# BIICK-Bench: A Bengali Benchmark for Introductory Islamic Creed Knowledge in Large Language Models

**Umar Hasan**
Department of Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
umar.hasan@northsouth.edu

## Abstract

Large Language Models (LLMs) are increasingly used as information sources globally, yet their proficiency in specialized domains for non-English speakers remains critically under-evaluated. This paper introduces the Bengali Introductory Islamic Creed Knowledge Benchmark (BIICK-Bench), a novel, 50-question multiple-choice benchmark in the Bengali language, designed to assess the foundational Islamic knowledge of LLMs. Crucially, this work is an evaluation of knowledge retrieval and does not endorse seeking religious verdicts (fatwas) from LLMs, a role that must remain with qualified human scholars. Addressing the digital language divide, BIICK-Bench provides a vital tool for the world's second-largest Muslim linguistic community. Fourteen prominent open-source LLMs were evaluated, ranging from 2.5B to 8B parameters. The fully automated evaluation reveals a stark performance disparity, with accuracy scores ranging from 0% to a high of 64%. The results underscore that even state-of-the-art models struggle with Bengali Islamic knowledge, highlighting the urgent need for culturally and linguistically specific benchmarks to ensure the safe and reliable use of AI in diverse communities.

## 1 Introduction

The proliferation of Large Language Models (LLMs) has democratized access to information on an unprecedented scale. For many, these models serve as de facto encyclopedias for topics ranging from science to religion. Within the global Muslim community, individuals increasingly turn to LLMs for quick answers on Islamic history, principles, and practices. However, this convenience is not distributed equally. The vast majority of development and evaluation resources are English-centric, creating a digital language divide that leaves major linguistic communities underserved and exposed to potentially unreliable AI-generated information [2].

This paper addresses this gap by focusing on Bengali, the second most spoken language among Muslims worldwide, by introducing the Bengali Introductory Islamic Creed Knowledge Benchmark (BIICK-Bench), the first automated benchmark designed to evaluate the foundational Islamic knowledge of LLMs specifically in the Bengali language. This work is motivated by the need to provide Bengali-speaking users with a clear understanding of the capabilities and limitations of current AI models on a topic of significant cultural and spiritual importance.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 5th Muslims in ML Workshop.

A critical delineation must be made: this research evaluates the capacity of LLMs to serve as a repository of general Islamic knowledge, not as a source for issuing religious verdicts (fatwas). Within Islamic jurisprudence, a fatwa is a binding legal opinion that requires deep scholarly qualifications and contextual understanding—qualities that current LLMs do not possess. It is firmly maintained that seeking fatwas from AI is impermissible and that this role must remain with qualified human scholars.

The primary contribution of this work is the 50-question, four-option multiple-choice question answering (MCQA) dataset, BIICK-Bench. A comprehensive evaluation of fourteen open-source LLMs was conducted leveraging this benchmark. The findings reveal that no model performs reliably, with the top score being only 64%. This underscores the critical need for language-specific evaluation tools to foster accountability and guide the responsible development of AI for all communities.

## 2 Related Work

The evaluation of LLMs is a vast and active research area. This work is situated at the intersection of general LLM evaluation and the specialized subfield of Islamic and non-English Natural Language Processing (NLP).

**General LLM Benchmarks:** A significant body of work focuses on creating comprehensive benchmarks to probe the general capabilities of LLMs. The most prominent is MMLU [4], which tests models across dozens of subjects in English. Other widely-used benchmarks test for commonsense reasoning, such as HellaSwag [7], or a model's propensity to generate falsehoods, like TruthfulQA [5]. While essential for gauging overall progress, their English-centric nature means they do not address the performance disparities that exist across languages.

