Evaluating Large Language Model Robustness using Combinatorial Testing

Jaganmohan Chandrasekaran*, Ankita Ramjibhai Patel‡, Erin Lanus†, Laura J. Freeman†,
* Sanghani Center for Artificial Intelligence & Data Analytics, Virginia Tech, Arlington, VA, USA
† National Security Institute, Virginia Tech, Arlington, VA, USA
‡ Independent Researcher, USA

***Abstract*—Recent advancements in large language models (LLMs) have demonstrated remarkable proficiency in understanding and generating human-like text, leading to widespread adoption across domains. Given LLM's versatile capabilities, current evaluation practices assess LLMs across a wide variety of tasks, including answer generation, sentiment analysis, text completion, and question and answers, to name a few. Multiple choice questions (MCQ) have emerged as a widely used evaluation task to assess LLM's understanding and reasoning across various subject areas. However, studies from the literature have revealed that LLMs exhibit sensitivity to the ordering of options in MCQ tasks, with performance variations based on option sequence, thus underscoring the robustness concerns in LLM performance.**

This work presents a combinatorial testing-based framework for systematic and comprehensive robustness assessment of pre-trained LLMs. By leveraging the sequence covering array, the framework constructs test sets by systematically swapping the order of options, which are then used in ascertaining the robustness of LLMs. We performed an experimental evaluation using the Measuring Massive Multitask Language Understanding (MMLU) dataset, a widely used MCQ dataset and evaluated the robustness of GPT 3.5 Turbo, a pre-trained LLM. Results suggest the framework can effectively identify numerous robustness issues with a relatively minimal number of tests.

***Index Terms*—Testing AI; Combinatorial Testing; Testing LLM; LLM Robustness; LLM Evaluation; Option Order Swapping;**

I. Introduction

Large Language Model (LLM), a type of Artificial Intelligence (AI), is primarily developed to accomplish natural language processing tasks. LLMs have emerged as a transformative technology making significant strides and exhibiting remarkable capabilities in understanding, reasoning, and generating human-like text. Successful adoptation of LLMs across domains depends on the ability to comprehensively test and evaluate (T&E) LLMs, thereby guaranteeing their performance. Given LLM's versatile capabilities, the current practice involves evaluating its capabilities across subject areas through various tasks [1]–[3]. Among such tasks, multiple choice questions (MCQ)-based tasks have emerged as a prominent evaluation method for assessing the LLM's understanding and reasoning abilities [2]. In the MCQ task, the LLM under evaluation is provided with a set of questions, with each question consisting of a set of typically four options, and the LLM's response option to each question is compared against the correct option called the "ground truth." From the comparison, the overall model accuracy is calculated, and LLMs achieving a higher score are considered to exhibit better performance, and thus considered to have better understanding and reasoning abilities. MCQ tests are typically drawn from benchmark datasets such as the Measuring Massive Multitask Language Understanding (MMLU) benchmark [4], a widely used MCQ benchmark dataset for evaluating LLMs. For example, Gemini, a state-of-the-art widely popular LLM, highlights that their model achieved an accuracy score of 90.0% in MMLU benchmark demonstrating the widespread use of MCQ-based tasks in LLM evaluation [5]. The MCQ-based tasks have gained prominence, and MCQ is one of the widely used tasks used among the LLM evaluation community [1]–[3], [5]–[8].

While these MCQ-based evaluations offer benchmarking capabilities, they suffer from significant limitations. Firstly, the inherent probability of an LLM choosing/predicting correct answers through random guessing. For example, in the case of an MCQ test set with four options, there is a 1 in 4 chance (25%) for an LLM's response matching with the ground truth through random guessing. Consequently, while an LLM might achieve a high score on an MCQ test set, it does not necessarily indicate better performance. The possibility of an LLM achieving a higher score through random chance necessitates further evaluation to robustly assess the LLM's understanding and reasoning capabilities. Secondly, an LLM can exhibit memorization behaviors, learning to recognize patterns in the question or answer set rather than truly understanding the underlying concepts.

Furthermore, recent studies from the literature have demonstrated that LLM performance in MCQ tasks is sensitive to the order of the answer options [9]–[14]. That is, upon changing the order of options in MCQ, the performance of an LLM tends to fluctuate, and in some cases, results in a significant drop in prediction accuracy [9]. These observations highlight the potential vulnerabilities in MCQ-based evaluation methods and underscore the need for robustness assessments of LLM

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense, Director, Operational Test and Evaluation (DOTE) through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(RE)) under Contract HQ0034-24-D-0023. The Systems Engineering Research Center (SERC) is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense nor DOTE nor ASD(RE).

Corresponding author: jagan@vt.edu

979-8-3315-3467-7/25 © 2025 IEEE

performance in MCQ tasks that cover all possible scenarios.

A comprehensive assessment necessitates testing all possible orderings ($n!$ for n options), resulting in an exhaustive test set. While exhaustive testing might enable a thorough assessment, the computational complexity makes this approach impractical for large-scale evaluations. For example, a question with four options will require testing 24 permutations (4!), and an MCQ dataset with 100 questions will require executing 2400 test cases. To address this challenge, this paper presents a combinatorial testing-based approach to evaluate the robustness of LLMs in MCQ tasks. Combinatorial testing, a pseudo-exhaustive test generation strategy, has proven effective in testing traditional software systems by achieving comprehensive coverage with a relatively minimal number of test cases [15]–[17]. Specifically, we investigate the applicability of sequence covering arrays in constructing test sets to perform robustness assessments of LLMs in MCQ tasks.

Recall that, for the robustness assessment of LLMs, the objective is to test all possible order combinations ($n!$ for n options). By employing sequence-covering arrays, we can systematically generate tests to evaluate the impact of combinations of options and, more importantly, the order in which the options are related. Furthermore, given n options, generating tests using a sequence covering array guarantees all n options will be tested in every t-possible order [18], [19]. Consider a question with four options $\{$A, B, C, and D$\}$ and correct answer C. Upon providing the question and answer sequence ABCD, assume that the LLM correctly predicts C. Robustness testing approaches that rely on random swapping primarily focus on testing by varying the position of the predicted option and the impact on the LLM's performance (e.g., CABD or ACBD or ABDC). They are most likely not to prioritize evaluating how the ordering of non-predicted options impacts the LLM's performance. On the contrary, using sequence covering arrays to generate test cases guarantees that the test scenarios will be a combination of both – testing the impact of the ordering of predicted options as well as the ordering of non-prediction options, including permutations like BACD and DACB, thereby enabling a relatively comprehensive robustness assessment.

Given an MCQ test set, the approach presented in this paper generates additional K question-option sets for each question-option from the benchmark test set based on the 3-sequence covering table from the sequence covering array library [19]. For an MCQ with four options, there are 4! = 24 possible orderings of options. However, by leveraging a 3-sequence covering array, our test generation approach generates six additional questions, a 75% reduction in test set size compared to all permutations. We present an experimental evaluation of our approach. We use GPT 3.5 Turbo [20], a pre-trained LLM, as our subject model. Eight datasets were selected at random from the MMLU benchmark.

