Turing-test Interview Emulation

Anonymous ACL submission

Abstract

The paper presents a novel artificial intelligence evaluation methodology grounded in the principles of the classic Turing test. We propose the **Turing-test Interview Emulation** (TiE) framework, which simulates a structured dialogue for the behavioral assessment of model capabilities. In contrast to the original test, our methodology employs a sequential question-answer format across diverse thematic categories, requiring the model to maintain dialogue context and perform comparative analysis of candidate responses. The model is tasked not with selecting a single correct answer but with identifying the most appropriate option from several alternatives, thereby complicating the decision-making process and introducing an additional reasoning step. The paper introduces both text and multimodal versions of the TiE dataset in English and Russian. Using this benchmark, we conduct a comprehensive comparative evaluation of a range of open-source and proprietary large language models (LLMs) and vision-language models (VLLMs).

1 Introduction

The paper "Computing Machinery and Intelligence" (1950) presents Turing's foundational exposition of the well-known imitation game, commonly referred to as the Turing Test (Turing, 2021). This test is often interpreted as a benchmark for a machine's ability to engage in plausible dialogue, leading to frequent claims of its "passing" by various conversational systems. The original formulation, however, establishes critical constraints: the judge must be aware that one participant is a machine, and the human participant must aim to help the judge discern the truth. Emerging from the seminal 1930s discourse on computability and mind (fueled by the works of Gödel, Herbrand, Church, Kleene, Rosser, and Turing himself), the test is theoretically grounded in the Church–Turing principle (Copeland, 1997). At its core, it is a comparison of the problem-solving capabilities of two systems: the human mind and a machine under evaluation. The formal completeness of natural language allows the judge to formulate any task, thereby testing the functional equivalence of these systems. While selecting appropriate dialogue responses is one valid task, the judge's scope is unbounded, ranging from solving logical puzzles and anagrams to evaluating humor, spatial reasoning, or causal inference.

Despite the remarkable progress of modern Large Language Models (LLMs), it remains possible to identify specific problem classes where they perform significantly worse than non-expert humans, though constructing such discriminative tasks presents a non-trivial challenge. This persistent gap is evidenced by the rapid, continual emergence of new benchmarks. The act of "passing" one benchmark often merely reveals the next layer of capability disparity, prompting the community to devise yet another, more challenging test (for example, Humanity Last Exam (Phan et al., 2025)). This work leverages this foundational understanding to design targeted, multi-modal evaluations that probe these persistent gaps in machine intelligence.

To this end, we present the **Turing-test Interview Emulation (TiE)**, a novel evaluation framework that simulates the conversational dynamics of a full Turing test interview. The associated dataset emulates a coherent, multi-turn dialogue where a subject is presented with a series of questions spanning diverse thematic categories relevant to the test's core constructs. Crucially, the dialogue is stateful: questions may presuppose context from earlier exchanges, requiring the maintenance of discourse coherence.

The principal contributions of this work are threefold:

- We propose a novel methodology for AI evaluation based on dialogue-emulated Turing-imitation tests.

- We introduce and publicly release [1] the TiE datasets, comprising both text-based and multimodal versions, available in English and Russian to facilitate broader investigation.
- We provide a comprehensive comparative evaluation and analysis of a range of open- and closed-source large language models (LLMs) and vision-language models (VLLMs) using the proposed benchmark.

2 Related work

Alan Turing's thought experiment (1950) (Turing, 2021) became a programmatic episode in the field of AI. The question of the ability of LLM to pass the Turing test in one form or another is regularly raised; this is due to the popularity of the test. Among the most recent works, one can recall (Murugesan, 2025), (Carlson, 2025), the work on the "reverse Turing test" (Sejnowski, 2023).

TURINGBENCH (Uchendu et al., 2021) operationalizes the classical Turing Test, evaluating a model's global human-likeness through open-ended, adversarial human-judged conversations. DialogBench (Ou et al., 2024) decomposes human-like dialogue into 12 specific, task-based capabilities for granular diagnostic evaluation. X-TURING (Wu et al., 2025) extends this paradigm into a long-term, multimodal "stress test," assessing sustained coherence and engagement over extended interactions. Collectively, they progress from testing holistic imitation to diagnosing discrete skills and, finally, to evaluating persistent, complex conversational intelligence.

However, a significant gap exists for equivalent, comprehensive evaluation resources in Russian. Existing open datasets primarily consist of annotated conversational collections, such as Telegram dialogues with relevance labels (Petrov, 2023), or tests focused on narrow facets like dialogue role tracking. These do not constitute a holistic, task-oriented evaluation framework. This scarcity extends to the spoken domain; although high-quality corpora of spontaneous Russian speech exist [2], they are typically proprietary and not accessible for open academic research.

[1] All examples from the development set are publicly available with their corresponding answers and can be used for few-shot evaluation or as example of the task. The test set is kept private to mitigate the risk of data contamination.

[2] https://defined.ai/datasets/russian-spontaneous-dialogue#russian-spontaneous-dialogue-form

3 Methodology

3.1 Idea

We draw inspiration from the applied reinterpretation of Alan Turing's "imitation game". The focus is no longer on determining whether a system can be mistaken for a human, but on whether an LLM can act as a reliable and useful conversational agent. Our benchmark evaluates the model's ability to produce responses that are coherent, context-sensitive, goal-directed, and appropriate to the user's intent. This shift reflects current practical demands: in real-world applications, the key criterion is not the identity of the interlocutor but the quality and utility of the interaction. Thus, we prioritize communicative adequacy over surface-level human mimicry, assessing how well a model can handle diverse intents, domains, and dialogue dynamics. Accordingly, success is defined not by whether a model appears human-like, but by whether it can consistently meet users' communicative needs.

