What are Foundation Models Cooking in the Post-Soviet World?

Anton Lavrouk, Tarek Naous, Alan Ritter, Wei Xu
Georgia Institute of Technology
{antonlavrouk, tareknaous}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

Abstract

The culture of the Post-Soviet states is complex, shaped by a turbulent history that continues to influence current events. In this study, we investigate the *Post-Soviet cultural food knowledge* of foundation models by constructing BORSCH, a multimodal dataset encompassing 1147 and 823 dishes in the Russian and Ukrainian languages, centered around the Post-Soviet region. We demonstrate that leading models struggle to correctly identify the origins of dishes from Post-Soviet nations in both text-only and multimodal Question Answering (QA), instead over-predicting countries linked to the language the question is asked in. Through analysis of pretraining data, we show that these results can be explained by misleading dish-origin co-occurrences, along with linguistic phenomena such as Russian-Ukrainian code mixing. Finally, to move beyond QA-based assessments, we test models' abilities to produce accurate visual descriptions of dishes. The weak correlation between this task and QA suggests that QA alone may be insufficient as an evaluation of cultural understanding. To foster further research, we will make BORSCH publicly available at github.com/alavrouk/BORSch.

1 Introduction

The Post-Soviet states have long held their own cultural and linguistic identities. During the Soviet era, these identities were pressured through forced assimilation under the Russian language and culture (Silver, 1974). Now, 33 years after the collapse of the Soviet Union, the Post-Soviet world continues to repair the damage inflicted by this so-called "Sovietization" (Rutland, 2023).

As foundation models continue to gain prominence, it is important that they are able to represent each Post-Soviet state. Yet, when examined via *food dishes*, an important element in every culture (Anderson, 2014), we find that they lack crucial knowledge. For example, кывырма (pronounced

Figure 1: A map of predictions from various popular consumer-facing models for the dish кывырма (*kivirma*), a traditional pastry from the Moldovan region of Gagauzia, which is home to a significant Russian-speaking population. The models were prompted in Russian, with the English translation also shown.

"ky-vyr-MA") is a dish from Gagauzia, a region of Moldova where Russian is commonly spoken (Mayer, 2014). However, Figure 1 shows that when asked in Russian, multilingual models fail to identify the origins of this dish, with each one predicting an incorrect Post-Soviet nation.

In order to further investigate these deficiencies, we conduct a detailed exploration of food culture understanding in foundation models across both text and image modalities. We focus our study on the Russian and Ukrainian languages, analyzing how the push of Ukrainian "de-Sovietization" (Boman, 2023) and the pull of historical Russian interference on Ukrainian culture (Boychuk et al., 2023) impacts the cultural perceptions of foundation models. Overall, our contributions are as follows:

- We construct BORSCH[1] (**B**enchmark **O**f **R**egional di**S**hes), a dataset for evaluating foundation models on multimodal food culture understanding in Russian and Ukrainian

[1]Named after the famous Ukrainian dish, борщ.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20699–20721
November 4-9, 2025

(§3). BORSCH is constructed via a bootstrapped entity extraction approach for collecting culturally relevant food dishes from web crawl corpora, with human-in-the-loop validation (§3.2).

- To compare models prompted in Russian and Ukrainian, we perform text and vision country of origin Question Answering (QA) using dishes in BORSCH. Beyond varying performances across countries, we find that models prompted in these languages over-predict their respective countries of origin (§4.1).
- To gain a more nuanced understanding of how models perform on Post-Soviet dishes in Ukrainian, we examine how the Russian-Ukrainian pidgin[2] *surzhyk* influences Ukrainian QA and VQA performance (§4.2).
- Through pretraining data analysis, we find many instances where BORSCH dishes co-occur with non-origin countries, harming model QA performance. In contrast to English corpora, these issues in Russian and Ukrainian largely stem from poor web-scraping (§4.3).
- Finally, we conduct an experiment which queries models for descriptions of dishes in BORSCH, which we then evaluate using a modality transition from text to image. We find this to be a challenging task with limited correlation to QA experiments (§5).

2 Related Work

Cultural Knowledge Bases. Recent interest regarding cultural-knowledge in foundation models has led to numerous studies attempting to quantify it (Hershcovich et al., 2022; Adilazuarda et al., 2024; Liu et al., 2024). Some studies construct multilingual knowledge bases of cultural assertions (e.g., *"In Bhutan, there is a tradition of wearing "Khyenkhor Robes" woven with threads infused with blessings from Buddhist monks"*) (Nguyen et al., 2023, 2024; Fung et al., 2024). Other works craft benchmarks of culturally-specific questions (e.g., *"What is the story of the series Al-Manassa?"*) (Yin et al., 2022; Myung et al., 2024; Shen et al., 2024; Arora et al., 2024). Further research expands on such directions to support multimodality (Ramaswamy et al., 2023; Liu et al., 2023; Libovický et al., 2025). There are also additional studies which focus exclusively on vision-based tasks such as culturally informed image generation (Bhatia et al., 2024; Karamolegkou et al., 2024; Kannen et al., 2024), visually grounded reasoning (Schneider and Sitaram, 2024), and image transcreation (Khanuja et al., 2024). While some of these works include food as part of their overall assessment, they mainly focus on broad cultural understanding. Meanwhile, we offer a more in-depth analysis on the nuances of cultural *food* knowledge.

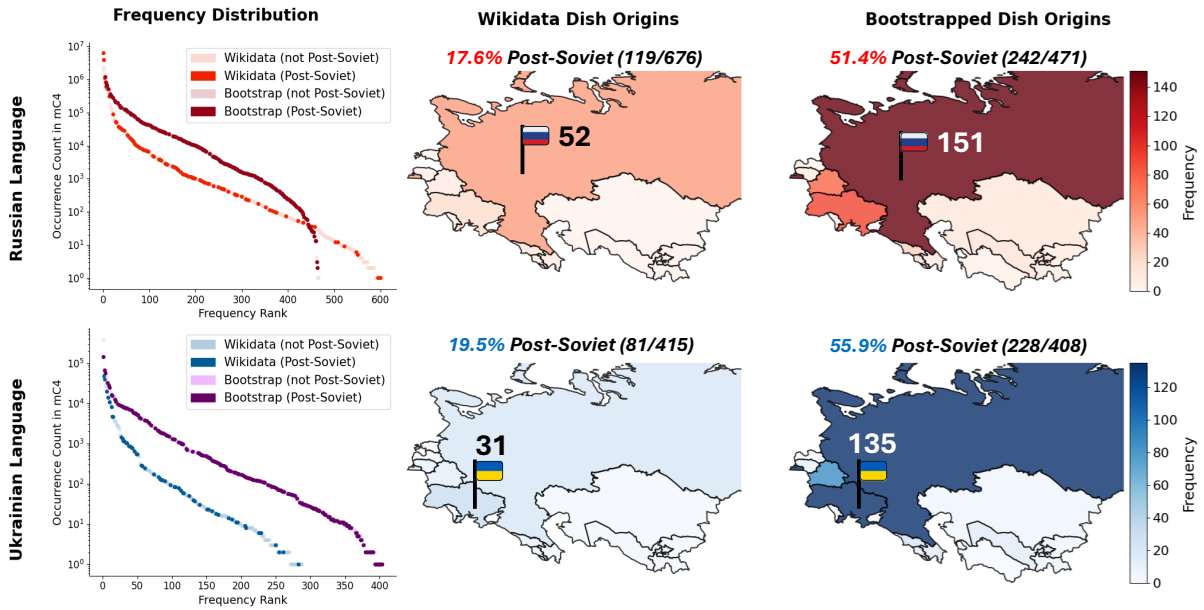
Cultural Food Knowledge. Food knowledge is a key element of culture, and is thus frequently evaluated in foundation models. Some studies assess model comprehension of culinary practices or dishes through pragmatic questioning (e.g., "*While eating, when does one drink Cantonese seafood soup?*") (Palta and Rudinger, 2023; Yao et al., 2023; Putri et al., 2024; Li et al., 2024c). Another line of reasoning uses food to attribute cultural generations to pretraining data (Li et al., 2024b). Finally, a group of works measures food culture understanding in foundation models by testing them on a culturally diverse set of food-dish entities. However, the food-dishes used for evaluating models in past work are obtained solely from either Wikidata (Zhou et al., 2024) or Wikipedia (Winata et al., 2024), which we show leads to missing out on many culture-specific dishes in non-English languages (§3.1). For example, the food-dishes originating from Russia and Ukraine in WORLDCUISINES (Winata et al., 2024) cover only 20.8% of the dishes originating from Russia and Ukraine that we provide in BORSCH.

Russian and Ukrainian Culture in LLMs. Existing cultural studies on the Russian language in foundation models focus on social/gender biases (Grigoreva et al., 2024; Kuznetsov, 2024; Li et al., 2024a) or image generation (Vasilev et al., 2025). For the Ukrainian language, Kharchenko et al. (2024) explores cultural values of foundation models. From our understanding, our study is the first to perform a large scale exploration of Post-Soviet cultural knowledge using the Russian and Ukrainian languages in parallel.

3 Constructing BORSCH

To enable a more in-depth assessment of models' understanding of Russian and Ukrainian food cultures, we focus on collecting both the popular and less commonly known dishes relevant to those cultures. We achieve this by first extracting all avail-

[2] A pidgin is a simplified language for communication between people with different native tongues (Romaine, 2017).



(a) (b)

Figure 2: **(a)** Frequency in mC4 vs frequency rank in BORSCH. Culturally relevant bootstrapped dishes are both common and long-tail, while Wikidata dishes are less frequent overall. **(b)** Countries of origin of dishes in BORSCH, which were obtained from multilingual Wikidata (§3.1) and commonly used web-crawled corpora (§3.2). While there are less bootstrapped dishes, they are more likely to originate from a Post-Soviet nation.

able dishes in Wikidata (§3.1), then expanding on this initial set through a bootstrapped extraction approach from web-crawled data with human-in-the-loop (§3.2). We also annotate countries of origin (§3.3), collect images for the dishes (§3.4), and create a Post-Soviet, parallel sub-dataset of dishes (§3.5), enabling many multimodal evaluations.

