Generating multiple choice questions from a textbook: LLMs match human performance on most metrics

Andrew M. Olney[1]

[1]*Institute for Intelligent Systems, University of Memphis, 365 Innovation Drive, Suite 303, Memphis, TN 38152, USA*

Abstract

Multiple choice questions are traditionally expensive to produce. Recent advances in large language models (LLMs) have led to fine-tuned LLMs that generate questions competitive with human-authored questions. However, the relative capabilities of ChatGPT-family models have not yet been established for this task. We present a carefully-controlled human evaluation of three conditions: a fine-tuned, augmented version of Macaw, instruction-tuned Bing Chat with zero-shot prompting, and human-authored questions from a college science textbook. Our results indicate that on six of seven measures tested, both LLM's performance was not significantly different from human performance. Analysis of LLM errors further suggests that Macaw and Bing Chat have different failure modes for this task: Macaw tends to repeat answer options whereas Bing Chat tends to not include the specified answer in the answer options. For Macaw, removing error items from analysis results in performance on par with humans for all metrics; for Bing Chat, removing error items improves performance but does not reach human-level performance.

Keywords

multiple choice, question generation, LLM, Bing Chat, Macaw

1. Introduction

Multiple choice questions are widely used in education. In contrast to free response questions, multiple choice questions are scored objectively and at great speed, simply by checking the selected answer code, e.g. A-D, against an answer key. However, the ease of grading multiple choice questions (MCQs) comes at a nontrivial cost of creating them, with the greatest effort spent on creating distractor answer options [1, 2].

Automatic generation of MCQs has received increasing research interest over the past two decades. Early approaches had little training data and so approached the MCQ generation task as four subtasks in a pipeline architecture: sentence selection, answer selection from selected sentences, question generation using the sentence and answer, and distractor generation [3]. Each of these subtasks can be addressed using NLP approaches developed outside the MCQ literature, like summarization techniques for sentence selection, keyword extraction techniques for answer selection, general question generation [4], and semantic similarity approaches for

AIEDLLM1: Empowering Education with LLMs, July 7, 2023, Tokyo, Japan
aolney@memphis.edu (A. M. Olney)
https://olney.ai/ (A. M. Olney)
0000-0003-4204-6667 (A. M. Olney)

distractor generation. This general approach sidesteps the need for MCQ training data by leveraging data from other tasks, but as a result inherits biases from those datasets, e.g. a news-trained summarizer will not select important sentences in a science text [5].

More recently, deep learning approaches have been applied to MCQ generation, but older neural techniques tend to predominate. For example, one simple approach to generating distractors is to use an encoder-only model like BERT to predict masked tokens. This approach has been used to create single-word distractors for vocabulary MCQs [6] as well as multiword distractors when applied sequentially [7]. Additional work has used bidirectional LSTMs to generate distractors for reading comprehension MCQs by conditioning on inputs like text, question, and answer [8, 9, 10] in order to produce distractors that are more relevant.

In contrast, a small amount of work has used large language models (LLMs) for MCQ tasks, often in conjunction with fine-tuning. The encoder-decoder T5 model [11] has been used to generate distractors, either by using its pretraining objective to fill a span of masked tokens [12], similar to the BERT work above, or by fine-tuning T5 to generate distractors conditioned on text and question [13]. The encoder-only GPT-2 model has also been used to generate distractors conditioned on text and question [14], analogous to the LSTM work above.

Limited MCQ training data has been problematic for LLM approaches that use fine-tuning. The recently proposed Macaw model [15], based on T5, leverages diverse datasets by representing various question answering/generation tasks as angles. Each angle consists of slots like A (answer), Q (question), M (multiple choice options), and C (context) as well as a mapping from input to output slots. Macaw's angle-based approach allows a large number of question-oriented datasets to be used as training data by representing them as angles and training on them all simultaneously. For example, a question-answering dataset can be used to both answer questions ($Q \rightarrow A$) and generate questions from answers ($A \rightarrow Q$), and more complex datasets for reading comprehension can be used to answer a question based on text ($CQM \rightarrow A$) as well as reverse mappings that generate MCQ elements. By training on a large number of datasets and angles, Macaw promises more general and robust performance for a variety of question-related tasks.

An evaluation of Macaw compared Macaw-generated MCQs to textbook MCQs in a human evaluation [16]. That study found that Macaw's performance did not improve across three angles that systematically reduced the complexity of the task, $C \rightarrow QMA$, $AC \rightarrow QM$, and $QAC \rightarrow M$, but rather performed best with $AC \rightarrow QM$. The most common error was an inability to generate four distinct answer options, which was addressed by paraphrasing C up to 10 times and running $AC \rightarrow QM$ on each paraphrase to generate more diverse M. This method improved the generation success of $AC \rightarrow QM$ from 83% to 97.5%. Finally, the human evaluation, which measured question meaningfulness and fluency, answer correctness and presence in the options, distractor distinctness and non-overlap with the correct answer, and overall quality, found that Macaw MCQs were rated significantly lower than textbook MCQs on 5 of 7 metrics, but that Macaw was rated highly in absolute terms, e.g. 85% compared to textbook 94% on overall quality. One limitation of this evaluation is that while the Macaw questions and textbook questions were on the same general topic, they were not precisely aligned. Thus it is possible that some of the differences in ratings are due to differences in concepts being covered rather than the questions themselves, e.g. perceived difficulty by the human raters.

