REVEAL-Bangla: A Dataset for Cross-Lingual Multi-Step Reasoning Evaluation

Khondoker Ittehadul Islam **Gabriele Sarti**
Center for Language and Cognition (CLCG), University of Groningen
k.i.islam@student.rug.nl g.sarti@rug.nl

Abstract

Language models have demonstrated remarkable performance on complex multi-step reasoning tasks. However, their evaluation has been predominantly confined to high-resource languages such as English. In this paper, we introduce a manually translated Bangla multi-step reasoning dataset derived from the English REVEAL dataset, featuring both binary and non-binary question types. We conduct a controlled evaluation of English-centric and Bangla-centric multilingual small language models on the original dataset and our translated version to compare their ability to exploit relevant reasoning steps to produce correct answers. Our results show that, in comparable settings, reasoning context is beneficial for more challenging non-binary questions, but models struggle to employ relevant Bangla reasoning steps effectively. We conclude by exploring how reasoning steps contribute to models' predictions, highlighting different trends across models and languages. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable versatility across a wide spectrum of natural language processing tasks (Radford et al., 2019). A pivotal breakthrough in enhancing their complex reasoning capabilities has been the introduction of Chain-of-Thought (CoT) prompting (Wei et al., 2022), which encourages models to generate intermediate reasoning steps before arriving at final answers, yielding substantial performance improvements (White et al., 2024; Wang et al., 2022). Despite these advances, the evaluation of LLM reasoning capabilities remains heavily skewed toward high-resource languages, creating significant gaps in our understanding of how these models perform across linguistically diverse

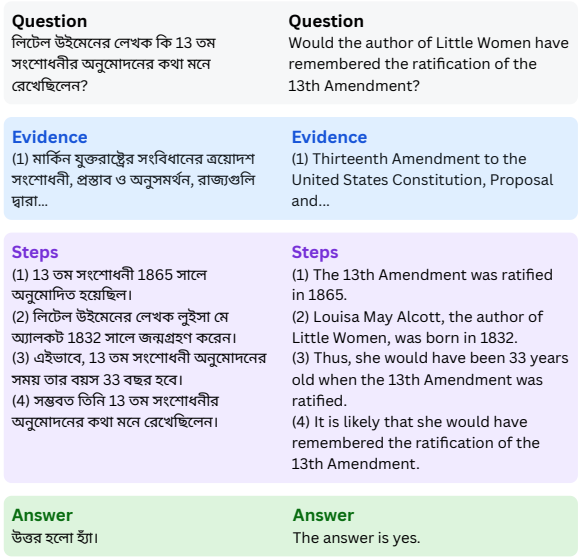

Figure 1: A Row instance of **REVEAL-Bangla** containing translated Question, Evidence, Reasoning Steps and Answer from REVEAL.

contexts. In this work, we focus specifically on the Bangla language, which boasts 268 million speakers and ranks as the sixth most spoken language globally[2], particularly for its computationally challenging morphological richness (Choudhury et al., 2007; Das et al., 2010). As the native language of Bangladesh and the second most prominent Indo-Aryan language after Hindi (Eberhard et al., 2021), Bangla represents a critical case study for cross-lingual reasoning evaluation. The growing technological transformation in densely populated and economically emerging regions where Bangla is spoken (Rahman, 2024) underscores the urgent need for developing faithful AI technologies that can enhance social welfare and economic opportunities. While recent work in Bangla focused on simple extractive or multiple-choice question answering Ekram et al. (2022); Shafayat et al. (2024);

[1] Dataset: https://huggingface.co/datasets/khondoker/reveal-bangla, licensed CC-BY-ND 4.0

[2] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 31–43

Rony et al. (2024), to our knowledge, no datasets with human-validated reasoning steps are available for this language. This lack of resources hinders our ability to assess and improve the reasoning capabilities of LLMs in the Bangla language.

In this work, we address this gap by introducing **REVEAL-Bangla**, a manually translated Bangla version of a subset of the English REVEAL dataset, containing annotated multi-step reasoning chains with gold answers. We exploit our resource and its original English counterpart to evaluate the abilities of two small language models—both proficient in Bangla and English, but one predominantly English-centric, and the other mainly Bangla-centric—in exploiting reasoning step to produce the correct answers given a query, following recent work showing how non-English languages can harm reasoning abilities in LLMs (Qi et al., 2025).

Moreover, recent cross-lingual studies have revealed that generated reasoning chains often exhibit inconsistencies and produce misleading intermediate steps, raising questions about their explanatory reliability (Lanham et al., 2023; Paul et al., 2024). To address these concerns, post-hoc attribution techniques have emerged as valuable tools for analyzing models' internal processes by assigning importance scores to context elements such as summarization (Varun et al., 2024) and retrieved documents (Qi et al., 2024; Cohen-Wang et al., 2024), thereby revealing their contribution to final predictions. We exploit a similar methodology using the CONTEXTCITE method (Cohen-Wang et al., 2024) to examine how reasoning steps contribute to model answers in English and Bangla, highlighting different patterns of importance across the two languages.

