LlaMa meets Cheburashka: impact of cultural background for LLM quiz reasoning

Mikhail Lifar[1] **Bogdan Protsenko**[1] **Daniil Kupriianenko**[1] **Nazar Chubkov**[1]
Kulaev Kirill[1,2] **Alexander Guda**[1] **Irina Piontkovskaya**[3]
[1]The Smart Materials Research Institute, Southern Federal University, Rostov-on-Don, Russia
[2]Skolkovo Institute of Science and Technology, Moscow, Russia
[3]AI Foundation and Algorithm Lab, Moscow, Russia
Correspondence: guda@sfedu.ru, piontkovskaya.irina@huawei.com

Abstract

Quiz games is the type of intellectual competition which are well suited for testing LLMs reasoning and problem solving skills. Indeed, a good quiz puzzle requires not only factual knowledge, but also the ability to analyze clues given in question, generate hypothesis, and choose the best one using logical reasoning and subtle hints. Recently, modern LLMs have made significant progress in general reasoning tasks, making this kind of evaluation extremely interesting. In this paper, we address a major limitation in the current LLMs' assessment: the models are usually evaluated on English language, or on the multi-lingual benchmarks reflecting English-centric culture, obtained by the translation from the English originals. In the contrary, we test the ability of the modern LLM to deal with the questions of real human quiz games from non-English-speaking society. Namely, we apply LlaMa3-405B to solve the quiz tasks created by the "What?Where?When?" Russian-speaking intellectual gaming community. First, we show, that although the LLM demonstrates strong reasoning and linguistic proficiency in Russian language, the performance diminishes significantly because of the poor knowledge of culture-specific facts. Second, we show the importance of the reasoning strategy choice for answering medium-difficulty questions, for which the model "posses" the necessary knowledge, but the correct answer cannot be given immediately. Evaluating several single- and multi-agent approaches, we obtain 6% improvement in the overall accuracy comparing to the baseline step-by-step reasoning.

1 Introduction

Most of the existing NLP benchmarks were developed with an objective to provide fully automatic and statistically valuable measurement for a given task type ([1]–[3]). This means that each dataset typically consists of the large amount of examples of the same format, derived from a narrow data distribution. Such a kind of the controllable testing is well suited for a task-specific ML algorithms, but does not reflect the abilities of strong modern general-purpose AI models. Another type of LLM evaluation, like Chatbot-Arena ([4]), where models "compete" to each other on the arbitrary tasks provided by users, suffers from excessive variability in results, and provides little feedback about separate model's skills.

Intellectual games offer an appealing alternative, combining well-defined rules and evaluation criteria with the diversity of individual cases, and pose major challenges for AI algorithms. Jeopardy [5], Chess and Go [6] wins against human champions marked important milestones in the AI development.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Nowadays, when the reasoning abilities of LLMs are experiencing explosive development, it is extremely interesting to test their competitiveness in logical thinking and analysis by intellectual games.

Another issue of the existing evaluation is the prevalence of English language, and, more importantly, lack of the tasks requiring different cultural background. Indeed, although multi-lingual benchmarks exists, many of them are obtained by the translation from the English originals [7], [8]. For example the scores on translated MMLU benchmark [9] reflects the ability of the model to answer questions about American school program in other languages, like Yaruba or Hindi, for various topics, including law, literature and history, which does not correspond to the real situation of the language use. Recent works support the presence of such cultural gap [10], [11].

All of the above highlights the importance of novel ideas introduction for general LLM testing, coming from different cultural environments. In this paper, we are looking at one unique cultural phenomenon, namely, Russian-language "What?Where?When?" game, for which archives of thousands of tournaments are publicly available. Using the data from this database, we analyze qualitatively and quantitatively the performance of state-of-the-art LlaMa3-405B model [12]), and propose reasoning strategies suitable for this task.