3.1 Extracting Food Entities from Wikidata

As a starting point, we acquire an initial list of dishes from the Wikidata multilingual knowledge base[3] in the Russian and Ukrainian languages. We extract all entities that are registered under the class "food" in Wikidata, which encompasses many food-related sub-classes (e.g., *sweets*, *fast food*, etc.). We then manually select culturally relevant food dishes that are attributable to a specific country/countries of origin, and discard beverages and more generic food entities (e.g., globally common dishes such as *grilled chicken*, branded goods such as *kitkat*, etc.). The resulting coverage of food entities in Wikidata is relatively poor for both Russian (676 dishes) and Ukrainian (415 dishes) languages. Moreover, only 119 dishes (17.6%) in Russian Wikidata and 81 dishes (19.5%) in Ukrainian Wikidata are associated with origins in any Post-Soviet state, which indicates that the existing coverage is not only sparse but also lacks cultural relevance. We address this multilingual coverage gap in Wikidata by collecting additional dishes from web-crawl data.

3.2 Bootstrapped Extraction from Corpora

Previous research by Naous et al. (2024) demonstrated that culturally-relevant food dishes can be collected from large web-crawl corpora. Their approach relies on extracting unigrams and bigrams appearing after a set of manually crafted patterns which likely occur before the mention of a food dish (e.g., *recipe of* ___, *how to cook* ___ , etc.). This was followed by human annotation to filter out erroneous extractions. We build on this method by performing bootstrapped pattern-based extraction with a human-in-the-loop to iteratively collect food dishes from the Russian and Ukrainian portions of the mC4 web-crawl corpus (Xue et al., 2021).

We start with the dishes obtained from Wikidata (§3.1) as a seed list, which we use to search the corpus of each respective language and retrieve all 3-gram and 4-gram patterns that precede any dish. We then ask a human annotator to select five from the 100 most frequent patterns. Using the selected patterns, we search the corpus again and extract every unigram that appears after a 3-gram pattern and every bigram that appears after a 4-gram pattern. This strategy exploits the exponential growth in

[3] www.wikidata.org/wiki/Wikidata:Main_Page

n-gram counts: rare 4-gram patterns limit bigram extraction volume to make manual review more feasible, while more common 3-gram patterns ensure sufficient unigram extraction volume.

Finally, we de-duplicate the extracted unigrams and bigrams, which results in up to 10k extractions that we give to human annotators to manually filter for food dishes. We repeat this bootstrapping process for two more rounds. Detailed statistics regarding extractions during the bootstrapping process are located in Appendix A. During the filtering process, a random sample of 3770 extractions (2015 in Russian and 1755 in Ukrainian) underwent double annotation, yielding substantial annotator agreement with Cohen's Kappa (κ) values of 0.73 and 0.77 for Russian and Ukrainian respectively.

3.3 Determining a Dish's Country of Origin

In order to enable the evaluation of models' food culture understanding, we annotate each collected dish for its associated country/countries of origin. Two college educated annotators, one fluent in Russian and one fluent in Ukrainian, conducted independent research using web resources on each dish and manually labeled each dish's country of origin. In cases where dishes were found to have multiple countries of origin (20% of dishes in Russian, 27% in Ukrainian), particularly for areas that predate modern country borders, annotators were asked to label all relevant countries. An example is чак-чак (*chak-chak*), a popular cake originating from Central Asia which existed before the Soviet Union. Its origins were labeled as Russian (Tatar and Bashkir)[4], Kazakh, Tajik, Kyrgyz, and Uzbek as it is a common delicacy in all of those nations.

Figure 2b compares the origins of food dishes extracted from Wikidata (§3.1) vs. the bootstrapped process (§3.2) on the map. We find that bootstrapping retrieves more dishes that are common in the Post-Soviet region. Furthermore, as shown in Figure 2a, the bootstrapping process helps cover Post-Soviet dishes in Russian and Ukrainian that are both highly frequent and long-tail in corpora, while Post-Soviet dishes obtained from Wikidata consist of mostly popular dishes.

3.4 Dish Image Collection

To facilitate vision-languages analyses, we collect up to 5 images for each dish in BORSCH. We first searched for images in Wikimedia Commons[5], a collection of freely usable media files. This process enabled image retrieval for 74% of dishes in Russian and 68% in Ukrainian. For the remainder of the dishes which did not have images in Wikimedia Commons, we used the Google Custom Search API[6] and queried for images with a Creative Commons (CC) open license. All retrieved images were then manually filtered to remove irrelevant content (see filtering interface in Appendix B). In total, we collected 5285 images in Russian (3.64 images per dish on average), and 2907 images in Ukrainian (3.53 images per dish on average). Additionally, we inspect potential VLM pretraining data contamination involving these images in Appendix C.

3.5 Aligning Russian & Ukrainian BORSCH

To enable direct comparisons between languages, we manually translate (and transliterate, when necessary) dishes originating from Post-Soviet countries in the Russian set to Ukrainian and vice-versa, resulting in a parallel sub-dataset of 433 dishes with names in both Russian and Ukrainian. Of these dishes, 174 (40%) appear in both datasets, while 126 (29%) are unique to the Russian dataset and 133 (31%) are unique to the Ukrainian dataset. The distribution of origins for the parallel sub-dataset can be found in Appendix D. We validated the dish origin annotations in this parallel sub-dataset by engaging a second annotator, achieving substantial agreement with a Cohen's Kappa (κ) of 0.879. Furthermore, we note that each dish in this sub-dataset can now have up to 10 images (if it was originally a part of both Russian and Ukrainian BORSCH).

4 LLM Performance: Dish QA & VQA

To begin, we test foundation models prompted in Russian and Ukrainian on QA and VQA tasks focusing on dish origins within the parallel sub-dataset of BORSCH (§4.1). Then, to further understand these results, we explore the effect of Russian code mixing on Ukrainian QA and VQA (§4.2). Finally, we investigate dish-country co-occurrences in the underlying pretraining data as a factor influencing both Russian and Ukrainian QA (§4.3).

4.1 Parallel Country of Origin QA & VQA

We first assess a model's ability to predict a dish's country/countries of origin in two setups: **(i)** standard text-based Question Answering (QA) where the model is provided the dish name and **(ii)** Visual

[4]These are two minority ethnic groups in Russia.

[5]commons.wikimedia.org/wiki/Main_Page

[6]developers.google.com/custom-search/v1/

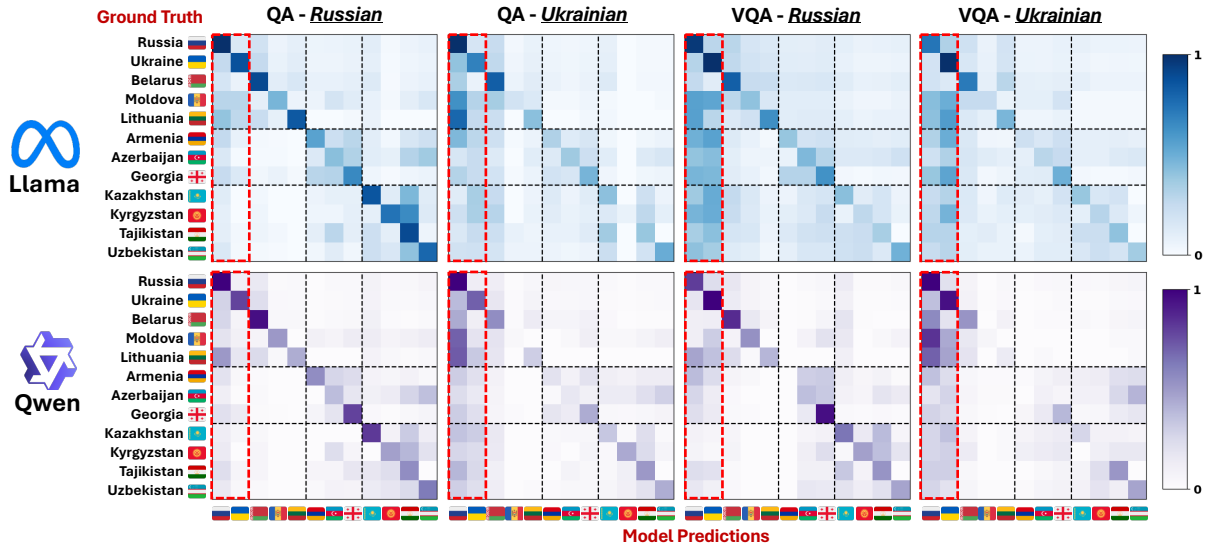


Figure 3: Confusion matrices for country-of-origin QA/VQA on dishes in BORSCH. Models exhibit a low recall on Russian and Ukrainian dishes, and struggle with Post-Soviet countries in the Caucasus (Armenia, Azerbaijan, Georgia) and Central Asia (Kazakhstan, Kyrgyzstan, Tajikistan, Uzbekistan). Estonia, Latvia, and Turkmenistan are excluded due to low sample size, and the confusion matrices are row (truth) normalized, as each country is not equally represented in BORSCH.

Question Answering (VQA) where the model is provided an image of the dish. We use the **Post-Soviet parallel sub-dataset** of BORSCH, enabling fair, cross-lingual comparisons.