2 Related Work

Large Language Models (LLMs) operate as probabilistic sequence predictors, estimating the likelihood of the next token given previous context (Vaswani et al., 2017; Radford et al., 2019). For practical application, explicit training on instructions was found to further improve answer quality (Sanh et al., 2022; Wang et al., 2022). Recently, eliciting reasoning from LLMs, e.g. via step-by-step Chain of Thought reasoning (CoT, Wei et al., 2022), was found to further improve the response accuracy for complex queries.

Some popular reasoning datasets in English include STRATEGYQA (Geva et al., 2021), featuring reasoning- and knowledge-intensive yes/no queries; FERMI (Kalyan et al., 2021), comprising estimation questions that require numerical answers and a blend of knowledge and reasoning; MUSIQUE (Trivedi et al., 2022), which includes multi-hop reasoning questions with free-text entity answers, generated from Wikipedia paragraphs; and SPORTS UNDERSTANDING (Srivastava et al., 2022), consisting of yes/no questions that demand reasoning about sports players, leagues, and maneuvers. Jacovi et al. (2024) combined the aforementioned datasets and human-annotated each LLM-generated step in terms of **attribution** relative to provided Wikipedia paragraphs and **logical coherence** in light of previous reasoning steps. The resulting dataset, dubbed REVEAL, was used to prompt capable LLMs in a chain-of-thought setting and analyzed using Natural Language Inference (NLI) classifiers to evaluate model-generated responses. In our work, we manually translate a subset of REVEAL into Bangla and adopt their setup to evaluate cross-lingual English-Bangla models.

Recently, Jin et al. (2024) explored small and large parameterized models, revealing a linear relationship between accuracy and the number of reasoning steps. We conduct a similar analysis, focusing particularly on small language models (SLMs) with a manageable size (1B parameters). SLMs were recently found capable of high-quality answers in RAG setups (Huang et al., 2024; Liu et al., 2024b), with relevant input information compensating for their limited reasoning abilities. In this work, we use annotated CoTs produced by larger LLMs to investigate whether SLMs can effectively leverage reasoning information in English and Bangla.

3 Development of REVEAL-Bangla

3.1 Data Collection

We start by selecting a subset of the REVEAL dataset.[3] Provided we want to test the ability of SLMs to obtain the correct answer given valid reasoning chains, we focus specifically on examples having all reasoning steps as either logical or fully attributable to the provided Wikipedia paragraphs. Furthermore, among the three models considered by the REVEAL authors to generate answers, we decided to choose the two models with the most answers, i.e., Flan-UL2-20B (Tay et al., 2022) and

[3] https://huggingface.co/datasets/google/reveal

GPT-3 (text-davinci-003, Brown et al., 2020).[4]

We obtain a total of 104 unique questions, with 188 evidence paragraphs and 355 reasoning steps. While only 60% of all the steps are fully attributed to context, all steps are logically relevant. The dataset contains 70% yes–no *binary* questions, making it especially fitting for verifying the relevance of reasoning steps towards a simple atomic answer.[5]

Translation The English→Bangla translation of the selected subset (751 texts) was performed by a native Bangla-speaking graduate student. During the translation process, some digits and certain terms were left unchanged, for instance *76ers* (a basketball team in the NBA), *g/dL* (Grams per decilitre), /kævənd/ (pronunciation of Henry Cavendish), *Équipe d'Haïti de football* (French spelling of the Haitian National Football Team), *inter alia*. As an additional analysis, we assess the quality of automatic translations from Google Translate on the same subset, finding high-quality outputs for health and historical data, but subpar performance on the Sports Understanding subset (examples in Appendix D). Generally, automatic translations were of higher quality when performed one sentence at a time. We employ only the manually translated subset in our evaluation.

4 Evaluation

Model Selection For our evaluation on Reveal and our Bangla variant, we use Llama-3.2-1B-Instruct (or *EngLlama*) (Grattafiori et al., 2024) and BanglaLLama-3.2-1b-bangla-alpaca-orca-instruct-v0.0.1 (or *BenLlama*) (Zehady et al., 2024). *EngLlama* is a popular English-centric multilingual SLM, while *BenLlama* is a Bangla-centric model fine-tuned from *EngLlama* using Bangla-Alpaca-Orca, a collection of instruction tuning examples including the popular Alpaca and OpenOrca datasets (Taori et al., 2023; Lian et al., 2023) automatically translated into Bangla. Importantly, despite their different language focus, both models maintain answering capabilities in both English and Bangla, motivating our cross-lingual analysis.

Prompting Setup We experiment our methods on two main settings: (1) gen_ans, where the model produces an answer without any reasoning step and (2) w_cot_gen_ans, where we provide the model with the annotated reasoning steps from Reveal and Reveal-Bangla.[6] Both models were tested on the English and Bangla Reveal subsets containing the same examples, using a prompt including the query and relevant evidence paragraphs, plus the reasoning steps in the w_cot_gen_ans setting.[7] We test our models on Nvidia A100 GPU, using greedy decoding for reproducible results, and limiting output length to 256 tokens. Reasoning steps in the w_cot_gen_ans setting are appended to the assistant portion of the chat, using continue_final_message = True to let the model complete the generation by producing a final answer. We leave the remaining generation parameters unchanged.

