The Breeze 2 Herd of Models: Traditional Chinese LLMs Based on LLaMA with Vision-Aware and Function-Calling Capabilities

Breeze Team @ MediaTek Research*

Abstract

Breeze 2 is a suite of advanced multi-modal language models, available in 3B and 8B parameter configurations, specifically designed to enhance Traditional Chinese language representation. Building upon the LLaMA 3, Breeze 2 continues pretraining on an extensive corpus to enhance the linguistic and cultural heritage of Traditional Chinese. It incorporates vision-aware capabilities through a visual encoder and a bridge module, and supports function-calling via prompt templates and post-training on function-calling data. The effectiveness of Breeze 2 is benchmarked across various tasks, including Taiwan general knowledge, instruction-following, long context, function calling, and vision understanding. Furthermore, we showcase the capabilities of the its 3B model in a mobile application. We are publicly releasing all Breeze 2 models under the Llama 3 Community License[2].

1 Indroduction

In recent years, the advent of foundation models has revolutionized the field of artificial intelligence, enabling a wide array of tasks across different modalities such as language and vision. These models, which undergo extensive pre-training and post-training stages, have demonstrated remarkable capabilities in understanding and generating natural language. However, a significant gap remains in the representation and performance of these models concerning Traditional Chinese, a language predominantly used in Taiwan, Hong Kong, and Macau.

Traditional Chinese, with its rich linguistic and cultural heritage, has been underrepresented in the development of large language models (LLMs). Prominent models such as GPT-4o[3], Claude 3[4], and Llama 3 [Dubey et al., 2024] have shown exceptional proficiency in English and Simplified Chinese. Nevertheless, they often fall short in accurately capturing the nuances and complexities inherent in Traditional Chinese. This misalignment not only affects the accuracy of language generation but also overlooks the cultural significance embedded within Traditional Chinese texts.

While previous works [Hsu et al., 2024, Lin and Chen, 2023] have attempted to address this gap, they lack vision-aware and function-calling capabilities, both of which are crucial for commercial applications. Vision-aware capabilities enable models to interpret and generate content that integrates text and images, essential for industries like e-commerce and virtual assistants. Function-calling capabilities allow models to interact with external systems, perform specific tasks, and provide accurate responses, enhancing applications like automated customer support.

*All contributors to this work have made equal contributions. We list all contributors in alphabetical order by first name, as follows: Chan-Jan Hsu, Chia-Sheng Liu, Meng-Hsi Chen, Muxi Chen, Po-Chun Hsu, Yi-Chang Chen, and the supervisor Da-Shan Shiu. The author contributions are detailed in the appendix.

[2] `https://huggingface.co/collections/MediaTek-Research/breeze-2-67863158443a06a72dd29900`

[3] `https://openai.com/index/hello-gpt-4o/`

[4] `https://www.anthropic.com/news/claude-3-family`

This work aims to address the underrepresentation of Traditional Chinese in LLMs by developing a model that integrates vision-aware and function-calling capabilities. The result of our efforts is **Breeze 2**: a suite of two multi-modal language models with 3B and 8B parameters (Section 2). **Breeze 2** is built upon the impressive foundation of Llama 3 [Dubey et al., 2024] and continued pre-trained on 900 GB of high-quality Traditional Chinese corpus (Section 4.1) to enhance the linguistic and cultural heritage of Traditional Chinese. To enable vision-aware capabilities, we adapted the design from LLaVA [Liu et al., 2023] by integrating a visual encoder and a bridge module, and employed a two-stage training process: vision alignment and visual instruction tuning (Section 3). To enable function-calling capabilities, we adapted the design from Breeze-FC [Chen et al., 2024b] by integrating prompt templates for function calling (Section 5.1) and conducted post-training on function-calling data (Section 5.4).

To comprehensively evaluate the effectiveness of **Breeze 2**, we conducted extensive benchmarking across a wide range of tasks, including Taiwan general knowledge (Section 6.1), instruction-following (Section 6.1), long context handling (Section 6.2), function calling (Section 6.3), and vision understanding (Section 6.4). Additionally, the 3B parameter model is designed to be more adaptable for mobile use, and we demonstrate **Breeze 2**'s capabilities in a mobile application (Section 7.3).

2 Model Architecture

Building upon our goals to address the underrepresentation of Traditional Chinese in LLMs, Breeze 2 adopts the widely-used "ViT-MLP-LLM" architecture, originally introduced by LLaVA [Liu et al., 2023], which has become a common configuration in open-source multimodal large language models. This architectural choice allows Breeze 2 to effectively leverage the strengths of both vision and language models, while also incorporating improvements inspired by subsequent works such as InternVL 1.5 [Chen et al., 2024c].

As illustrated in Figure 1, the implementation of this architecture in Breeze 2 involves the integration of two key components:

- A pre-trained Llama 3 model, serving as the foundation for language understanding and generation.
- A pre-trained InternViT-300M-448px model, which acts as the vision encoder.

These two components are connected via a multi-layer perceptron (MLP) projector, which serves as a bridge between the visual and linguistic features.

The choice of the InternViT-300M-448px model as the vision encoder was made for two significant reasons:

1. **Visual Processing Capabilities:** The 448px input resolution allows for detailed image analysis, which is crucial for tasks requiring fine-grained visual understanding.
2. **Enhanced Chinese Language Understanding:** Notably, the InternVL pretrained model demonstrates a higher level of comprehension for Chinese characters compared to other vision models. This enhanced capability in processing Chinese text within images gives Breeze 2 a distinct advantage in Traditional-Chinese-specific visual-language tasks.

Breeze 2 is available in two model sizes, each tailored for different use cases:

- **Breeze 2 3B:** This variant is based on the Llama 3.2 3B model, offering a balance between performance and computational efficiency, making it suitable for resource-constrained environments.
- **Breeze 2 8B:** This larger variant is based on the Llama 3.1 8B model, providing enhanced capabilities for more complex tasks that require higher capacity and performance.

The choice of these specific Llama 3 variants as the base models for Breeze 2 allows for a range of applications, from more resource-constrained environments to scenarios requiring higher capacity and performance.

The MLP projector plays a critical role in aligning the visual features extracted by the InternViT model with the linguistic features processed by the Llama model. This alignment is essential for enabling seamless multimodal reasoning and generation tasks.