Setup. We evaluate `Qwen2-72B-Instruct` (Yang et al., 2024) and `Llama-3.1-70B-Instruct` (Dubey et al., 2024) on text-only QA tasks, and their vision-enabled counterparts `Qwen2-VL-72B-Instruct` and `Llama-3.2-90B-Vision-Instruct` for VQA tasks[7]. We prompt these models with open-ended questions to align with real-world applications (Röttger et al., 2024). To identify countries in each model's response, we use `spaCy`'s multilingual NER tool, known to be highly effective for location recognition (Honnibal et al., 2020). After extracting named entities, we search them for country names or aliases (e.g., *Czech Republic* vs. *Czechia*). This two-step approach allows us to detect any missing aliases by manually analyzing named entities not in our country/alias list. Models are prompted using five different variants of the same question, each containing placeholders for dish names (see Appendix F). We attach all (up to 10) available images to each VQA prompt.

Results. Confusion matrices[8] for the country-of-origin QA/VQA experiments are presented in Figure 3. We focus on Post-Soviet predictions for cultural specificity, with frequent non-Post-Soviet errors listed in Appendix G. Furthermore, in Appendix H, we report the results of a modified VQA task where models are additionally provided dish names alongside images. This additional information improves performance on most dishes. Overall, we find that when prompted in Russian and Ukrainian, models frequently over-predict Russia and Ukraine as dish origins. In Figure 3, this is evident from the wide distributions in the Russian and Ukrainian prediction columns. We investigate this further by focusing on dishes whose origins include both Russia and Ukraine (35% of the parallel corpus). Figure 4 shows that when prompted in Russian, models are more likely to predict Russia as an origin for these dishes, while models prompted in Ukrainian are more likely to predict Ukraine. This mainly impacts models prompted in Russian and affects QA tasks more than VQA.

4.2 Impact of Code Mixing on QA & VQA

One potential factor influencing model QA/VQA performance in Ukrainian is Russian code mixing. In the period following Ukraine's establishment as a sovereign state in 1991, many Russian speakers in Ukraine transitioned to speaking the Ukrainian language (Fomenko, 2023). Understandably, this linguistic adjustment was not instant, which is why *surzhyk*, referring to various mixed Russian-Ukrainian codes, has since gained

[7] More information in Appendix E regarding model choice.

[8] As models can generate multiple predictions, we utilize the methodology in Heydarian et al. (2022) and Krstinić et al. (2020) to construct "multi-label confusion matrices."

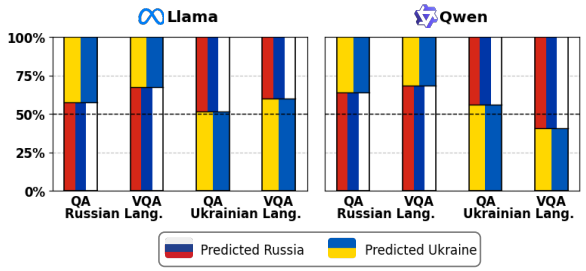


Figure 4: For dishes originating from both Russia and Ukraine, Russian models tend to predict Russia more often than Ukraine, and Ukrainian models tend to predict Ukraine more often than Russia. The proportions shown exclude non-Russian/Ukrainian predictions.

a foothold in Ukraine (Kurapov et al., 2024). For example, in surzhyk, sentences that are otherwise fully Ukrainian often incorporate Russian words, particularly nouns (Podolyan et al., 2005). On the other hand, the reverse phenomenon is far less common. Using mC4 as a representative corpus of text data (more on this later), we find that Ukrainian code mixing accounts for 17% of BORSCH dish occurrences in Russian mC4, which is far less than Russian dish occurrences in Ukrainian mC4 (41%).

To study surzhyk in pretraining corpora, we use BORSCH, particularly since the dish names in our parallel Ukrainian-Russian sub-dataset only differ by an average edit distance of 2.3 characters. This similarity makes the dishes in BORSCH especially susceptible to code mixing (e.g., Ukrainian пир**і**г versus Russian пир**о**г)[9].

Setup. Using the Russian and Ukrainian dish names in the parallel sub-dataset of BORSCH (§ 3.5), we search for instances of Russian code-mixing in Ukrainian corpora. In particular, we search mC4, the most widely used and extensively studied open multilingual corpus for pretraining (Kreutzer et al., 2022). While more recent corpora contain newer, higher quality data, they do not contain multilingual components, which are crucial for our analysis. We then quantify surzhyk by analyzing the *difference in mC4 occurrences of the Ukrainian dish name and the Russian dish name*.

To assess model performance, we calculate the Jaccard score per dish as each dish may have a varied number of countries as its origin. The Jaccard score (Jaccard, 1901) measures the similarity between the predicted and ground truth country of origin sets as the size of their intersection divided by the size of their union. Importantly, this *set overlap* Jaccard score accounts for countries predicted by the model which are not part of the gold set. Additionally, to enable the analysis of surzhyk in VQA, we modify the VQA experiment by asking the model to *name* a dish given its image. We evaluate model performance on this experiment using *exact match accuracy over 5 prompts* with a tolerance of 1 edit (Levenshtein) distance. While exact match is a common QA metric (Rajpurkar, 2016), we introduce the 1 edit distance tolerance due to Russian and Ukrainian noun declension (Press and Pugh, 2015; Comrie, 2018). Full, per-country results for this experiment can be found in Appendix I.

[9]Directly translates to "pie," but colloquially represents a specific class of Eastern European pastries.

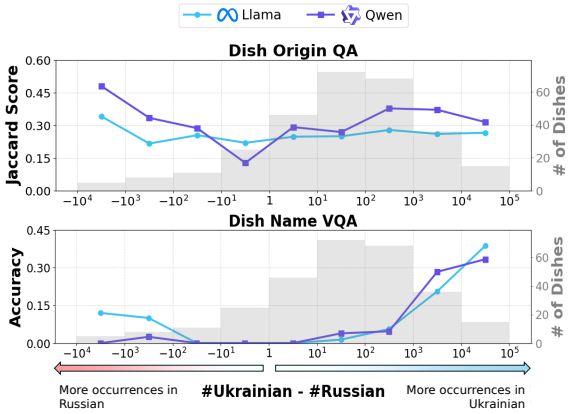


Figure 5: Dish origin QA and dish name VQA tasks suffer when prompting Qwen and Llama in Ukrainian because dish names lack standardization in Ukrainian corpora, which can occur due to the use of *surzhyk* (a mixed Russian-Ukrainian code). Dishes with identical Russian and Ukrainian names are not analyzed.

Results. We begin by calculating the average QA Jaccard score and VQA exact match accuracy across nine log-spaced bins. These bins are determined by occurrence differences in Ukrainian mC4 (#Ukrainian - #Russian) for the 66.5% of dishes in the parallel dataset that have distinct Russian and Ukrainian names. The results are located in Figure 5. For dish origin QA, we observe the poorest performance when the pretraining data contains an approximately equal mix of Russian and Ukrainian names, while dishes standardized to either a Russian or Ukrainian name perform better. Similarly, for dish name VQA, we find that it is important for the dish name to be standardized, although preferably using the Ukrainian name.

4.3 Impact of Co-Occurrences on QA

Previously, we demonstrated that QA performance in Ukrainian can be affected by Russian code mixing (§4.2). We now turn our attention to **incorrect**

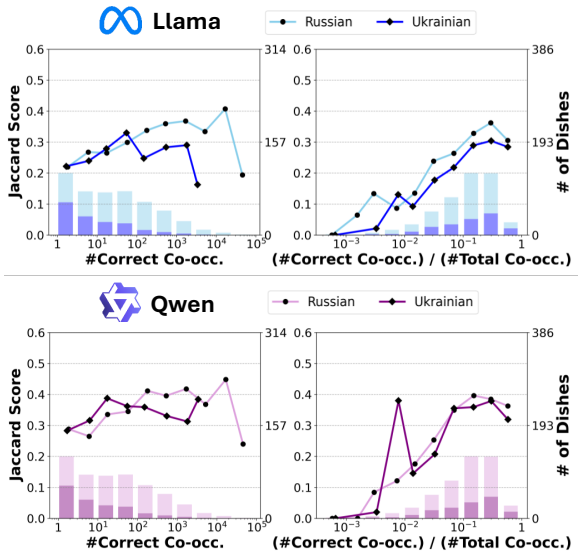


Figure 6: While a higher number of correct dish-country co-occurrences (`#Correct Co-occ.`) supports better text-only QA performance, the ratio of correct v.s. total co-occurrences (`#Correct Co-occ./#Total Co-occ.`) proves even more crucial in both Llama and Qwen.

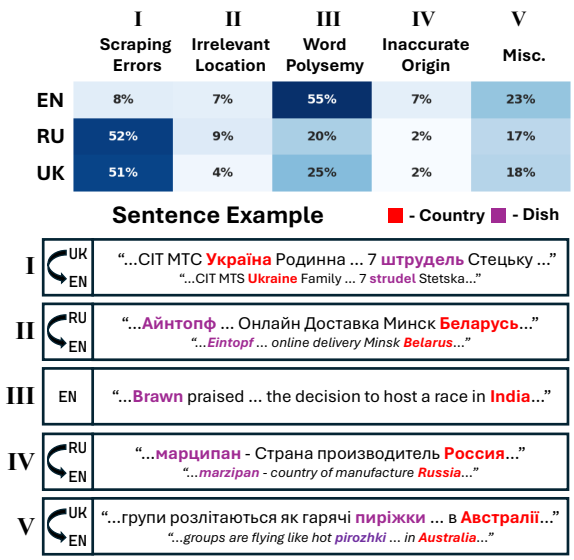


Figure 7: An inspection of 400 randomly sampled incorrect dish-country co-occurrences in English, Russian, and Ukrainian mC4 reveals that Russian and Ukrainian data suffers disproportionately from poor web scraping.

dish-country co-occurrences in both Russian and Ukrainian pretraining data.

