Benchmarking Multimodal Models for Ukrainian Language Understanding Across Academic and Cultural Domains

Yurii Paniv
Ukrainian Catholic University
paniv@ucu.edu.ua

Artur Kiulian
OpenBabylon
akiulian@gmail.com

Dmytro Chaplynskyi
lang-uk initiative
chaplinsky.dmitry@gmail.com

Mykola Khandoga
OpenBabylon
mkhandoga@gmail.com

Anton Polishko
OpenBabylon
anton.polishko@gmail.com

Tetiana Bas
Minerva University
tetiana@uni.minerva.edu

Guillermo Gabrielli
OpenBabylon
guillermo.gabrielli.fer@gmail.com

Abstract

While the evaluation of multimodal English-centric models is an active area of research with numerous benchmarks, there is a profound lack of benchmarks or evaluation suites for low- and mid-resource languages. We introduce ZNO-Vision, a comprehensive multimodal Ukrainian-centric benchmark derived from the standardized university entrance examination (ZNO). The benchmark consists of over 4300 expert-crafted questions spanning 12 academic disciplines, including mathematics, physics, chemistry, and humanities. We evaluated the performance of both open-source models and API providers, finding that only a handful of models performed above baseline. Alongside the new benchmark, we performed the first evaluation study of multimodal text generation for the Ukrainian language: we measured caption generation quality on the Multi30K-UK dataset. Lastly, we tested a few models from a cultural perspective on knowledge of national cuisine. We believe our work will advance multimodal generation capabilities for the Ukrainian language and our approach could be useful for other low-resource languages.

1 Introduction

Vision-language models (VLMs) have expanded LLM capabilities into more domains, allowing for models to work with plenty of new tasks such as OCR (Liu et al., 2024), image captioning, visual question answering and many more.

While numerous benchmarks (Li et al., 2024) evaluate VLMs performance across a range of multimodal tasks, these resources primarily serve English-language models, underscoring a critical gap for evaluating VLMs in less-resourced languages. This absence is especially pronounced for Ukrainian, where multimodal benchmarks are exceedingly scarce.

Our work addresses this gap by introducing a suite of Ukrainian-specific benchmarks and presenting benchmarking results for leading proprietary and open-source VLMs. To estimate academic knowledge, we developed a new benchmark based on the External Independent Evaluation (ZNO) - national university entrance and teacher certification exam (ZNO, 2024), which includes a large selection of questions across various fields, such as chemistry, mathematics, Ukrainian language and literature, etc. Besides that, we evaluated all models using Multi30K-UK (Saichyshyna et al., 2023), one of the few existing Ukrainian multimodal benchmarks. Additionally, for the culture test, we developed a new multimodal benchmark, UACUISINE, based on 20 popular Ukrainian dishes.

We believe that our effort would advance the development of VLMs applications for the Ukrainian language across academic and business sectors worldwide, wherever it's being used.

Code, evaluation scripts, and datasets are available at this link: https://github.com/lang-uk/mmzno-benchmark.

2 Related Work

Recent years have seen significant development in multimodal benchmarks for evaluating VLMs. Existing benchmarks can be broadly categorized into three groups. General visual understanding benchmarks include VQA (Antol et al., 2015) (1M+ question-answer pairs), GQA (Ainslie et al., 2023) (compositional reasoning), and MMMU (Yue et al., 2024) (broad domain reasoning). Cultural and multilingual benchmarks are represented by CulturalVQA (Nayak et al., 2024) (11 countries), WorldCuisines (Winata et al., 2024) (30 languages), and MaXM (Changpinyo et al., 2023) (7 languages). Visual reasoning benchmarks feature CLEVR (Johnson et al., 2016) (compositional reasoning), A-OKVQA (Schwenk et al., 2022) (external knowledge), and Visual7W (Zhu et al., 2016) (semantic

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 14–26

understanding).

While these benchmarks provide comprehensive evaluation frameworks, they predominantly focus on English language capabilities. Recent multilingual benchmarks often rely on translations rather than culturally-grounded content, highlighting a critical gap for evaluating VLMs in underrepresented languages like Ukrainian. Translation-based benchmarks like xGQA (Pfeiffer et al., 2022) (9,670 questions in 7 languages) often introduce artifacts and fail to capture cultural nuances (Park et al., 2024). Current cultural evaluations are either too limited in scope (CulturalVQA: 2,378 questions across 11 countries) or too narrow in focus (WorldCuisines: food-specific across 30 languages).

