Linguística. Revista de Estudos Linguísticos da Universidade do Porto, vol. 19, 2024
https://doi.org/10.21747/16466195/ling19a7

Analyzing GPT-4 Misinterpretations of Russian Grammatical Constructions[1]

Timofei Plotnikov
UiT The Arctic University of Norway (Norway)
timofei.plotnikov@uit.no

Recebido: 18/07/2024
Aceite: 02/11/2024
Publicado: 23/12/2024

ABSTRACT: Generative Pre-trained Transformers (GPTs), which are Large Language Models (LLMs) primarily trained on datasets dominated by English texts, can incorporate inaccuracies into their final outputs. The training data bias presents challenges for accurate performance in non-Roman script languages, such as Russian. Challenges are specifically expected with grammatical constructions in Russian which are colloquial, idiomatic, and non-compositional. Although the empirical investigation reveals that, overall, GPT-4 performs well with the majority of Russian grammatical constructions, it still encounters limitations with low-frequency constructions due to insufficient training data, a lack of context, and the influence of the English language.

KEYWORDS: Large Language Models; GPTs; Grammatical Constructions; Russian.

Introduction

Artificial Intelligence (AI) chatbots, which represent practical applications of LLMs, are seeing widespread use due to their advanced capabilities. These models can perform a variety of language-specific tasks such as improving, translating, interpreting text, responding to context, and engaging in meaningful conversations. One of the types of LLMs is GPT, predominantly trained on a vast amount of English data from the Internet (Brown *et al*., 2020; Lai *et al*., 2023; Ray, 2023; Wendler *et al*., 2024). The most renowned AI chatbot, ChatGPT, which implements GPT-4 at the moment, was launched in November 2022 and played a crucial role in advancing research into human-LLM interaction.

While LLMs exhibit human-like capabilities, they are still prone to errors. Assessing the accuracy of the responses from LLMs remains a complex and

[1] Special thanks for assistance with the Python coding and processing are dedicated to the Department of Computer Science, UiT The Arctic University of Norway.

insufficiently explored issue, especially in languages other than English (Papadimitriou *et al*., 2022; Lai *et al*., 2023; Ray, 2023; Hendy *et al*., 2023; Wendler *et al*., 2024; Liu, 2024). OpenAI (the company behind GPTs and ChatGPT) states[2] that GPTs are proficient in working with the English language but can struggle with other languages, especially those that use a non-Roman script. This is why investigating Russian, which is not only one of the most spoken languages in the world but also uses a Cyrillic script, is important for the human-LLM interaction research.

The present study specifically focuses on analyzing the performance of GPT-4, which is implemented in the AI chatbot ChatUiT (https://chat.uit.no/), on a dataset that includes 2,277 Russian grammatical constructions collected from the Russian Constructicon (https://constructicon.github.io/russian/). The Russian Constructicon is an open-access platform featuring a searchable database of Russian multiword grammatical constructions designed specifically for researchers and second language (L2) learners (Bast *et al*., 2021). Considering the potential of AI chatbots that implement GPTs to assist researchers and L2 learners of Russian in language-specific tasks, the following research questions arise: *how well does GPT-4 interpret Russian grammatical constructions? What are the factors contributing to GPT-4's misinterpretations with certain Russian grammatical constructions?*

In order to address the research questions, the primary objective of this study is to conduct an experiment exploring language-specific issues that users can face while using AI chatbots for tasks involving Russian. The final goal is to demonstrate AI performance in languages other than English while acknowledging limitations, thereby contributing to the future development of LLMs.

This article consists of 6 sections. Section 1 introduces the topic. Section 2 provides the reader with the theoretical foundation relevant to the research: an overview of the performance of LLMs applicable to the current study, and an introduction to grammatical constructions and Russian Constructicon. Section 3 describes the experimental setting. In Section 4, the obtained results are interpreted and analyzed. Section 5 summarizes the entire investigation and suggests ideas

[2] https://openai.com/index/chatgpt-can-now-see-hear-and-speak/ (21.03.2024).

based on the findings. Section 6 highlights the limitations of the study and proposes future areas of research.

1. Background

1.1. Previous research on GPTs and their performance

Generative pre-trained transformer is a type of large language models developed by OpenAI, based on a deep learning architecture for handling natural language processing (NLP) tasks, primarily for text generation (Zhu & Luo, 2022; Yenduri *et al*., 2024). GPTs are initially pre-trained on a large dataset of diverse texts and can perform a broad range of language-related tasks, enabling GPTs to generate human-like texts based on a given input. The dataset includes a wide variety of content sourced from a broad spectrum of internet platforms, notably including materials from Wikipedia and Google Books (Brown *et al*., 2020). GPTs are designed to discover statistical patterns between words and their contexts in the training dataset using a technique called autoregressive language modeling, which is a building block of GPTs (Yenduri *et al*., 2024). Autoregressive language modeling attempts to predict the next word based exclusively on the previous words sequentially processing the context from left to right (Brown *et al*., 2020; Zhu & Luo, 2022; Pannatier *et al*., 2024). This technique allows GPTs to generate accurate responses based on statistical predictions. Statistical predictions rely on statistical patterns learned from the training dataset and the frequent occurrence of words, aiming to capture a broader context (Gong *et al*., 2019; Yenduri *et al*., 2024). GPTs heavily depend on contextual information provided in the input, specifically on semantically similar examples, where phrases and sentences, despite their differences, share similar meanings within a single text (Liu *et al*., 2021). By providing more context for GPT models, their performance will be more accurate, coherent, and less biased (Yenduri *et al*., 2024). Without sufficient context, GPTs tend to misinterpret users' requests. This can lead to weak and untruthful performance and predisposes GPTs to frequency/statistical bias, especially with non-frequent and rare words (Brown *et al*., 2020; Dudy & Bedrick, 2020: 13).

