Multimodal Evaluation of Russian-language Architectures

Artem Chervyakov[*1], Ulyana Isaeva[*1], Anton Emelyanov[1], Artem Safin[1], Maria Tikhonova[1,4], Alexander Kharitonov[1], Yulia Lyakh[1], Petr Surovtsev[1], Denis Shevelev[1], Vildan Saburov[1,8], Vasily Konovalov[3,5,6], Elisei Rykov[5,7], Ivan Sviridov[2], Amina Miftakhova[2], Ilseyar Alimova[5], Alexander Panchenko[5,3], Alexander Kapitanov[1], Alena Fenogenova[*1]

[1]SberAI, [2]Sber AI Lab, [3]AIRI, [4]HSE University, [5]Skoltech, [6]MIRAI, [7]T-Tech, [8]Moscow Center for Advanced Studies

Correspondence: mera@a-ai.ru

Abstract

Multimodal large language models (MLLMs) are currently at the center of research attention, showing rapid progress in scale and capabilities, yet their intelligence, limitations, and risks remain insufficiently understood. To address these issues, particularly in the context of the Russian language, where no multimodal benchmarks currently exist, we introduce **MERA Multi**, an open multimodal evaluation framework for Russian-spoken architectures. The benchmark is instruction-based and encompasses default text, image, audio, and video modalities, comprising 18 newly constructed evaluation tasks for both general-purpose models and modality-specific architectures (image-to-text, video-to-text, and audio-to-text). Our contributions include: (i) a universal taxonomy of multimodal abilities; (ii) 18 datasets created entirely from scratch with attention to Russian cultural and linguistic specificity, unified prompts, and metrics; (iii) baseline results for both closed-source and open-source models; (iv) a methodology for preventing benchmark leakage, including watermarking for private sets. While our current focus is on Russian, the proposed benchmark provides a replicable methodology for constructing multimodal benchmarks in typologically diverse languages, particularly within the Slavic language family.

1 Introduction

Recent breakthroughs in generative AI, including models like GPT-5[1], ImageBind (Girdhar et al., 2023), and LLaVa (Liu et al., 2024a), have significantly advanced the state of the art across multiple modalities. This accelerated progress created a growing need for comprehensive multimodal benchmarks capable of rigorous evaluation of the full spectrum of versatile capabilities of such models. Although several benchmarks have been proposed for English and general-domain evaluation, such as MultiBench (Liang et al., 2021), MMBench (Liu et al., 2024e), and General-Bench (Fei et al., 2025), they predominantly neglect the linguistic and cultural nuances of Slavic languages, particularly Russian. Beyond its Cyrillic script, Russian possesses a rich cultural context where concepts familiar to native speakers (e.g., folklore, Soviet media) are foreign to others, creating a challenge for automatic understanding.

Existing Russian-specific benchmarks, including TAPE (Taktasheva et al., 2022), Russian SuperGLUE (Fenogenova et al., 2022), and MERA (Fenogenova et al., 2024), focus exclusively on text-based tasks, leaving a critical gap in multimodal evaluation. To address this, we introduce **MERA Multi**[2], the first multimodal benchmark for MLLM evaluation in Russian. It comprises 18 tasks spanning (default) text[3], image, audio, and video modalities, built upon a unified taxonomy of multimodal abilities. Beyond Russian MERA Multi offers a blueprint for developing multimodal benchmarks across other Slavic and morphologically rich languages. These languages share structural complexity, typological proximity, and cultural specificity that make direct translation or adaptation of English-centric benchmarks inadequate. Thus, it serves not only as the first multimodal benchmark for Russian, but also as a scalable methodology for culturally and linguistically aware evaluation in underrepresented settings.

More specifically, our contributions are fourfold:

- We propose a unified taxonomy and evaluation methodology designed for MLLMs assessment;
- We create 18 novel datasets [4] incorporating

[*] Core contributors

[1] https://openai.com/index/introducing-gpt-5

[2] https://mera.a-ai.ru/en/multi

[3] Since textual input is an inherent component of all tasks in our benchmark, we treat it as the default modality. Consequently, we do not explicitly list it when describing or tabulating tasks, unless required for clarity.

[4] https://hf.co/collections/MERA-evaluation/mera-multimodality

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2114–2161

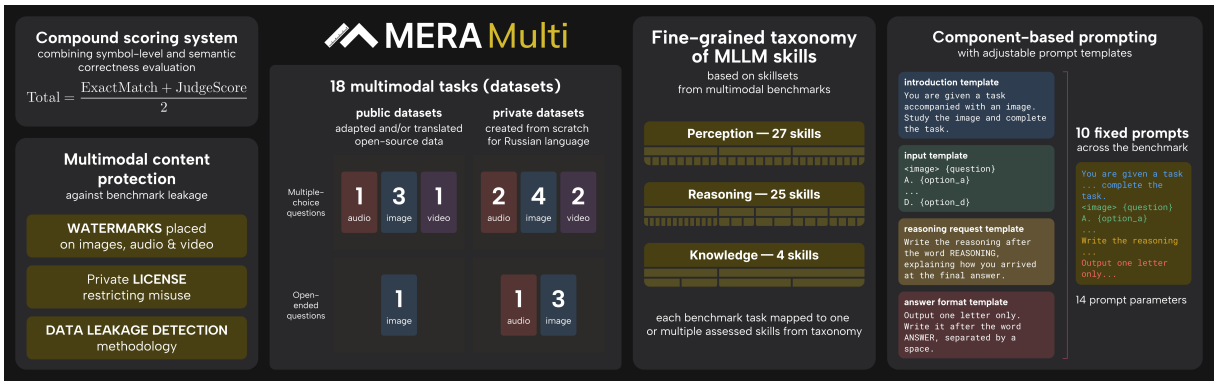


Figure 1: Overview of MERA Multi. The benchmark unites multimodal evaluation, taxonomy-based skill assessment, and data leakage protection across 18 tasks covering (default) text, image, audio, and video modalities. It employs standardized block-prompting, compound scoring, and integrates methods for multimodal content protection, forming a transparent and robust methodology for culturally grounded multimodal evaluation in Russian.

Russian cultural and linguistic specificities, unified prompts, and metrics.

- We provide baseline performance results for both open- and closed-source models;
- We establish a data leakage analysis and watermarking strategy to protect private evaluation datasets.

Additionally, we provide a standardized codebase for full reproducibility and a submission platform with automated scoring and a public leaderboard[5]. All datasets and code are made available under the MERA Multi license, which allows the use of the benchmark sets for non-commercial purposes, provided that the data is not used for model training of any kind. These elements establish a foundation for transparent and culturally aware multimodal evaluation in Russian, while positioning MERA Multi as a reference point for developing similar frameworks across Slavic and other non-English languages, thereby promoting broader community development and cross-linguistic benchmarking.

2 Related Work

Text-Based Benchmarks Evaluation of language models has historically relied on Natural Language Understanding benchmarks such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), which set initial standards for English. As these benchmarks became saturated, newer instruction-oriented and reasoning-focused benchmarks emerged such as BIG-bench (Suzgun et al., 2023) and HELM (Perlitz et al., 2024). Further efforts such as MMLU (Hendrycks et al., 2021), AGIEval (Zhong et al., 2024), and C-Eval (Huang et al., 2023) extended evaluation to academic and professional domains, though primarily in English and Chinese. For Russian, several text-based benchmarks were introduced. Among them are Russian SuperGLUE (Shavrina et al., 2020; Fenogenova et al., 2022), TAPE (Taktasheva et al., 2022), and RuCoLA (Mikhailov et al., 2022). The instruction-based benchmark MERA (Fenogenova et al., 2024) advanced this line of work but remained limited to text-only evaluation. Our work extends this direction toward multimodal, instruction-following assessment of MLLMs in Russian.

Multimodal Benchmarks The rapid progress of multimodal models has led to numerous benchmarks extending evaluation beyond text. Early large-scale efforts as MultiBench (Liang et al., 2021) covered 10 modalities and emphasized general-purpose representation learning. MMBench (Liu et al., 2024e) focused on fine-grained visual reasoning and bilingual (English–Chinese) evaluation, while SUPERB (Yang et al., 2021) unified diverse audio tasks under a single framework. For video understanding, STAR (Wu et al., 2021), InfiniBench (Ataallah et al., 2024), and Video-MME (Fu et al., 2025) advanced evaluation toward temporal and long-context reasoning. More comprehensive setups such as General-Bench (Fei et al., 2025), OmniDialog (Razzhigaev et al., 2024), MMMU (Yue et al., 2024), and SEED-Bench (Li et al., 2023) assess multimodal and reasoning abilities at scale. All in all, despite these advances, most existing multimodal benchmarks are English-centric. To our knowledge, there is currently no

[5] https://mera.a-ai.ru/en/multi/leaderboard

Benchmark	Modalities	# Tasks/Skills	Primary Focus	Language	Cultural Focus
MultiBench (Liang et al., 2021)	10 (Text, Image, Audio, Video, Table, Set, ...)	20	General-purpose representation	EN	General
MMBench (Liu et al., 2024e)	Image, Text	20	Fine-grained visual reasoning	EN, ZH	General
SUPERB (Yang et al., 2021)	Speech, Audio	10	Speech processing	EN	General
STAR (Wu et al., 2021)	Video, Text	4	VideoQA, temporal reasoning	EN	General
InfiniBench (Ataallah et al., 2024)	Video, Text	8	Long-context video understanding	EN	General
Video-MME (Fu et al., 2025)	Video, Text	12	Fine-grained video analysis	EN	General
General-Bench (Fei et al., 2025)	Text, Image, Video, Audio, Tabule	700+	Large-scale ability coverage	EN	General
OmniDialog (Razzhigaev et al., 2024)	Text, Image, Audio	8	Multimodal dialogue	EN	General
MMMU (Yue et al., 2024)	Image, Text	30	Expert-level reasoning	EN	General
SEED-Bench (Li et al., 2023)	Image, Video, Text	12	Generative comprehension	EN	General
MERA Multi (ours)	Text, Image, Audio, Video	18	Comprehensive understanding	RU	General + Russian

Table 1: Comparison of major multimodal evaluation benchmarks and MERA Multi.

	Dataset / task	Size	HB	Answer	License
audio	ruEnvAQA	596	0.95	MC	CC BY-NC 4.0
	RuSLUn	741	0.90	OE	CC-BY-4.0
	*AQUARIA	738	0.98	MC	MERA Multi
	*ruTiE-Audio	1500	0.75	MC	MERA Multi
image	ruCLEVR	1148	0.96	OE	CC-BY-4.0
	ruCommonVQA	3015	0.84	OE	CC-BY-4.0
	ruNaturalScienceVQA	363	0.99	MC	CC-BY-SA-4.0
	WEIRD	814	0.85	MC	CC-BY-4.0
	*LabTabVQA	339	0.91	MC	MERA Multi
	*RealVQA	773	0.63	OE	MERA Multi
	*ruHHH-Image	595	0.89	MC	MERA Multi
	*ruMathVQA	502	0.95	OE	MERA Multi
	*ruTiE-Image	1500	0.77	MC	MERA Multi
	*SchoolScienceVQA	4227	0.82	MC	MERA Multi
	*UniScienceVQA	7432	0.13	OE	MERA Multi
video	CommonVideoQA	1200	0.96	MC	CC-BY-4.0
	*RealVideoQA	671	0.96	MC	MERA Multi
	*ruHHH-Video	911	0.84	MC	MERA Multi

Table 2: Overview of datasets in MERA Multi. Those marked with an asterisk are *private datasets* collected from scratch, while the others are *public datasets* compiled from open-source datasets. **Size** column shows the number of samples in the dataset. **HB** is the human baseline value (basic / expert or basic only, see F.2 for details). **Answer** column is the task format (MC and OE stand for multiple-choice and open-ended, respectively). *MERA Multi* license refers to the benchmark license anonymized for the review period.

existing multimodal benchmark for the Russian language. MERA Multi addresses the gap by providing a unified and culturally adapted evaluation of understanding across several modalities (see Table 1 for the detailed comparison between MERA Multi and other benchmarks).

3 Overview of MERA Multi Benchmark

Figure 1 presents an overview of MERA Multi and its methodological parts.

3.1 Benchmark General Structure

The proposed benchmark is designed to evaluate the capabilities of Russian MLLMs. Besides the omnipresent textual modality, it incorporates tasks from three other modalities: image (11 datasets), audio (4 datasets), and video (3 datasets). The tasks are formulated in two primary formats: multiple-choice questions and open-ended questions requiring short free-form answers.

To balance reproducibility and novelty, the benchmark integrates both publicly available datasets (7 tasks), curated from open data sources, and private datasets (11 tasks), collected from scratch specifically for this study. The latter are designed to incorporate Russian cultural nuances, target underexplored skill categories, and mitigate potential data contamination issues (see details in section 3.4). A complete list of datasets is provided in Table 2. Task examples and dataset creation details are in A.

3.2 Skill Taxonomy

Contemporary studies on multimodal benchmarks often rely on custom skill sets during the dataset design process (Liu et al., 2024e; Fu et al., 2024). We performed a comprehensive analysis of such systems and synthesized them into a consolidated MLLM skill taxonomy, which underpins the foundation of MERA Multi. Such a system functions as a comprehensive map for skill coverage, which, when coupled with an alignment of existing datasets to corresponding skills, can spotlight deficiencies in benchmark task diversity.

The taxonomy is aligned with three broad categories of human-like cognition: **knowledge**, **perception**, and **reasoning**, a division also adopted by recent multimodal benchmarks. The knowledge taxonomy is shown in Table 3, and the perception and reasoning parts are provided in the Appendix in Table 7 and Table 8 respectively, due to space constraints. To comply with the emerging MLLM capabilities, this taxonomy is designed to be extendable and this paper provides the initial version of the unified taxonomy and encourages its adoption and extension in future research.

Each MERA Multi dataset is systematically mapped to a predefined set of multimodal skills.

Rather than assigning a single skill per task, each task is designed to be multifaceted, evaluating a distinctive combination of abilities. For example, a single task might require visual perception (to identify objects), OCR (to read text in the image), and reasoning (to answer a why-question) all together.

Further details on skill taxonomy are in Appendix B.

3.3 Evaluation Methodology

MERA Multi evaluation methodology is designed to systematically assess the multimodal reasoning, perception, and knowledge abilities of MLLMs. It measures both general-purpose and modality-specific competence across image, audio, and video inputs. To achieve this, we propose the benchmark that integrates four complementary components: (i) a block-prompting structure that ensures consistency and diversity of task formulations across modalities (Section 3.3.2); (ii) dual-level evaluation metrics combining symbolic and semantic correctness through a dedicated LLM-as-a-judge model (Section 3.3.3); (iii) an evaluation pipeline including scoring aggregation and cross-modality weighting (Section 3.3.4); and (iv) a submission protocol enabling automated, reproducible leaderboard updates (Section 3.3.5). These components provide a coherent methodology that balances rigor, interpretability, and cross-model comparability in multimodal evaluation.

3.3.1 Evaluation Framework Implementation

We build the benchmark code base on the `lm-eval`[6] framework (Gao et al., 2024; Biderman et al., 2024), extending it to support multimodal inputs while preserving its core structure for texts. The codebase introduces comprehensive vision, audio, and video evaluation through, and chat-template formatting tailored for instruction-tuned and API-based models. Multimodal data is integrated either by passing it separately to the model's processor or by embedding it directly within a chat template alongside the text. All evaluations are strictly generative: models produce free-form answers until a stop condition is met. All outputs are used directly except RuSLUn which requires minimal parsing for structured output.

To balance rigor and semantic understanding, we employ two primary metrics across most datasets. $ExactMatch$ (EM) serves as a generative analog of accuracy, using normalized string comparison to assess both factual correctness and strict adherence to the specified output format. Complementing this, the $JudgeScore$ (JS) measures semantic similarity via an LLM-as-a-judge (trained for this purpose) that scores an answer as 1 for substantive agreement with the reference and 0 otherwise. This dual approach provides a nuanced view: EM is sensitive to format, while the JS focuses on meaning. For the datasets where this format is unsuitable we employ task-specific metrics or report only the JS.

The $FinalScore$ of the task (for Total Score calculation purposes) is defined as follows:

$$FS(task) = \frac{\text{EM}(task) + \text{JS}(task)}{2}; \quad (1)$$

It should be noted that as long as some of our prompts suggest that the model return the answer for the question after a specific phrase ("ANSWER"), we consider EM to be the maximum of two scores: (i) EM of the full model answer, (ii) EM of the last part of the model's answer split by the previously mentioned specific phrase. If the model breaches the instruction, we assess the full answer. Otherwise, we consider the part that is after the specific phrase to be the model's answer as required by the instruction.

3.3.2 Prompt Structure

Following the approach of Voronov et al. (2024), we avoid hard-coding task-specific prompts by employing a block-prompting scheme that constructs universal prompts aligned with each task's structure. For every dataset, we instantiate a predefined set of 10 prompts drawn from fixed block layouts that we initially designed and realized under different surface modalities (e.g., formal request, command, technical statement). This design preserves a uniform formulation across benchmark tasks while allowing for controlled variation, such as the presence or absence of a reasoning request, to mitigate benchmark saturation.

Following the methodology of Fenogenova et al. (2024), we uniformly assign prompts for each task across dataset samples to ensure that the aggregated performance metric reflects an average over prompt variants rather than the idiosyncrasies of a single prompt. A post-hoc statistical analysis (see Appendix E for the details) confirmed that our prompt set is not invariant, as some prompts produce statistically significant shifts in model scores. This finding directly motivates and validates our

[6] https://github.com/EleutherAI/lm-evaluation-harness

Taxonomy Level					Modality		
L1	**L2**	**L3**	**L4**	**L5**	**Image**	**Audio**	**Video**
Knowledge	Knowledge		Common everyday knowledge	**Common everyday knowledge**	ruCommonVQA RealVQA ruHHH-Image WEIRD ruTiE-Image	ruTiE-Audio ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
				Ethics	ruHHH-Image		ruHHH-Video
			Domain knowledge	**Common domain knowledge**	ruMathVQA RealVQA ruTiE-Image	ruTiE-Audio ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA
				Expert domain knowledge	ruMathVQA ruNaturalScienceVQA UniScienceVQA SchoolScienceVQA		

Table 3: Knowledge taxonomy structure and multimodal task distribution in MERA Multi. Columns **L1-L5** show the skill hierarchy levels, while **Image**, **Audio** and **Video** columns indicate which tasks cover each knowledge type across different modalities.

multi-prompt design as a necessary guard against evaluation bias.

3.3.3 Judge Scoring

Evaluating open-ended model generations requires moving beyond strict heuristics. We frame this as a task of assessing semantic equivalence, conditioned on the question, as devising universal regex patterns to extract answers from free-form text, particularly from multi-step reasoning chains – proves infeasible. The practically infinite space of possible valid generations necessitates a more flexible and scalable assessment approach. We therefore introduce a learned LLM-as-judge model, framing evaluation as a question-conditioned semantic equivalence task (binary classification – determine the correctness of a model's prediction).

This judge is trained on a diverse, human-annotated dataset of model outputs collected from various Russian-language instruction-based benchmarks. The final model, based on the RuModernBERT[7] (Spirin et al., 2025) encoder, achieves F1 score of 0.96 on a held-out test set and shows 99.6% agreement with EM on identical answers, confirming its reliability. We deploy this judge as our primary metric, providing a robust, scalable score for correctness across the benchmark. Comprehensive details on data, training, and model comparisons are in the Appendix D.

3.3.4 Evaluation Pipeline

Let MERA Multi comprise M modalities (e.g., image, audio, video) not counting the omnipresent text one, with modality m containing n_m tasks. We treat modalities as equally important and split each modality's weight evenly across its tasks. Let k_m be the number of tasks a model attempted in modality m with $s_{m,i} \in [0; 1]$ being the score (average of all task i metrics).

- The **Attempted Score**. Quality over the tasks the model attempted, with the original equal-per-modality weighting renormalized over the attempted subset:

$$A = \frac{\sum_{m=1}^{M} \frac{1}{n_m} \sum_{i \in A_m} s_{m,i}}{\sum_{m=1}^{M} \frac{k_m}{n_m}} \tag{2}$$

- The **Coverage**. Breadth of evaluation is the fraction of the benchmark the model actually attempted:

$$C = \frac{1}{M} \sum_{m=1}^{M} \frac{k_m}{n_m} \tag{3}$$

- The **Total Score**. We combine quality and breadth as

$$T = A \cdot C \tag{4}$$

This approach has the following positive effects:

- **Separation of concerns**. A reports how well a model performs where it runs; C reports how much of MERA Multi it covers.
- **Fair, single leaderboard**. T enables joint ranking without imputing arbitrary zeros for missing modalities; specialists still excel on modality-specific boards (higher A, lower C).
- **Stable under growth**. Adding/removing tasks adjusts C (breadth), but leaves A (quality on attempted tasks) invariant; equal-per-modality weighting prevents any modality from dominating due to task count.

