Research article
Submitted: 5 June 2024
Accepted: 9 July 2024

TESTING CHATGPT'S CAPABILITIES AS AN ENGLISH-CROATIAN MACHINE TRANSLATION SYSTEM IN A REAL-WORLD SETTING: ETRANSLATION VERSUS CHATGPT AT THE EUROPEAN CENTRAL BANK

Krešimir Jozić

University of Zagreb

Abstract

The latest innovations in the field of natural language processing are large language models such as ChatGPT, which can perform various language tasks, including machine translation (MT). This study explores ChatGPT's performance as an English-Croatian MT system in a realistic professional setting. An excerpt from a European Central Bank (ECB) document was translated with ChatGPT and eTranslation, the more conventional neural MT system used by ECB translators. A human evaluation, an automatic evaluation and an error analysis were performed in order to determine whether ChatGPT would be more useful than eTranslation for the ECB's Croatian translation unit, and whether ChatGPT is better at dealing with context-related issues. The findings suggest that ChatGPT would presently not be as useful as eTranslation for English-Croatian MT at the ECB. Its performance in terms of context-related issues is still unclear. Future research could further explore this topic, as the use of ChatGPT for English-Croatian MT remains underexplored.

Keywords: neural machine translation, large language model, ChatGPT, eTranslation, English-Croatian translation, context-related issues

1. Introduction

Since the creation of the digital computer, researchers have attempted to use them for automatic translation (Hutchins 2006, 376). Over the years, many new machine translation (MT) systems have been developed, using a variety of architectures and approaching the translation problem from different perspectives (Koehn 2020, 33-40). New machine translation systems were often accompanied by great enthusiasm and promises of human-quality translation, yet they mostly failed to live up to these unrealistic expectations (Koehn 2020, 29-30). Currently, transformer-based large language models such as ChatGPT are garnering much attention due to their remarkable abilities in performing a variety of language-related tasks, including text summarization, creative writing and machine translation (Hughes 2023). The aim of the present study is to explore the capabilities and usefulness of such large language models, specifically ChatGPT, as an English-Croatian machine translation tool in a professional setting, and to compare its performance with a state-of-the-art neural machine translation system. For this purpose, I was able to use my traineeship at the European Central Bank (ECB): I asked translators with experience working in the ECB's Croatian translation unit, and access to ECB's in-house resources, to help me evaluate how useful ChatGPT's translations would be for the Croatian unit's daily work, and to compare ChatGPT's translations with those produced by eTranslation, the more conventional neural machine translation system regularly used by translators at the ECB and other EU institutions.

In the first section of this paper, titled Background, the history of approaches to machine translation is briefly recounted, highlighting the issues and innovations that each subsequent model brought. This section is divided into two subsections, covering approaches to machine translation before and after the rise of neural machine translation, as cutting-edge MT models are currently based on neural networks. A brief subsection describing how machine translation is generally used in the Croatian unit is also included in the Background section. In the Aims and hypotheses section, I formulate the exact research questions and expectations I had regarding how ChatGPT and eTranslation would perform in the tasks. The

Methodology section explains how the evaluation of the two machine translation systems was performed. The results of this study are then presented in the Results section. Finally, the Discussion section reflects on the results and compares them against initial expectations.

2. Background

2.1 Approaches to machine translation before the neural turn

Different authors date the origins of machine translation to different times, some dating it as far back as the 17th or even the 9th century, as this was when the foundations were laid for the cryptographic and mathematical methods that would later be used for machine translation (DuPont 2018; Hutchins 2006, 375). The first patents for translation machines were filed in the 1930s, though these were relatively crude machines that used paper bands or belts with perforations as their memories (Hutchins 2004, 1-7). It was a decade later, in the mid-1940s, when scientists first started discussing the potential use of the newly invented digital computer for automatic translation (Hutchins 2006, 376). In 1949, machine translation pioneer Warren Weaver published a memorandum which proved influential and sparked interest in machine translation research in the United States (Koehn 2020, 34; Hutchins 2006, 376). After a public demonstration of a Russian-English machine translator in New York in 1954, the doors were opened to large-scale funding in the USA; many other countries were inspired to also start developing MT systems (Hutchins 2006, 376).

