

VISION-LANGUAGE MODELS UNDER CULTURAL AND INCLUSIVE CONSIDERATIONS

Anonymous authors

Paper under double-blind review

ABSTRACT

In an era with increasing integration of AI applications into our lives, it is important to consider cultural perspectives when evaluating such systems. In image captioning, the automated process of generating textual descriptions for images, current evaluation datasets might include implicit cultural images without providing culture-related captions. To address this problem, we create a small culture-centric evaluation benchmark by filtering VizWiz, an existing dataset with images taken by people who are blind or visually impaired. We then evaluate different vision-language models, investigating their current cultural knowledge. Lastly, we include a case study from integrating such models as a digital visual assistant for people seeking sighted support. While the evaluation results for state-of-the-art models seem promising, we identified some weak spots: data limitations, hallucinations, and problems with conventional evaluation metrics. Both our data, code, and model outputs will be publicly available and we hope to extend our culture-specific dataset with more images, captions, and tasks.

1 INTRODUCTION

Culture is an evolving concept generally described as the shared beliefs, values, customs, traditions, language, and general way of life of a collective group of people (Ren, 2012). It encompasses elements that shape the way individuals within a society think and interact (Birukou et al., 2013). As AI is fundamentally transforming how we perceive and interact with the world, it is essential to ensure that these advancements are culturally representative and inclusive (Hershcovich et al., 2022). AI applications are becoming more and more accessible to people. An example for this is ‘Be My Eyes’, a software used to support blind and visually impaired individuals seeking sighted support by connecting them with volunteers through live video, which recently introduced ‘Be My Eyes Virtual Volunteer’, a visual AI assistant powered by OpenAI’s GPT-4V vision-language model (OpenAI, 2023).

Despite considerable efforts to evaluate vision-language (VL) models such as GPT-4V from diverse perspectives, the extent of their cultural awareness remains unexplored (Cao et al., 2024). Existing evaluation benchmarks for VL models focus primarily on English with few, implicit cross-cultural references. While there exist multilingual evaluation datasets such as MARVL (Liu et al., 2021) and XM3600 (Thapliyal et al., 2022) that aim to have a cultural component to their images (e.g., a traditional wedding costume), these datasets tend to also include a large proportion of images with little cultural significance (e.g., a bag of carrots). As such, these datasets may not representatively measure the cultural knowledge of VL models, despite being useful for evaluating their multilingual capabilities.

To address these challenges, we first filter an existing English image captioning dataset identifying implicit cross-cultural concepts. We then use this filtered dataset as a starting point of a benchmark to evaluate image captioning performance on cultural images of state-of-the-art VL models, such as BLIP-2 (Li et al., 2023a), Gemini Pro Vision (Google, 2023), and GPT-4 Vision. We evaluate the same models also on an existing multicultural dataset to further investigate their current cultural knowledge. Lastly, we perform a case study with GPT-4 Vision as a virtual assistant integrated into Be My Eyes. With these experiments we investigate how AI applications, such as image captioning, can incorporate a global culture, fostering a more inclusive and culture-aware experience for all individuals.

2 BACKGROUND

Current vision-language models are trained without consideration for the subjective perspectives and cultural influences of those who provided the image descriptions (Ye et al., 2023), necessitating carefully curated sources of data and annotation paradigms that are more culturally aware and inclusive (Arora et al., 2023). In line with this need, several vision-language datasets have recently been introduced. Liu et al. (2021) introduced MaRVL, a visiolinguistic reasoning dataset across five languages and cultures (Indonesian, Mandarin Chinese, Swahili, Tamil, and Turkish). Thapliyal et al. (2022) created Crossmodal-3600 (XM3600), a geographically diverse set of 3600 images annotated with human-generated reference captions in 36 languages. Liu et al. (2023a) built a cross-cultural dataset of textual prompts to generate images in Chinese cultural style with GPT-4V. Similarly, Liu et al. (2023b) introduced a cross-cultural understanding benchmark for text-to-image synthesis representing eight distinct cultural backgrounds. Lastly, Ventura et al. (2023) proposed prompt templates and evaluation techniques across multiple languages to unlock and assess cultural knowledge within text-to-image generation. While the overall trend is promising, these existing cross-cultural datasets appear to focus on the cross-lingual aspect, and few of them can be used for image captioning. Moreover, the existing multicultural datasets that can be used for image captioning (Liu et al., 2021; Thapliyal et al., 2022) primarily focus on the multilingualism of the captions rather than the cultural concepts depicted in the images. In particular, both datasets include images depicting general objects in the real world that lack inherent cultural significance, e.g. carrots, spoons, trousers, tables, etc. Cao et al. (2024) were the first to perform cultural concept filtering of the MaRVL dataset by asking native Chinese speakers to distinguish culturally relevant and neutral images. They also found that GPT-4V provides more culturally relevant captions than MaRVL’s original human annotations.

