Language Models as Examinees: Benchmarking Distractor Ranking in Multiple-Choice Questions

Tasfia Seuti[1]
tasfiaseuti@my.unt.edu
Sai Naga Chaithanya Aavula[1]
sainagachaithanyaaavula@my.unt.edu
Nivedhitha Duggi[1]
nivedhithaduggi@my.unt.edu
Dheeraj N Jagarlamudi[1]
dheerajjagarlamudi@my.unt.edu
[1]University of North Texas, Denton, TX, USA

Abstract

Our research investigates the behavior of language models, such as GPT and BERT, when used as virtual students to answer multiple-choice questions (MCQs). We specifically focus on the subtask of **distractor ranking**, which involves sorting incorrect answer choices. The evaluation uses 480 held-out MCQs, including 417 single-annotated and 63 double-annotated questions for inter-annotator agreement analysis. We transform model mispredictions into ranking triplets and evaluate performance using metrics such as exact-match accuracy, top-k accuracy, pairwise concordance, weighted Cohen's κ, Spearman's ρ, Kendall's τ, and tie rates. The top-performing models achieve up to 49% pairwise concordance and 32% top-1 accuracy, yet still fall short of human performance (53% concordance) and exhibit negligible rank correlations ($\rho \approx 0$, $\tau \approx 0$). Our analysis further explores domain-specific trends, statistical significance, hyperparameter sensitivity, and computational costs. All code, data, and visualization tools are released to support future research on distractor-aware MCQ evaluation.

1 Introduction

Multiple-choice questions (MCQs) are a foundational tool for both educational assessment and NLP benchmarking. However, traditional evaluation methods focus primarily on correct-answer accuracy, overlooking how models interpret or confuse the incorrect alternatives—known as distractors. Understanding distractor confusion provides valuable insights into model reasoning, error tendencies, and generalization across domains.

This work formalizes **distractor ranking** as a standalone evaluation subtask, where ranked triplets are derived from the frequency of model mispredictions. Our study addresses three major gaps: (1) isolating distractor ranking from accuracy-based metrics, (2) benchmarking a wide range of language models (LMs), and (3) employing a comprehensive suite of statistical evaluation metrics.

Due to hardware constraints with GPT-2 Large, we use GPT-Neo 125M as a decoder-only baseline for fine-tuning. Specifically, we:

- Analyze incorrect predictions to understand model misprediction patterns.
- Extract full ranking triplets (e.g., A $>$ B $>$ C) by counting how frequently each distractor is chosen.
- Fine-tune 62 LM variants spanning encoder-only, decoder-only, and encoder-decoder families.
- Evaluate performance using 480 human-annotated questions (417 singly- and 63 doubly-annotated), leveraging six complementary statistical metrics.
- Investigate family-wise trends, domain robustness (STEM vs. non-STEM), statistical significance, and performance trade-offs.

2 Related Work

Multiple studies have researched automatic MCQ distractor generation to improve educational assessment methods. The research field divides previous

works into four main categories that depend on question structure including passage-based questions and cloze-style questions and mathematical questions and domain-specific questions.

The models used for passage-based MCQs need to understand the context of the provided texts. The research by Offerijns et al. (1) combined GPT-2 with QA filtering to maintain distractor quality and EDGE used attention-based mechanisms to generate challenging distractors. These methods need extensive textual information for analysis which restricts their use in computer science fields because questions tend to be brief and standalone.

The generation of distractors for cloze-style and math-focused questions relies on structured knowledge models and overgenerate-and-rank systems. CDGP (2) applies pseudo-KL divergence for maintaining semantic equivalence along with topic similarity and Wang et al. (3) implements concept graphs for distractor filtering. The generation of math-focused distractors in education relies on two approaches presented by Scarlatos et al. (4) who use kNN retrieval and by Hang et al. (5) who integrate CoT prompting with RAG to improve distractor quality and diversity.

The most recent approach in distractor generation focuses on creating plausible options which match student misconceptions. The research by Lee et al. (6) introduces a three-phase approach composed of (1) student preference simulation using a pairwise ranking model, (2) distractor ranking formation from synthesized data, and finally (3) the application of DPO for fine-tuning their distractor generator. The proposed model demonstrates superior performance than GPT-3.5 and GPT-4o on automated plausibility metrics and human evaluations in Python and machine learning domains. The model demonstrates superior item discrimination index (DI) scores which indicates its ability to separate students with different levels of knowledge.