Setup. To start, we analyze how frequently each BORSCH dish (full dataset) appears alongside country names and aliases in the Russian and Ukrainian subsets of mC4. For each dish, we define two metrics. The **correct co-occurrence count** is the number of mC4 documents that mention both the dish and its true country of origin. The **correct co-occurrence ratio** divides this count by the total mentions of that dish with any country (correct or incorrect). We then examine how these metrics impact model country of origin QA performance.

Results. Figure 6 shows text-only QA performance for Qwen and Llama averaged over dishes which are grouped into 10 log-spaced bins based on their correct co-occurrence counts or ratios. We find an improvement in text-only QA across both Russian and Ukrainian languages as dishes occur more frequently with the correct country in pretraining data. More importantly, the improvement in Jaccard score is steeper when looking at the correct co-occurrence ratio, indicating that it is critical for a dish not to co-occur frequently with incorrect countries of origin.

Qualitative Analysis. Furthermore, we also seek to identify why food dishes co-occur with countries irrelevant to their origin to begin with. To answer this, we sample 400 sentences in each language where dishes co-occur with countries other than their true origin. To ensure a more informative sample, we only allow a unique dish to be selected three times. To see if incorrect co-occurrences differ between Russian/Ukrainian and English languages, we extract English Wikidata dishes (§3.1) and annotate their origins (§3.3), resulting in 2348 total dishes (which we will release along with the rest of BORSCH). Then, we similarly sample 400 incorrect co-occurrences for these dishes.

Figure 7 shows the distribution of incorrect co-occurrence cases across the Russian, Ukrainian, and English languages. We find that Russian and Ukrainian corpora suffer greatly from web scraping errors, while the English corpus does not. Other notable issues include word polysemy (previously studied in Naous and Xu 2025), irrelevant geographic mentions, inaccurate dish origins, and incidental occurrences where a dish and a country appear together but are unrelated (miscellaneous).

5 Dish Description Generation

Finally, we introduce a new generation-based task for food cultural understanding that goes beyond conventional QA setups. In particular, we focus specifically on a model's ability to describe the *appearance* of a dish in the BORSCH parallel dataset.

Setup. We begin by prompting the models to produce textual descriptions of a dish's appearance

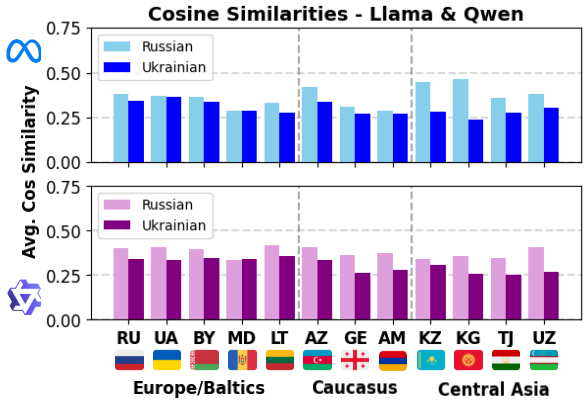


Figure 8: Models prompted in Russian are better equipped to describe Post-Soviet, culturally relevant dishes compared to models prompted in Ukrainian.

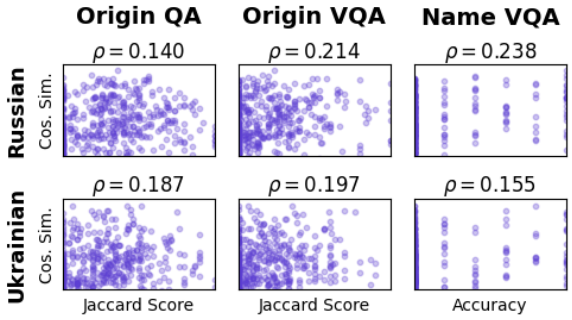


Figure 9: The Spearman's ρ between dish-description performance and QA tasks on **Qwen** models shows a small, positive correlation. Llama models show a similar trend (Appendix J.3). All axes span from 0 to 1.

given its name (which is redacted in the rare case where it is part of a model's output). As in §4.1 and §4.2, we use the dishes in the parallel subdataset of BORSCH. To assess the accuracy of a generated description, we translate it to English[10] and use it to prompt `FLUX.1-dev`[11] to generate an image of the dish. We measure similarity between the generated and ground-truth images collected in BORSch (§3.4) using the `DiNOv2-giant` (Oquab et al., 2023) image encoder. Following the approach used by `DiNOv2`'s creators and recent work by Khanuja et al. (2024), we extract `[CLS]` token embeddings from generated and ground-truth images and compute their cosine similarity. To create a frame of reference for these scores, we utilize the fact that the dishes in BORSCH have up to five different images. We find that intra-dish image embeddings exhibit a mean cosine similarity of 0.52, whereas inter-dish embeddings average 0.07.

[10] `cloud.google.com/translate`

[11] `huggingface.co/black-forest-labs/FLUX.1-dev`, currently ranked #1 on GenAI Arena (Jiang et al., 2024).

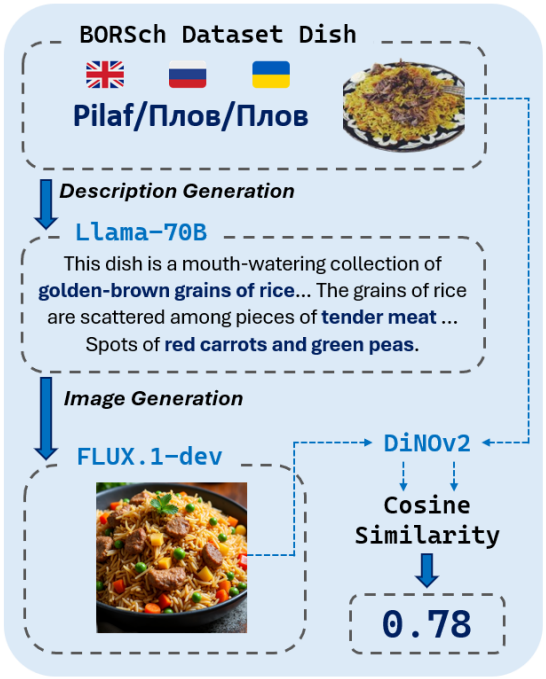


Figure 10: Our dish description evaluation pipeline for the dish *pilaf*. `Llama-3.1` generates a valid, detailed description (which we abridge to include key points).

Results. Figure 8 presents the average cosine similarities for each language-region pair. Additionally, Figure 10 displays a qualitative example of our image description evaluation pipeline. More examples spanning many cosine similarities can be found in Appendix J.1. Overall, we find that Russian models outperform Ukrainian models, even on dishes originating from Ukraine.

To ensure the veracity of our evaluation pipeline, we perform human evaluation on a random sample of 400 model generated dish descriptions. A fluent annotator rates the quality of the *pre-translation* descriptions from 0 to 1, scoring how well they visually describe the ground truth images. The Pearson correlation coefficient (r) between human ratings and cosine similarities from our evaluation pipeline showed strong agreement, with values of 0.78 (Russian) and 0.82 (Ukrainian) for `Qwen`, and 0.81 (Russian) and 0.82 (Ukrainian) for `Llama`. Additionally, in Appendix J.2, we test alternative correlation measures to ensure consistent results.

Finally, we present a scatter plot of embedding cosine similarities vs. models' QA performances (§4.1, §4.2) in Figure 9. We observe weak positive correlations, indicating that our two designed evaluations are complementary, each capturing different aspects of cultural food knowledge. Further details regarding this claim are located in Appendix J.4.

6 Conclusion

We present BORSCH, a dataset targeted at evaluating cultural food knowledge in the Russian and Ukrainian languages. Through BORSCH, we identify significant gaps in model knowledge and investigate pretraining data to uncover the causes of these shortcomings. We hope our insights into incorrect co-occurrences and language contamination in pretraining data will contribute to building more culturally aware models and pretraining corpora.

Limitations

Our study is (intentionally) narrow in its scope, focusing exclusively on Post-Soviet dishes in Russian and Ukrainian. This narrow scope allows us to explore insights which are relevant to the two languages and the culture surrounding them, such as the Russian-Ukrainian code switching known as surzhyk. This is important due to the historically complex relations between the Russian and Ukrainian languages (Kulyk, 2024).

Additionally, using the Russian and Ukrainian languages yields a dataset that is skewed towards dishes originating from countries that predominantly speak either of these languages. To accommodate this, our experiments mainly focus on the distinctive relationship between Russia and Ukraine, while also gathering dishes from other post-Soviet nations to reveal smaller insights and lay the groundwork for future research.

Finally, we annotated and conducted QA exclusively on dish origin and name. However, dishes have other subjective/varying characteristics worth exploring, such as taste or smell. We chose to use origin/name as they are a more objective measure that directly measures model knowledge, not opinion. Future work can focus on model preferences/opinions on food dishes in Russian vs. Ukrainian, but this is outside the scope of this study.

Acknowledgments

The authors would like to thank Oleksandr Lavreniuk, Dennis Pozhidaev, and Jad Matthew Bardawil for their valuable discussion and annotation; Kartik Goyal for their valuable discussion. In memory of Sergei Novikov.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Eugene Newton Anderson. 2014. *Everyone eats: Understanding food and culture*. NYU Press.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*.

Björn Boman. 2023. The coexistence of nationalism, westernization, russification, and russophobia: facets of parallelization in the russian invasion of ukraine. *International Politics*, 60(6):1315–1331.

Yuliana Boychuk, Rory Dumelier, Yevheniya Fau, Khrystyna Mysiv, Anastasiya Sereda, Alison Toews, and Courage White. 2023. The effect of russian colonialism on ukrainian cultural identity.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Bernard Comrie. 2018. Russian. In *The world's major languages*, pages 282–297. Routledge.

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

Olena Fomenko. 2023. Brand new ukraine? cultural icons and national identity in times of war. *Place Branding and Public Diplomacy*, 19(2):223–227.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition and lm benchmarking.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. Rubia: A russian language bias detection dataset. *arXiv preprint arXiv:2403.17553*.