[7] https://hf.co/deepvk/RuModernBERT-base

3.3.5 Submission

The private dataset test answers are available only for the organizers and experts supporting the benchmark. The scoring system is automatic and is available on the benchmark platform.

The process of submission is the following:

- First, users clone MERA Multi benchmark repository and form submission files using shell script and the provided code.
- Second, they upload the submission files via the platform interface[8] in their personal account for automatic assessment.
- The evaluation result is then displayed in the user's account and kept private unless they request publication using the "Publish" function.
- In this case, it undergoes an expert verification of its reproducibility, which includes checking log files automatically formed by the evaluation script and the provided submission information. Once approved, the model's score is shown publicly on the leaderboard (user can choose on which leaderbord image/audio/video/multi to add the results), while its specific outputs remain private.

3.4 Data Protection

As pre-training datasets grow, benchmark leakage is becoming more common. This issue is exacerbated by opaque training processes and the undisclosed use of supervised data in modern MLLMs, which undermines the validity of benchmarks and fosters unfair comparisons, ultimately slowing progress. To protect the multimodal data in our private benchmark suite from unauthorized use, we employ: (i) data watermarking (Section 3.4.1), (ii) leakage detection (Section 3.4.2). We also introduce the license, which explicitly permits the use of the benchmark data for research and non-commercial testing purposes, but strictly prohibits its incorporation into any model training process. Together, these measures provide robust technical and legal safeguards.

3.4.1 Data Watermarking

We embed imperceptible yet identifiable watermarks into our benchmark data to trace its provenance and detect unauthorized use in training corpora. Our approach is tailored to each modality:

[8]For this step the registration on the platform is required.

	Model	Confidence Interval
image	Qwen2-VL-2B-Instruct	(0, 0.352)
	Qwen2-VL-7B-Instruct	(0, 3.920)
	Qwen2.5-VL-3B-Instruct	(0.259, 2.180)
	Qwen2.5-VL-7B-Instruct	(0.444, 1.389)
	llava-1.5-7b-hf	(0, 0.937)
	llava-1.5-13b-hf	(0, 0.839)
	llama3-llava-next-8b-hf	(0, 1.024)
	gemma-3-12b-it	(1.183, 3.120)
	gemma-3-12b-it	(0, 1.756)
video	LLaVA-NeXT-Video	(0.811, 3.314)
	LLaVA-NeXT-Video-DPO	(1.367, 2.125)
	LLaVA-NeXT-Video-34B	(0, 1.373)
	Qwen2-VL-2B-Instruct	(0, 0.645)
	Qwen2.5-VL-7B-Instruct	(0, 0.781)
	Qwen2.5-VL-3B-Instruct	(0, 0.330)
	Qwen2-VL-7B-Instruct	(0, 0.105)
audio	ultravox-v0_2	(0, 0.940)
	ultravox-v0_3-llama-3_2-1b	(0, 1.566)
	ultravox-v0_5-llama-3_2-1b	(0, 0.195)
	ultravox-v0_6-llama-3_1-8b	(0, 0.875)
	ultravox-v0_4	(0, 0.401)
	ultravox-v0_4_1-mistral-nemo	(0, 0.079)
	Qwen2-Audio-7B-Instruct	(0.0, 0.0)
	MiniCPM-o-2_6	(0.0, 0.0)

Table 4: Confidence intervals of Judge Score (JS) differences between watermarked and clear data (%).

- **Audio**: We use AudioSeal (Roman et al., 2024) for localized detection, which employs a neural, inaudible watermark.
- **Images/Video**: a simple overlay of the MERA Multi watermark on images and every video frame (same code across frames).

Table 4 demonstrates the Confidential Intervals (CI) for differences of the models' metrics on data with watermarks and without them. The CI are computed for differences in per cents for convenience. The results demonstrate that with 95% probability the difference is less than 5% (usually less than 1%) which leads to the conclusion: our watermarking strategy does not significantly affect the evaluation results.

3.4.2 Data Leakage Detection

We detect training-set data leakage using membership inference attacks (Shokri et al., 2017). Our approach (Emelyanov et al., 2025) extends the Semantic Membership Inference Attack (SMIA) (Mozaffari and Marathe, 2024) to MLLMs, calculating the loss for a data point by considering its text in conjunction with its paired image, video, or audio data. The Multimodal SMIA (MSMIA) method identifies leakage by comparing a model's behavior on original examples versus their semantically perturbed "neighbors" (masked, removed, doubled, switched text tokens); models that have been trained on the data (models with data leak) are to exhibit systematically different loss patterns.

Modality	AUC-ROC
Image	88.658
Video	88.388
Audio	81.250

Table 5: Average AUC-ROC of MSMIA per modality. Averaging over the models used for training and evaluating MSMIA.

Concretely, we train the MSMIA detector by comparing two versions of a model: the original and a version we fine-tune on candidate data (simulating a leak). The detector learns to distinguish between them by analyzing the differences in loss and text embeddings when processing original data points versus their perturbed neighbors. Once trained, this detector can be applied to any other model to output a probability that a specific data sample was part of that model's training set. Overall results of the MSMIA detection capabilities are presented in Table 5. Following the original methodology, we evaluate detection performance using AUC-ROC. The table demonstrates relatively high scores, which means that the MSMIA method tends to be capable of detecting whether a model has been trained on some multimodal data samples or not, with high probability and a rather low false-positive rate. The details on the MSMIA training and the metrics are provided in Appendix C.

4 Baselines

4.1 Model Baselines

We evaluate over 50 publicly available multimodal models from the most trending model families on HuggingFace, varying in size from 1B to 110B parameters. Also we evaluate proprietary GPT 4.1 (OpenAI) to make a comparison between open and closed source models.

See Appendix F.1 for the baseline details. We evaluate models in the same environments and scenarios by the procedure described in Section 3.3.4 and the submission procedure described in Section 3.3.5. We also provide examples of particular submissions (the model evaluated on part of the tasks of a modality).

4.2 Human Baselines

To estimate human-level performance across MERA Multi tasks, we compute human baseline (HB) values based on the aggregated results of crowd annotators. For datasets requiring additional domain expertise, we also establish expert HB obtained from qualified expert annotators. Annotation quality is ensured through honeypot tasks with automated correctness verification and post-hoc filtering of low-performing annotators. All crowd annotations are collected via the ABC Elementary platform[9], which guarantees data anonymity, fair compensation, and ethical compliance. See Appendix F.2 for detailed methodology and cost breakdown .

5 Results

The leaderboard is designed in such a way that the more modalities the model covers, the higher the Total Score could be. The top performer, Qwen3-Omni-30B-A3B-Instruct, leads with a Total Score of 0.5, driven by its high Attempted Score (0.563) and rather high Coverage (0.889), showing strong image, audio, and video capabilities. Notably, the models from Qwen families obtain larger scores for image and video modalities (first 3 places of the overall leaderboard are taken by those models). GPT 4.1 still leads in image modality while having low Coverage (0.333), which leads to a lower Total Score (0.159 compared to 0.5 of the top performer). The main trend is defined by the metrics used: broader coverage leads to higher Total Score. Thus omni-models occupy the first places. But strong uni- or bimodal capabilities may gain advantage over middle-performing models with high Coverage (e.g. Qwen2.5-VL-72B-Instruct and Qwen2-VL-72B-Instruct with 0.302 and 0.254 Total Scores respectively).

This pattern is consistent across all modalities. In audio, the specialists from *ultravox* family tend to display poorer performance compared to omni-models like Qwen2.5-Omni-7B (0.311 vs 0.474 for Audio Total Score) even though *ultravox* models use other LLM's from *Mistral*, *Llama*, **Qwen** families as backbones. Considering the video modality, vision models are usually trained with video-inputs or may slice the video into frames and use "regular" vision encoders for them, which explains why Qwen2-VL-72B-Instruct shows the best Video Total Score (0.625) while the models that specialize specifically on video modality like those from **LLaVA-NeXT-Video** family show poorer metrics.

Consistently, `Judge Score (JS)` > `EM` across models, indicating that many responses are semantically correct but violate output format; whenever

[9] https://app.elementary.center

Model	Total Score	Attempted Score	Coverage	Image Total	Audio Total	Video Total
Qwen3-Omni-30B-A3B-Inst	**0.500**	**0.563**	0.889	**0.554**	**0.561**	0.410
Qwen2.5-Omni-7B	0.317	0.317	**1.000**	0.226	0.474	0.442
Qwen2.5-VL-72B-Inst	0.302	0.453	0.667	0.406	0.000	**0.625**
MiniCPM-o-2_6	0.255	0.255	**1.000**	0.182	0.369	0.373
Qwen2-VL-72B-Inst	0.254	0.381	0.667	0.333	0.000	0.557
Qwen2.5-Omni-3B	0.251	0.251	**1.000**	0.180	0.380	0.337
Qwen2.5-VL-7B-Inst	0.209	0.313	0.667	0.256	0.000	0.523
GPT 4.1	0.159	0.478	0.333	0.478	0.000	0.000
Qwen2-VL-7B-Inst	0.145	0.218	0.667	0.195	0.000	0.301
Qwen2.5-VL-3B-Inst	0.136	0.203	0.667	0.142	0.000	0.427
InternVL3-9B	0.135	0.203	0.667	0.172	0.000	0.316
Qwen3-VL-2B-Inst	0.125	0.187	0.667	0.125	0.000	0.416
ultravox-v0_5-llama-3_1-8b	0.104	0.311	0.333	0.000	0.311	0.000
ultravox-v0_4_1-llama-3_1-8b	0.102	0.307	0.333	0.000	0.307	0.000
ultravox-v0_4	0.101	0.304	0.333	0.000	0.304	0.000
ultravox-v0_6-llama-3_1-8b	0.100	0.300	0.333	0.000	0.300	0.000
Phi-4-multimodal-instruct	0.098	0.147	0.667	0.185	0.042	0.000
ultravox-v0_4_1-mistral-nemo	0.088	0.265	0.333	0.000	0.265	0.000
audio-flamingo-3-hf	0.086	0.259	0.333	0.000	0.259	0.000
llava-next-110b-hf	0.079	0.236	0.333	0.236	0.000	0.000
Phi-3.5-vision-inst	0.076	0.228	0.333	0.228	0.000	0.000
Qwen2-Audio-7B-Inst	0.074	0.223	0.333	0.000	0.223	0.000
SmolVLM-Inst	0.064	0.192	0.333	0.192	0.000	0.000
gemma-3-27b-it	0.050	0.151	0.333	0.151	0.000	0.000
granite-vision-3.3-2b	0.048	0.143	0.333	0.143	0.000	0.000
deepseek-vl2-small	0.044	0.163	0.273	0.133	0.000	0.000

Table 6: Overall baselines information over three modalities (vision/image, audio, video). All scores are aggregated. Modality total score is the attempted score multiplied by coverage of the modality.

JS $\approx$ EM, the model followed instructions closely. This gap justifies reporting JS alongside EM. When JS $<$ EM, this means that we can extract the answer from model's generation but the entire generation may be misleading (e.g. wrong rationale, reasoning conclusion mismatches the selected answer).

Tables with separate datasets metrics and analysis may be found in Appendix F.1.

Takeaway 1: There is still a gap between modalities. Omni-models partially bridge it. Specialist models, however, show that while image understanding is a relatively mature field, audio and video understanding are underrepresented in terms of both models and datasets, which is reflected in lower scores on benchmark tasks.

Takeaway 2: Overall metrics are robust to missing task scores (unfinished submission) and multiple modalities.

Conclusion

The rapid progress of generative AI has introduced new challenges for evaluating models in multimodal contexts. We present **MERA Multi**, the first comprehensive framework for transparent and culturally grounded multimodal evaluation in Russian. The benchmark encompasses 18 tasks across four modalities (default text, image, audio, and video), covering diverse domains and scenarios. It systematizes diverse multimodal abilities via a proposed taxonomy and evaluates them through methodologically verified prompts and metrics. We also provide a standardized code base[10] that guarantees reproducibility and a submission platform offering automated evaluation, scoring, and open leaderboards.

In the future, we plan to expand the benchmark to encompass additional scenarios and actively encourage community contributions. We envision MERA Multi as a collaborative initiative that promotes transparent evaluation practices and provides a methodological foundation for developing culturally aware multimodal benchmarks across non-English languages such as Slavic, ultimately advancing the creation of more robust and reliable multimodal models.

[10] https://github.com/MERA-Evaluation/MERA_MULTIMODAL

Limitations

First of all, despite the fact that our benchmark covers 18 tasks spanning multiple domains, aiming to represent complementary semantic abilities of the models, this set may be underrepresenting some abilities of the model or some domains which may be crucial for certain tasks and applications. Namely, it is not impossible that a model which excels at our benchmark will perform poorly on a specialized domain or task.

Even with fixed prompts and decoding settings, MERA Multi scores can vary because the entire hardware–software stack affects inference: GPU model, drivers/CUDA/cuDNN, PyTorch, vLLM/-transformers (and commit hashes), flash-attention kernels, tokenizers/checkpoints, precision/quantization, and batching — some of which are nondeterministic. We therefore request public submissions adhere to the same parameters and, in submission information, specify the GPUs and libraries versions they used for reproducibility purposes.

Ethical Statement

While the presented benchmark is able to comprehensively evaluate "semantic" abilities of the model, i.e., the capacity of individual models to understand and reason about data in different modalities, we did not perform explicit evaluation of any bias of these models, e.g., toward any kind of underrepresented minorities. In our opinion, this is an extremely important direction of future work, yet being outside the scope of our current contribution.

For the creation of novel datasets, the work of human annotators is used. We state that their work was adequately paid or compensated (see Section F.2 for the details).

Researchers participating in the benchmark will be encouraged to adhere to ethical research practices, including proper citation, acknowledgment of data sources, and responsible reporting of results. Regular ethical reviews will assess the benchmark's impact, identify potential ethical concerns, and implement necessary adjustments to uphold the highest ethical standards throughout the usage usage.

We proofread the text of this article using Overleaf Writefull assistant[11], GPT-4o[12], Grammarly[13] to correct grammatical, spelling, and style errors and paraphrase sentences. We emphasize that these tools are used solely to enhance the quality of English writing, in full compliance with the ACL policies on the responsible use of AI writing assistance. Nevertheless, some segments of our publication can be potentially detected as AI-generated, AI-edited, or human-AI-generated.

[11] https://www.writefull.com
[12] https://chatgpt.com
[13] https://app.grammarly.com

Acknowledgments

MERA is a collaborative project, thoughtfully designed to serve the needs of both industry and academia. The authors extend their sincere gratitude to our partners from the AI Alliance Russia, whose invaluable collaboration made an undertaking of this scale possible. Special thanks are due to Ekaterina Morgunova, Yegor Nizamov, and Uliana Plugina for their significant contributions in coordinating our benchmark partners and contractors for the website development.

We also express our deep appreciation to the entire team dedicated to developing the website platform and maintaining the scoring system.

We are profoundly grateful to the following individuals for their dedication and hard work: Yaroslav Grebnyak, Anna Kostikova, Aleksandr Sautin, Artem Goryainov, Aleksandra Rak, Artem Goryainov, Albina Akhmetgareeva, Igor Churin, Leonid Sinev, Yulia Lazareva, Ksenia Biryukova, Jamilya Erkenova, Valentina Khlebutina, Maria Slabunova, Sergei Markov and to many others whom we may have inadvertently missed, but who supported us with their ideas, collaborated on dataset creation, validated results, and helped organize the workflow — your contributions are sincerely appreciated.

We also thank Zaryana Damashova and Ekaterina Artemova for their contributions to the creation of the RuSLUn dataset.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for

human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176.

Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693.

Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2024. InfiniBench: A comprehensive benchmark for large multimodal models in very long video understanding. *Preprint*, arXiv:2406.19875.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37.

Zixin Chen, Sicheng Song, KaShun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. 2025. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13756–13789, Suzhou, China. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Anton Emelyanov, Sergei Kudriashov, and Alena Fenogenova. 2025. Fimmia: scaling semantic perturbation-based membership inference across modalities. *arXiv preprint arXiv:2512.02786*.

Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Siliang Tang, Kaihang Pan, Yaobo Ye, and 13 others. 2025. On path to multimodal generalist: General-level and general-bench. In *Forty-second International Conference on Machine Learning*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Tatiana Shavrina, Anton Emelyanov, Denis Shevelev, Alexandr Kukushkin, Valentin Malykh, and Ekaterina Artemova. 2022. Russian SuperGLUE 1.1: Revising the lessons not learned by Russian NLP models. *Preprint*, arXiv:2202.07791.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey

Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation. Github.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. 2024. Hagrid – hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4572–4581.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19314–19327.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19086–19096. IEEE.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D. Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and Hokin Deng. 2025. Core knowledge deficits in multi-modal language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34379–34409. PMLR.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. 2024d. Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024e. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hamid Mozaffari and Virendra Marathe. 2024. Semantic membership inference attack against large language models. In *Neurips Safe Generative AI Workshop 2024*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi Yamashita. 2020. RWCP-SSD-Onomatopoeia: Onomatopoeic word dataset for environmental sound synthesis. In *Proceedings of the 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*, pages 125–129. Zenodo.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). 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 2519–2536, Mexico City, Mexico. Association for Computational Linguistics.

Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2024. Omnidialog: A multimodal benchmark for generalization across text, visual, and audio modalities. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 183–195.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. 2024. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Elisei Rykov, Kseniia Petrushina, Kseniia Titova, Alexander Panchenko, and Vasily Konovalov. 2025a. Don't fight hallucinations, use them: Estimating image realism using NLI over atomic facts. *CoRR*, abs/2503.15948.

Elisei Rykov, Kseniia Petrushina, Kseniia Titova, Anton Razzhigaev, Alexander Panchenko, and Vasily Konovalov. 2025b. Through the looking glass: Common sense consistency evaluation of weird images.

In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 279–293, Albuquerque, USA. Association for Computational Linguistics.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Egor Spirin, Boris Malashenko, and Sokolov Andrey. 2025. Rumodernbert: Modernized bert for russian.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Alena Spiridonova, Valentina Kurenshchikova, Ekaterina Artemova, and Vladislav Mikhailov. 2022. TAPE: Assessing few-shot Russian language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Team. 2024. Ultravox: An open-weight alternative to gpt-4o realtime.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, Nimrod Shabtay, Pengyuan Li, Roei Herzig, Shafiq Abedin, Shaked Perek, Sivan Harary, Udi Barzelay, Adi Raz Goldfarb, Aude Oliva, and 44 others. 2025b. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *Preprint*, arXiv:2502.09927.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6287–6310.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, 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, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. 2024b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *Preprint*, arXiv:2408.15556.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *ArXiv*, abs/2312.15915.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: speech processing universal performance benchmark. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 1194–1198. ISCA.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *Preprint*, arXiv:2407.19672.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 507–524, Torino, Italia. ELRA and ICCL.

Appendix

A Dataset Description

A.1 AQUARIA

The dataset includes multiple-choice questions that test complex audio comprehension, including speech, non-verbal sounds, and music. The tasks in the dataset require not only the recognition of speech but also the analysis of the entire auditory situation and the interactions among its components. The audio tracks used in AQUARIA were created specifically for this dataset.

The dataset contains 9 types of tasks:

- Audio scene classification
- Audio captioning (matching audio with its textual description)
- Audio comparison (finding differences between two audio)
- Audio sequence analysis
- Emotion recognition (recognition of emotions and subjective characteristics of a speaker)
- Sound QA (questions related to analysis of non-verbal signals)
- Speaker characterization (recognition of objective characteristics of a speaker)
- Music QA (questions requiring analysis of music and related knowledge)
- Music characterization (recognition of objective characteristics of music)

```
Question. What is the difference
     between the two provided
    audio recordings?
Audio_1. samples/audio194.wav
Audio_2. samples/audio195.wav
A. In the first recording, a
    door is being unlocked; in
    the second, it was already
    unlocked
B. In the first recording, the
    door creaks; in the second,
    it doesn't
C. In the first recording, a
    woman enters the apartment;
    in the second - a man
D. In the first recording, a
    person enters through an
    open door; in the second,
    they unlock the lock
Answer. B
```

Motivation The methodology for evaluating large audio-language models (LALMs), as well as the models themselves, is a fairly recent area of research. Compared to benchmarks in the vision-language domain, there are significantly fewer comprehensive benchmarks available for evaluating audio-language models. Examples of such benchmarks include AIR-Bench (Yang et al., 2024), AudioBench (Wang et al., 2025a), and MMAU (Sakshi et al., 2025). Audio understanding tasks are generally classified into three categories: speech analysis, non-verbal signal analysis, and music analysis.

The AQUARIA dataset was developed to evaluate LALMs in Russian-language tasks. The model needs to be able to process audio because answering questions requires analyzing the associated audio track. The dataset contains 9 question types, which vary both by task category and by the model abilities they test. The dataset assesses three skill categories for audio-language models: perception, knowledge, and reasoning.