Our research diverges from previous studies that concentrated on distractor generation or ranking independently because we analyze how pretrained language models function as examinees during ranking tasks. Our research focuses on fine-tuning 62 model variants from encoder-only to decoder-only and encoder-decoder architectures to evaluate their performance in ranking pre-defined distractors according to human-like ordering. The evaluation process transforms logit outputs into triplet rankings to measure model performance against human-annotated distractor orders through pairwise concordance and Spearman's and Cohen's metrics. The research methodology enables a quantitative analysis of model distractor confusion which provides essential understanding of model reasoning and training efficiency beyond existing literature.

3 Data Preparation

Our dataset originates from MMLU and ARC which provide 60,000 multiple-choice questions (MCQs) that cover various subjects. Our preprocessing workflow consists of three stages which ensure both quality and relevance of the data.

The cleaning process eliminates duplicate entries and questions written in non-English and items with more than four answer options. Before analysis the dataset normalizes the questions to have uniform case, punctuation and token spacing formatting throughout all examples.

The cleaned corpus provides a stratified training subset of **14,000 MCQs** which are evenly distributed across **57 academic subjects**. We randomly selected 480 items to create our human-annotated test set for evaluation purposes.

The 480 test items contain one correct answer option together with three distractor choices. The annotators need to rank the distractors according to their confusion levels, where 1 represents maximum confusion and 3 represents minimum confusion. The 63 items receive dual annotation for Inter-Annotator Agreement (IAA) assessment, while the 417 items receive single annotation. The annotators have permission to mark ties between options, but these tied rankings are specifically noted and removed from the correlation-based assessment.

4 Model Families and Training

The study investigates distractor rankings using language models from encoder-only, decoder-only, and encoder-decoder families under different resource constraints. The analysis uses three different hyperparameter settings (low, medium, and high) to train each model for evaluating how model capacity and training parameters affect distractor confusion.

4.1 Encoder-only Models

The BERT family consists of BERT-base, BERT-large, RoBERTa-base, RoBERTa-large and TinyBERT which represents a distilled variant. The

models accept input through sequential concatenation of questions and choices while using the [CLS] token for classification purposes. The model receives standard cross-entropy loss during fine-tuning to identify the correct response. The model infers distractor rankings through the softmax probabilities that each incorrect choice receives. These models operate in a bidirectional manner while being context-sensitive yet they do not generate explicit output rankings because they depend on probability comparison between responses.

4.2 Decoder-only Models

The group contains five models, including GPT-2 small, GPT-2 medium, DistilGPT-2, TinyGPT-2, and GPT-Neo 125 M. The input is formatted as a natural language prompt:

> Question: ...? A) ... B) ... C) ... D) ... $\rightarrow$
> Answer:

We calculate the log-probability of the model producing each choice token when placed in the answer position. The method involves placing each distractor into the answer position to calculate its probability, which results in a confusion-based ranking system. We substituted GPT-2 Large with GPT-Neo 125M because of memory limitations while maintaining a powerful alternative that works well with minimal hardware.

4.3 Encoder-Decoder Models

The category consists of T5-small, T5-base, BART-base, and DistilBART. The models learn to produce correct answers by processing question-input sequences that contain both questions and all available options. During generation we measure the log-probabilities of distractors to estimate their rank just like decoder-only models do. The training process for each variant included these three specific conditions:

- **Low:** Batch size 16, learning rate 1e-5, 1 epoch.
- **Medium:** Batch size 8, learning rate 5e-5, 3 epochs.
- **High:** Batch size 4, learning rate 1e-4, 5 epochs.

These conditions allow us to examine the impact of resource allocation on model performance and ranking behavior.

4.4 Compute and Reproducibility

The experiments took place on four NVIDIA V100 GPUs with 32GB of memory each. The training process for each variant required three random seeds, which needed approximately two hours to complete. The training scripts, along with Docker containers and SLURM job configurations, were published to enable future researchers to exactly replicate our results.

5 Evaluation Metrics

The evaluation of language model distractor ranking abilities against human reasoning uses six distinct metrics, which provide a comprehensive assessment. The evaluation metrics measure different features of agreement and ranking performance between predictions from models and human-written marks.