**Islamic and Arabic NLP:** There is a growing interest in developing resources for Islamic and Arabic NLP. This includes creating large-scale Arabic question-answering datasets like ArabicaQA [1]. More specific to religious texts, researchers have developed extractive QA datasets for the Qur'an and Hadith, such as QUQA and HAQA [3]. BIICK-Bench contributes to this area by providing the first fully automated multiple-choice benchmark specifically designed to test foundational Islamic knowledge in Bengali. It offers a rapid, objective evaluation method that complements existing resources and serves as a necessary tool for auditing the factual accuracy of models for a major, non-Arabic-speaking Muslim community.

## 3 The BIICK-Bench Benchmark

### 3.1 Motivation and Design

Existing benchmarks for LLMs almost exclusively use English, failing to capture the performance deficits that often appear in other languages. BIICK-Bench was designed to address this gap for the Bengali language with three core principles:

1. **Linguistic and Cultural Specificity:** The benchmark is created entirely in Bengali, using questions from a curriculum designed for native speakers, ensuring cultural and linguistic authenticity.
2. **Theological Consistency:** The questions and answers are derived from a single, consistent curriculum based on the mainstream Islamic theological framework.
3. **Automated and Reproducible Evaluation:** The MCQA format allows for rapid, objective, and reproducible evaluation, eliminating the need for subjective qualitative analysis.

### 3.2 Data Curation and Access

The 50 questions in BIICK-Bench are sourced directly from the final examination of the "Introduction to Islamic Creed" (ইসলামী আকীদার পরিচয়) course offered by Taibah Academy, a well-regarded online platform for Bengali-speaking Muslims. This approach ensures that the questions are expert-validated, theologically consistent, and directly relevant to the target community.

**Reproducibility and Data Access:** The benchmark questions are proprietary to Taibah Academy's course. Thus, they have not been publicly released with this work. However, to ensure this research is transparent and reproducible, clear instructions have been provided for other researchers to access the questions for replication purposes.[1] This approach maintains research integrity while respecting the terms of use for the course materials.

## 4 Experimental Setup

### 4.1 Models Evaluated

A diverse set of fourteen prominent open-source LLMs was evaluated, all accessible on the Hugging Face Hub [6]. All models were run on free-tier Google Colab notebooks (T4 GPU) with 4-bit quantization. The selected models are grouped into three categories:

- **Mainstream Instruct Models:** Llama-3-8B, Mistral-7B, Gemma-7B, Qwen1.5-7B, DeepSeek-7B, Aya-23-8B, Granite-8B, Zephyr-7B-beta.
- **Smaller Foundational Models:** Phi-2 (2.7B), GPT-2-XL (1.5B).
- **Multilingual and Niche Models:** Sarvam-1 (2.5B), Apollo-1-8B, XGLM-4.5B, BLOOM-7B.

### 4.2 Evaluation Protocol

The evaluation was fully automated. Each of the 50 questions from BIICK-Bench was formatted into a standardized Bengali prompt instructing the model to respond with only the letter of the correct option. Model-specific chat templates were applied where appropriate to ensure optimal performance. The model's generated output was parsed to extract the first valid letter (A, B, C, or D). This prediction was then compared against the ground-truth correct answer to determine accuracy.

## 5 Results and Analysis

The performance of the fourteen models on BIICK-Bench is visualized in Figure 1. The results reveal a significant performance gap and highlight the challenges models face with specialized, non-English content.

Gemma-7B is the clear top performer, achieving 64% accuracy, significantly outperforming all other models. This strong result suggests its pre-training data may have contained a more substantial and higher-quality Bengali corpus than its peers.

The second-highest score comes from Sarvam-1, a model specifically optimized for ten Indic languages, including Bengali. Its 44% accuracy is respectable, yet it falls short of expectations for a language-specialized model, performing 20 percentage points below the general-purpose Gemma-7B. This highlights the difficulty of the benchmark's specialized religious domain, suggesting that language optimization alone is insufficient without adequate domain-specific knowledge.

