knowledge comprehension to answer the provided questions successfully. We follow the same evaluation procedure as in A-OKVQA, using the log-probabilities of the models to select the answer and accuracy as the evaluation metric.

For each close-ended benchmark, we have created a subset of 2.5k randomly sampled examples. We translate the three benchmarks with the Latxa-Llama-3.1-Instruct 70B (Sainz et al., 2025), feeding the model with the original question and all possible answers for a given instance. We created specific translation prompts for each benchmark, using two-shot examples to follow a specific format. The specific prompts will be found in Appendix B. As a result of this translation process, we created the Basque evaluation benchmarks $VQAv2_{Eus}$, A-

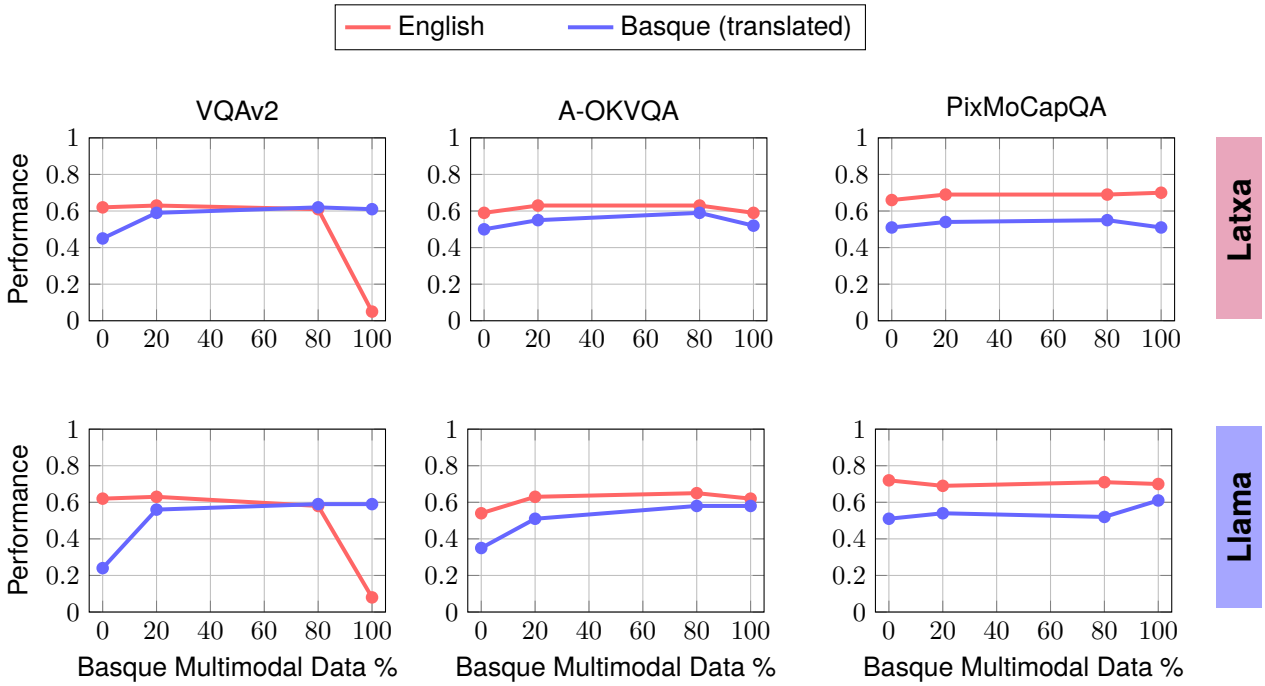

Figure 2: Accuracy across multimodal benchmarks of Latxa-based (top row) and Llama-based (bottom row) MLLMs trained with different percentages of Basque Multimodal Instruction Data. The models are evaluated on the English (original) and Basque (translated) versions of close-ended benchmarks.

OKVQA$_{Eus}$, and Pixmo-CapQA$_{Eus}$, and their English parallel versions. In this way, we can evaluate MLLMs both in English and Basque using the same instances. Evaluations were conducted using the LLMs-eval (Zhang et al., 2025) suite. We extended its support to include the *A-OKVQA* and *PixMo-CapQA* datasets in addition to the already supported *VQAv2*.