LLMs, including GPTs, are predominantly trained on massive corpora of English text (Lai *et al*., 2023; Ray, 2023; Wendler *et al*., 2024). According to Brown *et al*. (2020: 6), GPT-3 predominantly implements the Common Crawl text corpus, which is 93% English by word count and 7% non-English; Russian, one of the most spoken languages in the world, comprises just 0.2% of this corpus. Beginning with GPT-3.5, OpenAI has not disclosed the details of the datasets used to train the GPT models. However, more advanced GPTs still retain a practical bias towards the English language in their outputs (Lai *et al*., 2023). Since the training data in the text corpus is heavily disproportionate, LLMs exhibit decreased performance in languages other than English. This also results in English being employed as an internal pivot language in LLMs; inputs in languages other than English are initially translated into English, processed, and then translated back into the original language, which leads to misinterpretations and failure to convey the intended meaning accurately (Wendler *et al*., 2024). The English-language-biased data in LLMs not only leads to grammatical inaccuracies by imposing English syntactic patterns but also results in the inappropriate application of lexicon to other languages, where it may be irrelevant or inapplicable (Papadimitriou *et al*., 2022). Besides, the biased data may introduce unexpected elements in the output, such as punctuation insertions, which result in an incorrect linguistic representation of other languages (Hendy *et al*., 2023).

GPTs can also struggle with the interpretation of abstract concepts and non-literal language, which includes idioms and sarcasm (Ray, 2023; Yenduri *et al*., 2024). Responses from GPTs tend to be overly literal, leading to misunderstandings of the original meaning and the desired tone. This limitation can result in incorrect translations and interpretations of idiomatic expressions, which are deeply embedded with cultural values. Johnson *et al*. (2022) report that translations from Russian through GPTs often fail to convey cultural values due to the underrepresentation of Russian in the training dataset. In contrast, GPTs tend to reflect American values more effectively due to the dominance of data available mostly in American and, to a lesser extent, British English (Dodge *et al*., 2021; Johnson *et al*., 2022; Jackson, 2023: 21; Liu, 2024: 114); this reveals that GPTs are primarily pre-trained to represent cultural-linguistic concepts within the

Anglosphere. As a result, the prevalence of American English texts can sometimes be visible in outputs generated in another language with a significant change in embedded cultural values (Johnson *et al*., 2022). Therefore, it is important to note that English serves as a pivot language not only lexically but also semantically (Wendler *et al*., 2024), since the role of LLMs is not limited to straightforward translation only. LLMs may fail to accurately convey cultural values in another language, but the biased output can still appear somewhat grammatically correct and acceptable to non-experts. Furthermore, the official blog [3] of OpenAI (the company behind GPTs) cautions users that the model's outputs may not always be entirely accurate, since it is not an expert and does not take full responsibility for the provided content – only human experts can determine the relevance and accuracy of the output.

According to Tan *et al*. (2023), the GPT LLM family can struggle to generate relevant and accurate outputs for more complex tasks. However, not only GPTs but also many other LLMs exhibit significant challenges when compared on tasks of varying complexity levels that require reasoning abilities such as causal understanding, logical deduction, and counterfactual reasoning; these cognitive processes are complex and often associated with human-like consciousness, which LLMs do not possess (Gemini Team, 2023: 23; Lai *et al*., 2023). Even GPT-4, a more advanced version of the GPT LLM family, does not consistently show improved performance in tasks requiring high-level reasoning abilities. Consequently, the lack of the reasoning abilities is critical for the precise interpretation of user inputs. This limitation is evident when dealing with grammatical constructions in Russian, which often include idiomatic expressions that are not amenable to a logical, compositional analysis due to their complex pragmatic and cultural nuances (Janda *et al*., 2018). This issue is compounded by the nature of Russian grammatical constructions which often exhibit non-compositionality, i.e., the meaning of a construction cannot be interpreted solely from its individual components (Janda *et al*., 2020: 163; Rakhilina *et al*., 2022). Besides, some of these constructions are primarily idiomatic and colloquial and cannot be successfully interpreted based on

[3] https://help.openai.com/en/articles/6783457-what-is-chatgpt (21.03.2024).

statistical predictions alone. Therefore, the concept of grammatical constructions and their components is being elaborated further in more detail.

1.2. The concept of grammatical constructions and a constructicon

Grammatical constructions are defined as ''learned pairings of form and function, including words and idioms as well as phrasal linguistic patterns'' (Goldberg & Suttle, 2010: 469). Learned pairings of form and function include common and minor patterns. Examples of common patterns encompass passive, topicalization, and relative clauses; minor patterns cover words and idioms. While varying in their generality, size, and complexity, both patterns are still considered grammatical constructions. The main aspects of constructions primarily include their semantic attributes, information and discourse characteristics, as well as usage conditions (e.g., register and genre) (Fillmore *et al*., 1988; Goldberg, 2006; Goldberg & Suttle, 2010).

The linguistic theory that focuses on grammatical constructions is Construction Grammar (CxG). One of the main products that CxG provides is a *constructicon* as it supports the main idea of this theory that language is constructed according to a set of grammatical rules to convey meaning (Fillmore *et al*., 1988; Goldberg, 2006; Goldberg & Suttle, 2010; Hilpert, 2014; Lyngfelt *et al*., 2018; Janda *et al*., 2024). A constructicon is a large network of various grammatical constructions of a specific language, where the constructions are grouped into families, clusters, and networks according to their semantic and syntactic characteristics; grammatical constructions are typically demonstrated with examples from corpora and labeled according to the CEFR levels of language proficiency (Hilpert, 2014; Lyngfelt *et al*., 2018; Janda *et al*., 2020). In practice, a constructicon represents a distinctive dictionary of grammatical constructions containing thorough annotations. The annotations of constructions usually include definitions, valid lexical references from phraseological/idiomatic dictionaries, and grammar references explaining the syntactic/semantic types of the construction, as well as usage labels (Janda *et al*., 2020: 163).