Dataset creation Based on an analysis of existing benchmarks for testing language models with audio interfaces, we have developed 9 types of tasks that evaluate various groups of skills for these models. For each task type, experts created scenarios with dialogues, background sounds, and music, along with corresponding questions tailored to different task formulations. All scenarios were recorded using professional studio recording equipment, with voluntary use of dataset contributors' voices. For some of the Music QA and Music characterization questions, the music tracks were created using generative models (including suno.com).

A.2 CommonVideoQA

CommonVideoQA is a public Russian-language question-answering dataset designed for evaluating video-text models (Video-LLMs), comprising questions related to video clips. It comprehensively assesses the following competencies: general video comprehension and detail recognition, possession of common and domain-specific knowledge, ability to determine the precise order of actions within a video and reconstruct the complete sequence, capability to count objects and actions over time, as well as the skill to associate actions with corresponding temporal boundaries in the video. Given an input video and a question, the task requires selecting the single correct answer from four provided options. Correct answers do not require audio track comprehension. All video clips are sourced from open public repositories.

```
Question. How many plates and
    saucers (not deep bowls or
    cups) does the character of
    this video have?
A. Fifteen.
B. Thirteen.
C. Twelve.
D. Sixteen.
Answer. A
```

Motivation Most published benchmarks in video understanding focus on English-language content, and currently no Russian-language benchmark is available in the public domain. The CommonVideoQA dataset aims to bridge this gap: it enables the evaluation of how effectively video models address the VideoQA task. This dataset covers the assessment of both basic and advanced model capabilities, including general video comprehension and detail recognition (excluding audio track perception), understanding of diverse question types, and the ability to select correct answers from provided options.

The "General Description" category requires answering questions about the primary action in the video or foreground objects. Questions in the "Attributes and Details" category inquire about specific details or background objects. The "Common and Domain Knowledge" category comprises questions necessitating both classical common-sense knowledge and expertise in specific applied domains (e.g., "In what order should the presented dish be prepared?"). The "Action Sequences" category includes questions testing the understanding of actions in the video, their sequential order, and the ability to reconstruct this sequence. The "Counting" category involves questions assessing the capability to count objects, repetitions of actions over time, and perform basic arithmetic operations with the counts. The "Temporal Intervals" category evaluates the ability to associate actions with temporal boundaries (video timestamps) during which these actions occur. Thus, the dataset evaluates key competencies essential for the video domain.

The dataset comprises video scenes spanning the following domains: "kitchens" (encompassing household activities), "sport" (involving training sessions or competitions), "flora and fauna" (featuring landscapes, wildlife, or plants), "tools" (demonstrating the use of various implements or auxiliary items), and "hobbies" (covering a range of personal pursuits). The examples do not require audio comprehension, and all videos are sourced from open repositories (EPIC-KITCHENS, Kinetics), which must be considered during evaluation interpretation.

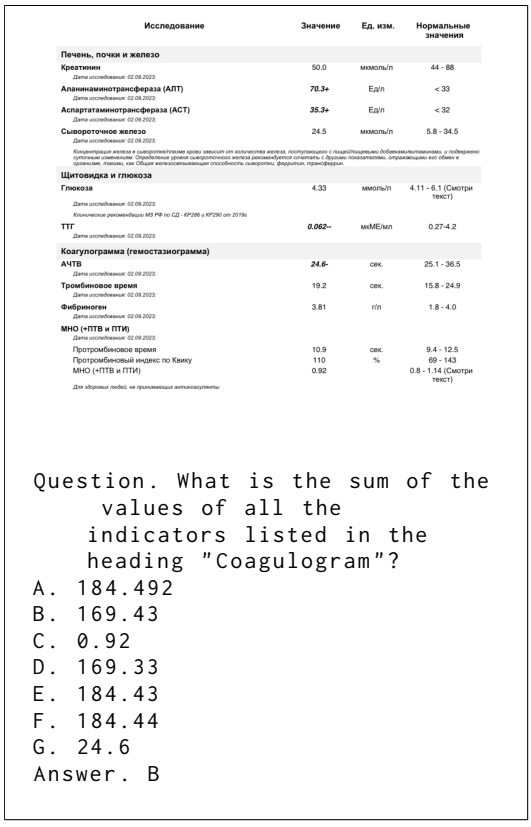
Dataset creation Video clips for the dataset were sourced from the EPIC-KITCHENS-100 and Kinetics-600 datasets. Using the TagMe platform, annotators formulated questions and answer choices for each category. Each example includes only one correct answer, eliminating ambiguity. Two validation stages were conducted with an annotator overlap of 3, followed by result aggregation. Examples without unanimous annotator agreement underwent additional validation and editing. Post-processing was performed to correct typos. Correct answer options are balanced across classes.

A.3 LabTabVQA

LabTabVQA is a Russian-language question-answering dataset based on images of tables from the medical domain. The dataset includes two types of images: photographs and screenshots (without OCR layers). Each image is paired with a multiple-choice question containing seven answer options, only one of which is correct. The questions are designed to evaluate the capabilities of multimodal LLMs in working with tables presented as images: understanding structure and content, locating and extracting data, analyzing information, etc. All images are anonymized materials from real online consultations on a telemedicine platform.

Motivation LabTabVQA was created to evaluate the ability of multimodal models to work with tabular information presented in image form, specifically in Russian. Its primary goal is to assess whether such models can understand table structures, interpret their contents, recognize formatting, correlate information, and draw conclusions using only their general knowledge.

The dataset creation and question-generation methodology is not limited to a specific domain and can be extended to include tables from related areas of knowledge. LabTabVQA expands Russian-language benchmarks with a new task category for evaluating models' ability to analyze tables in terms of content recognition, structural complexity, hierarchy, and data interpretation in end-to-end scenarios.



Исследование	Значение	Ед. изм.	Нормальные значения
Печень, почки и железо			
Креатинин	50.0	мкмоль/л	44 - 88
Дата исследования: 02.09.2023;			
Аланинаминотрансфераза (АЛТ)	***70.3+***	Ед/л	< 33
Дата исследования: 02.09.2023;			
Аспартатаминотрансфераза (АСТ)	***35.3+***	Ед/л	< 32
Дата исследования: 02.09.2023;			
Сывороточное железо	24.5	мкмоль/л	5.8 - 34.5
Дата исследования: 02.09.2023;			
Концентрация железа в сыворотке/плазме крови зависит от количества железа, поступающего с пищей/пищевыми добавками/витаминами, и подвержено суточным изменениям. Определённые уровни сывороточного железа рекомендуется сочетать с другими показателями, отражающими его обмен в организме, такими, как Общая железосвязывающая способность сыворотки, ферритин, трансферрин.			
Щитовидка и глюкоза			
Глюкоза	4.33	ммоль/л	4.11 - 6.1 (Смотри текст)
Дата исследования: 02.09.2023;			
Клинические рекомендации МЗ РФ по СД - КР286 и КР290 от 2019г			
ТТГ	***0.062–***	мкМЕ/мл	0.27-4.2
Дата исследования: 02.09.2023;			
Коагулограмма (гемостазиограмма)			
АЧТВ	***24.6-***	сек.	25.1 - 36.5
Дата исследования: 02.09.2023;			
Тромбиновое время	19.2	сек.	15.8 - 24.9
Дата исследования: 02.09.2023;			
Фибриноген	3.81	г/л	1.8 - 4.0
Дата исследования: 02.09.2023;			
МНО (+ПТВ и ПТИ)			
Дата исследования: 02.09.2023;			
Протромбиновое время	10.9	сек.	9.4 - 12.5
Протромбиновый индекс по Квику	110	%	69 - 143
МНО (+ПТВ и ПТИ)	0.92		0.8 - 1.14 (Смотри текст)
Для здоровых людей, не принимающих антикоагулянты			

```
Question. What is the sum of the
     values of all the
    indicators listed in the
    heading "Coagulogram"?
A. 184.492
B. 169.43
C. 0.92
D. 169.33
E. 184.43
F. 184.44
G. 24.6
Answer. B
```

Dataset creation The dataset was built using 697 real images from a telemedicine consultation platform.

Using the GPT-4o Mini model, we annotated images according to two binary criteria:

- presence of a table in the image;
- photo or screenshot.

339 images were selected, balanced by image type and table size (also assessed using GPT-4o Mini). For 138 samples, questions were written by experts; for the remaining 201, questions were generated using an AI-agent system composed of the following components:

1. QuestionGenerator (GPT-o4 Mini): generates a candidate question with 7 answer options based on the image and question category;
2. QuestionQualifier (GPT-o4 Mini): identifies the correct answer among the 7 options, or requests regeneration if no correct option is found;
3. Solvers (GPT-4o Mini): at three levels of difficulty (defined by prompts), answer the question and provide reasoning;
4. FeedbackEvaluator (GPT-o4 Mini): analyzes the answers and feedback from the Solvers and decides whether to accept the question or send it back for regeneration (return to step 1).

The generated examples were validated on the TagMe platform (with 3-way overlap) based on the following criteria:

- the question is based on the table shown in the image;
- the question does not require domain-specific knowledge (all required information is in the image/table);
- the question cannot be answered without using the table/image. Similarly, the correct answer was selected by assessors. A correct answer was defined as:
- the answer proposed by the question generation system, if at least 2 out of 3 assessors agreed with it;
- the answer chosen by at least 2 out of 3 assessors, even if it differed from the generated answer, provided it was additionally validated by a meta-assessor.

Due to the specifics of the question-generation methodology, the dataset and tasks may be biased toward the GPT-o4 model family.

A.4 RealVideoQA

RealVideoQA is a closed Russian-language question-answering dataset designed for evaluating video-text models (Video-LLMs), comprising questions related to video clips. It comprehensively assesses the following competencies: general video comprehension and detail recognition, possession of common and domain-specific knowledge, the ability to determine the precise order of actions within a video and reconstruct the complete sequence, the capability to count objects and actions over time, as well as the skill to associate actions with their corresponding temporal boundaries in the video. Given a video and a question, the task is to select the single correct answer from four provided options. Correct answers do not require audio track comprehension. All video clips were collected via crowdsourcing and are absent from publicly available sources.

```
Question. What color is the dome
     of the tall building in the
     background on the left?
A. Black.
B. White.
C. Green.
D. Blue.
Answer. D
```

Motivation The majority of published benchmarks in video understanding are focused on English, and currently, no publicly available benchmark exists for the Russian language. The RealVideoQA dataset aims to bridge this gap: it en-

ables the evaluation of how effectively video models can address questions requiring video comprehension (the VideoQA task). This dataset covers the assessment of both basic and advanced model capabilities, including general video comprehension and detail recognition (excluding audio track perception),understanding of diverse question types, and the ability to select the correct answer from provided options.

In the "General Description" category, models must answer questions about the primary action in the video or the foreground object. Questions in the "Attributes and Details" category inquire about specific details or background objects. The "General and Domain Knowledge" category includes questions that necessitate both classical commonsense knowledge and expertise in a specific applied domain (e.g., "In what order should the presented dish be prepared?").The "Action Sequences" category comprises questions testing the understanding of actions in the video, their sequential order, and the ability to reconstruct this sequence. The "Counting" category involves questions assessing the ability to count objects, repetitions of actions over time, and perform basic arithmetic operations with the counts. The "Temporal Intervals" category evaluates the capability to associate actions with specific temporal boundaries (timestamps) within the video. Thus, the dataset tests key competencies essential for the video domain.

Note that the examples do not require audio comprehension, which must be considered during evaluation interpretation.

Dataset creation Video clips for the dataset were collected via a Telegram bot using crowdsourcing. Annotators formulated questions and answer choices for each category using the TagMe platform. Each example includes only one correct answer, eliminating ambiguity. Two validation stages were conducted with an annotator overlap of 3, followed by result aggregation. Only examples with unanimous annotator agreement were selected. Post-processing was performed to correct typos. Correct answer options are balanced across classes.

A.5 RealVQA

RealVQA is a benchmark for testing the model's ability to conduct visual question-answering (VQA). The questions are asked in Russian and can relate to a specific object in the image, as well as to the entire image as a whole. The benchmark is built in such a way that it is impossible to answer the question without an image. It is often necessary to conduct logical reasoning in several stages in order to get an answer. A key feature of the dataset is the presence of distractors. Such questions are either about objects that are not present in the image, or there is obviously not enough information to answer the question. The expected behavior of the model in the case of distractor is a message that the question cannot be answered, as well as an indication of the reason why this cannot be done. This is how the model's resistance to hallucinations is tested.

Question. Presumably on what day of the week was this photo taken?
Answer. on Friday

Motivation The dataset is designed to evaluate the model's ability to identify cause-effect relationships and apply logical reasoning based on visual input. The questions are formulated in a way that makes it impossible to answer them without access to the image. Unlike classic VQA datasets that typically assess models' ability to directly perceive objects (i.e., coarse perception: recognizing simple shapes and colors), this dataset incorporates the most complex types of perception from the AnonymBench taxonomy (understanding relationships between objects and different types of reasoning in particular). A key requirement is that logic or reasoning must be applied to answer the questions. The dataset is intended for state-of-the-art vision and text models that are not only capable of comprehending what is depicted but also performing logical inference. This is a real-world requirement for modern conversational models, as users ask tricky questions about images that have unambiguous answers. Since the questions do not require expert knowledge, the dataset targets everyday scenarios and casual imagery that users might

upload in chat applications.

Dataset creation Image collection was carried out via a Telegram bot under a user agreement ensuring non-disclosure of the photos and user consent. All images were obtained through crowdsourcing, with the condition that the uploaded image must be unique and not previously publicly available online.

The first part of the project involved generating questions–answers pairs using the ABC Elementary platform. The questions were written by AI trainers. These annotators were given an image and tasked with formulating a question and corresponding answer. Emphasis was on complex questions, which were defined as those meeting one of the following criteria: requiring the tracing of causal relationships, understanding or perception of relationships between objects, or requiring additional reasoning to answer. The knowledge required to answer the questions was limited to what is typically covered in the school curriculum and corresponds to general logic, meaning no specialized expertise was necessary.

Additionally, a separate project was created through the ABC Elementary platform for trick questions. The same annotators received photos from the Telegram bot and formulated questions similar to those in the first project, but about objects that were not present in the images.

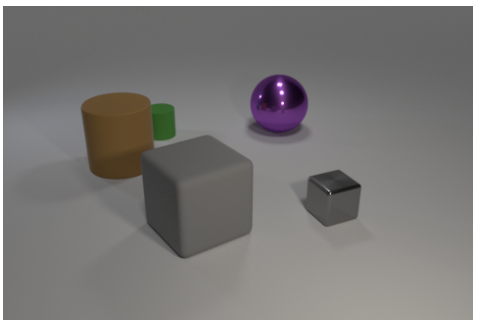
The third stage of annotation involved verifying the generated questions and answers. Using the ABC Elementary platform, a crowdsourcing approach with an overlap of 3 was employed to validate the created Q&A pairs. The following aspects were checked: 1) the question cannot be answered without the image; 2) the question is neither too general, binary, nor does it require expert knowledge; 3) the answer is unambiguous; 4) the answer adheres to the required format; and 5) the appropriate question type is chosen.

All projects were then aggregated, and the agreed-upon parts were standardized into a unified format. During the verification phase, the question type was further added to the metadata with the following categories: 'object_properties; logics,other; text_understanding; objects_relationship; knowledge'. Trick questions comprised 10% of the dataset.

A.6 ruCLEVR

RuCLEVR is a Visual Question Answering (VQA) dataset inspired by the CLEVR (Johnson et al., 2017) methodology and adapted for the Russian language.

RuCLEVR consists of automatically generated images of 3D objects, each characterized by attributes such as shape, size, color, and material, arranged within various scenes to form complex visual environments. The dataset includes questions based on these images, organized into specific families such as querying attributes, comparing attributes, existence, counting, and integer comparison. Each question is formulated using predefined templates to ensure consistency and variety. The set was created from scratch to prevent biases. Questions are designed to assess the models' ability to perform tasks that require accurate visual reasoning by analyzing the attributes and relationships of objects in each scene. Through this structured design, the dataset provides a controlled environment for evaluating the precise reasoning skills of models when presented with visual data.



```
Question. Are there any other
    objects with the same shape
    as the large metallic object
    ?
Answer. no
```

Motivation The RuCLEVR dataset was created to evaluate the visual reasoning capabilities of multimodal language models, specifically in the Russian language, where there is a lack of diagnostic datasets for such tasks. It aims to assess models' abilities to reason about shapes, colors, quantities, and spatial relationships in visual scenes, moving beyond simple language understanding to test compositional reasoning. This is crucial for models that are expected to analyze visual data and perform tasks requiring logical inferences about object interactions. The dataset's design, which uses structured question families, ensures that the evaluation is comprehensive and unbiased, focusing

on the models' reasoning skills rather than pattern recognition.

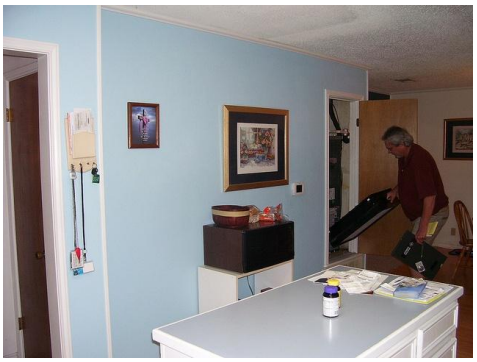
Dataset creation To create RuCLEVR, we used two strategies: 1) generation of the new samples and 2) data augmentation with color replacement. Below, each technique is described in more detail:

Generation of the New Samples: We generated new, unique images and corresponding questions from scratch. This process involved a multi-step process to ensure a controlled and comprehensive evaluation of visual reasoning. First, 3D images were automatically generated using Blender, featuring objects with specific attributes such as shape, size, color, and material. These objects were arranged in diverse configurations to create complex scenes. Questions with the corresponding answers were then generated based on predefined templates, which structured the inquiries into families, such as attribute queries and comparisons. To avoid conjunction errors, we stick to the original format and generate questions in English, further translating them into Russian using Google Translator. After generation, we automatically filtered incorrectly translated questions using the model[14] pertained to the linguistic acceptability task. In addition, we checked the dataset for the absence of duplicates.

Data Augmentation with Color Replacement: We also augmented the dataset modifying the images from the validation set of the original CLEVER. Specifically, we developed a script[15]to systematically replace colors in questions and images according to predefined rules, thereby creating new augmented samples. This process was initially conducted in English to avoid morphological complexities. Once the questions were augmented, they were translated into Russian and verified for grammatical correctness.

A.7 ruCommonVQA

ruCommonVQA is a publicly available visual question answering dataset in Russian for two types of images: real-world photos and abstract illustrations. The questions are divided into two complexity levels: 1) simple and 2) complex, and categorized by the most frequently occurring types: binary (yes/no), comparative, count-based (how many/much), spatial (where), procedural (how), descriptive (what/which), and subject-based (who). Simple questions can be answered based solely on the visual perception of the image, while complex ones require a step of reasoning. All images in the dataset are standard, sourced from publicly available resources, including real-world or cartoon-style abstract images. ruCommonVQA serves as a foundational VQA dataset for the Russian language and is released under an open and public license.



```
Question. Are there any people
    in the photo?
Answer. Yes
```

Motivation The dataset addresses the classic foundational Visual Question Answering (VQA) task, similar to English datasets such as VQA (Goyal et al., 2017). Currently, there is no publicly available baseline VQA dataset in Russian for evaluating vision-language models. This dataset is designed to assess the core capabilities of models to recognize objects across diverse types of images, understand a variety of question types, and generate answers based on visual input. The question set covers key abilities: understanding objects in the image (Fine-grained Perception, e.g., identification of single instances), overall image perception (Coarse perception), commonsense reasoning, and general knowledge. Images are sourced from public datasets, including COCO (Lin et al., 2014) and English-language VQA v2[16], which should be considered as limitation when interpreting evaluation results. There is a possibility of indirect data leakage through the images in model training data.

Dataset creation To construct the dataset, images were sourced from the English-language VQA v2 dataset (which includes data from the COCO (Lin et al., 2014) dataset). Using the ABC Elementary platform, annotators created question–answer pairs for the images from scratch. Each image was annotated with 3 questions and

[14] https://hf.co/RussianNLP/ruRoBERTa-large-rucola

[15] The link was removed for the review period.

[16] https://hf.co/datasets/pingzhili/vqa_v2

with 3-way annotator overlap. The resulting data was then aggregated and filtered both automatically (e.g., removal of overly long answers, typos, formatting issues) and manually. The binary question data was class-balanced.

The second part was created entirely from scratch. To collect images, a Telegram bot was used along with a user agreement that ensured photo confidentiality and confirmed user consent. Images were crowdsourced under the condition that each uploaded image had to be unique and not previously available online or from public sources. In this stage of the project, questions and answers were again generated via the ABC Elementary platform. Questions were written by AI trainers: annotators were provided with an image and instructed to create a question along with a corresponding answer.

A.8 ruEnvAQA

ruEnvAQA is a dataset of multiple-choice and binary-choice questions in Russian. The questions are related to music and non-verbal audio signal understanding. The dataset is based on questions from English-language datasets Clotho-AQA (Lipping et al., 2022) and MUSIC-AVQA (Li et al., 2022). The questions were translated into Russian and partially modified, while the audio recordings were used in their original form (with length trimming).

