Amharic LLaMA and LLaVA: Multimodal LLMs for Low Resource Languages

Michael Andersland*
mike.andersland@gmail.com

Abstract

Large Language Models (LLMs) like GPT-4 and LLaMA have shown incredible proficiency at natural language processing tasks and have even begun to excel at tasks across other modalities such as vision and audio. Despite their success, LLMs often struggle to perform well on low-resource languages because there is so little training data available. This shortcoming is especially prevalent with open source models. In this work, we explore training LLaMA-2 to speak Amharic, a language which is spoken by over 50 million people world wide, but has orders of magnitude less data available than languages like English. We employ methods previously used for training LLMs on other languages with data scarcity, and use open source translation models to perform data augmentation and grow our dataset from millions of tokens to billions. We further enhance the capabilities of our model by connecting an image encoder and training on a translated visual instruction tuning dataset in the same manner as LLaVA, resulting in a multimodal Amharic LLM that can understand images along with text. We introduce an Amharic version of a popular benchmarking dataset to evaluate our work. Our models and dataset are open sourced and available on GitHub[1].

1 Introduction

The field of natural language processing (NLP) has seen a massive transformation in recent years, spurred primarily by the development of the transformer architecture (Vaswani et al., 2023) and its subsequent application to language modeling via large self supervised neural networks, often with tens or hundreds of billions of parameters trained on trillions of tokens. Recent prominent models include the GPT series (Brown et al., 2020), PaLM(Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023), among others. While state of the art performance is still generally achieved only by proprietary models as of late 2023, the release of LLaMA has given way to a vibrant open source community that has produced near-SOTA open models like Mistral (Jiang et al., 2023) and Mixtral (Jiang et al., 2024). Though LLaMA can only process text inputs, recent projects have augmented it with multimodal understanding for images and video by aligning pretrained vision or audio encoders with LLaMA(Zhang et al., 2023).

One of the most valuable aspects of these models is their ability to perform few-shot or zero-shot adaptation to novel tasks and instructions without the need for additional training. This is enabled by the massive scale of model parameters and training data. For languages like English, there is an abundance of public data on the internet, enough to provide trillions of tokens for pretraining. For low resource languages that lack high quantities of data, most LLMs fall short and either fail to perform tasks at a high level, or cannot understand the language at all.

In this work, we focus on developing an open source multimodal language model that can perform NLP tasks in Amharic and understand images. Amharic, the official language of

*Work performed at Garri Logistics
[1]https://github.com/iocuydi/amharic-llama-llava

Ethiopia, is a semitic language spoken by over 50 million people worldwide. However, it is vanishingly rare on the public internet. Several Amharic NLP datasets exist, but do not typically exceed 1 million tokens, and are often focused on specific tasks (Tonja et al., 2023). Less than 0.1% of CommonCrawl is Amharic, and even when combining open source datasets without deduplication, we find that less than 500 million tokens of Amharic are available. In addition, the content of this data tends to be biased towards a small set of topics like news and politics.

Previous work on language modelling for less represented languages has shown promising results on Chinese (Cui et al., 2023) and Finnish (Luukkonen et al., 2023), with varied approaches including extended pretraining of open source LLMs, development of new pretrained models from scratch, and tokenizer extension. The datasets used to train the models for Chinese and Finnish had tens and hundreds of billions of tokens, respectively. We base our work on the approach used for Chinese Llama and Alpaca, and continue the pretraining of LLaMA-2 with an extended tokenizer.

Even compared to other lower resource languages, the amount of Amharic data available is especially small, with orders of magnitude less data than the prior work. To increase the size and diversity of our data, we apply machine translation to create billions of diverse synthetic Amharic tokens from English text in the RedPajama dataset (Computer, 2023). Until recently, open source translation models were not accurate for Amharic, but work such as Seamless M4T (Barrault et al., 2023) has yielded results approaching and in some cases exceeding the accuracy of proprietary translation models.

Data generation via translation has been effective for translation tasks (Sawai et al., 2021), and we apply this technique with the expectation that the quality of Amharic spoken by the model will not exceed the translation quality, but can still offer an improvement over the small dataset for general NLP and image understanding tasks.