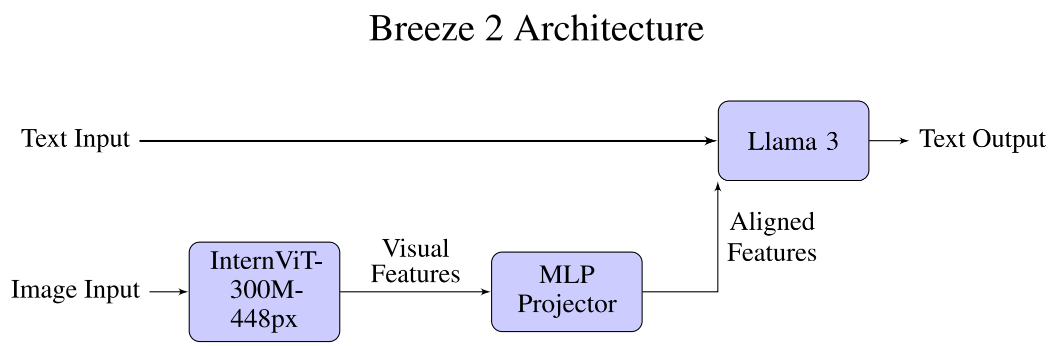

Figure 1: Overview of Breeze 2 architecture, showing the integration of the InternViT vision encoder, MLP projector, and Llama 3 language model.

3 Training Method

The training process for Breeze 2 involves a comprehensive multi-stage approach, designed to enhance the model's capabilities in both textual and multimodal domains, with a specific focus on Traditional Chinese and Taiwan-centric content. As illustrated in Figure 2, the training method consists of three main stages:

1. **Extended Text-to-Text Pre-training:**
   - Starting with either the Llama 3.2 3B or Llama 3.1 8B as the base model
   - Extended pre-training to improve Traditional Chinese language understanding and generation
   - Incorporation of Taiwan-specific knowledge and contexts
2. **Vision-Alignment Pre-training:**
   - Integration of the InternViT-300M-448px vision encoder
   - Aligning visual features with the language model through the randomly initialized MLP projector
   - Emphasis on multimodal understanding, particularly for Traditional Chinese text in images
3. **Post-training:** This final stage combines three critical components, with data from all components integrated for comprehensive training:
   - **Text-to-Text Instruction Tuning:** Further refinement of the model's ability to follow text-based instructions and enhancing task-specific performance in Traditional Chinese
   - **Visual Instruction Tuning:** Joint training on text and image inputs, improving the model's capability to generate coherent responses based on both textual and visual information, with a focus on Taiwan-related multimodal tasks
   - **Fine-tuning for Function Calling:** Enhancing the model's ability to determine whether and which functions to call in response to Traditional Chinese language prompts

This comprehensive training approach ensures that Breeze 2 not only excels in Traditional Chinese language processing but also demonstrates strong performance in multimodal tasks and function calling, particularly those relevant to Taiwan-specific contexts.

4 Pre-training

Building upon the architectural foundation described in the previous section, the Breeze 2 series leverages the powerful Llama 3 series models as its core. This chapter delves deeper into the

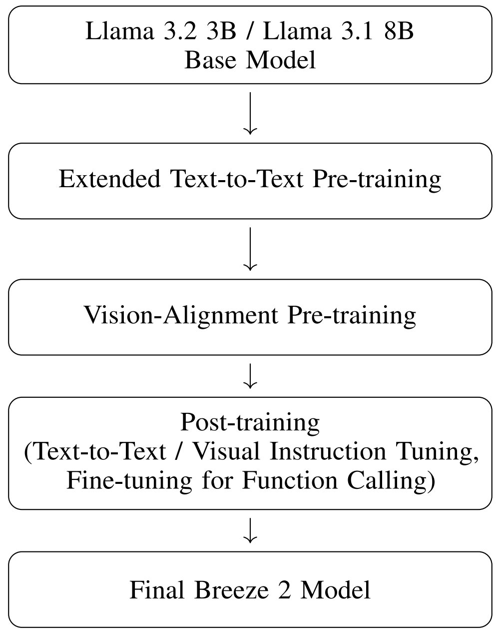

Figure 2: Training Process of Breeze 2. This figure illustrates the comprehensive multi-stage training approach for Breeze 2, starting from the Llama 3.2 3B or Llama 3.1 8B base model. The process includes Extended Text-to-Text Pre-training to enhance Traditional Chinese language understanding, Vision-Alignment Pre-training to integrate visual features, and Post-training for refining text and visual instruction tuning and function calling capabilities. The final output is the Breeze 2 model, optimized for Traditional Chinese and Taiwan-specific contexts.

sophisticated pre-training process that not only elevates Breeze 2 from a basic language model to a versatile multimodal system but also significantly enhances its capabilities in Traditional Chinese language processing and understanding of Taiwan-specific contexts.

The extended pre-training strategy of the Breeze 2 series is meticulously designed to strengthen the model's proficiency in Traditional Chinese, with a particular focus on Taiwan-related knowledge. This process consists of two distinct stages:

1. **Text-to-Text Training:** This initial stage focuses on enhancing the language understanding and generation capabilities inherited from the Llama 3 base model. Special emphasis is placed on improving the model's grasp of Traditional Chinese linguistic nuances and Taiwan-specific terminology and concepts.
2. **Vision Alignment:** The second stage involves a crucial pre-training step that bridges the text-to-text base model with the InternViT image encoder, as introduced earlier. This stage is key to enabling Breeze 2's multimodal functionalities, with particular attention to visual understanding in a Traditional Chinese and Taiwanese cultural context.

Furthermore, this chapter will provide a comprehensive overview of the carefully curated datasets employed in each stage of pre-training. These datasets are specifically selected to include a rich variety of Traditional Chinese texts and images relevant to Taiwan's cultural, social, and historical landscape. We will also detail the training methods and techniques utilized to optimize the model's performance across both textual and visual domains, with a consistent focus on Traditional Chinese and Taiwan-centric content.

4.1 Text-to-text Pre-training Data for Localization

For the text-to-text pre-training stage focused on Traditional Chinese and Taiwan-specific content, we curated a diverse dataset from various sources. Table 1 summarizes the composition of this dataset.