The constructicon approach to constructions can be conceptualized as a complex grammar network model in which, according to Diessel (2019), different linguistic elements are linked and related. Within this theoretical model, constructions exhibit taxonomic and horizontal relations. Taxonomic relations are schemas in which components of constructions are generalized and categorized in order to be reused for the production of novel instances, while horizontal relations refer to how constructions are interconnected and used (Diessel, 2023). Thus, one of the types of grammatical constructions that can be represented in a constructicon contains a variable slot and a fixed part (anchor). The variable (open, non-fixed) slot can be filled with a set of lexemes (fillers) which represent words and phrases that are suitable for the grammatical construction (Janda *et al*., 2020). Meanwhile, the fixed part, which is partially schematic, has a strict structure (e.g., word order and/or specific words) that is duplicated precisely and remains unchanged due to lexical and semantic constraints (Janda *et al*., 2018; Bast *et al*., 2021). Although the variable slot is ''open'', it can still be subjected to certain lexical and semantic constraints. This idea is specifically implemented in *Russian Constructicon (RusCon),* a one-of-a-kind searchable database of grammatical constructions with detailed annotations (Bast *et al*., 2021):

ID1712 NP/VP *za komp'juterom*
Rabotat' *za* *komp'juterom*
[work.INF behind computer.INS]
'To work at the computer' (Bast *et al*., 2021)

This example demonstrates that noun phrase/verb phrase (NP/VP) can vary and instead of *rabotat'* 'to work', there are also various lexemes such as *sidet'* 'to sit', *provodit' vremja* 'to spend time', *zasypat'* 'to fall asleep', *žizn' prohodit* 'life passes', *prosiživat'* 'to sit out', etc. that can fill the slot of NP/VP, while *za komp'juterom* 'at the computer' is the fixed part of the construction and cannot be changed.

The RusCon database, a collection of 2,227 entries, was created through a collaboration between UiT The Arctic University of Norway, the University of Gothenburg, and the National Research University Higher School of Economics by professional linguists, researchers, BA, MA, and PhD students (Janda *et al*., 2018;

Bast *et al*., 2021). Each entry includes one grammatical construction, featuring a fixed part and a variable slot. Entries also include a unique ID number (Janda *et al*., 2024) that ranges from 1 to 2,227, a definition, examples (usually five additional examples featuring the same grammatical construction), CEFR levels (A1, A2, B1, B2, C1, C2), equivalents in other languages (e.g., translations of the grammatical construction into English), syntactic aspects (type, structure, function), dependency structure (the illustration of grammatical relationships between different parts of a sentence), communicative types, usage labels, extra comments, and scholarly references.

The 2,227 grammatical constructions of the RusCon are predominantly colloquial and underrepresented in written discourse, primarily manifesting idiomatic rather than schematic, transparent meanings (Janda *et al*., 2020; Rakhilina *et al*., 2022). Many of these constructions are not included in dictionaries or other reference works and have never been documented in scholarly literature before. The RusCon project attempted to primarily focus on high-frequency constructions. Therefore, phrases that potentially represent constructions but had fewer than 27 attestations in *the Russian National Corpus* were considered infrequent and usually excluded (Janda *et al*., 2018). However, since many of the constructions are non-compositional, they typically had to be collected, translated, and described manually, due to their opaque forms and meanings not yet being comprehensible to computational systems. Additionally, there is an issue with the poor comparability of phraseological units across languages (Rakhilina *et al*., 2022). Despite the existence of many bilingual dictionaries for major languages, bilingual phraseological dictionaries remain rare. Therefore, many items for RusCon were collected from idiomatic and phraseological dictionaries intended for native speakers; these dictionaries often lack detailed descriptions of the semantic and syntactic components of non-transparent expressions (Janda *et al*., 2018). Such descriptions are crucial for non-native speakers and L2 learners to fully grasp the rich pragmatic and cultural implications of idiomatic constructions. Thereby, according to Janda *et al*. (2018: 167), RusCon is an important resource in addressing this need from the complementary perspectives of non-native language users.

In light of the extensive scope, non-compositionality, and detailed annotations of the RusCon database, the current investigation aims to utilize its 2,227 entries to provide insights into GPT-4's performance, considering that AI chatbots are used for translation and interpretation in various languages. Thus, the findings may be applicable to other languages, particularly those that use the Cyrillic script, and can potentially interest linguists, L2 learners, and NLP developers.

2. Data and methodology

In order to test how well GPT-4 interprets Russian grammatical constructions, a dataset of 2,227 examples was gathered from Russian Constructicon. Each of the 2,227 entries in the resource includes a grammatical construction with its own ID number followed by an illustrative phrase, e.g., ID3 (*a*) *kak že* NP-Nom? – *A kak že mama*? 'And what about mom?'; (*a*) *kak že* is the fixed part, while NP-Nom? is the variable slot. The fixed part of grammatical constructions is more critical for the focus of this study, as it potentially indicates that similar issues could occur in other examples featuring the same grammatical construction. Therefore, the results where the issue was located elsewhere rather than in the fixed part are less relevant to this investigation and are not examined in more detail.

All 2,227 illustrative phrases were subjected to testing through the AI chatbot ChatUiT, which implements the most advanced model of the GPT LLM family at the time the experiment was conducted: GPT-4. Prior to each example the question in English ''What do you think of the following sentence?'' was asked, followed by automated processing of the dataset using the Python programming language. Although the phrasing of the question might seem informal and clumsy, it was deliberately chosen to simulate a conversational manner similar to everyday human interactions, representing how non-scientists potentially interact with AI chatbots. Additionally, the task was assigned in English because RusCon is designed primarily for non-native speakers with annotations for those who are proficient in English but might not be fluent enough to understand linguistic descriptions in Russian. Thereby, conducting the experiment in English helps to reveal the accuracy and depth of GPT-4's translations and interpretations of Russian grammatical constructions.

RusCon also provides CEFR levels of language proficiency (A1, A2, B1, B2, C1, C2) for each entry, allowing the data to be divided across these six levels.

