The Journal of Open Source Software

MM-PoE: Multiple Choice Reasoning via. Process of Elimination using Multi-Modal Models

Sayak Chakrabarty [1] **and Souradip Pal** [2]

1 Northwestern University **2** Purdue University

DOI: 10.21105/joss.07783

Software

- Review
- Repository
- Archive

Editor: Fabian-Robert Stöter

Reviewers:

- @jorgedch
- @hoechenberger

Submitted: 22 October 2024
Published: 19 March 2025

Summary

This paper introduces Multiple Choice Reasoning via. Process of Elimination using Multi-Modal models, also known as Multi-Modal Process of Elimination (MM-PoE), is a method to enhance vision language models' performance on multiple choice visual reasoning tasks by employing a two-step scoring system that first eliminates incorrect options and then predicts from the remaining ones. Our experiments across three question-answering datasets show the method's effectiveness, particularly in visual reasoning tasks. This method addresses one of the key limitations of the paper (Ma & Du, 2023) by extending to tasks involving multi-modalities and also includes experimentation techniques for few-shot settings.

Statement of Need

Large Language models (LLMs) excel at in-context learning for multiple-choice reasoning tasks but often treat all options equally, unlike humans who typically eliminate incorrect choices before selecting the correct answer. The same is true for vision language models (VLMs) in case of visual question-answering tasks with multiple choices. This discrepancy can limit the effectiveness of vision language models in accurately solving such tasks. To address this, we introduce Multi-Modal Process of Elimination (MM-PoE), a two-step scoring method designed to enhance VLM performance by mimicking human reasoning strategies in multi-modal settings.

In the first step, the method evaluates and scores each option, systematically eliminating those that appear incorrect. The second step involves masking these eliminated options, allowing the VLM to focus solely on the remaining viable choices to make a final prediction. Our zero-shot experiments across three datasets demonstrate MM-PoE's effectiveness, particularly excelling in logical reasoning scenarios. Additionally, MM-PoE proves adaptable to few-shot settings and is compatible with the current state-of-the-art vision language models (VLMs).

Using this tool, researchers and practitioners can experiment and significantly improve the accuracy and reliability of VLMs in multiple choice reasoning tasks, making it a valuable tool for advancing machine learning models for visual reasoning.

State of the Field

A common strategy for answering multiple-choice questions, especially under examination conditions, involves a process of elimination where incorrect answers are systematically discarded to narrow down the choices to the most likely correct ones. This approach, grounded in everyday test-taking strategies(Zhang et al., 2023), contrasts with how current language models (LMs) and vision language models (VLMs) handle multiple-choice reasoning tasks. Typically, VLMs evaluate each option independently or collectively without actively discarding less likely answers, potentially reducing their effectiveness in distinguishing the best choice from plausible distractors.

This paper argues that vision language models can benefit from an explicit two-step reasoning process akin to human problem-solving techniques. The proposed method, known as Multi-Modal Process of Elimination (MM-PoE), enhances the decision-making process by first scoring and then eliminating options that are seemingly incorrect before focusing on selecting the correct answer from the remaining choices. This method is designed to align with natural human reasoning by replicating how individuals often approach multiple-choice questions, particularly under the constraint of time and accuracy, as frequently experienced in academic testing environments.

Our hypothesis posits that vision language models, when equipped with a mechanism to discard implausible answers systematically, can achieve better performance on multiple-choice visual reasoning tasks. This is particularly relevant in the context of logical reasoning, where the elimination of clearly incorrect options can simplify the decision process and potentially lead to more accurate outcomes. This idea is supported by previous work demonstrating the effectiveness of LMs in various reasoning tasks when adapted to more human-like reasoning methods (Holtzman et al., 2021).

In the development of MM-PoE, we draw inspiration from the established capabilities of LMs to handle complex reasoning tasks (Brown et al., 2020) and the known strategies that humans employ in test-taking scenarios as depicted in (Ma & Du, 2023). The approach builds on the foundational work in language modeling likelihood (Brown et al., 2020), which demonstrates the LMs' ability to perform in-context learning. By incorporating a structured process to eliminate unlikely choices in a multi-modal setting, MM-PoE aims to refine this capability, making it more targeted and efficient in dealing with the nuanced challenges presented by multiple-choice questions.

The effectiveness of this approach is underscored through zero-shot and few-shot experiments across a diverse set of reasoning datasets, illustrating that the integration of human-like elimination strategies can significantly enhance the performance of vision language models. This paper aims to show that by mimicking human reasoning processes, we can make VLMs not only perform better on standardized visual reasoning tasks but also behave in ways that are more interpretable and aligned with human cognitive processes.