To evaluate the robustness of GPT 3.5 Turbo, we implement a systematic assessment framework. For each question (referred to as the base question) from the dataset, using our approach, we generate six additional questions (referred to as variants). The LLM's response to both base and its variants is recorded and analyzed. We conducted the robustness assessment at the question level: the LLM is considered to exhibit robust behavior if the LLM's response remains consistent across the base and all its variants (regardless of correctness of response). Conversely, if any variant resulted in an outcome different from the LLM's response to the base question, the LLM failed the robustness assessment. Finally, we quantify the overall robustness by calculating the proportion of questions that pass the robustness assessment across the dataset.

Our results indicate that the variants generated using our approach can detect several robustness vulnerabilities in the LLM. In several cases, for 50% or more variants, the LLM produced a different response from the base question, highlighting severe concerns about the overall robustness of the LLM.

This paper makes the following contributions:

- A CT-based test generation method for evaluating the robustness of LLMs in MCQ tasks by systematically swapping the order of answer options.
- A robustness assessment framework for a granular question-level analysis of LLMs performance in MCQ tasks.
- A preliminary comparative study between CT-based option-swapping and exhaustive test sets for detecting LLM robustness vulnerabilities.

The remainder of this paper is organized as follows. Section II presents background. Section III describes our approach. Section IV presents our experiments, including the design of our experiments, followed by the results and discussion. Section V discuss related work. Finally, in Section VI, we discuss our concluding remarks and directions for future work.

II. Background

A. Testing Large Language Models

AI practitioners leverage deep learning techniques to build an LLM. These models are trained on large text datasets, analyze the underlying relationship from their massive training datasets, and develop capabilities to perform various tasks with minimal or no human intervention. Recent advancements in LLMs have demonstrated their potential to understand the input, reason based on the provided input, and generate output. This ability has enabled LLM to perform numerous tasks such as text generation, sentiment analysis, summarization, information retrieval, text completion, translation, and named entity recognition, to name a few. Furthermore, users interact with LLMs through natural language, similar to human communication.

Given its versatile capabilities, testing of LLMs is broadly focused on evaluating its understanding, reasoning, and generation capabilities. To facilitate a systematic and comprehensive assessment, current practices in testing LLM aim to evaluate its understanding, reasoning abilities, and generation capabilities. The LLM evaluation community accomplishes this goal by testing holistically across various tasks and datasets.

Testing LLMs presents unique considerations. Firstly, the mode of access to the LLM can influence the testing process, including the associated testing costs. Secondly, the design and formulation of inputs to interact with the LLMs, commonly referred to as prompts, influence the LLM's overall behavior. The mode of access and prompts are briefly discussed next.

- Access mode: Pre-trained LLMs are distributed using two common access modes: (1) Pre-trained LLMs that can be downloaded and executed locally. (2). Pre-trained LLMs accessed using an Application Programming Interface (API). Executing an LLM on a local machine requires significant computational resources. Accessing LLMs via an API requires strong internet connectivity and involves usage-based costs, but does not require significant computational resources at the user end.
- Prompt: Users interact with LLM through natural language prompts. A prompt, provided as input to an LLM, is a set of instructions on actions to be performed and desired behavior that guides the LLM's response generation.

B. Combinatorial Testing

Combinatorial testing (CT) is a black-box test generation technique that focuses on systematically testing the interactions among a system's input parameters. In a system with five input parameters, each with three possible values, testing all possible input value combinations would require $3^5 = 243$ tests. CT, on the other hand, constructs tests that cover multiple input value combinations per test, thereby resulting in a significantly reduced test suite while maintaining comparable fault detection capabilities to exhaustive testing. For example, a 2-way combinatorial test suite can effectively test all pairwise combinations of five inputs (2-way parameter interactions), using only 15 test cases, a 94% reduction compared to the exhaustive test set, while retaining the fault detection effectiveness.

Sequence covering array (SCA) is a type of combinatorial test design that focuses on testing the order of events. Unlike traditional combinatorial test design, which primarily examines the effect of interactions between input parameters simultaneously, SCA-based test design guarantees that all t-way permutations will be covered by the test suite. In the context of systematic order-swapping, this unique characteristic of SCA – "any t events will be tested in every possible order" [18], [19] – can be leveraged to systematically construct additional variants (or tests) from a given MCQ dataset. For example, given an MCQ question with four options, all possible 3-way orderings can be tested using the SCA presented in Table I.

III. Approach

In this section, we introduce a combinatorial testing-based approach to evaluate the robustness of LLMs in MCQ tasks. Our approach consists of three steps: First, we generate additional questions (variants) from each question (base question) in the MCQ dataset using a 3-way sequence covering array. Next, in step 2, both the base question and its variants are provided as input to the LLM, and the LLM responses are recorded. Finally, in step 3, we perform a question-level robustness assessment of LLMs in MCQ tasks. To illustrate our approach, we employ a running example.

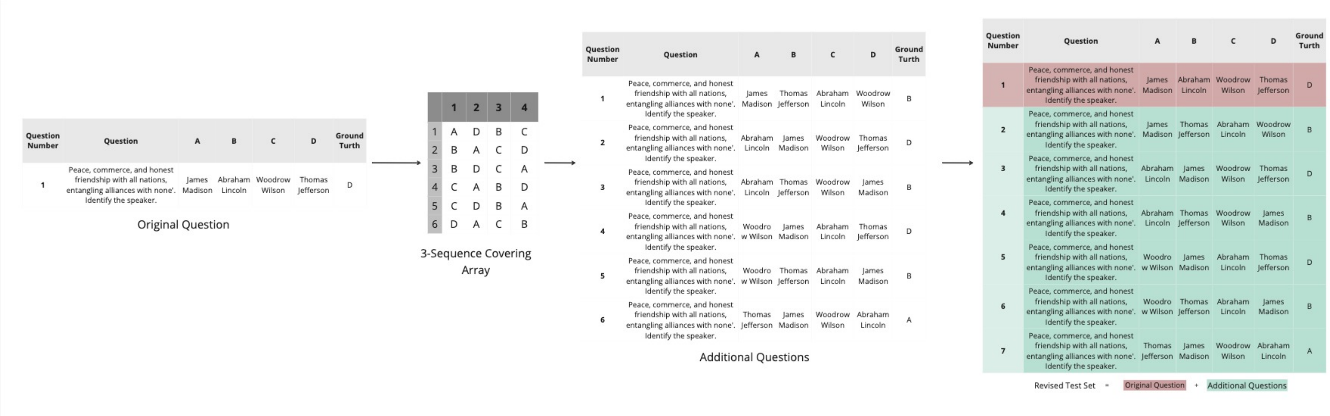
TABLE I: 3-Sequence covering array for testing four events [19]

	1	2	3	4
1	A	D	B	C
2	B	A	C	D
3	B	D	C	A
4	C	A	B	D
5	C	D	B	A
6	D	A	C	B

In step 1, the goal is to generate variants from each base question by systematically swapping the order of options. Each base question from the MCQ dataset is assumed to have its set of options and ground truth. We begin the variant generation process by selecting an individual question from the MCQ test set, after which variants are generated using the 3-way sequence covering array table. Consider the example presented in Figure 1. The original question consists of 4-options and its corresponding ground truth, which serves as our base question. Using the 3-Sequence Covering Array table, six additional questions (variants) are generated by systematically swapping the order of options. Furthermore, when options are reordered, the ground truth is adjusted accordingly. For example, in the case of variant 1, the ground truth is adjusted from D (original question) to B (reordered). This process is repeated to all the questions from the MCQ dataset, generating the robustness assessment test dataset. Overall, given an MCQ test set with N questions, our approach using a 3-way sequence covering table for questions with four options generates six variants per question, yielding a total of $N + (6N)$ questions. In step 2, we execute the test cases and record the LLM's response. We classify each test outcome as either *passing test* (LLM prediction matches ground truth) or *failing test* (prediction differs from the ground truth).