Analyzing the WorldCuisines, we found three critical limitations regarding Ukrainian cuisine: (1) representation was restricted solely to location identification tasks without deeper cultural assessment, (2) the selection of dishes failed to capture the breadth of Ukrainian culinary traditions, and (3) several dishes were incorrectly categorized as Ukrainian while featuring Russian-language captions and representing Russian cuisine variants.

2.1 Ukrainian Multimodal Benchmarks

As it has been mentioned in the introduction, the Ukrainian benchmarks for multimodal LLMs are scarce. This subsection describes what's available to the best of our knowledge.

M5 (Schneider and Sitaram, 2024): a multilingual benchmark that includes 41 languages and 5 different MLLM tasks. However, it is only the image captioning that actually spans over the 41 languages and includes Ukrainian. The image captioning dataset contains 143600 questions. M5 employs professional annotators to ensure high-quality annotations across all languages.

ALM(Vayani et al., 2024) benchmark consists of diverse 22763 VQA questions, translated into 100 languages using machine translation and then edited by native speakers of the corresponding languages.

Both M5 and AML benchmarks fulfill an important task of expanding the linguistic diversity of multimodal large language model (MLLM) benchmarks. However, as their focus is on broad multilingual coverage, they naturally lack specificity in evaluating Ukrainian multimodal capabilities.

The Ukrainian Visual Word Sense Disambiguation Benchmark (Laba et al., 2024) is designed to evaluate the ability of multimodal language models to resolve visual word sense ambiguity in Ukrainian, particularly with homonyms. The task requires selecting the correct meaning of an ambiguous word from a set of images, highlighting challenges related to low-resource languages, hallucinations, and representation gaps. Results show that multilingual retrieval models struggle with Ukrainian, often retrieving images corresponding to the more frequent meaning of a homonym instead of the intended one. Additionally, image generation models exhibit similar biases, defaulting to dominant meanings rather than reasoning through context. The benchmark reveals a significant performance gap between Ukrainian and English multimodal understanding, underlining the need for language-specific retrieval fine-tuning and better alignment of multilingual embeddings.

The Multi30K-UK benchmark (Saichyshyna et al., 2023) is an adaptation of the Multi30K dataset (Elliott et al., 2016) for Ukrainian, created via a combination of machine translation and human editing. It is primarily designed for image captioning and machine translation.

3 Datasets & Methodology

ZNO multi-choice questions. External Independent Evaluation (abbr. "ZNO" in Ukrainian) is a national Ukrainian test for high school graduates (ZNO, 2024). This test is challenging for LLMs even in a text-only setting (Romanyshyn et al., 2024). We gathered questions from the Osvita portal (Osvita, 2024), where an image is required for the answer. The dataset consists of 4306 question-pairs in 13 categories (overview in Appendix B): Math, Geography, Ukrainian language and literature, Teaching, History, Spanish, German, French, English, Chemistry, Physics, Biology, and Other (for a small portion of unclassified questions). From our source dataset, we filtered out questions with multiple images, images as answers,

Subset	# Questions	Visual-Only	Visual %
Dev	491	235	47.86%
Validation	490	233	47.55%
Test	3325	1864	56.06%
Total	4306	2332	54.16%

Table 1: Distribution of ZNO Dataset by subset. The Dev and Validation subsets each represent 10% of all data that can be used during model training.

Category	Total	Visual-Only	Visual-Only %
Chemistry	1021	946	92.65%
Mathematics	821	771	93.91%
Physics	661	595	90.02%
History	434	0	0.00%
Geography	374	0	0.00%
Biology	332	0	0.00%
English language	204	0	0.00%
French language	199	0	0.00%
Kindergarten teaching	134	0	0.00%
Ukrainian language and literature	56	0	0.00%
Other	31	0	0.00%
Spanish language	22	20	90.91%
German language	17	0	0.00%

Table 2: Distribution of ZNO questions by category. As we can see, STEM categories represent more than half of the dataset, even having more than 90% of all visual-only questions (a typical question has a text and image, but in a visual-only setting, the model has to perform OCR to answer the question).

and choice-matching questions to streamline the benchmark setting, leaving only questions that require a single letter (e.g., B) as an answer.

Multi30K-UK. We evaluated models for the caption generation task on the Multi30K-UK dataset. We use Flickr2017 and Flickr2018 datasets as dev and test subsets, respectively.