Table 1: Composition of Text-to-text Pre-training Dataset

Category	Size (GB)	Description
Web Crawl	421.3	CommonCrawl data (CC100 ZHTW [Conneau et al., 2020], C4 ZHTW [Dodge et al., 2021], FineWeb-zhtw [Lin et al., 2024]) and other miscellaneous web data
Academic	94.0	Theses, research plans, and reports
Code	110.0	Programming-related content (TheStack [Kocetkov et al., 2022])
News	131.0	News articles
Dialogue	37.0	Transcribed conversation, legislative records
Blog	20.7	Blog content
Forum	60.0	Q&A content, BBS, forum discussions
Judicial	21.4	Legal judgments and documents
Books	1.6	Book excerpts and summaries
Wiki	1.1	Wikipedia content in Traditional Chinese
Total	**898.1**	

This carefully curated dataset encompasses a wide range of content types, ensuring comprehensive coverage of Traditional Chinese language usage and Taiwan-specific knowledge. The dataset includes web crawl data, academic publications, code repositories, news articles, dialogues, social media content, forum discussions, legal documents, and literary works. This diverse collection aims to enhance the model's understanding of various domains and linguistic styles prevalent in Taiwan's cultural and social context.

4.2 Training Details of Text-to-text Pre-training

The text-to-text pre-training stage for Breeze 2 was conducted using 5,760 H100-GPU hours for the 3B version and 16,620 H100-GPU hours for the 8B version. This training process allowed for comprehensive learning of the model's language understanding and generation capabilities, particularly for Traditional Chinese and Taiwan-specific contexts.

Key training parameters and configurations include:

- **Parallelism Strategy:** For the 3B version, only data parallelism was employed during training. In contrast, the 8B version required a tensor parallelism size of 2, in addition to data parallelism, to efficiently manage the larger model size.
- **Precision and Optimization:** The model utilized bfloat16 precision and Transformer Engine's FP8 training support, as provided by NVIDIA's NeMo Framework [Kuchaiev et al., 2019].
- **Batch Size and Sequence Length:** A global batch size of 1,024 was employed, with a maximum sequence length of 8,192 tokens. This configuration allowed for maintaining a balance between computational efficiency and model performance.
- **Learning Rate and Scheduler:** The training utilized a distributed fused Adam optimizer with an initial learning rate of 1×10^{-5}. A cosine annealing learning rate scheduler was implemented, featuring a warmup period of 500 steps and a minimum learning rate of 1×10^{-6}.

These training details provide insight into the pre-training process of Breeze 2, leveraging the capabilities of NVIDIA's NeMo Framework. The difference in training hours and parallelism strategies between the 3B and 8B versions reflects the adaptations necessary for training larger language models efficiently.

4.3 Pre-training Data for Visual Alignment

To achieve effective visual alignment, we implemented a comprehensive 2-phase pre-training approach. This method ensures that our models are well-prepared for a variety of visual tasks.

In the first phase, we focused on training only the MLP layers. This initial step allows the model to establish a foundational understanding of the data, utilizing approximately 3.42M data points.

Following this, we transitioned to the second phase, which involves full weights pre-training. During this phase, all components of the model, including ViT, MLP, and LLM, are trainable, utilizing approximately 10.89M data points. This holistic training approach ensures that the model can leverage the strengths of each component to achieve superior performance.

The pre-training dataset for our Breeze 2 model is sourced from a diverse array of publicly accessible datasets. These datasets cover a wide range of tasks, such as captioning, visual question answering, detection, grounding, and OCR (Optical Character Recognition). In Tables 2 and 3, we detail the specific datasets used in each phase of our pre-training process.

Table 2: Datasets Used for MLP Pre-training

Category	# of Data	Datasets
Captioning	2.8M	ALLVA [Chen et al., 2024a], ShareGPT4V [Chen et al., 2023], ShareCaptioner [Chen et al., 2023], Traditional Chinese version of LAION [Schuhmann et al., 2022], Traditional Chinese version of COYO [Byeon et al., 2022], Traditional Chinese version of GRIT-Caption [Peng et al., 2023]
General QA	0.22M	VQAv2 [Goyal et al., 2017]
OCR	0.4M	Traditional Chinese version of SynthDoG [Kim et al., 2022]
Total	**3.42M**	

Table 3: Datasets Used for ViT + MLP + LLM Pre-training

Category	# of Data	Datasets
Captioning	3.42M	ShareGPT4V, ShareCaptioner, ALLVA, ShareGPT4o [Chen et al., 2024c], TaiTravel [Chang, 2024], Datasets from the cauldron [Laurençon et al., 2024], Traditional Chinese version of ShareGPT4V
General QA	1.02M	VQAv2, Datasets from the cauldron
Detection	0.2M	AS-V2 [Wang et al., 2024]
OCR	2.05M	ArT [Chng et al., 2019], COCO-Text [Veit et al., 2016], Datasets from the cauldron, In-house OCR dataset, SynthDoG, Traditional Chinese version of SynthDoG
Reasoning	0.39M	Datasets from the cauldron
Science	0.02M	Datasets from the cauldron
Chart	3.17M	MMC-Inst [Liu et al., 2024a], Datasets from the cauldron
Knowledge	0.52M	WIT [Srinivasan et al., 2021], In-house Taiwan-knowledge caption dataset
Total	**10.89M**	

By leveraging these diverse datasets, our pre-training process equips the Breeze 2 model with the ability to handle a wide range of visual tasks, ensuring robust performance across different applications.

4.4 Training Details of Visual Alignment Pre-training

The visual alignment pre-training process consists of two sequential phases to enhance training efficiency and performance based on our experimental results:

1. **Phase 1: MLP Phase**: The MLP projector was trained from scratch for one epoch. The AdamW optimizer was employed with a base learning rate of 1×10^{-5}. The learning rate schedule consisted of a 10% linear warm-up period, followed by a constant learning rate. This phase utilized a batch size of 1024.
2. **Phase 2: Full-weight Phase**: Subsequent to the MLP phase, full-weight training was conducted for 2 epochs. The AdamW optimizer was again utilized, this time with a base learning rate of 1×10^{-6}. The learning rate schedule mirrored that of the MLP phase, incorporating a 10

5 Post-training

Breeze 2 employs a tuning-based methodology to achieve both instruction-following and function-calling capabilities. This approach involves fine-tuning pre-trained base models using specifically designed prompt templates, as detailed in Section 5.1. We have curated three distinct post-training datasets: one for instruction following (Section 5.2), one for visual instruction following (Section 5.3), and one for function calling (Section 5.4). We applied supervised fine-tuning (SFT) to the base model using a combination of these three post-training datasets. The details of the post-training are shown below (Section 5.5)

5.1 Prompt Template Design

For consistency, we primarily adopted the prompt template from Llama 3, which is based on the Chat Markup Language (ChatML), a widely adopted format introduced by OpenAI. We then made slight adjustments to fit the needs for enabling vision-aware and function-calling capabilities. The key differences are as follows:

- **Image placeholder.** We introduce the special tokens `<|start_img|>`, `<|img|>`, and `<|end_img|>`. A sequence of `<|img|>` tokens represents the patches of a provided image, while the `<|start_img|>` and `<|end_img|>` tokens enclose this sequence to indicate the position of the image placeholder.
- **Bounding box placeholder.** We introduce the special tokens `<|start_bbox|>` and `<|end_bbox|>`. These tokens enclose the list of coordinates of the bounding box. The coordinates are represented by four values: the x-coordinate of the top-left point, the y-coordinate of the top-left point, the x-coordinate of the bottom-right point, and the y-coordinate of the bottom-right point. All coordinates are normalized between 0 and 1000 and represented as integers.
- **Providing function descriptions.** Based on the investigation in [Chen et al., 2024b], functions provided in a dedicated role and the system role exhibit similar capabilities in terms of function-calling accuracy, but relevance detection is superior when functions are provided in the dedicated role. For consistency with Llama 3, we finally chose to provide function descriptions in the system role and compromised on relevance detection by introducing the Decision Token proposed by [Chen et al., 2024b]. To provide the function descriptions for enabling function calling, we prepend the list of the function description following the "Customized Functions:".
- **Decision Token.** We introduce the special tokens `<|use_tool|>` and `<|answer|>` as the Decision Token [Chen et al., 2024b]. The Decision Token is the first generated token when functions are given, determining the action of models. If the model chooses to provide a direct answer, it outputs `<|answer|>` first; if it chooses function calling, it outputs `<|use_tool|>` first. The Decision Token also facilitates the creation of non-function-call data from function-called data, significantly enhancing relevance detection (see Section 5.4 for details).
- **Supporting function calls and responses in parallel.** Llama 3 supports only a single function call and response. We extend this to support function calls and responses in parallel. The list of function calls is provided following the special token `<|python_tag|>`. The list of function responses is put in the `ipython` role, with the order item by item corresponding to function calls.

More details are provided in the GitHub repository[5]. To facilitate developers' access to Breeze 2, we also provide a Python package[6] to handle the complex conversion from conversation turns to the conditional prompt given to the model.

5.2 Post-training Data for Instruction Following

We employ SlimOrca [Lian et al., 2023, Mukherjee et al., 2023, Longpre et al., 2023] for post-training fine-tuning. SlimOrca, a refined subset of OpenOrca, incorporates an additional filtering mechanism

[5] `https://github.com/mtkresearch/mtkresearch/blob/main/doc/format_prompt_v3.md`
[6] `https://github.com/mtkresearch/mtkresearch`

to exclude potentially mislabeled data points. Similar to Orca, SlimOrca comprises an augmented FLAN dataset that emphasizes solving questions through reasoning chains. The diverse system prompts within the corpus enhance the model's attentiveness to descriptions. Our training utilizes the entire dataset, which includes 500k data points in English. Preliminary experiments indicate that the fine-tuned model, extensively trained on English instances, demonstrates strong performance in Traditional Chinese. We attribute this performance to the model's knowledge transfer ability, which is enhanced by intensive pre-training on Traditional Chinese.

5.3 Post-training Data for Visual Instruction Following

To enhance the model's ability to follow visual instructions, we utilized a comprehensive set of datasets for post-training. This includes all datasets from the 2-phase pre-training, each with a selected portion to ensure a balanced and diverse training set.

In addition to the pre-training datasets, we incorporated several high-quality chat datasets to further improve the model's chat capabilities. These additional datasets include:

- LLaVA [Liu et al., 2023]
- LLaVAR-2 [Zhou et al., 2024]
- High-quality chat datasets from Canbrain-1 [Tong et al., 2024]

To specifically enhance the model's ability to handle image-based Chinese chat, we also included an in-house Taiwan-knowledge chat dataset. This comprehensive post-training dataset comprises a total of 6M data points, ensuring that the model is well-equipped to handle a wide range of visual instruction tasks.

By leveraging these diverse and high-quality datasets, our post-training process significantly boosts the model's performance in visual instruction following, making it more robust and versatile in real-world applications.

5.4 Post-training Data for Function Calling

To enable function calling capabilities, we curated high-quality function-calling datasets, including:

- `FC-110k`: This dataset comprises 110k high-quality instances, filtered from a combination of APIGen Liu et al. [2024b] and the glaive-function-calling-v2 dataset[7].
- `FC-TC-19k`: We generated 19k function-calling instances in Traditional Chinese using synthetic translation methods proposed by Chen et al. [2024b]. Instances were selected from `FC-110k` while striving to maintain a balanced distribution of function types among the called functions.
- `FC-NF-10k`: Using synthetic methods proposed by Chen et al. [2024b], we generated 10k instances of non-function-call data from `FC-110k` and `FC-TC-19k`. The ratio between English and Traditional Chinese instances is 9:1. Specifically, to generate non-function-call data, consider an example with three functions: `func_A`, `func_B`, and `func_C`. If `func_A` is helpful and called in the original data, we create non-function-call data by removing `func_A` and only using `func_B` and `func_C`. This ensures that function calling is not triggered.

5.5 Training Details of Post-training

We employed the AdamW optimizer with a learning rate of 1×10^{-6}. The learning rate was warmed up during the first 10% of the training steps. Gradient clipping was applied with a maximum norm of 1. The model was trained for a total of 4 epochs with a batch size of 1024.

6 Benchmarks and Results

Breeze 2 underwent comprehensive evaluation across numerous established benchmarks. These assessments cover a wide spectrum of tasks, including genera knowledge, long context capabilities, function calling, and visual comprehension.

[7] `https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2`

6.1 Taiwan General Knowledge and Instruction-Following Benchmarks

For general knowledge of Taiwan, we tested on TMMLU+ [Tam and Pai, 2023, Hsu et al., 2023]. The dataset composed of 67 subjects covering disciplines from STEM to humanities, and includes professional domain knowledge as well. Additionally, to measure chatting abilities, we utilize MT-Bench-tw, which is a translated version of MT-Bench [Zheng et al., 2024] with detailed manual rewriting. MT-Bench-tw along with MT-Bench highlights multi-turn dialogue capabilities of the language model.

Table 4 demonstrates the evaluation results.

Table 4: Comparison of different models across various benchmarks. The judge model on MT-Bench-tw and MT-Bench is GPT-4o.