Methodology

The Multi-Modal Process of Elimination (MM-PoE) introduced in this paper operates on a two-step mechanism(Datta & Chakrabarty, 2024) designed to enhance the decision-making capabilities of vision language models (VLMs) in multiple-choice visual reasoning tasks. This method employs a novel approach to option elimination followed by a focused prediction phase. The strategy is rooted in the belief that separating the elimination of clearly incorrect options from the choice of the best remaining option will improve overall task performance.

Problem Setting

Given a multiple-choice visual reasoning task, we define the problem setting as follows:

- Let x be the question or context provided.
- Let h be the image provided.
- Let $Y = \{y_1, y_2, \dots, y_n\}$ be the set of multiple-choice options available.
- Let y be the correct answer from Y.

The goal is to develop an in-context learning method that accurately selects y from Y given x and h.

Two-Step Scoring Method

Step 1: Elimination

In the first step of the MM-PoE method, each option y_i is scored based on a specified metric. The score function, score(x, h, y_i), evaluates each option's plausibility given the question x and image h. The scores are used to eliminate options deemed less likely to be correct. Specifically, options whose scores are below the average score are eliminated. This is calculated as follows:

$$s_i = \text{score}(x, h, y_i)$$

$$Y_{\text{wrong}} = \{y_i | s_i < \text{avg}(s_1, \ldots, s_n)\}$$

This elimination strategy intuitively aligns with how humans often discard options that seem clearly incorrect before carefully considering the remaining choices.

Step 2: Prediction

The second step involves making the final choice from the non-eliminated options. This step utilizes a binary mask to exclude the eliminated options during the prediction phase. The mask for each option y_i is defined as follows:

$$m_i = \begin{cases} 0 & \text{if } y_i \in Y_{\text{wrong}} \\ 1 & \text{otherwise} \end{cases}$$

The masked context x_{mask} is then constructed by modifying the original context x to include only the options for which $m_i = 1$. Each option is scored again, but this time within the context that explicitly excludes the eliminated options, possibly by using a template T that masks out Y_{wrong} in the presentation of the options:

$$x_{\text{mask}} = T(x, Y, \text{mask})$$

The final predicted answer $\hat{y}$ is then the option with the highest score among the remaining options:

$$\hat{y} = \arg\max_{i|m_i=1} \text{score}(x_{\text{mask}}, h, y_i)$$

Experimental Setup

To evaluate the effectiveness of the Multi-Modal Process of Elimination (MM-PoE), we designed an experimental framework that tests the method across a diverse set of visual reasoning datasets. This setup aims to compare MM-PoE with existing scoring methods to highlight its potential improvements in accuracy and reasoning capability. Our experiments primarily focused on a zero-shot setting to evaluate the generalization capabilities of MM-PoE without any task-specific tuning. Accuracy was used as the main metric for performance evaluation, with results averaged over multiple seeds to ensure robustness.

To further explore the versatility of MM-PoE, we also examined its performance in few-shot settings by incorporating examples into the model's input, aiming to observe any changes in effectiveness when provided with context-specific demonstrations.

Data

Our experiments were conducted on three different multiple-choice visual reasoning datasets - Visual Question Answering(VQA) (Antol et al., 2015), ScienceQA (Lu et al., 2022), and Diagram Understanding(AI2D) (Kembhavi et al., 2016), selected to cover a broad spectrum of reasoning types and complexities. These tasks include both traditional visual reasoning tasks and more specialized ones designed to test specific reasoning skills. To ensure a comprehensive evaluation, we used train sets from established benchmarks when available; otherwise, we utilized development sets. In case of varying number of options in the multiple-choice answers for SceinceQA and AI2D datasets, we filtered questions containing image context and exactly four options.

Dataset	#Options	Train	Dev	Test
VQA	18	248,349	121,512	244,302
ScienceQA	4	12726	4241	4241
AI2D	4	3921	982	-

Model

For the core experiments, we utilized the GIT and BLIP models, chosen for their balance between computational efficiency and performance in instruction-tuned vision language tasks. These models have demonstrated strong capabilities in handling various multi-modal tasks and serve as a robust platform for evaluating our MM-PoE method.

Baselines

We compared MM-PoE against five baseline scoring methods to assess its relative performance:

1. **Language Modeling (LM):** This baseline uses the raw vision language modeling likelihood as the scoring function.
2. **Average Language Modeling (AVG):** This method averages the log probabilities across all tokens in the option.
3. **Calibration:** This involves adjusting the VLM scores based on calibration techniques that aim to correct for the model's confidence.
4. **Channel:** Channel methods score each option based on how likely the question is given the option, which reverses the typical conditional probability used in LMs.
5. **Multiple Choice Prompting (MCP):** This approach formats the input by presenting the question followed by all options, prompting the model to select the most likely option.