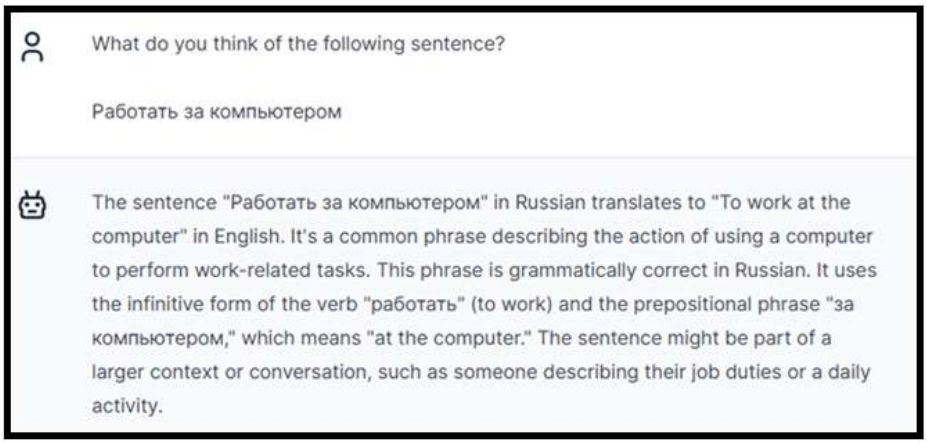
Most of the 2,227 illustrative phrases featuring a grammatical construction are accompanied by five additional examples in RusCon. The five additional examples of constructions that presented challenges for GPT-4 were also subjected to testing (more details in *Section 4*). The identical task was assigned to GPT-4, where the gathered data was automatically processed through Python, and the same question ''What do you think of the following sentence?'' was asked before each example.

Following the testing phase, I performed a manual evaluation of GPT-4 misinterpretations involving detailed comparative analyses with various elements of the annotations from RusCon. Further, data from the largest web corpora of Russian: *the Russian Web 2011 corpus (ruTenTen 2011), the Russian National Corpus*, and *Google Books*, which are incorporated into the training dataset of GPTs, are being employed in this investigation to ascertain whether low frequency contributes to the misinterpretation of certain examples. Additional resources, such as dictionaries and relevant scholarly articles, are also being employed, given that annotations from RusCon sometimes do not provide sufficient information that is relevant to the analysis of misinterpretations.

3. Results

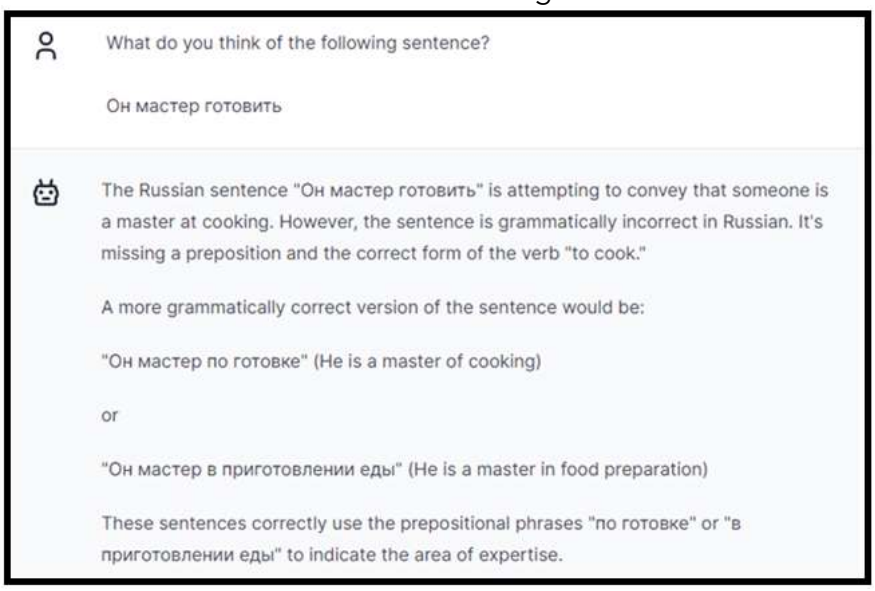
GPT-4 responses are predominantly heterogeneous across all 2,277 outputs, describing the meanings of illustrative phrases along with comments on their grammatical structures and contextual usages. Below, Figure 1 provides a GPT-4 response with a correct interpretation, while Figure 2 provides one with an incorrect interpretation.

FIGURE 1 – The response with a correct interpretation by GPT-4 in ChatUiT of the grammatical construction: ID1712 NP/VP *za komp'juterom* – *Rabotat' za komp'juterom* 'To work at the computer'



(Source: ChatUiT (2024), https://chat.uit.no/)

FIGURE 2 – The response with an incorrect interpretation by GPT-4 in ChatUiT of the grammatical construction: ID577 NP-Nom Cop NP-Nom VP-Inf – *On master gotovit'* 'He is a master at cooking'



(Source: ChatUiT (2024), https://chat.uit.no/)

Out of 2,277 tested examples, GPT-4 generally performs well but struggles with the correct interpretation of 18 examples (GPT-4 outputs in text file format for all 2,277 inputs are available online[4]). All the 2,277 examples from RusCon adhere to standard Russian language usage; therefore, the 18 outputs classified as misinterpretations are deviations from these established norms due to various

[4] Plotnikov, Timofei, 2024, "Replication Data for: Analyzing GPT-4 Misinterpretations of Russian Grammatical Constructions", https://doi.org/10.18710/8CAPJM, DataverseNO, V1

reasons, which are elaborated upon further. However, not all the 18 examples where GPT-4 struggles are due to misinterpretation of the fixed part of grammatical constructions. In some of the examples, there are other issues with the illustrative phrase, e.g., glitches, typos, and errors within the variable slot (a brief explanation of these misinterpretations is provided in *Section 4.2*.) While omitting the results that do not have issues with the fixed part of grammatical constructions, GPT-4 struggles with the correct interpretation of nine illustrative phrases across various language reference levels.

TABLE 1 – Incorrect interpretations of all examples versus incorrect interpretations of the examples with the fixed part of grammatical constructions assigned to the CEFR levels of language proficiency in RusCon

Language reference level	Number of grammatical constructions in RusCon	Overall misinterpretation of the examples	Misinterpretation of the examples with the fixed part of grammatical constructions
A1	86	0	0
A2	302	1	1
B1	654	5	1
B2	697	10	7
C1	428	1	0
C2	110	1	0
Percentage		*0.8%*	*0.4%*

(Source: Plotnikov, Timofei, 2024, "Replication Data for: Analyzing GPT-4 Misinterpretations of Russian Grammatical Constructions", https://doi.org/10.18710/8CAPJM, DataverseNO, V1)

While looking at the table, there were no issues with the interpretation of grammatical constructions within A1 level, just one inaccuracy within A2, one within B1, and seven within B2; there were additionally no issues within C1 and C2. Besides, the C2 level provides 110 construction examples, while B2 provides 697, which is nearly seven times more. Naturally, the larger dataset would also contain more misinterpretations in GPT's outputs, and that is what is observed between the B2 and C2 levels.