2 'What?Where?When" game and dataset

"What? Where? When?" (Russian abbreviation is spelled as CheGeKa) is a very popular form of intellectual leisure of the international Russian-speaking community. During the game, players should answer the prepared questions, given a short time for brainstorming. The database consisting of the history of as much as 4390 tournaments is publicly available. For the September 2024, there are 337110 questions in total. Each question supposes a short answer, usually consisting of a single entity or concept. The questions are open-ended, no answer options are provided. A new question set is prepared by volunteers for each game. A good question cannot be answered simply by factual knowledge; question authors try to include non-obvious clues to make searching for the answer fun. When the correct hypothesis is made in the players' mind, there is a feeling that all the parts of the question have fallen into place (players call this a "click").

For CheGeKa dataset ([13], [14]), the authors selected a subset of questions from the database with more factuality and shorter reasoning chains, to make it simpler for modern LLM algorithms. The dataset consists of 29376 questions for train and 520 for the test set. According to MERA leaderboard [15], this dataset is one of the most difficult. Many multilingual models of relatively small size cannot hit even 1% quality threshold. This ability to solve it emerges in larger models, with the best F1 scores of 0.55 for GPT4o and 0.5 for LlaMa3-405B. Significant 0.27 level of Russian-focused GigaChat-7B model demonstrates the importance of the larger target-language pre-training.

Next, we analyze the types of thinking, involved in the process of solving the questions (see more examples in Appendix, Table 2).

Factual or commonsense knowledge. is enough for answering some of the questions. At the same time, for the most of them, human players can find the answer even without direct fact knowledge. E.g., for the question *Say "deed" in Sanskrit* the player can remember concepts related to Indian religious domain, and choose the most fitting one.

Culture-specific knowledge. A large amount of questions involve the knowledge which is common for the most of Russian-speaking people, but almost not known outside of this cultural environment. Such a knowledge may include popular songs, movies, local news or historical events, or traditions. For example, the question *This composer wrote music for the cartoons"Little Raccoon", "Cheburashka", "Mother for a Baby Mammoth", "Shake! Hello" and many others* is one of the easiest for humans, because the songs from these famous cartoons is a part of typical kids playlists, and Russian-speaking people remember the songwriter's name from their childhood; but for the other world, the name of Vladimir Shainskiy doesn't mean anything.

Question analysis. It is of importance to extract all direct and indirect clues from the question. Consider the question: *In Krylatskoye there is a cycling track, in the Druzhba hall there is a swimming pool, and in Mytishchi?* Here, we are given the sequence of pairs "location – sports", where the sports are different, and all the locations are not far from each other. The answer should contain the sports facility located in Mytishchi, the city close to Moscow.

	EI	SC	SD	SDwCA
Mean LLM score on all questions	0.52	0.52	0.55	0.56
All information considered	0.93	0.93	0.96	0.78
Correct reasoning sequence	not relevant	0.81	0.8	0.65
Hallucinations	0.18	0.43	0.27	0.05

Table 1: Important reasoning properties for each of the different prompting approaches used, together with the LLM-estimated score.

Hypotheses generation and ranking. In the question above, it is not clear which object to choose. For example, there is a big Ice Stadium in Mytishchi. But this answer does not "click", because it does not explain the choice of the examples in the question. To get the correct answer, the player should come up with the idea that the list of the facilities for different sports resembles Olimpic Games, and think about Moscow Olimpic Games of 1980. Indeed, both objects in the question were built for this event, which supports the guess. The last step is to recall which competitions took place in Mytishchi. Note that this knowledge is non-trivial, but for local players it can be considered of medium difficulty.

3 Method

We apply the following reasoning strategies:

Method 1. ExtractInfo (EI) A CoT-like approach with multi-agent elements. The first agent tries to extract as much information as possible about the answer from the question text (whether the answer is an object, an action, or a property, what time epoch it may belong to, and so on). The second agent generates a final answer to the question based on the information from the first agent.

Method 2. SelfConsistency (SC) The multiple response generation, among which the most popular response among the generated responses is given as the final response of the model. We used Self Consistency in combination with Chain-of-Thoughts.