	TMMLU+	MMLU	MT-Bench-tw	MT-Bench
Breeze 7B	42.7	61.7	5.43	6.49
Breeze-2 3B	43.2	59.9	4.33	5.59
Breeze-2 8B	46.4	66.6	5.64	6.81

6.2 Long Context Benchmarks

We use the synthetic passkey retrieval task to evaluate long context capabilities of the model. The passkey retrieval task is designed to assess the model's ability to accurately retrieve and utilize a passkey or specific information embedded within lengthy irrelevant text. This task is particularly relevant in scenarios where the model needs to process and understand large volumes of text to extract critical information.

To thoroughly evaluate the model's performance across different parts of the 128k context, we divide the context into 16 bins, each covering an 8k token range. Bin 0 represents the first 8k tokens (0-8k), while Bin 15 represents the last 8k tokens (120k-128k). The passkey is placed at different positions within these bins to test the model's retrieval capability throughout the entire context length. The results are listed in Table 5.

Evaluation results show the model's awareness of information within the long context at different relative positions throughout the 128k range.

Table 5: Comparison of Breeze-2 3B and 8B models on passkey retrieval task across all 16 bins (0-128k tokens): Both models show strong performance, with 8B maintaining perfect accuracy in later bins

Model	Retrieval accuracy-128k (%)							
	Bin 0 (0-8k)	Bin 1 (8-16k)	Bin 2 (16-24k)	Bin 3 (24-32k)	Bin 4 (32-40k)	Bin 5 (40-48k)	Bin 6 (48-56k)	Bin 7 (56-64k)
Breeze-2 3B	100	100	100	100	100	100	95	95
Breeze-2 8B	100	100	95	100	95	100	100	100

Model	Retrieval accuracy-128k (%)							
	Bin 8 (64-72k)	Bin 9 (72-80k)	Bin 10 (80-88k)	Bin 11 (88-96k)	Bin 12 (96-104k)	Bin 13 (104-112k)	Bin 14 (112-120k)	Bin 15 (120-128k)
Breeze-2 3B	95	95	100	100	100	100	95	90
Breeze-2 8B	100	100	100	100	100	100	100	100

6.3 Function Calling Benchmarks

To evaluate the performance of function calling, we utilized the Berkeley Function Calling Leaderboard (BFCL) [Yan et al., 2024] for English contexts. Additionally, we employed the Function Calling Leaderboard for ZHTW [Lee et al., 2024], a translated version of the BFCL, to assess performance in Traditional Chinese contexts. The evaluation metrics are described below.

AST Accuracy (%): This metric evaluates the structural correctness of language model outputs for function-calling tasks by comparing the Abstract Syntax Tree (AST) representations of generated and target function calls. It includes four problem types—Simple Function, Multiple Function, Parallel

Function, and Parallel Multiple Function—categorized based on the combination of the number of provided functions and function calls.

Executable Accuracy (%): This metric assesses performance by executing the generated APIs and comparing the results with the expected outcomes. It also includes the same four problem types as AST Accuracy.

Relevance Detection (%): This metric measures the success rate of not generating any function call when none of the provided functions are relevant. This scenario helps determine whether a model will hallucinate its functions and parameters when the provided functions are irrelevant to the user's query.

Table 6 and Table 7 present the evaluation results.

Table 6: Results of BFCL in English contexts. S., M., P., and P.M. represent Simple Function, Multiple Functions, Parallel Functions, and Parallel Multiple Functions, respectively. The version of GPT-4o-mini is 2024-07-18.

	Overall Accuracy	**AST Accuracy**				**Executable Accuracy**				**Relevance Detection**
		S.	M.	P.	P.M.	S.	M.	P.	P.M.	
Breeze-FC 7B	87	90	93	84	84	100	92	88	78	76
Gorilla-v2 7B	86	94	96	87	86	97	96	80	75	60
GPT-4o-mini	?	91	93	90	84	?	?	?	?	77
Breeze-2 3B	82	91	93	83	77	93	92	80	75	59
Breeze-2 8B	87	91	93	89	79	97	92	84	63	80

Table 7: Results of the Function Calling Leaderboard for ZHTW in Traditional Chinese contexts. S., M., P., and P.M. represent Simple Function, Multiple Functions, Parallel Functions, and Parallel Multiple Functions, respectively. The version of GPT-4o-mini is 2024-07-18.

	Overall Accuracy	**AST Accuracy**				**Executable Accuracy**				**Relevance Detection**
		S.	M.	P.	P.M.	S.	M.	P.	P.M.	
Breeze-FC 7B	78	82	86	77	67	88	88	80	60	73
Gorilla-v2 7B	76	85	87	73	68	92	92	62	73	54
GPT-4o-mini	81	86	89	86	78	88	86	70	58	67
Breeze-2 3B	75	83	84	79	63	88	80	76	65	53
Breeze-2 8B	80	86	87	86	69	89	84	76	60	69

6.4 Vision Understanding Benchmarks

To evaluate the vision capability of our models, we have tested them on MMMU [Yue et al., 2024] and our proposed TMMBench[8]. MMMU is a vision-language benchmark that contains questions from college-level exams and textbooks. It includes diverse image types and focuses on complex reasoning tasks that require specialized knowledge. To evaluate the comprehension of traditional Chinese and Taiwan-specific knowledge, we have developed an additional benchmark named TMMBench. TMMBench consists of traditional Chinese questions covering topics related to Taiwan, such as Taiwanese attractions, daily life, tables or diagrams containing traditional Chinese, and questions from Taiwan's university entrance exams. We plan to release TMMBench in the near future. The evaluation results are shown in Table 8.

Table 8: Comparison of different models across various vision-language benchmarks

	MMMU	TMMBench
GPT-4o-mini	59.4	49.3
Breeze-2 3B	35.8	38.3
Breeze-2 8B	44.0	45.9

[8] https://github.com/mtkresearch/TMMBench

7 Demonstration

7.1 Case Study: Text-to-text Instruction Following

To demonstrate the enhanced capabilities of Breeze 2 3B in handling Taiwan-specific contexts and following complex instructions, we conducted two comparative case studies. The first focuses on knowledge about Taiwanese night markets, while the second examines the model's ability to generate more formal and polite text based on given instructions.