Since there were issues with the fixed part in nine grammatical constructions, five additional examples featuring the same grammatical construction were subsequently examined. However, only seven out of the nine grammatical constructions, where the issue occurred in the fixed part, included these five

additional examples. The initial assumption was that testing these extra examples could help draw more insightful conclusions regarding the misinterpretations of the nine grammatical constructions. Yet, these 35 additional examples are typically extensive and unambiguous, or they are a combination of two sentences, which leads to almost always correct and accurate interpretations due to the sufficient contextual information provided. Nevertheless, GPT-4 exhibits challenges in accurately interpreting just one additional example of the grammatical construction ID1652, which is discussed in more detail later

3.1. Misinterpretations of the fixed part

(1) ID131 NP-Nom Cop *zdorov* VP-Inf
On *zdorov* *vrat'*
[he.NOM healthy lie.INF]
'He is good at lying'

GPT-4 states ''the grammar and word choice are incorrect''[4]. However, GPT-4 identifies the meaning of the construction accurately but still insists that the word *zdorov* would not be used in the context of lying in standard Russian. Upon examination through Google Book Search, this example appears in eight attestations, including books and dictionaries; in all eight attestations, the construction is used in the context of lying. However, RusCon indicates that the construction is colloquial, signifying its rarity in formal written language. This results in insufficient training data with this grammatical construction in the context of lying, which is why GPT-4 struggles.

(2) ID217 *neznamo* PronInt
Neznamo *kak* *ja* *vernulsja* *domoj*
[unknown how I return.PST.M homeward]
'Without knowing how I came back home'

GPT-4 identifies the usage of the grammatical construction with the adverb *neznamo* as an error, suggesting the substitution with *neizvestno* instead. According to *Bol'šoj tolkovyj slovar' russkogo jazyka* (Kuznecov, 2000: 623), *neznamo* is colloquial rather than formal, and thus, it is not commonly used in written discourse. Furthermore, data from the Russian Web 2011 corpus reveals that *neznamo* appears in 3,292 attestations, while *neizvestno* is found in 366,697 attestations. The significantly lower frequency of *neznamo* explains why GPT-4 recognizes it as an error, due to the scarcity of data involving this word, which is rare in written texts.

(3) ID587 VP *po* NP-Dat.Pl
Oni ljubili často ezdit' po znakomym
[they.NOM love.PST.PL often visit.INF by acquaintances.DAT]
'They often liked to visit their acquaintances'

Although GPT-4 interprets the meaning of the example correctly, it classifies the example as ''somewhat ambiguous''[4] due to the use of the preposition *po* followed by the word *znakomym*, which can function as both a plural noun and an adjective in varying contexts. Consequently, ambiguity arises, since GPT-4 interprets *znakomym* as an adjective and thus it anticipates an additional word after it.

(4) ID603 NP-Pl *odin drugoj*-Gen Adj-Cmp
Devočki odna drugoj strojnee
[girls.NOM one.F another.DAT slimmer.COMP]
'The girls are each slimmer than the other'

GPT-4 claims that the example is ''grammatically incorrect and may cause confusion as to what exactly is being compared''[4]. However, additional examples in RusCon employing this grammatical construction incorporate either a dash or a comma before *odin drugoj*. Therefore, introducing a punctuation mark (a dash in this case in this particular example) could clarify the intended meaning by indicating that *devočki* is the subject of the sentence and the rest are related elements. Nevertheless, the absence of a dash is not a sufficient basis for GPT-4 to state the

grammatical structure entirely incorrect. Additionally, GPT-4 mentions that it is ''difficult understand the intended meaning without additional context''[4].

(5) ID1275 VP *vvidu* NP-Gen
On ne prišël vvidu bolezni
[he.NOM not come.PST.M due.to illness.GEN]
'He did not come due to illness'

GPT-4 suggests that the usage of *vvidu* is outdated and a more common way to use *iz-za* in this case, because *vvidu* is not used to indicate a cause or a reason in modern Russian and overall is considered too formal. However, according to *Slovar' trudnostej russkogo jazyka* (Rozental & Telenkova, 2005: 76-77) *vvidu* is used both to indicate a reason expected in the future, and to designate reasons that are timeless or related to the present or the past. Nevertheless, in the Russian Web 2011 corpus, *vvidu* can be found in 433,466 attestations, while *iz-za* appears in 3,987,006 attestations. The usage of *iz-za* is indeed more common, but *vvidu* is not outdated and still widely used in formal and more nuanced contexts. GPT-4's misinterpretation of this example stems from an overemphasis on the frequency of usage, which oversimplifies language nuances.

(6) ID1586 *kogda* XP, (***a***) *kogda* XP
Kogda vovremja pridët, kogda opozdaet
[when on.time come.3SG when late.arrive.3SG]
'Sometimes he/she/it arrives on time, sometimes he/she/it arrives late'

GPT-4 states that the structure of the example is incorrect. In the revised examples proposed by GPT-4, pronouns such as ''someone'' or ''you'' are incorporated, since grammatical patterns in English typically require a subject before a verb, which is not always necessary in Russian. Furthermore, GPT-4 removes the fixed part of the grammatical construction, thereby altering the intended meaning:

- *Kto-to pridët vovremja, a kto-to opozdaet* 'Someone will arrive on time, while someone else will be late'
- *Inogda prihodiš' vovremja, inogda opazdyvaeš'* 'Sometimes you arrive on time, sometimes you are late'

While attempting to convey the intended meaning, GPT-4 appears to impose an English syntactic structure, since the absence of a subject makes interpretation more confusing and complicated for GPT-4. In addition, GPT-4 once again indicates that it requires more context for a more accurate interpretation.