Method 3. SuggesterDiscriminator (SD) The first agent – the generator – produces an answer. The second one evaluates it on a 10-point scale. If the score is below the threshold, the generation-evaluation cycle is repeated.

Method 4. SDwCA Suggester-Discriminator with Critical Analysis. Similar to **SuggesterDiscriminator**, but if the score given to the answer is too low, another agent – the critic – explains why this answer is not suitable. Then the generator gives a new answer taking into account all the previous answers and the critique.

Baseline (AsIs) Besides, we implement the baseline method, asking the model to give an answer without the clarification of a reasoning method. In this case, the model utilises the default approach learnt during instruction tuning and alignment phase. We refer it as **AsIs**.

4 Experiments

Experimental details. The proposed prompting approaches were evaluated on the subset of 416 randomly chosen questions from the CheGeKa dataset. From these 416, 50 more challenging questions were selected for more thorough analysis. Llama-3.1 405B was the language model used in our tests. Parameters of response generation were the following: temperature 0.6, top-k 50, top-p 0.9. All the questions were asked, and all the responses were obtained in Russian language. The prompting approaches have the following hyperparameters. For the SC method, the number of generations was set to 4. For SD and SDwCA methods, we set the maximum number of generations to 7, and the threshold of the score to 9.

Evaluation method. Following [15] and [13] benchmarks, we estimate Exact Match (EM) score and token-wise F1 score (by ruGPT tokenization [16]). Besides, we perform LLM-based evaluation, sending to LLM the task and the whole solution together with the ground truth answer, and asking to evaluate the correctness of the answer.

Results. Our main results are presented at Fig. 1. First, we can see that F1 and EM scores significantly underestimate the results, compared to the LLM score. Although the typical answer to the question is

a single word or a short phrase, the standard evaluation methods often cannot capture the variability of the possible correct answers. The examples of such issues can be found in Table 2 in Appendix. Although there are a few cases of LLM evaluation failure, our inspection of the results shows that LLM evaluation is more reliable.

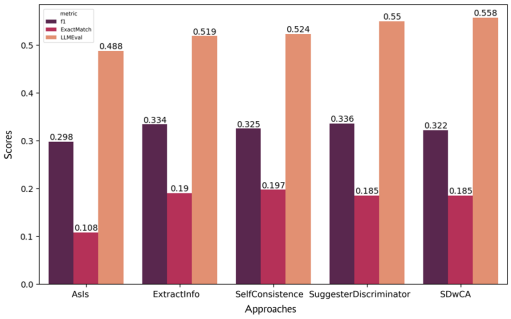

Figure 1: F1, ExactMatch and LLM-estimated scores achieved by different prompting approaches

Second, reasoning-based methods outperform simple prompting by all the metrics. This means that all our prompting strategies lead to improvement upon default LlaMa reasoning; multi-agent approaches with a critic are the best, providing 7% improvement upon "AsIs" and 3% improvement upon single-agent Information Extraction and Self-Consistency methods, as measured by LLM evaluation. Interestingly, that all our reasoning strategy significantly improve Exact Match score (by 8%)

We also measured the entropy of the results across methods and observed the high variability of their output (the results are presented in Appendix).

To understand more deeply the strengths and weaknesses of different methods, we selected 50 questions of medium difficulty, where at least one method succeeded, but not all of them. We manually investigated the solutions, checking the quality of question analysis, the presence of the reasoning chain leading to correct answer, and the presence of hallucinations, i.e. completely made up facts (Table 1). The first interesting observation is a trade-off between reliability and creativity, depending on the choice of the agent-critic. Indeed, SDwCA, in which a verbose critical analysis provided, demonstrates remarkably low level of hallucinations at the cost of the quality of the analysis: this method is less able to understand all the clues in the question, and less likely to generate the correct reasoning chain. From the other hand, SD method with simple score-based critic demonstrates the excellent question analysis results.As for the methods without critic, they both unperformed on question analysis. Moreover, SelfConsistency hallucinates in 43% of cases, which is much higher then other methods. This means that multiple hypotheses generation should be accompanied by critic or other filtering approach.

