Foodies also Need Their Own GPTs: Evaluating Language Models in Visual Question Answering on the WorldCuisines Dataset

Genta Indra Winata, Emmanuele Chersoni

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Foodies also Need Their Own GPTs: Evaluating Language Models in Visual Question Answering on the `WorldCuisines` Dataset

Genta Indra Winata
Capital One AI Foundations
`gentaindrawinata@gmail.com`

Emmanuele Chersoni
The Hong Kong Polytechnic University
`emmanuelechersoni@gmail.com`

Food is universally acknowledged as an important medium of expression of cultures, and at the same time a way for different people and traditions to connect to each other (Wahlqvist, 2007). Dishes reflect local identities and can tell stories, and those can be shared in turn across countries that share a culinary heritage (Anderson, 2014). There can be slight variations on similar recipes, often difficult to distinguish for the unexperienced eye; at the same time, with the global increase of food tourism all over the world (Ellis et al., 2018), there will be soon the need for new technologies to fill the gap between the image of a dish on a menu and the curiosity of a tourist willing to taste something new. Recognizing a dish and providing information about it are not trivial challenges even for multimodal large language models (MLLMs), given the large variety of existing dishes and recipes. In order to stimulate research on food-related visual question answering, we developed the WorldCuisines dataset (Winata et al., 2025), a multilingual and multicultural benchmark comprising text-image pairs of dishes for 30 different languages, for a total of more than 1 million data points. The dataset supports different tasks, such as identifying specific food types and their origins from their pictures. Our preliminary evaluation shows that performance varies a lot across different models. MLLMs generally perform better when they are provided with information about a typical location for a given dish in the prompt (e.g. I am in Hanoi and I am about to eat this now, together with an image showing a *bánh mì*), while they struggle when the prompt contains a location that is not typical (e.g. consider the same prompt as before, but with Hokkien fried rice instead of *bánh mì*). Such findings suggest that there MLLMs can be easily misled by unexpected textual information about the location of the user, and therefore they still have a lot of room for improvement.

References

Eugene Newton Anderson. 2014. *Everyone Eats: Understanding Food and Culture*. NYU Press.

Ashleigh Ellis, Eerang Park, Sangkyun Kim, and Ian Yeoman. 2018. What Is Food Tourism? *Tourism Management*, 68:250–263.

Mark L Wahlqvist. 2007. Regional Food Culture and Development. *Asia Pacific Journal of Clinical Nutrition*, 16:2.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025. WORLDCUISINES: A Massive-scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. In *Proceedings of NAACL*.