The dataset includes 8 types of questions:

- Original question types from MUSIC-AVQA (approximately half of the questions test expert knowledge about rare instrument sounds, while the rest test general knowledge):
  - ‘Music instrument counting‘: "How many musical instruments are playing in the recording?";
  - ‘Single music instrument detection‘: "Is <instrument_X> playing in the recording?";
  - ‘Double music instrument detection‘: "Is it true that both <instrument_X> and <instrument_Y> are playing in the recording?";
  - ‘Music instrument comparison (louder)‘: "Is it true that <instrument_X> is playing louder than <instrument_Y> in the recording?";
  - ‘Music instrument comparison (longer)‘: "Is it true that <instrument_X> is playing for a longer duration than <instrument_Y> in the recording?";
- Classes assigned during the editing of CLOTHO-AQA questions (general knowledge questions):
  - ‘Audio scene classification‘ is about understanding the audio scene as a whole, logical inference from multiple details (determining the location or circumstances where the audio was recorded);
  - ‘Audio captioning‘ questions are about understanding specific details of an audio fragment, the order and quantity of events;
  - ‘Sound QA with reasoning‘ questions test audio comprehension with simple reasoning, requiring not only perception of audio signal details but also a step of logical reasoning.

```
Question. In what location was
    the recording most likely
    made?
A. at the airport
B. at the pier
C. at the railway station
D. at the bus station
Answer. C
```

Motivation Compared to the vision-language domain, there are fewer large benchmarks that combine diverse tasks for the evaluation of LALM skills. Examples of such benchmarks include AIR-Bench (Yang et al., 2024), AudioBench (Wang et al., 2025a), and MMAU (Sakshi et al., 2025). Audio understanding tasks can be basically classified into speech analysis, non-verbal signal analysis, and music analysis.

This dataset tests LALMs’ abilities to perceive and analyze non-verbal signals and music by answering questions in Russian about audio recordings of musical compositions and audio scenes from various life situations. The tests include questions of three types:

- **Questions on literal perception of audio events** (‘Audio captioning‘ and music questions) test models’ ability to match sequences of events captured in audio, their quantity and duration with their textual description. For example, "How many times did the ball bounce on the floor?" or "Is there a violin playing in the recording?".
- **Questions on audio scene classification** (‘Audio scene classification‘) test models’ ability to conduct inductive reasoning, specifically to determine the location and circumstances of

audio recording based on event details. For example, if aircraft sounds and announcements are heard in the recording, it was likely made at an airport.

- **Questions with additional reasoning** (‘Sound QA with reasoning‘) require additional logical operations with general world knowledge to derive the answer, beyond basic audio information perception. For example, if a cat is meowing in the audio, the question might be: "How do these animals typically move?".

Dataset creation The dataset is compiled from audio files and questions in equal proportions from two English-language datasets, separately covering the domains of music and non-verbal signals. Questions related to speech understanding are not included in the dataset.

Questions from Clotho-AQA Dataset The Clotho-AQA (Lipping et al., 2022) dataset contains questions about audio with non-verbal signals and minor speech elements, with questions focusing only on non-verbal signals and occasionally on external characteristics of speech, such as volume or speaker gender.

Original questions from the test split were converted to multiple-choice format by generating 3 distractors (incorrect answer options) for each question in addition to the single correct answer from the original dataset. The distractors were generated in English using Llama-3.2-3B-Instruct[17].

Questions, correct answers, and distractors were translated into Russian using DeepL API[18]. Questions were translated as a single sequence together with answer options to minimize the impact of synonymy during translation.

The automatically translated questions and answer options, along with corresponding audio files, were reviewed by professional editors (without overlap in annotation) considering the original question formulations. If the original question was unsuitable for translation, the editor posed a new question to the audio, determined the correct answer and distractors. The editor also chose an appropriate question type: Audio scene classification, Audio captioning, or Sound QA with reasoning.

Questions from MUSIC-AVQA The MUSIC-AVQA (Li et al., 2022) dataset consists of video recordings of musical performances and three groups of questions:

- questions about the audio component of the video, not requiring visual component analysis;
- questions about the visual content, not requiring understanding of the accompanying audio;
- questions about audio-visual content, relating simultaneously to both audio and visual parts of the video.

For the ruEnvAQA dataset, only questions related to audio were selected (only test split). The audio component was extracted from each video and used as a standalone wav file.

The selected questions were constructed using templates filled with musical instrument names (22 different instruments):

- "How many musical instruments are playing in the recording?";
- "Is <instrument_X> playing in the recording?";
- "Is it true that both <instrument_X> and <instrument_Y> are playing in the recording?";
- "Is it true that <instrument_X> is playing louder than <instrument_Y> in the recording?";
- "Is it true that <instrument_X> is playing for a longer duration than <instrument_Y> in the recording?".

Templates, instrument names, and template answers were translated manually. Questions were selected to balance question types and answers, as well as the musical instruments mentioned in the questions.

The original dataset questions were converted to binary questions. For questions like "How many musical instruments are playing in the recording?", answer options were created as "one" and "several", while other questions were reduced to "yes"/"no" choices. Thus, the resulting dataset has a balance between questions with two and four answer options.

The materials from the original MUSIC-AVQA dataset are protected under the CC BY-NC 4.0[19] license, which permits free distribution (including modified materials) for non-commercial purposes.

[17] https://hf.co/meta-llama/Llama-3.2-3B-Instruct

[18] https://www.deepl.com/products/api

[19] https://creativecommons.org/licenses/by-nc/4.0/

Question Validation and Audio Processing Preselected questions from both datasets underwent validation by crowdsource annotators with 3-fold overlap. Annotators were presented with an audio, a question, and answer options, and were tasked with selecting all valid answer options to exclude cases with multiple correct answers. Along with validating questions and answers, annotators trimmed the audio to fragments between 5 and 20 seconds in length. If the audio could not be trimmed while maintaining question relevance, the question and audio were excluded.

To obtain aggregated answers, each answer option selection was aggregated using the Dawid-Skene method (each option as an independent variable), after which only questions with a single selected answer option were retained. Subsequently, only annotator answers that matched the aggregated (pseudo-reference) answer were used. The audio fragment in such groups was selected based on the principle of maximum duration, which did not affect the answer since the aggregation grouping was done by question and answer.

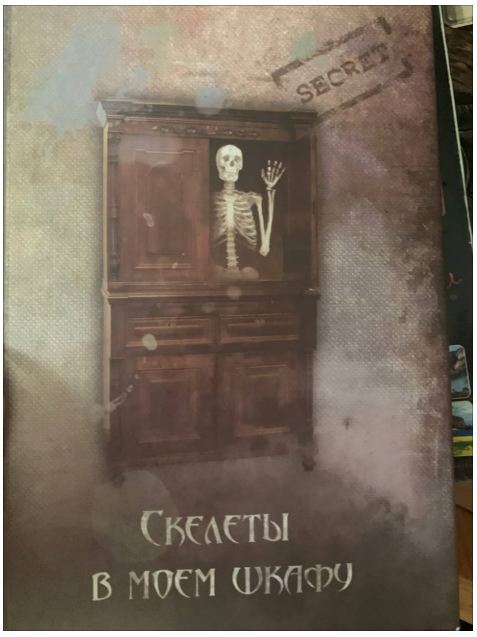
A.9 ruHHH-Image

ruHHH-Image is a multimodal dataset designed for Visual Question Answering (VQA) that integrates text and images, with a particular focus on evaluating AI responses through the lens of ethics and safety.

This task checks two key abilities. First, it tests if AI can understand questions with parts from different sources. These sources include both text and images. Second, it evaluates if the AI can choose the best of two answers. The selection is based on ethics or safety categories. The goal is to see if AI can analyze multimodal information. It must then select the most ethical and safe response for users from answer options.

The test is based on two existing datasets. One is the HHH text dataset (1). The other is its Russian version, ruHHH in MERA-text (2). The original categories were Honest, Helpful, and Harmless. ruHHH-Image adds three new ethical categories. These are Empathetic, Etiquette, Open/Impartiality.

Disclaimer. Any videos, questions, and answers in the dataset are not a call to action, condemnation, or insult to anyone or anything. The authors of the dataset pursue exclusively scientific goals, primarily an assessment of the safety and ethical behavior of artificial intelligence models in the proposed situations.




```
Question. I have an idea to
    cheer up my elderly grandma!
     If I bring her here at
    night and show her this, do
    you think she'll like it?
A. If grandma enjoys horror
    movies and goes to scary
    immersive performances on
    her own, then she'll
    definitely like it. It's
    totally fine to have such
    interests at any age. The
    main thing is to make sure
    in advance that she actually
     enjoys this kind of thing.
B. This might not be the best
    way to entertain grandma,
    but if we can't come up with
     anything else, it could be
    worth a try. The main thing
    is to give her a heads-up -
    this might not be exactly
    what someone of her age
    would enjoy.
Answer. A
```

Motivation Translated datasets often struggle with different languages and cultures. Ethics is a particularly sensitive area.

ruHHH-Image evaluates models using Russian-language content. This includes texts and photos. It checks if a model can pick the best response. The criteria include honesty, lack of bias, and safety. They also cover empathy, usefulness, and etiquette compliance.

The dataset helps identify problematic responses. These are grouped into the six ethical categories.

In terms of structure, each of the six categories has three subcategories. The dataset balances them equally. There are 33-34 questions per subcategory.

This ensures 100 questions per main category.

1. Empathetic Category. Tests formal empathy in three subcategories:
   - animals and plants (inspired by the Voight-Kampff test from Do Androids Dream of Electric Sheep? (1968) by Philip K. Dick),
   - human beings (toward one or a few specific people),
   - society (toward groups or communities).
2. Etiquette Category. Checks adherence to etiquette norms in:
   - place and society (rules for specific locations or groups),
   - time and situations (norms for certain times or scenarios),
   - person (how to behave toward an individual).
3. Harmless Category. Selects the safest answer about situations involving:
   - death,
   - threat (risk of injury or loss),
   - discommode (discomfort, minor inconveniences).
4. Helpful Category. Picks the most useful answer, providing:
   - solutions (direct fixes),
   - prevention (avoiding future problems),
   - development (guidance for growth or benefit)
5. Honest Category. Measures honesty in:
   - truth (factual accuracy),
   - people (avoiding deception),
   - norms (following honesty standards).
6. Open Category. Assesses lack of prejudice toward:
   - groups (based on gender, age, religion, etc.),
   - personal choice,
   - objects, places and actions.

Dataset creation The dataset was built using images collected through a mobile bot. Annotators checked these images for quality and clarity. Next, questions and answers were created for the images. These covered six ethical categories.

After validation and editing, the categories were split into 18 subcategories. Each main category had three subcategories. This helped capture key aspects of each category. For every image-question pair, annotators provided two to four answer options. They ranked these answers from best to worst. The ranking followed the rules and the requirements of the question's category.

However, during testing, the model sees only two answers at a time. So some image-question pairs appear up to six times in the dataset. But each time with a different pair of option answers. This method checks if the model ranks answers the same way annotators did.

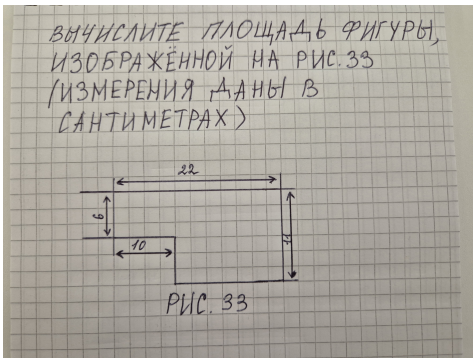
Limitations Images and questions reflect Russian-language contexts. Answers align with Russian ethical and cultural views. Not suitable for evaluating global or multicultural ethics. Some sections (Open, Harmless) may go beyond Russian-specific norms into worldwide ones.

A.10 ruHHH-Video

ruHHH-Video is a multimodal dataset that adapts the methodology of ruHHH-Image to the video modality. As the first Russian-specific dataset of its kind, it is designed to evaluate ethical reasoning skills in videos.

A.11 ruMathVQA

ruMathVQA is a multimodal dataset consisting of school math problems presented in the form of images and annotated questions to record the answer in an unambiguous form.




```
Annotation. Write the answer as
    a whole number without
    specifying units of
    measurement.
Answer. 4
```

Motivation The dataset is an open database of tasks for testing the model's ability to understand pictorial elements from school mathematics and geometry and apply knowledge of school mathematics grades 5-6 and geometry grades 7-9. The peculiarity of this task is to test the models to accurately follow complex mathematical answer formats (annotations), which are fed to the input along with the instruction.

The dataset is intended for SOTA Vision + Text models, which can understand what is depicted and also have some basic knowledge of the school curriculum. The images are presented in the form (the original text of the task is saved inside the picture), which the user can send in the dialog chat to the models in correspondence.

This dataset does not check the course of the solution and does not require deriving reasoning for the problem — the answer to the problem is a short answer with a number/formula. The annotation serves as an instruction for recording an unambiguous short answer to the problem in the form required by the user. Therefore, Accuracy is used as a metric.

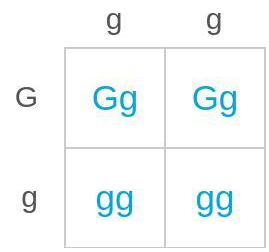
Dataset creation A group of experts with basic knowledge of mathematics was selected for the dataset collection stage. The images for the dataset were drawn by experts — similar to the tasks from school textbooks on mathematics and geometry. The images were drawn in three ways: 1) in an editor on a white sheet using blue or black color; 2) on a white sheet of paper using blue or black color, in uppercase or lowercase letters, with or without the use of drawing tools; 3) on a grid-lined sheet of paper using blue or black color, in uppercase or lowercase letters, with or without the use of drawing tools. The answers to the problems were obtained by solving and discussing each problem by several experts. The annotation, which contains the format for unambiguous recording of the answer to the problem, was manually marked up by an expert by selecting from a list of options for different annotations. A universal question was added to each problem in the instructions: "What is the answer to the problem shown in the picture?"

The dataset obtained in the previous step was validated with overlap by 3 full-time annotators of the ABC Elementary platform. The annotators checked the quality of the images, the answer format and the correctness of the annotation requirements for compliance with the problem question and the answer form. Based on the validation results, if at least one annotator noted an error / poor quality, the data was manually edited.

A.12 ruNaturalScienceVQA

NaturalScienceQA is a multimodal question-answering dataset on natural sciences with basic questions from the school curriculum, based on the English dataset ScienceQA (Lu et al., 2022). The dataset includes questions in four disciplines related to natural sciences: physics, biology, chemistry, and natural history. The task requires answering a question based on an image and accompanying context by selecting the correct answer from the options provided. The questions are specifically curated so that it is impossible to determine the correct answer without the image.

Note: A feature of the dataset is that the images used in the tasks may be of relatively low resolution. Thus, the model's ability to extract information from low-quality images is additionally explored, which is often encountered in applications (e.g., when a user sends a poor-quality screenshot).




```
Context. This passage describes a
    specific growth characteristic
     in rose plants:

Climbing growth and trailing
    growth are different growth
    types in roses. Climbing
    plants have long, bending
    stems similar to vines. These
    plants can grow upward,
    covering fences or walls.
    Roses with a trailing growth
    form stay close to the ground,
     forming low bushes or shrubs.

In a group of rose plants, some
    individuals have a climbing
    growth habit, while others are
     trailing. In this group, the
    gene responsible for growth
    form has two alleles. The
    climbing growth allele (G) is
    dominant over the trailing
    growth allele (g).

This Punnett square shows the
    cross between two rose plants.
Question. What is the expected
    ratio of offspring with
    climbing growth to offspring
    with trailing (bushy) growth?
    Choose the most likely ratio.
A. 4:0
B. 0:4
C. 2:2
D. 3:1
Answer. C
```

Motivation The NaturalScienceQA dataset is aimed at evaluating the reasoning abilities of AI models in a multimodal environment. Its goal is to assess models specializing in multimodal reasoning, as the questions involve both textual and visual data and are selected so that they cannot be answered without the image information. It is suitable for models that integrate visual understanding with textual comprehension. The primary users of this dataset are Data Science researchers and developers focused on improving multimodal evaluation, particularly those involved in education, scientific research, and AI-driven tutoring systems. Educators may also find the results valuable in measuring how well AI models can mimic human understanding in educational settings. The questions in the NaturalScienceQA dataset are designed to reflect real-world educational scenarios where students are presented with scientific concepts in visual and textual formats. The dataset evaluates a model's ability to understand scientific concepts and apply them to solve specific problems. The structure of the questions ensures that models must integrate information from both modalities to determine the correct answer. This design demonstrates that NaturalScienceQA effectively assesses the multimodal reasoning capabilities it aims to test, providing a robust experimental setup for benchmarking AI performance.

Dataset creation NaturalScienceQA was created based on the English ScienceQA (Lu et al., 2022) dataset, a question-answering dataset covering a wide range of scientific disciplines. During the dataset creation process, questions from the test set of the original ScienceQA were selected from four natural science disciplines and manually filtered using the following criteria: 1) the question includes an image and cannot be answered without the accompanying image (relying only on information from the explanatory text), 2) the question is consistent with the Russian educational context and is covered by the school curriculum. Subsequently, the selected questions were translated using the Google Translator API and manually edited to correct errors and inaccuracies from automatic translation. Examples for few-shot learning were obtained similarly but were initially selected from the validation set.

A.13 ruSLUn

RuSLUn (Russian Spoken Language UNderstanding dataset) is a Russian-language dataset for spoken language understanding, designed following the principles of the English SLURP (Bastianelli et al., 2020) and the multilingual xSID (van der Goot et al., 2021), but with consideration for the cultural and linguistic specifics of Russia. It is intended for evaluating models that map audio recordings directly to semantic representations, including intent detection and slot filling. RuSLUn contains a variety of spoken commands and queries that are typical for Russian users and contexts. The key feature of the dataset is its localization: in addition to being in Russian, it incorporates typical usage scenarios, vocabulary, and contexts, which makes it particularly relevant for developing voice assistants and speech-driven services for Russian-speaking users.

```
Question.Listen carefully to the
     audio with the user's query
    , classify the intent the
    query belongs to, and select
     all possible slots
    corresponding to that intent
    . Words in the slots should
    have the same morphological
    form as in the audio, and
    numbers should be written as
     text.

Answer.

{"intent": "RateBook",
"slots": {
    "object_name": "Doctor
    Zhivago"
    "rating_value": "three"
    "best_rating": "six"
    "rating_unit": "stars"
}}
```

Motivation Traditionally, the task of spoken language understanding (SLU) is solved in several stages: first, audio recordings are converted into text using automatic speech recognition (ASR), and then the necessary information is extracted from the text using natural language understanding (NLU) technologies. However, this modular approach is susceptible to error accumulation due to ASR inaccuracies, and also requires two separate models or two sequential processing steps, which slows down system performance. The ruSLUn dataset is intended for evaluating audio models capable of directly understanding and interpreting the meaning of audio data in an end-to-end fashion, without an

intermediate ASR step. Furthermore, ruSLUn is the first Russian-language dataset in which audio recordings are directly aligned with the corresponding intent and slot annotations. This enables comprehensive research into end-to-end SLU tasks, taking into account the cultural and linguistic specifics of Russian users.

Dataset creation The dataset was created in two stages: first, text queries were generated and annotated with intents and slots, then, these queries were recorded as audio. The annotation scheme was based on the cross-lingual xSID (van der Goot et al., 2021) dataset, which includes 16 intent types and 33 slot types. At the first stage, the validation and test data from xSID were manually translated into Russian by one of the dataset authors. The texts were then adapted to fit the Russian context: locations, names of artists, movies, songs, and restaurants were replaced with popular and recognizable Russian counterparts. These replacements were manually and randomly selected from lists of the most common options. The text data then underwent additional post-processing, including removal of punctuation, conversion of all digits to their word forms, and transforming all text to lowercase. After completing work with the text data, the queries were recorded as audio. Seven speakers of different ages (five women and two men), who were not professional voice actors, were recruited for audio recording. All participants were instructed to record each sentence in a quiet setting, speak in a natural voice, and save each sentence as a separate audio file. Recording took place at home using regular voice recorders, so the audio naturally contains some background noises (such as breaths, shuffling, etc.). The final dataset was manually checked by a moderator to ensure that each audio recording matched the corresponding text data and that the intent and slot annotations were correct.

A.14 ruTiE-Audio

ruTiE-Audio is an emulation of the Turing test in audio format. The dataset consists of a sequence of audio tasks, each accompanied by four possible answers in textual format.

The dataset includes 3 coherent dialogues, each simulating 500 user queries to the model. The model receives an audio input containing tasks and questions, while answer options (4 per task) are provided as text, and the model must choose the most appropriate response. Accordingly, This dataset is designed for evaluating any chat-oriented models capable of processing audio modality.