Each method provides a different approach to scoring options, allowing for a comprehensive comparison of how each interacts with the structure and strategy of MM-PoE.

Implementation

The effectiveness of MM-PoE hinges on the robustness of the scoring function and the accuracy of the elimination step. The scoring function can be any VLM-based likelihood estimator, such as vision language modeling likelihood, or any of its alternatives like average log probability or calibrated log probability. Our implementation tests multiple such scoring functions to identify the most effective ones in both eliminating implausible options and accurately selecting the final answer.

The MM-PoE method is designed to be model-agnostic, meaning it can be implemented using any existing VLM capable of scoring text options, and it is flexible enough to be adapted to different types of multiple-choice visual answering questions across various domains. The scoring functions were carefully chosen based on their theoretical alignment with the two-step elimination and prediction philosophy of MM-PoE. We conducted extensive parameter tuning and optimization to maximize the performance of both the elimination step and the final prediction accuracy.

This experiment setup was designed to rigorously test the effectiveness of MM-PoE across a range of visual reasoning tasks and compare its performance against standard baseline methods. The results of these experiments are intended to demonstrate the potential benefits of integrating a process of elimination approach into vision language model reasoning strategies for multiple-choice questions.

Results

MM-PoE consistently outperformed or matched the best-performing baselines across all datasets, showing particular strength in logical reasoning. The method's effectiveness in separating elimination and prediction tasks was crucial to its success.

Model	Dataset	LM	AVG	Calibration	Channel	MCP	PoE
microsoft/git-base-vqav2	ScienceQA	27.4	17.8	23.2	24.6	25.8	27.2
microsoft/git-base-vqav2	AI2D	25.4	26.2	26.4	25.4	25.3	26.5
microsoft/git-base-textvqa	ScienceQA	21.8	20.4	25.8	23.4	23.6	28.2
microsoft/git-base-textvqa	AI2D	26.5	27.6	20.8	26.2	24.2	26.8

Table 1: Comparison of Multiple-Choice Prompting (MCP) and Process of Elimination (PoE) accuracy scores on 2 visual question answering datasets for the `microsoft/git-base-vqav2` and `microsoft/git-base-textvqa` models in the zero-shot settings. Each dataset has a different number of answer choices. PoE mostly outperforms MCP on all the visual reasoning tasks for the two multi-modal models mentioned.

Examples

ScienceQA Example

Question: Which of these states is farthest north? **Options**: West Virginia, Louisiana, Arizona, Oklahoma **Ground Truth Option**: West Virginia

Predicted Masks: West Virginia, Louisiana, [MASK], [MASK] **Predicted Option**: West Virginia

AI2D Example

Question: Are phytoplankton predators or prey in this food chain? **Options**: producer, predator, prey, NA **Ground Truth Option**: prey

Predicted Masks: [MASK], predator, prey, NA **Predicted Option**: prey

Conclusion

MM-PoE demonstrates a significant improvement in handling multiple choice visual reasoning tasks by mimicking a human-like process of elimination approach. Future work will focus on enhancing its generalizability and efficiency, possibly extending to handle better masking strategies.

Ethics Statement

While this method uses publicly available data and models, users should be aware of potential biases in the data and model outputs.

Acknowledgements

We would like to extend our sincere gratitude to Northwestern University for providing access to their servers and GPU resources, which were instrumental in conducting this research. The computational power and infrastructure made available by the university enabled the efficient processing and analysis of large datasets, significantly contributing to the success of the project. Without this support, the research would not have been possible at the scale or speed required. We deeply appreciate the university's commitment to fostering a collaborative research environment and supporting technological innovation.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. *International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/ICCV.2015.279

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Datta, A., & Chakrabarty, S. (2024). On the consistency of maximum likelihood estimation of probabilistic principal component analysis. *Advances in Neural Information Processing Systems*, *36*. https://doi.org/10.48550/arXiv.2311.05046

Holtzman, A., West, P., Shwartz, V., Choi, Y., & Zettlemoyer, L. (2021). Surface form competition: Why the highest probability answer isn't always right. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7038–7051. https://doi.org/10.18653/v1/2021.emnlp-main.564

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 235–251. https://doi.org/10.1007/978-3-319-46493-0_15

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. *The 36th Conference on Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2209.09513

Ma, C., & Du, X. (2023). POE: Process of elimination for multiple choice reasoning. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4487–4496). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.273

Zhang, Y., Chakrabarty, S., Liu, R., Pugliese, A., & Subrahmanian, V. (2023). SockDef: A dynamically adaptive defense to a novel attack on review fraud detection engines. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2023.3321345