Verifiers To verify the accuracy of the model-generated final answer against the actual final answer, we choose mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 (Laurer et al., 2022) model as it is the only NLI model that supports both English and Bangla. We consider *entailment* labels as correct answers and *contradict* as otherwise. As this NLI model additionally verdicts *neutral*, authors manually verify the response to classify it as valid or not. Furthermore, as language detection tools such as langdetect (Shuyo, 2010) do not support Bangla, we manually assign *contradict* to answers generated in scripts that do not match English or Bangla in the respective settings. We provide our *hypothesis* and *premise* NLI template in the Appendix C.1. We also present additional limitations of the multi-lingual NLI model in the Appendix C.2, to foster research on cross-lingual NLI comprising Bangla.

Results Figure 2 shows the accuracy of tested models in both languages. Unsurprisingly, we find both models performed better on their respective main languages. Moreover, despite their small size, both models were generally found to effectively use the provided reasoning steps to further improve their accuracy. However, we observe that *EngLlama* obtains worse performances when given w_cot_gen_ans steps in Bangla (35.6% →

[4] We do not include Flan-PaLM-540B (Longpre et al., 2023) due to our limited evaluation resources.

[5] We present the counts and tokens distribution of steps and evidence in Figures 5 and 6 in the Appendix.

[6] A third setting prompting the model to generate its own reasoning steps, gen_cot_ans, was not included due to the poor performance of SLMs on CoT reasoning.

[7] Examples of prompt templates for each setting are available in Appendix B.

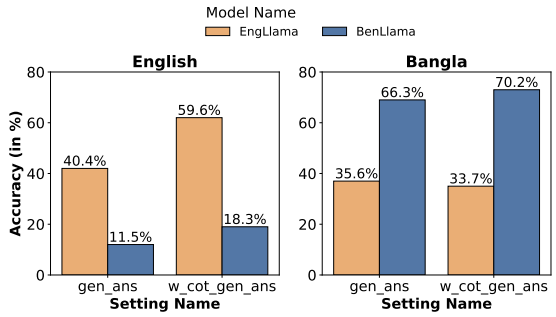

Figure 2: Accuracy of *EngLlama* and *BenLlama* for the gen_ans and w_cot_gen_ans settings on English and Bangla REVEAL subsets.

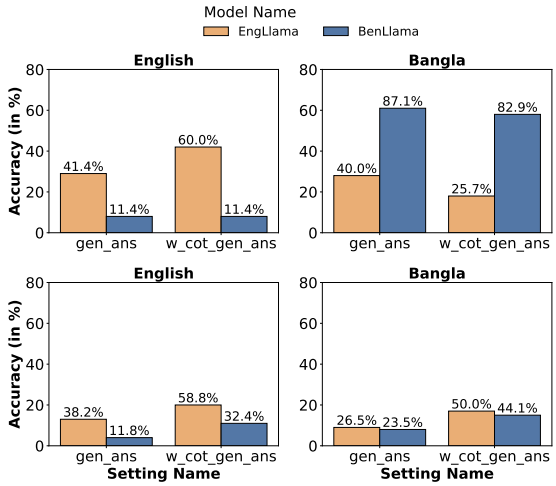

Figure 3: *EngLlama* and *BenLlama* accuracy on REVEAL Binary (top) and Non-Binary (bottom) questions.

33.7%), and find that CoT gains for the *BenLlama* model in Bangla are much milder than for the *EngLlama* model in English (+3.9% vs. +19.2%). These results confirm that, in the less-resourced Bangla setting, *additional relevant reasoning information may not be sufficient to mitigate the limited language capabilities of the tested SLMs*, especially when a Bangla-specific tuning was not performed, as was the case for *EngLlama*.

We further examine model performances across on binary and non-binary questions in the selected REVEAL subset in Figure 3. We find that the *EngLlama* model excels in non-binary questions across both languages, outperforming the *BenLlama* model in both gen_ans and w_cot_gen_ans settings, even in Bangla by a narrow margin. The stronger performance of *BenLlama* in the aggregate case is largely motivated by binary questions, in which the model obtains accuracy $> 80\%$. We also find that while CoT steps have an uneven effect on binary questions, they are consistently beneficial for non-binary ones, across

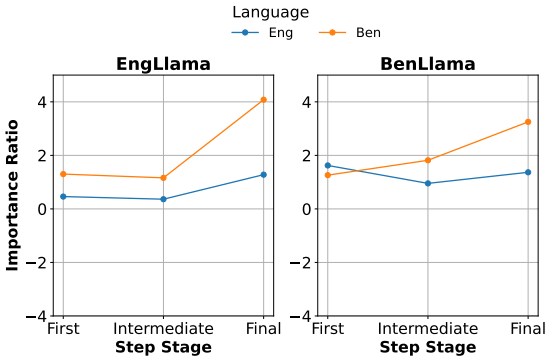

Figure 4: Importance ratio for *EngLlama* and *BenLlama* on w_cot_gen_ans reasoning steps between -4 (lowest) and $+4$ (highest).

both models and languages. This confirms previous findings on the limited effectiveness of CoT in simpler settings by Liu et al. (2024a), and suggests the benefits of CoT generalize even to less-resourced languages.