In contrast to developments in LLM fine-tuning, recent work has illustrated how LLMs can generalize to new tasks without traditional input/output training data. This approach

was popularized by GPT-3, which was shown to successfully complete various tasks without explicit training as long as it was given task instructions [17] (so-called zero-shot learning), with performance typically improving if additional demonstrations were provided (so-called few-shot learning). Remarkably, GPT-3 was able to exceed fine-tuned state-of-the-art performance on some benchmark tasks using this approach. Succeeding work has investigated instruction-tuning, which fine-tunes an LLM to follow instructions written in natural language for various tasks, and has found improvements over GPT-3 on benchmark tasks using models with fewer parameters [18, 19], but of course, even greater gains have been found using instruction tuning on the largest LLMs (>100B parameters) [20, 21]. The success of instruction-tuned models has led to commercialization successes like ChatGPT and Microsoft's Bing Chat, which present an instruction-following LLM that can follow directions across conversational turns (so-called chatbot interaction). While these recent instruction-tuned LLMs are impressive in casual interactions, it is not clear how well they can generate MCQs compared to fine-tuned LLMs.

The present paper attempts to clarify the state of the art in MCQ generation by conducting a carefully-controlled human evaluation of three conditions: the fine-tuned augmented version of Macaw [16], instruction-tuned Bing Chat with zero-shot prompting, and human-authored questions from a college anatomy and physiology textbook [22]. In order to control for concept coverages, conditions are content-aligned, i.e. AI conditions generate MCQs based on the same input. Our primary research questions are (1) how well do the MCQs produced by the LLMs compare to textbook MCQs in a human evaluation study, (2) what errors do the LLMs make in the task, and (3) can LLM errors explain the human performance gap.

2. Human evaluation study

A human evaluation study was conducted to compare the fine-tuned augmented version of Macaw (Macaw+) [16], instruction-tuned Bing Chat with zero-shot prompting (Bing Chat), and human-authored questions from a college anatomy and physiology textbook (Textbook).

2.1. Design

The evaluation study used a within-subjects design with Macaw+, Bing Chat, and Textbook as conditions. Conditions were presented using a 6 x 3 balanced Latin square (i.e., $3! = 3 \times 2 \times 1$) to counterbalance condition order and prevent carryover effects between conditions. However, the underlying context of each MCQ (i.e., the input source sentence and correct answer for the LLMs) was not counterbalanced. This design decision means that in a fully-used Latin square, a context in a particular location would be paired with each condition, making fatigue effects equivalent across conditions. The human ratings were analyzed using mixed-effects beta regression with random intercepts for rater and rating question using the `glmmTMB` R package [23]. Beta regression is appropriate for continuous bounded outcome variables, unlike linear regression, which isn't suitable for bounded outcomes, and logistic regression, which can be used for proportions, but only when the proportion is a ratio of two counts [24]. Because beta regression is defined on the open interval (0,1), we use a standard transformation to squeeze our closed interval outcome variables to the open interval [25]. We conducted statistical tests

Write a multiple choice question using the following sentence and answer. Convert the sentence into a question that matches the answer. Use JSON format.
Sentence: <sentence>
Answer: <answer>

Figure 1: Prompt used to generate MCQs using Bing Chat. Tokens marked by < > were replaced with their respective input strings.

at $\alpha = .05$ to address our research questions. If beta regression revealed a significant effect of condition, post hoc tests were conducted to determine differences between conditions.

2.2. Participants

Raters ($N = 16$) were recruited through the Amazon Mechanical Turk (AMT) marketplace from April to May of 2023 using the CloudResearch platform [26]. Raters were recruited using an occupation screener that paid 1 cent for their reporting of occupation. Raters were required to reside in the U.S., Canada, New Zealand, United Kingdom, Ireland, or Australia, and be employed as a nurse ($n = 9$), doctor ($n = 5$), or allied health provider with anatomy and physiology expertise ($n = 2$). The educational and occupational constraints we designed to ensure raters were experts in the evaluation subject domain: they had passed anatomy and physiology in their studies and used this knowledge on a daily basis. Demographic constraints are enforced by CloudResearch based on rater responses to previous demographic surveys. Raters were further required to have completed at least 100 previous AMT tasks with at least a 95% approval rating. Raters were paid $12 regardless of reliability, based on an estimated 120 minutes to complete the task. In addition, raters were paid up to $50 in bonuses for passing quality checks determined by intra-rater reliability: a $5 bonus for passing each check, and an additional $20 bonus for passing a comprehensive check.

2.3. Materials

A textbook on anatomy and physiology [22] from OpenStax was used as the source of 120 textbook questions. Questions were web scraped from the OpenStax website[1], manually checked, and aligned with the answer key accessible by registering as an instructor. The 120 questions for this evaluation were drawn from the first 4-5 questions from each of the textbook's 28 chapters.

MCQs for the three conditions were produced as follows. The Textbook condition used the MCQ as it appeared in the textbook. The LLM conditions both required sentence and answer as input, so the textbook questions were manually transformed into sentences, and these sentences and original answers were used as inputs to the LLM models. For example, the textbook question "Which of the following specialties might focus on studying all of the structures of the ankle and foot?" with associated answer "regional anatomy" was transformed into the sentence "Regional anatomy might, for example, focus on studying all of the structures of the ankle and foot." Each sentence/answer was input to the fine-tuned, augmented Macaw model described in [16] to create MCQs for the Macaw+ condition (see Section 1 for key details).