After pretraining on this augmented dataset, we train a small MLP projection (Liu et al., 2023a) to connect a CLIP encoder (Radford et al., 2021) to our pretrained network, and then apply supervised fine tuning on Amharic instruction tuning data obtained by translating English instruction tuning datasets in a similar manner. Instruction tuning includes pure text conversation pairs as well as visual instruction data (Liu et al., 2023b) containing image features from CLIP. We evaluate our models with Amharic-MMLU, our Amharic version of the popular LLM benchmark dataset MMLU (Hendrycks et al., 2021). We apply translation again to create Amharic-MMLU from the standard English MMLU.

2 Models

LLaMA-2 is an open source foundational language model that has rivaled the performance of similar proprietary models. LLaVA is an open source multimodal model that adds a CLIP vision encoder to LLaMA and trains end to end in order to align image encodings with LLaMA and enable visual understanding and reasoning. LLaMA-2 has variants with 7B, 13B, and 70B parameters. Chat variants tuned for multi turn dialogue settings are available for each model size.

Due to limited resources we use the 7B standard (not chat-tuned) variant for all experiments. As in the previous work, we extend LLaMA-2 pretraining for one epoch, during which the model is trained via next token prediction on unstructured Amharic text. We align a CLIP encoder with the pretrained model by training a small MLP mapping between the two with a translated image captioning dataset. This is followed by one epoch of fine tuning on our multimodal Amharic instruction dataset.

The LLaMA tokenizer is poorly suited to Amharic data. Because Amharic text is so rare relative to other languages in public text, Amharic characters do not have dedicated tokens as more common languages might. For example in English, a single word might map to a single token. For the Ge'ez characters that make up Amharic (and other rare characters) the LLaMA tokenzier and others solve this issue by representing them with a combination of multiple generic byte tokens. In some cases this means that a single Amharic word may be encoded to 10+ tokens, while an equivalent English word might only require a single token.

Table 1: Tokenization Comparison (excluding start and stop tokens)

Tokenizer	**Text**	**Tokenized Text**
LLaMA	Hi, how are you?	6324, 29892, 920, 526, 366
Amharic LLaMA	Hi, how are you?	6324, 29892, 920, 526, 366
LLaMA	**ሰላም፣ እንዴት ነህ?**	29871, 228, 139, 179, 228, 139, 142, 228, 139, 160, 228, 144, 166, 29871, 228, 141, 168, 228, 141, 152, 228, 142, 183, 228, 140, 184, 29871, 228, 141, 147, 228, 139, 136
Amharic LLaMA	**ሰላም፣ እንዴት ነህ?**	46702, 32547, 35199

In addition to impeding learning, this tokenization scheme increases the sequence length for any text being processed, which is problematic for the scaling of the transformer with respect to sequence length, and greatly reduces the effective context window and processing speed (this effect can even be observed on proprietary models, which often stream responses more slowly when queried with rare characters). Following the approach of Chinese LLaMA, we use SentencePiece (Kudo and Richardson, 2018) to learn a tokenization scheme for Amharic from our public (untranslated) Amharic data, resulting in an Amharic token vocabulary of 19008 tokens. We combine this with the LLaMA tokenizer of 32000 tokens for a new vocabulary size of 51008. We leave the original embeddings unchanged, though they can update during training when English tokens occasionally appear in the Amharic data. Table 1 illustrates the improved tokenization with the new vocabulary.

3 Data

For our pretraining task, we use a combined dataset consisting of 436 million tokens from public sources including CommonCrawl, Azime and Mohammed (2021) and various web scrapes, along with an additional 3.348 billion Amharic tokens translated from the RedPajama dataset, specifically the sections containing text from Wikipedia and from various books. Table 2 contains the exact proportions.

Table 2: Composition of our Amharic pretraining dataset

Data Source	**Percentage**	**Tokens**
Translated RedPajama Wikipedia	48%	1826m
Translated RedPajama Books	40%	1522m
Real Amharic Text	12%	436m

To translate the English data, we apply the Seamless M4T large model with text to text translation, specifying English to Amharic. For the image encoder alignment and visual instruction tuning, we use Seamless M4T to translate the BLIP caption dataset (Li et al., 2022) and the LLaVA visual instruction tuning dataset.

