Evaluation of BERT and ChatGPT models in inference, paraphrase and similarity tasks

Radmir Kim\textsuperscript{1}, Anton Kotsenko\textsuperscript{1}, Aleksandr Andreev\textsuperscript{1}, Anastasiia Bazanova\textsuperscript{1}, Dmitry Aladin\textsuperscript{1}, David Todua\textsuperscript{1}, Aleksei Marushchenko\textsuperscript{1}, and Oleg Varlamov\textsuperscript{1*}

\textsuperscript{1}Bauman Moscow State Technical University, 2-ya Baumanskaya Street, 5/1, Moscow, 105005, Russian Federation

\textbf{Abstract.} The purpose of this paper is to study the application of ChatGPT and BERT models in the field of mechanical engineering. In the context of machine learning, the ChatGPT and BERT models can be applied to various natural language processing tasks such as analyzing technical documentation and building instructions according to a particular version of the documentation, diagnosing malfunctions or customer service. The paper discusses the fundamental features of BERT and ChatGPT models, their origin, and also investigates the main architectural features and identifies the main advantages and disadvantages of the models. The paper analyzes and selects various natural language processing tasks to test the models' ability to understand natural language in the context of machine learning. The selected criterion tasks are divided into semantic groups to identify the capabilities of ChatGPT and BERT models in each of three areas: logical inference tasks, paraphrasing tasks, and text similarity tasks. The paper also discusses the concept of operational design, which involves developing inputs that guide the models to produce desired outputs. The paper quantitatively analyzes and compares the performance of BERT and ChatGPT based models. The reasons for the bottlenecks of ChatGPT model in natural language understanding tasks are discovered and investigated. Possible improvements of ChatGPT model performance using the mivar approach are considered.

1 Introduction

The creation of new intelligent natural language processing technologies for mechanical engineering artificial intelligence (AI) is of great importance. Already, intelligent dialogue systems solve many business problems. We need to explore which methods will better solve natural language understanding (NLU) problems when creating complex AI systems.

Mivar technologies of logical artificial intelligence have been developed for quite some time [1]. For tasks in the format of production networks, they allow finding a solution with linear computational complexity. Recently, quite a few mivar expert systems have been created in various fields: systems for plant care [2], for managing educational programs [3]

* Corresponding author: ovarlamov@gmail.com
at the university [4], for intelligent management systems [5] of vehicles [6] and robots [7],
system of making decisions about the safety of thermolabile blood components [8], for
planning of robot actions [9] and for many other areas. Mivar technologies are also used in
hybrid intelligent information systems (HIIS). HIIS can be used, for example, for
judicial practice. In addition to the mivar approach, metagraph representation [13] of data
sets [14] can be used in HIIS to overcome the limitations [15] of existing knowledge bases
[16].

The HIIS approach also includes neural network methods. Neural networks can be used
for generating questions in Russian [17], for sentimental analysis of multilingual texts [18],
for creating a brief summary of judicial acts [19] or for bypassing text CAPTCHAs [20]
using convolutional neural networks [21]. The use of LiDAR has allowed solving a list of
new tasks: evaluation of the quality of segmentation of an individual tree [22],
determination of the location of trees and estimation of their diameters [23], inventory of
trees [24] and classification of tree species [25] based on LiDAR data [26]. With the help of
satellite imagery analysis [27] has been possible to solve tasks of segmenting trees damage
[28] and identifying deforested areas [29] using neural networks [30]. In addition to the
tasks described above, neural network technologies in HIIS allow solving a huge number of
completely different tasks: management of data lakes [31], processing and visualization of
tomography signals for making decisions on COVID [32], detection of energy theft in
intelligent networks using explainable attention maps [33], for removing noise in audio and
video [34], as well as for indoor navigation [35].

Thus, the topic of the work is relevant and important for the creation of new intelligent
natural language processing technologies in mechanical engineering.

2 The ChatGPT and BERT models

A revolution in the world of natural language processing has become transformers –
language models trained in self-learning mode on large amounts of data in the form of
unprepared text. The transformer consists of two main elements: an encoder and a decoder.