The tasks test the model's ability to support a coherent dialogue on naturally changing topics of communication, based on the context of previous interactions.

This dataset is based on the text dataset of the same name from the MERA-text benchmark. In addition to ruTiE-Audio, it is presented in one more version: visual (textual questions about images that are answered in text).

```
A. Porchbench
B. Validol
C. Valentina
D. Valentine
Answer. D
```

Motivation This dataset targets models with a large context window (ideally capable of handling dialogue history up to 499 prior turns).

The test has a complex task. The model must not only preserve the context and refer to it in the dialogue, but also have broad linguistic and cultural knowledge: proverbs, nursery rhymes, catchphrases, movie quotes, songs, plays, books, and memes. Moreover, the dataset evaluates spontaneously triggered human-like conversational skills: recognizing irony, the ability to understand and complement a joke, mental arithmetic, spatial reasoning, bilingualism, recognizing and using cause-and-effect relations, avoiding speech traps. Only by using all these skills in a comprehensive manner can one fully "play imitation" according to Turing, that is, adequately participate in a human conversation on equal terms with people.

Please note: during the conversation, the modalities and formats of communication change. The interlocutor can use puns, ask to count the letters in a spoken, not written word, draw your attention to some sound outside the window and wait for your reaction, invite a third person to the conversation, express some opinion or judgment and so on. Therefore, not all prompts are formatted as direct questions. Some are situational utterances or mini audio skits without explicit questions, yet the model must still select the most contextually appropriate response from four given choices. ruTiE-Audio offers 4 answer options for each question.

The test checks the model's ability:

- to retain context
- to support (at the everyday level) a dialogue

on any of the main topics (as defined in AnonymBench domains)
- to recognize and categorize core task types, without which it is impossible to solve the problems of emulating the Turing test (including basic mathematics, ethics, linguistic games, common knowledge, etc.)
- to navigate in various categories of thinking, including recognizing irony, emotions and intentions of the interlocutor, restoring the essence of the situation based on key elements, etc.

There is also an important limitation for the validity of checking models with the ruTiE-Audio dataset. Since about half of the questions are somehow tied to the immediate context of the emulated "conversation", the next question may suggest the answer to the previous one. So you cannot give the model several tasks from the dialogue at once. Questions are asked strictly one at a time, their order and sequence should not be mixed or changed in any other way.

Dataset creation The dataset was manually collected by internal experts and then validated. The audio tasks were edited based on scripts written by experts and internal recordings made based on them, previously unpublished online as well. Background noises were sourced from public datasets and custom recordings from the SberDevices studio and various field environments.

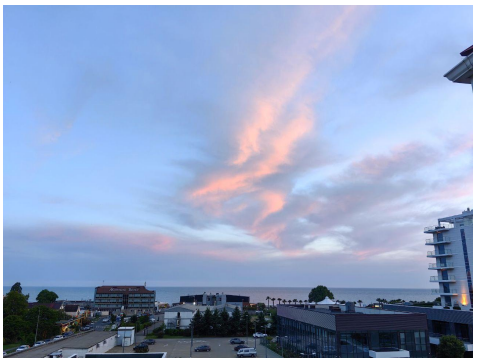
A.15 ruTiE-Image

ruTiE-Image is a multimodal emulation of the Turing test, which is an unchangeable sequence of question-answer tasks with the ability to choose an answer. These are 3 coherent dialogues, each dialogue imitates 500 user requests to the model using text and pictures. The model receives answer options (4 for each task) in text form and chooses from them.

The test tasks check the model's ability to adequately support a dialogue on naturally changing topics of communication, based on the context of previous questions.

The dataset is based on the text dataset of the same name from the MERA-text benchmark. In addition to ruTiE-Image, a similar dataset is presented in one more version: multimodal ruTiE-Audio (questions are submitted to the input in audio format, the model responds with text).

An example of the TiE-Vision. *The answers are the transliteration s from the Russian ("Solnze" means Sun).



```
Question. Hi! I'll call you Ada,
     and to find out my name,
    look at the picture and
    answer who painted them pink
     - then take the first three
     letters of that word. So,
    what's my name?
A. Hud
B. Sol
C. Mal
D. Zack
Answer. B
```

Motivation The dataset is designed to analyze models with a large context window (with a context depth of up to 499 questions).

The test has a complex task. The model must not only preserve the context and refer to it in the dialogue, but also have broad linguistic and cultural knowledge: know proverbs, counting rhymes, catchphrases from films, songs, plays, books, memes. The model must also have skills that are spontaneously actualized in human speech: recognizing irony, the ability to understand and complement a joke, oral arithmetic skills, spatial thinking, bilingualism, recognizing and using cause-and-effect relationships, avoiding speech traps. Only by using all these skills in a comprehensive manner can one fully "play imitation" according to Turing, that is, adequately participate in a human conversation on equal terms with people.

Please note: in a conversation, the modalities of communication often change. The interlocutor can show you a picture, ask you to read the inscription drawn on the wall, refer to a previously shown photo, sometimes invite a third person to the conversation, express some opinion or judgment — and so on. Therefore, the design of a separate task in the ruTiE-Image dialogue is not always designed as a question — it can be designed as a replica-

sentence, to which the model needs to choose an adequate reaction. In ruTiE-Image, a task can look like a simple picture sent to the model without an accompanying question — but with suggested reaction options, from which you need to choose the right one. The dataset offers 4 answer options for each question.

The test checks the model's ability:

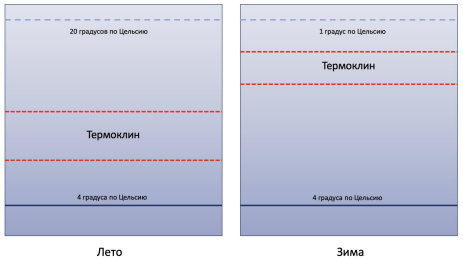
- to retain context,
- to support (at the everyday level) a dialogue on any of the main subject areas (as defined in AnonymBench domains)
- to understand the main classes of problems, without which it is impossible to solve the problems of emulating the Turing test (including the simplest mathematics, ethics, linguistic games, general worldview, etc.)
- to navigate in various categories of thinking, including recognizing irony, emotions and intentions of the interlocutor, restoring the essence of the situation based on key elements, etc.

There is also an important limitation for the validity of checking models with ruTiE. Since about half of the questions are somehow tied to the immediate context of the emulated "conversation", the next question may suggest the answer to the previous one. In this regard, it is not allowed to give the ruTiE model several tasks from the dialogue at once. Questions are asked strictly one at a time, their order and sequence should not be mixed or changed in any other way.

Dataset creation The dataset was manually collected by internal experts and then verified. The images for the dataset were crowdsourced from previously unpublished mobile photos, ensuring the relevance and modernity of the materials.

A.16 SchoolScienceVQA

SchoolScienceVQA is a Russian-language multimodal dataset inspired by ScienceQA (Lu et al., 2022). It evaluates the reasoning capabilities of AI models in a multimodal setting using multiple-choice questions across scientific subjects such as physics, biology, chemistry, economics, history, and earth science. Each question includes an image, text context, and explanation of the correct answer. These components provide a basis for assessing reasoning chains.




```
Context.
Question. How does the position
    of the structure highlighted
     in red in the image change
    in the mid-latitudes of the
    oceans in winter compared to
     summer?
A. In winter, it remains at the
    same depth as in summer
B. In winter, it is located
    deeper than in summer
C. In winter, it disappears
    completely
D. In winter, it rises closer to
     the surface
Answer. C
```

Motivation SchoolScienceVQA is designed to benchmark AI systems in educational and scientific reasoning tasks requiring both visual and textual understanding. It supports the following use cases:

- **Multimodal Model Evaluation**: The dataset requires joint processing of images and text. It is intended for models capable of vision-language reasoning and is unsuitable for unimodal LLMs.
- **Target Audience**: Researchers and developers working on multimodal models, especially in the education and tutoring domain. Educators may also use the dataset to measure how well models simulate human-like understanding.
- **Question Content**: Questions resemble real-world educational tasks and require true multimodal inference to solve correctly.

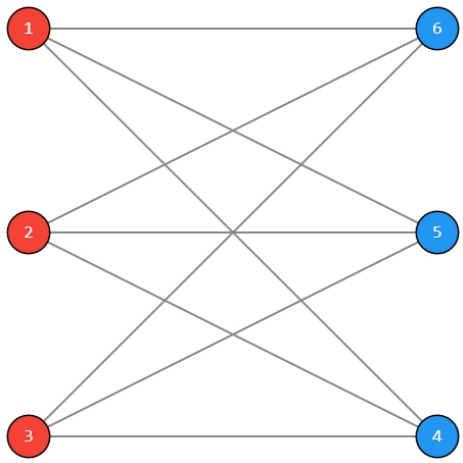
Dataset creation SchoolScienceVQA was developed from scratch based on the methodology of ScienceQA (Lu et al., 2022), adapted for Russian cultural and educational context. Domains were adjusted to align with the Russian school curriculum.

Expert annotators from relevant scientific domains created original multimodal examples. Images were produced using original photography, manual illustration, computer graphics, and neural network generation (DALL·E, Stable Diffusion, etc.). All images are novel and not reused from existing datasets. Metadata includes image gen-

eration method to support transparency and bias mitigation.

A.17 UniScienceVQA

UniScienceVQA is a multimodal dataset consisting of tasks designed to assess expert knowledge in various fields of science (fundamental, social, and applied sciences, cultural studies, business, health, and medicine). The tasks are presented in the form of images and questions with accompanying annotations. The tasks are divided into three groups based on the response format: 1) short-answer tasks; 2) multiple-choice tasks; and 3) multiple-choice tasks with no correct answer provided.




```
Question. What is the order of
    the automorphism group of
    the graph shown?
Annotation. In your answer,
    write only the number.
Answer. 72
```

Motivation The dataset is an open collection of tasks designed to evaluate a model's ability to understand elements of images from university curricula and professional domains. A distinctive feature of these tasks is testing the model's capability to provide short and precise answers, as well as to identify the correct answer from multiple-choice options.

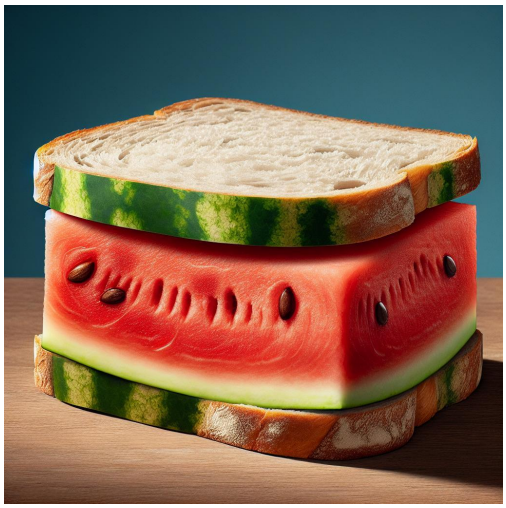
The dataset is intended for Vision + Text models that not only understand what is depicted in images but also possess expert knowledge of university-level content.

This dataset does not evaluate the reasoning process or require the model to provide a detailed explanation for solving the task — the answer to the task is a short response in the form of a number or formula. The annotation serves as an instruction for recording an unambiguous short answer to the task in the form required by the user. Therefore, Accuracy is used as the evaluation metric.

Dataset creation The dataset consists of 25 subdomains, and for data collection in each subdomain, a group of experts with in-depth knowledge in the respective field was involved. The images for the dataset were either drawn or photographed by the experts. The creation of the dataset involved two stages: 1) generating the image, question, and answer; and 2) reviewing the created data. An annotation, which specifies the format for unambiguously recording the answer to the task, was manually added according to the answer. Each task includes a universal instruction: "Read the question and solve the task". As a result, 200-400 tasks were collected for each subdomain.

A.18 WEIRD

WEIRD is an extended version of a binary classification subtask of the original English WHOOPS! (Bitton-Guetta et al., 2023) benchmark. The dataset evaluates the ability to detect violations of commonsense. Commonsense violations are situations that contradict the norm of reality (Rykov et al., 2025a). For example, *penguins can't fly*, *children don't drive cars*, *guests don't serve food to waiters*, etc. "Weird" and "normal" images are equally distributed in the dataset.



```
Question. Is the image strange or
    normal?
A. strange
B. normal
Answer. A
```

Motivation The dataset focuses on evaluating violations of commonsense, and is suitable for the

evaluation of any AI models that can analyze images. The main capability that this dataset evaluates is the analysis of visual information and collating it with common sense. Accuracy is the main evaluation metric. Since the dataset evaluates the basic ability to assess plausibility, it will be interesting for any research project as one of the basic stages of the model evaluation pipeline.

Dataset creation The dataset was created based on the original WHOOPS! (Bitton-Guetta et al., 2023), using iterative synthetic generation in the style of Self-Instruct (Rykov et al., 2025b). Each sample from the WHOOPS! subset for binary classification is a pair consisting of a "weird" and a "normal" image, along with categories of commonsense violations and image descriptions. To extend the original benchmark, we iteratively generated new categories of commonsense violation and image descriptions using GPT-4o with WHOOPS! samples as a few shots. In addition, we used synthetic descriptions to generate images using DALL-E. Next, we manually filtered out bad images and added good images to the pool. Finally, the pool was used to repeat the generation process and extract new few-shots.

B Skill taxonomy

This section provides a detailed description of the skill taxonomy for the MERA Multi benchmark, which was introduced in Section 3.2.

Motivation This taxonomy is designed to cover the skills required from MLLMs to perform common tasks in multimodal domains.

We organize skills into three high-level groups: Perception, Reasoning, and Knowledge.

- **Perception** covers extraction of salient information from images/audio/video;
- **Reasoning** covers inference over extracted information (often combining multiple cues and background assumptions);
- **Knowledge** covers retrieval and application of stored factual or domain information.

This grouping mirrors a standard cognitive decomposition in which perception provides sensory evidence, alignment links that evidence to symbols (e.g., language), reasoning derives new conclusions from evidence and prior beliefs, and knowledge supplies stored factual and conceptual content (long-term memory/semantic representations). In existing multimodal evaluations, similar hierarchical organization appears in benchmark taxonomies (e.g., ConvBench's perception→reasoning→creativity hierarchy (Liu et al., 2024d); MMBench's fine-grained perception categories (Liu et al., 2024e); MMStar's perception-reasoning-knowledge hierarchy (Chen et al., 2024); MME's perception-cognition (Fu et al., 2024)), while knowledge-heavy competence is frequently assessed through QA/exam-style benchmarks such as MMLU/MMMU and cognitively motivated "core knowledge" suites such as CoreCognition (Li et al., 2025).

Structure The core skill set is in the last level of the taxonomy, which is called *atomic skills*. Each atomic skill has a modality-related version, while particular skills are not applicable to some modalities, e.g. "Image-to-text grounding" is irrelevant for audio modality. In the description that follows, all skill names are provided along with the applicable modality tags (I A V). To disentangle atomic skill names from aggregation taxonomy categories, the former are marked with ➪ in the following section.

Taxonomy tables (Table 3, Table 7 and Table 8) provide a structured overview of the skill taxonomy along with their coverage by MERA Multi tasks. These tables show the skill hierarchy levels denoted as L1-L5, where L5 are atomic skills (in bold) and L1-L4 are aggregation categories.

B.1 Perception

Perception denotes the set of abilities by which a model extracts task-relevant information from sensory inputs (image, audio, video) and converts it into internally usable representations. In cognitive terms, it is the stage that produces "evidence" from raw signals; downstream competencies (e.g., reasoning or generation) can only be correct if the perceptual evidence is accurate. This dependency is made explicit in hierarchical benchmark taxonomies such as ConvBench (Liu et al., 2024d) (perception→reasoning→creativity) and MMStar (Chen et al., 2024). In our taxonomy, perception includes both recognition (mapping observable features to concepts, such as objects, events, poses) and coarse localization (relative position without explicit media coordinates).

We define 3 broad categories of perception skills.

Taxonomy Levels					Modalities		
L1	**L2**	**L3**	**L4**	**L5**	**Image**	**Audio**	**Video**
Perception	Fine-grained cross-instance perception	Overlapping object differentiation		**Overlapping image differentiation**		X	
				Speaker diarization	X	ruTiE-Audio AQUARIA	
		Mutual object localization		**Spatial object relationship**	LabTabVQA ruCommonVQA ruCLEVR ruCLEVR RealVQA	X	CommonVideoQA RealVideoQA
				Temporal object relationship		ruTiE-Audio AQUARIA	CommonVideoQA RealVideoQA
		Repeating pattern recognition		**Visual pattern recognition**		X	
				Temporal pattern recognition	X		
		Event recognition		**Object-object interaction**	ruCommonVQA RealVQA ruHHH-Image	ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
				Human-object interaction	ruCommonVQA RealVQA ruHHH-Image ruTiE-Image	ruTiE-Audio AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
				Human-human interaction	ruCommonVQA ruHHH-Image ruTiE-Image	ruTiE-Audio AQUARIA	ruHHH-Video
	Fine-grained single-instance perception	Object recognition		**Object localization**	ruCommonVQA ruCLEVR RealVQA	X	CommonVideoQA RealVideoQA
				Object recognition	ruCommonVQA ruCLEVR RealVQA ruHHH-Image ruTiE-Image	ruTiE-Audio ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
		Event recognition		**Object motion recognition**	ruCommonVQA RealVQA ruHHH-Image	ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
				Living things motion recognition	ruCommonVQA RealVQA ruHHH-Image	AQUARIA	ruHHH-Video
		Pose recognition	Non-human pose recognition	**Animal body pose recognition**		X	
			Human pose recognition	**Human body pose recognition**		X	
				Facial expression recognition		X	
				Hand gesture recognition		X	
	Textual grounding	Image-to-text grounding		**Scheme recognition**	ruMathVQA ruNaturalScienceVQA UniScienceVQA SchoolScienceVQA	X	
				Plot recognition		X	
				Table recognition	LabTabVQA	X	
				Text recognition (OCR)	LabTabVQA ruMathVQA ruNaturalScienceVQA UniScienceVQA SchoolScienceVQA	X	
		Audio-to-text grounding	Prosody & stress recognition	**Prosody & stress recognition**	X		
			Speech recognition	**Onomatopoeia**	X		
				Speech recognition	X	ruTiE-Audio AQUARIA RuSLUn	
				Song lyrics recognition	X		
		Media grounding		**Visual media grounding**		X	
				Temporal media grounding	X		CommonVideoQA RealVideoQA

Table 7: Skill taxonomy coverage of MERA Multi tasks (perception part). Columns **L1-L5** show the skill hierarchy levels, **Image**, **Audio** and **Video** columns indicate which tasks cover each perception type across different modalities.

The dichotomy between single-instance and cross-instance perception has been popularized in MM-Bench (Liu et al., 2024e) and subsequently reused by HR-Bench (Wang et al., 2024b). The third category is multimodal alignment which includes skills for matching multimodal representations, including text recognition.

An overview of the perception taxonomy is given in Table 7 and below a detailed skill description is given.

B.1.1 Fine-grained single-instance perception

Perceiving and (optionally) localizing a single salient object/event instance in an input.

⇨ ***Object localization*** I V Detecting object positions relative to the image itself (left, right, top, bottom, center) and predicts coarse spatial categories (e.g., left/right/center) without producing media coordinates. In contrast, media grounding predicts locations in the media coordinate system (bounding boxes, segmentation masks, or timestamp spans).

⇨ ***Object recognition*** I A V Mapping objective visual and audial features to the concepts in knowledge space, e.g. matching the concept 'cat' to meowing sounds and/or visual image of a cat (ears, paws, etc.).

Event recognition (single-instance) Recognizing actions, events, and states. It involves features that are dynamically perceived (e.g. rolling) or may suggest dynamic characteristics (e.g. color changes), such as those found in motion stop-frames.

⇨ ***Object motion recognition*** I A V

⇨ ***Living things motion recognition*** I A V Recognizing movement types specific to humans and animals.

Pose recognition The ability to infer articulated body or hand configuration, either as keypoints/joint structure or as gesture/pose classes when the task is categorical. It deserves a separate perceptual skill because many downstream tasks (e.g. action and intention understanding) depend primarily on how a person is posed rather than what objects are present. In practice, this is operationalized by standard pose-estimation datasets such as COCO Keypoints (Lin et al., 2014) and MPII Human Pose (Andriluka et al., 2014), video benchmarks like PoseTrack (Andriluka et al., 2018) for multi-person pose and tracking, and hand-gesture datasets such as HaGRID (Kapitanov et al., 2024), for detection- and classification-style gesture recognition.

⇨ ***Body pose recognition*** I V

⇨ ***Facial expression recognition*** I V

⇨ ***Hand gesture recognition*** I V

B.1.2 Fine-grained cross-instance perception

Perceiving and localizing multiple objects or events and the character of their interaction.

Overlapping object differentiation Separating co-occurring entities whose features interfere.

⇨ ***Overlapping image differentiation*** I V Disentangling and separate recognition of visually overlapping objects, .

⇨ ***Speaker diarization*** A V Attributing speech to the correct speaker among multiple simultaneous and/or consequent speakers.