7.1.1 Case Study 1: Taiwanese Night Markets

We presented both models with the following query in Traditional Chinese:

> 請改寫以下短文，使其具備關於台灣夜市的具體細節，例如著名的夜市名稱。
> 台灣的夜市文化是台灣人生活中不可或缺的一部分。夜市不僅提供各種美食，還有各式各樣的商品和娛樂活動。每當夜幕降臨，夜市便開始熱鬧起來，吸引了大量的遊客和當地居民。夜市裡的攤販們熱情地招呼顧客，讓人感受到濃厚的人情味。夜市裡的美食總能滿足不同人的口味。夜市不僅是購物和品嚐美食的好地方，也是體驗台灣文化的絕佳場所。
>
> *English translation:*
> *Please rewrite the following short passage to include specific details about Taiwanese night markets, such as the names of famous night markets.*
> *Night market culture is an integral part of Taiwanese life. Night markets not only offer a variety of delicious food but also various goods and entertainment activities. As night falls, these markets come alive, attracting numerous tourists and locals alike. The vendors in the night markets enthusiastically greet customers, creating a warm and friendly atmosphere. The food at night markets can satisfy different tastes. Night markets are not only great places for shopping and tasting delicious food but also excellent venues for experiencing Taiwanese culture.*

The responses from both models were then analyzed for accuracy and richness of Taiwan-specific information. Below are the original responses from both models:

> **Breeze 2 3B:** 台灣著名的夜市包括士林夜市、饒河街夜市、寧夏夜市、通化街夜市、逢甲夜市、六合夜市、南機場夜市、公館夜市、瑞豐夜市、花園夜市、羅東夜市、基隆廟口夜市等。這些夜市各有特色，吸引了不同的人潮。夜市不僅是台灣人生活中不可或缺的一部分，也是吸引遊客和當地居民的重要景點。每當夜幕降臨，夜市便開始熱鬧起來，吸引了大量的遊客和當地居民。夜市裡的攤販們熱情地招呼顧客，讓人感受到濃厚的人情味。夜市裡的美食總能滿足不同人的口味。夜市不僅是購物和品嚐美食的好地方，也是體驗台灣文化的絕佳場所。
>
> **Llama 3.2 3B Instruct:** 台灣的夜市文化是台灣人生活中不可或缺的一部分。夜市不僅提供各種美食，還有各式各樣的商品和娛樂活動。每當夜幕降臨，夜市便開始熱鬧起來，吸引了大量的遊客和當地居民。夜市裡的攤販們熱情地招呼顧客，讓人感受到濃厚的人情味。如，台北的夜市如士林夜市、電信夜市、世貿夜市等，都是人氣的熱門夜市之一。夜市裡的美食總能滿足不同人的口味，從海鮮到甜點，還有各種特殊的夜市特產。夜市不僅是購物和品嚐美食的好地方，也是體驗台灣文化的絕佳場所。

The responses reveal a significant difference in the depth and accuracy of knowledge about Taiwanese night markets between the two models. Breeze 2 3B demonstrated a more comprehensive and accurate understanding of the topic. It correctly named twelve well-known night markets across Taiwan, including Shilin Night Market (士林夜市), Ningxia Night Market (寧夏夜市), Fengjia Night Market (逢甲夜市), Ruifeng Night Market (瑞豐夜市) and others. These night markets span multiple cities and regions in Taiwan, showcasing Breeze 2 3B's broad understanding of the country's night market culture.

In contrast, Llama 3.2 3B Instruct only correctly identified one night market, Shilin Night Market (士林夜市). Moreover, it generated two non-existent night markets, "Telecom Night Market" (電信夜市) and "World Trade Night Market" (世貿夜市), which do not exist in reality.

7.1.2 Case Study 2: Formalizing Email Content

For this case study, we presented both models with the following task in Traditional Chinese:

> 使以下郵件的風格更加正式，並使用敬語。
> 嗨，John，
> 我想跟你確認一下我們下週的會議時間。你方便的話，能不能告訴我你什麼時候有空？另外，我們需要討論一下新項目的進度。
> 謝謝，Tom
>
> *English translation:*
> *Make the following email more formal and use honorific language.*
> *Hi John,*
> *I want to confirm our meeting time for next week. If it's convenient for you, could you tell me when you're available? Also, we need to discuss the progress of the new project.*
> *Thanks, Tom*

The responses from both models were then analyzed for formality, politeness, and overall improvement in language. Below are the original responses from both models:

> **Breeze 2 3B:**
> 親愛的John，
> 我想跟您確認一下我們下週的會議時間。您方便的話，能否告訴我您什麼時候有空？
> 另外，我們需要討論一下新項目的進度。
> 謝謝，Tom
>
> **Llama 3.2 3B Instruct:**
> 嗨，John，
> 我想跟您確認一下我們下週的會議時間。您是否有空的話，能否告訴我您什麼時候有空？
> 另外，我們還需要討論一下新項目的進度。望您能夠提供相關的資訊。
> 謝謝，Tom

The responses reveal a notable difference in the quality of formalization and use of honorific language between the two models. Breeze 2 3B demonstrated a more natural and appropriate transformation of the original text:

1. It changed the greeting from "嗨" (Hi) to "親愛的" (Dear), which is more formal in Traditional Chinese business correspondence.
2. It consistently used the honorific "您" instead of "你" throughout the email.
3. The sentence structure was improved for clarity and formality without unnecessary additions.

In contrast, Llama 3.2 3B Instruct made some improvements but also introduced issues:

1. It retained the informal greeting"嗨" (Hi), which is inconsistent with the request for a more formal tone.
2. While it did use the honorific "您", it created a redundant and awkward sentence: "您是否有空的話，能否告訴我您什麼時候有空？" (If you are available, could you tell me when you are available?), which is both repetitive and illogical.
3. It added an unnecessary sentence at the end: "望您能夠提供相關的資訊。" (Hope you can provide relevant information.), which wasn't part of the original message and doesn't significantly enhance formality.

7.1.3 Analysis and Implications

These case studies underscore the effectiveness of Breeze 2 3B's specialized pre-training in capturing nuanced, location-specific knowledge and handling complex language tasks in Traditional Chinese. In the first case study, the model demonstrated a comprehensive understanding of Taiwanese night markets, accurately naming multiple locations across different cities. This showcases its enhanced capability in handling queries that require deep understanding of local cultural contexts.

In the second case study, Breeze 2 3B exhibited superior ability in refining language for formal communication. The model demonstrated a better understanding of formal language conventions and was able to make appropriate adjustments without altering the core message or introducing redundancies.

These case studies highlight Breeze 2 3B's advanced capabilities in handling Taiwan-specific contexts and complex language tasks in Traditional Chinese. The model's ability to provide accurate, contextually appropriate responses demonstrates its potential for enhancing a wide range of applications that require nuanced understanding of local culture and language use.

7.2 Case Study: Visual Instruction Following

To demonstrate the visual instruction-following capability of Breeze 2 8B, we provide an example below.