UACUISINE Benchmark. In this dataset, we addressed the issues with the WorldCuisine dataset mentioned in section 2. The UACUISINE benchmark consists of seven question types across three categories: (1) dish identification (three variants), (2) text generation (ingredients and recipe), and (3) characteristic classification (temperature and taste). The identification questions were adapted from WorldCuisines and translated into Ukrainian, while preparation and classification questions were newly introduced to assess deeper culinary understanding. We curated a dataset of 20 most typical Ukrainian dishes and annotated each with 7 question types in Ukrainian, generating 140 question-answer pairs.

Evaluation Framework. We adapted our benchmarks to the lmms-eval framework (Zhang et al., 2024) to reuse correct implementations of Vision-Language model inference, where the format of the prompt and image processing differs from model to model.

For the ZNO benchmark, the model is given an image and a natural language question about the image. The expected answer is a letter, e.g., A/B/C/D. Options consist of Ukrainian letters, except for English, Spanish, German, and French tests. The dataset contains 491/490/3325 (dev/validation/test) samples, each comprising an image, a question, and multi-choice answers encoded as letters.

The 10/10/80 dev/val/test split follows the MMMU (Yue et al., 2024) paradigm, where the dev set is used for few-shot in-context learning, the validation set is employed for hyperparameter tuning and prompt optimization, and the test set, which constitutes the majority of the data, is reserved for benchmarking. 54% of questions are pure visual questions to test OCR capabilities for models.

For benchmarking, similar to MMMU, we provided the same setting for all models by adding the same suffix prompt to all questions. We selected our prompt based on average performance across different models on the dev set of the benchmark, making benchmarking standardized across a diverse set of open-source and proprietary models.

For easier answer extraction, we experimented with direct prompt instructions to output answer in a specific format, such as дай відповідь на питання і напиши варіант відповіді в квадратних дужках, наприклад: "[A]" (answer this question and write the answer in quadratic braces, for example "[A]"). In our experiments, models struggled with specific format instructions, so we removed references to format and relied on a set of rules to extract a correct answer from the defined selection of options. As a result of our prompt tuning, the resulting prompt is Дай відповідь буквою-варіантом відповіді з наданих варіантів. (Answer by choosing the letter option

Model Name	ZNO Val	ZNO Test
anthropic/claude-3.7-sonnet	0.75	0.72
google/gemini-2.5-pro-preview-03-25	0.64	0.69
openai/gpt-4o	0.62	0.63
qwen/qwen2.5-vl-7b-instruct	0.54	0.56
meta-llama/llama-4-maverick	0.53	0.53
qwen/qwen-2.5-vl-72b-instruct	0.51	0.52
meta-llama/llama-4-scout	0.48	0.49
qwen/qwen2.5-vl-3b-instruct	0.44	0.40
qwen/qwen2-vl-7b-instruct	0.42	0.39
google/gemma-3-27b-it	0.42	0.38
google/gemma-3-12b-it	0.41	0.39
qwen/qwen2.5-vl-32b-instruct	0.36	0.33
meta-llama/llama-3.2-90b-vision-instruct	0.35	0.33
mistral-community/pixtral-12b	0.31	0.31
qwen/qwen2-vl-2b-Instruct	0.30	0.31
cohereforai/aya-vision-8b	0.29	0.31

Table 3: Accuracy scores on ZNO dataset across different models for validation and test subdatasets. The bottom part of the table contains models for which the results are approximately the same as for text-only measurement (meaning it's the same as a random guess). Claude 3.7 Sonnet shows the strongest performance across all models, while Qwen2.5-VL-7B-Instruct and meta-llama/llama-4-maverick are the best open-source models for this particular task. More detailed breakdown by category could be found in Appendix A.

from the provided options).

Evaluation for UACUISINE consists of three metrics. For the dish name prediction and characteristic classification, we use exact match score (EM) - the specific dish name should be present in the resulting output. For the ingredients generation, we use a matching score called the Intersection Match (IM). We calculate IM by calculating the percentage of dish ingredients mentioned in the resulting output.

For recipe generation evaluation, we use BERT score (Zhang et al., 2020) using "bert-base-multilingual-cased" model (Devlin et al., 2018) (which is a default choice for Ukrainian in the reference implementation) to capture semantic similarity.

For Multi30K-UK, we use SacreBLEU (Post, 2018) and the same BERT score as well. We prepend every request with a prompt "Опиши зображення одним реченням." (Describe image in one sentence).