Next, in step 3, we analyze the LLM's response and perform a question-level robustness assessment. As shown in Figure 2, an LLM demonstrates robust behavior when its responses remain consistent across the base question and all its variants. Conversely, if any variant resulted in an outcome different from the LLM's response to the base question, the LLM failed the robustness assessment. That is, inconsistency between base and variant responses indicates a robustness failure.

Figures 3 and 4 illustrate the robust behavior for the question presented in the example. In both scenarios, the LLM's response to the original question and all its variants remains constant. In Figure 3, the LLM's response to the original question (Q1) and all its variants (Q2-Q7) remains consistent and matches the ground truth (pass). Conversely, Figure 4 shows consistent incorrect responses (fail) across the base and all its variants (incorrect responses indicated in red font). Both scenarios indicate robust behavior, as our approach evaluates response consistency across the base question and its



Question Number	Question	A	B	C	D	Ground Turth
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D

Original Question

	1	2	3	4
1	A	D	B	C
2	B	A	C	D
3	B	D	C	A
4	C	A	B	D
5	C	D	B	A
6	D	A	C	B

3-Sequence Covering Array

Question Number	Question	A	B	C	D	Ground Turth
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A

Additional Questions

Question Number	Question	A	B	C	D	Ground Turth
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A

Revised Test Set = Original Question + Additional Questions

Fig. 1: Example - Test Generation using CT-based Approach

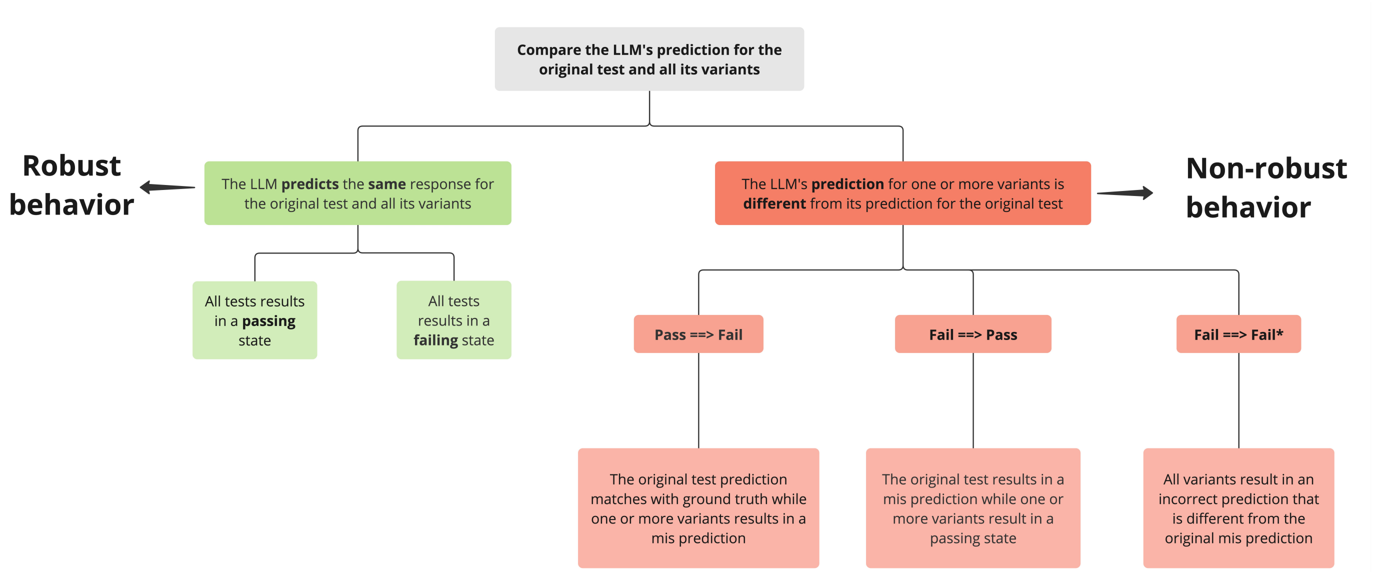

Compare the LLM's prediction for the original test and all its variants
Robust behavior
The LLM predicts the same response for the original test and all its variants
All tests results in a passing state
All tests results in a failing state
Non-robust behavior
The LLM's prediction for one or more variants is different from its prediction for the original test
Pass ==> Fail
Fail ==> Pass
Fail ==> Fail*
The original test prediction matches with ground truth while one or more variants results in a mis prediction
The original test results in a mis prediction while one or more variants result in a passing state
All variants result in an incorrect prediction that is different from the original mis prediction

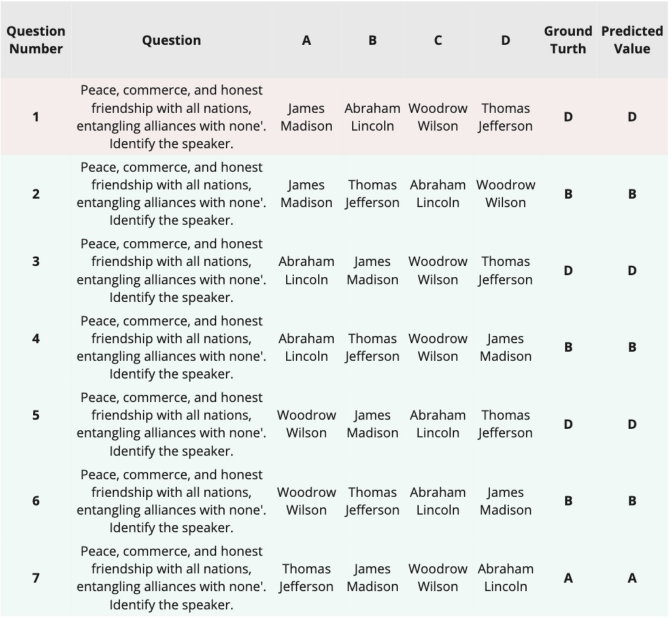
Fig. 2: Overview of the Robustness Assessment Framework

variants independent of accuracy.