Communicative adequacy is a complex construct that requires not only standard knowledge but also pragmatic and cultural awareness, including the ability to navigate spontaneous topic shifts and fluctuating conversational tones (e.g., from serious to playful). It involves sustaining dialogue by providing contextually fitting responses, even when no definitively correct answer exists.

3.2 Datasets

This work introduces three distinct datasets: 1) a text-based test set (available in English and in an adapted Russian version), 2) an audio dataset (presented in Russian), and 3) a visual test set. All datasets share a common methodological framework, detailed in the following section.

The TiE task is structured as a fully-formed, instructional dialogue. Each conversational turn is appended to the model's previous response, thereby constructing a continuous context (see Algorithm 1). The dialogue context is composed of the previous questions and the answer options chosen by the model in prior steps. Crucially, the context does not contain information about all possible answer options for the current step.

Within this format, the model is required to evaluate the given answer options and select the more appropriate one, based on a holistic assessment of both. This design inherently assumes that the model retains and utilizes the entire dialogue history, which may include references to earlier ex-

Algorithm 1: Sequential TiE dialog construction.

Input : Dialogue with questions $\{q_n\}_{n=1}^{N}$, options $\{\mathcal{O}_n\}_{n=1}^{N}$, evaluation prompt P, evaluated model LLM

Output : Per-turn records: inferred choice $\hat{o}_n$ (and optionally raw output y_n)

$H \leftarrow \emptyset$ `// accumulated history shown to the evaluated LLM`

for $n \leftarrow 1$ **to** N **do**

- $q \leftarrow q_n$; $\mathcal{O} \leftarrow \mathcal{O}_n$
- $x \leftarrow P \,\|\, (q, \mathcal{O}, H)$
- $y \leftarrow \mathrm{LLM}(x)$ `// free-form text output`
- $\hat{o} \leftarrow$ `AnswerExtractor`(y)
- **save**$(n, q, \hat{o}, y)$
- $H \leftarrow$ `SimplePrompt` $\|\, (q, \mathcal{O}_n)$

changes. A distinctive feature of the dataset is its departure from binary correctness; instead, it requires the model to process and comparatively judge both provided options to determine the most suitable response.

Throughout the dialogue, the dataset explores three main aspects, which are also established in its metadata:

- **The length of the model's context and memory.** Each question in the dataset is tagged with a $use\_context$ metadata flag. A value of "True" signifies the question is context-dependent, while "False" denotes it is self-contained and can be asked in isolation.
- **Turing imitation**: core reasoning capabilities for sustaining adequate dialogue. Each utterance in the dataset is annotated with a binary $turing\_imitation$ flag in its metadata. The distribution of this label is intentionally imbalanced within each dialogue, reflecting the natural, non-uniform occurrence of such utterances in spontaneous human conversation. Such categories[3] are:
  1. $sentiment$ — emotional coloring;
  2. $intent$ — the intentions of the participants in the dialogue or the characters described in the question;
  3. $style$ — the style of the text; for example, it belongs to the clerical style, certain authors' style, etc.;
  4. $humor$ — the presence of humor, the ability to determine how funny the text is;
  5. $irony$ — irony and its detection;
  6. $facts$ — factual accuracy, honesty;
  7. $profanity$ — profane/obscene vocabulary;
  8. $adult\_content$ content for the adults, sensitive topics;
  9. $text\_metrics$ / $text'n'pic\_metrics$ / $text'n'audio\_metrics$ — simple symbolic/mathematical operations, count the number of letters, consonants, vowels, voiced, deaf, count words with the letter "o", solve the simplest mathematical example given in the text or digital form, etc.);
  10. $language\_structure$ / $language'n'audio\_structure$ / $language'n'image\_structure$— ability to perceive word forms and structural-formative relations in a sentence: inflections, text consistency, spelling/syntax, etc.;
  11. $topic\_modelling$ — ability to determine the subject of the text;
  12. $multilanguage$ — cross-lingual and multilingual tasks;
  13. $algorithmic\_transformations$ — different text shifters, sorting characters, adding/removing parts, duplications, and so on.
- **Category**: the ability of the model to distinguish between the basic classes of problems that are necessary to solve the emulation of the Turing test. The category is consistently annotated meta-data for each utterance across all three datasets, with its classes explicitly balanced per dialogue.
  1. $world$ – knowledge about the world;
  2. $math$ — symbolic calculations, mathematics, logic;
  3. $memory$ — activation of the directed long-term memory function of the model, including some information and a question in memory, extracting some information from long-term memory;
  4. $reasoning$ — conclusions, causal relationships;
  5. $strings$ — operations with strings: anagrams, sub-sequence counting, etc.;
  6. $spell$ — questions related to spelling and the composition of words;
  7. $gamesandrules$ — the ability to handle systems based on rules: games, including chess problems, traffic rules, puzzles,

[3] all categories are identical for the three datasets of different modalities; where there are differences, the modality-specific tags are listed separated by "/".

and similar systems;

8. *sound* — text questions on sound modality and audio form of words, sounds, accents, rhyme, and audio on text;
9. *shape* (questions on associative connections, "awareness" of the forms of the real world through symbolic systems and graphic objects);
10. *lexis* (knowledge of the language system, linguistic knowledge, word formation: hyperonyms/hyponyms, kinship terms, etc;
11. *emotion* — emotion recognition;
12. *ethics* — ethical tasks;
13. *trap* — trick questions, contextual or logical-linguistic traps leading to the wrong answer, knocking off the course of the dialogue.

3.3 Datasets creation

The datasets statistics is presented in Table 1.

Text The text dataset was created from scratch by two professional editors using linguistic games, jokes, memes, set phrases, and elements of the cultural code, with Russian-specific references adapted for English-speaking users (the preliminary translation was done by the "deepl" library [4]). The dataset was manually compiled by internal experts (the details of the dataset creation are presented in the Appendix A.2).