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. 2022. Challenges and strategies in cross-cultural nlp.

Mohammadreza Heydarian, Thomas E Doyle, and Reza Samavi. 2022. Mlcm: Multi-label confusion matrix. *Ieee Access*, 10:19083–19095.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. 2024. Genai arena: An open evaluation platform for generative models.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *arXiv preprint arXiv:2407.06863*.

Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. *arXiv preprint arXiv:2407.06177*.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *arXiv preprint arXiv:2404.01247*.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions.

Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. One ruler to measure them all: Benchmarking multilingual long-context language models. *arXiv preprint arXiv:2503.01996*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Damir Krstinić, Maja Braović, Ljiljana Šerić, and Dunja Božić-Štulić. 2020. Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, 1(2020):1–14.

Volodymyr Kulyk. 2024. Language shift in time of war: the abandonment of russian in ukraine. *Post-Soviet Affairs*, 40(3):159–174.

Anton Kurapov, Oleksandra Balashevych, Christoph Bamberg, and Pawel Boski. 2024. Cutting cultural ties? reasons why ukrainians terminate or continue to interact with russian culture despite the ongoing russian-ukrainian war. *Journal of Cross-Cultural Psychology*, 55(5):553–571.

Aleksey Kuznetsov. 2024. Machine learning and culture: An analysis of sociocultural biases in large language models using gigachat and yandexgpt (russian).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024a. Uncovering differences in persuasive language in russian versus english wikipedia.

Huihan Li, Arnav Goel, Keyu He, and Xiang Ren. 2024b. Attributing culture-conditioned generations to pretraining corpora.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024c. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture.

Jindřich Libovický, Jindřich Helcl, Andrei Manea, and Gianluca Vico. 2025. Cus-qa: Local-knowledge-oriented open-ended question answering dataset.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. On the cultural gap in text-to-image generation. *arXiv preprint arXiv:2307.02971*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Milan Mayer. 2014. Gagauz people–their language and ethnic identity.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. *arXiv preprint arXiv:2501.04662*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1774–1784. ACM.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision.

Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2024. Benchmarking multimodal models for ukrainian language understanding across academic and cultural domains. *arXiv preprint arXiv:2411.14647*.

Ilona E Podolyan et al. 2005. How do ukrainians communicate?(observations based upon youth population of kyiv). *Journal of Intercultural Communication*, 5(2):1–14.

Ian Press and Stefan Pugh. 2015. *Ukrainian: A comprehensive grammar*. Routledge.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese. *arXiv preprint arXiv:2402.17302*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. Geode: a geographically diverse evaluation dataset for object recognition.

Suzanne Romaine. 2017. *Pidgin and creole languages*. Routledge.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Peter Rutland. 2023. Thirty years of nation-building in the post-soviet states. *Nationalities Papers*, 51(1):14–32.

Florian Schneider and Sunayana Sitaram. 2024. M5–a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *arXiv preprint arXiv:2407.03791*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Brian Silver. 1974. Social mobilization and the russification of soviet nationalities. *American Political Science Review*, 68(1):45–66.

Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. 2025. RusCode: Russian cultural code benchmark for text-to-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7641–7657, Albuquerque, New Mexico. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Benchmarking llm-based machine translation on cultural awareness. *arXiv preprint arXiv:2305.14328*.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge.

Round	Metric	RU	UK
Round 1	#Seeds	688	416
	#Extractions	6409	6462
	#Dishes	371	386
Round 2	#Seeds	371	386
	#Extractions	11 724	5442
	#Dishes	195	92
Round 3	#Seeds	195	92
	#Extractions	3429	3440
	#Dishes	44	179

Table 1: The **#Seeds** (initial dishes given to the algorithm), **#Extractions** (total algorithm extractions), and **#Dishes** (total dishes in the extractions) for each round of bootstrapping in Russian and Ukrainian.

A Bootstrapping Statistics

In Table 1, we list the number of seed dishes, the number of extracted potential dishes, and the number of dishes that were annotated as real dishes in the potential dishes list. The reported values can contain duplicates, and once all extractions were acquired from every round of bootstrapping, they were de-duplicated (based on edit distance and confirmed manually) before being added to BORSCH.

B Image Extraction Annotation Interface

We use a custom interface to choose which images to extract during the dataset creation step. Figure 17 shows this interface in use, as well as an example of why it is necessary; automatically pulling down the images shown in the interface would result in images of all word senses, not just the dish.

C Image Pretraining Data Contamination

Given our QA/VQA tasks, contamination becomes problematic when an image is directly captioned with a dish name or origin. We combed through the leading large image-caption dataset, `relaion2B-multi-research-safe` (Schuhmann et al., 2022), which contains 2 billion image/caption pairs in various languages, to find instances of BORSch dishes. To estimate contamination, we assume an upper bound where every text dish occurrence is linked to a BORSch image. We find that 46% of dishes have captions. In reality, the number of these occurences which are tied to BORSch images is most likely far lower. We further note that Ramaswamy et al. (2023) explored crowd-sourcing

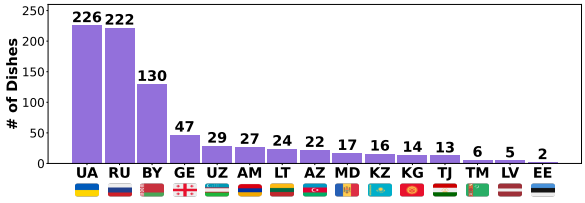


Figure 11: Countries of origin of dishes in the parallel, Post-Soviet sub-dataset of BORSCH. Dishes are heavily focused around Russia and Ukraine.

novel images for a cultural/geographical task, and found that each image cost $1.08 to ensure photographers and quality assurance annotators are fairly compensated for their time. With the 8192 images in BORSch, this would amount to around $8800 for the whole corpus, which is not feasible, so we fall back to collecting open source data.

D Parallel Corpus Origin Distribution

Figure 11 displays the origin distribution of the dishes in the parallel, Post-Soviet sub-dataset of BORSCH. We note that if a dish has multiple origins, it counts as a dish for each of those origins.

E Model Choice Justification

In our work, we conduct QA/VQA, as well as a novel description evaluation generation experiment, using the `Llama-3.1-70B-Instruct`, `Llama-3.2-90B-Vision-Instruct`, `Qwen2-72B-Instruct`, and `Qwen2-VL-72B-Instruct` models. We choose the `Llama-3.1` and `Qwen-2` model families as they have vision-enabled counterparts which are directly built on top of the original text-only models (Dubey et al., 2024; Yang et al., 2024; Wang et al., 2024), importantly allowing for a fair cross-modality comparison. Both `Qwen2-72B-Instruct` and `Llama-3.1-70B-Instruct` are 98th (1159 elo) and 68th (1220 elo) respectively in Russian LLMArena (Chiang et al., 2024), which is on the level of most capable multilingual models of that size. Both models score higher than `GPT4`, which has been shown to perform well on the Russian subset of MMLU (Achiam et al., 2023). In Ukrainian, there exists no popular LLM-arena equivalent to the best of our knowledge. Nonetheless both `Qwen-2` and `Llama-3.1` are used in existing studies evaluating Ukrainian NLP tasks, performing well (Kim et al., 2025; Paniv et al., 2024).

F Prompts For Each Experiment

Tables 5 and 6 contain the prompts used in the QA, VQA, VQA with dish (§4.1), dish name VQA (§4.2), and image generation experiments (§5) for Russian and Ukrainian respectively.

G Incorrect, Non-Post Soviet Predictions

In Figure 3, we construct confusion matrices of dish-origin model predictions, focusing on cases where both true and predicted countries are Post-Soviet. However, models can predict other countries for Post-Soviet originating dishes as well. Figure 12 shows the most common incorrect, non-Post-Soviet predictions for Post-Soviet dishes.

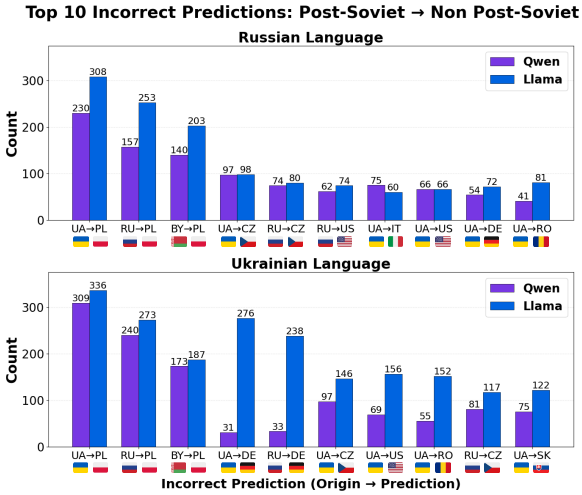


Figure 12: The top 10 most common incorrect, non-Post-Soviet predictions for BORSCH dishes that Llama and Qwen make when prompted in Russian and Ukrainian.

H Dish Origin VQA with Dish Name

In §4.1, we prompt text models for a dish's country of origin given its *name*, while at the same time prompting vision models for a dish's country of origin given its *images*. A logical continuation would be to give vision models both the dish's name **and** images. Figure 13 shows the overall results of this experiment, which exhibit very similar trends to the results in §4.1. Furthermore, Figure 14 shows that generally, adding the dish name to the VQA prompt improves performance for most dishes compared to just having the images. However, there are some dishes where adding the name actually harms performance.

I Dish Name VQA Results

We report the full results of the dish name VQA experiment introduced in Section 4.2. We provide

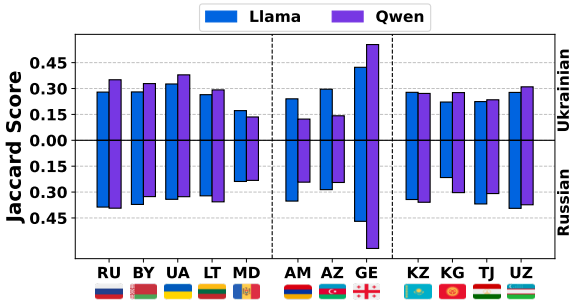


Figure 13: Both Llama and Qwen vision models can quite accurately predict the country of origin of Post-Soviet dishes when prompted with an image of the dish and its name. At the region and language level, trends do not change much from what is observed in §4.1.