We provide typical examples of questions and LLM answers in Appendix A

5 Discussion and Limitations

In this paper, we investigate how multi-lingual LLM can answer quiz questions created by Russian-speaking authors to challenge the acuity of the human mind. Our study clearly demonstrates that for the real-life tasks, language proficiency cannot be considered in isolation from the cultural background, just as the reasoning skills hardly can be separated from the language and the world knowledge. We observe that LlaMa model can operate Russian language quite well, with excellent understanding of question details, and generate sound reasoning steps and explanations. From the other hand, the model sometimes struggles with questions which are trivial for the most of human native speakers; such situations demonstrate the pitfall of the model training, but also force the model to activate its reasoning skills. We conclude that culture-specific games is the important and understudied testbed for LLM's abilities.

Finally, we got interesting insights about generator-critic reasoning strategies. We demonstrate that the criticism is a double-edged sword: although in general it improves the reasoning quality, it is able to "demotivate" the model, suppressing the creativity and making the model less focused during the answer analysis. From the other hand, if the critic provides only the score without explanation, there could be the opposite influence, forcing the model too much to output plausible answer based on completely made up facts (hallucinations), resembling KPI-based approach in HR management.

As the limitation of our work, we should mention that we test only one SOTA model. Besides, although Russian is a relatively high-resource language in LlaMa pre-training, the alignment work was

not completed for it. Hence our finding could be considered as the direction for further investigation rather then the final results.

Acknowledgment: ML, BP and AG acknowledge the financial support of the Strategic Academic Leadership Program of the Southern Federal University ("Priority 2030"). Additionally, we acknowledge Skoltech SMILES-2024 Summer School of Machine Learning.

References

[1] A. Wang, Y. Pruksachatkun, N. Nangia, *et al.*, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[2] A. Srivastava, A. Rastogi, A. Rao, *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *arXiv preprint arXiv:2206.04615*, 2022.

[3] D. Hendrycks, C. Burns, S. Basart, *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[4] W.-L. Chiang, L. Zheng, Y. Sheng, *et al.*, "Chatbot arena: An open platform for evaluating llms by human preference," *arXiv preprint arXiv:2403.04132*, 2024.

[5] A. K. Baughman, W. Chuang, K. R. Dixon, Z. Benz, and J. Basilico, "Deepqa jeopardy! gamification: A machine-learning perspective," *IEEE transactions on computational intelligence and AI in games*, vol. 6, no. 1, pp. 55–66, 2013.

[6] D. Silver, T. Hubert, J. Schrittwieser, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.

[7] V. Dac Lai, C. Van Nguyen, N. T. Ngo, *et al.*, "Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback," *arXiv e-prints*, arXiv–2307, 2023.

[8] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *International Conference on Machine Learning*, PMLR, 2020, pp. 4411–4421.

[9] O. AI, *Multilingual massive multitask language understanding (mmmlu)*, 2024. doi: `https://huggingface.co/datasets/openai/MMMLU`.

[10] F. Koto, N. Aisyah, H. Li, and T. Baldwin, "Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu," *arXiv preprint arXiv:2310.04928*, 2023.

[11] C. C. Liu, F. Koto, T. Baldwin, and I. Gurevych, "Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings," *arXiv preprint arXiv:2309.08591*, 2023.

[12] A. Dubey, A. Jauhri, A. Pandey, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[13] E. Taktasheva, T. Shavrina, A. Fenogenova, *et al.*, "Tape: Assessing few-shot russian language understanding," *arXiv preprint arXiv:2210.12813*, 2022.

[14] E. Mikhalkova, "A russian jeopardy! data set for question-answering systems," *arXiv preprint arXiv:2112.02325*, 2021.

[15] A. Fenogenova, A. Chervyakov, N. Martynov, *et al.*, "Mera: A comprehensive llm evaluation in russian," *arXiv preprint arXiv:2401.04531*, 2024.