Audio The audio for TiE-Audio tasks was generated using original scripts, authored by specialists, and recorded in-house. To ensure realism, background noise was incorporated from public datasets, as well as from custom studio and controlled field recordings. The scripts were voiced by 9 non-professional speakers, selected to maximize the diversity of vocal characteristics, and supplemented by 2 synthetic voices. The synthetic voices were generated using a proprietary voice synthesis platform. The resulting audio was then post-processed in the REAPER digital audio workstation (DAW)[5]. Within REAPER, the vocal tracks were enhanced using the DAW's native flanger effect, followed by parametric equalization for precise spectral refinement. A summary of the voice characteristics (gender, timbre, prosody, etc.) for all dialogue participants is provided in Table 4 in the Appendix.

[4] deepl library

[5] https://www.reaper.fm/

Vision Images for the picture dataset, collected via a crowdsourcing bot within the same community, were required to be original photographs that were not found online. Thus, the TiE-Vision dataset was constructed using both photographs of real situations and images that contain allusions to cultural elements such as popular song lyrics, common idiomatic expressions, and scenarios from anecdotes and established narratives. All images were resized to be at most 1920 pixels on the largest side and watermarked with the 90% transparent organization logo.

Table 1: Structural and quantitative characteristics of the TiE benchmark datasets. $N/Dial.$ = Number of dialogues; $N/Quest.$ = Number of questions in one dialogue; $Ans.Options$ = answer options (2 or 4); $Modality$ = the dataset class, text-based of multimodal.

Name	N/Quest	N/Dial.	Ans. Options	Modality
TiE-Text	500	10	2	Text (Ru/En)
TiE-Vision	500	3	4	Text + Image
TiE-Audio	500	3	4	Text + Audio

3.4 Evaluation Procedure

We propose to evaluate models using the `lm-eval` framework [6] (Gao et al., 2024; Biderman et al., 2024), an established open-source codebase. Since `lm-eval` does not natively support the TiE evaluation protocol, we additionally provide a dedicated TiE-adapter a fully compatible evaluation harness that replicates the required functionality [7] that supports two key evaluation modes: LLM-as-a-judge (see also 5.1 for reference) and regex-based answer extraction. In this work, we use the generative protocol: the model generates until EOS, after which the required discrete answer is extracted either by parsing or by a judge model; this is crucial because each answer is appended to the dialog history for subsequent turns, and including full unconstrained generations (e.g., long reasoning) would inflate context length and introduce noise that can degrade scores. Context size is controlled either via tokenizer max_length/truncation or, more conveniently, by the number of previous questions included in the custom runner (with an analogous parameter for media size in multimodal variants). Metrics are reported as `Accuracy/Exact Match`, where EM is computed on the extracted discrete options, making the evaluation robust to superficial formatting differences.

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

[7] to be released publicly upon publication

4 Evaluation

Baselines Table 2 summarizes the baseline model suite used in the experiments. We evaluate 17 instruction-tuned LLMs spanning a wide range of scales (2B–685B parameters) and context capacities, and covering text-only as well as multimodal (text+image, text+audio, and text+image+audio) settings. Model selection was driven by two constraints: (i) ensuring sufficient coverage of each modality (at least several representative models per modality) and (ii) prioritizing systems with demonstrated Russian-language competence, since three of the four dataset variants are Russian. To anchor the benchmark against strong contemporary systems, we additionally include a small set of closed-source proprietary models with state-of-the-art text, vision, and audio understanding. For text datasets, we use system instruction that essentially provides the same information about answer format that the user prompt.

Metrics We report baselines on four TiE datasets with three scores per model: `EM` / `RegExp EM` / `Judge EM`. `EM` is a strict Exact Match between the raw generation and the target answer. `RegExp EM` extracts an option from the generation[8] and then applies strict EM. `Judge EM` uses an LLM-as-a-judge (see also 5.1 for reference) to score the generation against each answer option, selects the highest-probability option as the predicted answer, and then applies strict EM. For Audio and Image datasets, the text-only LLMs are evaluated by removing the multimodal content from input data and keeping the textual dialogs scaffold. `RegExp` and `Judge` scores are provided to separate capability from instruction-following/formatting.

Results The Table 3 demonstrates the baseline results for all four TiE datasets. Human performance (see Appendix A.3 for details) is high across all datasets (0.75 audio, 0.77 image, 0.95 text-ru, 0.984 text-en), leaving substantial headroom. The data proves being strong resistant to overfitting (the variance of the metrics between the separate dialogs within one dataset is < 0.03). Because Audio/Image are 4-way multiple choice, scores near 0.25 correspond to random guessing, and many models, including several multimodal systems, cluster around this level on Audio after Regexp/Judge normalization, suggesting that the audio tasks are rather hard for the models. Audio also shows large formatting/instruction-following gaps: models such as Ultravox, Idefics3, and Phi-3.5-vision have near-zero raw EM but rise to ~0.25 under any post-processing, implying they often produce free-form answers that fail strict option-token matching.

In contrast, Image exhibits a clearer capability separation: the strongest multimodal models (gemini-2.0-flash ~0.73 and gpt-4o ~0.705 raw EM) substantially outperform text-only LLMs evaluated without images. The text-only baselines typically sit around ~0.25–0.35 on Image, indicating limited inference from the dialog scaffold alone. Mid-tier multimodal VLMs (e.g., Qwen3-VL-8B, Qwen2.5-Omni) fall between these groups, while some smaller multimodal models show improvements mainly after extraction, again pointing to output-format issues.