The pure text instruction pairs in our fine tuning task include Amharic translated versions of the Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), and OpenAssistant (Köpf et al., 2023) datasets. We use the Google Translate API to translate Alpaca and Dolly, and Seamless M4T to translate OpenAssistant. We prune the OpenAssistant conversation trees to ensure that only highly rated responses are used in our dataset. In addition to these Amharic datasets, we create mixed English and Amharic datasets to help the model leverage existing knowledge of English tokens to learn more about the new Amharic tokens and their relation to concepts that the model may already understand well in English. We replace either the human or AI role in an instruction pair with the original untranslated data, and add a specification to the prompt indicating the language in which the AI role is expected to answer. We further augment this data with a translation task in which either the human or AI role of a translated instruction pair has its English and Amharic versions inserted into a new synthetic instruction pair specifying a translation.

Our Seamless M4T translations are performed on an A100 GPU over several weeks. Because Seamless M4T performance can suffer with long sequences, the text is translated in chunks of a few sentences at a time, not exceeding a fixed token limit, with very long sentences excluded entirely. To speed up translation we use batch inference and map sentences into different sized batches depending on length before rearranging them in the original order after translation.

4 Experiments

We follow the experimental setup and hyperparamter configurations used in the 7b parameter model subset of the Chinese LLaMA experiments. For visual instruction tuning, we follow the experimental setup used in LLaVA-1.5. When combining visual and text-only instruction tuning, we use the LLaVA-1.5 setup.

We use LoRA (Hu et al., 2021) for the attention layers but do not train with quantization. We train each model for one epoch on a single A100 GPU, which takes 1-4 weeks for pretraining and 2-7 days for finetuning. Pretraining is more computationally expensive, and with our limited resources we choose to perform most of our experiments at the finetuning stage. We perform one pretraining run with 436m tokens gathered from public sources, and another with our augmented 3784m token dataset including mostly synthetic translated data. We use the 3784m model as a base for most of our finetuning experiments.

We finetune with different versions of our dataset to explore the effects of including English data and the translation task versus pure Amharic data. We compare tuning on just Amharic data, English data followed by Amharic data, and all the data together.

We compare the effect of visual instruction tuning on pure text tasks, and explore the order in which the model is exposed to visual data relative to other training phases. We finetune with visual and text-only instruction data and compare this to omitting text-only data.

5 Results and Evaluation

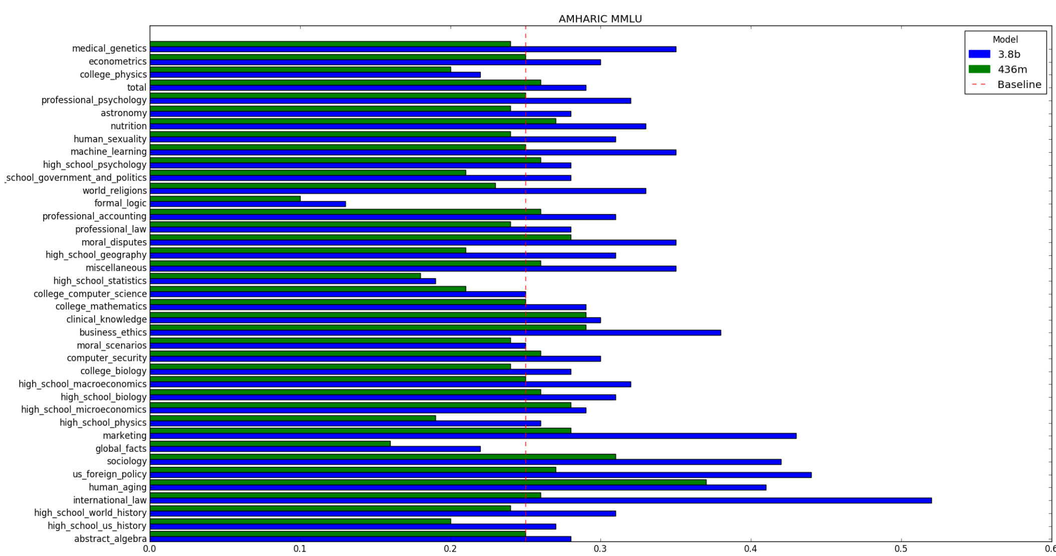

Figure 1: Amharic MMLU Subject Performance, 436m vs 3784m token dataset

Quantitative evaluation is difficult for low resources languages. As with training data, well measured benchmarks and even baselines against which to benchmark are scarce. We use SeamlessM4T again to create a rudimentary Amharic version of the widely used English MMLU dataset for language task evaluation. We test the text understanding and world knowledge of our models with MMLU by asking multiple choice questions. Across most topics, the variants pretrained with the augmented dataset outperform those using the smaller datasets, with some notable exceptions. Both models fail to even exceed a baseline of a random guess on several STEM topics like math, logic, and physics, as shown in Figure 1. We suspect that this may be due to the nature of the questions, where the mistranslation of a single character could completely alter the meaning of the question and the answer. In contrast, the models tend to significantly outperform the baseline on topics like law and ethics. Table 3 compares models across all Amharic MMLU topics, and again with STEM topics excluded. Qualitatively, our models perform well at a variety of tasks