4 Experimental Setup

For each benchmark evaluation, we used their specific metrics with fixed random seeds for Python, NumPy, and Torch. For ZNO, we used a temperature of 1 and a maximum output tokens equal to 1024; we noticed that proprietary models produced many tokens before generating an answer. For Multi30k, we adopted Flickr30k (Young et al., 2014) evaluation methodology as presented in the lmms-eval framework for standardized multimodal evaluation (Zhang et al., 2024), employing temperature of 0 and a maximum output tokens equal to 64. For the UACUISINE benchmark, we employed the same temperature of 1 and a maximum output tokens equal to 512.

We evaluated both proprietary and open-source multimodal language models to provide a comprehensive assessment of current capabilities on Ukrainian language tasks. Besides standard setting, we measured the same question without images provided in the text-only setting to measure contamination. The lowest theoretical baseline evaluation of ZNO is to select the first choice in each question, getting a 22% accuracy score. As part of a benchmark, the model falls back on a randomly chosen answer from options if it fails to provide an answer. That's why, as a baseline, we evaluated all models in a text-only setting without images provided and treated similar scores in both settings as failing to beat a baseline. Most text-only evaluations score approximately 34%, values close to what we treat as a baseline.

Model Name	Multi30k 2017		Multi30k 2018	
	BERT	BLEU	BERT	BLEU
openai/gpt-4o	0.74	3.54	0.74	3.39
meta-llama/llama-4-scout	0.72	1.82	0.72	1.68
anthropic/claude-3.7-sonnet	0.71	1.40	0.72	1.78
meta-llama/llama-4-maverick	0.71	1.82	0.71	1.85
meta-llama/llama-3.2-90b-vision-instruct	0.71	1.96	0.71	2.03
mistral-community/pixtral-12b	0.71	1.48	0.71	1.97
qwen/qwen2.5-vl-7b-instruct	0.71	1.37	0.71	1.49
google/gemma-3-12b-it	0.71	1.53	0.71	1.77
google/gemma-3-27b-it	0.70	1.61	0.71	1.65
qwen/qwen2-vl-7b-instruct	0.70	0.89	0.70	1.08
qwen/qwen2.5-vl-32b-instruct	0.69	1.19	0.70	1.23
qwen/qwen2.5-vl-3b-instruct	0.69	0.61	0.69	0.19
qwen/qwen2-vl-2b-instruct	0.68	0.17	0.68	0.21
cohereforai/aya-vision-8b	0.65	0.64	0.66	0.62
qwen/qwen-2.5-vl-72b-instruct	0.32	1.86	0.59	1.51
google/gemini-2.5-pro-preview-03-25*	0.00	0.00	0.00	0.00

Table 4: Average SacreBLEU and BERT scores on the Multi30k-UA dataset. As we can see with the SacreBLEU score, there is a great difference between reference captions and generated captions (we provide examples of references and generation in Appendix D). The best performing is GPT-4o, as shown by both BERT Score and SacreBLEU in particular, indicating that those texts are closer to the benchmarked target domain of texts. Nevertheless, most of the models provide good enough captions to capture what's happening in the image. Qwen models tend to generate long descriptions even if prompted to provide them in short sentences, resulting in low scores for Qwen2.5-VL-72B-Instruct model. Gemini 2.5-pro-preview-03-25 refused to generate captions for provided images with a standard prompt.

5 Results & Discussion

In Table 3, Gemini 2.5 Pro (Georgiev et al., 2024), Claude 3.7 Sonnet (Anthropic, 2024), and GPT-4o (OpenAI et al., 2024) demonstrated the best results on ZNO benchmark, with Qwen2.5-VL-7B (Wang et al., 2024) being the strongest open source model alongside LLama 4 Maverick (Meta, 2025). Surprisingly, LLaMA 3.2 (Dubey et al., 2024) and Pixtral (Agrawal et al., 2024) failed to even beat a baseline. Even though Paligemma-3B-mix-224 (Beyer et al., 2024) showed some promising performance on caption generation, we didn't include it in our final evaluation because it is a base model. It was not tuned to provide output in a closed caption test setting. The detailed breakdown of the model's performance per question category is provided in Appendix A.

As for the UACUISINE benchmark, the leaderboard is close to a ZNO one, except for Gemini 2.5 Pro, which failed to generate recipes and dish ingredients.

As shown in Table 4, testing the caption generation task on the Multi30K-Uk dataset did not provide a way to evaluate model performance confidently. There are a couple of reasons for that: 1) the target domain is too different (the model frequently used synonyms, which affects the SacreBLEU score), 2) the model did not follow the instructions to provide an answer in one sentence only, 3) confidence that BERT Score model is a good fit to measure semantic similarity in Ukrainian. As we show in Appendix D, models generate captions correctly, but describe different details, making direct string comparison difficult.