For the binary text datasets, the language contrast is stark. Text Ru yields high raw EM for top models (e.g., gemini ~0.916, gpt-4o ~0.895, deepseek-v3.2 ~0.871), whereas Text En has raw EM of 0 for all listed models and only reaches random guessing chances under post-processing (~0.48–0.51). Which is mainly caused by generation artifacts like excessive punctuation or wrong answers (see example of English prompt in Appendix A.4). Since Text Ru and Text En are translations, this pattern indicates strong language- and instruction-following sensitivity in the Russian setting rather than a purely content-driven difficulty shift, which makes sense.

5 Diagnostic Analysis

The experiments are set up to make the key variables in the research questions explicitly controllable in a dialog setting. By encoding each turn with the accumulated chat history and sweeping `context_size`, we treat "growing context" as a direct experimental factor (RQ1). The dataset labels attribute context effects along two axes: unified task categories (to compare communicative functions; RQ1.1) and Turing-imitation facets (to relate performance to human-likeness phenomena; RQ1.2). Finally, because longer histories inherently include more interlocutor involvement (persona/stance/argumentation) and previous model outputs, the same manipulation exposes when interactional structure helps or confounds model success (RQ1.3).

[8] We use rather simple patterns: `\b([ABCD])\b` for letter options, `\b([12])\b` for digit options. If no matches are found, put "-1" as a fallback option

Model	Context Length	Modalities	Parameters	Link	Paper
gemini-2.0-flash-001	1000k	Text, Image, Audio	N/A	models/gemini-2.0-flash	N/A
gpt-4o	128k	Text, Image	N/A	models/gpt-4o	(Hurst et al., 2024)
Qwen3-VL-8B-Instruct	262k	Text, Image	9B	Qwen/Qwen3-VL-8B-Instruct	(Yang et al., 2025)
Qwen3-VL-2B-Instruct	262k	Text, Image	2B	Qwen/Qwen3-VL-2B-Instruct	(Yang et al., 2025)
Qwen3-4B-Instruct-2507	262k	Text	4B	Qwen/Qwen3-4B-Instruct-2507	(Yang et al., 2025)
Qwen2.5-Omni-7B	32k	Text, Image, Audio	11B	Qwen/Qwen2.5-Omni-7B	(Xu et al., 2025)
Qwen2.5-Omni-3B	32k	Text, Image, Audio	6B	Qwen/Qwen2.5-Omni-3B	(Xu et al., 2025)
ultravox-v0_4	131k	Text, Audio	8B	fixie-ai/ultravox-v0_4	N/A
ultravox-v0_3	131k	Text, Audio	8B	fixie-ai/ultravox-v0_3	N/A
Phi-3.5-vision-instruct	131k	Text, Image	4B	microsoft/Phi-3.5-vision-instruct	(Abdin et al., 2024)
Idefics3-8B-Llama3	131k	Text, Image	8B	HuggingFaceM4/Idefics3-8B-Llama3	(Laurençon et al., 2024)
Olmo-3-7B-Instruct	65k	Text	7B	allenai/Olmo-3-7B-Instruct	N/A
gpt-oss-20b	131k	Text	22B	openai/gpt-oss-20b	(OpenAI et al., 2025)
T-lite-it-1.0	32k	Text	8B	t-tech/T-lite-it-1.0	N/A
gemma-3-4b-it	131k	Text, Image	4B	google/gemma-3-4b-it	(Team et al., 2025)
Meta-Llama-3.1-8B-Instruct	131k	Text	8B	meta-llama/Llama-3.1-8B-Instruct	N/A
deepseek-v3.2	163k	Text	685B	deepseek-ai/DeepSeek-V3.2	(DeepSeek-AI et al., 2025)

Table 2: Summary of baseline model specifications and architectural characteristics.

Model	ruTiE Audio	ruTiE Image	ruTiE Text En	ruTiE Text Ru
Human Baseline	0.75	0.77	0.984	0.95
gemini-2.0-flash-001	0.11 / 0.181 / 0.044	**0.73** / **0.703** / 0.605	0.0 / 0.492 / 0.492	**0.916** / **0.916** / **0.917**
gpt-4o	0.055 / 0.077 / 0.028	0.705 / 0.69 / **0.623**	0.0 / 0.489 / 0.489	0.895 / 0.896 / 0.896
Qwen3-VL-8B-Instruct	**0.257** / 0.265 / 0.261	0.393 / 0.399 / 0.408	0.0 / 0.498 / 0.498	0.805 / 0.806 / 0.807
deepseek-v3.2	0.159 / 0.239 / 0.13	0.338 / 0.392 / 0.333	0.0 / 0.481 / 0.481	0.871 / 0.872 / 0.874
Qwen3-4B-Instruct-2507	0.255 / 0.263 / **0.264**	0.349 / 0.36 / 0.351	0.0 / 0.486 / 0.487	0.777 / 0.777 / 0.781
Meta-Llama-3.1-8B-Instruct	**0.257** / 0.259 / 0.259	0.313 / 0.32 / 0.32	0.0 / 0.495 / 0.496	0.729 / 0.73 / 0.732
gemma-3-4b-it	0.254 / **0.269** / 0.263	0.311 / 0.327 / 0.319	0.0 / **0.506** / **0.507**	0.683 / 0.695 / 0.697
T-lite-it-1.0	0.137 / 0.158 / 0.148	0.316 / 0.331 / 0.328	0.0 / 0.483 / 0.483	0.776 / 0.802 / 0.804
gpt-oss-20b	0.174 / 0.183 / 0.175	0.326 / 0.331 / 0.331	0.0 / 0.214 / 0.217	0.865 / 0.866 / 0.872
Olmo-3-7B-Instruct	0.235 / 0.259 / 0.257	0.253 / 0.269 / 0.271	0.0 / 0.494 / 0.494	0.577 / 0.578 / 0.579
Qwen2.5-Omni-3B	0.121 / 0.166 / 0.187	0.203 / 0.237 / 0.24	0.0 / 0.498 / 0.5	0.696 / 0.706 / 0.707
ultravox-v0_4	0.037 / 0.263 / 0.245	0.099 / 0.192 / 0.197	0.0 / 0.496 / 0.497	0.729 / 0.73 / 0.732
ultravox-v0_3	0.044 / 0.249 / 0.217	0.099 / 0.192 / 0.197	0.0 / 0.496 / 0.497	0.729 / 0.73 / 0.732
Qwen3-VL-2B-Instruct	0.163 / 0.241 / 0.241	0.027 / 0.231 / 0.248	0.0 / 0.504 / 0.506	0.621 / 0.622 / 0.627
Qwen2.5-Omni-7B	0.166 / 0.166 / 0.191	0.27 / 0.273 / 0.297	0.0 / 0.172 / 0.178	0.758 / 0.76 / 0.76
Idefics3-8B-Llama3	0.008 / 0.257 / 0.257	0.035 / 0.345 / 0.345	0.0 / 0.489 / 0.493	0.0 / 0.701 / 0.703
Phi-3.5-vision-instruct	0.0 / 0.251 / 0.251	0.0 / 0.265 / 0.265	0.0 / 0.492 / 0.496	0.0 / 0.6 / 0.604

