# Our model *shocks* the industry with a novel and innovative approach WOW!

**Wumbolo, Tegla, Nikola Kraljević**

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{Wumbolo,Tegla,Nikola.Kraljevic2}@fer.hr

**Abstract**

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 multi-class 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 binary decisions, evaluating pairs of answers at a time. This approach, inspired by the efficiency of One-vs-One and One-vs-Rest strategies, aims to enhance the model's reasoning capabilities and accuracy . Our results showcase... ako budu dobri rezultati tu cemo flexat, xD

## 1. Introduction

Question answering tasks have long been a central challenge within NLP. The competition, named "*SemEval 2024 BRAINTEASER: A Novel Task Defying Common Sense*" described in the paper (Jiang et al., 2023) is here to push modern models to their limits, putting them to a quite adversarial setting. The paper makes a distinction between vertical and lateral thinking. Vertical thinking, also known as linear, logical, convergent, 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 more of a divergent and creative process, where the question might not make sense when reading it at first and to come up with an answer you need to explore mutliple angles. While LLMs show good vertical thinking capabiliteis, they are notorious for hallucinating answers. An even more adversarial task for LLMs is answering questions which 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 [1].

In this paper, we tackle the second subtask from the competition "Sentence puzzles" with an alternative approach which we believe will make it easier for the LLM to reason about all the possible answers. In Section 3. we will describe the form of the so called "Sentence puzzles". Our approach and the reasoning behind it will be described in Section 4. Results of our experiment will be presented in Section 5. and Section 6. respectivley.

## 2. Related Work

Reasoning in NLP is a hot topic, with LLMs in the center of attention. It is no secret that LLMs are far from perfect and it takes only one session of asking questions about a topic one is well versed in to see inconsistencies, but also be amazed at times. Knowing where a model fails is important and leads to improvement. Many benchmarks have been made with the purpose of challenging models with more intricate reasoning like "Commonsense QA" (Talmor et al., 2019), a dataset consisting of questions which are easy for humans and requrire no prior knowlege like a specific document or context, just common sense. Similarly the "BRAINTEASER" (Jiang et al., 2023) is a benchmark dataset consisting of questions which are constructed in such a way that it is required to consider multiple approaches when answering them.

As this was a competition dataset, there were some previous works that tackled this problem. A team of researchers from the National Technical University of Athens published their submission (Panagiotopoulos et al., 2023) named "*Transformer Models for Lateral Thinking Puzzles*". They showed promising results, significantly outpreforming baselines reported in (Jiang et al., 2023). Their approach consisted of lightweight tuning of the models for the original formulation of the problem as well as transformation of the problem to a binary classification problem. The transformation took form of taking each question with four candidate answers and transforming it to three questions which required a binary label, signaling if the answer was correct or not [2]. 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 subjecting a model to only one candidate answer takes away information. When one is answering a question with candidate answers, it is a good idea to eliminate candidates that are somewhat obviously wrong. That is why we believe an approach where we force a model to choose between two candidate answers at a time might be helpful, this is described in detail in Section 4.

## 3. Dataset

Dataset

## 4. Pairwise approach

Pairwise approach

[1] These accuracies are for one of the subtasks of the competition - "Sentence puzzles", which we will describe in more detail in the following chapters

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

## 5. Results

Eval Results

## 6. Conclusion

Conclusion

## 7. Acknowledgements

Acknowledgements: (Jiang et al., 2023) semeval, (Panagiotopoulos et al., 2023) ails-lab

## References

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. Brainteaser: Lateral thinking puzzles for large language models.

Ioannis Panagiotopoulos, Giorgos Filandrianos, Maria Lymperaiou, and Giorgos Stamou. 2023. Ails-ntua at semeval-2024 task 9: Cracking brain teasers: Transformer models for lateral thinking puzzles.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.