The quality of the translations is crucial for evaluation benchmarks, so four Basque native speakers validated our benchmarks. For each benchmark, samples were uniformly sampled for annotation. Each annotator evaluated 40 question–answer pairs per benchmark. From which 20 are from a shared common set and 20 from their own independent set. Across all annotators, each benchmark received a total of 100 annotations. The inter-annotator agreement is the mean of accuracy between the four annotators on each image of the shared common set. Table 1 shows high acceptance rates and agreement, confirming the quality of the translations and the suitability of our benchmarks for MLLM evaluation.

4.2.2. Open-ended benchmarks

Close-ended benchmarks are a good option to assess the multimodal understanding of models. However, we also want to measure the quality and coherence of the generated textual output for multimodal scenarios. Thus, we also use an open-ended generation benchmark.

WildVision (Lu et al., 2024) is a recent benchmark consisting of 500 text-image pairs obtained from the WildVision-Arena; a platform that collected human preferences to evaluate MLLMs. Due to numerous samples requiring specific multimodal capabilities such as OCR, a human annotator has filtered the dataset to only incorporate those pairs based on general multimodal capabilities. The filtered dataset resulted in a total of 199 samples. We translated those samples to Basque using Latxa-Llama-3.1-Instruct 70B (Sainz et al., 2025) and manually review the quality of all the samples. As a result, we created the WildVision$_{Eus}$ dataset, with 199 images and open-ended questions in Basque. Recall that the original benchmark does not provide any ground-truth answers, which complicates the evaluation process (§5).

5. Experiments and Findings

We evaluate our Basque MLLMs for close-ended and open-ended generation tasks, both in English and Basque. To train our models, we have built upon the training codebase of Llava (Liu et al., 2023), following the recommended hyperparameters by Pangea (Yue et al., 2025). The codebase will be public, and the specifics of the training configuration and infrastructure will be explained in detail in Appendix A.