Table 3: Performance evaluation of the baselines using different Exact Match (EM) criteria: 1) raw EM, 2) EM after regex-based normalization, and 3) EM with human adjudication.

Research Questions

RQ1: How does the incremental accumulation of dialog context influence the performance of language models across diverse communicative tasks?

RQ1.1: Does extended context uniformly improve performance, or are there specific task categories (e.g., reasoning, memory, intent detection) where its impact is negative or negligible?

RQ1.2: How does model performance correlate with the Turing emulations of the dialogue as context length increases?

RQ1.3: To what extent does the interlocutor's involvement (e.g., persona consistency, argumentative turns) within the growing context serve as a catalyst or a confounding factor for model success?

5.1 Experimental Setup

We ran experiments on four dataset variants: Russian text, English text, and multimodal Russian in image and audio forms, repeating runs while varying `context_size` (how much prior dialog history is shown), `model_name` (the evaluated LLM), and `media_size` (image resolution or audio sample rate, included to control for modality scaling and better isolate other effects). Each run used an LLM-as-a-judge setup: the judge model[9] receives the question, a target answer option, and the model's free-form generation, and returns a probability that the generation matches the target; for multiple-choice tasks, we call the judge once per option and select the closest option, avoiding brittle parsing while keeping generative outputs. All runs were merged into one table and

[9]The model was taken from (Chervyakov et al., 2025). See huggingface.co/MERA-evaluation/MERA_Answer_judge

analyzed with a GLM [10] (Gill et al., 2019) to estimate coefficients and significance for the factors. We also fit separate GLMs per dataset to study language/modality-specific patterns. The target variable is EM score for each question (binary), the explanatory variables are `model_name` (name of the LLM), `modality` (one of "audio", "image", "text_ru", "text_en"), `context_size` (size of the history the LLM can observe), `media_size` (size in pixels or sample rate of the multimodal content), `turing_imitation` (Turing test-like categories, common among four datasets), `unified_category` (unified question categories, common among four datasets), `use_context` (indicator whether the question requires knowing the history of previous questions). The Binomial model also includes intercept. Clustering by individual questions was modeled.

5.2 Relative Influence of Model Factors: Wald Statistics

The Figure 3 shows joint Wald χ^2 for each main term by modality (symlog scale), revealing strong modality–factor differences. The size of the dialog history (`context_size`) dominates everywhere, but it is much larger in text (especially English, then Russian) than in image/audio, indicating that a longer dialog history matters most when the input grows purely as text. The indicator of questions requiring knowledge of the previous questions (`use_context`) is also mainly a text effect (large for Russian/English, smaller for audio, near-zero for image), suggesting that the single-turn to multi-turn shift changes the task primarily in linguistic settings. The unified question category (`unified_category`) remains substantial across modalities, again strongest in text, meaning coarse question types consistently modulate difficulty. The Turing Test-like category (`turing_imitation`) is strong in text and visible in multimodal Russian, while the extreme English value likely reflects an unstable estimate (e.g., sparse categories/separation or the effect is absorbed by some other factor), so its magnitude should be interpreted cautiously.

5.3 Turing-Imitation Associations

The Figure 1 breaks the Turing-imitation term into categories and shows their joint Wald χ^2 by modality on a symlog scale. The main pattern is stark: text (ru/en) carries more category signal (often $\chi^2 \approx 10^0 - 10^2$), while image and audio show less interdependence, implying that these Turing-like distinctions meaningfully separate outcomes mainly in linguistic settings. Several categories are consistently strong in text (e.g., intent, multilanguage, and text metrics), suggesting robust conversational phenomena that affect human-likeness judgments. The figure also shows language asymmetries, most notably profanity (and adult content) being far more influential in English than Russian, plausibly due to stronger English-centric safety/alignment behavior and lexical cue coverage. Overall, category effects appear dominated by linguistic variation, whereas multimodal runs are likely driven more by media/grounding factors than by these fine-grained Turing-imitation categories.