3 EXPERIMENTS

We focus on two lines of experiments: (1) We filter the VizWiz dataset for implicit cultural concepts. VizWiz is a widely used visual question answering and image captioning dataset representing a real-world use case, where examples consist of images and questions submitted by people who are blind, together with crowdsourced answers and image captions (Gurari et al., 2020). The selection of this dataset serves two main purposes. Firstly, it is a challenging dataset specifically tailored to real-world challenges faced by people seeking to access visual information. Secondly, VizWiz might contain implicit cultural references that are currently not captured due to the lack of culture-specific captions. (2) We evaluate the cross-cultural performance of state-of-the-art close-sourced and open-sourced models in image captioning, using our filtered VizWiz dataset and XM-3600 (Thapliyal et al., 2022). We selected the latter dataset since it includes 3600 images annotated with human-generated captions in 36 languages, offering geographical and cultural diversity. We compare the annotated captions and the model-generated captions using both the COCO evaluation package¹ and human evaluation.

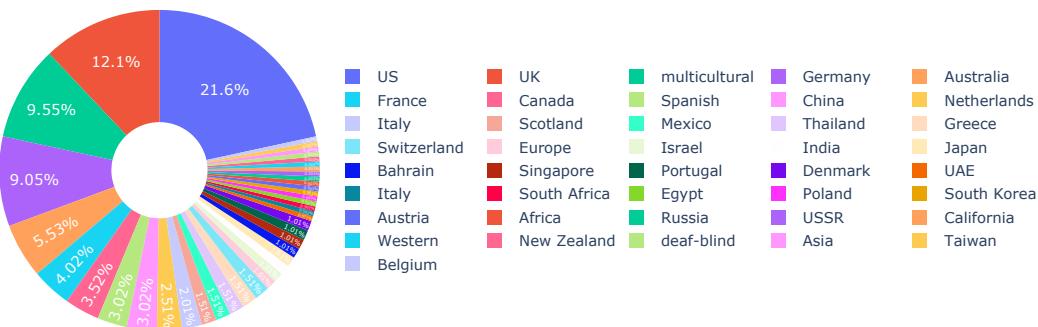


Figure 1: Distribution of the cultural concepts identified in the VizWiz dataset by the annotators.

¹<https://github.com/tylin/coco-caption>

3.1 DATA FILTERING

To filter the data we hired annotators through the Prolific platform². We asked them to identify images from the Vizwiz dataset visualizer³ related to their cultural background, provide the image name, the reason they think the image is culture-related, and their preferred caption from the dataset (VizWiz provides five different image captions per image). We also gave them the option to suggest a better caption that encompasses cultural aspects. After four rounds of surveys, we collected all the culture-specific candidate images and proceeded to a second step of verification. In this step, we retained only those images that had received consensus agreement from at least two individuals. We collected a total of 200 images spanning 43 different identified cultures. We present the distribution of the different cultures identified by the annotators in Figure 1. It should also be noted that the annotators suggested a cultural revision of the original captions in all 200 images. For further information about our data filtering approach and results please refer to the Appendix.

3.2 MODELS AND EVALUATION

We conduct experiments on the image captioning task in the zero-shot setting, in which a pretrained model is queried to produce a textual description for an image without having been exposed to images from the same dataset. We rely on two commonly used open-access models:⁴ BLIP-2 6.7B with OPT (Zhang et al., 2022) as LLM backbone and finetuned on the COCO Captions dataset (Li et al., 2023b), and InstructBLIP 7B with Vicuna (Chiang et al., 2023) as LLM backbone (Dai et al., 2023); and two state-of-the-art closed-access models: GPT-4(Vision) (OpenAI, 2023), and Gemini Pro Vision (Google, 2023). For all of these models, we experimented with three different prompt types including a continuation prompt and an interrogative prompt taken from (Dai et al., 2023). We evaluate the model-generated captions in two ways: with the COCO evaluation suite and with human annotators. The COCO evaluation suite was first introduced by (Chen et al., 2015) as a framework to assess image captions using numerous automatic metrics, including BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE (Lin, 2004), and METEOR (Denkowski & Lavie, 2014). Both of our evaluation datasets—culture-filtered VizWiz and XM3600—have two human-annotated caption alternatives that are both taken into account in these metrics. For the human evaluation, we ask annotators to select their preferred caption (human-generated, model-generated, or both) and determine if a caption is relevant, coherent, and accurate (on a binary scale) given the corresponding image. We did not make the annotators aware that one caption was model-generated to minimize bias. For further details on the human evaluation please refer to the Appendix.

It should also be noted that the evaluation is performed by comparing an English model-generated caption with the English reference captions. The main reason we do not evaluate the performance of VL models in generating multilingual captions lies in the motivation behind this study which is the integration of English-speaking VL models as virtual assistants.

4 RESULTS

The results after evaluating the human reference and the model-generated captions using the COCO evaluation suite are presented in Table 1. As mentioned above, the results are the average scores across three different model prompts, and the different cultures/languages.

We find that BLIP-2 outperforms InstructBLIP on both datasets, which is likely a result of the BLIP-2 checkpoint having been finetuned for image captioning on COCO Captions. Gemini Pro Vision and GPT-4V exhibit comparable performance. Their performance is much higher than that of the BLIP models on VizWiz. Since Vizwiz is naturally noisy due to the high ratio of low-quality, blurry images, the increased scale and overall multimodal reasoning capabilities of the closed-source models appear to give a significant advantage. On XM3600, however, BLIP-2 outperforms Gemini, which may be a result of the XM3600 annotations being initialized from an automatic captioning model trained on COCO Captions (the same dataset BLIP-2 was finetuned on), thus having a similar captioning style.