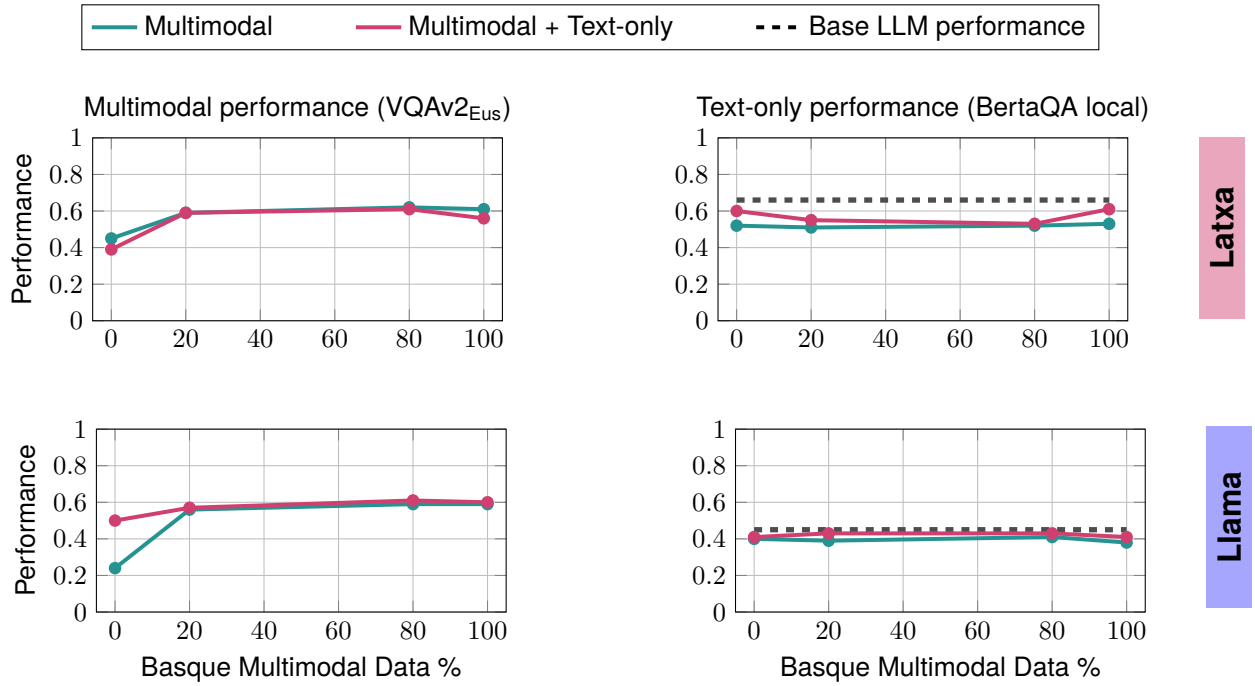

Figure 3: Accuracy across Basque multimodal (left) and text-only (right) benchmarks of Latxa-based (top row) and Llama-based (bottom row) MLLMs trained with different percentages of Basque Multimodal Instruction Data. The models are evaluated on the English (original) and Basque (translated) versions.

All MLLM configurations share the same Stage 1 training recipe and data. We trained the two Stage 1 models in a mix of the translated and original CC3M dataset. However, due to the lack of previous studies on the optimal proportions of multilingual data for this stage, we followed the recommendations for Stage 2 (Yue et al., 2025) while focusing on the Basque language. Particularly, we have chosen a ratio of 80% of Basque and 20% English samples. As for the hyperparameters, we used the AdamW optimizer together with a cosine learning rate scheduler with an initial learning rate of 10^{-3} and a warm-up ratio of 0.03. We use mixed-precision training and a sequence length of 8192 to achieve a global batch size of 32 samples.

For Stage 2, we have used the Pixmo-AMA dataset to explore data mixture strategies for multimodal instruction tuning (§5.1), how text-only performance behaves compared to the original backbone LLM (§5.2), and whether a Basque instructed LLM is required for a strong Basque MLLM (§5.3). Contrary to Stage 1, we initialized the learning rate scheduler at 2×10^{-5}, with a weight decay of 0.01 and a warm-up ratio of 0.03. We use mixed-precision training and a sequence length of 8192 to achieve a global batch size of 128 samples.

5.1. Exploring multimodal data mixture strategies

To study how the training data mixtures affect the MLLM performance, we trained Latxa-based and Llama-based architectures on the Pixmo-AMA dataset using four Basque–English sample ratios: 0:100, 20:80, 80:20, and 100:0. It should be noted that each split contains the same underlying samples, with only their languages being swapped between English and Basque.

We train a total of eight configurations (2 architectures on 4 training splits), which we evaluate using the three multimodal close-ended benchmarks described in Section 4: VQAv2, A-OKVQA and PixMoCapQA, both in English and Basque. The results are presented in Figure 2. As can be observed, the best configurations for both backbone LLMs coincide using 80% Basque data and 20% English. Furthermore, Latxa and Llama backbone LLMs, obtain a similar average accuracy of 0.62 and 0.61 points across the three benchmarks in both languages, respectively. This has been the highest average among the four data-mixture ratios, although with very low margins.

Finding 1: Low ratios of Basque multimodal data are enough to obtain solid results in Basque benchmarks. In fact, the performance of the models in Basque remains comparable when at least 20% of Basque multimodal data is used to

	Human			GPT-5		
	Latxa	**Tie**	**Llama**	**Latxa**	**Tie**	**Llama**
Winning %	23.11	53.79	20.45	31.44	36.74	31.82

Table 2: Multimodal open-ended performance comparison between top-performing Llama-based and Latxa-based MLLMs across human and automated evaluation paradigms. Human evaluations resulted in inconclusive outcomes for 3% of samples due to annotation discrepancies.

train the model. In VQAv2, which is the most informative among the three benchmarks, the Llama-based configuration trained on the split composed of 100% Basque multimodal data surpasses the version trained with only 20% by only 0.03 absolute points. This gain is even smaller for the Latxa-based configuration, whose improvement is only 0.02 points. Our finding is also supported by the Pangea study (Yue et al., 2025), which reports a similar trend where performance in multimodal tasks reached a plateau once the proportion of multilingual data exceeded 20%.