5.4 Unified Categories Impact

The Figure 2 shows the joint Wald χ^2 for unified question categories by modality (symlog scale). The main result is that category effects are driven primarily by text: most categories have a sizeable effect in textual Russian/English datasets, while image and audio show relatively smaller effects, suggesting that coarse content type explains much more variance when the evidence is primarily linguistic. In text, categories like spell, sound/rhymes, and shape/features stand out, consistent with strong sensitivity to surface-form and descriptive reasoning demands in Turing-style dialogs. There are also language asymmetries — e.g., world (and lexis) appears stronger in English, while some affective/interactional categories (e.g., emotions) look relatively stronger in Russian — plausibly reflecting differences in training coverage, lexical resources, and cultural/idiomatic grounding across languages. In Image, category importance is small yet slightly elevated for categories like shape/features (and to a lesser extent sound/rhymes), which plausibly reflects residual grounding demands tied to describing attributes rather than pure world knowledge. In audio, effects are also tiny and noisier, with faint signals in more form-sensitive categories such as spell/strings/lexis, consistent with speech pipelines (ASR/normalization) reducing orthographic cues and compressing category separability.

6 Conclusion

This paper introduces the Turing-test Interview Emulation (TiE), a novel evaluation framework that

[10] statsmodels.org/stable/glm.html

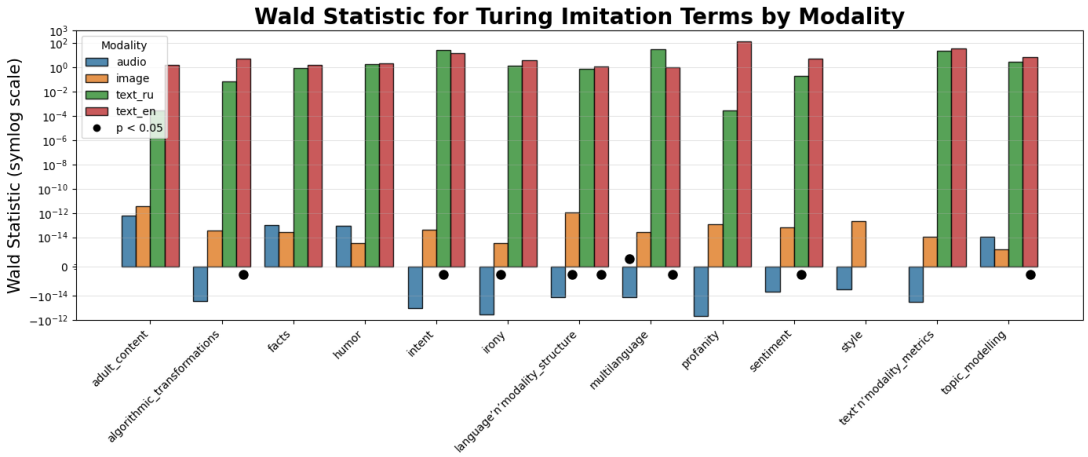


Figure 1: Joint Wald χ^2 statistics for **Turing-imitation category** effects estimated by the GLM, shown separately for audio, image, Russian text, and English text dataset variants (symmetric-log scale). Bars quantify the relative importance of each imitation facet within a modality; black markers indicate categories with $p < 0.05$.

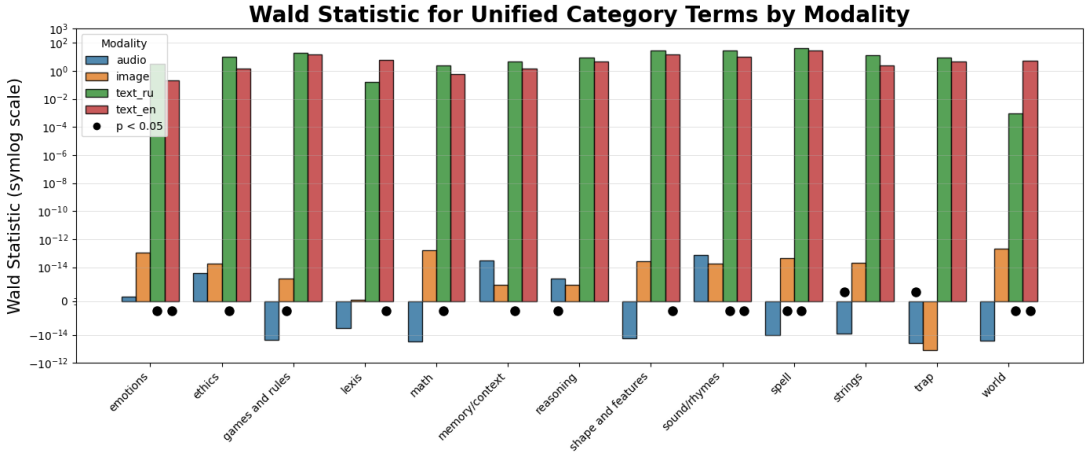


Figure 2: Joint Wald χ^2 statistics from the fitted GLM for **unified question-category** effects, stratified by dataset modality (audio, image, Russian text, English text) and plotted on a symmetric-log scale. Bars indicate the relative contribution of each category to explaining metric variance within each modality; black markers denote categories significant at $p < 0.05$.

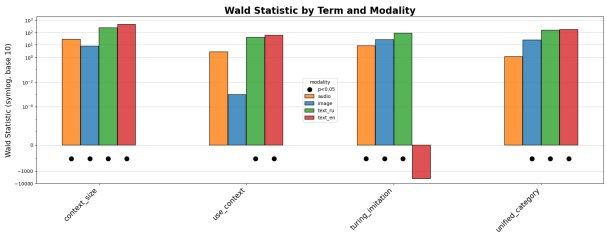


Figure 3: Joint Wald χ^2 statistics for the **main GLM terms** reported separately for audio, image, Russian text, and English text dataset variants (symmetric-log scale). Bars summarize each term's relative importance within a modality; black markers denote effects significant at $p < 0.05$.

adapts the classic Turing test into a structured dialogue format to assess models as adequate conversational agents. We propose TiE dataset versions in both text and multimodal formats, exemplified in English and Russian. Through a comparative evaluation of leading LLMs and VLLMs, we validate the framework's utility and analyze how models perform across diverse categories of "imitation". Our work establishes a new assessment paradigm that prioritizes contextual reasoning and pragmatic appropriateness, offering both a methodological framework and empirical insights into current model capabilities.

