трудов. Вклад педагога и наставника в дело образования, науку, воспитание неоценим, так как достоин не только уважения, но и изучения, использование педагогического опыта Николая Александровича в современных условиях высшего образования. На методологических семинарах и практических занятиях, проводимых со студентами и аспирантами, имеет смысл изучать наследие ученого и трудов его учеников.

Литература

1. Андреев, В.И. Педагогическая эвристика для творческого саморазвития многомерного мышления и мудрости : монография / В.И. Андреев. – Казань : Центр инновационных технологий, 2015. – 288 с.
2. Бордовская, Н.В. Диалектика педагогического исследования : Логико-методологические проблемы. – Санкт-Петербург : Издательство РХГИ, 2001. – 512 с.
3. Дружинина, М.В. Педагогический дизайн в профессиональном образовании : монография / М.В. Дружинина; Северный (Арктический) федеральный университет имени М.В. Ломоносова. – Архангельск : САФУ, 2021. – 168 с.
4. Краевский, В.В. Методология педагогики : новый этап : учеб. пособие для студ. высш. учеб. заведений / В.В. Краевский, Е.В. Бережнова. – Москва : Издательский центр «Академия», 2006. – 400 с.
5. Хуторской, А.В. Педагогическая инноватика : учеб. Пособие для студ. Высших учеб. Заведений / А.В. Хуторской. – Москва : Издательский центр «Академия», 2008. – 256 с.

А.В. Зайда, Я.А. Согуляк
Национальный исследовательский
Томский политехнический университет

Problem of cheating with neural networks in language learning

This study explores the performance of commercially available large language models in common language learning tasks. Structure and working principles of neural networks were considered to hypothesize which tasks would perform better. Experiments were conducted to verify the assumptions. Several variants of task adaptations were compared in tests to discover the most resistant to cheating.

Keywords: language learning; cheating; large language models; task evaluation; task adaptation

Over the course of the previous year, large language models (LLMs), a specific category of neural networks, have been receiving an immense amount of attention. This burgeoning interest can be attributed in part to numerous advances in language modeling, which have notably empowered these models to achieve a level of text comprehension that closely resembles human understanding. However, arguably the most pivotal factor driving widespread public interest in this domain has been the emergence of commercial-grade services, offering the capabilities of such models in a chat bot format. This development has effectively facilitated nearly unrestricted usage of LLMs in a wide array of natural language processing applications, including, regrettably, their exploitation for illicit purposes such as cheating in language learning exercises.

This work aims to explore capabilities of commercially available LLMs in regards to cheating in various language learning tasks. Therefore, following tasks were put forward:

– Explore working principles of LLMs;

– Identify how susceptible different types of exercises are to being solved by LLMs;

– Explore ways of mitigating cheating perpetrated with use of LLMs.

Modern LLMs fundamentally solve language modeling tasks. In essence, text generation by such neural networks is an exercise in building a conditional probability distribution over entire dictionary. In other words, model just estimates how likely each word is to be the continuation of the sentence, given certain words as the context. There were countless developments in the mechanism of identifying context: from considering all other words in a sequence as context to only considering part of earlier encountered words. Yet all of them work with the proximity of context words to the predicted one. Therefore, LLMs do not inherently possess any capability for reason or logic, as they do not exercise any reasoning in tracking contextual relations of words.

The probability of each word being the continuation of generated text is computed with attention [4] mechanism. This mechanism learns how different words correlate with each other in training dataset of text. Key factor in this process is the virtually unlimited capability to track similarity between word even across vast distances, unlike earlier approaches to language modelling, namely recurrent neural networks with gated recurrent units or long short-term memory. Although, every real application of LLM has a practical limit of this trackability. Even the oldest relevant model – the transformer in its 2017 conception has a theoretical limit of up to 2000 words [4], likely significantly lower in practice. Whenever generated text exceeds maximum length of sequence, a phenomenon known as catastrophic forgetting [2] occurs, characterized by inability of the model to accurately judge the next word in sequence based on earlier outputs. This manifests in a long illogical generated

text, usually with references to nonexistent statements. Even though modern LLMs are known for resolving catastrophic forgetting by significantly increasing the word limit over earlier neural networks, they still are susceptible to this phenomenon by their design.