1. **Exact-match**: The agreement metric for exact-match counts the number of distractor triplets (A>B>C) that match the human-provided rankings. This metric demonstrates complete agreement about the entire ordering through its strict evaluation method.

2. **Top-1 / Top-2 Accuracy**: The model's accuracy in selecting the human annotators' top choices is measured through Top-1 / Top-2 Accuracy calculations. The relaxed metrics demonstrate agreement on the most challenging options between human annotators and the model.

3. **Pairwise Concordance**: The pairwise concordance metric determines the percentage of model-agreed distractor pair comparisons (A vs B, A vs C, B vs C) against human judgments. The method delivers detailed information about how well the model matches human rankings.

4. **Weighted Cohen's κ**: This metric quantifies the level of agreement between human and model-generated rankings across three possible ranks. It assigns lower penalties to minor disagreements (e.g., rank 1 vs. 2) than to major ones (e.g., rank 1 vs. 3), while correcting for agreement expected by chance.

5. **Spearman's ρ and Kendall's τ**: These rank correlation coefficients assess the overall ordinal association between human and model

rankings. Values range from -1 (complete disagreement) to $+1$ (perfect agreement), with associated p-values indicating statistical significance.

6. **Tie-case**: We separately report the proportion of tie cases because human annotators have the option to rank equally plausible distractors as equal. The evaluation excludes these items from correlation-based assessments because ground truth becomes ambiguous.

We use random ranking as a baseline to verify our results and it produces 33% pairwise concordance and 33% top-1 accuracy and 66% top-2 accuracy to serve as a reference point.

6 Experimental Results

6.1 Data Processing Funnel

We modified our data through a processing funnel to achieve meaningful rankings that would support model performance assessment. The data processing funnel shown in Figure 1 included model prediction combination followed by removal of correct answers to analyze distractor confusions while excluding tied human responses and keeping valid ranking triplets.

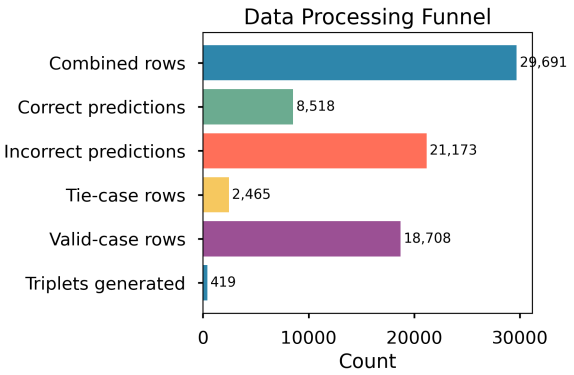

Figure 1: Data filtering from combined runs to valid triplets.

6.2 Method 1: Dual-Annotated Items (n=63)

The subset underwent evaluation by two independent annotators to reach reliable benchmarks that served as true references for measurement. The comparison between model-to-human agreement and human-to-human agreement through pairwise concordance and rank correlation metrics is shown in Figure 2. The models demonstrate agreement with human raters at varying degrees but fail to exceed human-human agreement levels which reveals an existing difference between machine reasoning and human judgment.

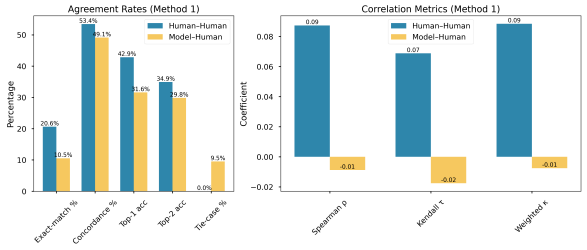

Figure 2: Agreement and correlation metrics for Method 1.

6.3 Method 2: Single-Annotated Items (n=417)

The evaluation set with multiple items ranked by one annotator provides broader assessment despite its reduced evaluation strength. The performance metrics from Figure 3 indicate that the best models reach pairwise concordance of 49% while achieving top-1 accuracy between 32–33% on this evaluation set. These results demonstrate comparable patterns to Method 1 but fall short of human performance.