Table 3: Amharic MMLU Scores with and without STEM topics. Sequential refers to splitting the full text dataset and training on the mixed English-Amharic data before any pure Amharic data. Full text refers to the entire finetuning dataset except the multimodal visual examples.

Pretrain	**Finetune**	**Amharic MMLU Score**	**Amharic MMLU Non-STEM Score**
436m	Full Text	0.26	0.30
3784m	Amharic Only	0.27	0.32
3784m	Sequential	0.28	0.35
3784m	Full Text	0.29	0.37
3784m	Full Text + Visual	**0.30**	**0.38**
3784m	Visual Only	0.26	0.27

including text generation, summarization and named entity recognition, as well as visual question answering and analysis of multiple pictures. Figures 2 and 3 contain examples. Though both visual models were able to provide coherent answers about images, they excel in different areas. The model that saw only visual instruction data and no pure text data tends to give better responses to multimodal queries, but falls short on pure text queries. The model that saw the full extent of visual and text data outperforms all pure text variants and has the best overall performance on Amharic MMLU, but offers less detailed responses to multimodal queries and is more likely to make mistakes about details in images. Both visual models can respond to queries containing multiple images as shown in Figure 4, but performance deteriorates rapidly as the number of images increases, making the models unable to understand video. Further qualitative examples, comparisons to GPT-4 (OpenAI et al., 2023) and failure cases are included in the appendix.

Figure 2: Example output for a writing task. An approximate translation in English is included for both the prompt and response.

Prompt	**Response**
ከኢትዮጵያ ወደ አሜሪካ ለሚሄድ ሰው የቱሪስት ቪዛ ማመልከቻ ደብዳቤ ይጻፉ	**ውድ ቪዛ ኦፊሰር፣ ይህ ደብዳቤ በደንብ እንደሚያገኝህ ተስፋ አደርጋለሁ። በቅርቡ ወደ አሜሪካ ለመጓዝ እና ለሦስት ወራት ያህል ለመቆየት እጓጓለሁ። ለጉዞዬ ቪዛ ለማውጣት ወደ ኤምባሲው ማመልከቻዬን ለማቅረብ እጽፍልሃለሁ። ለጉዞዬ ወጪዎች ለመሸፈን እና ራሴን ለመደገፍ ወደ አሜሪካ በምሄድበት ወቅት ለመስራት የሚያስችል በቂ ገንዘብ እና ንብረት እንዳለኝ ማረጋገጥ እፈልጋለሁ። በተጨማሪም በጉዞዬ ወቅት ምንም አይነት የወንጀል ክስ፣ የጤና ችግር ወይም ተላላፊ በሽታ እንደሌለኝ አረጋግጣለሁ። በጉዞዎቼ ወቅት፣ የአሜሪካን ህግ እና ስርዓት በማክበር እና የአሜሪካን ማህበረሰብ ወክዬ እንድወክል ተመኘሁ። ማመልከቻዬን ስላጤንከኝ አመሰግናለሁ።**
Write a tourist visa application letter for someone visiting the USA from Ethiopia	Dear Visa Officer, I hope this letter finds you well. I am looking forward to traveling to America soon and staying for three months. I am writing to you to submit my application to the embassy to issue a visa for my trip. I want to make sure I have enough money and assets to work while I'm in America to cover my travel expenses and support myself. I also certify that I have no criminal charges, health problems or communicable diseases during my trip. During my travels, I aspire to uphold American law and order and represent the American community. Thank you for considering my application.