After the pretraining process that aims to build probabilistic language model is finished, commercially available model usually undergoes additional conditioning and fine-tuning. Services that mimic chat bots will be trained on additional data – actual chat-like messages. Other types of fine-tuning include additional training on data from a specific field in order to from actual expertise in LLM. This step is majorly responsible for any semblance of formal logic or consciousness that LLM might appear to possess. Usually, this results in a model that functions on a query-response principle.

To evaluate the performance of LLMs in language learning exercises it is necessary to consider these exercises from the perspective of model. Therefore, two large groups of exercises can be identified: fill in the gap and sequence generating exercises.

Fill in the gap types of exercises are straightforward for LLMs and can be expected to be solved very accurately. After all, prediction of a word based on its context is the very working principle of language models. It is possible that vocabulary type of exercises of this kind are more prone to errors. However commercially available chat-like LLMs are fundamentally capable of ingesting target vocabulary for a given exercise, increasing the accuracy of prediction.

Sequence generating exercises include the prediction of more words in a sequence, like a sentence. Such exercises range from completing the sentences with 2-5 words to essay writing. These types of tasks are more prone to errors even from the probabilistic standpoint, since subsequent sampling of words from a conditional probability distribution tend to accumulate inaccuracies. Furthermore, features of LLMs such as catastrophic forgetting and lack of formal logic increase the risk of failure in such exercises. Still, generation of short sentences and sentence parts is expected to be performed flawlessly by LLMs.

It was decided to choose one exercise type from each of the 2 discussed groups. Common exercises of choosing one word for a gap and essay writing were chosen as they can be easily generalized. Following experiments were conducted using commercially available service ChatGPT, which provides a chat-like interface for state-of-the-art GPT large language models. For the fill in the gap type of exercise. 20 sentences were constructed with vocabulary roughly matching B1 level of English proficiency and above. From each of these sentences, one word was intentionally omitted to form the gap. Then those sentences were fed to a language model with a prompt to fill the gaps in them. The described method should also work with a cloze type of exercises, but independent sentences with gaps should provide better variety and difficulty for LLM.

As expected, the model performed well on these tasks. Even though some filled words did not match those that authors supposed, all the guesses were still plausible given limited context and polymorphism of the language.

Sequence generation exercises are more promising in identifying non-human written text. To conduct the experiment, it is necessary to choose a specific exercise.

In the realm of artificial intelligence, the selection of an essay as a testing ground serves as a means to examine the cognitive capabilities of a neural network. Essays, being rich in thematic content and requiring a profound understanding of context, demand the network to not only comprehend the topic at hand but also to dynamically produce relevant and coherent responses. This presents a unique challenge that can help evaluate the network's adeptness at contextual comprehension and information synthesis. Furthermore, the act of crafting an essay transcends mere content generation; it necessitates the logical structuring of ideas, coherence in thought progression, and the articulation of compelling narratives. Additionally, it is imperative to acknowledge the temporal investment involved in the essay-writing process. The arduous nature of this endeavor underscores the appeal of leveraging neural networks for such tasks, as they offer the potential to expedite the creation of high-quality compositions. Thus, it is reasonable to anticipate the eagerness of students to actively incorporate neural networks into their writing workflows, aiming to streamline and enhance their essay composition experiences.

It was decided to evaluate the essay according to the criterion of the first appearance of a logical error. Logical errors often lead to incoherent or ambiguous writing. Essays that lack logical flow or contain contradictory statements can confuse and mislead readers. Also, logical errors can signal a lack of understanding or misinterpretation of the essay's topic. A neural network may generate sentences or paragraphs that are factually incorrect, fail to address the main ideas, or reach flawed conclusions. And finally, writing an essay involves analyzing evidence, reasoning, and critically evaluating arguments. Logical errors can demonstrate the neural network's incompetence in these areas as well.

To make the assessment clearer, it was decided to use the number of the sentence in which the logical error first appears. The word number is not suitable as a criterion, since a logical error in a sentence is not always expressed in one specific word.

Four series of experiments were conducted, in each of which the neural network generated an essay with special conditions. In the first series of experiments, the topic of the essay was asked directly. Perhaps, this is the easiest of all conditions, so the neural network is expected to handle it best. The second series, in addition to directly setting the topic, involves a link to an article [1], information from which the neural network should rely on to write an essay. In

the third series, there is again a link to the article [3], but this time the topic is not directly indicated, it is only asked to write an essay on the topic raised by the author of the article. It is assumed that this condition is the most difficult and the largest number of errors will be made here. And finally, in the fourth series the topic of the essay is again set directly, but the word count range increases from 200-250 to 275-300. It is expected that more words will lead to more errors.

