# Holding BERTs Hand Through ABCD Questions

**Martin Bakač, Martin Josip Kocijan, Nikola Kraljević**

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{Martin.Bakac,Martin-Josip.Kocijan,Nikola.Kraljevic2}@fer.hr

**Abstract**

Question answering tasks have long been a central challenge within NLP. This paper presents our approach to tackling the Sentence Puzzle task in the SemEval-2024 Task 9 competition: "BRAIN-TEASER: A Novel Task Defying Common Sense." Diverging from the conventional multiclass classification framework, we introduce a novel pairwise comparison methodology. Instead of forcing the model to choose the correct answer from four options, we reformulate the task into multiple decisions, evaluating triplets of answers at a time. This approach aims to make the model more robust to distractors and enhance the model's reasoning capabilities and accuracy. We compare results using this method and previously explored reformulations of multiple choice as a multiclass classification problem and a series of binary classification choices. We aim to find out which of these ways of *asking* BERT a multiple choice question works best on what task – SemEval SentencePuzzle and WordPuzzle. Our results showcase that the pairwise approach yields results that are worse than the multiclass classification approach, but better than the binary classification approach on both problems.

## 1. Introduction

The competition, named "*SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense*" described in the paper (**?**) is here to push modern models to their limits, putting them to quite an adversarial setting. That paper makes a distinction between vertical and lateral thinking. Vertical thinking, also known as linear, logical or convergent thinking, is mostly a sequential and analytical process, requiring direct memory recall or a few logical steps to come to a sensible conclusion. Lateral thinking, also known as *thinking outside the box*, is a more divergent and creative process, where the question might not make sense at first, and in order to come up with an answer, it is necessary to explore multiple angles. An even more adversarial task for transformer-based models is answering questions that require lateral thinking. State-of-the-art models such as ChatGPT show an accuracy of 60%, humans show a 90% accuracy, while random guessing gets close to 25% accuracy as there are four answer candidates for each question (**?**). In this paper, we tackle the "BRAINTEASER" puzzles with an alternative approach that we believe will make it easier for transformer models to reason about all the possible answers, focusing on BERT-like models (**?**). The puzzles are structured as multiple-choice questions with one question being *"None of the above"*. Our goal is to compare different ways of formulating a multiple-choice question by fine-tuning BERT-like models on these formulations. In Section 3. we will describe the form of the two types of puzzles. Our approach, the reasoning behind it, and other formulations of the multiple choice problem will be described in Section 4.. Results of our experiment will be presented in Section 5..

## 2. Related Work

Reasoning is a hot topic in NLP, with transformer-based models in the center of attention. It is no secret that transformer-based models are far from perfect, and that it takes just one session of asking questions about a topic that one is well-versed in to see inconsistencies but also to be amazed at times. Knowing where a model fails is important and may lead to improvement. Many benchmarks have been performed with the purpose of challenging models with more intricate reasoning, like "Commonsense QA" (**?**), a dataset comprised of questions that are easy for humans and that require no prior knowledge such as a specific document or context, and require just common sense. Similarly, "BRAINTEASER" (**?**) is a benchmark dataset comprised of questions that are constructed in such a way that it is required to consider multiple approaches when answering them. As this was a competition dataset, there was previous work that tackled this problem. **?**) showed promising results, significantly outperforming baselines of 60% reported in (**?**). Their approach consisted of fine-tuning models with a straightforward approach, treating the problem as a multiclass classification task. Additionally they performed a transformation of the problem to a binary classification problem. The transformation worked by taking each question with four candidate answers and mapping it to three questions that required a binary label, signalling whether the answer was correct or not [1]. Their results showed that the transformation was not quite useful, as the same models had somewhat worse accuracies when the transformation was applied. This inspired us to explore another transformation of the problem, as we believe that presenting a model with only one candidate answer takes away critical information. When one is answering a question with candidate answers, it is a good idea to eliminate candidates that are obvious outliers. That is why we believe that an approach where we force a model to choose between two candidate answers at a time might be helpful, and this is described in detail in Section 4..

## 3. Dataset

As we have mentioned before, the "BRAINTEASER" dataset consists of two types of puzzles, *Sentence puz-*

[1] The fourth candidate answer was always "None of the above" so they discarded it.