Figure 3: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Prompt	Response
በዚህ ምስል ውስጥ ምን እየሆነ ነው?	በሥዕሉ ላይ አንድ የሰዎች ቡድን በትልቅ ሰማያዊ የጭነት መኪና ጀርባ ላይ ቆሞ ትልቅ ከረጢት ወይም ከረጢት እየጫነ ነው።የጭነት መኪናው አልጋው ላይ በርካታ ከረጢቶችያሉት ሲሆን አንዳንዶቹም በሰዎች እየተጫኑ ነው።የጭነት መኪናው እና የሰዎች ቡድን ከቤት ውጭ ያለ ይመስላል ፣ ምናልባትም እንደ የግንባታ ቦታ ወይም ከቤት ውጭ ክስተት ባሉ ቦታዎች ላይ።የጭነት መኪናው እና የሰዎች ቡድን መኖራቸው የጭነት መኪናው ትልቅ ጭነት ለማጓጓዝ ጥቅም ላይ እየዋለ መሆኑን ያመለክታል
What is happening in this picture?	In the picture, a group of people is standing on the back of a big blue truck loading a large bag or sack. The truck has several bags on the bed, some of which are being loaded by people. The truck and group of people appear to be outdoors, perhaps at a construction site or outdoor event. The presence of a truck and a group of people indicates that the truck is being used to transport large loads

Figure 4: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Prompt	Response
ሁለቱ ሥዕሎች ምን የሚያመሳስላቸው ነገር አለ? በመካከላቸው ያለው ልዩነት ምንድን ነው?	በሁለቱ ሥዕሎች መካከል ያለው የጋራ ነገር ሁለቱም እንስሳት ፒያኖንእየተጫወቱ መሆናቸውነው። ይሁን እንጂ በሁለቱ ሥዕሎች መካከል ያለውልዩነት ፒያኖው ላይ ያለው እንስሳ ነው።አንድ ሥዕል አንድ ፈረስ ፒያኖውንሲጫወት የሚያሳይ ሲሆን ሌላኛውደግሞ የዶሮ ፒያኖውን ሲጫወት የሚያሳይነው። ። ሁለቱም እንስሳት ፒያኖውን ለመጫወት ሲሞክሩ የሚያሳዩት ትዕይንትአስደሳችና አዝናኝ ነው።
What do the two pictures have in common? What is different between them?	The common denominator between the two pictures is that both animals are playing the piano. However, the difference between the two pictures is the animal on the piano. One picture shows a horse playing the piano and the other shows a chicken playing the piano. The scene of both animals trying to play the piano is hilarious.

6 Conclusion

We have explored solutions for improving pretrained language models with an understanding of a low resource language. We have presented data augmentation via translation as a way to rapidly grow the available dataset by at least an order of magnitude while significantly increasing the diversity, yielding improved performance on text and visual instruction tasks after additional fine tuning. We applied a similar methodology to create a baseline benchmark and measure the performance of our models.

6.1 Limitations

These models have limitations that should be considered when extending or deploying them.

- They may exhibit bias or generate harmful content.
- The models are brittle and can hallucinate or deliver unpredictable outputs, especially with high temperature and sampling.
- Our finetuning data mostly consists of single turn pairs, and performance degrades in conversational chat settings as the dialogue grows.
- Further evaluation and tuning with highly accurate real world Amharic datasets is needed before production deployments.
- Though their understanding of Amharic characters has improved, the models lack cultural knowledge that might be expected to come along with understanding of Amharic, as the majority of the Amharic they saw was synthetic data from translated English.

6.2 Future Work

The area of applying large models to low resource languages is ripe for exploration and stands to benefit many populations that cannot make effective use of today's models. Future projects could include:

- Creating high quality human-informed Amharic datasets for fine tuning and evaluation, which can significantly exceed the quality of the translated datasets and may offer outsized gains when applied during the fine tuning step, and more comprehensive and accurate quantitative evaluation.
- Performing additional tuning steps like Reinforement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) to improve performance and alignment .
- Experimenting with LLaMA-2-Chat model variants, other model sizes, quantization, and new architectures like Mixtral.
- Applying these techniques to other poorly represented languages on which today's models struggle.
- Expanding the experiments done in this paper with larger synthetic datasets to learn about how this approach scales.
- Adding Amharic OCR capabilities to the model by procedurally generating images of noisy Amharic text, finetuning CLIP if necessary, and augmenting the visual instruction tuning dataset with LLM-generated instruction sequences that reference the images.
- Improving the efficiency of our translation batching algorithm and the cleaning and management of translated data including removal of poorly translated sentences and sentences translated from unexpected source languages.
- Experimenting with Retrieval Augmented Generation (RAG) (Lewis et al., 2021) and Amharic document embeddings.
- Adding video or audio understanding to the model .