Articles for conditions 2 and 3 were selected based on the considerations that a student with an English level of at least B1 would most often write essays on similar topics and that these articles would provide a good balance between complexity and readability.

For each condition, 20 essays were generated by neural network. All of them were examined manually and the results of testing for logical errors are shown in the table below (Table 1). The essay number is represented as X.Y, where X is the number of the series of experiments, and Y is the number of the essay in the given series. The «sentence» column indicates the number of the sentence in which the logical error was first encountered. If one is not found, the symbol «-» is inserted. If the generated text is not an essay at all, it is set to 0.

Table 1

Raw experiment data

Number	**Sentence**	**Number**	**Sentence**	**Number**	**Sentence**	**Number**	**Sentence**
1.1	–	2.1	12	3.1	0	4.1	–
1.2	–	2.2	2	3.2	5	4.2	–
1.3	–	2.3	–	3.3	4	4.3	13
1.4	–	2.4	10	3.4	–	4.4	–
1.5	4	2.5	8	3.5	4	4.5	15
1.6	–	2.6	22	3.6	4	4.6	18
1.7	5	2.7	15	3.7	3	4.7	12
1.8	3	2.8	13	3.8	2	4.8	–
1.9	–	2.9	9	3.9	4	4.9	15
1.10	15	2.10	6	3.10	3	4.10	–
1.11	–	2.11	2	3.11	2	4.11	9
1.12	3	2.12	5	3.12	6	4.12	–
1.13	–	2.13	8	3.13	8	4.13	11
1.14	–	2.14	–	3.14	8	4.14	14
1.15	7	2.15	14	3.15	3	4.15	–
1.16	–	2.16	–	3.16	0	4.16	17
1.17	11	2.17	11	3.17	7	4.17	–
1.18	–	2.18	7	3.18	4	4.18	–
1.19	–	2.19	–	3.19	2	4.19	–
1.20	3	2.20	4	3.20	11	4.20	19

Next, the data needs to be normalized. To accommodate different sentence sizes. The «sentence» column is now presented in decimal format showing the ratio of the number of the sentence in which a logical error was found to the total number of sentences.

For convenience, it was decided to introduce indexing of sentences starting from 0, i.e. the first sentence is numbered 0, the second is numbered 1, etc. If no logical errors are found in the essay, the cell in “sentence” column is set to 1. The resulting value can be interpreted as the continuous part of essay that does not contain logical errors, with the value of 1 meaning that it is necessary to read the whole essay before encountering first logical mistake. This will also help distinguish between regular cases and cases where the error appeared in the last sentence, as well as cases where the error did not appear at all. The normalized data is represented in Table 2.

Table 2

Normalized experiment data

Number	**Sentence**	**Number**	**Sentence**	**Number**	**Sentence**	**Number**	**Sentence**
1.1	1.0000	2.1	0.6875	3.1	0.0000	4.1	1.0000
1.2	1.0000	2.2	0.0560	3.2	0.2860	4.2	1.0000
1.3	1.0000	2.3	1.0000	3.3	0.2140	4.3	0.6320
1.4	1.0000	2.4	0.4500	3.4	1.0000	4.4	1.0000
1.5	0.1875	2.5	0.3680	3.5	0.2140	4.5	0.9333
1.6	1.0000	2.6	0.8750	3.6	0.2500	4.6	0.9444
1.7	0.2500	2.7	0.7778	3.7	0.1430	4.7	0.6111
1.8	0.1540	2.8	0.6667	3.8	0.0625	4.8	1.0000
1.9	1.0000	2.9	0.4210	3.9	0.2310	4.9	0.8240
1.10	1.0000	2.10	0.2778	3.10	0.1540	4.10	1.0000
1.11	1.0000	2.11	0.0710	3.11	0.0710	4.11	0.4000
1.12	0.1333	2.12	0.2500	3.12	0.3125	4.12	1.0000
1.13	1.0000	2.13	0.4375	3.13	0.4375	4.13	0.5556
1.14	1.0000	2.14	1.0000	3.14	0.4667	4.14	0.7222
1.15	0.4290	2.15	0.7650	3.15	0.1250	4.15	1.0000
1.16	1.0000	2.16	1.0000	3.16	0.0000	4.16	0.8420
1.17	0.6250	2.17	0.5260	3.17	0.4290	4.17	1.0000
1.18	1.0000	2.18	0.3750	3.18	0.1875	4.18	1.0000
1.19	1.0000	2.19	1.0000	3.19	0.0777	4.19	1.0000
1.20	0.1540	2.20	0.1766	3.20	0.6667	4.20	0.9000

In order to clearly show in which cases the neural network performed better and in which it did worse, it is necessary to calculate the arithmetic mean value of the «sentence» column as well as its standard deviation in each

experiment series. First logical error position can be considered a random variable, because text generation by neural network is basically a sampling from a conditional probability distribution, where presence of each word is a random event.