The majority of mainstream instruction-tuned models performed poorly. Llama-3-8B (24%) and Mistral-7B (26%) scored at or near the random-chance baseline of 25%, a concerning result for models often considered to be state-of-the-art. Similarly, massively multilingual models like Aya-23-8B (14%) and BLOOM-7B (14%) also failed, indicating that broad language coverage does not guarantee competence in specialized domains. As expected, the

[1] Replication instructions: 1. Navigate to the Taibah Academy website (`https://taibahacademy.com/course/3`). 2. Create a free account and enroll in the course titled "ইসলামী আকীদার পরিচয়". 3. Complete the course modules to gain access to the final exam, which contains the questions used in our BIICK-Bench. 4. The code is available at: `https://github.com/umarbhasan/biick-bench`.

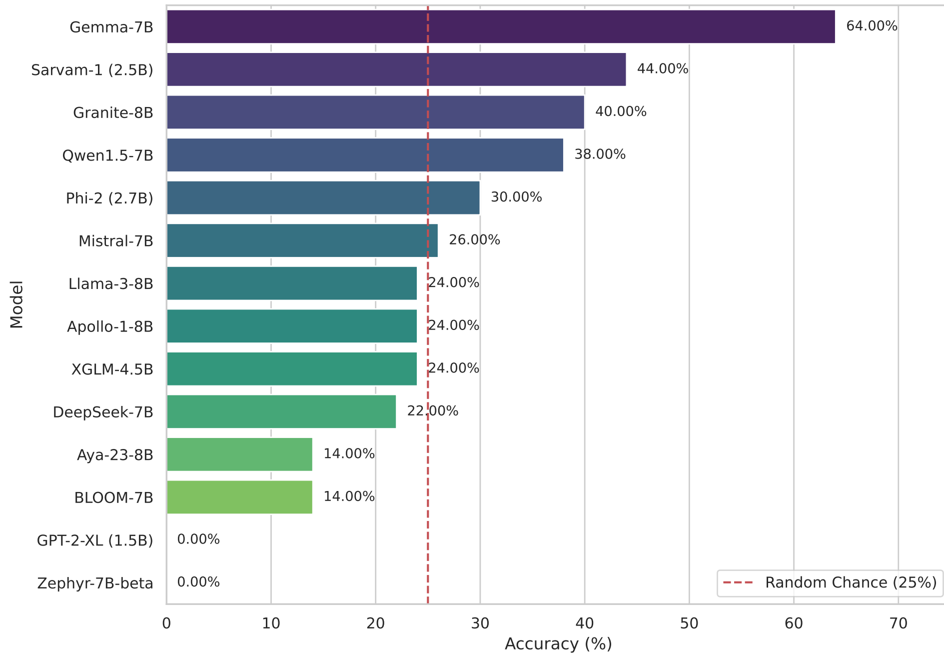


Figure 1: BIICK-Bench accuracy scores visualized. Gemma-7B is a clear outlier. Most models, including top-tier instruction-tuned ones like Llama-3 and Mistral, perform poorly, scoring near the random chance baseline of 25%.

older foundational model, GPT-2-XL, failed completely. Notably, the more recent Zephyr-7B-beta also scored 0%, underscoring the benchmark's difficulty even for modern instruction-tuned models.

## 6 Discussion

The findings have important implications for the Bengali-speaking Muslim community. The starkly low performance across most models, including leading ones, demonstrates that LLMs cannot be considered reliable sources for general Islamic knowledge in Bengali. The results validate the need for community-specific benchmarks like BIICK-Bench to hold developers accountable and to inform users about the risks of uncritical AI adoption.

**Limitations:** This study has limitations. Firstly, BIICK-Bench is small, with only 50 questions. Secondly, the evaluation is limited to a set of publicly available open-source models that could run on free-tier compute. The performance of larger, proprietary models like GPT-5 or Claude 4 remains an open question.