In scenarios where the LLM produces a different outcome for one or more variants, then the LLM is considered to fail the robustness assessment tests. As illustrated in Figure 2, non-robust behavior manifests in three distinct patterns: (1) Correct base prediction with one or more variant mispredictions, (2) LLM incorrectly predicted for the original question (fail). However, one or more variants' prediction matches their respective ground truth (pass), and (3) incorrect predictions throughout (both the original questions and all variants), but with inconsistent error patterns. That is, the LLM's response to one or more variants results in an incorrect response that differs from the LLM's incorrect response to the base question.

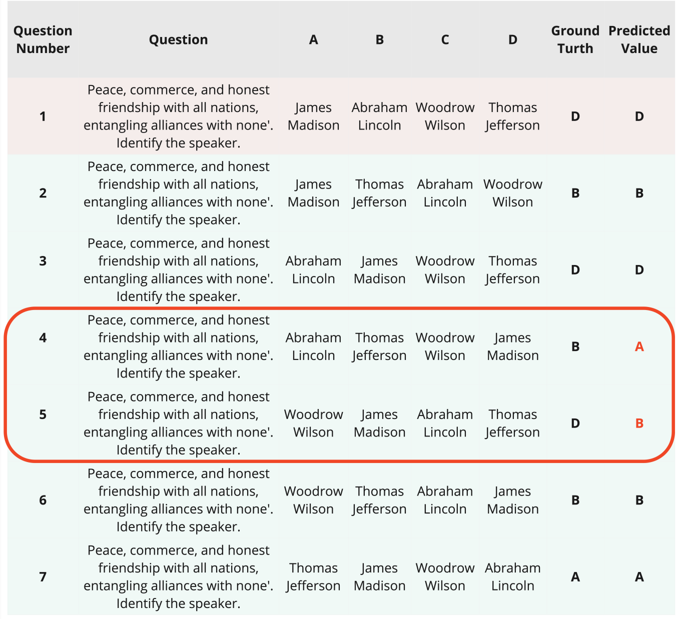
In the event of deviation of LLMs response between the original question and its variants, three scenarios are possible:

- Scenario 1 - the LLMs correctly predicted the original question (pass) but resulted in a misprediction for one or more variants (fail). In Figure 5, the base question and all four out of six variants resulted in a correct prediction. However, for Questions 4 and 5, the LLM predicted an incorrect response.
- Scenario 2 - LLM incorrectly predicted for the original question (fail). However, one or more variants' prediction matches their respective ground truth (pass). For the example in Figure 6, the LLM predicted the same incorrect response for both the base and five out of six variants. However, for Question 3, the LLM predicted a response that matches the ground truth, thus failing the robustness assessment.
- Scenario 3, where the original questions and all variants result in incorrect prediction (fail). However, one or more variants exist whose incorrect response differs from the LLM's incorrect response to the original question. Consider Questions 5 and 6 in Figure 7. The LLM predicted incorrect responses for both questions, similar to the rest. However, the LLM responses to questions 5



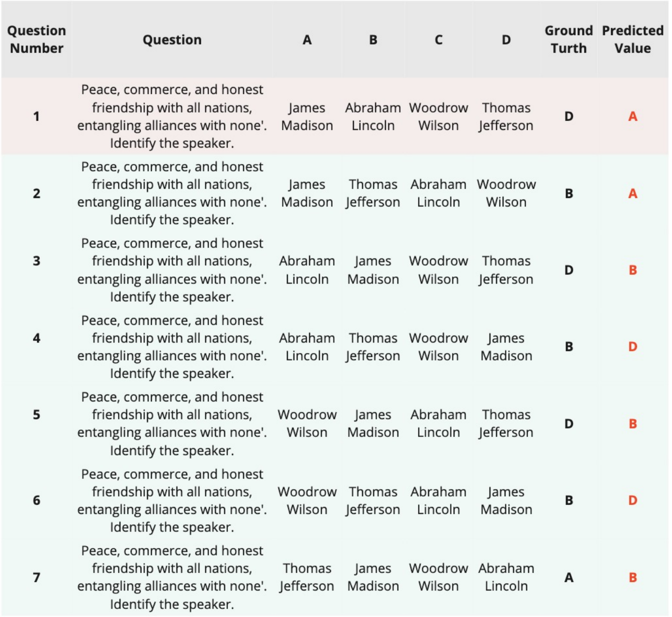
Question Number	Question	A	B	C	D	Ground Turth	Predicted Value
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D	D
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B	B
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D	D
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B	B
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D	D
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B	B
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A	A

Fig. 3: Robust Behavior - Example 1



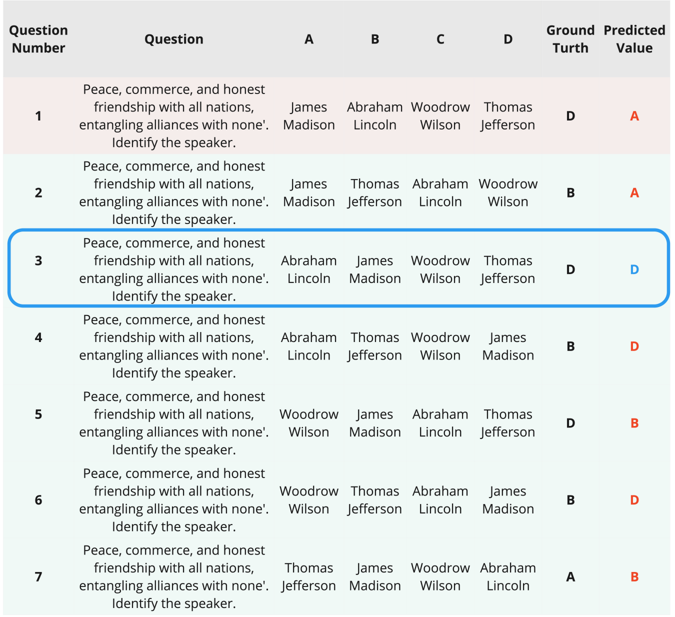
Question Number	Question	A	B	C	D	Ground Turth	Predicted Value
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D	D
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B	B
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D	D
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B	A
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D	B
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B	B
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A	A

Fig. 5: Non Robust Behavior - Scenario 1



Question Number	Question	A	B	C	D	Ground Turth	Predicted Value
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D	A
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B	A
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D	B
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B	D
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D	B
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B	D
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A	B

Fig. 4: Robust Behavior - Example 2



Question Number	Question	A	B	C	D	Ground Turth	Predicted Value
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D	A
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	James Madison	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B	A
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	James Madison	Woodrow Wilson	Thomas Jefferson	D	D
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	James Madison	B	D
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	James Madison	Abraham Lincoln	Thomas Jefferson	D	B
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	Abraham Lincoln	James Madison	B	D
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	James Madison	Woodrow Wilson	Abraham Lincoln	A	B

Fig. 6: Non Robust Behavior - Scenario 2

(Woodrow Wilson) and 6 (Abraham Lincoln) are different from the rest (James Madison).

The aforementioned assessment is performed for all the questions from the MCQ dataset to determine the robustness of the LLM in MCQ tasks.