6.3 Acknowledgements

Thanks to Daniel Temesgen for insightful discussions on the Amharic language and the challenges of language modelling and data preparation for low resource languages. Thanks to the Google for Startups and Google Cloud teams for assisting with consistent access to A100 GPUs and cloud credits.

References

Israel Abebe Azime and Nebil Mohammed. An amharic news text classification dataset, 2021.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t: Massively multilingual multimodal machine translation, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

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

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. Fingpt: Large generative models for a small language, 2023.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Ranto Sawai, Incheon Paik, and Ayato Kuwana. Sentence augmentation for language translation using gpt-2. *Electronics*, 10(24), 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. Natural language processing in ethiopian languages: Current state, challenges, and opportunities, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude

Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023.

A Appendix

We present additional qualitative examples of model performance with image and text instructions. Figure 5 includes an intentionally confusing scene where the model partially understands but confuses the act of ironing with washing clothes. Both this and Figure 6 include comparisons to GPT-4's Amharic visual reasoning capabilities via the GPT-4-vision-preview API. Amharic LLaVA generally demonstrates improved Amharic visual reasoning over GPT-4, which is impeded by a poor understanding of Amharic. GPT-4's Amharic has improved rapidly since its release, and will likely continue to get better.

Figure 7 shows a longer text generation instruction in which the model provides a mostly valid answer with some erroneous repetition on its suggested third step of the response. Figure 8 shows an example of model shortcomings on Amharic text. The model infers some general meaning of the text from the context, but fails to accurately decipher the text. This is likely due to the rarity of visual Amharic text data in the CLIP training set as well as the complete absence of Amharic text from the images in our visual finetuning data, but may be easily solved with a procedurally generated synthetic Amharic OCR dataset.

Figure 9 shows the model unsuccessfully attempting to interpret a video presented as a series of frames stitched together. The model understands the frames in isolation but is unable to identify the activity performed across them.

Figure 5: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Model	Prompt	Response
Amharic-LLaVA	**በዚህ ምስል ላይ ያልተለመደው ምንድን ነው?**	**በዚህ ምስል ውስጥ ያልተለመደ ገጽታ አንድ ሰው ቢጫ ታክሲ ጀርባ ላይ የልብስ ማጠቢያ መስመርን በመጠቀም የልብስ ማጠቢያውን ማጠብ ነው። ይህ የተለመደ እይታ አይደለም ፣ ምክንያቱም ሰዎች በተለምዶ የልብስ ማጠቢያ መስመሩን በቤት ውስጥ ወይም በንግድ ተቋማት ውስጥ ይጠቀማሉ። ሰውየው በልብስ ማጠቢያ መስመሩ ላይ ተንጠልጥሎ መገኘቱ ትዕይንቱን አስደሳችና ያልተለመደ ያደርገዋል**
	What is unusual about this picture?	An unusual aspect in this picture is that a man is washing his clothes using a clothesline on the back of a yellow taxi. This is not a common sight, because people usually use the clothesline at home or in business establishments. The fact that the man is hanging on the clothesline makes the scene interesting and unusual.
GPT-4 (2023)	**በዚህ ምስል ላይ ያልተለመደው ምንድን ነው?**	**በምስሉ ላይ የታየውን በመሆኑ ሊመረጥ አይችልም እንጂ እንደ አዲስ ያልታወቀ አስቤ ነገር የሚታየው፣ አንድ ሰው መኪና ተሸክመው ትንሹ የመጥበሻው ማዕከላዊ ክፈፍ ላይ ማንጠባጠብ አገናኝ ነው። ይህ የመንገድ ላይ እንደሚከናወን ወይም የሞክር ሥራ ባለመሆኑ እና ሰውች ለመሳቀፍ ይችላል። የተሳሳተ መስሪያ መጠቀም፣ ከመንገድ ላይመኪና አፍነጫ መከታተል፣ ወይም አነስተኛውን ናዳ ለመቁመር የመጠቀም ስምንት ሊታይ ይችላል። በእውነት ምስሉ አነስተኛውን መንገድ ፋና እስከነው አንድ ብልኅ ሥራ ወይም የአስገራሚ የተንኮል ስጦታ መሆኑ እና የታዘዘውን ሃላፊነት መቻል ሊኖር ይችላል።**
	What is unusual about this picture?	It can't be selected because of what's shown in the picture, but what appears to be a new unknown object, a person carrying a car, is a link dripping onto the central frame of the pan. This is done on the street or because it's not a test job and people can get yelled at. Using the wrong device, tracking the car off the road, or using the smallest number to count eight can be seen. Indeed, as long as the picture is taken in the least possible way, it is possible that it is a clever work or a wonderful gift of cunning, and that there can be a responsibility that is commanded.