Table 1: A **word puzzle** question (sub-task B). The correct answer is in bold.

| Question | Candidates |
|---|---|
| What part of London is in France? | **The letter N.**<br>The letter O.<br>The letter L.<br>None of the above. |

Table 2: A **sentence puzzle** examples (sub-task A). The correct answer is in bold.

| Question | Candidates |
|---|---|
| A man shaves everyday, yet keeps his beard long. | **He is a barber.**<br>He wants to maintain his appearance.<br>He wants his girlfriend to buy him a razor.<br>None of the above. |

*zles* and *Word puzzles*. Word puzzles focus more on letter composition, defying the original meaning of the words at hand, as shown in Table 1. The second type of puzzles, so-called *Sentence puzzles*, are shown in Table 2. The three-step process of constructing the "BRAINTEASER" dataset is described in great detail in (**?**). The process consists of collecting data (Internet scraping), sampling distractors and generating reconstruction examples. As some pre-trained models could have included these questions in their training corpus, steps were taken to ensure that the questions were novel to the models (**?**). The dataset consists of the original data, the semantic reconstruction data, and the context reconstruction data. The semantic reconstruction used an open-source rephrasing tool[2] and human annotators as quality control to rephrase the questions. The context reconstruction process used GPT-4, as well as human annotators, to initially shift the context of the question. In the following sections we will refer to these distinct types of questions as **Ori.**, **Sem.** and **Con.** for original questions and semantic and context reconstructions. Finally, the data was split into $507/120/120$ questions for the *train/test/validation* datasets, respectively.

## 4. Methods

In this section we will discuss the different approaches to asking multiple-choice questions and describe the training pipeline. The source code is available on GitHub[3].

### 4.1. Reframing multiple choice

This section will discuss the main point of this paper – different formulations of a multiple-choice question for BERT. We will first describe the already explored multiclass and binary classification approaches and introduce the pairwise classification approach.

#### 4.1.1. Multiclass classification

For the multiclass classification approach, the question and answers are concatenated and the problem is framed as multiclass classification where the classes are answers $A$, $B$, $C$, and $D$ (None of the above). The main advantage of this approach is that the model sees all the answers in each pass.

#### 4.1.2. Binary classification

In reframing the problem as a binary classification task we take every question and answer pair and label it $1$ if the answer is the correct one, and $0$ otherwise. When determining the final chosen answer for each question, we group the pairs by question, and, if the model outputs $1$ for only one answer, we chose that answer, otherwise $D$ (None of the above) is chosen. One advantage of this is that we artificially have more training examples, i.e., question-answer pairs.

#### 4.1.3. Pairwise approach

For the new pairwise approach, we transform the questions in the following way. Each tuple comprised of a question and its candidate answers

$$(Q, A, B, C, D)$$

is transformed into three tuples:

$$(Q, A, B, D),$$
$$(Q, A, C, D),$$
$$(Q, B, C, D).$$

In case the correct answer was $C$, the ground-truth label in the tuple $(Q, A, B, D)$ is $D$ (None of the above). For each set of tuples, *votes* for each answer are counted and the answer with the most votes is chosen. We aim to see if this approach sampling distractors multiple times for each example will yield better results.

### 4.2. Training process

In line with the approach taken by (**?**) we load a pre-trained BERT-like model from Hugging Face. The data evaluation function is modified according to the reformulation and a BERT-like model is fine-tuned to the data.

For this work we decided to use general pre-trained models RoBERTa-large[4] (**?**) and DeBERTa-v3-base[5] (**?**) as our goal was to compare different formulations, and was not to achieve the highest score. For future work, this approach could be taken with less general, fine-tuned BERT-like models.

[2] https://quillbot.com/

[3] https://github.com/MBakac/BRAINTEASER

[4] https://huggingface.co/FacebookAI/roberta-large

[5] https://huggingface.co/microsoft/deberta-v3-base

Table 3: Sentence puzzle results across different formulations of multiple-choice question answering. The systems were evaluated using accuracy as the metric. The evaluation was conducted on the test split of the dataset. The columns labeled "Original," "Semantic," and "Context" represent accuracy on the respective question types. "Ori. + Sem." and "Ori. + Sem. + Con." represent grouped accuracy scores across multiple reconstructions of the same question. The group metrics ("Ori. + Sem." and "Ori. + Sem. + Con.") indicate the model's accuracy across different reconstructions of the same question, where a model's prediction is considered accurate if it predicts correctly for both the original and semantic reconstructions in "Ori. + Sem." and for all three reconstructions in "Ori. + Sem. + Con."