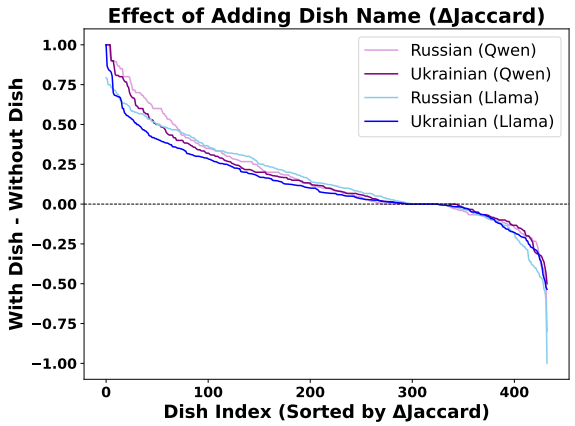


Figure 14: Adding the name of the dish to the country of origin VQA prompt increases model performance on Post-Soviet dishes in Russian and Ukrainian. However, there are a select few dishes where this is not the case.

a language model images of a dish and a prompt querying for the name of the dish displayed in the images (Appendix F). If the result contains the dish name within one edit distance, we consider this a success. We measure the success rate (accuracy) over five prompts to get the average **exact match accuracy**, and we report these results in Table 2.

J Dish Description Evaluation Pipeline

J.1 Additional Examples

We provide additional examples of dish descriptions (translated into English), images generated from these descriptions, ground truth images, and resulting cosine similarities in Figure 19.

J.2 Human Evaluation

We calculate more correlation metrics between annotator scores (reflecting how accurately a generated description matches the ground-truth image)

	Llama-3.2		Qwen2	
Country	RU	UK	RU	UK
Estonia	0.00	0.00	0.00	0.00
Latvia	0.12	0.16	0.20	0.20
Lithuania	0.06	0.09	0.13	0.11
Russia	0.11	0.07	0.12	0.10
Ukraine	0.07	0.07	0.10	0.08
Belarus	0.06	0.07	0.12	0.10
Moldova	0.05	0.05	0.04	0.05
Armenia	0.18	0.15	0.16	0.15
Azerbaijan	0.14	0.08	0.10	0.14
Georgia	0.17	0.16	0.17	0.09
Kazakhstan	0.15	0.14	0.10	0.16
Kyrgyzstan	0.10	0.10	0.11	0.11
Tajikistan	0.22	0.18	0.14	0.14
Turkmenistan	0.10	0.03	0.00	0.00
Uzbekistan	0.14	0.12	0.12	0.11

Table 2: For both Llama and Qwen vision models, providing a name of a dish given its image is a difficult task. While there are some country/region/language/model trends, the overall trend that applies everywhere is that performance is poor, rarely reaching above 20%.

	Llama-3.2		Qwen2	
	RU	UK	RU	UK
Pearson r	0.82	0.82	0.78	0.81
Spearman ρ	0.84	0.83	0.80	0.81
Kendall τ	0.68	0.67	0.63	0.65
Lin's ρ_c	0.79	0.76	0.74	0.79

Table 3: Correlations between human ratings and `DiNOv2` encoding cosine similarities of the ground-truth and `FLUX.1-dev` generated dish images. Using different metrics does not change the result much.

and cosine similarities of the `DiNOv2` encodings of the ground-truth and `FLUX.1-dev` generated dish images. The results can be found in Table 3. We also present a scatterplot of our human ratings and the encoding cosine similarities in Figure 15.

J.3 Llama QA Correlations

In Figure 9, we show how the cosine similarity from the image generation experiment correlates with all of our various QA tasks for the Qwen vision and text models. We show the same results, but for Llama vision and text models, in Figure 16

J.4 Set Similarity Metrics Ablation

One of our findings regarding the dish description evaluation experiment is that the produced cosine similarities have a small, positive Spearman correlation with QA/VQA Jaccard scores. To confirm these findings, we report in Table 4 the Pearson correlation r, Spearman correlation ρ, and Kendall τ for three different set similarity metrics:

$$\text{Jaccard Score} = \frac{|P \cap T|}{|P \cup T|}; \quad (1)$$

$$\text{Dice Coeff.} = \frac{2|P \cap T|}{|P| + |T|}; \quad (2)$$

$$\text{Overlap Coeff.} = \frac{|P \cap T|}{\min\big(|P|, |T|\big)}; \quad (3)$$

where T is the set of ground-truth countries and P is the predicated set of countries. We do not observe any noticeable deviation from our originally reported results.

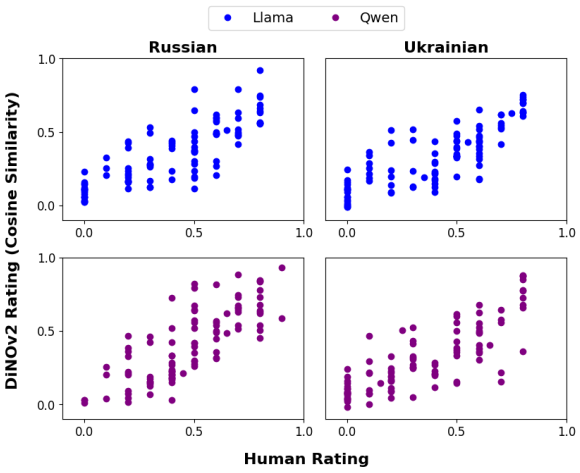


Figure 15: Scatterplot of human ratings vs `DiNOv2` encoding cosine similarities of the ground-truth and `FLUX.1-dev` generated dish images.

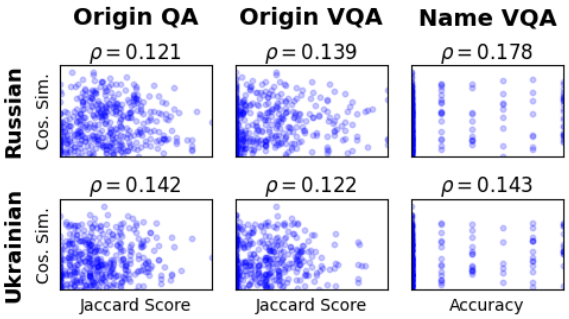


Figure 16: The Spearman's ρ between dish-description performance and QA tasks on **Llama** models shows a small, positive correlation. All axes span from 0 to 1.

K Annotator Instructions

The following were the instructions given to annotators in each task that required human evaluation/annotation. Annotators were fluent in the

	Llama-3.2		Qwen2	
	RU	UK	RU	UK
QA Jaccard r	0.08	0.12	0.09	0.12
QA Jaccard ρ	0.12	0.14	0.14	0.19
QA Jaccard τ	0.08	0.09	0.09	0.13
QA Dice r	0.11	0.15	0.13	0.15
QA Dice ρ	0.12	0.16	0.14	0.19
QA Dice τ	0.08	0.10	0.09	0.13
QA Overlap r	0.12	0.19	0.17	0.20
QA Overlap ρ	0.09	0.20	0.14	0.23
QA Overlap τ	0.06	0.14	0.10	0.16
VQA Jaccard r	0.09	0.06	0.16	0.15
VQA Jaccard ρ	0.14	0.12	0.21	0.20
VQA Jaccard τ	0.10	0.08	0.15	0.14
VQA Dice r	0.10	0.08	0.18	0.16
VQA Dice ρ	0.14	0.13	0.21	0.20
VQA Dice τ	0.09	0.09	0.15	0.14
VQA Overlap r	0.08	0.14	0.16	0.16
VQA Overlap ρ	0.11	0.17	0.18	0.19
VQA Overlap τ	0.08	0.12	0.13	0.13

Table 4: Correlation metrics (Pearson, Spearman, Kendall) between image generation cosine similarity (§5) and QA/VQA (§4.1) measured using three set overlap metrics: Jaccard score (used in the main paper), Dice coefficient, and Overlap coefficient.

necessary languages, college educated, paid a rate of $18 an hour, and recruited from the university. All annotators were fully informed of the study's aims and methods from the outset. We held frequent meetings to ensure annotator understanding of experiments their annotations were used in, and all annotators had the option to withdraw their contributions at any point if they wished.

K.1 Dish Filtering

Please label the following as dishes (T) or not dishes (F). A "dish" meets the following criteria

- Made up of multiple ingredients (e.g. *cucumber* does not meet this criteria).
- Culturally specific in some way (e.g. *grilled chicken* does not meet this criteria, as it is globally common).
- Something that people eat/is edible.
- Although drinks can meet these criteria, we exclude them.

If you are unsure for a certain dish, please mark/highlight it and we will discuss it.

K.2 Origin Annotation

Please label the countries which the given dish originates from (use the Alpha-2 country code of the country, which can be found at`https://www.iban.com/country-codes`). There can either be one or multiple origins for the dish. Try to find multiple sources corroborating your decisions; this is as much a research task as it is an annotation task. If you cannot find multiple corroborating sources or do not feel confident with your annotation, please mark the dish and we will exclude it.

K.3 Image Filtering

Please mark the image if it meets the following criteria:

- There is no food dish in the image.
- There are multiple food dishes in the image (a side, like bread, is fine as long as it is not the main focus).
- There are people in the image.
- The image is of poor quality (blurry, too small, etc.).
- There is text in the image.
- There is anything personally identifying in the image (documents, names, etc.).