²<https://www.prolific.com/>

³https://vizwiz.cs.colorado.edu/VizWiz_visualization/view_dataset.php

⁴ Accessible without any cost on the platform provided by Hugging Face.

	VizWiz			XM3600		
	B@4	Cider	ROUGE-L	B@4	Cider	ROUGE-L
InstructBLIP	3.30	28.22	20.10	3.74	30.80	24.25
BLIP-2	4.36	35.70	22.80	5.87	46.42	29.82
Gemini-PV	7.82	85.10	65.99	3.43	30.14	24.74
GPT-4V	6.54	67.33	59.42			

Table 1: Results from evaluating different models on generating image captions for three multicultural datasets.

The results of the human evaluation are shown in Figure 2. In line with the automatic metrics, our human annotators tend to prefer the captions produced by closed-access models, Gemini Pro Vision and GPT-4V, over those from the BLIP-family models. Moreover, the former are rated as highly accurate, while the latter are deemed inaccurate in more than half of the cases. Nevertheless, our preference comparison shows that the human-annotated captions are still preferred over all of the models.

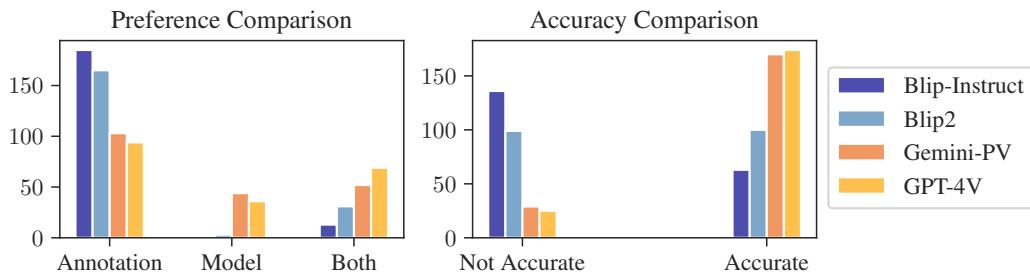


Figure 2: Results of the human evaluation of the model-captions for the filter VizWiz dataset. The right plot presents the preference evaluation (i.e. if the annotator preferred the original annotated caption, the model-generated caption, or both). The left plot presents the accuracy evaluation (i.e. how many times the model responses were accurate).

Overall, we see promising results after evaluating VL models in zero-shot settings for generating captions about images from culturally diverse scenarios. We can further verify our results and enrich our insights with a comprehensive error analysis of the model-generated captions.

5 ERROR ANALYSIS

Following our evaluation, we manually inspected 30 random model-generated responses for every language of the XM3600 dataset and all 200 generated captions for our filtered VizWiz dataset. It seems that InstructBLIP and BLIP-2 captions are very short, lack a lot of information, and are often irrelevant hallucinations. This is, to an extent, expected as we perform zero-shot captioning, so the models are not necessarily accustomed to the desired captioning style. In this case, few-shot prompting or finetuning the models would improve model performance (Brown et al., 2020; Mañas et al., 2023; Ramos et al., 2023). The closed-access models, in contrast, largely provide further or more useful and culture-specific details about the image than given by the human captioners. They also seem to provide more accurate captions compared to the open-sourced models. These points may explain why Gemini-PV and GPT-4V were overall preferred in our human evaluation. There are, however, cases where the captions by these two models are anecdotally too long⁵ and contain hallucinations. To further examine their capabilities and weaknesses, we provide a more comprehensive error analysis of Gemini-PV.

⁵For example whenever there is an image depicting alcohol GPT-4V generates a caption and then a responsible alcohol consumption advice: *This is a bottle of Dewar's White Label Blended Scotch Whisky. Dewar's is a well-known brand of whisky that originates from Scotland. The "White Label" is one of their standard and popular products. Please note that when discussing alcoholic beverages, responsible drinking and understanding the laws and regulations regarding alcohol consumption in one's region are important [...].*

Some selected examples from the XM3600 dataset⁶ are presented in Table 2. It seems that in the majority of the cases, the models provide accurate captions, but they use synonyms or emphasize different aspects (example 1). Often, the models also provide additional information not mentioned in the original captions (example 2) or the models are more accurate than the original captions (example 3). In some instances, the model gives a caption with information present in the caption of other languages, but not in the English caption, suggesting a possibility that the model has already seen this dataset during pretraining. Lastly, Gemini-PV occasionally produces inaccuracies and hallucinations mostly by mistaking cultural places or figures (example 5 where the image is actually ‘The Wondering Philosopher by George M. Groutas’ and not Zeus). Apart from those examples, we also encountered a few cases where the model outputs a generic message (e.g. *Sorry, I can’t help with identifying or making assumptions about content in images.*). For a detailed performance across cultures and metrics please refer to the Appendix.