| System | Original | Semantic | Context | Ori. + Sem. | Ori. + Sem. + Con. | Overall |
|---|---|---|---|---|---|---|
| **Multiclass classification problem** | | | | | | |
| Human | .907 | .907 | .944 | .907 | .889 | .920 |
| ChatGPT | .608 | .593 | .679 | .507 | .397 | .627 |
| RoBERTa-L | .435 | .402 | .464 | .330 | .201 | .434 |
| microsoft/deberta-v3-base | .775 | .775 | .700 | .775 | .675 | .750 |
| FacebookAI/roberta-large | .850 | .850 | .775 | .850 | .700 | .825 |
| **Binary classification problem** | | | | | | |
| microsoft/deberta-v3-base | .650 | .650 | .525 | .625 | .625 | .608 |
| FacebookAI/roberta-large | .125 | .125 | .125 | .125 | .125 | .125 |
| **Pairwise** | | | | | | |
| microsoft/deberta-v3-base | .650 | .570 | .650 | .650 | .525 | .623 |
| FacebookAI/roberta-large | .675 | .700 | .725 | .650 | .600 | .700 |

Table 4: Word puzzle results across different formulations of multiple-choice question answering. Refer to the caption of Table 3 for further details.

| System | Original | Semantic | Context | Ori. + Sem. | Ori. + Sem. + Con. | Overall |
|---|---|---|---|---|---|---|
| **Multiclass classification problem** | | | | | | |
| Human | .917 | .917 | .917 | .917 | .900 | .917 |
| ChatGPT | .561 | .524 | .518 | .439 | .292 | .535 |
| RoBERTa-L | .195 | .195 | .232 | .146 | .061 | .207 |
| FacebookAI/deberta-v3-base | .625 | .688 | .625 | .594 | .375 | .646 |
| FacebookAI/roberta-large | .312 | .375 | .438 | .250 | .188 | .375 |
| **Binary classification problem** | | | | | | |
| microsoft/deberta-v3-base | .000 | .000 | .000 | .000 | .000 | .000 |
| FacebookAI/roberta-large | .000 | .000 | .000 | .000 | .000 | .000 |
| **Pairwise** | | | | | | |
| microsoft/deberta-v3-base | .562 | .469 | .500 | .469 | .250 | .510 |
| FacebookAI/roberta-large | .250 | .281 | .406 | .219 | .125 | .250 |

## 5. Results

We compared the three described formulations of multiple-choice question answering across two BERT variants and looked at the model accuracy overall as well as specifically for *semantic* and *context* subsets separated for the Sentence puzzle and Word puzzle subtasks (Table 3 and Table 4). Instance- and group-based performance metrics for our models as well as the human, ChatGPT and RoBERTa-L baselines are presented. The results are presented in terms of accuracy. The individual metric columns *Original*, *Semantic* and *Context* in the tables represent the results of respective reconstructions in the dataset, while the group metric columns *Ori. + Sem* and *Ori. + Sem. + Con.* represent accuracy scores across multiple reconstructions of the same question.

### 5.1. Zero accuracy in the binary classification formulation

A clear outlier in Table 4 are the accuracy results in the binary classification formulation of multiple choice. This is due to the *harsh* criteria for deciding on an answer with this approach where only one out of three binary questions has to be an *A*, *B* or *C* and the other two *D (None of the above)*. This, combined with the rather small dataset, and also the word puzzle task being harder overall (our models yielded worse overall accuracy scores for this task), meant that there were no cases where this approach worked, hence

the accuracy is zero. Furthermore, we note that the models have been trained to have a strong bias toward answering negatively to all instances. This is partially caused by the fact that the dataset used to train the models for binary classification is imbalanced, with a ratio of 1:3 between positive and negative instances. This is a significant issue as the models are not able to learn the patterns of the positive instances, which are the ones that we are interested in finding out. The same poor results for this formulation were obtained by Panagiotopoulos et. al. (**?**).

### 5.2. Pairwise approach evaluation

Our pairwise approach showed performance on par with the baseline ChatGPT and RoBERTa-L models(**?**) for individual metrics, and better performance for group metrics for both subtasks. Comparing our approach with the multiclass classification approach, our approach performed worse across both subtasks. Our approach showed better performance than the binary classification approach that showed miserable results across the board. Evidently, training the model on questions and all answers (i.e., using the multiclass approach) yields the best results.

## 6. Conclusion

To conclude, we have implemented and tested a new method for reformulating multiple-choice questions. Our pairwise approach performed better than the binary classification reformulation but worse than the multiclass classification reformulation. For further research we suggest looking into other possible reformulations, expanding the dataset and using more fine-tuned models.