K.4 Image Description Human Evaluation

Please rate how well the dish description visually matches the provided image of the same dish from 0 to 1. Please note that a good description is both accurate and concise. Just as an insufficient description should receive a poor score, a description that states a lot of extra, unnecessary information should also be penalized.

L Wikidata Query

Our food dish SPARQL query (Figure 18) retrieves information about food items, including their English labels and a list of country codes representing their countries of origin. The `?fid` variable identifies each food item, and `?food_en` provides its English name. The query uses `GROUP_CONCAT` to aggregate unique country codes (`?countryOfOriginCode`) for each food item into a single, comma-separated list (`?countryOfOriginList`). The filter condition ensures that only English labels are selected, while an `OPTIONAL` block allows for cases where a country of origin may not be specified, making that part of the data retrieval non-mandatory. Finally, the results are grouped by `?fid` and `?food_en` to return

distinct food items with their respective country origins. We only provide this English query (even though we also used two more queries for Russian and Ukrainian), as it is trivial to modify the query for Russian/Ukrainian by simply modifying the language code.

M Full Dataset Composition and Statistics

We give the detailed country of origin composition of the Russian and Ukrainian subsets of BORSCH in Tables 7 and 8 respectively. Note that if a dish has multiple countries of origin, this counts as one dish for each of those origins. For example, if a dish traces its origins to both Bulgaria and North Macedonia, then this would count as a dish for Bulgaria **and** as a dish for North Macedonia.

N Reproducibility and Hyperparameters

N.1 Compute Set-Up

- Experiments were run on eight A40 GPUs, and evaluation was distributed among them using huggingface.
- Per language, every non-vision enabled experiment would take around 8 GPU hours to run (1 hour across 8 GPUs).
- Per language, every vision-enabled experiment would take around 24 GPU hours to run (3 hours across 8 GPUs).
- Inference was conducted using vLLM (Kwon et al., 2023).

N.2 `Llama-3.1-70B-Instruct`

- https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.
- 70.6 billion parameters.
- do_sample = False, context_length = 4096, max_tokens = 200.
- We adhered to the license and intended use of this model (www.llama.com/llama3_1/license/).

N.3 `Llama-3.2-90B-Vision-Instruct`

- https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct.
- 88.6 billion parameters.
- do_sample = False, context_length = 8000, max_tokens = 200.
- We adhered to the license and intended use of this model (www.llama.com/llama3_2/license/).

N.4 `Qwen2-72B-Instruct`

- https://huggingface.co/Qwen/Qwen2-72B-Instruct.
- 72.7B billion parameters.
- do_sample = False, context_length = 4096, max_tokens = 200.
- We adhered to the license and intended use of this model (huggingface.co/Qwen/Qwen2-72B-Instruct/blob/main/LICENSE).

N.5 `Qwen2-VL-72B-Instruct`

- https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct.
- 73.4 billion parameters.
- do_sample = False, context_length = 8000, max_tokens = 200.
- We adhered to the license and intended use of this model (huggingface.co/Qwen/Qwen2-VL-72B-Instruct/blob/main/LICENSE).

N.6 `DiNOv2-giant`

- https://huggingface.co/facebook/dinov2-giant
- 1.14 billion parameters.
- All hyperparameters are default.
- We adhered to the license and intended use of this model (Apache License 2.0).

N.7 `FLUX.1-dev`

- https://huggingface.co/black-forest-labs/FLUX.1-dev
- 12 billion parameters.
- All hyperparameters are default.
- We adhered to the license and intended use of this model (github.com/black-forest-labs/flux/blob/main/model_licenses/LICENSE-FLUX1-dev).

N.8 `spaCy` NER

- **Russian:** `ru_core_news_sm`. More info at https://spacy.io/models/ru.
- **Ukrainian:** `uk_core_news_sm`. More info at https://spacy.io/models/uk.
- We adhered to the license and intended use of this model (MIT License).

O AI Assistants

We used ChatGPT for `GPT-4o` and `o1` as grammar/spell checkers.

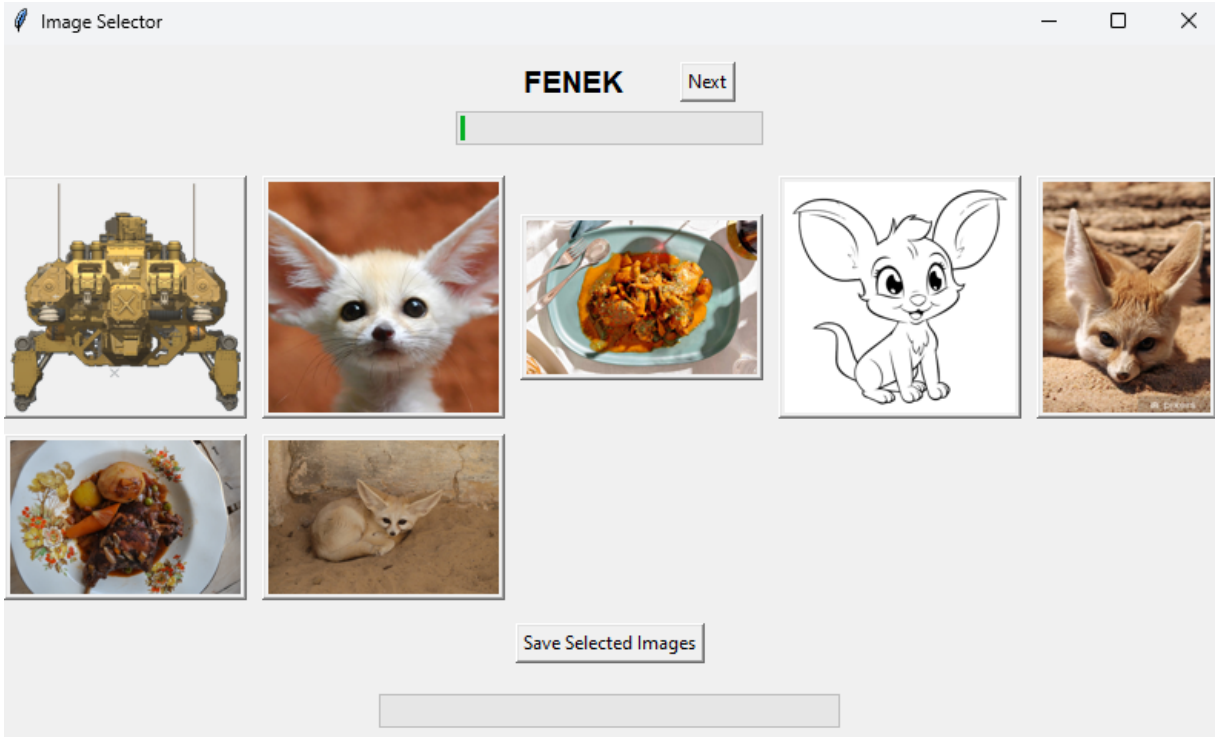


Figure 17: An example of the annotation interface for filtering the custom google image search results for dishes in BORSCH. As you can see, "fennek" has three word senses shown here. Some sort of robot, a fox, and a Maltese dish. Dishes like this necessitate filtering by a human annotator.

Wikidata SPARQL Query

```
SELECT ?fid ?food_en (GROUP_CONCAT(DISTINCT ?countryOfOriginCode; SEPARATOR = ", ")
AS ?countryOfOriginList) WHERE {
    ?fid (wdt:P31/(wdt:P279*)) wd:Q746549;
         rdfs:label ?food_en.
    FILTER((LANG(?food_en)) = "en")
    OPTIONAL {
        ?fid wdt:P495 ?countryOfOriginEntity.
        ?countryOfOriginEntity wdt:P297 ?countryOfOriginCode.
    }
}
GROUP BY ?fid ?food_en
```

Figure 18: The English SPARQL query used to retrieve Wikidata food items, both for the dataset as well as for the seedlist in our algorithm.

Russian Prompt Table	
Prompt Type	**Prompt**
QA	Из какой страны или стран происходит блюдо [DISH]? В какой стране или странах возникло блюдо [DISH]? В каких странах принято подавать блюдо [DISH]? Какие страны считаются родиной блюда [DISH]? В каких странах блюдо [DISH] популярно или традиционно?
VQA	Из какой страны или каких стран происходит блюдо на фотографиях? В какой стране или в каких странах возникло блюдо на фотографиях? В каких странах принято подавать блюдо на фотографиях? Какие страны считаются родиной блюда на фотографиях? В каких странах блюдо на фотографиях популярно или традиционно?
VQA + Dish	Из какой страны или каких стран происходит блюдо [DISH], показанное на фотографиях? В какой стране или в каких странах возникло блюдо [DISH], показанное на фотографиях? В каких странах принято подавать блюдо [DISH], показанное на фотографиях? Какие страны считаются родиной блюда [DISH], показанного на фотографиях? В каких странах блюдо [DISH], показанное на фотографиях, популярно или традиционно?
Dish Name VQA	Какие возможные названия могут быть у блюда на фотографиях? Какими именами может быть известно это блюдо на фотографиях? Как можно назвать блюдо, изображённое на фотографиях? Какие названия имеет блюдо на фотографиях? Какие имена имеет это блюдо на фотографиях?
Image Generation	Опишите блюдо [DISH], не используя его название.

Table 5: All prompts used for experiments in **Russian**. `[DISH]` is a template for the actual dish name.