After the transformer architecture based on the attention mechanism showed its
effectiveness, its individual parts acquired independent existence. First, a network called
Generative Pre-trained Transformer (GPT) was developed, which used a modified
transformer decoder. A model called Bidirectional Encoder Representations from
Transformers (BERT) was then created using the transformer encoder.

Besides the transformer, the new models are united by a learning strategy on a large
corpus of unlabeled texts. GPT predicts the next word of the text, and BERT predicts
“closed” words within a sentence. As a result of such training, a language model is formed,
which includes grammar, semantics and even certain knowledge. After preliminary
training, the model parameters are fine-tuned for a specific task using labeled data.

Standard language representation models that existed before BERT, also including
BERT, were unidirectional. This limited the choice of architectures that could be used for
pre-training. For example, in GPT, each token could only serve the previous token (from
left to right) in the internal attention layer of the model.

The GPT network is a transformer decoder with the second attention block removed,
after which it becomes like an encoder. However, the fundamental difference from the
encoder is the use of masked self-attention. Since the next word is predicted during the pre-
training process, each word in the input sequence can only “see” the words that come
before it, but not after it. Therefore, a mask is added to the self-attention weights in which
the elements above the diagonal are equal to minus infinity.
The ChatGPT model is a fine-tuning of the GPT-3.5 transform architecture. To train the model, the Reinforcement Learning with Human Feedback (RLHF) approach is used, which allows improving the basic GPT-3 model towards understanding more complex user requests or instructions, reducing the likelihood of generating unreliable and toxic information. The RLHF approach is to use a reward model calibrated according to expert judgment.

3 Natural language tasks to evaluate a model's understanding ability

As criteria for evaluation, we will take tasks from the GLUE benchmark, which includes the following ones:

1. CoLA (Corpus of Linguistic Acceptability) is a binary single-sentence classification task to determine whether a given sentence is linguistically “acceptable”.
2. SST-2 (The Stanford Sentiment Treebank) is a binary classification task for predicting the sentiment of a given sentence.
3. MRPC (Microsoft Research Paraphrase Corpus) is a task of predicting whether two sentences are semantically equivalent.
4. STS-B (Semantic Textual Similarity) is a task of predicting how similar two sentences are on a scale of 1-5 in terms of semantic meaning.
5. QQP (Quora Question Pairs) is a collection of question pairs from the Quora website to answer community questions. The task is to determine whether a pair of questions are semantically equivalent.
6. MNLI (The Multi-Genre Natural Language Inference Corpus) is the task of predicting whether a premise entails a hypothesis, contradicts a hypothesis, or neither, given a premise sentence and a hypothesis sentence.
7. QNLI (Question Natural Language Inference) is a binary classification task built on SQuAD that aims to predict whether a context sentence contains the answer to a question sentence.
8. RTE (Recognizing Textual Entailment) is the task of predicting whether a premise entails a hypothesis.

All tasks are single-sentence or sentence-pair classification tasks, with the exception of STS-B, which is a regression task. Also, tasks can be divided into three semantic groups:

1. Inference Tasks: MNLI, QNLI, RTE
2. Paraphrase Tasks: MRPC, QQP
3. Similarity Tasks: STS-B

We will use the following models for comparison:

1. ChatGPT
2. BERT-base
3. RoBERTa-large

Thus, using this set of models, we can determine the limits of the capabilities of ChatGPT models in comparison with BERT models. The lower bound can be determined using the standard medium-sized BERT-base model, and the upper bound using RoBERTa-large, which is SOTA for many natural language understanding tasks.

4 Inference tasks

The results of the models for logical inference tasks are presented in Table 1.
Table 1. The results of the models in logical inference tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>MNLI Entailment</th>
<th>MNLI Contradiction</th>
<th>MNLI Neutral</th>
<th>RTE Entailment</th>
<th>RTE Not_Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>88.0</td>
<td>88.0</td>
<td>72.0</td>
<td>76.0</td>
<td>64.0</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>84.0</td>
<td>92.0</td>
<td>88.0</td>
<td>92.0</td>
<td>76.0</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>92.0</td>
<td>96.0</td>
<td>90.0</td>
<td>96.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>

As can be seen from Table 1, ChatGPT performs well on logical inference tasks. In constructing hypothesis and premise connectivity inference (MNLI), ChatGPT outperforms even RoBERTa-large by 1.3%, and the standard BERT model by 6.6%. ChatGPT is also the winner in the task of finding a consequence of a hypothesis from a premise (RTE), outperforming the large BERT model by 4%, and the base one by as much as 18%. In the task of determining the presence of an answer in context (QNLI), RoBERTa-large outperforms ChatGPT by 10%, and the GPT model, in turn, has a result comparable to the basic BERT-base model.