[1] https://openstax.org/details/books/anatomy-and-physiology

Because Bing Chat uses prompt-based input and there is no known best prompt for generating MCQs, several different prompts were designed based on templates from existing datasets [18, 19] and informally evaluated using a handful of the above sentence/answer pairs. The best-performing prompt given in Figure 1 was used to generate all Bing Chat MCQs using the EdgeGPT API [27]. Note the prompt includes the same sentence/answer used in the Macaw+ condition. Thus all three conditions are aligned on each MCQ context.

Six surveys were created with Qualtrics, an online survey tool, using a balanced Latin square to define the order of conditions. Because each row of the Latin square only contains 3 orderings, each ordering was repeated 20 times in a survey for a total of 60 MCQs. The same ordering of 60 contexts was used in each survey; only the condition applied to each position of the ordering varied across surveys. Each question, correct answer, and answer options were formatted vertically in that order on a single survey page using the direct assessment methodology [28, 29]. These three elements each had two associated ratings, followed by an overall quality rating, for a total of seven ratings per question, as shown in Table 1. All ratings were in horizontal slider format and arranged in descending order. The 0-100 sliders had no numeric indicators and were initialized at the midpoint. The remaining sliders had numeric indicators and snapped to integer positions. Each survey had instructions at the beginning to explain the task and the seven ratings. Survey instructions and an example survey page are shown in Appendix A.

Following the direct assessment methodology, degraded items were created to evaluate the internal reliability of each rater [29, 28, 30]. Degraded items were created by copying the question, answer, and options on an existing survey page and then applying the following transformations. Questions were degraded by deleting a span of words [28], where the length of the span was determined by the equation $span_{length} = 0.21696 * word_{count} + 0.78698$ [31]. Degraded answers were created by replacing the correct answer with one of the other answer options selected at random. Degraded answer options were created by randomly selecting a remaining incorrect answer option and then duplicating it while removing another incorrect option at random. Thus each survey of 80 pages contained 60 distinct pages and 20 degraded versions of distinct pages. An example degraded item is shown in Appendix A.

We refer to a distinct page and its degraded version as a control pair. A sample size of 20 control pairs is sufficient to detect a large (.8 SD) effect using a Wilcoxon signed-ranks test for matched pairs at $\alpha = .05$ and .95 power on a one-tailed test. Thus if we do not detect a

Table 1
Ratings used in human evaluation study

Measure	Scale
The question contains correct information	0-100
The question is grammatical and fluent	0-100
The given correct answer is correct	0-100
The given correct answer is present in the answer options	0-100
Number of answer options that give a correct answer	0-4
Number of answer options that are distinct (no duplicates)	1-4
Quality of the question, given answer, and answer options combined	0-100

large effect between ratings of distinct pages and their degraded versions, we infer the rater is not reliable (i.e., poor intra-rater reliability). The degraded pages were in randomly assigned positions in each survey and were evenly distanced from their matched distinct pages, modulo 50. This ensured that pages in control pairs had 50 other items between them, making it less likely that raters would remember their rating on a previous item.

We additionally developed an occupation survey to help us find more qualified raters. The occupation survey consisted of two questions, a generic occupation question from the standard Qualtrics demographics library with 20 answer options, and a conditional branch question that only appeared if a respondent selected `healthcare` on the first question. The conditional branch question asked for a more specific healthcare occupation, with nine total options including six matching our recruiting criteria. This indirect approach to asking about specific healthcare occupations was designed to avoid demand characteristics (i.e., false responses) from asking such questions directly.

2.4. Procedure

Six surveys were sequentially with a default quota of four raters. If a survey had sufficiently good reliability with less than four raters, it was terminated early; this only happened with survey 3. Likewise, if a survey had insufficient reliability with four raters, it was extended with a quota of an additional four raters; this only happened for survey 2. Raters were allowed to participate in more than one survey if they passed the comprehensive quality check.

Raters accessed the surveys through AMT and completed the surveys using Qualtrics. Because the study is a system evaluation and not human subjects research, informed consent was not obtained. Raters saw the instructions for the survey twice, once as a preview on AMT before undertaking the survey, and again once they clicked on the survey link. On each following page, raters read the question, the correct answer, and the answer options, and then completed the ratings described in Table 1. Raters were paid upon completion of the survey and received bonuses based on the quality checks passed, i.e. based on their intra-rater reliability for each rating, with the final rating in Table 1 serving as the comprehensive check.

2.5. Results and Discussion

Median completion time across surveys was 133 minutes, giving approximately 100 seconds to read the question, answer, and options and make 7 judgments. Control checks were considered to be passed if $p < .05$ on the aforementioned Wilcoxon signed-ranks test. Seven raters failed to pass the comprehensive check on a given survey and were excluded from future surveys. Every check on each survey was passed by at least two raters with the exception of survey 2, where all but `question informative` and `question fluent` were passed by two raters.

Initially, the Cronbach's alpha inter-rater reliability statistic was calculated for raters passing control checks in each survey. However, inspection of the raw ratings revealed that while some raters used 0-100 sliders as a measure of confidence (as intended), other raters used it in a binary fashion, leading to nonlinearity and lack of variability among raters. Because Cronbach's alpha was invalid for these data, a new agreement metric was constructed based on the contrast between distinct and degraded items. If we consider that the distinct items are likely good and

Table 2
Inter-rater reliability per survey for included raters.