(7) ID1652 *ne segodnja zavtra* VP

Ne	*segodnja*	*zavtra*	*my*	*vsë*	*uznaem*
[not	today	tomorrow	we	everything	learn.1PL]

'One of the next few days we will found out everything'

GPT-4 claims that the example is incorrect. Upon examination, it appears that a more prevalent usage of this construction involves inserting a dash between *segodnja* and *zavtra* (i.e., *ne segodnja-zavtra*). For instance, the Russian Web 2011 corpus provides only 883 attestations for *ne segodnja zavtra* and 3,157 attestations for *ne segodnja-zavtra* when the dash is employed. However, RusCon states that the spelling employing a dash does not correspond to the modern spelling norm, despite the higher frequency in the corpus. As a correct example, GPT-4 recommends separating *ne segodnja* and *zavtra* with a comma (i.e., *ne segodnja, zavtra*), which can be accurate in some contexts but not in this specific case, as *ne segodnja zavtra* is an idiomatic phrase with its own distinct pragmatic meaning. According to RusCon, *ne segodnja zavtra* means that the action will take place in the next few days without an exact indication of a specific day, while *ne segodnja, zavtra* literally means 'not today, but tomorrow', which clearly indicates that the action will take place tomorrow.

In one of the five additional examples featuring this grammatical construction in RusCon *(k s"emkam fil'ma Petrov nameren byl pristupit' ne segodnja zavtra),* GPT-4 also suggests employing a dash or a comma in *ne segodnja zavtra*, which indicates that GPT-4's interpretation is influenced by frequency.

(8) ID1762 *čto-to nezametno, čto* Cl

Čto-to nezametno, čto ty toropiš'sja

[something not.noticeable that you.NOM hurry.2SG]

'It is not noticeable that you're in a hurry'

GPT-4 suggests that the current example is nonsensical. One of the revised examples proposed by GPT-4 is *kažetsja, ty nezametno toropiš'sja*, which removes the fixed part of the grammatical construction; the literal translation 'it seems you are hurrying unnoticeably' does not convey the intended meaning, indicating that GPT-4 has misinterpreted it. However, GPT-4 states that the example is not used in formal written discourse, and RusCon also indicates that the current construction's usage is colloquial. Additionally, the Russian National Corpus provides just three attestations with the fixed part of this grammatical construction, i.e., *čto-to nezametno, čto*, while ruTenTen 2011 does not provide any at all. Therefore, GPT-4 lacks sufficient data to more accurately interpret this grammatical construction.

(9) ID2272 *kakoj/kakoe* (*tam*) Adv/Adj!

Kakoe tam bystro!

[what.NOM there fast]

'Far from fast!'

According to RusCon, the grammatical construction in this example is colloquial and expresses disagreement. However, GPT-4 completely misinterprets it, claiming that this example is an attempt to express that something is fast, which is incorrect. The construction with the same adjective appears in 18 attestations across Google Books, which indicates insufficient frequency of usage. Besides, GPT-4 states that a better response is not possible without additional context or details.

3.2. Misinterpretations of other results

GPT-4 does not seem to encounter challenges in properly interpreting the fixed parts of the following grammatical constructions, as it predominantly struggles with the variable slots:

(10) ID280 NP-Nom *prinjat'* NP-Acc *za* NP-Acc
On prinjal togo mužčinu za svoego otca
[he.NOM took.PST that.ACC man.ACC for own.GEN father.ACC]
He mistook that man for his father

(11) ID1562 VP *vo ves' opor/vo vsju pryt'/vo ves' duh*
On bežal vo ves' opor
[he.NOM ran.PST in all strength]
'He ran at full speed'

(12) ID2040 *ot* NumCrd-Gen *otnjat'* NumCrd-Acc *ravno*/VP NumCrd-Nom/NumCrd-Dat
Ot vos'mi otnjat' tri budet pjat'
[from eight.GEN subtract.INF three be.FUT five]
'Subtracting three from eight equals five'

In all three cases, GPT-4 identifies the sentence structure as a mistake, yet it consistently generates identical outputs. This appears to be a glitch in GPT-4.

(13) ID356 *čem by* VP-Inf, VP (*by*)
Čem by učit'sja, on guljaet
[what would learn.INF he.NOM walk.3SG]
'Instead of studying, he is walking around'

In this example, GPT-4 actually encounters difficulties with the grammatical construction and claims that the example is incorrect because it is not standard

Russian. However, according to RusCon, the construction is classified as obsolete, indicating that it is no longer prevalent in contemporary standard Russian. Therefore, GPT-4's response can be considered partially correct, as it reflects contemporary Russian language usage.

(14) ID577 NP-Nom Cop NP-Nom VP-Inf
On *master* *gotovit'*
[he.NOM master cook.INF]
'He is a master at cooking'

GPT-4 states that this example is not correct, which is not the case. RusCon characterizes this grammatical construction as colloquial, which again reassumes infrequent occurrence in the training data. As a result, this particular example lacks attestations in both the Russian Web 2011 corpus and the Russian National Corpus, but there were ten attestations in Google Books, confirming the scarcity of data available online for GPT-4 to more accurately interpret this example.

(15) ID609 PronPers-Dat NP-Dat Cop Pred
Mne *ušam* *holodno*
[to.me ears.DAT cold]
'My ears are cold'

GPT-4 claims that this example is grammatically incorrect and subsequently suggests an alternative, *mne holodno uši*, which is also incorrect. Despite the grammatical accuracy of the original example, it lacks attestations in corpora and Google Books, indicating its low frequency. Considering the infrequent occurrence of the example, GPT-4 proposes an incorrect version as an alternative, which appears to be a phenomenon of AI *hallucinations*[5].

[5] *AI Hallucinations* are false responses generated by an LLM that may seem correct but are actually nonsensical at all (Rawte et al., 2023).

(16) ID808 VP *v rezul'tate* NP-Gen

Neskol'ko čelovek postradalo v rezul'tate vzryva

[several persons suffer.PST.PL in result.GEN explosion.GEN]

'Several people were injured as a result of the explosion'

GPT-4 states that *neskol'ko čelovek* is plural and that the verb should correspond in number with the subject. However, in Russian, verb-subject agreement in number can vary depending on the context and the speakers' intentions (Nesset & Janda, 2023).