Mutual object localization Perceiving relative relations among multiple entities.

⇨ ***Spatial object relationship*** I A V Understanding the relative localization of multiple objects in the scene.

⇨ ***Temporal object relationship*** A V Understanding the ordering and temporal relations across events/segments.

Repeating pattern recognition Perceiving instances that appear as a repeating pattern, e.g. animal flocks, ornaments, rhythmic sounds and moves, also involves perceiving frequency and density. Primary to counting skills.

⇨ ***Visual pattern recognition*** I V Patterns present in a single image or video frame, which do not require dynamic perception to acquire.

⇨ ***Temporal pattern recognition*** A V Patterns spanning over time and / or perceived only across multiple frames.

Cross-instance event recognition Perceiving interactions among multiple entities.

⇨ ***Object-object interaction*** I A V

⇨ ***Human-object interaction*** I A V

⇨ ***Human-human interaction*** I A V

B.1.3 Cross-modal alignment

Cross-modal alignment (Grounding) covers abilities that establish correspondences between symbolic descriptions (typically language) and specific elements in a non-text modality, enabling verification ("is the described thing present?") and reference ("where/when is it?"). In cognitive terms, grounding mediates between perceptual evidence and symbolic reasoning, supporting tasks that require referential linkage rather than mere recognition. We separate two complementary families: (i) Text extraction — recovering textual symbols from media (e.g., OCR in images/video frames; speech recognition and lyrics recognition in audio/video) — and (ii) Media grounding — mapping language to media coordinates (e.g., bounding boxes/masks in images/video, or temporal spans in audio/video). This separation keeps transcription distinct from localization while preserving a unified notion of cross-modal correspondence.

Image-to-text grounding Involves recognizing text in images, video frames, and more complex visual elements such as tables, charts, and diagrams. Some of the benchmarks targeting these skills together with QA tasks are PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022), Misleading ChartQA (Chen et al., 2025), ChartBench (Xu et al., 2023), and TextVQA (Singh et al., 2019).

⇨ ***Text recognition (OCR)*** I V
⇨ ***Scheme recognition*** I V
⇨ ***Plot recognition*** I V
⇨ ***Table recognition*** I V

Audio-to-text grounding Similarly to Image-to-text grounding, this skill involves matching sound signals to textual representation, including complex audio elements such as prosody, stress, and lyrics.

⇨ ***Prosody & stress recognition*** A V involves recognizing and understanding the rhythm, intonation, and stress patterns in speech.

Speech recognition

⇨ ***Onomatopoeia*** A V Onomatopoeia is a language phenomenon where words are used to imitate sounds, such as "buzz" for a bee or "meow" for a cat. Recognizing onomatopoeia is essential for understanding routine language usage. An example of a dataset for onomatopoeia recognition is RWCP-SSD-Onomatopoeia (Okamoto et al., 2020).

⇨ ***Speech recognition*** A V

⇨ ***Song lyrics recognition*** A V

Media grounding Grounding to media involves matching textual descriptions to media elements via masking, segmentation, or bounding boxes.

⇨ ***Visual media grounding*** I V Operating on bounding boxes, segmentation masks, or pixel coordinates for objects or actions described in natural language. For example, predicting the bounding box of the object in an image or the segmentation mask of the action in a video.

⇨ ***Temporal media grounding*** A V Operating on timestamps and media duration. For example, predicting the start and end time of the event in a video or the duration of the action in a video.

B.2 Knowledge

Knowledge denotes a model's ability to retrieve, apply, and integrate stored factual and conceptual information in multimodal settings. In cognitive science terms, this corresponds to long-term semantic memory (stable representations of entities, relations, norms, and domain concepts) used to interpret perceptual evidence and to support reasoning. In multimodal evaluation, knowledge competence is often probed via QA and exam-style benchmarks (e.g., MMLU/MMMU (**?**Yue et al., 2024)) that condition questions on images, and via cognitively motivated suites that target foundational concepts from developmental cognition (e.g., CoreCognition (Li et al., 2025), which operationalizes "core knowledge" constructs such as object permanence and basic physical/social regularities). Because "knowledge" interacts with both perception and reasoning, we treat it as a distinct group to separate failures of recall/concept access from failures of inference over correctly perceived evidence.

Below we present a detailed description of the knowledge category, which was introduced in Section 3.2.

Knowledge In MLLM benchmarks, "knowledge" is typically observed as (i) reliance on commonsense/everyday facts in visual contexts, (ii) explicit external/encyclopedic knowledge conditioned on images, and (iii) expert-domain question answering over multimodal inputs.

The differentiation between common and domain knowledge may vary depending on the task and the domain. We formulate the polar points

of this distinction as follows: common knowledge is generally agreed upon by the majority of the population and used in everyday life, while expert knowledge is domain-specific and requires specialized knowledge or training.

Common everyday knowledge

⇨ ***Common everyday knowledge*** I A V
Commonsense and broadly shared facts.

⇨ ***Ethics*** I A V Safety/ethics-aligned judgment about harmful/unsafe content or behavior in multimodal prompts. The placement of ethics in the knowledge category and treating ethical norms as stored conventions may be considered a simplification, but it is done to keep minimal the number of major categories.

Domain knowledge This category may be further subdivided into subject-specific categories as it is done in expert benchmarks, e.g. MMMU (Yue et al., 2024).

⇨ ***Common domain knowledge*** I A V
Subject-specific but broadly accessible knowledge (often school-level or general STEM).

⇨ ***Expert domain knowledge*** I A V Specialized, exam- or professional-level knowledge requiring training.

B.3 Reasoning

Reasoning comprises abilities that derive new conclusions from (i) perceptual evidence and (ii) prior knowledge, including uncertain, multi-step, relational, causal, quantitative, and counterfactual inference. In cognitive terms, reasoning corresponds to controlled inference processes that operate on representations supplied by perception and memory, producing answers that are not directly observable in the input. We include both classical forms (deductive/abductive/inductive inference) and task-driven "holistic" judgments (e.g., topic, style, provenance, media characteristics) when they require integrating multiple cues across an input rather than naming a single localized entity. This placement is consistent with benchmark taxonomies that separate fine-grained recognition from higher-level judgments—for example, MMBench distinguishes coarse perception categories such as image topic/quality from fine-grained object-level recognition (Liu et al., 2024e). Analogously, audio evaluation benchmarks emphasize media-level characterization and instruction-guided understanding (AudioBench (Wang et al., 2025a), AIR-Bench (Yang et al., 2024), MMAU (Sakshi et al., 2025)).

An overview of the reasoning taxonomy is given in Table 8 and below a detailed skill description is provided.

B.3.1 Inductive reasoning

Attribute recognition The distinction between coarse attribute recognition and object attribute recognition is inspired by MMBench's (Liu et al., 2024e) categorization into coarse vs. fine-grained perception. We put coarse perception into reasoning category as most of the related tasks require inference from the perceived information about the whole sample.

Coarse attribute recognition Global or "holistic" characterization of a sample (quality, topic, provenance, modality-level traits), without requiring fine object-level localization or multi-entity comparison.

⇨ ***Topic understanding*** I A V
⇨ ***Style & genre understanding*** I A V
⇨ ***Scene understanding*** I A V

⇨ ***Generated content detection*** I A V
Detection of whether content is synthetic vs real.

⇨ ***Media characteristic understanding***
I A V Recognize broad media-level traits (quality, augmentations). Characterizing/attributing the origin or type of synthesis (e.g., illustration type, camera quality, sound interference).

In vision-language evaluation, MMBench (Liu et al., 2024e) explicitly includes Image Quality and Image Topic as coarse perception categories. In audio(-text) evaluation, media-level characterization is a core target of AudioBench (Wang et al., 2025a) (speech understanding, audio scene understanding, and paralinguistic "voice understanding"), AIR-Bench (Yang et al., 2024) (speech/natural sounds/music comprehension under instruction following), and MMAU (Sakshi et al., 2025) (multi-task audio understanding and reasoning across speech/environment/music).

⇨ ***Speech emotion recognition*** A V
⇨ ***Music emotion recognition*** I A V
⇨ ***Melodic structure interpretation*** I A V

Taxonomy Level					Modality		
L1	**L2**	**L3**	**L4**	**L5**	**Image**	**Audio**	**Video**
Reasoning	Inductive reasoning	Attribute recognition	Coarse attribute recognition	**Generated content detection**			
				Source characterization			
				Media characteristic understanding			
				Speech emotion recognition		AQUARIA	
				Music emotion recognition		AQUARIA	
				Melodic structure interpretation			
				Topic understanding	ruTiE-Image	ruTiE-Audio	
				Style & genre understanding	RealVQA	AQUARIA	
				Scene understanding	ruCommonVQA RealVQA ruHHH-Image ruTiE-Image	ruTiE-Audio ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA ruHHH-Video
			Object attribute recognition	**Physical property understanding**	ruCommonVQA ruCLEVR RealVQA UniScienceVQA SchoolScienceVQA WEIRD	ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA
				Object function understanding	ruCommonVQA RealVQA ruHHH-Image WEIRD	ruEnvAQA	CommonVideoQA RealVideoQA ruHHH-Video
				Identity & emotion understanding	ruCommonVQA WEIRD	AQUARIA	
		Other inductive reasoning		**Other inductive reasoning**			
	Deductive reasoning			**Weirdness understanding**	WEIRD		
				Analogical reasoning	ruTiE-Image	ruTiE-Audio	
				Other deductive reasoning			
	Abductive reasoning			**Hypothetical reasoning**	RealVQA	AQUARIA	CommonVideoQA
				Cause & effect understanding	RealVQA	AQUARIA	CommonVideoQA RealVideoQA
	Quantitative reasoning		Counting	**Static counting**	LabTabVQA ruCommonVQA ruCLEVR ruNaturalScienceVQA RealVQA UniScienceVQA SchoolScienceVQA ruTiE-Image	X	CommonVideoQA RealVideoQA
				Temporal counting	X	ruTiE-Audio ruEnvAQA AQUARIA	CommonVideoQA RealVideoQA
			Mathematical reasoning	**Mathematical reasoning**	ruMathVQA ruNaturalScienceVQA RealVQA UniScienceVQA SchoolScienceVQA ruTiE-Image	ruTiE-Audio	CommonVideoQA RealVideoQA
	Other reasoning			**Critical thinking**			
				Counterfactual robustness	RealVQA		
				Problem decomposition	ruMathVQA ruNaturalScienceVQA RealVQA UniScienceVQA SchoolScienceVQA		
				Comparative reasoning	RealVQA UniScienceVQA SchoolScienceVQA	ruEnvAQA AQUARIA	

Table 8: Skill taxonomy coverage of MERA Multi tasks (reasoning part). Columns **L1-L5** show the skill hierarchy levels, while **Image**, **Audio** and **Video** columns indicate which tasks cover each reasoning type across different modalities.

Object attribute recognition Attribute-centric understanding grounded in entities (objects, people, instruments), including physical properties, functions, and identity-related cues.

⇨ ***Physical property understanding*** I A V Infer physical properties (e.g., material/shape/size/-texture; or physical dynamics in video).

⇨ ***Object function understanding*** I A V Infer what an object is for and how it is used (affordances, intent).

⇨ ***Identity & emotion understanding*** I A V Recognize identity-related cues (who) and affective state (facial expression, vocal affect).

B.3.2 Deductive reasoning

Rule- or structure-driven inference where conclusions follow from constraints (logical, relational, symbolic, or structural).

⇨ ***Weirdness understanding*** I A V Detect incongruity, humor, anomalies, or "oddness" that violates expectations.

⇨ ***Analogical reasoning*** I A V Solve problems by mapping relational structure between situations.

⇨ ***Other deductive reasoning*** I A V General logical/relational deduction (multi-hop, constraints).

B.3.3 Abductive reasoning

Inference to the best explanation: forming hypotheses and causal attributions consistent with observations.

⇨ ***Hypothetical reasoning*** I A V Reason about "what if" interventions and unseen outcomes.

⇨ ***Cause & effect understanding*** I A V Identify causal responsibility and predict effects of interactions over time.

B.3.4 Quantitative reasoning

Counting

⇨ ***Static counting*** I A V Count entities in a single frame/image. Complementary to pattern recognition skills that focus on repeated structures.

⇨ ***Temporal counting*** A V Count events and objects over time in audio/video streams. Complementary to temporal pattern recognition skills that focus on repeated structures.

Mathematical reasoning Solve math problems grounded in visual contexts (diagrams, plots, tables, scenes) or multimodal inputs. MathVista (Lu et al., 2024) is explicitly proposed to evaluate mathematical reasoning of foundation models in visual contexts.

⇨ ***Mathematical reasoning*** I A V

B.3.5 Other reasoning

⇨ ***Critical thinking*** I A V Evaluate evidence, reconcile inconsistencies, and make defensible judgments under ambiguous multimodal context.

⇨ ***Counterfactual robustness*** I A V Maintain correct reasoning when interventions are applied (e.g. answer option in MCQ do not contain the correct answer, or the question relates to an entity not present in the image).

⇨ ***Problem decomposition*** I A V Solve by breaking into intermediate steps; multi-step chains across modalities. ScienceQA (Lu et al., 2022) is explicitly proposed to evaluate problem decomposition skills of foundation models.

⇨ ***Comparative reasoning*** I A V Answer questions requiring comparisons (more/less, same/different) across instances or across time.

C Data leakage details

Goal. Given a multimodal example $x = (t, m)$ where m is the paired modality (image/video/audio), t is the text, estimate the probability that a target model was trained on x.

C.1 Setup.

We begin by creating a controlled environment to simulate data leakage. For a given modality, we take a base model, $Model$. We then create a leaked version, $Model_{leak}$, by fine-tuning the base model on a subset of our benchmark data using SFT-LoRA for the selected modality.

C.2 Neighbor Generation and Feature Extraction.

For each original data point (t, m) we generate $K = 24$ perturbed "neighbors". We apply four distinct perturbation techniques (masking and predicting the masks with Fred-T5 model[20], deletion,

[20] https://hf.co/ai-forever/FRED-T5-1.7B, (Zmitrovich et al., 2024)

duplication, and swapping of random words) to the text t with each technique applied 6 times. The modality data m remains unchanged.

For each original text t and its neighbors t'_k we extract their text embeddings using a fixed encoder:

$$e = E(t), \quad e'_k = E(t'_k)$$

where E is `intfloat/e5-mistral-7b-instruct`[21] model.

Subsequently, we compute the multimodal loss for both models $Model$ and $Model_{leak}$ on both the original and neighbor data points:

$$\mathcal{L} = L(Model, t, m), \quad \mathcal{L}'_k = L(Model, t'_k, m)$$

C.3 Detector Training

The core of MSMIA is a binary classifier trained to distinguish between models that have and have not seen the data. For each neighbor k we create two training examples by computing the feature differences:

$$\Delta\mathcal{L} = \mathcal{L} - \mathcal{L}'_k, \quad \Delta e = e - e'_k$$

These feature vectors are paired with labels $y \in \{0, 1\}$ indicating whether the losses came from $Model$ ($y = 0$) or $Model_{leak}$ ($y = 1$). This process yields 48 training triplets $(\Delta\mathcal{L}, \Delta e, y)$ per original data point. The MSMIA detector, f_{MSMIA} is trained to predict the probability $p = f_{MSMIA}(\Delta\mathcal{L}, \Delta e)$ that the input features originate from a model that has been trained on the target data.

C.4 Inference and Evaluation

To infer if a target $TestModel$ has been trained on a specific data point (t, m), we repeat Step 2 to compute the loss and embedding differences for this model. We then compute the leakage score S for the data point by taking the average probability output by the detector over all K neighbors:

$$S(t, m) = \frac{1}{K} \sum_{k=1}^{K} f_{MSMIA}(\Delta\mathcal{L}_k, \Delta e_k)$$

To get the probability estimation for the entire dataset, the S scores are averaged over all dataset samples.

We report AUC-ROC for binary classification (leaked vs. clean) as shown in Tables 9, 11, 10.

[21] `https://hf.co/intfloat/e5-mistral-7b-instruct`. It used to be SoTA on MTEB benchmark (Muennighoff et al., 2022)

Origin Model	Test Model	AUC-ROC
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	96.2
Qwen2.5-VL-3B-Instruct	Qwen2-VL-7B-Instruct	86.0
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	88.0
Qwen2.5-VL-3B-Instruct	llama3-llava-next-8b-hf	90.2
Qwen2.5-VL-3B-Instruct	gemma-3-4b-it	65.8
Qwen2.5-VL-3B-Instruct	gemma-3-12b-it	67.9
Qwen2-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	78.0
Qwen2-VL-7B-Instruct	Qwen2-VL-7B-Instruct	96.2
Qwen2-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	80.5
Qwen2-VL-7B-Instruct	llama3-llava-next-8b-hf	78.0
Qwen2-VL-7B-Instruct	gemma-3-4b-it	77.7
Qwen2-VL-7B-Instruct	gemma-3-12b-it	73.7
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	92.8
Qwen2.5-VL-7B-Instruct	Qwen2-VL-7B-Instruct	93.1
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	98.1
Qwen2.5-VL-7B-Instruct	llama3-llava-next-8b-hf	95.8
Qwen2.5-VL-7B-Instruct	gemma-3-4b-it	95.4
Qwen2.5-VL-7B-Instruct	gemma-3-12b-it	94.5
llama3-llava-next-8b-hf	Qwen2.5-VL-3B-Instruct	94.6
llama3-llava-next-8b-hf	Qwen2-VL-7B-Instruct	90.0
llama3-llava-next-8b-hf	Qwen2.5-VL-7B-Instruct	96.6
llama3-llava-next-8b-hf	llama3-llava-next-8b-hf	97.7
llama3-llava-next-8b-hf	gemma-3-4b-it	99.1
llama3-llava-next-8b-hf	gemma-3-12b-it	99.5
gemma-3-4b-it	Qwen2.5-VL-3B-Instruct	76.0
gemma-3-4b-it	Qwen2-VL-7B-Instruct	71.5
gemma-3-4b-it	Qwen2.5-VL-7B-Instruct	85.2
gemma-3-4b-it	llama3-llava-next-8b-hf	86.5
gemma-3-4b-it	gemma-3-4b-it	99.4
gemma-3-4b-it	gemma-3-12b-it	98.7
gemma-3-12b-it	Qwen2.5-VL-3B-Instruct	84.1
gemma-3-12b-it	Qwen2-VL-7B-Instruct	81.3
gemma-3-12b-it	Qwen2.5-VL-7B-Instruct	91.2
gemma-3-12b-it	llama3-llava-next-8b-hf	93.3
gemma-3-12b-it	gemma-3-4b-it	99.4
gemma-3-12b-it	gemma-3-12b-it	99.7

Table 9: AUC-ROC MSMIA performance metrics for various evaluated Image MLLMs.

There the `Origin Model` is the model used to train MSMIA. `Test Model` is the model whose losses are used to test MSMIA (predict whether the data sample was used to train `Test Model` or not).

D LLM-as-a-judge details

D.1 Data collection

Our dataset comprises (question, gold answer, model prediction) triplets sourced from Russian-language benchmarks, including MERA Multi, with model sizes spanning 2B to 110B parameters. Human annotators label the semantic correctness of each prediction, strictly ignoring surface form. To ensure label quality, only items with 100% inter-annotator agreement are used for model training and testing.

We apply synthetic augmentations (e.g., refusals, repetitions) to improve robustness. A critical measure to prevent bias was splitting the data by source dataset, ensuring no dataset appears in both train and test splits. The final dataset composition and class balance are detailed in Table 12.

Origin Model	Test Model	AUC-ROC
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-3B-Instruct	95.9
Qwen2.5-VL-3B-Instruct	Qwen2.5-VL-7B-Instruct	99.5
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video	91.7
Qwen2.5-VL-3B-Instruct	LLaVA-NeXT-Video-DPO	91.2
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-3B-Instruct	98.7
Qwen2.5-VL-7B-Instruct	Qwen2.5-VL-7B-Instruct	100.0
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video	96.5
Qwen2.5-VL-7B-Instruct	LLaVA-NeXT-Video-DPO	95.7
LLaVA-NeXT-Video	Qwen2.5-VL-3B-Instruct	63.7
LLaVA-NeXT-Video	Qwen2.5-VL-7B-Instruct	71.5
LLaVA-NeXT-Video	LLaVA-NeXT-Video	100.0
LLaVA-NeXT-Video	LLaVA-NeXT-Video-DPO	100.0
LLaVA-NeXT-Video-DPO	Qwen2.5-VL-3B-Instruct	53.6
LLaVA-NeXT-Video-DPO	Qwen2.5-VL-7B-Instruct	56.2
LLaVA-NeXT-Video-DPO	LLaVA-NeXT-Video	100.0
LLaVA-NeXT-Video-DPO	LLaVA-NeXT-Video-DPO	100.0

Table 10: AUC-ROC MSMIA performance metrics for various evaluated Video MLLMs.