Survey	Question informative		Question fluent		Answer correct		Answer in options		Correct options		Distinct options		Combined quality	
	κ	n	κ	n	κ	n	κ	n	κ	n	κ	n	κ	n
1	.83	2	.85	2	.73	2	.92	3	.80	3	.93	3	.77	3
2	.65	2	.70	2	.23	2	.78	3	.65	2	.80	2	.65	2
3	.68	2	.75	2	.73	2	.98	2	.55	2	.88	2	.80	2
4	.70	2	.75	2	.65	2	.87	3	.75	3	.78	3	.78	2
5	.80	2	.83	2	.70	2	.76	4	.68	3	.81	4	.65	3
6	.68	2	.78	2	.68	2	.92	3	.55	3	.88	3	.70	3

so should have ratings above 50 on most scales and that degraded items are likely bad and so should have ratings below 50 on the same scales, then we can threshold all ratings to be either 1 (good) or 0 (bad) and calculate balanced accuracy for true positives and true negatives. The same approach works for `correct options` if we assume that there is 1 correct option by default (good) and for `distinct options` if we assume that there are 4 distinct options by default (good) and that degraded items have all other values on these metrics. We will refer to these assumptions as pseudo-truth because we are assuming that the generated MCQs are generally good and that their degraded variants are generally bad.

Inter-rater reliability was calculated for each rating within a survey using the following method. First, all ratings were converted to 1/0 as described above, and their balanced accuracy based on pseudo-truth was calculated. The top two most accurate ratings were kept for inter-rater reliability (regardless of absolute accuracy) and any additional ratings with accuracy greater than .8 were also included. Inter-rater reliability was then calculated between included ratings using Fleiss's kappa, adjusted for unbalanced classes [32]. The net effect of this approach is that comparing to pseudo-truth was a stricter criterion of intra-rater reliability than using the Wilcoxon signed-ranks test (e.g. 5 raters passed the Wilcoxon signed-ranks test on survey 1 for `question informative`, but only 2 raters passed the pseudo-truth procedure), and using only these raters ensures high inter-rater reliability without sacrificing validity. Intra-rater reliabilities are shown in Table 2 in the same order as Table 1 but using abbreviated labels. Final kappas showed substantial agreement ($\kappa > .60$) on 39 of 42 ratings. Ratings shown in Table 2 were used in all further analyses.

To answer our research question of how MCQs produced by the LLMs compare to textbook MCQs, we ran separate mixed-effects beta regressions with random intercepts for rater and question, using the source of the question as the fixed effect (Bing Chat, Macaw+, or Textbook). Descriptive statistics and regression results are shown in Table 3, with associated p values from one-way ANOVA. Significant differences between conditions were found only for `answer in options` and `combined quality`. However, pairwise contrasts for `answer in options` revealed no significant differences between Bing Chat ($M = .83$, $SE = .02$) and Textbook ($M = .86$, $SE = .02$), $p = .068$, and no significant differences between Macaw+ ($M = .86$, $SE = .02$) and Textbook, $p = .083$. Pairwise contrasts for `combined quality`, however,

Table 3
Descriptive statistics and significance of mixed-effects beta regressions comparing Bing Chat, Macaw+, and Textbook conditions.

Rating	Bing Chat		Macaw		Textbook		p
	M	SD	M	SD	M	SD	
Question informative	93.49	15.05	91.45	19.90	93.95	14.56	.543
Question fluent	95.72	13.81	94.73	17.05	97.83	9.28	.494
Answer correct	83.18	32.31	86.32	28.88	91.93	21.73	.134
Answer in options	88.71	28.27	96.19	13.51	96.09	13.79	.043
Correct options	1.15	.74	1.12	.59	1.04	.41	.355
Distinct options	3.93	.33	3.77	.65	3.98	.17	.103
Combined quality	85.52	24.17	86.93	24.21	93.62	15.80	.006

revealed significant differences between Bing Chat ($M = .85$, $SE = .02$) and Textbook ($M = .88$, $SE = .02$), $p = .009$, and significant differences between Macaw+ ($M = .85$, $SE = .02$) and Textbook, $p = .038$. Altogether, the LLM conditions were not significantly different from Textbook on six of seven measures. However, `answer in options` and relatively low p-values of `answer correct` and `distinct options` warrant further investigation.

3. Error analysis

To answer our research question of what errors the LLMs make in MCQ generation, we conducted an error analysis for `distinct options` and `answer in options`. MCQs generated by Macaw+ and Bing Chat were automatically scored using exact string match to determine if the given answer appeared in the answer options and if the four answer options were distinct from each other. Errors detected by exact string match were then manually reviewed to determine if they were actual errors. For example, a failed string match where the only difference was punctuation or an article like "a" would not be considered an actual error.

Macaw+ had 111 questions with distinct options or 92.5%. This is notably lower than the previously reported success rate of 97.5% [16]. Six of the nine failures occurred when response options contained lists, e.g. "carbon, hydrogen, oxygen, and nitrogen" which may explain the difference in the previous result if such lists are resistant to paraphrasing. Macaw+ had 120 questions where the given correct answer was one of the answer options, i.e. 100%. So the primary failure mode of Macaw+ in the evaluation was a failure to generate distinct answer options in 7.5% of cases, with the majority of these caused by answer options that are lists.