Original caption	Model-generated caption
1. Green woods in a thickset forest with sunlight in background.	This is a picture of a coniferous forest.
2. Semla with cream on white plate.	This is a Swedish semla. It is a traditional pastry eaten during Lent[...]
3. A red coral underwater.	It is a flower spike from an amaranth plant. The amaranth plant is [...]
4. Comics pamphlets on a blue background.	The image is a collection of Yu-Gi-Oh! cards.
5. White plaster bust sculpture portrait of a man.	This is a marble statue of the Greek god Zeus.

Table 2: Examples comparing the original and Gemini-PV generated captions from the XM3600 dataset.

For the filtered VizWiz dataset, some example cases are provided in Table 3. We observe that Gemini-PV sometimes identifies cultural information (such as cultural food, and languages) that is not present in the original captions nor the corrected VizWiz captions (example 1: grass jelly, 2: Luxembourgish). There are also cases where useful details are transcribed in an optical character recognition (OCR)-manner, even if the language is not English. For instance, transcribing Greek medication labels, Chinese rice cooker instructions, Russian book covers, etc. (example 3). There are also cases where Gemini-PV provides too little information making the human-annotated caption a better fit for the image (example 4). Furthermore, the model hallucinates and either starts generating more information than necessary with details not present in the image or fails to identify the content of the image resulting in inaccurate captions (example 5). Lastly, there were also a couple of generic messages that failed to provide a caption for the image or requested a less blurry picture(‘To provide more information, I would need a clearer image or additional context’). Those cases, however, were not consistent across prompts, suggesting that the models could provide a caption in some of the runs. Generally, these problematic cases were also the ones where the model’s captions were not preferred in the human preference evaluation shown in Figure 2.

Across both datasets and prompts, we observe that landmarks, foreign currencies, flags, cultural foods, and historical items/people seem to be particularly challenging concepts. For example, all models confused the Italian with the Irish flag and failed to identify the Chinese historical figure of Sun Yat-sen or a picture with a traditional Dutch costume and tulips. There was also some confusion with Swiss and Australian beers, Brazilian and Canadian currencies, and Scandinavian Coffee brands.

Original caption	Model-generated caption
1. A can of Asian drink resting on a counter.	A can of grass jelly, a dessert drink found in China, Taiwan, and Southeast Asia, made from[...]
2. A German book is open with half of the page visible.	A close-up of a book page with three sections, one in Luxembourgish, one in German[...]
3. A cardboard box of Pevison skin cream in Greek.	Pewisone cream with 1% econazole[...]The image OCR is: 1% +0,1% χρέμα νιτρουχή εκοναζόλη[...]
4. Meiji banana choco banana flavor chocolate Japanese snack.	This is a box of Japanese candy.
5. A closeup of a Bahraini one dinar note.	Someone holding a one Canadian dollar bill with a pink ribbon wrapped around it. The person[...]

Table 3: Examples comparing the original and Gemini-PV generated captions from the XM3600 dataset.

⁶Associated images can be found by looking up the original captions in the XM3600 annotation visualizer.

6 DISCUSSION

Given the current integration of VL models as virtual assistants for people who seek sighted support, their cultural performance on image captioning seems promising. Through examples from our error analysis and case studies, we can gain insights and identify some weak spots.

In particular, using the COCO evaluation suite the scores for the models are overall low compared to existing studies evaluating VL models on those datasets (Gurari et al., 2020; Thapliyal et al., 2022). On the other hand, our human evaluation and error analysis show that the generated captions by Gemini-PV and GPT-4V are accurate and preferred in many cases. Apart from the problem of using fixed evaluation metrics, there also seems to be an extended hallucination problem. This is also an existing major challenge not only for Vision Language models (Li et al., 2023c) but across various Language Model applications (Bang et al., 2023; Ji et al., 2023). Lastly, there seems to be a data problem since in both datasets we used there are errors in the annotated captions and a lack of culture-specific information. Thus, a more fine-grained dataset is needed with diverse, and representative multicultural data accounting for the diversity of human cultures.

Through our analysis, we also gained insights into some weak spots regarding what cultures and cultural concepts are not generally recognized by the models. However, since we use a finite amount of data, there might be a data bias in identifying particular cultures or cultural concepts as problematic. Additionally, cultural complexities and variations make it difficult to develop a standardized approach to cultural inclusion in AI. We do, however, hope that our culture-centric approach in the data filtering and annotation process can serve as an initial step towards evaluating and understanding the cultural abilities of vision-language models.

7 CASE STUDY

We illustrate the value of cultural and inclusive VL models via a case study on evaluating a visual assistant integrated into the ‘Be My Eyes’ application powered by OpenAI’s GPT-4V model. In this case study, we took a random sample of 20 images from the MaRVL dataset and 20 cultural-related images from the VizWiz dataset. For the MaRVL dataset, the concepts were related to religion, weddings, food, and everyday life. After manually evaluating the results from the model, we could see that even though GPT-4V’s responses are very detailed and helpful, they sometimes lack cultural background information. For instance, even though the model identified Buddhist and Islamic images, it did not identify Hindu-related figures like Agamas. Moreover, it failed to identify traditional Tamil, Indonesian, and Swahili food but identified Chinese dishes and some of the Turkish ones. In particular, it recognized images of ‘baklava’, but not of doner kebab or ayran. For the VizWiz dataset, the concepts were mostly about item brands, currency, national flags, and cultural food. We found that the model in BeMyEyes could identify almost all of the images relating to the US, Canada, Australia, UK, Germany, France, and China. The model is also very helpful with optical character recognition, providing additional details for food packages, alcohol bottles, and cosmetics. There are, however, problematic cases with Japanese, and Thai. All the results from this case study can be found in the Appendix.