Finding 2: A low ratio of English data is required to avoid catastrophic forgetting in English benchmarks. As expressed in the VQAv2 benchmark, both backbone LLM architectures show a drastic decline in English performance when only Basque multimodal data is used for training, indicating catastrophic forgetting of their English capabilities. These results highlight the need for English multimodal samples at training time to maintain English performance.

5.2. Evaluating text-only performance

Multiple studies have reported that multimodal training tends to reduce the performance of backbone LLMs on text-only tasks (Beyer et al., 2024; Yue et al., 2025). Incorporating text-only instructions into the training process has proven to be an effective way to counteract this decline (Dai et al., 2024). To analyze this effect in a low-resource context: i) we augmented the four multimodal training splits described in Section 5.1 with text-only instructions, ii) we trained MLLMs using those new training configurations, and iii) we evaluated the resulting models on both multimodal and text-only benchmarks.

The text-only instructions were randomly sampled from the Magpie-Llama-3.1-8B-Instruct-Filtered-1M dataset, originally developed for training Latxa (Sainz et al., 2025). We added a total of 29k text-only examples, 20% of them being in English and 80% in Basque while maintaining the multimodal samples. Consequently, the final training datasets consist of 173k samples, from which 17% is text-only and 83% is multimodal data.

In addition to multimodal evaluation, we evaluate the models for text-only capabilities using the BertaQA (Etxaniz et al., 2025) benchmark, a Basque multiple-choice trivia dataset emphasizing local cultural knowledge. This benchmark has been chosen for its strong correlation with human evaluation, as noted in (Sainz et al., 2025).

Figure 3 presents the performance of our previous eight configurations, comparing them with the addition of the text-only training split to the multimodal one. In this case, the text-only training split always uses a fixed 80-20 Basque-English ratio, whereas the multimodal splits' ratios are changed according to these configurations. We report only multimodal results for $\text{VQAv2}_{\text{Eus}}$, as we find this benchmark to best represent the overall Basque multimodal performance.

Finding 3: Incorporating text-only data helps reduce the Basque text-only performance drop observed relative to the original backbone LLM. Although this addition leads to an overall improvement in text-only performance, the degradation compared to the base LLM remains considerable, especially in Latxa-based configurations. Notably, this degradation is smaller when the split consists exclusively of either English or Basque content. Given that previous studies, such as (Dai et al., 2024), have demonstrated that this degradation can be effectively addressed by adding text-only data, our results may be caused by insufficient text-only instruction samples.

Finding 4: Text-only data can improve multimodal performance. This is especially the case of the Llama-based configuration trained solely with English multimodal data. The inclusion of Basque text-only data has significantly improved its multimodal performance in Basque. This suggests that, in the absence of Basque multimodal data, Basque text-only instructions can provide a bridge to transfer the multimodal capabilities acquired in English to Basque. This finding indicates modality-transfer capabilities when no multimodal data is present in the language of evaluation. We find this to be of special interest to address the scarcity of multimodal data for low-resource languages.

5.3. Evaluating the impact of backbone LLMs

The results of the close-ended evaluations, Figures 2 and 3, show similar performance between Latxa-based and Llama-based architectures in most of the configurations, suggesting that a Basque instructed backbone LLM has no advantage over a mostly English model. As close-ended generation may not demand language proficiency, we have evaluated our MLLMs in the Wildvision open-ended benchmark, to get a clearer view of the importance of the backbone LLM.

We opt for human evaluation, and due to the low scalability of this procedure, we have only evaluated the best-performing Latxa-based and Llama-based configurations. That is, we evaluate the models trained with the ratio of 80-20 Basque-English multimodal instructions, in addition to the text-only dataset.

For each question, the answers generated by both configurations have been compared pairwise to determine the preferred response or whether the result was a tie. The annotation process considered three evaluation aspects ordered by importance.