[16] D. Zmitrovich, A. Abramov, A. Kalmykov, *et al.*, "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)*, 2024, pp. 507–524.

A Examples

Here we present example of questions and model's answers, illustrating various aspects of the typical issues (see Tab. 2). Besides LLM evaluation, we asked two native speakers to answer the set of 50 questions we focused on. We found that the questions are quite difficult for humans: we obtained 10 correct answers in average of 50. From the other hand, there are some questions, which look very simple for humans but LLM cannon solve them, e.g., Q159 from the Table 2.

Id	Question	LLM answer	Correct answer	Comment
207	[Переводы] Скажите на санскрите ‘‘деяние’’ или ‘‘поступок’’. [Translations] Say ‘‘deed’’ in Sanskrit.	Карма Karma	Карма Karma	All approaches give correct answer
284	[Баски] Историческую основу этого произведения составляет эпизод разгрома басками арьергарда отступавших из Испании войск Карла Великого. [Basques] The historical basis of this work is the episode of the Basques’ defeat of the rearguard of Charlemagne’s troops retreating from Spain.	Песнь о Роланде Song of Roland	Песнь о Роланде Song of Roland	All approaches give correct answer. It’s an example of the question, which is very easy for the model, but not so easy for humans.
249	[Холодная тема] Именно до такой температуры полагается охлаждать коньяк перед подачей на стол. [Cold topic] Cognac is supposed to be cooled to this temperature before serving.	(AsIs) 18 градусов (AsIs) 18 degrees Celsius	До комнатной; не надо охлаждать Up to room temperature; no need to cool	Difficult to evaluate
133	[И послушай] Именно эту стандартную формулу римского права диктовал аббат Фариа Эдмону Дантесу, объявив ее впоследствии пением бельканто. [And listen] It was this standard formula of Roman law that Abbot Faria dictated to Edmond Dantes, who later declared it to be bel canto singing.	Аудиатур эт альтера парс -	Audiatur et altera pars Audiatur et altera pars	An example of difficulty to evaluate by EM/F1 metrics. Here, the model’s answer is correct, but it transliterated the Latin phrase to Cyrillic letters
159	[Все могут короли] Песню - название этой темы исполняла именно она. [Kings can do everything] The song - the name of this topic was performed by her.	(Human,IE,SD) Алла Пугачева (SDwCA) ABBA (Human,IE,SD) Alla Pugacheva (SDwCA) ABBA	Алла Пугачева Alla Pugacheva	All humans answered correctly. SC generated multiple wrong hypotheses. SDwCA made incorrect translation to “Winner Takes it All” and predicted ABBA
292	[Почтовая тема] Именно с находки этой своеобразной почты начинается один из романов Жюля Верна. [Mail topic] With the discovery of this peculiar mail one of Jules Verne’s novels begins.	(SC)Дети капитана Гранта (other)Бутылка с письмом (SC) Children of Captain Grant (other)Letter in a bottle	Бутылка с письмом A bottle with a letter	All humans and all models generated the correct hypothesis, but SC couldn’t follow the instruction (generated novel name instead of the mail type)

Id	Question	LLM answer	Correct answer	Comment
98	[Песни] Именно из этого мечтал застрелиться при всех герой песни ‘‘Верка, Надька и Любка’’. [Songs] It was from this that the character of the song ‘‘Verka, Nadka and Lyubka’’ dreamed of shooting himself in front of everyone.	Пистолет, револьвер, наган Pistol, revolver	Царь-пушка Tsar-cannon	Tsar-cannon is a huge memorial cannon, situated in Kremlin. Verka, Nadka and Lyubka are short women’s names, which full versions are Faith, Hope and Love. The model cannot get this
85	[Язык до Киева доведет] Язык ЭТОГО ЖИВОТНОГО в два раза длиннее его собственного тела, но нам более известна некая особенность ЕГО кожи. [Tongue will bring to Kyiv] THIS ANIMAL’s tongue is twice as long as its own body, but we are better aware of a certain feature of HIS skin.	(SDwCA) Муравьед (Other) Хамелеон (SDwCA) Anteater (Other) Chameleon	Хамелеон Chameleon	It is the easy question, but SDwCA “overthought” it, trying to connect the answer to the topic name, which is Russian proverb about the success of talkative (“long-tongued”) people

Table 2: Examples of questions and model answers. Topic name is given in square brackets.