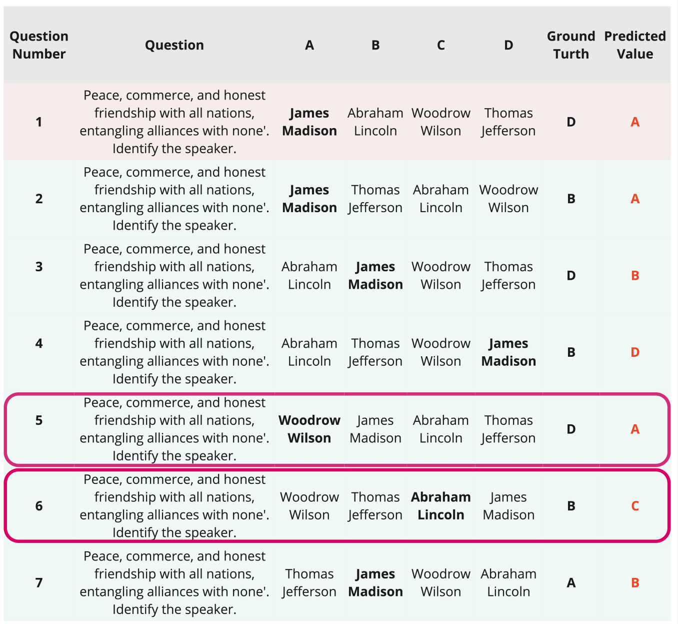
IV. Experiments

In this section, we first present the design of our experiments, including the research question, datasets, the LLM, steps in generating the tests, and the metrics used to measure the robustness of LLM. Next, we present and discuss the results of our experiments, followed by a discussion on threats to validity.

A. *Research Question*

Our experiments are designed to answer the following research question:

- How effectively can combinatorial testing-based approach with systematic option reordering identify response inconsistencies in LLMs when evaluated on MCQ tasks?



Question Number	Question	A	B	C	D	Ground Turth	Predicted Value
1	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	**James Madison**	Abraham Lincoln	Woodrow Wilson	Thomas Jefferson	D	A
2	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	**James Madison**	Thomas Jefferson	Abraham Lincoln	Woodrow Wilson	B	A
3	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	**James Madison**	Woodrow Wilson	Thomas Jefferson	D	B
4	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Abraham Lincoln	Thomas Jefferson	Woodrow Wilson	**James Madison**	B	D
5	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	**Woodrow Wilson**	James Madison	Abraham Lincoln	Thomas Jefferson	D	A
6	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Woodrow Wilson	Thomas Jefferson	**Abraham Lincoln**	James Madison	B	C
7	Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.	Thomas Jefferson	**James Madison**	Woodrow Wilson	Abraham Lincoln	A	B

Fig. 7: Non Robust Behavior - Scenario 3

B. Dataset

Measuring Massive Multitask Language Understanding (MMLU) is a widely used benchmark dataset for LLM evaluation [4]. It consists of MCQ test sets across 57 subjects in four categories namely Humanities, Other, Social Science and Science, Technology, Engineeering and Mathematics (STEM). The MMLU test set consists a total of 14070 multiple-choice questions, with a minimum of 100 questions per subject. Each question has four options to choose from, as well as the ground truth. For our experiments, we selected eight subjects by randomly selecting two subjects per category. Information about the datasets is presented in Table II.

TABLE II: Datasets Used in Experiments

Number	Category	Subject Area	Number of Questions
1	2*Humanities	Philosophy	311
2		Prehistory	324
3	2*Other	Business Ethics	100
4		Marketing	234
5	2*Social Science	Geography	198
6		US Foreign Policy	100
7	2*STEM	College Computer Science	100
8		Algebra	100

C. Model

Access mode: Hosting and executing LLMs on a local machine requires significant computational resources; therefore, we opted for an API-based access mode, which does not have similar computational requirements. We use OpenAI's GPT 3.5 Turbo, a widely used API-based pre-trained LLM, as our subject model [20]. We utilized GPT 3.5 Turbo without any fine-tuning or instruction-tuning modification, allowing us to evaluate the LLM's inherent robustness without the influence of additional training or optimization.

Prompt: Given that our experiments are performed on an MCQ dataset, we designed our prompt to explicitly instruct the LLM to return only its response (A, B, C, or D) and avoid any additional information in its output. Furthermore, we set the token size (the number of words a model will return as output) to 1 to ensure the LLM returns a one-word response, as we are expecting only the model's response to the multiple-choice question. The prompt used in our experiment is as follows: *"User will ask a question to you and provide four options. Please respond with the letter corresponding to your answer (e.g., A, B, C, or D)."*

D. Test generation and execution

For each subject from Table II, we generated a robustness assessment test set by generating variants for each MCQ using the 3-sequence covering array table, provided in [19]. Given that all eight datasets consist of MCQ with four options, we choose the sequence covering array to test all 3-sequences and it is presented in Table I. Following the steps described in the approach section (Section III), we generated eight test sets, one per subject. We used a Python script to generate the test set, execute the tests, interface with the LLM, and record the LLM's responses for analysis.

In some instances, the LLM generated outputs that did not correspond to the expected options (A, B, C, or D). When this occurred for the base question, both the base question and its associated variants were excluded from our analysis. This exclusion is necessary because in this scenario it is impossible to determine if the LLM's response to one or more variants has differed from its response to the base question. In other words, without a baseline, evaluating the LLM's behavior across the variants was not feasible. However, if the LLM predicts an acceptable value for the base question but predicts an out-of-bounds value for one or more variants, this scenario is considered a mismatch or deviation.

E. Metrics

The effectiveness of our approach in performing robustness assessments of LLM in MCQ tasks is measured in terms of the number of prediction inconsistencies (non-robust behavior) the CT-based test set can uncover. As illustrated in Figure 2, our robustness assessment framework operates at two levels: Level 1 examines whether the LLM's responses differ between base test and its variants, while Level 2 provides granular analysis of failure patterns. This study focuses on Level 1 robustness assessments, while Level 2 granular assessment remains a direction for future work.

For each test set, we perform a question-level analysis. Using the systematic evaluation framework, we analyze whether the LLM's response remains consistent across the base question and all its variants. A mismatch between the LLM's response for the base question and any of its variants is considered a prediction inconsistency. The number of questions in a test set exhibiting inconsistent behavior serves as an indicator of robustness or lack thereof. For example, if there exist multiple questions for which the LLM's response for three or more variants (representing $\geq 50\%$ of the total variants) does not match with the LLM's response for the base question, then it indicates a significant robustness issue in LLMs (in MCQ tasks).

F. Results and Discussion

Next, we present and discuss our experimental results. Table III presents the results of the robustness assessment on GPT 3.5 Turbo in MCQ tasks across eight subject datasets.