Limitations

The present study acknowledges several methodological and dataset constraints that warrant consideration when interpreting the results.

Constrained Dialog Context The core dataset comprises dialogs with a fixed context length of 500 conversational turns. This finite, bounded context is a significant limitation, particularly for multimodal analyses, as it does not capture the dynamics of open-ended, potentially infinite conversations.

Structural Homogeneity of the Dataset The dialogues' formats, diversity, and topical range are intrinsically constrained by the dataset's original structure and its underlying "Turing test" imitation classes. This may limit the generalizability of findings to more organic, unstructured, or thematically varied conversational contexts in practice.

Subsampling for Domain-Specific Analysis While the dataset's metadata allows for segmentation into smaller, topic-specific dialogs (e.g., for domain-specific evaluations based on 20-question samples), this approach inherently fragments the conversational flow. Although the integrity of individual segmented dialogs is preserved, this method cannot fully replicate performance in the original, longer conversational context.

Evaluation Constraints The experimental runs and benchmark assessments were subject to specific technical limitations. These include constraints on the maximum image size processed by the multimodal components [11] and the use of a limited context window, specifically 5-shot examples, for in-context learning evaluations. These factors directly influence the model's performance ceiling in the reported experiments.

Ethical Consideration

The multimodal data (comprising images, audio recordings, and texts) was assembled explicitly with the consent of the original authors and contributors. All data were collected for designated research purposes under agreed-upon terms, ensuring respect for creator rights and autonomy. The images used in this study were not indiscriminately scraped from the public internet. Instead, they were sourced from specific, consented collections, contributing to a more controlled and ethically sound dataset.

Use of AI-Assisted Tools During the preparation of this work, the authors used DeepSeek, ChatGPT and Grammarly for assistance with translation between Russian and English, as well as for improving grammatical clarity and formatting consistency. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

[11] E.g. MLLMs with image comprehension with 256k context length usually to process all 500 questions of a dialog require the images to be no more than 300 pixels on the longest side.

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, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

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, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models. *Preprint*.

Kristen W Carlson. 2025. Large language models pass the turing test. *SuperIntelligence-Robotics-Safety & Alignment*, 2(2).

Artem Chervyakov, Ulyana Isaeva, Anton Emelyanov, Artem Safin, Maria Tikhonova, Alexander Kharitonov, Yulia Lyakh, Petr Surovtsev, Denis Shevelev, Vildan Saburov, Vasily Konovalov, Elisei Rykov, Ivan Sviridov, Amina Miftakhova, Ilseyar Alimova, Alexander Panchenko, Alexander Kapitanov, and Alena Fenogenova. 2025. Multimodal evaluation of Russian-language architectures. *Preprint*, arXiv:2511.15552.

B Jack Copeland. 1997. The church-turing thesis.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abdellah Fourtassi. 2023. Understanding children's multimodal conversational development: Challenges and opportunities.

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, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation. Github.

Jeff Gill, Michelle Torres, and Silvia Michelle Torres Pacheco. 2019. *Generalized linear models: a unified approach*, volume 134. Sage Publications.

Syed Zohaib Hassan, Pål Halvorsen, Miriam S Johnson, and Pierre Lison. 2025. Evaluating llms on generating age-appropriate child-like conversations. *arXiv preprint arXiv:2510.24250*.

Johan Huizinga. 1971. *Homo ludens: A study of the play-element in culture*. Beacon press.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

San Murugesan. 2025. The turing test at 75: Its legacy and future prospects. *IEEE Intelligent Systems*, 40(1):20–24.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu,

Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. 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 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.

Denis Petrov. 2023. Russian dialogues dataset for conversational agents.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Terrence J Sejnowski. 2023. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Alan M Turing. 2021. Computing machinery and intelligence (1950). *Mind*, 59(236):33–60.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. Turingbench: A benchmark environment for turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016.

Weiqi Wu, Hongqiu Wu, and Hai Zhao. 2025. X-TURING: Towards an enhanced and efficient Turing test for long-term dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5874–5889, Vienna, Austria. Association for Computational Linguistics.

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.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Appendix

A.1 Examples

An example of the TiE dataset is presented below:

- **instruction:** *You are given a dialogue that you need to continue. Considering the dialog context, choose the best answer for the last question.*
  {context}
  {question}
  1. {choice1}
  2. {choice2}
  Which answer is most correct?
- **context:** *How many legs does a human have? Two.*
- **question:** *And what about an ant?*
- **choice1:** *Six.*
- **choice2:** *Also two.*
- **outputs** (golden answer): *1*

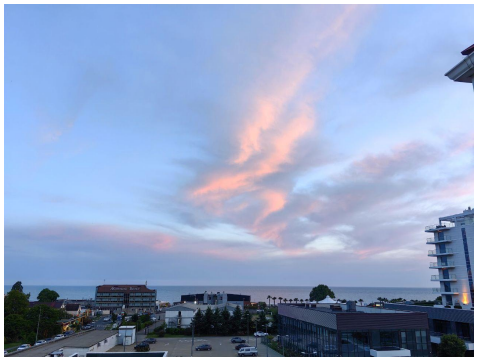
An example of the Vision-TiE dataset, the translated to English script of the audio:

> **Initial question.**
> (*Sound-effects: owl hooting – 4 times, dog barking – 1 time*).
> GIRL'S VOICE. Chatbot, tell me, is it

true that these sounds frighten people? (The general meaning of the answer options: “No, they inspire,” “People cannot hear these sounds,” “People don’t care,” “Yes, people are scared by such sounds.”)