1. **Relevance and language**: Whether the response is satisfactory and is written in Basque.
2. **Quality**: Overall quality of the response in terms of correctness.
3. **Language proficiency**: The proficiency of the response in the Basque language.

For the annotation process, four native Basque speakers rated the 199 samples of the Basque Wildvision dataset (§4). All annotators evaluated the same 44 responses to calculate the inter-annotator agreement, and we used majority voting to decide the final rating for those samples. Given the lack of scalability of this evaluation approach, we also conducted the same evaluation following an MLLM-as-a-judge paradigm. More concretely, we use the `GPT-5-2025-08-07` model, which is prompted with the question, the image, and the two responses generated by Llama-based and Latxa-based MLLMs. We also prompt GPT-5 with the set of evaluation criteria to guide its assessment.

The results obtained by both evaluation methods can be found in Table 2. Notice that we will focus on human evaluation results for the analysis of the performance of MLLMs. The inter-annotator agreement among humans, measured as Cohen's Kappa, is of 0.43, which is considered as moderate.

Finding 5: A Basque backbone LLM is not required to develop a strong Basque MLLM. In fact, Latxa-based and Llama-based configurations exhibited comparable performance in 54% of the open-ended benchmarks, confirming the minimal differences observed between the two models in the close-ended benchmarks. Although the Latxa backbone shows a slight performance advantage, the difference is not significant given the low number of test samples (199), leading us to conclude that both models perform similarly.

Finding 6: Judge models may offer a good way to evaluate Basque open-ended generations, but we cannot prescind from human evaluation. Although there is a fair agreement between GPT-5 and human annotators (Cohen's Kappa 0.33), GPT-5 tends to produce significantly fewer ties, even when explicitly instructed to do so in the prompt. This reflects a documented bias in LLM-based evaluators, as discussed in prior research (Chen et al., 2024). This means that, at least for Basque, MLLM-as-a-judge is a promising method to evaluate open-ended generation at bigger scales, but as some differences with human annotators are still observed, it cannot be used in isolation yet.

6. Conclusions

In this work, we create a total of 2 multimodal training datasets and 4 evaluation benchmarks for Basque, a low-resource language. We use these resources develop the first Multimodal Large Language Model for Basque, exploring several training strategies. We find that low ratios of Basque multimodal data are already enough to perform well for Basque multimodal benchmarks. More importantly, we also find that the inclusion of Basque text-only data can enhance the multimodal performance of MLLMs in Basque, showcasing the cross-lingual transfer capabilities of these systems. Furthermore, we see that the performance of English-centric backbone LLMs are close to Basque-centric LLMs. Those three findings together indicate that training an English-centric LLM with minimal multimodal data for the target language combined by text-only instructions in that target language can provide a feasible pathway for developing MLLMs for many low-resource languages.

Many directions are open for future work. One of the important aspects that could be explored in the future is training and evaluating on multimodal cultural knowledge. As this work relies on machine translation for training and evaluation data, no Basque cultural knowledge is included. Another interesting direction is measuring how far we can get without using any multimodal data in the target language, by leveraging already trained MLLMs and cross-lingual and cross-modality transfer. Finally, more and better multimodal training and evaluation data are needed, especially for specific skills such as OCR or table/chart understanding.

Limitations

This work explores various approaches for building MLLMs for low-resource languages. However, due to the computational cost of experiments and the human annotations required, we had to limit the exploration.

One of the limitations is the choice of a single language. Although Basque is representative of many low-resource languages in the number of resources available, it could be the case that our results do not fully generalize to other low-resource languages.

Another limitation is our choice of architecture and model, where we focus on a single architecture and two different LLMs. This decision was motivated by the availability of Latxa. However, other approaches might indeed be promising, such as leveraging a powerful MLLM and training it in the target language.

Finally, due to the lack of resources available for Basque, we had to rely on machine translation for both training and evaluation. Although we validated the translated benchmarks with human annotations, this method has its inherent limitations (Artetxe et al., 2020).