While the former factor is a limitation of our work, the latter is a manifestation of cultural and linguistic bias by the models.

Instruction-following Issues. The most prevalent challenge observed across models was inconsistent instruction following in Ukrainian. Even high-performing models like GPT-4o and Gemini frequently failed to respond in the expected format. A notable case is a meta-llama/llama-3.2-90b-vision-instruct, which, instead of Ukrainian letters for answers, responds in English ones. We have observed models replying in a much more

Model Name	BERT Score	Exact Match (EM)	Intersection Match (IM)
google/gemma-3-27b-it	0.71	0.00	0.69
cohereforai/aya-vision-8b	0.70	0.00	0.49
anthropic/claude-3.7-sonnet	0.69	0.25	0.73
meta-llama/llama-4-scout	0.68	0.08	0.53
google/gemma-3-12b-it	0.67	0.03	0.69
openai/gpt-4o	0.67	0.00	0.73
meta-llama/llama-3.2-90b-vision-instruct	0.65	0.00	0.43
qwen/qwen-2.5-vl-72b-instruct	0.65	0.19	0.44
qwen/qwen2.5-vl-32b-instruct	0.65	0.15	0.40
qwen/qwen2.5-vl-7b-instruct	0.65	0.21	0.11
meta-llama/llama-4-maverick	0.63	0.11	0.69
qwen/qwen2.5-vl-3b-instruct	0.58	0.21	0.14
qwen/qwen2-vl-2b-instruct	0.00	0.23	0.01
google/gemini-2.5-pro-preview-03-25*	0.00	0.35	0.01

Table 5: UACUISINE Evaluation Metrics Across Models. The best model overall is Claude 3.7 Sonnet, having high scores across all categories. Unfortunately, no model scored high on the simple task of naming a dish in the photo, with only a Gemini scoring a 35% of right answers. Across open source models, LLama 4 Maverick and Gemma-27B-it are the strongest ones. Gemini refused to generate a recipe and name ingredients.

verbose way than expected by Multi30K, therefore we modified prompts with an extra instruction to reply with a sentence for Multi30K, but issue persisted.

Code-switching issues. Besides instruction following, we've observed major issues with code-switching and language confusion. This behavior was particularly pronounced in open-ended tasks like recipe generation and ingredient listing in the UACUISINE benchmark, where models would be prompted in Ukrainian but switch to English, Chinese or Russian for response. This suggests that current VLMs experience the same code-switching issues that are known to happen in text-only multilingual LLMs (Kiulian et al., 2024). We have also observed the same issues of broken grammar "Куряче супу з лапшой"(chicken soup with spaghetti), non-existing words generation (Хаширо-ітамэ, Курицики, Кулібино) and tokenization artifacts (Рисotto "Risotto").

Cultural misattribution. A key issue was cultural appropriation, notably when Ukrainian Borsch (UNESCO-recognized cultural heritage (UNESCO, 2022)) was mislabeled as "Russian Red Borscht." This pattern extended to other Ukrainian dishes, with models defaulting to English or Russian translations even when prompted in Ukrainian. The misattribution went beyond labeling - in recipe generation, models often suggested Russian rather than traditional Ukrainian preparations. This systematic bias points to training data issues that risk reinforcing narratives diminishing Ukrainian cultural identity. Addressing this requires both improved Ukrainian language capabilities and better integration of accurate cultural knowledge in model training.

6 Conclusions

In this work, we introduced a suite of benchmarks to evaluate VLMs in Ukrainian, addressing a critical gap in resources for low- and mid-resource languages. ZNO benchmark enables researchers to estimate model performance objectively for Ukrainian multimodal generation using expert-made questions. To the best of our knowledge, there were no prior public evaluations of the Multi30k-Uk dataset for the caption generation task, which we hope will be useful for other researchers in estimating model proficiency for Ukrainian in a multimodal setting. As for UACUISINE, we hope to highlight cultural issues in Vision-Language Models with our research, providing a framework to measure them objectively in a Ukrainian-specific setting. Future research directions should focus on extending benchmarks to include more diverse, language-specific tasks, addressing the culture gap. Beyond Ukrainian, the methodologies introduced here could serve as a template for advancing multi-

modal language modelling in underrepresented languages, enabling more inclusive access to AI instruments.