Bing Chat had 120 questions with distinct options, i.e. 100%. Bing Chat had 109 questions where the given correct answer was one of the answer options or 91%. Of the remaining 11 mismatches, 8 could be considered to be valid questions overall, in the sense that one of the answer options was the correct answer, but that answer differed from the given correct answer in a nontrivial way. For example, 2 failures used a letter (A–D) to indicate the correct answer rather than using the given correct answer in the prompt, and another 3 failures used either a wider or narrower scoping of the answer than was presented in the options, e.g. answer option

"Axons from the retinal ganglion cells in the retina" is a narrower scoping of given correct answer "retinal ganglion cells." If such errors are judged leniently, then 8 of 11 of the errors can be viewed as a failure to precisely follow the prompt's instructions. So the primary failure mode of Bing Chat in the evaluation was a failure to include the given correct answer among the answer options in 9% of cases, with the majority of these cases being otherwise valid questions. Example errors for Macaw+ and BingChat are shown in Appendix B.

4. Explaining the human performance gap

The error analysis in Section 3 potentially explains the pattern of results in the human evaluation in Section 2.5. Recall `answer in options` was significant with Bing Chat having the lowest mean score, tracking Bing Chat's failure to include the given correct answer among the answer options in 9% of cases. The same error could potentially explain the relatively low p-values of `answer correct`, since the mismatch between the given correct answer and the answer options could lower the confidence of human raters that the given answer is correct. Similarly for Macaw+, `distinct options` had a relatively low p-value and Macaw+ had the lowest mean score, which tracks Macaw+'s failure to generate distinct answer options in 7.5% of cases.

To answer our research question of how LLM errors explain the human performance gap, we reanalyzed the human evaluation data to determine the effect of the above errors on `combined quality`, which was the only measure for which pairwise significant differences were found between Bing Chat, Macaw+, and Textbook. Two analyses were conducted for Bing Chat and Macaw+. First, we used the rating of their primary error type to predict their `combined quality`. Second, we tested the difference between LLM and Textbook `combined quality` when MCQs marked as bad were excluded. For example, if a rater marked a Macaw+ MCQ as having 4 distinct options (good), then their `combined quality` rating for that MCQ would be included in the analysis, otherwise it would be excluded. Both analyses used separate mixed-effects beta regressions with random intercepts for rater and question.

Mixed-effects beta regression for Bing Chat ratings using `answer in options` to predict `combined quality` revealed a significant effect, $p < .001$. When `answer in options` is 0, estimated `combined quality` is low ($M = .39$, $SE = .07$), and when `answer in options` is 100, estimated `combined quality` is high ($M = .89$, $SE = .02$). An additional mixed-effects beta regression was conducted comparing Bing Chat to Textbook using only MCQs where `answer in options` was scored highly (above 50). This filtering procedure removed 34 ratings from the Bing Chat condition and 5 ratings from the Textbook condition out of 360 total ratings. Mixed-effects beta regression revealed that with these errors removed, `combined quality` for Bing Chat ($M = .92$, $SE = .02$) was still rated lower than Textbook ($M = .93$, $SE = .01$), $p < .012$. A follow-up simulation analysis on the choice of 50 as a threshold revealed that the significant difference between conditions remained up to a threshold of 98. These results suggest that while `answer in options` errors strongly influence `combined quality` for Bing Chat, they do not fully explain the human performance gap.

Mixed-effects beta regression for Macaw+ ratings using `distinct options` to predict `combined quality` revealed a significant effect, $p < .001$. When `distinct options` is not 4 (bad), estimated `combined quality` is low ($M = .59$, $SE = .06$), and when `distinct`

options is 4 (good), estimated combined quality is high ($M = .86$, $SE = .02$). An additional mixed-effects beta regression was conducted comparing Macaw+ to Textbook using only MCQs where distinct options was scored highly (equal to 4). This filtering procedure removed 45 ratings from the Macaw+ condition and 7 ratings from the Textbook condition out of 360 total ratings. Mixed-effects beta regression revealed no significant difference between Macaw+ ($M = .91$, $SE = .02$) and Textbook ($M = .92$, $SE = .01$), $p = .149$. These results suggest that distinct options errors may explain the human performance gap for Macaw+.

5. Discussion

The goal of the present work was to clarify the state of the art in MCQ generation by comparing two LLMs, the fine-tuned augmented version of Macaw [16] and instruction-tuned Bing Chat with zero-shot prompting, to human-authored questions in a carefully-controlled human evaluation. Our results indicate that on six of seven measures tested, both LLM's performance was not significantly different from human performance. These six measures relate to individual components of the MCQ, specifically the question stem, the answer, and the answer options, and are very fine-grained, so the lack of significant difference is particularly notable. Only on the overall measure of combined quality was a significant difference found in favor of the human-authored questions.

Analysis of LLM errors indicates that Macaw and Bing Chat have different failure modes for this task: Macaw tends to repeat answer options whereas Bing Chat tends to not include the specified answer in the answer options. Each of these error types is strongly predictive of combined quality ratings. For Macaw, removing error items from analysis results in combined quality ratings that are not significantly different from human-authored questions on combined quality. For Bing Chat, removing error items improves combined quality, but resulting ratings remain significantly different from human-authored questions. Altogether, these results suggest that the LLMs are both remarkably capable of creating MCQs, and the error analyses suggest future research directions for each LLM on this task.