Ukrainian Prompt Table	
Prompt Type	**Prompt**
QA	З якої країни або яких країн походить страва [DISH]? Які країни вважаються батьківщиною страви [DISH]? У яких країнах страва [DISH] є традиційною або популярною? У яких країнах готують страву [DISH]? Назви країну або країни, у яких їдять страву [DISH]?
VQA	З якої країни або яких країн походить страва на фотографіях? Які країни вважаються батьківщиною страви на фотографіях? У яких країнах страва на фотографіях є традиційною або популярною? У яких країнах готують страву на фотографіях? Назви країну або країни, у яких їдять страву на фотографіях?
VQA + Dish	З якої країни або яких країн походить страва [DISH] на фотографіях? Які країни вважаються батьківщиною страви [DISH] на фотографіях? У яких країнах страва [DISH] на фотографіях є традиційною або популярною? У яких країнах готують страву [DISH] на фотографіях? Назви країну або країни, у яких їдять страву [DISH] на фотографіях?
Dish Name VQA	Якими назвами може бути відома страва на фотографіях? Які назви можна використати для опису страви, зображеної на фотографіях? Під якими назвами може бути відома ця страва на фотографіях? Чи можна визначити можливі назви страви на фотографіях? Як прийнято називати цю страву, зображену на фотографіях, і які інші назви можуть бути для неї відомі?
Image Generation	Опиши страву [DISH], не використовуючи її назву.

Table 6: All prompts used for experiments in **Ukrainian**. `[DISH]` is a template for the actual dish name.

Dish Name	Lang.	Description	Generated image	Truth Image	Cos. sim.
БЛИНЫ	RU	This mouth-watering dish consists of a series of round, thin discs, usually between 15 and 20 centimeters in diameter. They range in color from pale yellow to golden brown, depending on how well they are cooked. Each circle has a slightly bubbly texture on one side and is smooth on the other. They are elastic enough to be rolled or folded, but still crispy around the edges. These circles are usually served in a stack, sometimes with a small cut in the center of each to reveal the filling, if any.			0.93
МАНТИ	UK	These are huge, soft bags that look like little pillows or wrapped in a thread, made of thin, elastic dough. They are usually white in color, but can have a slight yellow tint depending on the flour used. Their shape varies from perfectly round to slightly ellipsoidal, with soft, outer folds that open a window to their inner treasures. The inner filling is usually brown or dark green due to the use of meat and herbs. The filling has a thick texture that contrasts with the soft, melt-in-your-mouth dough.			0.87
ПЛЯЦОК	RU	This majestic dish is a large, round pie, like a flat disk, about 30 centimeters in diameter. Its top surface is covered with a golden-brown crust, which gets its color from baking in the oven for a long time. The crust has a slightly crisp texture, with small cracks that give it a unique character. Under the crust is a soft, moist filling, which can be either sweet or savory, depending on the recipe. If it is a sweet version, the filling can be light brown or creamy, with fruit or nuts in it. If it is a savory version, the filling can be greenish from herbs or brown from meat.			0.77
КЮФТА-БОЗБАШ	UK	It is a huge, deep bowl filled with a thick, multi-layered broth that ranges in hues from dark brown to golden brown. The broth is the base of this dish, thick and rich throughout, with small dots of rendered fat on the surface. In the bowl lie huge, soft balls made of meat filling. They are the size of a tube, which gives them the appearance of miniature dumplings. These balls have a luxurious brown color and a soft, uniform texture. Scattered around these meat balls are huge pieces of roots that look like small islands in a sea of broth. These pieces of roots are light brown in color and have a soft, delicate texture.			0.64
ЧАШУШУЛЯ	RU	This dish is a mouthwatering combination of vibrant colors and textures. The base is a deep red sea of tomato sauce, cooked to perfection. In it float chunks of tender meat, usually lamb or beef, that have acquired a deep brown color from long braising. The meat is framed by pieces of sweet green peppers and onions, giving the dish a fresh and crunchy texture. Scattered around the meat and vegetables are soft, almost melt-in-your-mouth tomatoes, which give the dish an extra juicy flavor.			0.57
ЕЧПОЧМАК	UK	This dish looks like a small, golden-brown crescent-shaped pie. Its crust is crispy on the outside, but tender and fluffy on the inside. It has a flat, smooth surface with several small cuts on the top, which allows steam to escape during baking. When you open this pie, you will see a generous filling consisting of meat, herbs and various seasonings. The filling is juicy and thick, with meaty pieces that melt in your mouth. This pie is usually served hot, and it is often accompanied by a sauce or salad.			0.43
ЗИРВАК	RU	This dish is a mouthwatering combination of vibrant colors and textures. The base is rich brown pieces of meat, slowly braised to perfect tenderness. The meat is covered in a thick, rich sauce, which gives it a deep, velvety color. Scattered around the meat are pieces of vegetables: golden slices of carrot, green peas, and deep red pieces of tomato. All these ingredients give the dish a picturesque look, like a painting on a plate.			0.36
САЛАТ ЛИСИЧКА	UK	This dish looks like a small, delicate sculpture that reflects the image of a fox. It is made of different layers of food that recreate different parts of the fox's body. The base of this dish is a large, round base that looks like the fox's body. It is made of a chicken fillet cut in half, which is steamed, which gives it a soft, tender texture. This layer is white in color, which contrasts with the other ingredients. The fox's head is made of half a tomato, which gives it a bright red color. Its legs are made of thinly sliced scallions, which gives the dish a little crunch. The fox's eyes are made of olive halves, which give the dish a deep, dark color			0.24
АФАРАР	RU	The dish is an appetizing composition, where the base is made up of delicate, slightly transparent petals, similar to thin sheets of tissue paper, but with a denser texture. They have slightly wavy edges and shimmer with a mother-of-pearl shade, turning into a pale pearl. In the center of these petals, small pieces of white meat are carefully placed, which resembles fresh cottage cheese in color and texture, but with a more elastic structure. The meat is seasoned with herbs, which give the dish bright green accents, and finely chopped vegetables, giving it a slightly crunchy texture.			0.13

Figure 19: Examples of the image description evaluation pipeline described in §5.

Country Name	Count	Country Name	Count	Country Name	Count
Russian Federation	203	Lebanon	10	Nigeria	2
Ukraine	112	Belgium	9	Bosnia and Herzegovina	2
France	98	Bulgaria	9	Bolivia	2
Belarus	87	Finland	8	Ireland	2
Germany	79	Palestine	7	Iceland	2
Italy	78	Denmark	7	Australia	2
United States	52	Argentina	7	Nepal	2
Türkiye	48	Tunisia	7	Myanmar	2
Japan	42	Egypt	6	Afghanistan	2
Georgia	42	Netherlands	6	Malta	2
Spain	37	Brazil	6	Colombia	2
Poland	33	Switzerland	6	Malaysia	2
Indonesia	31	Serbia	6	Mauritania	2
China	26	Czechia	6	Senegal	1
India	25	Norway	6	South Africa	1
Armenia	25	Philippines	5	Ecuador	1
United Kingdom	24	Croatia	5	Kenya	1
Greece	22	Algeria	5	Bahrain	1
Austria	22	Jordan	5	Bangladesh	1
Korea, Republic of	22	Turkmenistan	5	Slovenia	1
Mexico	19	Viet Nam	4	Estonia	1
Uzbekistan	19	Slovakia	4	Cuba	1
Hungary	17	Uruguay	4	Oman	1
Azerbaijan	17	Pakistan	4	Yemen	1
Lithuania	16	North Macedonia	4	Eritrea	1
Kazakhstan	15	Iraq	4	Congo	1
Syria	14	Peru	3	Gabon	1
Romania	13	Canada	3	Dominican Republic	1
Portugal	13	Taiwan	3	Mali	1
Kyrgyzstan	13	Bhutan	3	Venezuela	1
Moldova, Republic of	12	Haiti	3	Thailand	1
Tajikistan	11	Albania	3	Cambodia	1
Iran	11	Libya	3	Cyprus	1
Sweden	10	Paraguay	3	Unknown	1
Israel	10	New Zealand	3	Latvia	1
Morocco	10	Mongolia	3	Singapore	1
Lebanon	10	Ghana	2	Brunei Darussalam	1

Table 7: Countries of origin in the **Russian** subset of BORSCH.

Country Name	Count	Country Name	Count	Country Name	Count
Ukraine	166	Austria	8	Bolivia	2
Russian Federation	157	Switzerland	8	Peru	2
Belarus	100	Kyrgyzstan	8	South Africa	2
France	59	Sweden	8	Argentina	2
Italy	47	Iran	8	Iceland	2
Poland	36	Belgium	7	Iraq	2
Türkiye	34	Egypt	7	Philippines	2
Georgia	33	Czechia	7	Afghanistan	2
United States	30	Croatia	6	Paraguay	2
India	29	Latvia	6	Sri Lanka	2
Germany	27	Palestine	6	Thailand	2
Japan	27	Malaysia	5	Turkmenistan	2
Indonesia	24	Bosnia and Herzegovina	5	Singapore	2
Lithuania	20	Tunisia	5	Ecuador	1
Uzbekistan	19	Denmark	5	Senegal	1
Greece	19	Brazil	5	Niger	1
Armenia	18	Nepal	5	Kenya	1
Spain	18	Jordan	5	Bahrain	1
Azerbaijan	17	Slovakia	4	Myanmar	1
United Kingdom	16	Pakistan	4	Benin	1
Bulgaria	16	Israel	4	Chile	1
Hungary	15	Bhutan	4	Trinidad and Tobago	1
China	14	Albania	4	Haiti	1
Mexico	14	Libya	4	Monaco	1
Romania	14	Norway	4	Sudan	1
Syria	11	Taiwan	3	Saudi Arabia	1
Nigeria	11	Ghana	3	Yemen	1
Portugal	11	Viet Nam	3	Congo, The Democratic Republic of the	1
Tajikistan	10	Finland	3	Angola	1
Serbia	10	Morocco	3	Venezuela	1
Moldova, Republic of	10	Uruguay	3	Canada	1
Korea, Republic of	10	Slovenia	3	Cyprus	1
Netherlands	9	Algeria	3	Mauritania	1
Kazakhstan	9	Estonia	3	Unknown	1
Lebanon	9	Mongolia	3	Australia	1
Austria	8	North Macedonia	3	Montenegro	1

Table 8: Countries of origin in the **Ukrainian** subset of BORSCH.