Next question.
GIRL’S VOICE. And if the owl had hooted three times as many times, how many reasons to be scared would there have been?
(The general meaning of the answer options: “Six,” “Twelve,” “Twenty-two,” “Twenty-seven.”)

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

A.2 Dataset characteristics

Not forgetting that The Imitation Game is primarily a game, in forming the dataset, we focused on the ludic aspect of dialogue in the broadest Eurocentric sense of play (Huizinga, 1971). The involvement of the machine in dialogue or polylogue was designed not only according to the principle of developmental-testing dialogue with children (Fourtassi, 2023; Hassan et al., 2025), but also akin to the classic absurdist-playful dialogues from Lewis Carroll’s fairy tales. The element of surprise is highly significant for the game within communication patterns, and of all types: from intellectual to courtship[12].

When selecting tasks for textual tests, we adhered to the concept of maximum diversity in topics and types of play tasks, along with their constant variation, in an attempt to lead the interlocutor into a trap, which is characteristic of playful dialog (the same condition holds for TiE-Vision and TiE-Audio)

Furthermore, in TiE-Audio, we followed not only the principle of selecting maximally unexpected sound clips for recognition but also of embedding them within natural speech situations of their occurrence when writing dialogues. This transforms the recognition of each sound from a simple test ("know"/"don’t know") into a puzzle format for recognizing both the sounding object and the situation of its sound ("if it sounds like this and that one was said before this sound, and those one after, then this is the sound of this kind of objects").

When selecting and grouping actors for voicing the dialogues (see Table 4), we followed the principles:

- To represent in the total sample the most frequent voice types based on acoustic characteristics (clear, hoarse, husky, nasal, resonant, guttural, metallic) and speech analogs of all characteristic singing voices (from bass to soprano);
- Not to complicate voice recognition within a dialogue excessively for either a human or a machine (i.e., to ensure auditory consistency within each dialogue by avoiding extreme vocal contrasts, such as pairing harsh/guttural voices with a weak, unstable one);
- To avoid diction and voice defects (rhotacism, sigmatism, aphonia, etc.);
- To provide each voice with conditions natural to it in terms of speech tempo, intonation, etc., for maximum dialogue naturalness;
- To present each voice not only in its natural mode but also in a playful mode, where the

[12]The dynamic elements of play mentioned Huizinga (1971), such as the deliberate creation of obstacles, adornment, surprise, pretence, tension, etc., all belong to the process of flirting and wooing.

dialogue character impersonates someone else (a historical figure, a fairy tale or anecdote character).

A.3 Human Baseline details

Human Baseline was conducted on two platforms: TagMe [13], Elementary [14]. For each of the four datasets, only one dialog of 500 questions was labeled due to the dataset's design, which (i) limits annotators' capabilities to label the datasets [15], and (ii) assures that all dialogs within one dataset are constructed following the same methodology and category distribution. Each dataset was labeled with an overlap of 3 annotators. Concerning Inter-Annotator Agreement (IAA):

- TiE Audio - 0.61
- TiE Image - 0.64
- TiE Text Ru - 0.85
- TiE Text En - 0.98

A.4 English prompt example

An example of the TiE Text prompt in English version of the test.

```
Choose one of the proposed
answer options for the last
question in the dialogue
below, and write only the
digit 1 or 2 as your answer.
Dialogue:

How many legs does a person
have?
1. Four
2. Two
Answer: 2

What about the ant?
Answer options:
1. Six
2. Two
Answer:
```

[13] tagme.sberdevices.ru

[14] app.elementary.center

[15] Each annotator is required to label all 500 questions in a row.

Dlg.	Char.	Origin	Gender	Voice Timbre & Quality	Prosody & Dynamics	Perceptual Features
1	Marina	N	F	Mezzo-soprano, smooth, guttural	Strong, deep, narrator mode, fast	Loud, crisp, clear
1	Kostya	N	M	Bass-baritone, vibrant, chest-voice	Medium, deep, conversational, inconstant	Loud, fuzzy, soft, muttering
1	Vova	N	M	Tenor, sharp, nasal	Medium, shallow, actor mode, fast	Loud, crisp, hard, slightly hoarse
1	Prof. Z.	S	M	Tenor-altino, sharp, nasal, metallic	Strong, shallow, mechanical, emotionless, slow	Quiet, crisp, hard
2	Alya	N	F	Contralto, smooth, chest-voice	Strong, deep, narrator mode, slow	Loud, crisp, soft, clear
2	Ilya	N	M	Tenor, sharp, nasal	Medium, shallow, actor mode, fast	Loud, crisp, hard, nasal
2	Slava	N	M	Tenor, sharp, chest-voice	Weak, deep, conversational, inconstant	Quiet, fuzzy, soft, hoarse
2	Prof. Z.	S	M	Tenor-altino, sharp, nasal, metallic	Strong, shallow, mechanical, emotionless, slow	Quiet, crisp, hard
3	Fyodor	N	M	Dramatic tenor, sharp, guttural	Medium, deep, conversational, slow	Loud, crisp, medium-soft, gurgling
3	Sasha	N	F	Soprano, medium-sharp, laryngeal-nasal	Medium, shallow, storytelling, fast	Quiet, fuzzy, medium-soft, hoarse
3	Taya	N	F	Coloratura soprano, sharp, laryngeal-chest, shrill	Strong, deep, command mode, inconstant	Loud, crisp, hard
3	Max	N	M	Lyric tenor, smooth, nasal	Weak, shallow, conversational, slow	Quiet, fuzzy, soft, clear

Table 4: Voice characteristics per dialogue participant. $Dlg.$ = Dialogue, $Char.$ = Character, N = Natural, S = Synthetic, M = Male, F = Female.