8 CONCLUSION

We evaluate the cultural performance of various VL models on image captioning using two multicultural datasets. Even though the results for the state-of-the-art close-source models for image captioning are promising, there is still room for improvement. Through examples from our error analysis and case studies, we can gain insights into the performance of Vision Language models and identify their weak spots. The integration of fixed evaluation metrics might not be representative of their performance, therefore we encourage researchers to reconsider a more comprehensive assessment framework. We aim to extend our small filtered cultural dataset by including question-answering tasks with POV cultural questions. We will also enhance the cultural and multilingual aspects. We would like to initiate a collaboration with teams from VizWiz and BeMyEyes to curate a cross-cultural dataset encompassing diverse languages and cultures beyond English. Collaborating with diverse communities and stakeholders is crucial for validating evaluation pipelines and ensuring that they capture the nuances of cultural representation.

REFERENCES

- Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 114–130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.12. URL <https://aclanthology.org/2023.c3nlp-1.12>.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Love-nia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.ijcnlp-main.45>.
- Aliaksandr Birukou, Enrico Blanzieri, Paolo Giorgini, and Fausto Giunchiglia. *A Formal Definition of Culture*, pp. 1–26. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-5574-1. doi: 10.1007/978-94-007-5574-1_1. URL https://doi.org/10.1007/978-94-007-5574-1_1.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. Exploring visual culture awareness in gpt-4v: A comprehensive probing, 2024.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=vvoWPYqZJA>.
- Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia (eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL <https://aclanthology.org/W14-3348>.
- Google. Introducing gemini: our largest and most capable ai model. *blog post*, 2023. URL <https://blog.google/technology/ai/googlegemini-ai/>.

- Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 417–434, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58520-4.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL <https://aclanthology.org/2022.acl-long.482>.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, December 2023. URL <https://doi.org/10.1145/3571730>.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML’23. JMLR.org, 2023a.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023b. URL <https://doi.org/10.48550/arXiv.2301.12597>.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL <https://aclanthology.org/2023.emnlp-main.20>.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. On the cultural gap in text-to-image generation, 2023a.
- Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL <https://aclanthology.org/2021.emnlp-main.818>.
- Zhixuan Liu, Youeon Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset, 2023b.
- Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2523–2548, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.185. URL <https://aclanthology.org/2023.eacl-main.185>.
- OpenAI. Gpt-4v(ision) system card. *blog post*, 2023. URL <https://platform.openai.com/docs/guides/vision>.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2840–2849, 2023. doi: 10.1109/CVPR52729.2023.00278.
- Hai Ren. Culture. Oxford Bibliographies in Anthropology, 2012. URL <https://www.oxfordbibliographies.com/view/document/obo-9780199766567/obo-9780199766567-0039.xml>.
- Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 715–729, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.45. URL <https://aclanthology.org/2022.emnlp-main.45>.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.
- Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models, 2023.
- Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. Cultural and linguistic diversity improves visual representations, 2023.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

A APPENDIX

A.1 ETHICAL STATEMENT

The motivation behind this study is that Vision Language models are becoming increasingly available to all individuals, even those who seek sighted support and cannot assess model hallucinations or inaccuracies. The primary purpose of our experiments is to assess the performance of Vision Language models in the task of image captioning using multicultural datasets. To achieve our goal we are in the process of building a culture-specific dataset by filtering existing data resources. However, it is crucial to recognize that results from our current filtered dataset may not be indicative of model performance across cultures. Furthermore, our refined dataset might retain biases present in the original source dataset.

We find it improbable that our experiments and filtered dataset could be beneficial for those intending to create deceptive models for malicious purposes. Additionally, the VizWiz dataset may lack coverage of highly specific subjects, offering only a general overview of factual topics. People who intend to use our resources, however, should state their purpose of usage and be accountable for their own work.

A.2 VIZWIZ DATA FILTERING - HUMAN ANNOTATION

As mentioned in the experimental set-up section, to filter the data we created a survey through the Prolific annotation platform. All annotators were compensated with 18\$ per hour. We ran this survey 4 times asking for 40 participants each time.

We asked people to identify images from the VizWiz dataset based on their cultural background, provide an original and a corrected caption, and specify the reason they selected the image as culture-specific. We grouped the reasons that the annotators provided for selecting culture-specific images in Figure 3. The full annotation guidelines were the following:

Creating datasets that reflect a variety of cultures is a challenging task. This is why we will try to filter an existing dataset. Your task is to find culture-related images from a dataset called VizWiz. You need to:

- Visit the dataset website[link]. - Browse the dataset or use the search bars on the left side of the page and search key-terms related to your culture 'Within visual question', 'Within visual answer' or 'Within captions'. - Try to find an image that is related to your culture/cultural background (i.e. food brand, currency, books, culture-specific locations etc.) - Provide your answers to the 5 following questions.