(17) ID1162 Adv-Cmp *vsego* VP

Èti busy podhodjat k plat'ju lučše vsego

[these beads suit.3PL to dress better of.all]

'These beads match the dress best'

(18) ID2084 *bez* NumCrd-Gen (*minut*) NumCrd-Nom

Prihodi k bez pjatnadcati desjat'

[come.IMP to without fifteen ten]

'Come by fifteen to ten'

Even though GPT-4 misinterprets these examples, it does not claim that these examples are entirely incorrect. It suggests that the incorporation of the preposition *k* might be slightly ungrammatical, as this preposition is not typically required after the verbs in these examples.

4. Discussion

The analysis indicates that GPT-4 demonstrates a high level of proficiency in handling Russian grammatical constructions. Among 2,227 constructions, GPT-4 struggled with 18 constructions, representing less than 1% of the entire dataset (0.4% referring to the misinterpretation of the fixed part of grammatical

constructions). Consequently, the response to the first question in the introduction is that GPT-4 demonstrates effective performance.

The additional inquiry concerned factors causing GPT-4 misinterpretations. Overall, GPT-4 tends to struggle with grammatical constructions when there is insufficient and infrequent data available (8 out of 18 examples), which evidences the frequency bias mentioned in the studies by Dudy & Bedrick (2020) and Brown *et al*. (2020). For other misinterpretations (4 out of 18), GPT-4 required more context necessary for accurate input interpretation, since GPTs can struggle to perform effectively due to contextual deficits (Gong *et al*., 2019; Liu *et al*., 2021; Yenduri *et al*., 2024). Predominantly, the linguistic structures that challenge GPT-4 are those that are colloquial and infrequently represented in formal written discourse, thereby contributing to a scarcity of documented data available on the Internet.

Further, GPT-4 struggles with certain grammar patterns. It states that the grammar is incorrect in examples (1), (5), (16) and produces ''correct'' suggestions based on the frequency of the dataset rather than on grammatical accuracy. In example (6), GPT-4 appears to incorporate English syntactic structure while attempting to propose revised examples in Russian. GPT-4 also alters the original lexicon of the example (6), removing the fixed part of the grammatical construction; the same phenomenon is observed in example (8). This aligns with the findings of Papadimitriou *et al*. (2022) and Wendler *et al*. (2024) regarding the influence of the English language on the final output. In another example (7), GPT-4 inserts unnecessary punctuation marks in the output, confirming that LLMs can introduce irrelevant linguistic elements (Hendy *et al*., 2023).

Except for the incorrect punctuation insertion, the example (7) has an idiomatic meaning, which GPT-4 cannot identify correctly. In example (9), GPT-4 also interprets the meaning literally rather than figuratively. These findings show that GPTs can exhibit limitations to interpret idiomatic expressions and non-literal language (Ray, 2023; Yenduri *et al*., 2024). Furthermore, these examples demonstrate that LLMs encounter difficulties in representing cultural values in languages other than English (Johnson *et al*., 2022; Jackson, 2023; Liu, 2024), considering that idiomatic expressions are not merely linguistics items but also carries of cultural significance.

In example (15), GPT-4 claims that the example is incorrect and attempts to provide an alternative, which is neither sensible nor standard Russian. This indicates that GPTs struggle to provide an adequate result when the initial input contains infrequent and rare words (Brown *et al*., 2020; Dudy & Bedrick, 2020), potentially leading to AI hallucinations (Rawte *et al*., 2023).

Some GPT-4 outputs with errors were not entirely wrong. GPT-4 claims that example (13) is not correct, but the grammatical construction is indeed archaic and not used in modern Russian anymore. Examples (17) and (18) were considered slightly incorrect due to a preposition, which is rather misleading as these examples are grammatically correct but are less frequently used in the Russian language. Therefore, GPT-4 suggests avoiding the preposition in those examples prioritizing word frequency over actual accuracy (Dudy & Bedrick, 2020).

Although GPT-4 demonstrates proficient performance across the majority of the dataset, some challenges remain, indicating that AI chatbots are not yet capable of completely substituting for human expertise. Researchers, teachers, and learners, who are non-native speakers, cannot fully rely on LLMs, since biased and inaccurate outputs can convey wrong linguistic ideas and might seem correct. To avoid such issues, expertise from human professionals is still essential. However, one solution for achieving better outputs is to provide more context. My experiment particularly demonstrates this in the testing of five additional examples of grammatical constructions where issues occurred in the fixed part. The results were mostly appropriate due to the sufficient context provided.

Regarding LLM development, it might be hard to fix the frequency bias at the present stage, but one of the solutions is to have more balanced training data. When the data is predominantly in English and data in Russian, which is one of the most widely spoken languages, represents only 0.2% in the main text corpus implemented in GPTs, the situation is less than ideal and needs attention from LLM developers. Although GPT-4 interprets Russian grammatical constructions quite well, its performance might be significantly worse with less commonly spoken languages that have structures vastly different from English.

5. Limitations and future directions

In the present study, the experiment was conducted using the AI chatbot ChatUiT, which functions similarly to ChatGPT but is not available to the general public. However, the results are generalizable because ChatUiT implements the foundational model GPT-4, the same one that is being employed in ChatGPT at the time the experiment was conducted.

Additionally, a zero-shot experiment was conducted in this study. While examples (10), (11), and (12) appear to be products of AI hallucinations, GPTs' responses may potentially vary if semantically/syntactically similar examples are provided in advance before conducting an analogous experiment. Thus, the ''problematic'' dataset identified in this investigation can also be tested with such examples.

There were also a few limitations when comparing GPT-4 interpretations with annotations from RusCon. Some entries had minor errors, occasional typos, or lacked sufficient information, since the RusCon project was completed manually rather than computationally. Furthermore, the dataset in RusCon is not large; additional constructions could potentially reveal more inaccuracies in areas other than those already mentioned.

Overall, GPT models from OpenAI are not the only large language models available on the market. There are other prominent ones, such as Gemini from Google and Claude from Anthropic. The methods of data training and the settings differ among these LLMs. Therefore, it could be interesting to see and compare how various LLMs perform on the same task.