B Response diversity

To study the variability of the results, obtained by different reasoning methods, we plot F1 score against the entropy of the answers predicted by different methods (fig. 2). The smaller cluster in the top left corner correspond to the straightforward questions where the model is the most certain. The concentration of the data in the opposite (bottom right) corner reflects the large disagreement of the answers by different reasoning strategies in the most cases.

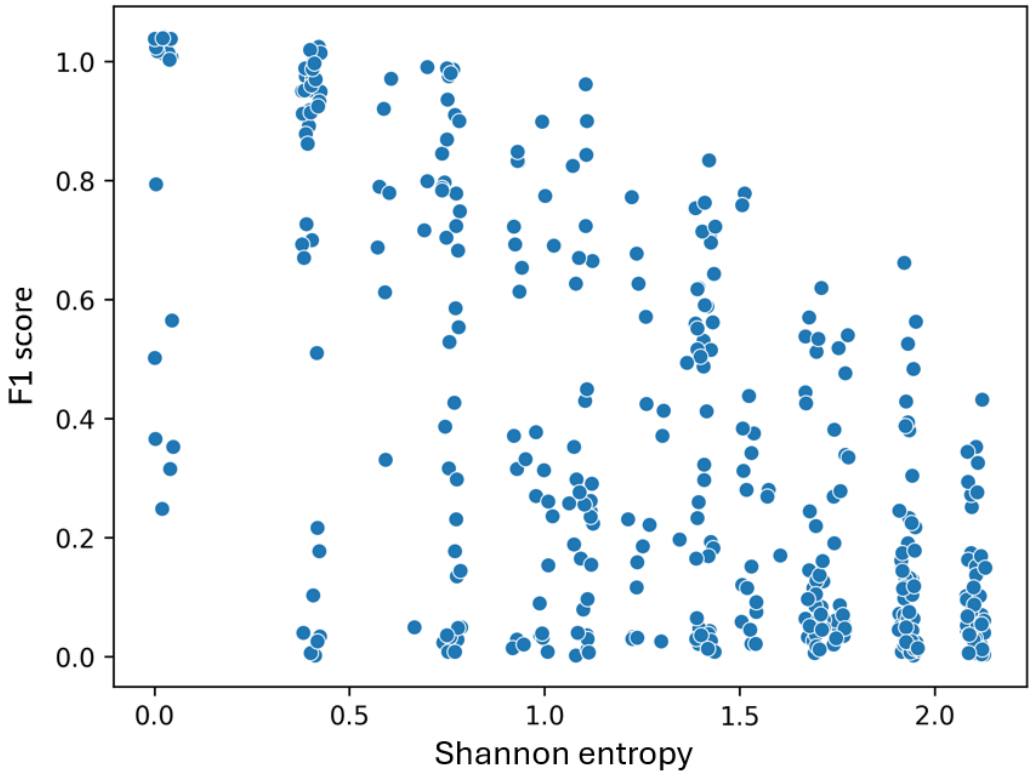

Figure 2: F1-score vs Shannon entropy of responses given by different prompting approaches (which could be thought as a measure of uncertainty of the model)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Clais in the abstract accurately reflect the paper's content

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: We report preliminary evaluation results showing interesting future direction.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theory provided

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We will publish code and data upon acceptance

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will publish code upon acceptance

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Paper describe the experimental setup

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: We focuse on qualitative analysis

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: We utilize inference procedure via expernal service

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We follow ethic rules and code of conduct

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our paper provide preliminary resultsor further analysis

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer:[NA] .

Justification: No risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No assets

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No assets

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer:[NA] .

    Justification: No IP involved

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.