1. Copy and paste the image name (VizWiz_train_**number**.jpg).
2. Based on your cultural background, specify what culture you think is the image related to.
3. Select a caption for the image from the suggested Image Captions.
4. Do you have a better suggestion for the image caption?
5. Provide a reason as to why the image is culture-specific.

After this, we collected information about the annotators' cultural backgrounds. We asked for both home-country of origin and current country location information since sometimes both can affect our cultural beliefs and practices. The distribution of the annotators counties of origin and location are presented in Figure 4a.

The last step is to answer some final questions about your cultural background, and age. We do not collect any other personal information. Your answers will only be used for statistical research purposes.

- What is your country of origin that you consider your 'home', influencing your cultural beliefs and other aspects of your identity?
- Is there a country in which you are currently located for a long period of time?
- How old are you? Fill in years in numbers.

After collecting all the responses, we kept only the images where at least two annotators agreed to select the image as culture-specific. After this extra validation, we resulted in a total of 200 images spanning 43 different identified cultures.

A.3 HUMAN EVALUATION

To conduct the human evaluation of the model generated responses we created a survey and hired annotators through the Prolific platform. We asked for a representative sample of countries and an even distribution of male and female participants. We provided the following instructions to the annotators for evaluating the captions:

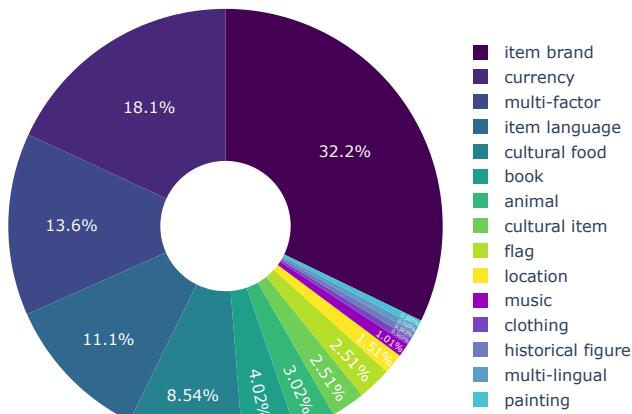


Figure 3: Distribution of the factors/indicators that lead the annotators to select a specific image as culture-related and specify the corresponding culture.

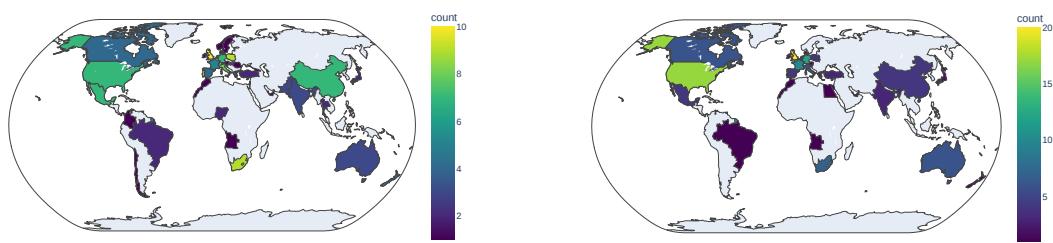


Figure 4: Plots as subfigures.

This study involves evaluating captions. To guide your ratings, imagine that you are describing the image to a visually impaired friend. Then consider:

How well does the caption describe the image to this friend? Does it take into account cultural considerations? You will be given two sets of captions describing an image.

1. Specify which caption you prefer for the given image (1, 2 or both).
2. Determine if each caption is accurate and relevant to the given image.

As a general guidance you should consider a caption as bad when it has one or more of the following issues:

- a) Caption misses the main topic of the image.
- b) Caption has major grammatical errors (such as being incomplete, words in the wrong order, etc). Please ignore the capitalization of words and punctuation.
- c) Caption includes hallucinations and mentions objects, activities, or relationships that are definitely not in the image.
- d) Caption is not as informative.
- e) Caption does not reflect the cultural information depicted in the image.

A.4 CROSS-CULTURAL PERFORMANCE

We also present the performance of our selected closed-source and open-source models across the 36 geographically diverse origins of the images in the XM3600 dataset. We cannot do the same for our filtered VizWiz dataset since mostly due to its size limitations we might get biased and not representative results. Upon reviewing the evaluation of captions across the 36 geographical origins, it becomes apparent that there is a lack of consistency. Diverse cases and models yield varying results, indicating that outputs fluctuate depending on models, metrics, prompts, reference caption comparison, and runs.

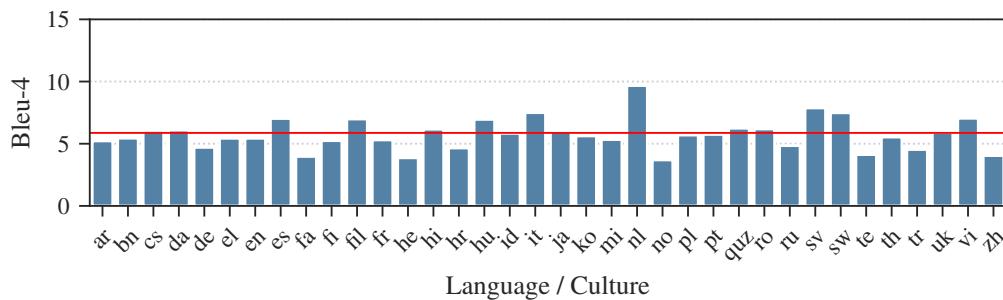


Figure 5: Bleu4 score between original and Blip2-generated captions.