7 Limitations

While we believe that our work is a step forward in evaluating the Ukrainian capabilities of VLMs, it has a number of limitations.

We have used the same prompt prefix for all queries to all the tested models. This prompt might introduce a bias in model comparison. We haven't evaluated in the chain of thought setting and reasoning models like o3 from OpenAI. We noticed that proprietary models are more likely to generate more than the maximum allowed 1024 tokens in the answer, which could impact the evaluation. We plan to address it in future work.

The ZNO dataset is heavily skewed towards STEM domains, having more than half of the questions in these categories. Also, STEM categories have the most visual-only questions (meaning that the model has to rely on its OCR capabilities to answer the question).

Multi30K provides both English and Ukrainian captions for the same image, which makes it suitable for testing a multi-modal translation task. We haven't performed such testing.

We rely on the "bert-base-multilingual-cased" model for BERT Score calculation, as it is a default choice for the BERT Score metric for the Ukrainian language (Zhang et al., 2020). We used several other top models for Ukrainian in the Retrieval task based on the MMTEB benchmark (Enevoldsen et al., 2025), but haven't found any meaningful differences in resulting scores against the default choice. This emphasizes the necessity for standardized metrics to evaluate semantic similarity model performance in the Ukrainian language.

8 Ethical Considerations

Several ethical considerations arise in developing and deploying multimodal benchmarks for Ukrainian language evaluation. Most critically, our work addresses questions of cultural representation and identity preservation, particularly salient given current geopolitical contexts. The systematic misattribution of Ukrainian cultural elements by AI models highlights risks of technological erasure of cultural identity. While our use of translated benchmarks enables comparative evaluation, this approach may inadvertently perpetuate biases and fail to capture uniquely Ukrainian contexts. Additionally, the observed tendency of models to default to Russian or English translations, even when prompted in Ukrainian, raises concerns about digital marginalization of Ukrainian language users. These considerations underscore the importance of developing culturally sensitive evaluation frameworks that can help ensure AI systems properly represent and serve Ukrainian language users.

Acknowledgments

This research was made possible through generous support from several organizations. We thank **The alliance of De Novo and MK-Consulting** for providing computational resources, and **ELEKS** for their grant in memory of Oleksiy Skrypnyk.

We also gratefully acknowledge **Amazon Web Services (AWS)** for cloud credits that enabled training and inference on H200 instances, and **Google Cloud Platform (GCP)** for credits supporting model training and inference.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Preprint*, arXiv:2305.13245.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Anthropic. 2024. Claude 3.5 and claude with code interpreter. Accessed: 2024-11-21.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby,

Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V. Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. *Preprint*, arXiv:2209.05401.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *Preprint*, arXiv:1605.00459.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, and 1115 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Preprint*, arXiv:1612.06890.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the ukrainian language representation. *Preprint*, arXiv:2404.09138.

Yurii Laba, Yaryna Mohytych, Ivanna Rohulia, Halyna Kyryleyza, Hanna Dydyk-Meush, Oles Dobosevych, and Rostyslav Hryniv. 2024. Ukrainian visual word sense disambiguation benchmark. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 61–66, Torino, Italia. ELRA and ICCL.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024. A survey on benchmarks of multimodal large language models. *Preprint*, arXiv:2408.08632.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Preprint*, arXiv:2305.07895.

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation — ai.meta.com. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. [Accessed 05-04-2025].

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *Preprint*, arXiv:2407.10920.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Osvita. 2024. Osvita.ua test portal. Accessed: 2024-11-03.

ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering. *Preprint*, arXiv:2406.02331.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. *Preprint*, arXiv:2109.06082.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

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. *Preprint*, arXiv:2407.03791.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

UNESCO. 2022. UNESCO - Culture of Ukrainian borscht cooking — ich.unesco.org. https://ich.unesco.org/en/USL/culture-of-ukrainian-borscht-cooking-01852. [Accessed 21-11-2024].

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *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, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *Preprint*, arXiv:2410.12705.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. *Preprint*, arXiv:1511.03416.

ZNO. 2024. External independent evaluation. Accessed: 2024-11-03.