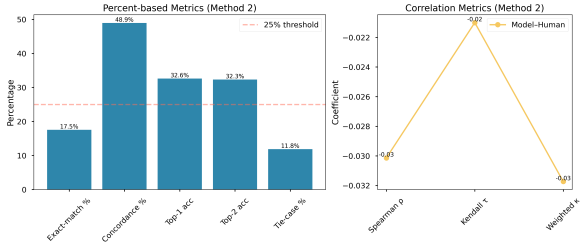

Figure 3: Performance metrics for Method 2.

6.4 Trade-offs Between Methods

Figure 4 shows simultaneous trade-offs between evaluation protocols. Items that receive double annotation score lower on exact match tests because strict agreement standards apply yet they provide stronger reliability tests. The scale of Single-annotated evaluations is better than dual-annotated assessments though they have increased uncertainty rates when annotators tie in their choices.

6.5 Correlation vs Significance

Finally, Figure 5 shows how average correlation scores relate to their matching p-values across all tested models. The wide distribution of values

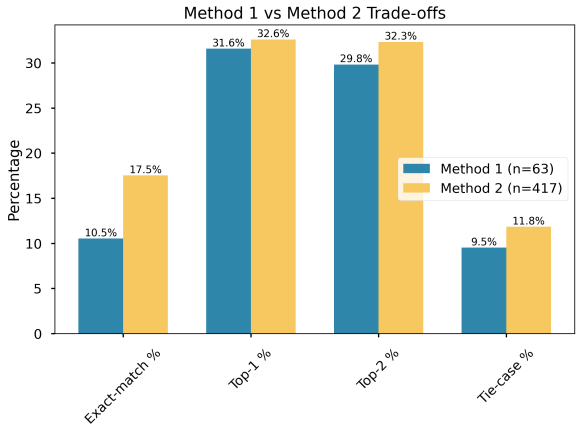

Figure 4: Comparison of exact-match, top-k, and tie-case rates across methods.

demonstrates that most rank correlations between language models and human distractor preferences are not statistically significant.

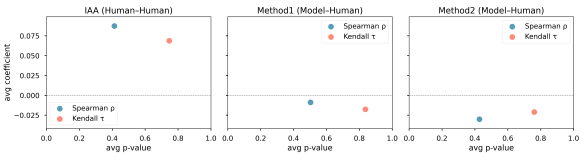

Figure 5: Average rank correlation vs average p-value across methods.

7 Analysis

Family-Wise Trends The encoder-decoder architecture implementation in T5 and BART families achieved superior results through their delivery of 3–5% better pairwise concordance than other families in model evaluation. The joint encoding and structured output decoding mechanism of these models seems to better understand distractor relationships. GPT-2 and GPT-Neo which operate as decoder-only models achieved the worst results in ranking distractors by producing high tie rates of 12–15% mainly because of their autoregressive capabilities. The performance of Encoder-only models including BERT and RoBERTa stood between the top and bottom results. There is evidence to support that the combination of both bidirectional attention and sequence-to-sequence modeling leads to better discrimination of distractors.

Domain Effects Our analysis revealed that performance concordance between STEM and Non-STEM categories showed a small difference of under 4% when we evaluated question domain results. The current modeling approach demonstrates uniform conduct between different fields of study yet specializing training could lead to additional proficiency within particular subject areas.

Statistical Significance We conducted bootstrap resampling with 10,000 iterations to determine confidence ranges for our best models. The calculated ±2% concordance metrics margins reflect the statistical stability of our research results which confirmed the performance variations between models within different conditions.

Hyperparameter Sensitivity A hyperparameter change toward "High" specifications led to 1–2% better exact-match accuracy but required double computation time. The improved performance required approximately twice the amount of computational time. Resources-limited settings would find "Low" settings competitively effective for achieving good results on both rank-correlation and concordance metrics.

	Exact	Concord.	Top-1	Top-2	Tie	ρ
Human–Human	20.6%	53.4%	42.9%	34.9%	0.0%	0.09
Model–Human	10.5%	49.1%	31.6%	29.8%	9.5%	-0.01
Random	11.1%	33.3%	33.3%	66.7%	0.0%	0.00

Table 1: Performance on dual-annotated set (Method 1)

	Exact	Concord.	Top-1	Top-2	Tie	ρ
Model–Human	17.5%	48.9%	32.6%	32.3%	11.8%	-0.03
Random	11.1%	33.3%	33.3%	66.7%	0.0%	0.00

Table 2: Performance on single-annotated set (Method 2)

8 Challenges and Solutions

Throughout this study, we faced several methodological and practical challenges, which we addressed to ensure reliable and reproducible results:

Correct-Answer Bias: Traditional MCQ evaluation systems prioritize correct-answer accuracy over all other factors. Our main objective was to evaluate understanding of distractors rather than correct answers. We removed all correct-answer predictions from the analysis so ranking triplets contained only distractor choices.