Attributing Answers to Reasoning Steps To conclude our analysis, we conduct a preliminary investigation into how CoT steps influence model answers. We employ ContextCite (Cohen-Wang et al., 2024) to attribute the final answer generated by the model to the provided reasoning steps in the w_cot_gen_ans setting using surrogate linear models, an approach similar to LIME (Ribeiro et al., 2016). Figure 4 presents an overview of our results for the two models across both languages. We observe that in most cases, later steps tend to have a larger influence on the model response. This suggests that the models place higher emphasis on answer-specific information located in later steps more than on understanding the context provided in earlier steps. This highlights the inherent limitations of these models in context comprehension, which is essential for answering complex questions. Future research could investigate whether this trend holds with larger model sizes. Additionally, we find that both models accord high importance to the Bangla language. We speculate that Bangla's morphological richness causes to assign larger values across attention layers. We leave the exploration of model interpretability in low-resource languages as an interesting direction for future work.

5 Conclusion

We presented REVEAL-Bangla, a manually translated portion of the popular English multi-step reasoning dataset REVEAL. Our cross-lingual analysis

of SLMs revealed limited performance gains from CoT reasoning in the less-resourced Bangla setting compared to English, with gains primarily involving more complex non-binary questions. Further investigation into attributing reasoning steps highlighted differences in importance across models and languages. These findings underscore the need for developing language-specific approaches to enhance reasoning capabilities in low-resource languages, rather than directly transferring techniques optimized for English.

Limitations

Dataset Scale and Coverage Our study is constrained by the relatively small scale of the translated dataset, comprising only 104 unique questions from the original REVEAL dataset. This limited sample size may not fully capture the diversity of reasoning patterns and linguistic phenomena present in Bangla. Additionally, the 70% skew toward binary questions may not accurately reflect real-world reasoning scenarios, potentially overestimating model performance on more complex, open-ended reasoning tasks.

Model Selection Constraints We restricted our evaluation to small language models with 1B parameters due to computational constraints. While this choice enables insights into resource-efficient deployment scenarios, it limits our understanding of how larger, more capable models might leverage Bangla reasoning steps. The exclusion of the gen_cot_ans setting, where models generate their own reasoning chains, further restricts our analysis to scenarios with gold reasoning steps, which may not reflect realistic deployment conditions.

Translation and Annotation Quality Although we employed manual translation by a native Bangla speaker, the translation was performed by a single annotator without inter-annotator agreement measures. This approach may introduce individual biases or inconsistencies in translation choices, particularly for domain-specific terminology in sports and medical contexts. The preservation of certain English terms and pronunciations, while necessary, may also affect how models process the hybrid text.

Evaluation Methodology Limitations Our reliance on the mDeBERTa-v3-base-xnli model for answer verification introduces its own limitations, as acknowledged in our appendix. The model's tendency to produce neutral verdicts required manual intervention, potentially introducing subjective judgments. Furthermore, the absence of Bangla-specific language detection tools necessitated manual script verification, which may not scale to larger evaluations.

Cross-lingual Generalization Our findings are specific to the English-Bangla language pair and may not generalize to other low-resource languages with different linguistic properties, writing systems, or relationships to English. The choice of *BenLlama*, which was fine-tuned on automatically translated instruction data, may also introduce artifacts from machine translation that affect our conclusions about Bangla reasoning capabilities.

Attribution Analysis Scope Our investigation into reasoning step attribution using ContextCite represents only a preliminary analysis. The surrogate linear model approach may not capture complex non-linear interactions between reasoning steps, and we did not explore alternative attribution methods that might reveal different patterns of step importance across languages.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Monojit Choudhury, Vaibhav Jalan, Sudeshna Sarkar, and Anupam Basu. 2007. Evolution, optimization, and language change: The case of bengali verb inflections. In *Proceedings of ninth meeting of the ACL special interest group in computational morphology and phonology*, pages 65–74.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Dipankar Das, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2010. Automatic extraction of complex predicates in bengali. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 37–45.

David Ms Eberhard, David M, Gary F. Simons, and Charles D. Fennig. 2021. Languages of the world.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wenyu Huang, Guancheng Zhou, Hongru Wang, Pavlos Vougiouklis, Mirella Lapata, and Jeff Z. Pan. 2024. Less is more: Making smaller language models competent subgraph retrievers for multi-hop KGQA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15787–15803, Miami, Florida, USA. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. *arXiv preprint arXiv:2402.00559*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. 2021. How much coffee was consumed during emnlp 2019? fermi problems: A new reasoning challenge for ai. *arXiv preprint arXiv:2110.14207*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and Teknium. 2023. Openorca preview1: A llama-13b model fine-tuned on small portion of openorcav1 dataset. https://huggingface.co/Open-Orca/OpenOrca-Preview1-13B.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024a. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *Preprint*, arXiv:2410.21333.

Suqing Liu, Zezhu Yu, Feiran Huang, Yousef Bulbulia, Andreas Bergen, and Michael Liut. 2024b. Can small language models with retrieval-augmented generation replace large language models when learning computer science? In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, ITiCSE 2024, page 388–393, New York, NY, USA. Association for Computing Machinery.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S. Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Riajur Rahman. 2024. Present technology strategy of bangladesh. *BDTask Blog*.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016*

Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 97–101, San Diego, California. Association for Computational Linguistics.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan, Sumon Kanti Dey, Amzad Hossain Rafi, Amzad Hossain Rafi, Ashraf Hasan Sirajee, and Jens Lehmann. 2024. Banglaquad: A bengali open-domain question answering dataset.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. 2024. BEnQA: A question answering benchmark for Bengali and English. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177, Bangkok, Thailand. Association for Computational Linguistics.