A ZNO Test Set Evaluation, Breakdown by Category

Model	Ukrainian*	History	English	French	German	Spanish	Teaching	Other
anthropic/claude-3.7-sonnet	66.67	70.69	87.20	86.25	60.00	85.71	79.55	60.00
cohereforai/aya-vision-8b	38.89	31.61	68.90	66.25	40.00	28.57	47.73	50.00
google/gemini-2.5-pro-preview-03-25	83.33	64.37	84.15	88.12	60.00	85.71	63.64	60.00
google/gemma-3-12b-it	66.67	48.85	79.88	70.00	40.00	71.43	47.73	70.00
google/gemma-3-27b-it	72.22	49.14	78.66	75.00	80.00	85.71	36.36	40.00
google/gemma-3-4b-it	44.44	28.74	54.88	56.88	60.00	28.57	27.27	20.00
openai/gpt-4o	83.33	67.53	89.02	80.00	60.00	100.00	63.64	60.00
meta-llama/llama-3.2-90b-vision-instruct	55.56	44.25	56.71	60.62	20.00	28.57	43.18	30.00
meta-llama/llama-4-maverick	50.00	58.33	83.54	76.88	60.00	85.71	63.64	50.00
meta-llama/llama-4-scout	50.00	46.84	81.71	78.12	60.00	28.57	43.18	30.00
mistral-community/pixtral-12b	38.89	36.49	65.24	61.25	20.00	28.57	36.36	20.00
qwen/qwen-2.5-vl-72b-instruct	50.00	54.60	62.80	32.50	20.00	42.86	59.09	40.00
qwen/qwen2-vl-2b-instruct	22.22	31.03	76.22	71.88	20.00	71.43	34.09	50.00
qwen/qwen2-vl-7b-instruct	55.56	42.24	85.98	86.88	40.00	85.71	54.55	50.00
qwen/qwen2.5-vl-32b-instruct	55.56	40.23	75.00	68.12	20.00	57.14	43.18	70.00
qwen/qwen2.5-vl-3b-instruct	22.22	37.07	84.15	86.25	40.00	71.43	50.00	40.00
qwen/qwen2.5-vl-7b-instruct	50.00	45.69	89.63	92.50	60.00	71.43	43.18	60.00

Table 6: Humanities results for ZNO benchmark. GPT-4o is the strongest model for the humanities, with Claude 3.7 Sonnet and Gemini 2.5 Pro being just behind it. The strongest open source model for the humanities is meta-llama/llama-4-maverick, with google/gemma-3-27b-it showing comparable performance.
* - "Ukrainian" contains evaluation for both language and literature knowledge.

Model	Biology	Chemistry	Geography	Mathematics	Physics
anthropic/claude-3.7-sonnet	73.68	71.85	71.33	72.60	65.03
cohereforai/aya-vision-8b	44.36	21.91	39.00	21.61	20.98
google/gemini-2.5-pro-preview-03-25	63.16	80.66	57.33	63.47	56.90
google/gemma-3-12b-it	53.01	29.50	43.00	26.64	26.47
google/gemma-3-27b-it	54.51	28.03	45.00	23.29	24.01
google/gemma-3-4b-it	34.59	23.99	28.67	23.14	20.79
openai/gpt-4o	69.17	68.79	73.00	43.84	54.63
meta-llama/llama-3.2-90b-vision-instruct	52.63	20.20	51.00	19.63	21.93
meta-llama/llama-4-maverick	72.18	48.35	58.33	44.75	37.43
meta-llama/llama-4-scout	58.65	47.86	47.33	42.31	35.73
mistral-community/pixtral-12b	38.72	26.32	40.33	22.37	21.55
qwen/qwen-2.5-vl-72b-instruct	65.41	58.26	63.00	42.62	43.10
qwen/qwen2-vl-2b-instruct	28.95	23.99	32.67	19.94	26.47
qwen/qwen2-vl-7b-instruct	48.87	32.44	42.00	21.31	32.70
qwen/qwen2.5-vl-32b-instruct	45.86	24.24	37.00	21.77	22.31
qwen/qwen2.5-vl-3b-instruct	46.24	35.37	38.67	32.42	30.43
qwen/qwen2.5-vl-7b-instruct	59.02	54.96	53.00	54.34	45.75

Table 7: STEM results for ZNO benchmarks. Claude 3.7 Sonnet is the strongest model overall in all categories, with the Qwen family being the strongest among open-source models.