5 Paraphrase tasks

The results of the models for paraphrasing tasks are presented in Table 2.

Table 2. The results of the models in paraphrasing tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRPC Accuracy</th>
<th>MRPC F1</th>
<th>QQP Accuracy</th>
<th>QQP F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>90.0</td>
<td>89.8</td>
<td>80.0</td>
<td>80.0</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>92.0</td>
<td>92.0</td>
<td>90.0</td>
<td>89.4</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>66.0</td>
<td>72.1</td>
<td>78.0</td>
<td>79.3</td>
</tr>
</tbody>
</table>

From the results we see that ChatGPT does a good job of identifying semantically equivalent questions (QQP), in this case the result is still worse, but comparable to the quality of the basic BERT model. RoBERTa-large on average shows 11% higher metrics on the same data.

In the case of identifying semantically equivalent sentences, even the basic BERT model is ahead of ChatGPT by a large margin – 24% for the Accuracy metric and 17.7% for the F1 metric. In general, this result of ChatGPT’s behavior can be explained by the fact that the model copes well with not very complex text structures, for example, questions from a user site in the QQP task, because questions are often very similar in structure, so ChatGPT is able to determine that two users are asking about the same thing and we see accuracy results similar to BERT-base. In the case of the more complex Microsoft Research Paraphrase Corpus dataset, ChatGPT’s performance drops significantly even compared to the simplest BERT model. In order to understand ChatGPT’s weak point for this task in more detail, let’s take a look at the model’s performance for each class. The results of the work by class are shown in Table 3.

Table 3. Accuracy metric by class in the MRPC task.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRPC Entailment</th>
<th>MRPC Not_Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-base</td>
<td>88.0</td>
<td>92.0</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>92.0</td>
<td>92.0</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>88.0</td>
<td>44.0</td>
</tr>
</tbody>
</table>
The table shows that ChatGPT achieves results matched in accuracy compared to BERT-base when evaluating the “Entailment” class, and a sharp decrease in performance is observed in the “Not_entailment” class, where the sentences in the pair are semantically unequal. This indicates that ChatGPT is insensitive to the semantic difference between a pair of sentences, which may be due to the lack of human feedback on this aspect during model training.

6 Similarity tasks

The results of the models for the similarity task are presented in Table 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>STS-B</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson’s Corr</td>
<td>Spearman’s Corr</td>
<td></td>
</tr>
<tr>
<td>BERT-base</td>
<td>83.0</td>
<td>81.9</td>
<td></td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>92.9</td>
<td>91.1</td>
<td></td>
</tr>
<tr>
<td>ChatGPT</td>
<td>80.9</td>
<td>72.4</td>
<td></td>
</tr>
</tbody>
</table>

Based on the results, you can see that BERT models show better results; if you compare ChatGPT with the base model, the gap is 9.5% in the case of Spearman’s correlation, and in comparison, with the large BERT model the difference is already 18.7%. Let's try again to identify areas where ChatGPT fails in performance and why. Since STS-B is a regression task, we sample multiple examples from a uniform distribution for text similarity ranging from 0 (no similarity) to 5 (equivalent in value), and show the absolute difference between the predicted and true values for ChatGPT and BERT-base, respectively.

The results are presented in Fig. 1, the horizontal axis shows the distribution values for similarity from 0 to 5, and the vertical axis shows the absolute difference in predictions. In most cases, ChatGPT is inferior to the basic BERT model because its predictions are usually far from true based on the absolute difference. We can also observe that ChatGPT performs worse when the sentences in the pair have lower similarity, such as 1.5 or 2.5, which correlates with our observations on the MRPC task where ChatGPT performed poorly in identifying semantically different sentences.

For ChatGPT, the most difficult is to predict similarity estimation for a pair of sentences within the decision boundary (about 2.5 similarity points). One reason is that ChatGPT is not fully trained for the STS-B task, so it cannot determine the correct decision boundary.