Nakatani Shuyo. 2010. Language detection library for java.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, and 1 others. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Yerram Varun, Rahul Madhavan, Sravanti Addepalli, Arun Suggala, Karthikeyan Shanmugam, and Prateek Jain. 2024. Time-reversal provides unsupervised feedback to llms. *Advances in Neural Information Processing Systems*, 37:29777–29806.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, and 1 others. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Abdullah Khan Zehady, Safi Al Mamun, Naymul Islam, and Santu Karmaker. 2024. Bongllama: Llama for bangla language. *arXiv preprint arXiv:2410.21200*.

Appendix

A Dataset

A.1 Sample

Question [E]	Can a Bengal cat survive eating only pancakes?
Question [B]	বেঙ্গল বিড়াল কি শুধু প্যানকেক খেয়ে বেঁচে থাকতে পারে?
Evidence [E]	1. Carnivore, Obligate carnivores: Obligate carnivores are diverse. The amphibian axolotl consumes mainly worms and larvae in its environment, but if necessary will consume algae. All felids, including the domestic cat, require a diet of primarily animal flesh and organs. Specifically, cats have high protein requirements and their metabolisms appear unable to synthesize essential nutrients such as retinol, arginine, taurine, and arachidonic acid; thus, in nature, they must consume flesh to supply these nutrients. 2. Pancake: A pancake (or hotcake, griddlecake, or flapjack) is a flat cake, often thin and round, prepared from a starch-based batter that may contain eggs, milk and butter and cooked on a hot surface such as a griddle or frying pan, often frying with oil or butter. Archaeological evidence suggests that pancakes were probably the earliest and most widespread cereal food eaten in prehistoric societies.
Evidence [B]	1. মাংসাশী, বাধ্য মাংসাশী: বাধ্য মাংসাশী বৈচিত্র্যময়। উভচর অ্যাক্সোলটল তার পরিবেশে প্রধানত কৃমি এবং লার্ভা খায়, তবে প্রয়োজনে শেওলা গ্রাস করবে। গৃহপালিত বিড়াল সহ সমস্ত ক্ষেত্রের জন্য প্রাথমিকভাবে পশুর মাংস এবং অঙ্গগুলির একটি খাদ্য প্রয়োজন। বিশেষত, বিড়ালদের উচ্চ প্রোটিনের প্রয়োজনীয়তা থাকে এবং তাদের বিপাক রেটিনল, আরজিনাইন, টাউরিন এবং অ্যারাকিডোনিক অ্যাসিডের মতো প্রয়োজনীয় পুষ্টি সংশ্লেষণ করতে অক্ষম বলে মনে হয়; এইভাবে, প্রকৃতিতে, এই পুষ্টি সরবরাহ করার জন্য তাদের অবশ্যই মাংস গ্রহণ করতে হবে। 2. প্যানকেক: একটি প্যানকেক (বা হটকেক, গ্রিডল কেক বা ফ্ল্যাপজ্যাক) একটি ফ্ল্যাট কেক, প্রায়শই পাতলা এবং গোলাকার হয়, যা একটি স্টার্চ-ভিত্তিক ব্যাটার থেকে তৈরি করা হয় যাতে ডিম, দুধ এবং মাখন থাকতে পারে এবং একটি গরম পৃষ্ঠে রান্না করা হয় যেমন একটি ভাজা বা ফ্রাইং প্যান, প্রায়শই তেল বা মাখন দিয়ে ভাজা হয়। প্রত্নতাত্ত্বিক প্রমাণগুলি থেকে জানা যায় যে প্যানকেকগুলি সম্ভবত প্রাগৈতিহাসিক সমাজে খাওয়া সবচেয়ে প্রাচীন এবং সর্বাধিক বিস্তৃত খাদ্যশস্য ছিল।
Steps [E]	1. Cats are obligate carnivores, meaning they need to eat meat to survive. 2. Pancakes are not a source of meat. 3. Thus, a Bengal cat cannot survive eating only pancakes.
Steps [B]	1. বিড়াল বাধ্যতামূলক মাংসাশী, যার অর্থ তাদের বেঁচে থাকার জন্য মাংস খেতে হবে। 2. প্যানকেক মাংসের উৎস নয়। 3. সুতরাং, একটি বেঙ্গল বিড়াল শুধুমাত্র প্যানকেক খেয়ে বাঁচতে পারে না।
Answer [E]	The answer is no.
Answer [B]	উত্তর হলো না।

Table 1: Samples from our dataset comprising of question, evidence, steps, and answer where **[E]** and **[B]** following them represents corresponding English and Bangla versions respectively.