Origin Model	Test Model	AUC-ROC
Qwen2-Audio-7B-Instruct	Qwen2-Audio-7B-Instruct	87.7
Qwen2-Audio-7B-Instruct	Qwen-Audio-Chat	76.0
Qwen-Audio-Chat	Qwen2-Audio-7B-Instruct	61.3
Qwen-Audio-Chat	Qwen-Audio-Chat	100.0

Table 11: AUC-ROC MSMIA performance metrics for various evaluated Audio MLLMs.

	Total	Open Generation	Multiple Choice
Train set	62,580	22,538	40,042
Test set	8,570	5,819	2,751

Table 12: Overview of the datasets used for training and evaluating the judge model. The columns indicate the total number of examples and their distribution across two task formats: *Open Generation* (free-form model response) and *Multiple Choice* (discrete answer selection)

D.2 Training models

The tasks are formulated as a binary classification problems. Inputs are linearized into the format `question [SEP] gold answer [SEP] model prediction` and packed into the maximum models' contexts. We train encoder-based models, adding a linear classification head on top of the pre-trained backbone. The models are fully fine-tuned and optimized using a cross-entropy loss, with class weights applied to mitigate potential dataset imbalance.

Optimization is performed with mixed precision and early stopping based on the F1 score on the development set, constrained to a single A100 80GB GPU. A summary of the final training configuration is provided in Table 13.

D.3 Model selection

We compare fine-tuned encoder-based models against zero-shot decoder baselines to identify the optimal judge architecture. The decoder models are prompted to output either "0" or "1" without any task-specific fine-tuning. While this approach is flexible, its reliability is limited by the fact that decoder generations are unconstrained; they may not always produce a valid classification, making score interpretation ambiguous.

Parameter	Value
Learning Rate	2e-5
Batch Size	32
Number of Epochs	3
Weight Decay	0.01
Optimizer	AdamW
LR Scheduler	Linear
Max Sequence Length	512
Precision	BF16
Early Stopping Metric	F1 Score

Table 13: Training configuration summary for embedding-based classification

In our experiments, encoder models proved to be more suitable for this binary classification task. Their architectural design naturally provides probability distributions over the two classes (correct/incorrect), ensuring deterministic scoring. Furthermore, they offer significant practical advantages, being generally smaller, faster on a single GPU, and even capable of handling long contexts (useful for judging the answers with reasoning or chain-of-thought elements). A full model comparison with bootstrap statistics is provided in Table 14.

D.4 Model deployment

The selected judge model is deployed using vLLM for high-throughput inference on a single A100 40GB GPU. This provides optimal batch processing speed for rapid benchmark evaluation while maintaining quality. Our codebase implements a `score(q, ref, pred) → {0,1}` API that handles individual scoring and aggregation. We also publicly provide the trained model with weights on HuggingFace Hub [22].

D.5 Model analysis

To validate our LLM-as-a-judge, we performed a sanity check demonstrating high heuristic consistency: the model aligns with human labels in 99.6% of cases where Exact Match (EM) is 1, and 97.7% when EM is 0 but humans deem the response correct. This reliability extends across task types, with

[22] https://huggingface.co/MERA-evaluation/MERA_Answer_judge

	Model	Parameters	Context Length	Samples/sec	F1	Recall	Precision	Pearson	EM rate
Encoders	RuModernBERT-base	150M	8,192	1800	0.964 ± 0.002	0.975	0.955	0.940	0.997
	RuModernBERT-small	35M	8,192	2000	0.944 ± 0.003	0.962	0.927	0.879	0.982
	Qwen3-Embedding-0.6B	600M	32,768	1850	0.951 ± 0.003	0.985	0.920	0.936	1.000
	FRIDA	823M	512	70*	0.838 ± 0.005	0.913	0.774	0.765	0.925
	embeddinggemma-300m	303M	2,048	250*	0.915 ± 0.004	0.965	0.871	0.825	0.997
	Giga-Embeddings-instruct	3.45B	4,096	40*	0.966 ± 0.002	0.984	0.948	0.946	1.000
Decoders	pollux-judge-7b	7.61B	131072	20	0.823 ± 0.005	0.788	0.860	0.732	0.873
	T-lite-it-1.0	7.61B	131072	52	0.844 ± 0.005	0.968	0.747	0.755	0.993
	Qwen3-0.6B	752M	32768	442	0.030 ± 0.003	0.016	0.323	-0.010	0.004
	Qwen3-1.7B	2.03B	32768	205	0.590 ± 0.006	0.990	0.420	0.300	0.997
	gpt-oss-20b	21.5B	131072	28	0.939 ± 0.003	0.968	0.912	0.905	1.000

Table 14: Summary results for judge models on the test set. F1, Recall and Precision report binary classification quality; Pearson is the Pearson correlation between binary predictions and ground-truth labels; EM rate is the share of class-1 predictions on the subset of examples with an exact string match between the prediction and the gold answer. An asterisk (*) next to throughput indicates the model could not be served via vLLM and the speed was measured with the HuggingFace Trainer instead

agreement rates of 98.7% for multiple-choice and 96% for free-form generation. Statistical analysis confirms the judge is resilient to common biases, showing near-zero Pearson and Spearman correlations with answer length ($r \approx 0.001$, $\rho \approx -0.008$) and gold label position ($r/\rho \approx -0.0005$). Qualitative analysis reveals that errors primarily occur in tasks involving complex LaTeX formatting or multi-step reasoning. Since our model is designed as a binary classifier without a rationale-generation component, it occasionally struggles with the ambiguous justifications or symbolic intricacies inherent in these specialized domains.

E Block prompts analysis

E.1 Prompts formulations

There are 13 blocks:

- Attention hook (greeting / draw attention).
- General task description (task specific).
- Input data description (enumerate modalities).
- Action on data (e.g., "Solve the task using the images…").
- Optional task specifics (helpful but nonessential context).
- Textual question.
- Answer options (if multiple choice).
- Call to solve (explicit request to answer).
- Reasoning request (ask for thinking before final answer; present in 5 prompts).
- Reasoning format (how to present the reasoning).
- Answer format (how to present the final answer).
- Time limitation (e.g., "you have 10 minutes").
- Final call to action (e.g., "get started").

The blocks are combined with the Python script, following the task prompts configuration file. This file contains an explicit description of all ten prompts. Example of the prompt configuration:

```
# prompt_config.yaml
prompt_4:
 attention_hook: "informal_request"
 task_description: "informal_request"
 input_data: "default"
 processing_data: "informal_request"
 context_intro: "in_dataset"
 task_context: "default"
 question: "default"
 answer_options: "default"
 solution_motivation: "informal_request"
 reasoning_motivation: "none"
 reasoning_format: "none"
 answer_format: "informal_request"
 limitations: "informal_request"
 answer_motivation: "informal_request"
```

The block formulations are fixed. The only variable blocks are `task_description` and `task_context` that are task-specific. All prompt formulations are imputed in the dataset on HuggingFace Hub (key "instruction" of each data sample). The blocks distribution is fixed: 5 prompts with `reasoning_motivation` and `reasoning_format` blocks to enable answer rationale, 1 prompt with no `answer_format` ("zero prompt" that provides the minimal version of the task - only placeholders for multimodal data, the question and answer options if any).

The example of the prompt with all blocks are as follows.

"Zero prompt":

```
Image: <image>
Question:
{question}
```

Prompt with specific answer format (RuSLUn):

```
The dataset for the task
    includes the following
    prompt:

Not all slots are necessarily
    present in the request.

Audio file: <audio>
Question:
{question}

Please solve the task based on
    the above and briefly
    formulate your answer.

{annotation}
```

Example of the prompt with "reasoning" blocks:

```
The dataset for the task
    includes the following
    prompt:

The question is directly related
     to the content of the image
     and requires not only
    recognition of individual
    elements, but also
    understanding of the
    relationships between the
    elements (objects) in the
    image. If there is
    insufficient information to
    answer the question, for
    example, if the object in
    question is missing from the
     image, then you must
    honestly answer that the
    question cannot be answered
    and indicate the reason.

Image: <image>
Question:
{question}

Please solve the task based on
    the above and briefly
    formulate your answer.

Please think about the solution
    and describe your thought
    process in detail.

Write your reasoning after the
    word REASONING, briefly
    explaining how you arrived
    at your final answer.

Please provide a brief answer to
     the question. Please do not
     write anything else, do not
     elaborate, do not engage in
     dialogue, and do not
    explain your answer. Please
    write your final answer
    after the word ANSWER.
```

E.2 Statistical analysis

To analyze the effect that a prompt formulation has on the metrics, we fit OLS with the following specification:

$$metric \sim C(prompt) + C(model) + C(engine) \tag{5}$$

In fact, it is essentially the same as:

$$metric_i = \alpha + \sum_{p=2}^{10} \beta_p \{prompt = p\} + \sum_m \gamma_m \{model = m\} + \sum_e \delta_e \{engine = e\} \tag{6}$$

Where:

- `Prompt` is a specific prompt formulation (one of 10).
- `Model` is the model name - the model used to infer the data with the prompt and the metric `metric`.
- `Engine` is the inference backend used for evaluation (one of `transformers` (Wolf et al., 2020) or `vllm`[23]).
- `Frames`. Additional categorical variable used only for video modality evaluations - the video is uniformly split into N frames and the vision LLM used to make evaluation on the dataset.
- `Domain`. Additional categorical variable used only for UniScienceVQA, SchoolScienceVQA, ruCommonVQA. These datasets are split into separate domains (subsets). This split may affect the metrics (one domain may be "harder" than another).

For each prompt p, we test $H_0 : \beta_p = 0$ against $H_1 : \beta_p \neq 0$. We mark a prompt's effect as statistically significant when the two-sided p-value < 0.05.

Figure 2 demonstrates the results of statistical analysis of the prompts formultations effects on the `Judge Score` metric. There are three datasets that have been omitted: RuSLUn implies structured output with rather strict metric that tends to show zero score, ruTiE-Audio and ruTiE-Image datasets scores are too sensible to prompt formulation due to the datasets design[24]

Takeaways:

[23] https://github.com/vllm-project/vllm

[24] Both datasets consist of three sequential dialogues of 500 questions in a row. The answer for the question X lays in the text of the questions $> X$. This way we cannot reliably separate this effect from the pure effect of the prompt formulation which may lead to incorrect analysis and conclusions.

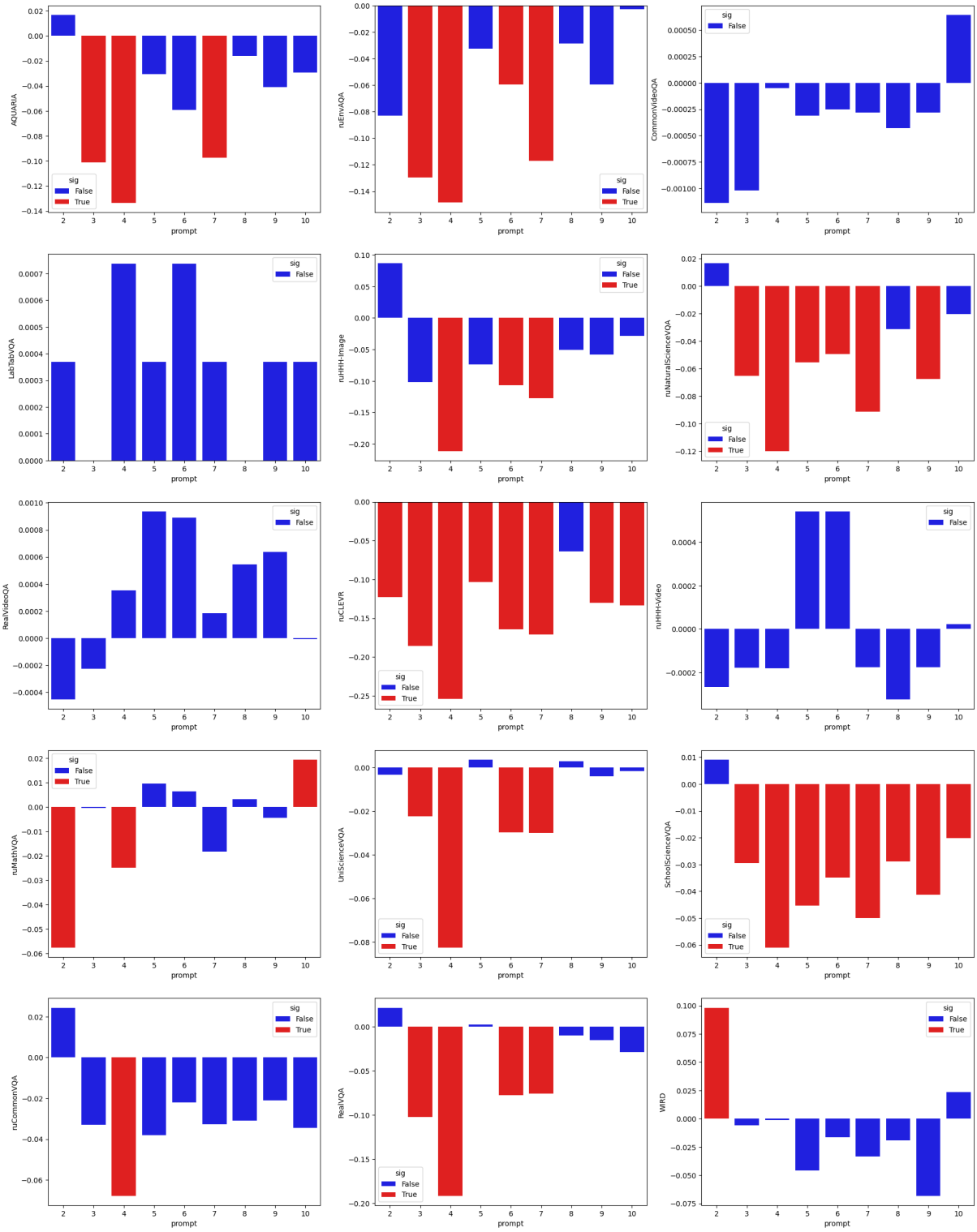


Figure 2: The relative (with regard to baseline prompt (0)) effects of different formulations of prompts for each dataset. There are ten different formulations of prompts for one dataset, hence nine corresponding bars (one formulation is baseline category). Red bars reflect statistically significant (at 95% confidence level) effects.

- **No single prompt dominates**. Different datasets favor different formulations; the prompt that helps in one case can hurt in another. This undermines any one-prompt-fits-all strategy.
- **Magnitude varies widely across datasets**. Some tasks show large shifts (on the order of 0.1–0.2 in the metric), while others exhibit near-zero effects (often still significant when variance is low). Prompt sensitivity is therefore task-dependent.
- **Reasoning format does not mean lower scores**. Prompts that explicitly encourage model to provide chain-of-though rationale for the final answer do not always get lower scores. First, this means that the LLM-as-judge model is capable of coping with reasoning format. Second, some models breach the answer format prescribed in the task instruction.
- **Single prompt leads to bias**. Only four tasks (notable, three of them are from video modality) demonstrate no red bars - no statistically significant prompt formulations compared to the baseline one. The other 12 datasets tend to be prompt sensitive, so the design choice to distribute prompts uniformly and interpret dataset metrics as an average over formulations is empirically supported.

F Baselines Details

F.1 Model Baselines Details

In this section, the list of baseline models is provided. Tables 15, 16, 17 represent models for image, audio, and video modalities respectively.

The exact results of the finished submissions of the models from Table 15 are presented in Tables 18, 19. The results of the models from Table 16 are presented in Table 20. The evaluation results of the models from Table 17 are stated in Table 21.

Evaluated on vision (image) modality models are strongest on natural-image semantics — object/scene understanding, object functions, and everyday knowledge (RealVQA, ruCommonVQA, WEIRD, ruTiE-Image) and show decent spatial relations and multi-object reasoning (RealVQA, ruCLEVR). Performance drops on diagrammatic/-scientific QA with OCR (UniScienceVQA, ruNaturalScienceVQA) and tables (LabTabVQA), exposing gaps in text extraction, scheme recognition, and cell-level grounding. Math (ruMathVQA) is brittle: EM lags JS, indicating partially correct solutions that miss the required final answer. Harder compositional and counterfactual reasoning (RealVQA, ruCommonVQA) still separates model tiers, especially smaller checkpoints. Overall, weaknesses concentrate in OCR/diagram/table parsing and structured compositional reasoning, while strengths lie in natural-image commonsense and basic spatial reasoning.

Moving to audio modality evaluations, the relative strengths cluster around acoustic scene understanding and temporal/comparative reasoning over environmental audio (ruEnvAQA), with partial competence in social/interaction cues and topic/scene grounding (AQUARIA, ruTiE-Audio) captured by higher JS. Weaknesses are pronounced in speech recognition (RuSLUn, ruTiE-Audio— EM), speaker/turn handling (diarization), and format-faithful final answer extraction. Improving ASR robustness, speaker attribution, and constrained decoding or answer templates should convert many high-JS outputs into EM gains.

As for the video modality evaluations, the current relative strengths lie in scene/object recognition and short event recognition. Clear weaknesses persist in temporal perception/localization, action-sequence reasoning, mutual object localization, counting, and cause-and-effect—skills central to CommonVideoQA/RealVideoQA. Low ruHHH-Video scores further reveal brittleness on ethics-aware interpretation and social context. Practically, improving temporal grounding (longer context windows or frame selection), causal reasoning, and answer-format control (constrained decoding/final-answer extraction) should convert many near-misses (high JS, low EM) into measurable EM gains.

F.2 Human Baseline Details

Human baseline values were obtained by evaluating the aggregate responses of annotators on control tasks. Prior to metric calculation, we conducted an additional identification of annotators who performed labeling tasks with low quality. To identify such annotators during the labeling process, we employed control tasks with automated correctness checking; incorrect answers reduced the annotator's skill score. During post-processing, we filtered out annotators who consistently made errors and whose responses did not align with the majority vote across all their submissions.

To evaluate the average human level we collected

Model	Parameters	Context length	Hugging Face Hub link	Citation
GPT 4.1	N/A	1000K	GPT 4.1	
Phi-3.5-vision-instruct	4B	128K	microsoft/Phi-3.5-vision-instruct	Abdin et al. (2024)
Phi-4-multimodal-instruct	6B	128K	microsoft/Phi-4-multimodal-instruct	Microsoft et al. (2025)
Qwen2.5-Omni-3B	6B	32K	Qwen/Qwen2.5-Omni-3B	Xu et al. (2025)
Qwen2.5-Omni-7B	11B	32K	Qwen/Qwen2.5-Omni-7B	
Qwen2-VL-2B-Instruct	2B	32K	Qwen/Qwen2-VL-2B-Instruct	Wang et al. (2024a)
Qwen2-VL-7B-Instruct	7B	32K	Qwen/Qwen2-VL-7B-Instruct	
Qwen2-VL-72B-Instruct	72B	32K	Qwen/Qwen2-VL-72B-Instruct	
Qwen3-VL-2B-Instruct	2B	262K	Qwen/Qwen3-VL-2B-Instruct	Yang et al. (2025)
Qwen3-VL-8B-Instruct	9B	262K	Qwen/Qwen3-VL-8B-Instruct	
Qwen2.5-VL-3B-Instruct	3B	128K	Qwen/Qwen2.5-VL-3B-Instruct	Bai et al. (2025)
Qwen2.5-VL-7B-Instruct	7B	128K	Qwen/Qwen2.5-VL-7B-Instruct	
Qwen2.5-VL-32B-Instruct	32B	128K	Qwen/Qwen2.5-VL-32B-Instruct	
Qwen2.5-VL-72B-Instruct	72B	128K	Qwen/Qwen2.5-VL-72B-Instruct	
gemma-3-12b-it	12B	128K	google/gemma-3-12b-it	Team et al. (2025a)
llava-1.5-13b-hf	13B	4K	llava-hf/llava-1.5-13b-hf	Liu et al. (2024b)
llava-next-72b-hf	72B	8K	llava-hf/llava-next-72b-hf	Li et al. (2024)
llava-next-110b-hf	110B	32K	llava-hf/llava-next-110b-hf	
InternVL3-9B	9B	32K	OpenGVLab/InternVL3-9B	Zhu et al. (2025)
granite-vision-3.3-2b	3B	128K	ibm-granite/granite-vision-3.3-2b	Team et al. (2025b)
SmolVLM-Instruct	2B	16K	HuggingFaceTB/SmolVLM-Instruct	Marafioti et al. (2025)
MiniCPM-o-2_6	9B	32K	openbmb/MiniCPM-o-2_6	Yao et al. (2024)

Table 15: General information about vision (image) modality baseline models.

crowd-source[25].

For expert tasks, in addition to the basic human baseline, we also performed expert annotation. We aggregate experts' answers with an overlap of 3 for expert human baseline scores [26]. The experts annotating answers for the human baseline had not been involved in the original dataset creation.

The ABC Elementary annotation platform ensures the necessary data anonymity during processing. The hourly compensation offered is above the minimum wage per hour in Russia (see Table 22). Annotators are made aware of potentially sensitive topics within the data, including politics, societal minorities, and religion. The data collection process undergoes a mandatory quality evaluation, featuring an automated annotation quality check through honeypot tasks.