Ethics Statement

MLLMs have the potential to impact society. While they can improve global information access and enable new forms of automation, they also present risks, including the enhancement of existing human biases and privacy issues for individuals. In this direction, the adoption of multilingual MLLMs for low-resource languages has the particular risk of extending this issue to local behaviors. Therefore, the development of these systems must prioritize responsible practices to address these challenges.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT1570-22 and IKER-GAITU-2 project), the Spanish Ministry of Science, Innovation and Universities (Molvi project PID2024-157855OB-C32, HumanAIze project AIA2025-163322-C61 and GeoR2-LLM project PCI2025-163286 by MICIU/AEI /10.13039/501100011033 and co-financed by the European Union) and the European Union's Horizon Europe research and innovation programme under Grant Agreement No 10113572, related to the LUMINOUS project. Julen Etxaniz holds a PhD grant from the Basque Government (PRE_2024_2_0028).

The models were trained on the Leonardo supercomputer at CINECA under the EuroHPC Joint Undertaking, project EHPC-EXT-2024E01-042. The authors also acknowledge the technical and human support provided by the DIPC Supercomputing Center.

Bibliographical References

Anthropic. 2024. Claude opus 4. https://www.anthropic.com. Large language model.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

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.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi,

Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2025. Aya vision: Advancing the frontier of multilingual multimodality.

Google DeepMind. 2025. Gemini 2.5 pro. https://ai.google.dev/gemini-api/docs/models. Advanced reasoning model. Model ID: gemini-2.5-pro.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2025. Bertaqa: how much do language models know about local culture? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Llama-Team. 2024. The llama 3 herd of models.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences.

OpenAI. 2025. Gpt-5. Large language model.

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 *International conference on machine learning*, pages 8748–8763. PmLR.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. Instructing large language models for low-resource languages: A systematic study for basque.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, page 146–162, Berlin, Heidelberg. Springer-Verlag.

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 87310–87356. Curran Associates, Inc.

Matteo Turisini, Giorgio Amati, and Mirko Cestari. 2023. Leonardo: A pan-european pre-exascale supercomputer for hpc and ai applications.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

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*, pages 11975–11986.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. LMMs-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, Albuquerque, New Mexico. Association for Computational Linguistics.

A. Multimodal Large Language Model Training

The two-stage training procedure requires different training configurations and infrastructure for each of the stages. For the Visual-Language Alignment stage, we trained the two configurations using a single node of 4 A100-sxm4 GPUs of the Hyperion cluster from the Donostia International Physics Center. Each training took a total of 80 GPU hours. This configuration led us to use a global batch size of 32 samples. We used the AdamW optimizer (Loshchilov and Hutter, 2019) together with a cosine learning rate scheduler starting on an initial learning rate of $1e^{-3}$ and warm-up ratio of 0.03. As for memory optimizations, we used gradient checkpointing and DeepSpeed ZeRO-3 (Rasley et al., 2020). These hyperparameters were based on (Yue et al., 2025), except that we opted for a context length of 8192 tokens, which results in a more VRAM-intensive training than the reference work.

For the Multimodal Instruction Tuning stage, training was conducted on the Leonardo supercomputer at CINECA's HPC infrastructure (Turisini et al., 2023) cluster. We utilize 8 nodes, providing a total of 32 A100 GPUs per training run. Following the Visual-Language Alignment phase, we maintained the context length of 8,192 tokens and employed gradient checkpointing for memory optimization. The training setup incorporated mixed-precision computation and DeepSpeed ZeRO-3 (Rasley et al., 2020) for fully sharded data parallelism across all GPUs. Given the batch size of 8 per node and gradient accumulation steps of 2, we achieve a global batch size of 128 samples. Following the Pangea methodology (Yue et al., 2025), we employed the AdamW optimizer with a cosine learning rate scheduler initialized at 2×10^{-5}, weight decay of 0.01, and a warm-up ratio of 0.03. Each model trains for approximately 3-4 hours, depending on the specific dataset configuration.