First, for each dataset, we analyzed the number of questions for which at least one variant whose response is different from the base question's response. Results indicate that the CT-based approach successfully identified robustness issues in GPT 3.5 Turbo across all eight subject areas. The findings are discussed next:

- Humanities: In Philosophy, 156 out of 311 questions (50%) exhibited order-dependent behavior. Similarly, in Prehistory, 146 out of 324 questions (45%) has at least one variant with a different response.
- Social Sciences: Approximately 30% of questions in Geography and US Foreign Policy datasets displayed order sensitivity.
- Other: In Marketing, 50 out of 234 questions (21%) exhibited robustness issues. In Business Ethics, 63% of questions failed the robustness assessment.
- STEM: The most pronounced robustness issues were observed in Computer Science (79% of questions failing the robustness assessment) and Algebra (98% of questions failing the robustness assessment). This may be attributed to these subjects' relatively low initial prediction accuracy, although further investigation is required.

A notable observation is that despite substantial initial prediction accuracy in most subjects (excluding Algebra at 21% and College Computer Science at around 50%), CT-based systematic option-swapping revealed significant robustness issues. For example, the marketing dataset, despite achieving a prediction accuracy of 88% (the highest among all subjects), the LLM failed the robustness assessments for 21% of questions (50 out of 234 total instances). Similarly, in the US Foreign Policy dataset, the LLM prediction was correct for 86 out of 100 questions. However, for 29 % of questions, the LLM produced a different outcome for at least one of the variants.

Next, to further understand the severity of robustness failures, we analyzed the number of questions where at least three out of the six generated variants produced different responses compared to the base question. We believe a 50% or more variant deviation threshold can be a key indicator of significant robustness concerns in LLM's response consistency in MCQ tasks. Across all eight subject areas, substantial robustness issues were observed. For example, the Pre-history dataset, the largest dataset (324 questions) in this study, has 85 questions (26%) for which three or more variants result in a different LLM prediction than their respective base question's response. Among all subjects, the LLM demonstrated better robustness with the Marketing dataset. However, 10% of its questions (23) demonstrated robustness issues where three or more of its variants have a different outcome compared to their base question's outcome.

The results suggest that the number of questions failing the robustness assessments decreases when comparing single-variant deviation to 50% or more variants deviation. Nonetheless, a substantial number of questions still fail the robustness assessment. From a testing standpoint, these results warrant further investigation of LLM's robustness. The findings demonstrate that the CT-based systematic order-swapping approach effectively identifies significant robustness vulnerabilities in LLM's response on MCQ tasks.

Overall, the result demonstrates the effectiveness of the CT-based test approach. With the generation of only six additional test cases per question, we identified a significant number of questions for which the LLM's response exhibits deviation, highlighting the approach's effectiveness in uncovering robustness issues in LLM-based MCQ evaluations.

Second, we conducted a preliminary comparison study between CT-based option-swapping test sets and exhaustive test sets to evaluate their relative efficacy in detecting robustness vulnerabilities in LLMs. For an MCQ dataset with four options, the exhaustive approach has 24 possible permutations. The first configuration (ABCD) is the base question; the remaining 23 represent variants. Recall that we access GPT 3.5 Turbo using an API. OpenAI, the provider of GPT 3.5 Turbo, enforces a daily rate limit on the number of requests per user [21]. Consequently, we are unable to complete the experiments for four datasets namely Geography, Marketing, Philosophy and Prehistory. Multiple execution attempts for these datasets resulted in timed-out errors. To address this limitation, we plan to conduct a comprehensive comparison study using a locally hosted LLM as a part of our future work.

For the remaining four datasets (US Foreign Policy, College Computer Science, Algebra, and Business Ethics), we compare the performance of both approaches on two scenarios: (1) the number of questions where at least one variant response differs from the base question's response and (2) the number of questions for which 50% or more variant responses differ from the base question's response. Note that, for scenario 2, we used a threshold of 3 or more variants for the CT-based approach and 12 or more variants for exhaustive testing.

Figures 8 and 9 presents the results from our comparison study. In both figures, the X-axis represents the dataset information, while the Y-axis represents the total number of questions. As presented in Table II, all four datasets consists of 100 questions. In scenarios where one or more variants' responses differ from the base question's response, we observed that the CT-based option-swapping test sets demonstrated detection capabilities comparable to exhaustive testing, with the latter exhibiting marginally better performance. For example, in the case of the business ethics dataset, seven additional questions (70 vs. 63) failed in the robustness assessment when tested using the exhaustive approach. However, despite this marginal

TABLE III: Model prediction results and variant response analysis across datasets

Name of the Dataset	Number of base questions	Model prediction for base questions		Questions with ≥1 variant response different from base		Questions with ≥ 3 variants response different from base	
		Correct	Incorrect	Count	%	Count	%
Geography	198	164	34	59	30%	27	14%
US Foreign Policy	100	86	14	29	29%	13	13%
College Computer Science	98	53	45	77	79%	50	51%
Algebra	100	21	79	98	98%	85	85%
Business Ethics	100	63	37	63	63%	39	39%
Marketing	234	206	28	50	21%	23	10%
Philosophy	311	224	87	156	50%	77	25%
Prehistory	324	236	88	146	45%	85	26%

performance difference, the CT-based test, with a significantly smaller number of tests (6) compared to the exhaustive approach (23), achieved a high detection rate. For the business ethics dataset, the CT-based dataset achieves a 90% detection rate (63/70). A similar pattern is observed for the other three datasets, with the CT-based achieving detection rates of 78%, 88%, and 100% for the US Foreign Policy, Computer Science, and Algebra datasets, respectively. For the other scenario, where 50% or more variants' responses differ from the base question's response, CT-based demonstrated marginally better performance than the exhaustive test set. These preliminary findings suggest that CT-based option-swapping is a promising approach for assessing the robustness of LLMs in MCQ tasks, offering a valuable alternative to computationally expensive exhaustive testing.

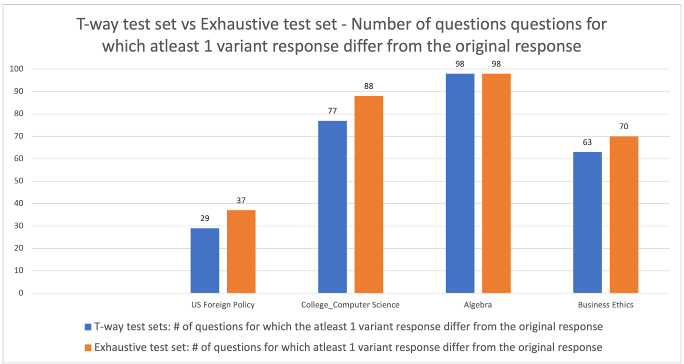


Fig. 8: T-way test set vs. Exhaustive test set - for atleast one variant

G. Threats to Validity

Threats to internal validity are factors that may be responsible for the experimental results without our knowledge. To mitigate the risk of human errors, we tried to automate as many tasks as possible, from generating the CT-based robustness assessment test set, interacting with LLM, recording its response (test execution), and analyzing the test results. Furthermore, we performed additional manual checks and used a pivot table to verify the validity of the results. For example, while analyzing the results of the CT-based test set, among the eight datasets, the LLM produced unexpected responses for two questions in the Computer Science dataset (Questions 77 and 80). These questions were subsequently removed from the analysis, resulting in a total of 98 questions considered instead of the original 100. Similarly, when analyzing the results from the exhaustive test set, the LLM generated unexpected responses for two base questions in the Algebra dataset (Questions 3 and 17), which were likewise excluded from the analysis.