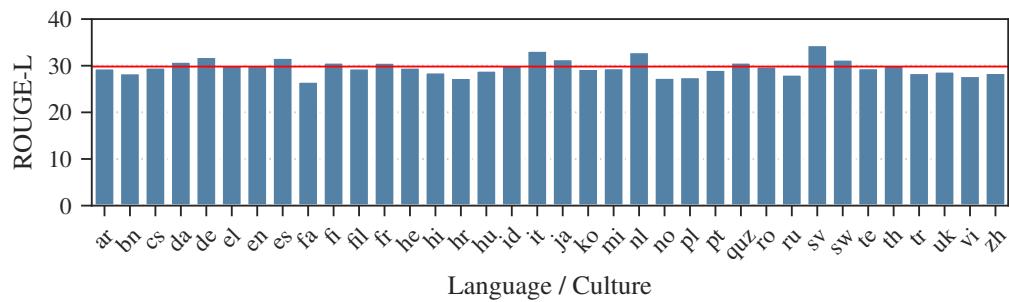


Figure 6: Rouge score between original and Blip2-generated captions.

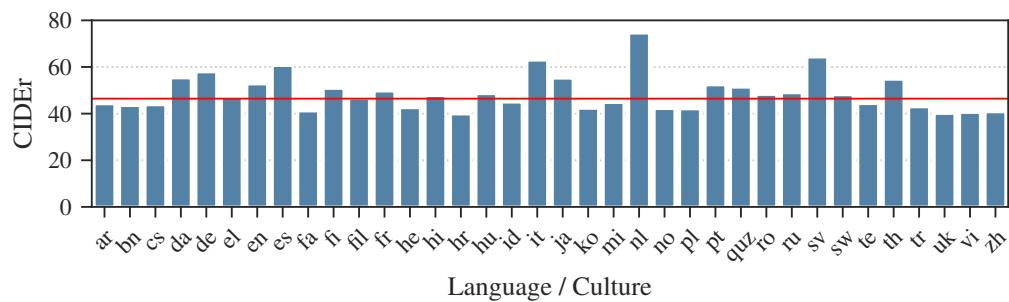


Figure 7: CIDEr score between original and Blip2-generated captions.

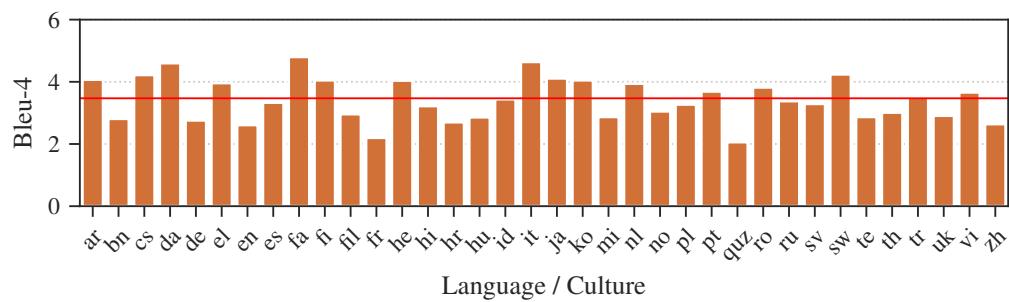


Figure 8: Bleu4 score between original and GeminiPV-generated captions.

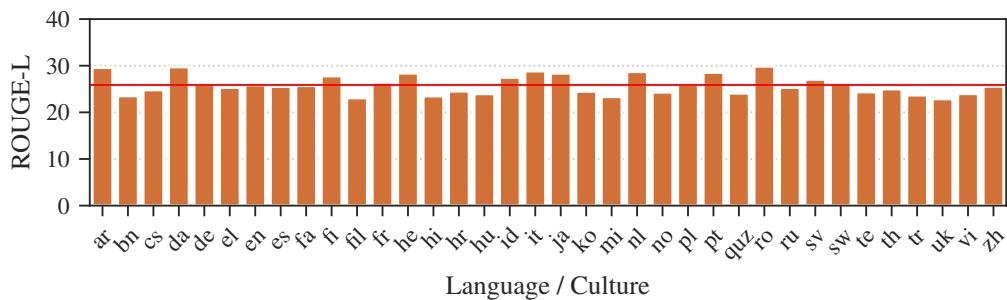


Figure 9: Rouge score between original and GeminiPV-generated captions.