B. Translation of the Datasets

Two translation methods have been used during this work. A sentence-level translation model for the CC3M dataset and an LLM-based translation procedure for the rest of the datasets and benchmarks. For the former, we have used the mt-hitz-en-eu model with the standard hyperparameters. Namely, no repetition penalty, temperature of 1.0, and top_p of 1.0. As for the latter datasets, we adopted a similar instruction translation procedure

used for text-only datasets in the Latxa work, using the Latxa-Llama-3.1-70B-Instruct LLM (Sainz et al., 2025). The model was prompted to translate English question-answer pairs into Basque using a 2-shot configuration. The detailed prompts for each dataset are shown in Figures 4, 5, 6 and 7.

For inference with the Latxa model, we have followed the recommended hyperparameter settings. That is, a temperature of 0.7, no repetition penalty, and top_p of 0.15. As for computational resources, we have used a single node of the Leonardo supercomputer at CINECA's HPC infrastructure (Turisini et al., 2023). The node is equipped with four NVIDIA Ampere A100 SXM4 GPUs (64GB HBM2e) interconnected via NVLink 3.0 and a 32-core Intel Xeon Platinum 8358 CPU (Ice Lake) with 512GB DDR4-3200 RAM. This node configuration allowed for a batch size of 64 question-answer pairs with a context length of 2048 tokens, which we found to be enough for all the samples.

Translation Procedure for **A-OKVQA$_{Eus}$**: 2-Shot Prompting for English-Basque

System Message

Role: System
Content: You are a helpful AI assistant that specializes in English to Basque translations.\n Your task is to translate instruction datasets from English to Basque.\n\n Here are some important guidelines:\n 1. Maintain the original meaning and intent of the instructions\n 2. Use standard Basque language (batua)\n 3. Keep technical terms that don't have widely accepted Basque translations\n 4. Preserve any code snippets, variables, or special characters exactly as they appear\n\n Please provide accurate Basque translations for all text fields.

Shot 1

Role: User
Content: Translate the following question and answer to Basque \n **Question**: What is the man by the bags awaiting? \n **Answer**: 1.skateboarder, 2.train, 3.delivery, 4.cab

Role: Assistant
Content: Galdera: Zer ari da poltsekin dagoen gizona itxaroten?**Erantzuna:** 1.skaterra, 2.trena, 3.Deliveria, 4.Taxia

Shot 2

Role: User
Content: Question: Where does this man eat pizza? **Answer:** 1.office, 2.cafe, 3.motel, 4.outside

Role: Assistant
Content: Galdera: Non ari da gizon hau pizza jaten? **Erantzuna:** 1.Ofizinan, 2.Kafetegian, 3.Motelean, 4.Kanpoan

New Query

Role: User
Content: Translate the following question and answers to Basque \n\n\n **Question**: {example['question']} **Answer**: {example['answer']}

Role: Assistant
Content: Galdera: ... \n **Erantzuna:** ...

Figure 4: Two-shot prompting procedure for English to Basque translation of the A-OKVQA$_{Eus}$ benchmark. example['question'] and example['answer'] correspond to the question and answer to be translated in the example.

Translation Procedure for **PixMo-CapQA$_{Eus}$**: 2-Shot Prompting for English-Basque

System Message

Role: System
Content: You are a helpful AI assistant that specializes in English to Basque translations.\n Your task is to translate instruction datasets from English to Basque.\n\n Here are some important guidelines:\n 1. Maintain the original meaning and intent of the instructions\n 2. Use standard Basque language (batua)\n 3. Keep technical terms that don't have widely accepted Basque translations\n 4. Preserve any code snippets, variables, or special characters exactly as they appear\n\n Please provide accurate Basque translations for all text fields.

Shot 1

Role: User
Content: Translate the following question and answer to Basque \n **Question**: Is this room decorated for the 1970s?

Role: Assistant
Content: Galdera: Gela hau 1970ean bezala dago diseinatuta?

Shot 2

Role: User
Content: Translate the following question and answers to Basque \n **Question:** Is there a mirror in the room?

Role: Assistant
Content: Galdera: Ba al dago ispilurik gelan?

New Query

Role: User
Content: Translate the following question and answers to Basque \n\n\n **Question**: {example['question']} **Answer**: {example['answer']}

Role: Assistant
Content: Galdera: ... \n **Erantzuna:** ...

