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

Problem

- Vision-language models (VLMs) are primarily evaluated on English-centric benchmarks
- Critical gap exists for evaluating VLMs in low- and mid-resource languages
- Ukrainian multimodal benchmarks are exceedingly scarce
- No comprehensive evaluation framework for Ukrainian language capabilities
- Cultural knowledge and representation issues

Our Contribution

- Introduced ZNO-Vision: comprehensive Ukrainian-centric benchmark
- First evaluation of multimodal text generation for Ukrainian language
- Evaluated Multi30k-Uk for caption generation task
- Created UACUISINE: cultural knowledge benchmark
- Evaluated both proprietary and open-source VLMs
- Provided methodology template for other low-resource languages
- Published code and datasets with our methodology to benefit community

ZNO-Vision Dataset

- Based on standardized university entrance examination and data from Osvita portal
- Covers mathematics, physics, chemistry, and humanities
- STEM categories dominate with 90%+ visual-only questions

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%

Additional Benchmarks

- UACUISINE: Cultural benchmark featuring 20 popular Ukrainian dishes
  a. Seven question types across three categories: identification, generation, classification
  b. Addresses cultural misrepresentation issues found in existing datasets
  c. Tests knowledge of national cuisine and cultural understanding
- Multi30K-UK (Saichyshyna et al., 2023): Image captioning evaluation using existing Ukrainian benchmark

Evaluation Framework

- Adapted “lmms-eval” framework for standardized evaluation
- 10/10/80 dev/validation/test split following MMMU paradigm
- Multiple metrics: Accuracy, BERT Score, SacreBLEU, Exact Match
- Standardized prompting across all models
- Temperature, token limits and other generation parameters optimized per benchmark

Models Evaluated

- Proprietary models: Claude 3.7 Sonnet, GPT-4o, Gemini 2.5 Pro
- Open-source models: Qwen2.5-VL series, LLaMA 4, Gemma 3
- Range from 2B to 90B parameters
- Comprehensive coverage of current state-of-the-art

Limitation: we haven’t measured reasoning models.

ZNO Benchmark Results

- Claude 3.7 Sonnet, Gemini 2.5 Pro, and GPT-4o lead performance
- Qwen2.5 and LLama 4 are the best open source models for Ukrainian MultiModal setting
- LLaMA 3.2 and Pixtral failed to beat random baseline
- Text-only evaluation of the same tasks lead is close to random baseline

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

Multi30k (Caption Generation) Results

- GPT-4o achieved highest BERT Score (0.74) and BLEU (3.54)
- Most models generated semantically correct but stylistically different captions
- Low BLEU scores due to synonym usage and different detail focus
- Models struggled with single-sentence instruction following
- Evaluation highlighted challenges in Ukrainian text generation assessment

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

UACUISINE Cultural Results

- Claude 3.7 Sonnet performed best overall across all categories
- No model scored high on simple dish naming task
- Gemini 2.5 Pro achieved 35% exact match but refused recipe generation
- LLaMA 4 Maverick and Gemma-27B-it are the strongest among open-source
- Significant room for improvement in cultural understanding

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

Challenge 1: Instruction Following and Performance

- Inconsistent instruction following in Ukrainian across all models
- High-performing models like GPT-4o frequently failed expected format, like “place answer inside [ ]”
- LLaMA 3.2-90b responded with English letters instead of Ukrainian
- Models generated verbose responses despite being instructed to be concise
- Format extraction rules needed due to instruction non-compliance

Challenge 2: Code-switching Issues

- Major issues with language confusion and code-switching
- Models prompted in Ukrainian switched to English, Chinese, or Russian
- Broken grammar generation: "Куряче супу з лапшой" (“chicken soup with noodles” in incorrect word form)
- Non-existing word creation and tokenization artifacts
- Same issues known in text-only ZNO setting

Challenge 3: Cultural Misattribution

- Ukrainian Borscht mislabeled as "Russian Red Borscht", despite being UNESCO-recognized Ukrainian cultural heritage
- Models defaulted to English or Russian responses when prompted in Ukrainian
- Recipe generation suggested incorrect preparations
- Systematic bias points to training data issues affecting cultural identity

Conclusion

- First objective framework for evaluating Ukrainian multimodal capabilities
- Revealed significant performance gaps and cultural representation issues
- Only top-tier models achieved reasonable performance above baseline
- Critical need for inclusive AI development for low-resource languages
- Open-source code and datasets available for community use

Code, datasets and leaderboard: github.com/lang-uk/mmzno-benchmark

Thanks for your attention!

This research wouldn’t be possible without support from:

- ELEKS for their grant in memory of Oleksiy Skrypnyk
- The alliance of De Novo and MK-Consulting for providing computational resources
- We also gratefully acknowledge Amazon Web Services (AWS) for cloud credits that enabled inference on H200 instances
- Google Cloud Platform (GCP) for credits supporting model inference