Fig. 1. Results of BERT-base and ChatGPT models for STS-B.
7 Opportunities for improving chatgpt using the mivar approach

As noted earlier, ChatGPT performs poorly in the tasks of determining the degree of similarity of sentences and paraphrasing; ChatGPT performed especially poorly in the MRPC task on the “Not_entailment” class, which indicates that it is insensitive to the semantic difference between sentences. A possible reason for this may be that ChatGPT knows many general facts about our world, but cannot qualitatively consider them in a certain context and, due to incomplete understanding, the model can find a logical link where there is none, for example, fail to identify a contradiction in sequential sentences and consider that they are logically related. Also, such instability of the model may be due to the lack of feedback from human during model training.

In a situation where the model exhibits instability of logical inference, mivar technologies can help. Mivar (Multidimensional Informational Variable Adaptive Reality) is the smallest structural element of a discrete information space. It is used in the creation of artificial intelligence for semantic analysis. The mivar model allows artificial intelligence to maintain dynamic balance and effectively overcome contradictions. Artificial intelligence created on mivar principles, if conditions change, solves the task in real time without the participation of a human operator.

In the case of ChatGPT, mivar technologies can be useful to improve query processing, that is, the generative model can use the mivar model to analyze questions and determine relationships between words and phrases. This can help the model better understand the context of the question and generate more accurate and coherent responses. The idea of improving ChatGPT is to replace human feedback with preprocessing of the source text using a mivar expert system, which will help the model automatically refine the context using a knowledge base. To improve the understanding of the text, you can accompany ChatGPT with hints, which can be generated using the mivar approach in the form of additional markup for sentences to identify contradictions, set expressions, clarify the meaning of a word in context, etc. Thus, the system gradually comprehends the text and achieves maximum understanding of natural language by artificial intelligence.

8 Conclusion

This paper examined the capabilities of BERT and ChatGPT models in mechanical engineering and natural language processing. It has been demonstrated that these models can be used to create more efficient and accurate production control systems, as well as to automate quality control processes. One of the main advantages of BERT and ChatGPT models is their ability to learn from large amounts of data, which allows them to achieve high accuracy in natural language processing tasks.

The research conducted in this paper identified and demonstrated the natural language understanding capabilities of BERT and GPT models in various NLU tasks. Through a series of quantitative studies, we found that ChatGPT performs well on inference tasks, but fails to perform paraphrasing and text similarity tasks, especially for negative examples. Overall, ChatGPT has comparable understanding ability to some pre-trained BERT models, but still fails to outperform the current best models on some NLU tasks.

In the future, BERT and ChatGPT models can be used to create more complex artificial intelligence systems that can process and analyze instructions from machine operators to provide specific instructions and guidance in various situations. However, despite all the advantages, using BERT and ChatGPT models can also face some problems such as training complexity, “hallucination” problems of generative models, and high computational requirements.
Overall, the BERT and ChatGPT models provide a powerful tool for natural language processing in mechanical engineering that can be used to create more efficient and accurate manufacturing control systems and other applications. However, to achieve the best results, it is necessary to continue research and development of these models, as well as take into account possible problems and limitations in their use. One of the possible directions for improving models, as proposed in this paper, could be natural language understanding systems based on a combination of generative and graph approaches, for example, a combination of the ChatGPT model and the mivar approach.

References

12. M. O. Taran, et.al., Studies in Computational Intelligence 1064, 466–475 (2023)
13. Y. E. Gapanyuk, attern Recognition and Image Analysis 33(3), 300–305 (2023)
16. V. Tarassov, Y. Kaganov, Y. Gapanyuk, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12948, 135–151 (2021)
17. M. A. Belyanova, A. M. Andreev, Y. E. Gapanyuk, Studies in Computational Intelligence 1008, 217–223 (2022)


20. V. Terekhov, V. Chernenky, D. Ishkov, Communications in Computer and Information Science 1573, 111–122 (2022)


22. I. A. Grishin, T. Y. Krutov, A. I. Kanev, V. I. Terekhov, Optical Memory and Neural Networks (Information Optics) 32, 270–276 (2023)