Figure 5: Two-shot prompting procedure for English to Basque translation of the PixMo-CapQA$_{Eus}$ benchmark. example['question'] and example['answer'] correspond to the question and answer to be translated in the example.

Translation Procedure for **VQAv2$_{Eus}$**: 2-Shot Prompting for English-Basque

System Message

Role: System
Content: You are a helpful AI assistant that specializes in English to Basque translations.\n Your task is to translate instruction datasets from English to Basque.\n\n Here are some important guidelines:\n 1. Maintain the original meaning and intent of the instructions\n 2. Use standard Basque language (batua)\n 3. Keep technical terms that don't have widely accepted Basque translations\n 4. Preserve any code snippets, variables, or special characters exactly as they appear\n\n Please provide accurate Basque translations for all text fields.

Shot 1

Role: User
Content: Translate the following question and answer to Basque \n **Question**: Explain the visual content of the image in great detail.\n **Answer**: In this outdoor wedding photo, a bride wearing a beautiful white dress is posing on a park bench surrounded by a group of five men dressed in suits. The groomsmen are wearing pink ties, adding a pop of color to the scene. The group is standing around the bride, smiling and enjoying the special occasion as they pose for the picture.\n \n. In the background, there are potted plants which further enhance the ambiance of the park setting. The bench, on which the bride is seated, is located close to the center of the image.

Role: Assistant
Content: Galdera: Azaldu irudia zehaztasun askorekin. **Erantzuna:** Ezkontzako argazki honetan, soineko zuri eder bat daraman emaztegaia parkeko banku batean dago, trajez jantzitako bost gizonez inguratuta. Neskameek gorbata arrosak daramatzate, eszenari kolorezko pop bat erantsiz. Taldea emaztegaiaren inguruan dago, irribarretsu, eta aukera bereziaz gozatzen dute koadrorako. Jarlekua, non emaztegaia eserita dagoen, irudiaren erditik hurbil dago.

Shot 2

Role: User
Content: Translate the following question and answers to Basque \n **Question:** What website copyrighted the picture? **Answer:** 1.foodiebakercom, 2.foodiebakercom, 3.foodiebaker, 4.foodiebakercom, 5.foodiebakercom, 6.http://foodiebakercom, 7.foodiebakercom, 8.foodiebakercom, 9.foodiebakercom, 10.foodiebaker

Role: Assistant
Content: Galdera: Zein webgunek du irudiaren copyrigheta? **Erantzuna:** 1.foodiebakercom, 2.foodiebakercom, 3.foodiebaker, 4.foodiebakercom, 5.foodiebakercom, 6.http://foodiebakercom, 7.foodiebakercom, 8.foodiebakercom, 9.foodiebakercom, 10.foodiebaker

New Query

Role: User
Content: Translate the following question and answers to Basque \n\n\n **Question**: {example['question']} **Answer**: {example['answer']}

Role: Assistant
Content: Galdera: ... \n **Erantzuna:** ...

Figure 6: Two-shot prompting procedure for English to Basque translation of the VQAv2$_{Eus}$ benchmark. example['question'] and example['answer'] correspond to the question and answer to be translated in the example.

Translation Procedure for **Wildvision$_{Eus}$**: 2-Shot Prompting for English-Basque

System Message

Role: System
Content: Cutting Knowledge Date: December 2023 Today Date: [2025-01-01]
You are a helpful assistant

Shot 1

Role: User
Content: Honako testu hau gaztelaniaz idatzita dago eta euskerara itzuli behar da. Itzulpena zehatza eta naturala izan behar da. Jarri ezazu bakarrik testua, ez egin erreferentziarik hau itzulpen bat denik.
Gaztelaniazko testua: example['question']
Euskerazko itzulpena:

Role: Assistant
example['answer']

Figure 7: Prompt for the Spanish to Basque translation of the Wildvision$_{Eus}$ benchmark. example['question'] and example['answer'] correspond to the question and answer to be translated in the example. We used a 0-shot strategy for open-ended settings.