Figure 6: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Model	Prompt	Response
Amharic-LLaVA	**በዚህ ሥዕል ላይ ምን እየሆነ ነው?**	**በሥዕሉ ላይ አንድ የሰዎች ቡድን በቀለማት ያሸበረቁ ባንዲራዎችን ይዘው እየተራመዱ ሲሆን አንዳንዶቹም ጃንጥላ ይዘው ነው። በሕዝቡ መካከል አንድ ሕፃን ወይም አንድ ትንሽ ልጅ እየተራመደ ሲሆን ምናልባትም ከቤተሰቡ ጋር እየተጓዘ ወይም እየተጓዘ ሊሆን ይችላል። ሕዝቡ ባንዲራዎችን እና ጃንጥላዎችን በመጠቀም የበዓል ወይም ሁኔታን ይፈጥራል**
	What is happening in this picture?	In the picture, a group of people are walking with colorful flags, some of them carrying umbrellas. In the middle of the crowd is walking a baby or a small child, possibly walking or traveling with his family. The people use flags and umbrellas to create a festive atmosphere
GPT-4 (2023)	**በዚህ ሥዕል ላይ ምን እየሆነ ነው?**	**በተመለከተው ሥዕል ላይ እንደሚታይበት ሰዎች ከባንዲራቸው ባለቤቶችና ከብርሃን መከላከያ አሞሮች ጋር መጐምጓም ይኖራቸዋል። እርስ በርስ በሚያጋጠሙ ከባንዲራቸው ባለቤቶች አኘዶ ኢትዮጵያውያን ወንዶች ናቸው። ብዙም ከእነሱ አብራ ከየይ ቢልጥና መከላከያ ትግጠኛን መዝጊያ አድርገው ተይዘው እንደሚሄዱ። ሰዎቹ ባለቤቶች የተመነጨ ልብስ የለበሱ የሚታይ ነበር ይበተንም። ሥዕሉ በኢትዮጵያ ውስጥ የተወሰደ እንደ መሆኑ አመጣጥ ይችላል፤ የሚታዩት የኢትዮጵያ ባንዲራን ወንበሪነት ያላቸውና አንዳንድ የተለያዩ አሞሮች በእጁ የያዙ ሰዎችን ማንበብ እንደሚቻል አይታይበታል። ትልቅ ባለቤቶችና ባለውነት የሚገልጹ ከብርሃን መከላከያ አሞሮች መጠቀም የሚታይም ነው፤ እስከ ዛሬም ድረስ በማኅበረሰብ ላይ ከወንበሩ ልዩ እንቅስቃሴ እንዳለበት ሊመስል አለበት።**
	What is happening in this picture?	As you can see in the picture above, people will have to go with their flag holders and light shields. They are Ethiopian men who are the owners of their flags facing each other. That they will be caught in a tight barrier if they are not far away from them. The owners of the people were seen wearing original clothes. It can be said that the picture was taken in Ethiopia. It is not seen that it is possible to read people who are seen with the Ethiopian flag as a chair and some different amours in their hands. The use of light protection amours is seen by large owners and those who express their wealth. To this day, it must seem that he has a special influence on society from his chair.

Figure 7: Example output for a writing task. An approximate translation in English is included for both the prompt and response.