A.2 Step Count and Token Distribution

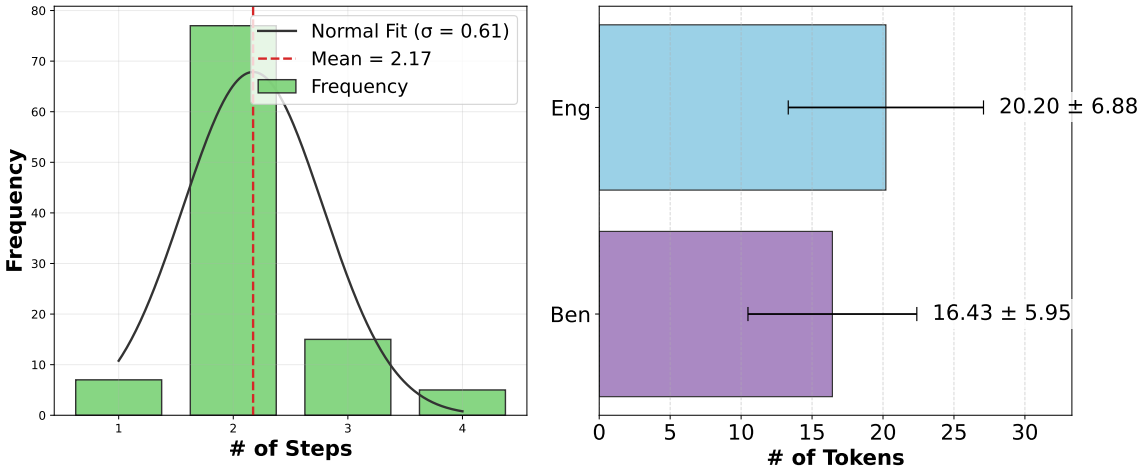

Figure 5: Distribution of Step Count and Token Distribution of Steps. Furthermore, interestingly, number of words required to describe a step in Bangla is less than of English.

A.3 Evidence Count and Token Distribution

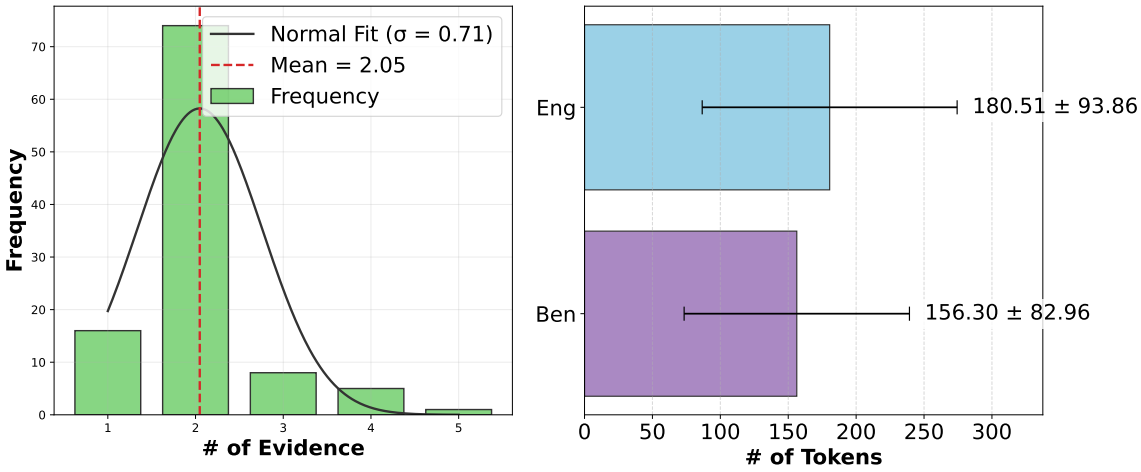

Figure 6: Distribution of Evidence Count and Token Distribution of Steps. On average, there were three evidences associated alongside the questions.

B Example of Chat Prompt Templates

B.1 Setting: gen_ans

En

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful assistant. Your goal is to respond to user queries using the provided evidence paragraphs. The final line must contain the word 'Answer:' followed by the answer to the user query. The response should contain ONLY the final response. If the question requires a yes/no answer, answer using only "yes" or "no". Do NOT provide any additional explanation or comments.<|eot_id|><|start_header_id|>user<|end_header_id|>

Evidence

1. Toilet paper, Description, Materials: Toilet paper is usually manufactured from pulpwood trees, but is also sometimes made from sugar cane byproducts or bamboo.

2. Logging: Logging is the process of cutting, processing, and moving trees to a location for transport. It may include skidding, on-site processing, and loading of trees or logs onto trucks or skeleton cars.

Question:

Would it be hard to get toilet paper if there were no loggers?<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>

Answer:

Bn

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

আপনি একজন উপকারী সহকারী। আপনার উদ্দেশ্য হল প্রদত্ত প্রমাণ অনুচ্ছেদ ব্যবহার করে ব্যবহারকারীর প্রশ্নের উত্তর দেওয়া। চূড়ান্ত লাইনে অবশ্যই 'উত্তর:' শব্দটি থাকবে এবং তারপরে ব্যবহারকারীর প্রশ্নের উত্তর থাকবে। প্রতিক্রিয়াটিতে শুধুমাত্র চূড়ান্ত উত্তর থাকবে। প্রশ্নটির যদি হ্যাঁ/না উত্তরের প্রয়োজন হয়, তাহলে শুধুমাত্র "হ্যাঁ" বা "না" ব্যবহার করে উত্তর দিন। কোন অতিরিক্ত ব্যাখ্যা বা মন্তব্য প্রদান করবেন না।<|eot_id|><|start_header_id|>user<|end_header_id|>