These results are based on a high-quality human evaluation. It is widely agreed that human evaluations provide the best evidence of system performance, yet as few as 20% of research papers on natural language generation include them [33]. We were careful to recruit raters whose profession required them to be highly knowledgeable in the MCQ content domain. Our evaluation includes both intra-rater reliability (can raters distinguish between actual items and intentionally degraded items) as well as inter-rater reliability (do raters agree with each other). Only raters with high intra- and inter-rater reliability were included in our analysis. Additionally, the evaluation was designed to minimize confounding effects of fatigue as well as individual rater characteristics like extreme responses.

Our study has two primary limitations. First, we only evaluated questions on the topic of anatomy and physiology. It is possible that the LLMs would perform differently on other topics, though neither model was trained specifically for this topic. Second, the task given to the LLMs simplifies the canonical task of generating MCQs from text by providing sentences and answers instead of selecting them [3]. Therefore, our results should not be taken as representative of MCQ generation from freeform text.

Acknowledgments

This material is based upon work supported by the Institute of Education Sciences under Grant R305A190448 and by the National Science Foundation under Grants 1918751 and 1934745.

References

[1] S. M. Downing, Selected-response item formats in test development, in: T. M. Haladyna, S. M. Downing (Eds.), Handbook of Test Development, Routledge, New Jersey, 2006, pp. 287–301.

[2] M. J. Gierl, O. Bulut, Q. Guo, X. Zhang, Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review, Review of Educational Research 87 (2017) 1082–1116. URL: https://doi.org/10.3102/0034654317726529. doi:10.3102/0034654317726529. arXiv:https://doi.org/10.3102/0034654317726529.

[3] D. C. Rao, S. K. Saha, Automatic multiple choice question generation from text: A survey, IEEE Transactions on Learning Technologies 13 (2020) 14–25. doi:10.1109/TLT.2018.2889100.

[4] V. Rus, B. Wyse, P. Piwek, M. Lintean, S. Stoyanchev, C. Moldovan, Overview of the first question generation shared task evaluation challenge, in: K. E. Boyer, P. Piwek (Eds.), Proceedings of QG2010: The Third Workshop on Question Generation, questiongeneration.org, Pittsburgh, 2010, pp. 45–57. URL: http://oro.open.ac.uk/22343/.

[5] A. M. Olney, Sentence selection for cloze item creation: A standardized task and preliminary results, in: T. W. Price, S. San Pedro (Eds.), Joint Proceedings of the Workshops at the 14th International Conference on Educational Data Mining, volume 3051 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. LDI–6.

[6] L. Gao, K. Gimpel, A. Jensson, Distractor analysis and selection for multiple-choice cloze questions for second-language learners, in: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 102–114. URL: https://aclanthology.org/2020.bea-1.10. doi:10.18653/v1/2020.bea-1.10.

[7] H.-L. Chung, Y.-H. Chan, Y.-C. Fan, A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies., in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4390–4400. URL: https://aclanthology.org/2020.findings-emnlp.393. doi:10.18653/v1/2020.findings-emnlp.393.

[8] Y. Gao, L. Bing, P. Li, I. King, M. R. Lyu, Generating distractors for reading comprehension questions from real examinations, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI Press, 2019, pp. 6423–6430. URL: https://doi.org/10.1609/aaai.v33i01.33016423. doi:10.1609/aaai.v33i01.33016423.

[9] Z. Qiu, X. Wu, W. Fan, Automatic distractor generation for multiple choice questions in standard tests, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain, 2020,

pp. 2096–2106. URL: https://aclanthology.org/2020.coling-main.189. doi:10.18653/v1/2020.coling-main.189.

[10] X. Zhou, S. Luo, Y. Wu, Co-attention hierarchical network: Generating coherent long distractors for reading comprehension, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI Press, 2020, pp. 9725–9732. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6522.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[12] K. Vachev, M. Hardalov, G. Karadzhov, G. Georgiev, I. Koychev, P. Nakov, Leaf: Multiple-choice question generation, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 321–328.

[13] R. Rodriguez-Torrealba, E. Garcia-Lopez, A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, Expert Systems with Applications 208 (2022) 118258. URL: https://www.sciencedirect.com/science/article/pii/S0957417422014014. doi:https://doi.org/10.1016/j.eswa.2022.118258.

[14] J. Offerijns, S. Verberne, T. Verhoef, Better distractions: Transformer-based distractor generation and multiple choice question filtering, 2020. arXiv:2010.09598.

[15] O. Tafjord, P. Clark, General-purpose question-answering with Macaw, 2021. URL: https://arxiv.org/abs/2109.02593. doi:10.48550/ARXIV.2109.02593.

[16] A. M. Olney, Generating multiple choice questions with a multi-angle question answering model, in: S. E. Fancsali, V. Rus (Eds.), Proceedings of the 3rd Workshop of the Learner Data Institute, 2022, pp. 18–23. doi:10.5281/zenodo.7761561.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, in: International Conference on Learning Representations, 2022, pp. 1–216. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

[19] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, X. Shen, Super-NaturalInstructions:

Generalization via declarative instructions on 1600+ NLP tasks, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5085–5109. URL: https://aclanthology.org/2022.emnlp-main.340.

[20] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[21] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2022, pp. 1–46. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[22] J. G. Betts, P. Desaix, E. Johnson, J. E. Johnson, O. Korol, D. Kruse, B. Poe, J. A. Wise, M. Womble, K. A. Young, Anatomy and Physiology, OpenStax, 2017.

[23] M. E. Brooks, K. Kristensen, K. J. Van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Machler, B. M. Bolker, glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling, The R Journal 9 (2017) 378–400.