## 7 Conclusion

This paper introduced BIICK-Bench, the first automated benchmark for evaluating the general Islamic knowledge of LLMs in the Bengali language. The evaluation of fourteen models reveals a significant gap in capability, with even top-tier models failing to perform reliably. This work serves as a crucial first step in bridging the digital language divide, providing a necessary tool for the evaluation and responsible development of AI for the global Muslim community.

**Future Work:** Future work can proceed along several directions. The size of BIICK-Bench can be expanded for more robust analysis. The benchmark could also be translated into other under-resourced languages spoken by large Muslim populations. Future efforts could also focus on developing an open-access subset of the benchmark. Finally, a qualitative analysis of common failure modes could provide deeper insights for model developers.

## References

[1] Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. ArabicaQA: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2049–2059, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314.

[2] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *CoRR*, abs/2101.05783, 2021. URL https://arxiv.org/abs/2101.05783.

[3] Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 90–97, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[5] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *CoRR*, abs/2109.07958, 2021. URL https://arxiv.org/abs/2109.07958.

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

[7] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019.

# 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: The abstract and introduction claim to introduce a new Bengali benchmark (BIICK-Bench) and evaluate several LLMs, which is what the paper does. The scope is clearly defined as "general Islamic knowledge" in Bengali and warns against use for fatwas.

   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: [Yes]

   Justification: The "Discussion" section (Section 6) explicitly discusses the small benchmark size and the constraints of the compute resources used, which limited model selection.

   Guidelines:

   - The answer NA means that the paper has no limitations, 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 when 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 not to 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: This paper is an empirical work focused on creating and using a benchmark; it does not present theoretical results that would require proofs.

   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 the 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: The paper fully discloses the models used (with Hugging Face identifiers) and the evaluation protocol. Crucially, Section 3.2 provides explicit, step-by-step instructions for researchers to gain access to the proprietary source questions for replication purposes.

   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 the supplemental material?

   Answer: [No]

   Justification: The evaluation code is released via a GitHub repository linked in the paper. However, the benchmark data itself is proprietary to a third party and cannot be released. The paper provides clear instructions for other researchers to access the data for replication, ensuring transparency while respecting ownership.

   Guidelines:

   - The answer NA means that the 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 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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 details the models tested, their versions, and the automated evaluation protocol. It specifies that models were run with 4-bit quantization on Google Colab.

   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 the 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: [No]

   Justification: As this is a single-run evaluation on a deterministic MCQA benchmark, only direct accuracy scores have been reported. Statistical significance tests are not applicable as there is no variance from multiple runs.

   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 rather 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: [Yes]

   Justification: It has been explicitly stated that all experiments were conducted on free-tier Google Colab notebooks (T4 GPU), providing a clear indication of the compute resources required.

   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: The research conforms to the NeurIPS Code of Ethics. The work is transparent, reproducible (via provided instructions), and includes a careful discussion of the societal context and potential misuse (i.e., using LLMs for fatwas).

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: [Yes]

Justification: The paper discusses the positive impact (providing an evaluation tool for an underserved community) and potential negative impacts (over-reliance on imperfect LLMs). It explicitly frames the work to mitigate harm by warning against using LLMs for religious rulings.

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 the 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: The asset being evaluated is a multiple-choice question dataset derived from an educational course, which does not pose a high risk for misuse.

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: [Yes]

    Justification: The paper uses open-source models accessible on the Hugging Face Hub, all of which have licenses permitting academic research. The source for the questions is also explicitly credited.

    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: [Yes]

    Justification: A new asset, the BIICK-Bench benchmark, has been introduced. Although the data cannot be directly released, clear documentation was provided on its curation, along with instructions for access and replication.

    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: This research does not involve crowdsourcing or human subjects.

    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: This research does not involve human subjects and therefore does not require IRB approval.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, a declaration is not required.

    Answer: [NA]

    Justification: The LLMs are the subject of our study, not an important, original, or non-standard component of the core methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.