প্রমাণ

1. টয়লেট পেপার, বর্ণনা, উপকরণ: টয়লেট পেপার সাধারণত পাল্পউড গাছ থেকে তৈরি করা হয়, তবে কখনও কখনও আখের উপজাত বা বাঁশ থেকেও তৈরি করা হয়।

2. লগিং: লগিং হল পরিবহনের জন্য গাছ কাটা, প্রক্রিয়াকরণ এবং স্থানান্তর করার প্রক্রিয়া। এতে স্কিডিং, অন-সাইট প্রক্রিয়াকরণ এবং ট্রাক বা এস্কেলেটন গাড়িতে গাছ বা লগ লোড করা অন্তর্ভুক্ত থাকতে পারে।

প্রশ্ন:

কাঠুরি না থাকলে কি টয়লেট পেপার পাওয়া কঠিন হবে?<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>

উত্তর:

Table 2: An example of a chat prompt template from gen_ans setting. **En** is of the corresponding English language and **Bn** is of the Bangla language.

B.2 Setting: w_cot_gen_ans

En	<|begin_of_text|><|start_header_id|>system<|end_header_id|> You are a helpful assistant. Your goal is to respond to user queries using the provided evidence paragraphs. The final line must contain the word 'Answer:' followed by the answer to the user query. The response should contain ONLY the final response. If the question requires a yes/no answer, answer using only "yes" or "no". Do NOT provide any additional explanation or comments.<|eot_id|><|start_header_id|>user<|end_header_id|> # Evidence 1. Toilet paper, Description, Materials: Toilet paper is usually manufactured from pulpwood trees, but is also sometimes made from sugar cane byproducts or bamboo. 2. Logging: Logging is the process of cutting, processing, and moving trees to a location for transport. It may include skidding, on-site processing, and loading of trees or logs onto trucks or skeleton cars. # Question: Would it be hard to get toilet paper if there were no loggers?<|eot_id|> <|start_header_id|>assistant<|end_header_id|> 1. Toilet paper is made from trees. 2. Loggers are responsible for cutting down trees. 3. Thus, without loggers, it would be difficult to get toilet paper. Answer:
Bn	<|begin_of_text|><|start_header_id|>system<|end_header_id|> আপনি একজন উপকারী সহকারী। আপনার উদ্দেশ্য হল প্রদত্ত প্রমাণ অনুচ্ছেদ ব্যবহার করে ব্যবহারকারীর প্রশ্নের উত্তর দেওয়া। চূড়ান্ত উত্তর দেওয়ার আগে, ধাপে ধাপে যুক্তি দিবেন, প্রতিটি যুক্তির ধাপকে একটি নতুন লাইনে সংখ্যাযুক্ত ভাবে তালিকাভুক্ত করবেন। চূড়ান্ত লাইনে অবশ্যই 'উত্তর:' শব্দটি থাকবে এবং তারপরে ব্যবহারকারীর প্রশ্নের উত্তর থাকবে। প্রতিক্রিয়াটিতে শুধুমাত্র সংখ্যাযুক্ত যুক্তির ধাপ এবং চূড়ান্ত উত্তর থাকবে। প্রশ্নটির যদি হ্যাঁ/না উত্তরের প্রয়োজন হয়, তাহলে শুধুমাত্র "হ্যাঁ" বা "না" ব্যবহার করে উত্তর দিন। কোন অতিরিক্ত ব্যাখ্যা বা মন্তব্য প্রদান করবেন না।<|eot_id|><|start_header_id|>user<|end_header_id|> # প্রমাণ 1. টয়লেট পেপার, বর্ণনা, উপকরণ: টয়লেট পেপার সাধারণত পাল্পউড গাছ থেকে তৈরি করা হয়, তবে কখনও কখনও আখের উপজাত বা বাঁশ থেকেও তৈরি করা হয়। 2. লগিং: লগিং হল পরিবহনের জন্য গাছ কাটা, প্রক্রিয়াকরণ এবং স্থানান্তর করার প্রক্রিয়া। এতে স্কিডিং, অন-সাইট প্রক্রিয়াকরণ এবং ট্রাক বা এস্কেলেটন গাড়িতে গাছ বা লগ লোড করা অন্তর্ভুক্ত থাকতে পারে। # প্রশ্ন: কাঠুরি না থাকলে কি টয়লেট পেপার পাওয়া কঠিন হবে?<|eot_id|> <|start_header_id|>assistant<|end_header_id|> 1. টয়লেট পেপার তৈরি হয় গাছ থেকে। 2. কাঠুরিরা গাছ কাটার জন্য দায়ী। 3. এইভাবে, কাঠুরি ছাড়া, টয়লেট পেপার পাওয়া কঠিন হবে। উত্তর:

Table 3: An example of a chat prompt template from w_cot_gen_ans setting. **En** is of the corresponding English language and **Bn** is of the Bangla language.