Table 3

Experiment summary

Series	Arithmetic mean value	Standard deviation
1	0.7466	0.3691
2	0.5591	0.3184
3	0.2664	0.2423
4	0.8682	0.1846

Arithmetic mean value shows the approximate accuracy of generated text, the higher the value, the better the model operates. The value of the standard deviation is related to how accurately the arithmetic mean value is determined, but does not indicate the accuracy itself. Basically, the lower the standard deviation, the greater the chance that the arithmetic mean is close to true mean.

Based on the data obtained, it safely can be said that the generation of essays with a direct specification of the topic, as expected, is more accurate than the generation with an indirect specification. Moreover, essays written with a determined topic were more coherent than the ones based on the article. Judging by the value of the standard deviation, the experiment with an increase in the number of words has the most accurate value of the arithmetic mean. Presumably, this may be due to the fact that logical errors for the most part occurred only in the second half of the generated text and were located closer to the average value.

In conclusion, it was identified that the essay writing is the least susceptible to cheating with language models exercise. Still, it is advisable to adapt these exercises by indirectly indicating the topic of essay, for example, by referencing some article as a source of arguments or the whole topic. This kind of adaptation leads to the least logically coherent result when generated by neural network while keeping the task engaging for honest students. Furthermore, increasing the number of words in essays is not advisable, as this action does not hamper cheating while also negatively impacting genuinely written works. On the other hand, shorter exercises with a single generated word proved to be the most vulnerable to cheating with no obvious ways to counteract LLM agent. Due to the discussed architectural features and trends of contemporary LLMs, this result is unlikely to change in the foreseeable future.

Литература

1. Bluth, K. How to Help Teens Put Less Pressure on Themselves : / greatergood, 2022. – URL: https://greatergood.berkeley.edu/article/item/how_to_help_teens_put_less_pressure_on_themselves (дата обращения: 22.10.2023). – Текст : электронный.

2. Luo, Y. et al. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning : / arXiv, 2023. – URL: https://arxiv.org/abs/2308.08747#:~:text=Catastrophic%20forgetting%20(CF)%20is%20a,continual%20fine-tuning%20of%20LLMs (дата обращения: 20.10.2023). – Текст : электронный.

3. Machowska, T. How fashion rules the world : / fibre2fashion, 2007. – URL: https://www.fibre2fashion.com/industry-article/2287/how-fashion-rules-the-world (дата обращения: 22.10.2023). – Текст : электронный.

4. Vaswani, A. et al. Attention Is All You Need : / arXiv, 2017. – URL: https://arxiv.org/abs/1706.03762 (дата обращения: 20.10.2023). – Текст : электронный.

Науч. рук.: Аксёнова Н.В., к-т филол. н., доц.

Т.А. Кинева

Северский технологический институт
Национального исследовательского ядерного университета «МИФИ»

Психолого-педагогические аспекты обучения иностранному языку в техническом вузе

В статье рассматриваются психолого-педагогические аспекты обучения иностранному языку в техническом вузе. Обсуждаются проблемы, с которыми сталкиваются студенты и преподаватели в учебном процессе, и предлагаются пути их решения. Обосновывается необходимость поиска релевантных методов обучения в условиях узкопрофильного технического вуза.

Ключевые слова: качество обучения; узкопрофильный вуз; иностранный язык в техническом вузе; структурно-логические схемы; познавательные процессы.

Необходимость усовершенствования в технических вузах языковой подготовки обусловлено профессиональными стандартами для будущих инженеров как условие их конкурентоспособности. Компетентностная модель требует готовности выпускников бакалавриата использовать полученные умения и навыки в профессиональной сфере. Научно-исследовательская деятельность, как учащихся магистратуры, так и аспирантов, также подразумевает высокий уровень знания иностранного языка.