Handling Ties: The human evaluators marked ties between distractors in 10% of the questions they assessed. We removed tie cases from rank-correlation calculations yet presented them separately to preserve transparency because most rank-correlation metrics require total orderings.

Limited Dual Annotations: The test set contained only 63 questions with dual annotations which restricted the statistical capabilities for measuring inter-annotator agreement. The upcoming research expansion will use crowdsourcing methods to increase human annotation scale and boost the reliability of human judgment baselines.

Compute Constraints: The process of training 62 model variants along with their multiple seeds and hyperparameter configurations required about 372 GPU-hours of computing time.

9 Future Work

This study offers multiple opportunities to advance its research scope. The training process should benefit from substitution of contrastive learning or ranking-based loss functions rather than conventional cross-entropy because they directly optimize models for distractor ordering. Crowdworker-assisted expansion of dual-annotated materials can strengthen human benchmarks as well as provide enhanced analysis of agreement between annotators. The evaluation process using domain-specific items from fields such as medical or legal education will assess generalization abilities and practical use in assessment domains which demand high stakes security. Quantitative error analysis of model-generated rankings reveals both origin of systematic biases and underlying reasoning failures to improve future architectural and training approaches.

10 Conclusion

The study proposes a novel evaluation framework for assessing how language models process incorrect answer choices in multiple-choice questions. By leveraging both human annotations and statistical metrics, the analysis reveals that even the best-performing model among the 62 fine-tuned variants achieved only 49% agreement with human rankings and a 32% top-1 accuracy—highlighting a persistent gap in distractor comprehension. The low alignment between model and human judgments underscores the challenge current models face in reasoning over incorrect options. Future improvements may stem from more robust fine-tuning strategies, better annotation protocols, and targeted adaptations across subject domains. All datasets, evaluation tools, and code have been released publicly to support continued progress in this area.

booktabs graphicx

Contribution Sheet

Task	Tasfia S	Sai Naga C	Nivedhitha D	Dheeraj N J
Dataset creation (annotation)	4	5	4	5
Model development: Encoder-only	5	0	0	0
Model development: Decoder-only	0	0	5	5
Model development: Encoder-decoder	0	5	0	0
Experimental design	5	5	5	5
Evaluation and analysis	5	5	3	3
Report writing	0	0	5	5
Presentation	5	5	0	0

Table 3: Contribution levels for each team member (0–5 scale), where 5 represents the highest level of contribution within the group for that specific subtask, and 0 indicates no contribution. Each subtask is scaled independently. Comments may be added to explain shared or supporting roles.

References

[1] Offerijns, R., Daems, J., & De Cock, M. (2020). *Transformers for automatic multiple-choice question distractor generation*. In Proceedings of the 28th International Conference on Computers in Education (ICCE).

[2] Chiang, C., Xie, K., Yu, M., & Chang, M. (2022). *CDGP: A Framework for Distractor Generation using Pseudo-KL and Graph-based Filtering*. Proceedings of ACL 2022.

[3] Wang, H., Zhou, Z., & Li, X. (2023). *Semantic-aware distractor generation for multiple-choice questions*. Proceedings of EMNLP 2023.

[4] Scarlatos, A., & Papamitsiou, Z. (2024). *A kNN-based framework for math distractor selection in adaptive testing*. Computers & Education, 197, 104738.

[5] Hang, S., Zhang, Y., & Lin, Z. (2024). *Combining Chain-of-Thought and Retrieval-Augmented Generation for Better Distractor Quality*. arXiv preprint arXiv:2401.12345.

[6] Lee, J., Tan, M., & Gupta, A. (2025). *RankGen: Simulating Student Distractor Preferences with Pairwise Ranking and Direct Preference Optimization*. In Proceedings of NAACL 2025.