B ZNO Dataset Overview

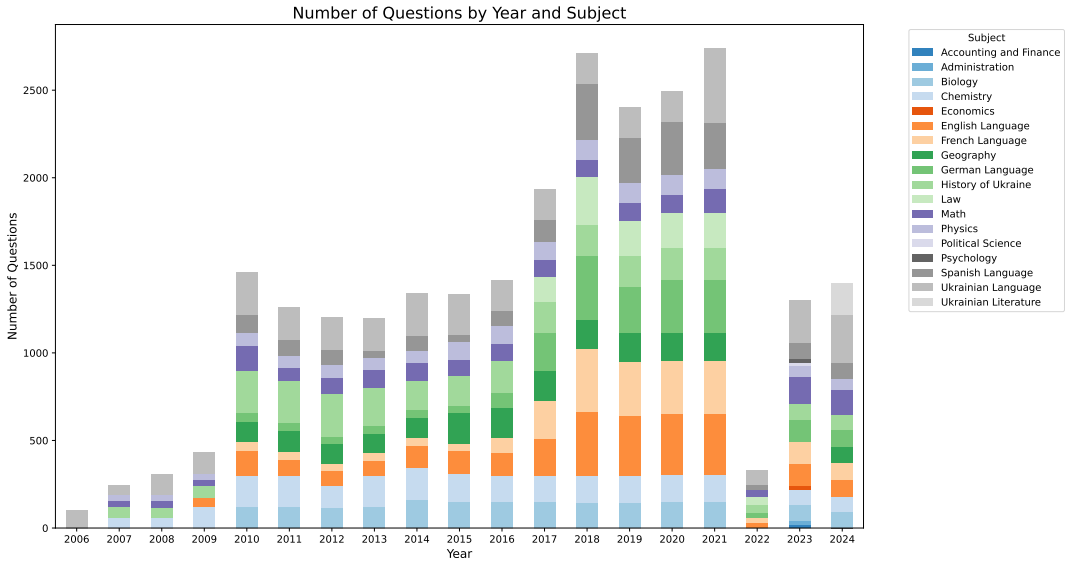

Figure 1: Distribution of ZNO questions by year and subject, showing a skewed yet diverse distribution of questions.

C UACUISINE Questions

Ukrainian Question	**English Translation**
1. Identification Questions	
Q1: Як називається ця страва?	Q1: What is this dish called?
Q2: Яка назва цієї страви в Україні?	Q2: What is the name of this dish in Ukraine?
Q3: Я обідаю в українському ресторані. Зараз я збираюся їсти цю страву. Як називається ця страва?	Q3: I'm having lunch at a Ukrainian restaurant. I'm about to eat this dish. What is this dish called?
2. Generation Questions	
Q4: Перелічи інгредієнти необхідні для приготування зображеної страви	Q4: List the ingredients needed to prepare the shown dish
Q5: Як приготувати цю страву. Опиши коротко покроковий рецепт.	Q5: How to prepare this dish. Describe briefly the step-by-step recipe.
3. Binary Classification Questions	
Q6: Яка ця страва на смак: солона чи солодка?	Q6: How does this dish taste: salty or sweet?
Q7: Ця страва подається холодна чи гаряча?	Q7: Is this dish served cold or hot?

Table 8: Ukrainian UACUISINE Questions with English Translations

D BERT Score Sentences Examples

Reference Ukrainian caption	Generated Caption
Людина в окулярах та капелюсі катається на санях. A person wearing goggles and a hat is sled riding.	Людина в зимовому одязі катається на санках по снігу. A man in winter clothes is sledding in the snow.
Людина в синьо-червоному вбранні стоїть у піщаній місцевості біля води. A person wearing a blue and red outfit is standing in a sandy area near water.	На зображенні людина стоїть на пляжі біля моря під час заходу сонця. The image shows a man standing on the beach at sunset.
Білий собака з коричневим обличчям сидить на цегляній доріжці. A white dog with a brown face is sitting on a brick walkway.	Собака на повідку сидить біля стіни з двома знаками "P". A dog on a leash is sitting next to a wall with two “P” signs.
Чоловічий матч з регбі, де один чоловік перекидається через іншого. Men’s rugby match where one man throws himself over another	Гравець у блакитній формі намагається забити спробу в регбі, тоді як суперник у зеленій формі намагається його зупинити. A player in a blue uniform tries to score a rugby try while an opponent in a green uniform tries to stop him.
Два футболісти в різних формах борються за контроль над м’ячем. Two soccer players in different uniforms struggle to take control of the ball	Двоє футболістів у червоній та білій формі борються за м’яч на полі. Two football players in red and white uniforms are fighting for the ball on the field.

Table 9: Predicted sentences for Multi30k-Uk 2018 subset. Captions are generated using GPT-4o model that scores 0.74 BERT Score and 3.39 SacreBLEU on this task. As shown with these examples, the generated caption correlates with what’s pictured in the image, but the model describes slightly different details.