References

Bast, R., Endresen, A., Janda, L. A., Lund, M., Lyashevskaya, O., Mordashova, D., Nesset, T., Rakhilina, E., Tyers, F. M., & Zhukova, V. (2021). *The Russian Constructicon. An Electronic Database of the Russian Grammatical Constructions*. https://constructicon.github.io/russian/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C.,

McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *ArXiv abs/2005.14165*. Web.

Diessel, H. (2019). *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*. Cambridge University Press.

Diessel, H. (2023). *The Constructicon: Taxonomies and Networks*. Cambridge University Press.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1286-1305). https://doi.org/10.18653/v1/2021.emnlp-main.98

Dudy, S., & Bedrick, S. (2020). Are some words worth more than others? *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 131-142). https://doi.org/10.18653/v1/2020.eval4nlp-1.13

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language, 64*(3), 501-538.

Gemini Team. Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., *et al*. (2023). Gemini: A family of highly capable multimodal models. *ArXiv preprint arXiv:2312.11805*.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Goldberg, A., & Suttle, L. (2010). Construction grammar. *WIREs Cognitive Science, 1*(4), 468-477. https://doi.org/10.1002/wcs.22

Gong, X.-R., Jin, J.-X., & Zhang, T. (2019). Sentiment analysis using autoregressive language modeling and broad learning system. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1130-1134).

Google Books. (n.d.). Retrieved July 15, 2024, from https://books.google.com/

Hendy, A., Abdelrehim, M. G., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M. A., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *ArXiv abs/2302.09210*.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh University Press.

Jackson, S., Beekhuizen, B., Zhao, Z., Zhao, Y. C., & McEwen, R. N. (2023). LLMs and linguistic competency: An exploration of GPT-4 and a non-hegemonic English variety. *Newhouse Impact Journal, 1*(1), 21-23. http://doi.org/10.14305/jn.29960819.2024.1.1.04

Janda, L. A., Endresen, A., Zhukova, V., Mordashova, D., & Rakhilina, E. (2020). How to build a constructicon in five years: The Russian example. *Belgian Journal of Linguistics, 34*, 162-175. https://doi.org/10.1075/bjl.00043.jan

Janda, L. A., Lyashevskaya, O., Nesset, T., Rakhilina, E., & Tyers, F. M. (2018). Chapter 6. A constructicon for Russian: Filling in the gaps. In B. Lyngfelt, L. Borin, K. Ohara, & T. T. Torrent (Eds.), *Constructicography: Constructicon development across languages* (pp. 165-181). John Benjamins Publishing Co. https://doi.org/10.1075/cal.22.06jan

Janda, L. A., Zhukova, V., & Endresen, A. (2024). Typology of reduplication in Russian: Constructions within and beyond a single clause. In M. Kopotev & K.

Kwon (Eds.), *Constructions with lexical repetitions in East Slavic* (pp. 71-96). De Gruyter Mouton. https://doi.org/10.1515/9783111165806-003

Johnson, R. L., Pistilli, G., Menédez-González, N., Dias Duran, L. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The ghost in the machine has an American accent: Value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*.

Kuznecov, S. A. (Ed.). (2000). *Bol'šoj tolkovyj slovar' russkogo jazyka*. Bukinist.

Lai, V. D., Ngo, N. T., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). Chatgpt beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Liu, C. (2024). The investigation of application related to ChatGPT in foreign language learning. *Applied and Computational Engineering, 35*, 110-115. https://doi.org/10.54254/2755-2721/35/20230376

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What makes good in-context examples for GPT-3? Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out. *arXiv preprint arXiv:2101.06804*.

Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018). *Constructicography: Constructicon development across languages*. John Benjamins. https://doi.org/10.1075/cal.22

National Corpus of the Russian Language. (2024). Retrieved May 12, 2024, from https://ruscorpora.ru/en/

Nesset, T., & Janda, L. A. (2023). A network of allostructions: Quantified subject constructions in Russian. *Cognitive Linguistics, 34*(1), 67-97. https://doi.org/10.1515/cog-2021-0117

Pannatier, A., Courdier, E., & Fleuret, F. (2024). GPTs: A new approach to autoregressive models. *arXiv preprint arXiv:2404.09562*.

Papadimitriou, I., Lopez, K., & Jurafsky, D. (2022). Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. *arXiv abs/2210.05619*.

Rakhilina, E., Zhukova, V., Demidova, D., Kudrjavceva, P., Rozovskaja, G., Endresen, A., & Janda, L. A. (2022). Frazeologija v rakurse Russkogo Konstruktikona [Phraseology in the light of the Russian Constructicon]. *Bulletin of the Russian Academy of Sciences: Studies in Literature and Language, 2*(32), 13-44. https://doi.org/10.31912/pvrli-2022.2.2

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M., Chadha, A., Sheth, A. P., & Das, A. (2023). The troubling emergence of hallucination in large language models - An extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rozental', D.È., & Telenkova, M.A. (2005). *Slovar' trudnostej russkogo jazyka*. Ajris-press.

Sketch Engine - Russian Web 2011 corpus (ruTenTen 2011). (2024). Retrieved May 12, 2024, from https://app.sketchengine.eu

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., & Qi, G. (2023). Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family. *International Semantic Web Conference* (pp. 348-367). Cham: Springer Nature Switzerland.

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do llamas work in English? On the latent language of multilingual transformers. *arXiv abs/2402.10588*, 1-29. https://arxiv.org/abs/2402.10588

Yenduri, G., Murugan, R., Govardanan, C., Supriya, Y., Srivastava, G., Reddy, P., Raj, D., Jhaveri, R., B, P., Wang, W., Vasilakos, A., & Gadekallu, T. (2024). GPT (Generative Pre-Trained Transformer) – A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access, 12*, 54608-54649. https://doi.org/10.1109/ACCESS.2024.3389497

Zhu, Q., & Luo, J. (2022). Generative pre-trained transformer for design concept generation: An exploration. *Proceedings of the Design Society, 2*, 1825-1834. https://doi.org/10.1017/pds.2022.185