Prompt	**Response**
እባክዎን የፒዛ ምግብ ቤት ለመክፈት 7 ምክሮችን ይስጡኝ	**1. ትክክለኛውን ቦታ ይምረጡ፡ የፒዛ ምግብ ቤት በጣም ጥሩ ቦታ ማግኘት አስፈላጊ ነው ምክንያቱም ደንበኞችዎ በቀላሉ ማግኘት እና ወደ ሬስቶራንቱ መምጣት አለባቸው።** **2. ጥሩ ድባብ ይፍጠሩ፡ የፒዛ ምግብ ቤት ድባብ ሞቅ ያለ እና እንግዳ ተቀባይ መሆን አለበት ይህም ደንበኞችን እንዲዝናኑ እና እንዲመገቡ ያደርጋል።** **3. ከፍተኛ ጥራት ያላቸውን ንጥረ ነገሮች ተጠቀም፡- የፒዛ ምግብ ቤት ደንበኞችን ለማስደሰት ከፍተኛ ጥራት ያላቸውን ንጥረነገሮች መጠቀም አስፈላጊ ነው። ትኩስ, ትኩስ እና ትኩስ የሆኑ ንጥረ ነገሮችን ይጠቀሙ እና ትኩስ, ትኩስ እና ትኩስ የሆኑ ንጥረ ነገሮችን ይጠቀሙ.** **4. ጥሩ የፒዛ ሼፍ ቀጥል፡- ፒዛ ሼፍ ደንበኞችን ለማስደሰት እና ከፍተኛ ጥራት ያላቸውን ምግቦች ለማቅረብ እውቀት ያለው እና ልምድ ያለው መሆን አለበት።** **5. ጥሩ አገልግሎት ያቅርቡ፡- የፒዛ ምግብ ቤት ሰራተኞች ለደንበኞችዎ ወዳጃዊ እና ቀልጣፋ አገልግሎት መስጠት አለባቸው።** **6. ጥሩ ዋጋዎችን ያቅርቡ፡- የፒዛ ምግብ ቤት ደንበኞችን ለመሳብ እና ለማቆየት ተወዳዳሪ ዋጋዎችን መስጠት አስፈላጊ ነው።** **7. ጥሩ ግብይት እና ማስተዋወቅ፡ የፒዛ ምግብ ቤት ደንበኞችን ለመሳብ እና ለማቆየት ጥሩ የግብይት እና የማስተዋወቂያ ዘመቻዎችን ይጠቀሙ። ይህ ማህበራዊ ሚዲያን፣ የመስመር ላይ ማስታወቂያዎችን እና የሀገር ውስጥ ዝግጅቶችን ሊያካትት ይችላል።**
Please give me 7 tips to open up a pizza restaurant	1. Choose the right location: Finding the best location for a pizza restaurant is important because your customers need to find it easily and come to the restaurant. 2. Create a good atmosphere: The atmosphere of a pizza restaurant should be warm and welcoming, which will make customers feel relaxed and fed. 3. Use high quality ingredients: In order to please the customers of a pizza restaurant, it is important to use high quality ingredients. Use fresh, fresh and fresh ingredients and use fresh, fresh and fresh ingredients. 4. Resume of a good pizza chef: A pizza chef must be knowledgeable and experienced to delight customers and serve high-quality dishes. 5. Provide good service: Pizza restaurant staff should provide friendly and efficient service to your customers. 6. Offer good prices: To attract and retain pizza restaurant customers, it is important to offer competitive prices. 7. Good Marketing and Promotion: Use good marketing and promotional campaigns to attract and retain pizza restaurant customers. This may include social media, online advertising and local events.

Figure 8: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Prompt	Response
ይህ ሥዕል ምንድን ነው?	**ምስሉ ሁለት ቀይ እና ነጭ የትራፊክ ምልክቶችን ያሳያል ፣ አንዱ ወደ ግራ እና ሌላኛው ወደ ቀኝ። ሁለቱም ምልክቶች "አደጋ" የሚል ቃል አላቸው ፣ ይህም አሽከርካሪዎች ጥንቃቄ እንዲያደርጉ እና ጥንቃቄ እንዲያደርጉ ማስጠንቀቂያ ነው።" ምልክቶቹ በትራፊክ ምልክቶች ላይ የተለመዱ ቀይ እና ነጭ ቀለሞች አሏቸው ፣ ይህም ለአሽከርካሪዎች እና ለእግረኞች በቀላሉ እንዲታዩ ያደርጋቸዋል**
What is this picture?	The image shows two red and white traffic signs, one to the left and the other to the right. Both signs have the word ”DANGER” on them, which is a warning to drivers to be careful and take care. And it makes them easily visible to pedestrians

Figure 9: Example output for an Amharic multimodal query. An approximate translation to English is included for both the prompt and response.

Prompt	Response
በቪዲዮው ውስጥ ምን እየሆነ ነው?	**በቪዲዮው ውስጥ አንድ ሰው በአንድ ክፍል ውስጥ በተለያዩ ቦታዎች ላይ ቆሞ በርካታ ፎቶግራፎችን ሲወስድ ይታያል**
What is happening in the video?	In the video, a person is seen standing in different places in the same room and taking several pictures