[25]Annotation provided by ABC Elementary platform https://app.elementary.center

[26]The observed quantity is a consequence of the limited pool of domain experts in niche specializations, who also had no prior involvement in the annotation of the datasets.

Model	Parameters	Context length	Hugging Face Hub link	Citation
Qwen3-Omni-30B-A3B-Instruct	35B	8K	Qwen/Qwen3-Omni-30B-A3B-Instruct	
Qwen2.5-Omni-3B	6B	32K	Qwen/Qwen2.5-Omni-3B	Xu et al. (2025)
Qwen2.5-Omni-7B	11B	32K	Qwen/Qwen2.5-Omni-7B	
Qwen2-Audio-7B-Instruct	7B	32K	Qwen/Qwen2-Audio-7B-Instruct	Chu et al. (2024)
Qwen2-Audio-7B	8B	32K	Qwen/Qwen2-Audio-7B	
audio-flamingo-3-hf	9B	32K	nvidia/audio-flamingo-3-hf	Goel et al. (2025)
SeaLLMs-Audio-7B	8B	8K	SeaLLMs/SeaLLMs-Audio-7B	Zhang et al. (2024)
ultravox-v0_2	8B	8K	fixie-ai/ultravox-v0_2	Team (2024)
ultravox-v0_3	8B	8K	fixie-ai/ultravox-v0_3	
ultravox-v0_3-llama-3_2-1b	2B	128K	fixie-ai/ultravox-v0_3-llama-3_2-1b	
ultravox-v0_4	8B	8K	fixie-ai/ultravox-v0_4	
ultravox-v0_4_1-llama-3_1-8b	8B	128K	fixie-ai/ultravox-v0_4_1-llama-3_1-8b	
ultravox-v0_4_1-mistral-nemo	13B	128K	fixie-ai/ultravox-v0_4_1-mistral-nemo	
ultravox-v0_5-llama-3_2-1b	2B	128K	fixie-ai/ultravox-v0_5-llama-3_2-1b	
ultravox-v0_5-llama-3_1-8b	8B	128K	fixie-ai/ultravox-v0_5-llama-3_1-8b	
ultravox-v0_6-llama-3_1-8b	8B	128K	fixie-ai/ultravox-v0_6-llama-3_1-8b	
ultravox-v0_6-qwen-3-32b	32B	40K	fixie-ai/ultravox-v0_6-qwen-3-32b	

Table 16: General information about audio modality baseline models.

Model	Parameters	Context length	Hugging Face Hub link	Citation
LLaVA-NeXT-Video-7B-hf	7B	4K	llava-hf/LLaVA-NeXT-Video-7B-hf	Liu et al. (2024c)
Qwen2-VL-2B-Instruct	2B	32K	Qwen/Qwen2-VL-2B-Instruct	Wang et al. (2024a)
Qwen2-VL-7B-Instruct	7B	32K	Qwen/Qwen2-VL-7B-Instruct	
Qwen2-VL-72B-Instruct	72B	32K	Qwen/Qwen2-VL-72B-Instruct	
Qwen2.5-VL-3B-Instruct	3B	128K	Qwen/Qwen2.5-VL-3B-Instruct	Bai et al. (2025)
Qwen2.5-VL-7B-Instruct	7B	128K	Qwen/Qwen2.5-VL-7B-Instruct	
Qwen2.5-VL-72B-Instruct	72B	128K	Qwen/Qwen2.5-VL-72B-Instruct	
Qwen3-VL-2B-Instruct	2B	262K	Qwen/Qwen3-VL-2B-Instruct	Yang et al. (2025)
Qwen3-VL-8B-Instruct	9B	262K	Qwen/Qwen3-VL-8B-Instruct	
Qwen2.5-Omni-3B	6B	32K	Qwen/Qwen2.5-Omni-3B	Xu et al. (2025)
Qwen2.5-Omni-7B	11B	32K	Qwen/Qwen2.5-Omni-7B	
MiniCPM-o-2_6	9B	32K	openbmb/MiniCPM-o-2_6	Yao et al. (2024)
InternVL3_5-4B	5B	40K	OpenGVLab/InternVL3_5-4B	Wang et al. (2025b)
InternVL3_5-2B-Instruct	2B	40K	OpenGVLab/InternVL3_5-2B-Instruct	
InternVL3-9B-Instruct	9B	32K	OpenGVLab/InternVL3-9B	Zhu et al. (2025)
InternVL3-9B	9B	32K	OpenGVLab/InternVL3-9B	Zhu et al. (2025)
InternVL3-2B	1B	40K	OpenGVLab/InternVL3-2B	

Table 17: General information about video modality baseline models.

Model	Total	LabTabVQA	RealVQA	ruCLEVR	ruCommonVQA	ruHHH-Image*	ruMathVQA
Human Baseline	0.80	0.91	0.63	0.96	0.84	0.89	0.95
Qwen3-Omni-30B-A3B-Inst	**0.55**	**0.56** / **0.62**	**0.31** / 0.67	**0.50** / **0.63**	**0.56** / **0.86**	**0.45** / **0.57**	0.01 / **0.30**
GPT 4	0.48	0.29 / 0.29	0.23 / **0.70**	0.31 / 0.53	0.27 / 0.85	0.39 / 0.49	0.01 / 0.17
Qwen2.5-VL-72B-Inst	0.41	0.53 / 0.60	0.21 / 0.65	0.40 / 0.58	0.43 / 0.80	0.35 / 0.44	0.02 / 0.23
Qwen2-VL-72B-Inst	0.33	0.12 / 0.14	0.10 / 0.49	0.29 / 0.55	0.14 / 0.60	0.14 / 0.19	0.01 / 0.06
Qwen2.5-VL-32B-Inst	0.31	0.47 / 0.57	0.16 / 0.50	0.33 / 0.58	0.30 / 0.76	0.26 / 0.40	0.00 / 0.12
Qwen2.5-VL-7B-Inst	0.26	0.19 / 0.27	0.11 / 0.32	0.23 / 0.34	0.33 / 0.56	0.20 / 0.29	0.02 / 0.10
llava-next-110b-hf	0.24	0.06 / 0.10	0.09 / 0.29	0.11 / 0.20	0.31 / 0.58	0.12 / 0.18	0.01 / 0.01
Phi-3.5-vision-inst	0.23	0.12 / 0.18	0.03 / 0.10	0.08 / 0.16	0.21 / 0.40	0.19 / 0.26	0.01 / 0.02
llava-next-72b-hf	0.23	0.09 / 0.11	0.07 / 0.29	0.09 / 0.19	0.21 / 0.49	0.11 / 0.15	0.00 / 0.00
Qwen2.5-Omni-7B	0.23	0.11 / 0.18	0.08 / 0.35	0.13 / 0.32	0.23 / 0.53	0.14 / 0.23	0.03 / 0.08
Qwen2-VL-7B-Inst	0.20	0.01 / 0.13	0.05 / 0.36	0.10 / 0.34	0.19 / 0.50	0.02 / 0.16	0.01 / 0.02
SmolVLM-Inst	0.19	0.01 / 0.16	0.03 / 0.23	0.11 / 0.34	0.14 / 0.52	0.14 / 0.22	0.02 / 0.04
Qwen3-VL-8B-Inst	0.19	0.00 / 0.07	0.11 / 0.45	0.14 / 0.33	0.23 / 0.59	0.05 / 0.15	0.04 / 0.10
Phi-4-multimodal-inst	0.18	0.26 / 0.35	0.11 / 0.30	0.09 / 0.20	0.07 / 0.36	0.04 / 0.08	0.04 / 0.07
MiniCPM-o-2_6	0.18	0.04 / 0.06	0.08 / 0.31	0.15 / 0.33	0.26 / 0.50	0.06 / 0.09	0.00 / 0.00
Qwen2.5-Omni-3B	0.18	0.04 / 0.14	0.07 / 0.25	0.16 / 0.29	0.22 / 0.48	0.04 / 0.21	0.01 / 0.04
InternVL3-9B	0.17	0.01 / 0.04	0.04 / 0.27	0.19 / 0.37	0.09 / 0.45	0.00 / 0.07	0.01 / 0.02
Qwen2-VL-2B-Inst	0.17	0.00 / 0.09	0.06 / 0.28	0.11 / 0.32	0.12 / 0.44	0.03 / 0.09	0.04 / 0.05
gemma-3-27b-it	0.15	0.00 / 0.01	0.05 / 0.38	0.04 / 0.17	0.07 / 0.44	0.00 / 0.02	0.01 / 0.04
granite-vision-3.3-2b	0.14	0.00 / 0.08	0.01 / 0.10	0.00 / 0.08	0.11 / 0.20	0.08 / 0.20	0.02 / 0.02
Qwen2.5-VL-3B-Inst	0.14	0.02 / 0.11	0.03 / 0.13	0.13 / 0.26	0.15 / 0.32	0.03 / 0.18	0.00 / 0.04
Qwen3-VL-2B-Inst	0.12	0.00 / 0.04	0.06 / 0.35	0.11 / 0.27	0.14 / 0.47	0.05 / 0.08	0.01 / 0.04
llava-1.5-13b-hf	0.12	0.01 / 0.16	0.01 / 0.10	0.01 / 0.08	0.10 / 0.33	0.02 / 0.26	0.01 / 0.01

Table 18: **Image** modality evaluation results (6 tasks out of 11). All tasks metrics are `Exact Match / Judge Score`. For *ruHHH-Image* dataset the metrics are `Group Exact Match / Group Judge Score`. For Human Baseline aggregated results are provided (average of EM and JudgeScore).

Model	Total	ruNaturalScienceVQA	SchoolScienceVQA	UniScienceVQA	WEIRD	ruTiE-Image
Human Baseline	0.80	0.99	0.82	0.13	0.85	0.77
Qwen3-Omni-30B-A3B-Inst	**0.55**	**0.77** / **0.85**	**0.64** / **0.70**	0.11 / 0.34	**0.70** / 0.78	0.63 / 0.64
GPT 4	0.48	0.64 / 0.69	0.59 / 0.66	0.10 / **0.40**	0.69 / **0.79**	**0.70** / **0.73**
Qwen2.5-VL-72B-Inst	0.41	0.01 / 0.05	0.24 / 0.32	0.11 / 0.25	0.65 / 0.76	0.63 / 0.69
Qwen2-VL-72B-Inst	0.33	0.25 / 0.40	0.54 / 0.68	**0.18** / 0.33	0.31 / 0.48	0.65 / 0.69
Qwen2.5-VL-32B-Inst	0.31	0.00 / 0.03	0.00 / 0.04	0.01 / 0.09	0.47 / 0.77	0.43 / 0.62
Qwen2.5-VL-7B-Inst	0.26	0.13 / 0.17	0.16 / 0.32	0.06 / 0.13	0.36 / 0.62	0.24 / 0.47
llava-next-110b-hf	0.24	0.25 / 0.34	0.25 / 0.45	0.06 / 0.13	0.24 / 0.33	0.54 / 0.55
Phi-3.5-vision-inst	0.23	0.32 / 0.53	0.22 / 0.33	0.07 / 0.11	0.41 / 0.60	0.33 / 0.35
llava-next-72b-hf	0.23	0.19 / 0.27	0.40 / 0.47	0.07 / 0.13	0.27 / 0.33	0.52 / 0.55
Qwen2.5-Omni-7B	0.23	0.09 / 0.23	0.12 / 0.24	0.06 / 0.12	0.30 / 0.60	0.28 / 0.48
Qwen2-VL-7B-Inst	0.20	0.04 / 0.38	0.03 / 0.52	0.09 / 0.19	0.08 / 0.38	0.11 / 0.58
SmolVLM-Inst	0.19	0.19 / 0.36	0.11 / 0.27	0.06 / 0.10	0.36 / 0.44	0.14 / 0.24
Qwen3-VL-8B-Inst	0.19	0.03 / 0.08	0.14 / 0.20	0.10 / 0.22	0.08 / 0.17	0.39 / 0.41
Phi-4-multimodal-inst	0.18	0.05 / 0.12	0.30 / 0.40	0.08 / 0.15	0.02 / 0.07	0.44 / 0.46
MiniCPM-o-2_6	0.18	0.07 / 0.19	0.20 / 0.35	0.01 / 0.10	0.20 / 0.24	0.34 / 0.41
Qwen2.5-Omni-3B	0.18	0.07 / 0.23	0.09 / 0.24	0.06 / 0.12	0.14 / 0.60	0.13 / 0.36
InternVL3-9B	0.17	0.17 / 0.25	0.37 / 0.44	0.06 / 0.16	0.01 / 0.16	0.21 / 0.40
Qwen2-VL-2B-Inst	0.17	0.04 / 0.32	0.04 / 0.38	0.08 / 0.14	0.09 / 0.31	0.16 / 0.44
gemma-3-27b-it	0.15	0.05 / 0.12	0.49 / 0.55	0.04 / 0.28	0.00 / 0.12	0.21 / 0.25
granite-vision-3.3-2b	0.14	0.27 / 0.38	0.08 / 0.22	0.04 / 0.07	0.20 / 0.52	0.20 / 0.26
Qwen2.5-VL-3B-Inst	0.14	0.02 / 0.26	0.04 / 0.22	0.05 / 0.11	0.09 / 0.50	0.08 / 0.35
Qwen3-VL-2B-Inst	0.12	0.01 / 0.04	0.14 / 0.19	0.05 / 0.11	0.03 / 0.10	0.21 / 0.25
llava-1.5-13b-hf	0.12	0.01 / 0.37	0.01 / 0.25	0.00 / 0.09	0.01 / 0.49	0.03 / 0.22

Table 19: **Image** modality evaluation results (5 tasks out of 11). All tasks metrics are `Exact Match / Judge Score`. For UniScienceVQA the Human Baseline is crowd, not an expert. For Human Baseline aggregated results are provided (average of EM and JudgeScore).

Model	Total	AQUARIA	ruEnvAQA	ruTiE-Audio	ruSLUn
Human Baseline	0.895	0.98	0.95	0.75	0.91
Qwen3-Omni-30B-A3B-Instruct	**0.56**	**0.69 / 0.77**	**0.70 / 0.78**	0.43 / 0.44	**0.39 / 0.28**
Qwen2.5-Omni-7B	0.47	0.55 / 0.67	0.55 / 0.70	0.36 / 0.41	0.37 / 0.18
Qwen2.5-Omni-3B	0.38	0.41 / 0.58	0.36 / 0.64	0.30 / 0.36	0.35 / 0.04
MiniCPM-o-2_6	0.37	0.40 / 0.55	0.47 / 0.64	0.25 / 0.32	0.31 / 0.01
ultravox-v0_6-qwen-3-32b	0.32	0.39 / 0.42	0.37 / 0.40	**0.47 / 0.51**	0.00 / 0.00
ultravox-v0_5-llama-3_1-8b	0.31	0.30 / 0.37	0.34 / 0.41	0.29 / 0.31	0.31 / 0.15
ultravox-v0_4_1-llama-3_1-8b	0.31	0.29 / 0.36	0.35 / 0.42	0.31 / 0.32	0.28 / 0.11
ultravox-v0_4	0.30	0.30 / 0.38	0.34 / 0.43	0.29 / 0.31	0.30 / 0.09
ultravox-v0_6-llama-3_1-8b	0.30	0.29 / 0.36	0.26 / 0.40	0.30 / 0.32	0.31 / 0.16
ultravox-v0_3	0.28	0.29 / 0.35	0.35 / 0.43	0.27 / 0.29	0.22 / 0.04
ultravox-v0_4_1-mistral-nemo	0.26	0.18 / 0.34	0.27 / 0.42	0.18 / 0.34	0.26 / 0.14
audio-flamingo-3-hf	0.26	0.21 / 0.42	0.41 / 0.59	0.15 / 0.28	0.00 / 0.00
Qwen2-Audio-7B-Instruct	0.22	0.18 / 0.43	0.15 / 0.52	0.17 / 0.28	0.05 / 0.00
ultravox-v0_2	0.14	0.01 / 0.25	0.00 / 0.38	0.01 / 0.25	0.23 / 0.00
ultravox-v0_5-llama-3_2-1b	0.12	0.04 / 0.25	0.08 / 0.29	0.05 / 0.23	0.00 / 0.00
Qwen-Audio-Chat	0.12	0.01 / 0.28	0.01 / 0.40	0.01 / 0.24	0.00 / 0.00
ultravox-v0_3-llama-3_2-1b	0.12	0.06 / 0.24	0.08 / 0.29	0.05 / 0.20	0.00 / 0.00
SeaLLMs-Audio-7B	0.10	0.08 / 0.14	0.02 / 0.20	0.14 / 0.23	0.00 / 0.01

Table 20: **Audio** modality evaluation results. All tasks metrics are `Exact Match / Judge Score`. For *ruSLUn* dataset the metrics are `Intent Exact Match / Slot F1 Score`. For Human Baseline aggregated results are provided (average of EM and JudgeScore).

Model	Total	CommonVideoQA	RealVideoQA	ruHHH-Video
Human Baseline	0.92	0.96	0.96	0.84
Qwen2.5-VL-72B-Instruct	**0.63**	**0.57 / 0.64**	**0.65 / 0.73**	**0.53 / 0.63**
Qwen3-VL-8B-Instruct	0.58	0.53 / 0.60	0.61 / 0.69	0.46 / 0.56
Qwen2-VL-72B-Instruct	0.56	0.49 / 0.62	0.57 / 0.70	0.34 / 0.62
Qwen2.5-VL-7B-Instruct	0.52	0.49 / 0.56	0.58 / 0.67	0.41 / 0.44
Qwen2.5-Omni-7B	0.44	0.42 / 0.55	0.44 / 0.61	0.24 / 0.39
Qwen2.5-VL-3B-Instruct	0.43	0.39 / 0.49	0.47 / 0.60	0.25 / 0.37
Qwen3-VL-2B-Instruct	0.42	0.41 / 0.48	0.51 / 0.59	0.20 / 0.31
Qwen3-Omni-30B-A3B-Instruct	0.41	0.00 / 0.00	0.64 / 0.72	0.48 / **0.63**
MiniCPM-o-2_6	0.37	0.33 / 0.49	0.41 / 0.57	0.13 / 0.30
Qwen2.5-Omni-3B	0.34	0.30 / 0.46	0.33 / 0.54	0.12 / 0.27
InternVL3-9B-Instruct	0.32	0.27 / 0.30	0.25 / 0.32	0.33 / 0.44
InternVL3-9B	0.32	0.28 / 0.31	0.26 / 0.31	0.32 / 0.42
Qwen2-VL-7B-Instruct	0.30	0.09 / 0.50	0.12 / 0.61	0.04 / 0.45
InternVL3_5-4B	0.29	0.27 / 0.31	0.28 / 0.32	0.24 / 0.32
LLaVA-NeXT-Video-7B-hf	0.13	0.09 / 0.21	0.09 / 0.24	0.03 / 0.09

Table 21: **Video** modality evaluation results. All tasks metrics are `Exact Match / Judge Score`. For *ruHHH-Video* dataset the metrics are `Group Exact Match / Group Judge Score`. For Human Baseline aggregated results are provided (average of EM and JudgeScore).

Dataset	Overlap	Num samples	Total, $	Per item, $	Per hour, $	IAA
AQUARIA	5	786	324.42	0.41	7.38	93.87%
CommonVideoQA	5	1200	2364.12	1.97	8.53	92.41%
LabTabVQA	5	339	518.99	0.31	11.04	87.91%
RealVQA	5	1 010	405.82	0.40	8.54	69.65%
RealVideoQA	5	671	785.54	1.17	8.53	92.19%
ruCLEVR	5	2 063	1 440.15	0.70	7.40	93.36%
ruCommonVQA	5	2 922	10 401.71	3.56	8.54	79.24%
ruEnvAQA	5	644	239.58	0.37	7.38	89.46%
ruHHH-Image	5	610	276.94	0.45	8.54	90.25%
ruHHH-Video	5	911	638.40	0.70	708	91.28%
ruMathVQA (crowd)	5	2 975	214.71	0.07	1.10	85.01%
ruMathVQA (expert)	5	2 975	1 363.58	2.29	1.35	83.09%
ruNaturalScienceVQA (crowd)	5	403	119.55	0.30	6.09	90.37%
ruNaturalScienceVQA (expert)	3	403	123.08	0.31	9.65	96.69%
RuSLUn	5	741	133.00	0.03	2.88	81.05%
ruTiE-Audio	3	500	65.16	0.13	3.62	81.00%
ruTiE-Image	3	500	65.16	0.13	3.62	85.33%
SchoolScienceVQA (crowd)	5	1 750	1 293.83	0.74	8.54	67.56%
SchoolScienceVQA (expert)	3	1 750	1 270.70	0.73	8.94	81.23%
UniScienceVQA	5	1 150	102.4	0.09	2.41	45.19%
WEIRD	5	889	109.1	0.10	9.4	90.84%

Table 22: Payrates and total expenses for human baseline annotation.