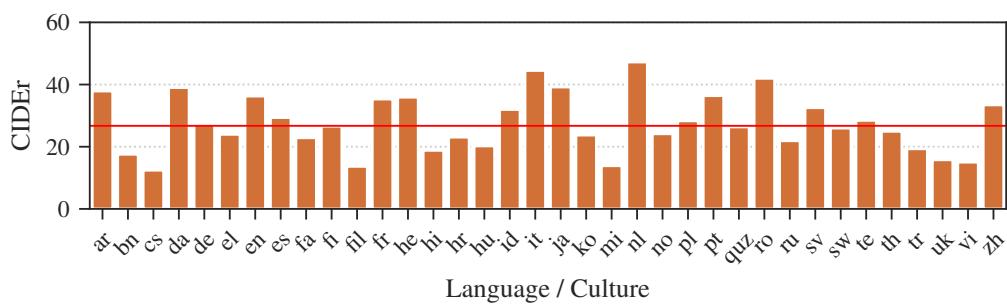


Figure 10: CIDEr score between original and GeminiPV-generated captions.

A.5 CASE STUDY

We illustrate the value of cultural and inclusive VL models via a case study on evaluating a visual assistant integrated into the ‘Be My Eyes’ application powered by OpenAI’s GPT-4V model. In this case study, we took a random sample of 20 images from the MaRVL dataset (Liu et al., 2021) and 20 culturally related images from our filtered VizWiz dataset. Here we provide a selection of images we tried in our case study. Each figure includes the target culture behind each image and the GPT-4 Vision output after loading the image in the Be My Eyes application.



Culture information: The Agamas (Devanagari: आगम, IAST: āgama) (Tamil: ஆகமம், romanized: ākamam) are a collection of several Tantric literature and scriptures of Hindu schools.



GPT4: The picture shows a decorative element that appears to be part of a larger structure, possibly a shrine or temple. At the center is a gold-colored figure with multiple arms in a symmetrical pose. The figure is adorned with a detailed halo-like element and is set against a dark, intricately carved background. Below the figure, there are offerings placed in a bowl, which include what looks like a coconut shell and other items wrapped in leaves. The background is a clear blue sky with some blurred tree branches at the top right corner.

Image description provided with Be My Eyes.

Figure 11: A picture extracted from MarVL depicting a statue of Agamas and the GPT-4V image description provided in BeMyEyes.



Culture information: Chinese Buddhism is a sinicized form of Mahāyāna Buddhism, which draws on the Chinese Buddhist Canon (大藏經, Dàzàngjīng, "Great Storage of Scriptures") as well as numerous Chinese traditions.



GPT4: The picture shows three statues that appear to be of Buddhist deities or figures, each seated on an ornate throne. They are positioned side by side against a backdrop that looks like a temple wall with intricate designs. The central figure is slightly elevated compared to the other two. All figures have a peaceful expression, with eyes closed as if in meditation. The statues are richly decorated with details on their clothing and headpieces. The photo is taken from a slight angle and the image is rotated 90 degrees to the right.

Image description provided with Be My Eyes.

Figure 12: A picture extracted from MarVL depicting Buddhist statues and the GPT-4V image description provided in BeMyEyes.



Cultural information: The leso, or kanga, is a symbol of Swahili culture since the very beginning. The long fabric made of cotton and printed with elaborate design (sometimes finished with a Swahili proverb) expresses emotion, femininity, beauty, and power.



The picture shows a display of various colorful fabric patterns hung on a wooden frame. The patterns are diverse, including geometric shapes, floral designs, and abstract prints in a multitude of colors like red, yellow, blue, orange, and black. The background is a bit cluttered, with a glimpse of a red vending machine and some items on a shelf. The image is taken indoors, and it appears to be a shop or a market stall.

Image description provided with Be My Eyes.

Figure 13: A picture extracted from MarVL depicting a Swahili traditional rectangular cloth and the GPT-4V image description provided in BeMyEyes.



Cultural information: Sambar, known as Kuzhambu in Tamil Nadu, is a lentil-based vegetable stew, cooked with pigeon pea and tamarind broth.



GPT4: The picture shows a bowl of soup with a variety of ingredients visible. The soup is in a terracotta bowl with a textured edge. The soup itself appears to be a broth-based one with chunks of different vegetables and possibly meats. Visible ingredients include what looks like red tomato pieces, green herbs or vegetables, and possibly some beans or small pasta. The colors suggest a hearty and possibly spicy dish. The bowl is sitting on a surface with a striped pattern, and the bottom edge of a laptop keyboard is visible in the foreground, indicating the photo was taken from above a desk or table.

Image description provided with Be My Eyes.

Figure 14: A picture extracted from MarVL depicting sambar, a traditional Tamil dish, and the GPT-4V image description provided in BeMyEyes.



Cultural Information: Doner kebab, also spelled as döner kebab, is a dish of Turkish origin made of meat cooked on a vertical rotisserie



GPT4: The picture shows a person, likely a chef, wearing a white apron and a white hat with a red trim, standing in front of a vertical rotisserie with a large piece of meat cooking on it. The meat is browned on the outside, suggesting it's being roasted. The chef appears to be in a restaurant or a kitchen, as there are other people and what looks like a menu in the background.

Image description provided with Be My Eyes.

Figure 15: A picture extracted from MarVL depicting a döner, a traditional Turkish dish, and the GPT-4V image description provided in BeMyEyes.