[24] R. Kieschnick, B. D. McCullough, Regression analysis of variates observed on (0, 1): percentages, proportions and fractions, Statistical Modelling 3 (2003) 193–213. URL: https://doi.org/10.1191/1471082X03st053oa. doi:10.1191/1471082X03st053oa. arXiv:https://doi.org/10.1191/1471082X03st053oa.

[25] M. Smithson, J. Verkuilen, A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables., Psychological Methods 11 (2006) 54–71.

[26] L. Litman, J. Robinson, T. Abberbock, TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences, Behavior Research Methods 49 (2017) 433–442. doi:10.3758/s13428-016-0727-z.

[27] A. Cheong, Edge GPT, 2023. URL: https://github.com/acheong08/EdgeGPT, original-date: 2023-02-09T16:07:42Z.

[28] Y. Graham, T. Baldwin, A. Moffat, J. Zobel, Is machine translation getting better over time?, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 443–451. doi:10.3115/v1/E14-1047.

[29] C. Federmann, O. Elachqar, C. Quirk, Multilingual whispers: Generating paraphrases with translation, in: Proceedings of the 5th Workshop on Noisy User-Generated Text, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 17–26. doi:10.18653/v1/D19-5503.

[30] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, P. Koehn, C. Monz, Findings of the 2018 Conference on Machine Translation (WMT18), in: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 272–303. doi:10.18653/v1/W18-6401.

[31] A. M. Olney, Generating response-specific elaborated feedback using long-form neural question answering, in: Proceedings of the Eighth ACM Conference on Learning @ Scale, L@S '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 27–36. URL: https://doi.org/10.1145/3430895.3460131. doi:10.1145/3430895.3460131.

[32] D. Marasini, P. Quatto, E. Ripamonti, Assessing the inter-rater agreement for ordinal data through weighted indexes, Statistical Methods in Medical Research 25 (2016) 2611–2633. URL: https://doi.org/10.1177/0962280214529560. doi:10.1177/0962280214529560. arXiv:https://doi.org/10.1177/0962280214529560, pMID: 24740999.

[33] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 355–368. doi:10.18653/v1/W19-8643.

A. Rating task

Each multiple choice question was presented to human raters using Qualtrics. Instructions at the beginning of the survey are shown in Figure 2. Each survey page had one MCQ with associated ratings as shown in Figure 3. Sliders were required to move from default positions to advance to the next page. A degraded item is shown in Figure 4.

Instructions

We would like you to rate 80 multiple choice questions for quality.

On each screen, you will be presented with a **question**, followed by a **correct answer** and **answer options**.

Please read these and make judgements about their quality:

Question

- Does it contain correct information?
- Is it grammatical and fluent?

By *correct information*, we mean that the question is answerable, has no false claims, and does not contradict itself. For example "What brand of car do elephants drive" has low correct information because elephants don't drive cars. Another bad example is "What breaks?" because it's not specifically answerable.

By *fluent*, we mean natural English. For example, "How many years do you have" may be grammatical, but it is not as fluent as, "How old are you." So even if the question is grammatical, we also want you to consider how fluent it is.

Correct answer

- Is it correct?
- Is it in the answer options?

You may find the correct answer is completely wrong or only partly right. If the question is not answerable, you should mark the given answer as wrong. You may find the given answer is identical to an answer option, similar to an answer option, or not in the options at all. Ideally, the given answer is correct and is present in the answer options.

Answer options

- Number that give a correct answer
- Distinct (no duplicates)

You may find multiple answer options are correct. You may also find that answer options are duplicates; for example "John F Kennedy", "JFK", "Kennedy", "Obama" has only two distinct options. Ideally, there are 4 distinct options and only one of them is correct.

Overall quality
Finally, we'd like you to give an overall rating to the multiple choice question's quality.

BONUSES
The survey has built-in quality checks. You will be awarded a $5 bonus for passing each technical check and a $20 bonus for passing the overall check -- **up to a $50 bonus combined**.

Please take your time and look things up if you aren't sure.

You may prefer to use a tablet to complete these ratings. A phone is possible but may be cramped.

That's it, please start when ready!

Figure 2: Survey instructions.