Figure 3: Image for demo.

We presented the model with the following visual query in Traditional Chinese using Figure 3:

> <image>
> 請問前三名共可獲得多少錢？
>
> *English translation:*
> *<image>*
> *How much money will the top three get in total?*

The response shown below is correct:

> **Breeze 2 8B:**
> 第一名可獲得30萬元，第二名可獲得20萬元，第三名可獲得15萬元。前三名共計65萬元。

7.3 Mobile APP

Motivation and Features of the App. To showcase the capabilities of Breeze 2 3B in a practical, user-friendly format and demonstrate its potential for real-world applications, we developed a mobile application tailored for Taiwan-localized knowledge and multi-modal interactions. The app, powered by the Breeze 2 3B model, excels in understanding Taiwan's cultural, historical, and geographical contexts in Traditional Chinese. With vision-aware capabilities, the app recognizes and interprets visual content such as landmarks, cultural artifacts, and signage, providing detailed explanations and insights. By supporting multi-modal inputs, users can combine text and images, enabling comprehensive descriptions, translations, and historical information, enhancing the experience for both local and international users.

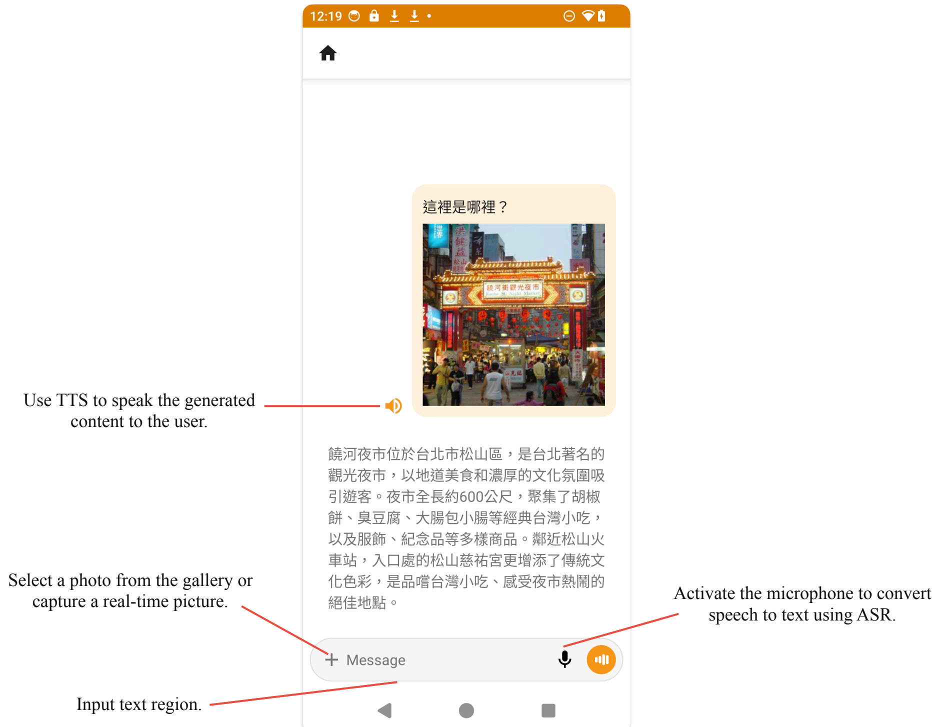

Figure 4: Overview of the App Interface Design. This figure illustrates the main user interaction elements in the application. Users can select a photo from the gallery or capture a real-time picture to initiate visual processing. A dedicated input text region enables users to provide textual commands. Additionally, the application integrates a microphone for automatic speech recognition (ASR) to convert speech into text, allowing for hands-free interaction. The system also supports text-to-speech (TTS) functionality, which vocalizes generated content back to the user, ensuring accessibility and an enhanced user experience.

Implementation. The Breeze 2 3B model, which integrates both language and vision capabilities, is integrated into the mobile application using ExecuTorch [9], a tool designed for preparing and deploying large-scale models on Android devices. For the purposes of this implementation, we utilize two configurations of the Breeze 2 3B model: a language-only version (referred to as LLM in this context) based on the Llama 3.2 3B architecture, and the full vision-language integrated version (referred to as VLM). ExecuTorch converts these models into .pte files and tokenizers into .bin files, optimized for on-device inference. The ExecuTorch API, accessed via a JNI layer, supports tasks such as model initialization, parameter adjustments, and inferencing with delegates like XNNPACK (CPU-based) and MediaTek AI Accelerators.

Developed in Android Studio [10], the app handles both LLM and VLM functionalities, enabling users to load models, adjust parameters, and input system or user prompts. For vision-aware tasks, users can upload images or take live pictures, combining them with textual prompts for multi-modal interactions.

The app has been tested with selected configurations, focusing on the language-only and full vision-language integrated versions of the Breeze 2 3B model. Initial results demonstrate that ExecuTorch

[9] https://pytorch.org/executorch/stable/index.html
[10] https://developer.android.com/studio?hl=zh-tw

effectively supports multi-modal interactions, confirming its suitability for deploying advanced AI models on Android platforms.

Performance. The Android application was developed and tested on a smartphone equipped with the MediaTek Dimensity 9400 chipset. Although this platform is equipped with a high-performance AI accelerator (NPU), we opted to use the CPU-based XNNPACK delegate for inference. This choice ensures a fair comparison of performance metrics across devices, as XNNPACK is a widely supported and standardized delegate for on-device AI computations.

Key performance metrics, including model loading time, memory usage, prefill rate (tokens per second), and text decoder rate, are summarized in Table 9 for the language-only configuration of Breeze 2 3B.

Table 9: Performance Metrics on Android Device with XNNPACK delegate for Breeze 2 3B (language-only configuration)

Performance Metric	Value
Model Loading Time (s)	2.48
Memory Usage (GB)	6.87
Prefill Rate (Tokens/s)	17.07
Text Decoder Rate (Tokens/s)	3.83

Note: The performance metrics for the full vision-language integrated version of Breeze 2 3B are currently being evaluated and will be reported in future updates.

Limitations. While the app demonstrates significant strengths, it does face some limitations. Despite the use of quantization, the 3B model remains resource-intensive, posing challenges for older devices with limited RAM and storage. Vision-aware functionalities, such as processing large image inputs or performing real-time visual analysis, can experience latency on mid-range devices, which may affect the responsiveness of the application. The app's memory requirements are relatively high, which could impact multitasking capabilities, especially on devices with constrained resources.