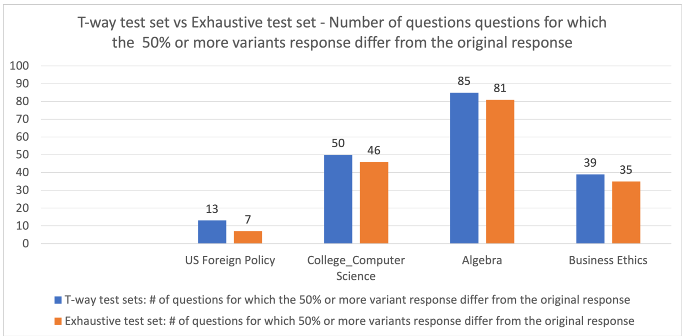


Fig. 9: T-way test set vs. Exhaustive test set - for 50% or more variants

Threats to external validity occur when the results from our experiments cannot be generalized to other subjects. GPT 3.5 Turbo, the LLM model used in our study, has been used in other similar studies [10], [12]. Furthermore, in our experiments, we randomly selected eight datasets across different subject areas from the MMLU dataset, thereby alleviating the risk of a lack of diverse subject areas in our study.

V. Related Work

In this section, we discuss existing work that is closely related to our work. First, we discuss prior research investigating LLM's sensitivity to option ordering in MCQ tasks.

Gupta et al. conducted a robustness assessment of LLMs in MCQ tasks using test-retest reliability, a measure used to assess consistency [9]. While their research objective is the same as ours – investigating the impact of option order permutations on LLM predictions using the MMLU benchmark, they employed random shuffling and evaluated LLMs using 5-shot in-context learning. Furthermore, they introduced a quantitative robustness metric measuring the consistency of correct answers between original and shuffled versions of the MMLU dataset, utilizing two randomly shuffled datasets for assessment. Their study, which evaluated various open-source LLMs, corroborated our findings regarding LLMs' sensitivity to option ordering, particularly noticeable poor performance in STEM subjects. While their work shares our fundamental research goals, our approach differs significantly by implementing a CT-based systematic swapping strategy, enabling more granular question-level analysis compared to their random shuffling approach.

Pezeshkpour et al. investigated the robustness of LLMs with respect to the sensitivity to option ordering in MCQs [11]. Their study examined the option order sensitivity and explored the factors contributing to this sensitivity in LLMs. Specifically, they investigated the robustness of two LLMs under various conditions (broader in scope) by examining the impact of model size, tuning mechanisms, and diverse MCQ test sets with varying numbers of options (3, 4, and 5 options per question). The authors demonstrated that GPT-4 and InstructGPT exhibit significant sensitivity to option order in zero-shot settings across multiple benchmarks. These findings align with our results, as all our experiments are conducted in zero-shot settings, meaning the LLMs are neither fine-tuned nor instruction-tuned and are used in their pre-trained state. Furthermore, their results indicate that few-shot settings do not effectively mitigate the lack of robustness to option order sensitivity. To address these challenges, they proposed calibration strategies aimed at improving robustness. While both their work and ours aim to understand the option-order sensitivity in LLMs, the scope of the studies differs. Their work comprehensively analyzes multiple factors influencing option-order sensitivity and proposed methods to enhance robustness. In contrast, our work serves as a pilot study, focusing on the applicability of combinatorial testing (CT)-based systematic test generation for assessing the robustness of LLMs.

Li et al. evaluated the effectiveness of using MCQ tasks in evaluating LLMs [12]. Their investigation revealed that LLMs are susceptible to option order swapping and prefer specific positions. Li et al. used bilingual datasets (English and Chinese), whereas this study focuses exclusively on English datasets. Furthermore, their study, which included GPT 3.5 Turbo (also used in our research), focused on two specific order swaps (ABCD and BACD). On the other hand, our approach generates six unique variants for each question, ensuring comprehensive coverage of all possible 3-way orderings of the options.

Parlapalli et al. explored the impact of option order sensitivity in LLMs [14]. Their findings indicate that LLM performance in MCQ tasks is significantly influenced by the order in which the options are presented, thus exhibiting selection and positional bias in their behavior. To address this issue, they proposed a bias mitigation strategy to minimize the selection bias in LLMs. Similarly, Zheng et al. demonstrated the inherent selection bias in LLMs, specifically how selection bias impacts the LLM behavior in MCQ tasks in option order swapping. Findings from their study indicate that token bias is the primary factor driving LLM's preference towards specific option orderings. Additionally, they presented a de-biasing approach to mitigate selection bias in LLMs. While our work also employs GPT 3.5 and MMLU (similar to [14]), our primary objective diverges significantly. The goal of our work is to explore the applicability of CT and construct a test set through systematic order swapping.

To the best of our knowledge, the work presented in this paper is the first to investigate LLM sensitivity to option order by constructing test sets that guarantee comprehensive coverage of all possible 3-way orderings of the options.

Next, we discuss the use of CT in testing AI systems. While the applicability of CT has been explored to address the various test & evaluation (T&E) challenges across machine learning-enabled software systems [22]–[28], to the best of our knowledge, there are only two prior works that explore the applicability of CT in testing LLM, and they are discussed next.

Garn et al. explored the applicability of CT in generating test oracles to evaluate semantic consistency in LLMs [29]. Given an original sentence, they break down the sentence into words, and synonyms for each word is identified. Next, the approach constructs an IPM by mapping each word from the sentence as a parameter, and synonyms for each word is mapped as their respective values. A pairwise test set is generated, and it is then used to evaluate LLMs for semantic consistency.

Perko et al. investigated the use of CT to evaluate LLM's response to diverse prompts in the medical domain [30]. Their work introduced a CT-based prompt generation pipeline to create variations of original prompts systematically. By using pairwise CT, they generated test cases (additional prompts) and evaluated LLM responses across these variations.

Similar to these studies, this work explores the applicability of CT in testing LLMs. However, the goal of this work is different from that of the prior work. We investigate the use of CT to assess the robustness of LLMs in MCQ tasks by systematically swapping the order of options.

VI. Conclusion and Future work

This paper presents a combinatorial testing-based approach to perform robustness assessments of LLMs in MCQ tasks. By leveraging CT, the approach constructs test sets to assess the robustness of LLMs to option order sensitivity in MCQ tasks. Our approach's key idea is to use a 3-way sequence covering array to generate test sets by systematically swapping the order of options. We performed an experimental evaluation of our approach by performing a robustness assessment on GPT 3.5 Turbo LLM. Datasets from eight different subject areas were randomly selected from the MMLU benchmark. Then, each dataset was converted into a robustness assessment test set using our approach, and question-level robustness assessments were performed. Results suggest that the test set can successfully identify numerous robustness issues across all subject areas. For example, in seven out of eight datasets, $\geq 25\%$ of questions failed in robustness assessments. Overall, the result indicates that the CT-based approach with a relatively minimal number of tests (75% fewer test cases compared to an exhaustive test set) can successfully detect a significant number of robustness issues in LLM performance on MCQ tasks.