C NLI

C.1 Structure Example

C.1.1 English

Hypothesis: Who does the actress that played mary poppins in the 1964 film play in princess diaries? The answer is Queen Clarisse Renaldi.
Premise: Who does the actress that played mary poppins in the 1964 film play in princess diaries? The answer is Julie Andrews.

Table 4: Example of the **Hypothesis** and **Premise** structure for English language. Here, **Hypothesis** incorporates ground answer and **Premise** incorporates model predicted answer.

C.1.2 Bangla

Hypothesis: নিউটন সর্বজনীন মহাকর্ষ ধ্রুবকের মান পরিমাপ করার জন্য পদার্থবিজ্ঞানীর কাজের ক্ষেত্রটি কী? সুতরাং উত্তর হলো পদার্থবিদ এবং রসায়নবিদ।
Premise: নিউটন সর্বজনীন মহাকর্ষ ধ্রুবকের মান পরিমাপ করার জন্য পদার্থবিজ্ঞানীর কাজের ক্ষেত্রটি কী? উত্তর হলো নিউটন সর্বজনীন মহাকর্ষ ধ্রুবকের মান পরিমাপ করার জন্য পদার্থবিজ্ঞানীর কাজের ক্ষেত্রটি হলো প্রমাণ।

Table 5: Example of the **Hypothesis** and **Premise** structure for Bangla language.

C.2 Example Cases of Limitations on Bangla

C.2.1 Entails Proper Noun Spelling Mistakes

Hypothesis	রেডক্যাপ অভিনেতা সদস্যের পত্নী কে? উত্তর হলো শিলা হ্যানকক।
Premise	রেডক্যাপ অভিনেতা সদস্যের পত্নী কে? উত্তর হলো শিলা হ্যানক
Ground Label	Contradiction
Predicted Label	Entailment

Table 6: Example of NLI model incorrectly predicting Entailment where the Premise differs with Hypothesis through a spelling mistake on proper noun শিলা হ্যানকক.

C.2.2 Labels *contradict* on model's elaborate correct answers on binary question

Hypothesis	ডরোথিয়া ওয়েন্ডিং কি পোর্শার উদ্ভব জায়গা থেকে এসেছেন? উত্তর হলো হ্যাঁ।
Premise	ডরোথিয়া ওয়েন্ডিং কি পোর্শার উদ্ভব জায়গা থেকে এসেছেন? উত্তর হলো ডরোথিয়া ওয়েন্ডিং পোর্শার উদ্ভব জায়গা থেকে এসেছেন।
Ground Label	Entailment
Predicted Label	Contradict

Table 7: Example of NLI model incorrectly judging Contradict where the Premise contained elaboration on the single word হ্যাঁ (”yes”) answer.

C.2.3 Labels *entailment* where the main final answer is missing

Hypothesis	মহম্মদ আত্তার গাড়ি কী, যে কোম্পানির ড্যাটসান তৈরি করা হয়েছে, এর নজির? সুতরাং উত্তর হলো নিসান আলটিমা।
Premise	মহম্মদ আত্তার গাড়ি কী, যে কোম্পানির ড্যাটসান তৈরি করা হয়েছে, এর নজির? উত্তর হলো নিসান মোটর কোং, লিমিটেড (হেপবার্ন: নিসান জিদোশা কাবুশিকি গাইশা) হল একটি জাপানী বহুজাতিক অটোমোবাইল প্রস্তুতকারক যার সদর দপ্তর নিশি-কু, ইয়োকোহামা, জাপানে। কোম্পানিটি নিসান, ইনফিনিটি এবং ড্যাটসুন
Ground Label	Contradict
Predicted Label	Entailment

Table 8: Example of NLI model falsely predicts Entailment where the actual proper noun answer ”নিসান আলটিমা” (Nissan Altima) is not present in the Premise.

D Google Translation Errors

D.1 Example on American Football Context

Source: DK Metcalf is an American football player. Hitting the wheel route is part of American football. So the answer is yes.
Manual Translation: ডি কে মেটকাফ একজন আমেরিকান ফুটবল খেলোয়াড়। হুইল রাউট করা আমেরিকান ফুটবলের অংশ। সুতরাং উত্তর হলো হ্যাঁ।
Google Translation: ডি কে মেটকাফ একজন আমেরিকান ফুটবল খেলোয়াড়। চাকা রুটে গাড়ি চালানো আমেরিকান ফুটবলের অংশ। তাহলে উত্তর হল হ্যাঁ।

Table 9: Example of **Google Translation**'s sub-par performance compared to **Manual Translation** when given the **Source** English text with the mistake underlined in pink (■).

D.2 Example on Basketball Context

Source: Ben Simmons is a basketball player. Calling for the screen is part of basketball. So the answer is yes.
Manual Translation: বেন সিমন্স একজন বাস্কেটবল খেলোয়াড়। স্ক্রিনের জন্য কল করা বাস্কেটবলের অংশ। সুতরাং উত্তর হলো হ্যাঁ।
Google Translation: বেন সিমন্স একজন বাস্কেটবল খেলোয়াড়। পর্দায় ডাকা বাস্কেটবলেরই অংশ। তাহলে উত্তর হল হ্যাঁ।

Table 10: Example of **Google Translation**'s sub-par performance on Basketball context with the mistake underlined in pink (■).