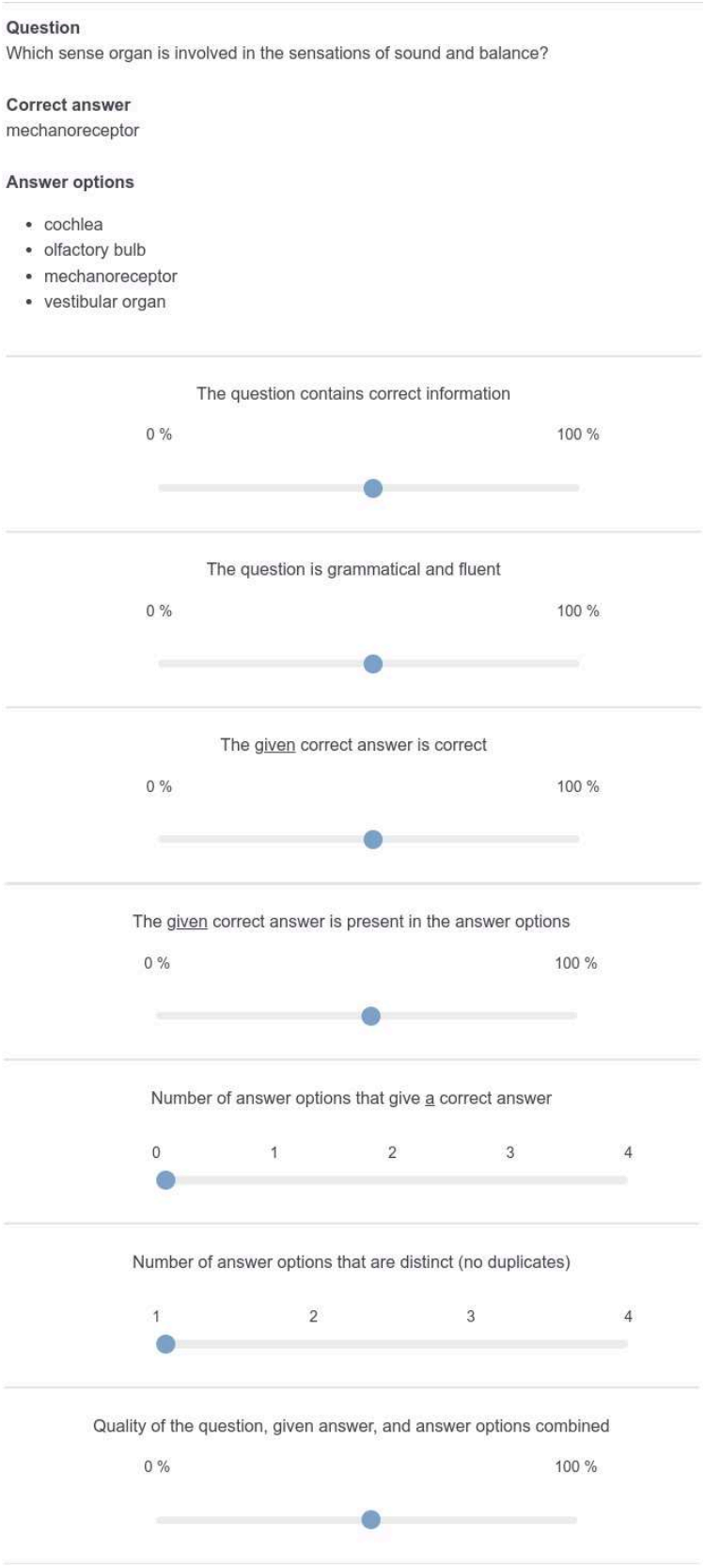

Figure 3: Survey page illustrating the rating task.

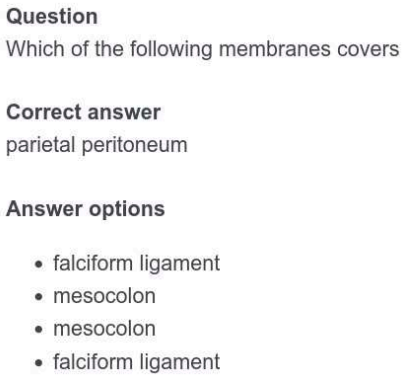

Figure 4: A degraded item. Rating options for degraded items are identical to Figure 3.

B. Error analysis

The LLMs differ in the types of errors they make during MCQ generation. The primary error mode of Macaw+ is to fail to make enough distinct options as distractors, especially when the distractors are lists, as shown in Figure 5. The primary error of Bing Chat is failing to follow the instructions of the prompt precisely. As shown in Figure 6, Bing Chat has generated a valid question if the given correct answer had been “fauces,” but the given correct answer was “oral cavity.”

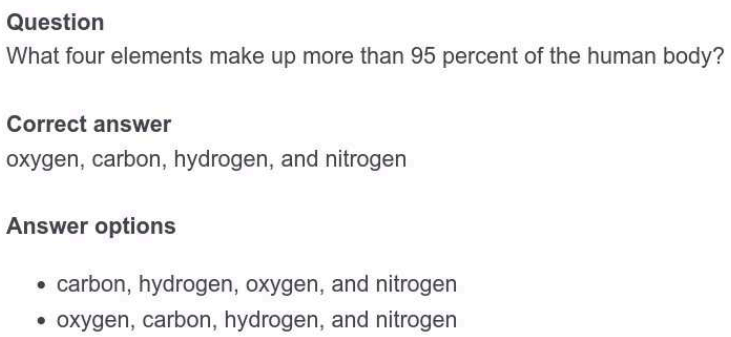

Figure 5: A typical Macaw+ generation error where insufficient distractors have been generated.

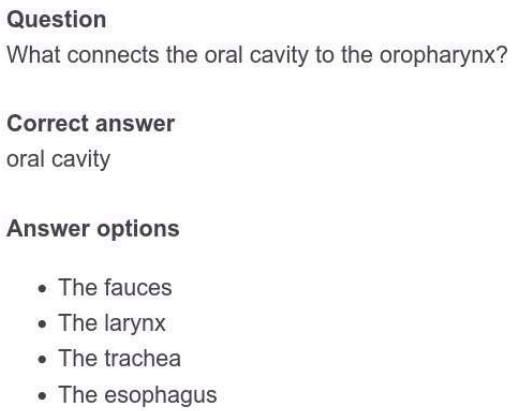

Figure 6: A typical Bing Chat generation error where the prompt was not followed precisely. The input sentence was “The fauces connects the oral cavity to the oropharynx." and the input answer was “oral cavity.” The respective prompt using these elements is given in Figure 1.