We plan to extend our work in the following directions as a part of future work. First, we aim to broaden our investigation by conducting robustness assessment experiments on additional open-source, locally hosted LLMs. Furthermore, we plan to examine the correlation between LLM's confidence (in its response) and response stability, specifically investigating whether base questions for which the LLM had lower confidence correlated with a higher number of deviations among its variants. Next, we will explore the scalability of a CT-based approach for MCQ with five or more options. Is a 3-way sequence covering array sufficient to detect robustness issues for MCQ datasets with five or more options? Findings from this pilot study have demonstrated the applicability of CT in identifying robustness issues (fault detection) in MCQ tasks across diverse subject areas. As a natural next step, we plan to develop CT-based fault localization techniques to understand LLM's robustness failures. The plan is to investigate whether specific combinations in option-order swapping influence LLM's behavior.

Last, we found through this work that LLMs did not always give a consistent response to the same option ordering. Combinatorial test sets are typically generated with the assumption of its application to deterministic software and thus each interaction must appear at least once but need not appear multiple times in the test set; the necessary number of occurrences is called the *index* and denoted with parameter λ. However, LLMs are based on statistical learning and inherently have some stochasticity. Thus repeated queries for the same prompt may be required to accurately evaluate their behavior. This necessitates an investigation into an appropriate redundancy level ($\lambda > 1$) within combinatorial test suites for effectively testing LLMs.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[2] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong *et al.*, "Evaluating large language models: A comprehensive survey," *arXiv preprint arXiv:2310.19736*, 2023.

[3] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang *et al.*, "Position: Trustllm: Trustworthiness in large language models," in *International Conference on Machine Learning*. PMLR, 2024, pp. 20 166–20 270.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=d7KBjmI3GmQ

[5] "Gemini," 2025, accessed: 2025-01-16. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/performance

[6] "Llama," 2025, accessed: 2025-01-16. [Online]. Available: https://www.llama.com

[7] "Helm Leaderboard," 2025, accessed: 2025-01-16. [Online]. Available: https://crfm.stanford.edu/helm/classic/latest//leaderboard

[8] "Helm MMLU," 2024, accessed: 2025-01-16. [Online]. Available: https://crfm.stanford.edu/helm/mmlu/latest/

[9] V. Gupta, D. Pantoja, C. Ross, A. Williams, and M. Ung, "Changing answer order can decrease mmlu accuracy," *arXiv preprint arXiv:2406.19470*, 2024.

[10] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang, "Large language models are not robust multiple choice selectors," in *The Twelfth International Conference on Learning Representations*, 2023.

[11] P. Pezeshkpour and E. Hruschka, "Large language models sensitivity to the order of options in multiple-choice questions," *arXiv preprint arXiv:2308.11483*, 2023.

[12] W. Li, L. Li, T. Xiang, X. Liu, W. Deng, and N. Garcia, "Can multiple-choice questions really be useful in detecting the abilities of llms?" *arXiv preprint arXiv:2403.17752*, 2024.

[13] Z. Zhang, Z. Jiang, L. Xu, H. Hao, and R. Wang, "Multiple-choice questions are efficient and robust llm evaluators," *arXiv preprint arXiv:2405.11966*, 2024.

[14] V. Parlapalli, B. S. Ingole, M. S. Krishnappa, and V. Ramineni, "Mitigating order sensitivity in large language models for multiple-choice question tasks," *International Journal of Artificial Intelligence Research and Development (IJAIRD)*, vol. 2, no. 2, pp. 111–121, 2024.

[15] D. R. Kuhn, R. N. Kacker, Y. Lei *et al.*, "Practical combinatorial testing," *NIST special Publication*, vol. 800, no. 142, p. 142, 2010.

[16] C. Nie and H. Leung, "A survey of combinatorial testing," *ACM Computing Surveys (CSUR)*, vol. 43, no. 2, pp. 1–29, 2011.

[17] D. R. Kuhn, I. Dominguez Mendoza, R. N. Kacker, and Y. Lei, "Combinatorial coverage measurement concepts and applications," in *2013 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2013, pp. 352–361.

[18] D. R. Kuhn, J. M. Higdon, J. F. Lawrence, R. N. Kacker, and Y. Lei, "Combinatorial methods for event sequence testing," in *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*. IEEE, 2012, pp. 601–609.

[19] "Testing event sequences," 2025, accessed: 2025-01-16. [Online]. Available: https://csrc.nist.gov/projects/automated-combinatorial-testing-for-software/combinatorial-methods-in-testing/event-sequence-testing

[20] OpenAI, "GPT-3.5 Turbo," https://platform.openai.com/docs/models, 2024, accessed: 2025-01-16.

[21] "Openai rate limit," 2025, accessed: 2025-01-16. [Online]. Available: https://platform.openai.com/docs/guides/rate-limits/overviewwhy-do-we-have-rate-limits

[22] J. Chandrasekaran, Y. Lei, R. Kacker, and D. R. Kuhn, "A combinatorial approach to testing deep neural network-based autonomous driving systems," in *2021 IEEE international conference on software testing, verification and validation workshops (ICSTW)*. IEEE, 2021, pp. 57–66.

[23] C. Gladisch, C. Heinzemann, M. Herrmann, and M. Woehrle, "Leveraging combinatorial testing for safety-critical computer vision datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 324–325.

[24] E. Lanus, L. J. Freeman, D. R. Kuhn, and R. N. Kacker, "Combinatorial testing metrics for machine learning," in *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2021, pp. 81–84.

[25] A. R. Patel, J. Chandrasekaran, Y. Lei, R. N. Kacker, and D. R. Kuhn, "A combinatorial approach to fairness testing of machine learning models," in *2022 IEEE international conference on software testing, verification and validation workshops (ICSTW)*. IEEE, 2022, pp. 94–101.

[26] T. Cody, E. Lanus, D. D. Doyle, and L. Freeman, "Systematic training and testing for machine learning using combinatorial interaction testing," in *2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2022, pp. 102–109.

[27] E. Lanus, B. Lee, L. Pol, D. Sobien, J. Kauffman, and L. J. Freeman, "Coverage for identifying critical metadata in machine learning operating envelopes," in *2024 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2024, pp. 217–226.

[28] J. Chandrasekaran, E. Lanus, T. Cody, L. J. Freeman, R. N. Kacker, M. S. Raunak, and D. R. Kuhn, "Leveraging combinatorial coverage in the machine learning product lifecycle," *Computer*, vol. 57, no. 7, pp. 16–26, 2024.

[29] B. Garn, L. Kampel, M. Leithner, B. Celic, C. Çulha, I. Hiess, K. Kieseberg, M. Koelbing, D.-P. Schreiber, M. Wagner *et al.*, "Applying pairwise combinatorial testing to large language model testing," in *IFIP International Conference on Testing Software and Systems*. Springer, 2023, pp. 247–256.

[30] A. Perko, I.-D. Nica, and F. Wotawa, "Using combinatorial testing for prompt engineering of llms in medicine," in *27th International Multiconference Information Society–IS 2024*, 2024, pp. 930–935.