8 Conclusion and Future Work

In this work, we introduced Breeze 2, a suite of multi-modal language models designed to address the underrepresentation of Traditional Chinese in large language models. By building upon the knowledge base of Llama 3 and incorporating extensive pre-training on Traditional Chinese corpora, we have developed models that excel in understanding and generating content relevant to Taiwan's linguistic and cultural context.

Our comprehensive multi-stage training approach, which includes extended text-to-text pre-training, vision-alignment pre-training, and post-training for instruction tuning and function calling, has resulted in models that demonstrate strong performance across a wide range of tasks. The evaluation results on various benchmarks highlight Breeze 2's capabilities in general knowledge, instruction following, long context handling, function calling, and vision understanding, particularly in Traditional Chinese contexts.

Key achievements of Breeze 2 include:

- Improved performance on Taiwan-specific knowledge tasks, as evidenced by the results on TMMLU+ and MT-Bench-tw.
- Preservation of long-context capabilities inherited from Llama 3, with both 3B and 8B models showing high accuracy in passkey retrieval tasks across 128k token contexts.
- Competitive function calling abilities in both English and Traditional Chinese, comparable to specialized models like Breeze-FC and Gorilla-v2.
- Promising results in vision-language tasks, particularly on the Taiwan-specific TMMBench.

While Breeze 2 represents an advancement in addressing the needs of Traditional Chinese language processing and Taiwan-specific applications, there are several avenues for future work:

1. **Further enhancement of vision-language capabilities:** While Breeze 2 shows promising results on vision-language tasks, there is room for improvement, especially when compared to larger models like GPT-4o-mini. Future work could focus on refining the vision-alignment training process and expanding the visual instruction tuning dataset.
2. **Optimization for mobile deployment:** Given the successful implementation of Breeze 2 3B in a mobile application, future work could focus on further optimizing the model for mobile environments, potentially exploring techniques like quantization and model compression to improve efficiency without sacrificing performance.
3. **Exploration of larger model sizes:** While the current 3B and 8B models show strong performance, investigating the potential benefits of larger model sizes for Traditional Chinese processing could yield interesting insights and potentially even better results.

In conclusion, Breeze 2 addresses the underrepresentation of Traditional Chinese in LLMs and incorporates vision-aware and function-calling capabilities. By focusing on Taiwan-specific applications, Breeze 2 contributes to the development of more inclusive and culturally aware AI applications. As we continue to refine and expand upon this work, we anticipate Breeze 2 will play a role in bridging the gap between advanced AI technologies and the unique linguistic and cultural needs of Traditional Chinese-speaking communities.

References

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

Yu-Ying Chang. Taitravel, 2024. URL `https://huggingface.co/datasets/benchang1110/TaiVision-TaiTravel`.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. URL `https://arxiv.org/abs/2311.12793`.

Yi-Chang Chen, Po-Chun Hsu, Chan-Jan Hsu, and Da-shan Shiu. Enhancing function-calling capabilities in llms: Strategies for prompt formats, data integration, and multilingual translation. *arXiv preprint arXiv:2412.01130*, 2024b. URL `https://arxiv.org/abs/2412.01130`.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024c. URL `https://arxiv.org/abs/2404.16821`.

Chee-Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, Chee Seng Chan, and Lianwen Jin. Icdar2019 robust reading challenge on arbitrary-shaped text (rrc-art), 2019. URL `https://arxiv.org/abs/1909.07145`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747/`.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL `https://aclanthology.org/2021.emnlp-main.98/`.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da shan Shiu. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite, 2023.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. Breeze-7b technical report. *arXiv preprint arXiv:2403.02712*, 2024. URL `https://arxiv.org/abs/2403.02712`.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022. URL `https://arxiv.org/abs/2111.15664`.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The stack: 3 tb of permissively licensed source code. *Preprint*, 2022.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. NeMo: a toolkit for building AI applications using neural modules, 2019. URL `https://arxiv.org/abs/1909.09577`.

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

Liang-Chieh Lee, Cheng-Wei Lin, Pei-Chen Ho, Chien-Yu Yu, Yi-Chang Chen, and Da-Shan Shiu. Function calling leaderboard for zhtw, 2024. URL `https://github.com/mtkresearch/function-calling-leaderboard-for-zhtw`.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023. URL `https://https://huggingface.co/Open-Orca/SlimOrca`.

Cheng-Wei Lin, Wan-Hsuan Hsieh, Kai-Xin Guan, Chan-Jan Hsu, Chia-Chen Kuo, Chuan-Lin Lai, Chung-Wei Chung, Ming-Jen Wang, and Da-Shan Shiu. Fineweb-zhtw: Scalable curation of traditional chinese text data from the web, 2024. URL `https://arxiv.org/abs/2411.16387`.

Yen-Ting Lin and Yun-Nung Chen. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*, 2023.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning, 2024a. URL `https://arxiv.org/abs/2311.10774`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf`.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, et al. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *arXiv preprint arXiv:2406.18518*, 2024b.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL `https://arxiv.org/abs/2306.14824`.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL `https://arxiv.org/abs/2210.08402`.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463257. URL `https://doi.org/10.1145/3404835.3463257`.

Zhi-Rui Tam and Ya-Ting Pai. An improved traditional chinese evaluation suite for foundation model. *arXiv*, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL `https://arxiv.org/abs/2406.16860`.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images, 2016. URL `https://arxiv.org/abs/1601.07140`.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. `https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html`, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Shijie Zhou, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. A high-quality text-rich image instruction tuning dataset via hybrid instruction generation, 2024. URL `https://arxiv.org/abs/2412.16364`.

Author Contributions

We list all contributors in alphabetical order of their first names, as follows: Chan-Jan Hsu (許湛然), Chia-Sheng Liu (劉佳昇), Meng-Hsi Chen (陳孟羲), Muxi Chen (陳沐希), Po-Chun Hsu (許博竣), Yi-Chang Chen (陳宜昌), and the supervisor Da-Shan Shiu (許大山).

Po-Chun Hsu and Yi-Chang Chen are the project leads of this work. Chan-Jan Hsu and Po-Chun Hsu contributed to the textual pre-training data. Po-Chun Hsu and Yi-Chang Chen contributed to the pre-training and post-training recipes. Yi-Chang Chen and Muxi Chen contributed to the experiments of function calling cabaility. Chia-Sheng Liu, Meng-Hsi Chen, and Yi-Chang Chen contributed to the experiments of vision-aware capability. Muxi Chen contributed to the mobile app.