From this point until the 1980s, most machine translation systems developed in the United States and around the world employed various methods sometimes collectively referred to as **rule-based approaches** (Hutchins 2006, 376; Stein 2016, 9). MT systems using rule-based approaches translate from one language into another by following a pre-determined set of rules written by linguists, relating to morphology, syntax, semantics or other linguistic properties of the source and target languages (Liu and Zhang 2015, 111-2). Though rule-based models were the standard in MT for a long time, they have many flaws: writing the rules is labor and

time intensive, covering the wide variety of linguistic phenomena in any given language is extremely difficult, and the quality of the translations often leaves much to be desired (Liu and Zhang 2015, 111; Stein 2016, 10). Finally, though it should theoretically be possible to write rules complex enough to consistently produce high-quality translations, in practice, adding more complexity only increases translation quality up to a certain point (Stein 2016, 10). This is due to the fact that new rules often contradict old ones, thereby producing new errors that need to be accounted for (Liu and Zhang 2015, 111; Stein 2016, 10).

It was only in the late 1980s that another approach to MT started to challenge the dominance of rule-based approaches: these were the **corpus-based** or **data-driven methods** (Koehn 2020, 36; Hutchins 2006, 380). These models utilize data from large text corpora to predict the most likely translation of the source text, rather than relying on prewritten rules (Stein 2016, 10-1). The most impactful data-driven method was the **statistical method,** which became the dominant MT paradigm from the 1990s to mid-2010s (Koehn 2020, 39-40). Statistical machine translation (SMT) systems usually employ two separate corpora: the first corpus contains large amounts of aligned source language and target language sentences and aims to represent all possible translations of a source sentence (Stein 2016, 11). The second corpus only contains target language sentences and aims to represent all valid sentences in the target language (ibid.). By recognizing patterns found in both corpora, statistical machine translation systems can predict what the most likely translation of a source sentence would be (ibid.). Statistical methods offer some advantages over rule-based methods: as no linguistic rules must be written manually, statistical systems generally require less labor force to create, which can be especially useful when creating MT systems for smaller languages (Stein 2016, 13). However, the quality of SMT output can vary: statistical systems generally fare better than rule-based systems when it comes to word choice and disambiguation, though they tend to underperform in other areas, such as word order or syntax (Stein 2016, 14).

2.2 Neural turn and beyond

The artificial neural network model was first proposed as early as the 1950s (Koehn 2020, 31), but it was only in the mid-2010s that **neural machine translation (NMT),** another data-driven MT method, took over as the dominant method in what is now often referred to as the neural turn (Koehn 2020, 39-40; Forcada 2017, 2). To illustrate how quickly NMT went from one of the competing methods to the new standard in MT, Koehn (2020, 40) compares NMT and SMT submissions at the Conference on Machine Translation, a major event for MT researchers: in 2015, only one pure neural machine translation system was submitted; in 2016, an NMT system won in almost all categories; in 2017, the vast majority of submissions were NMT systems. To this day, neural systems remain the standard in machine translation and natural language processing: major online MT services, such as Google Translate, DeepL and Microsoft Translator, use NMT (Wu et al. 2016, 1; Microsoft, n.d.; DeepL 2021.) and the most recent innovations, such as generative large language models, are also based on neural networks (IBM, n.d. a).

2.2.1 Artificial neural networks

NMT systems are in some ways similar to SMT systems: they analyze patterns in large bilingual corpora in order to predict the most likely translation of a source text (Forcada 2017, 2; Pérez-Ortiz et al. 2022, 148). What differentiates NMT systems from their statistical predecessors is their complex architecture based on artificial neural networks (Forcada 2017, 2). In this section, I will attempt to briefly summarize the most important features of artificial neural networks, including weights, word-embeddings and the attention mechanism. A more detailed explanation of neural network architecture is given in Pérez-Ortiz et al. (2022), which also served as the basis for this section.

As their name suggests, artificial neural networks are inspired by the structure of the natural neural networks in the human brain, which consist of interconnected cells called neurons (Pérez-Ortiz et al. 2022, 142-3). In an artificial neural network, a "neuron" can be thought of as a unit which can be **excited** or **inhibited** to a certain degree (Pérez-Ortiz et al. 2022, 143-4). This is called the **activation state** (ibid.). The activation state of a neuron depends on the signals it receives from the

neurons connected to it, and the strengths of those connections, which are represented by numbers called **weights** (ibid.).

Large numbers of such artificial neurons can be connected to form layered networks, which are capable of solving various computational tasks, including translation (Pérez-Ortiz et al. 2022, 146-7). The first layer is called the **input layer** and is made up of neurons that only take in stimuli from outside the neural network (ibid.). These stimuli represent the problem that the network needs to solve, e. g. a source text that the network needs to translate into the target language (Pérez-Ortiz et al. 2022, 146-8). Next, the input neurons stimulate the neurons in the second layer, called the **hidden layer** (Pérez-Ortiz et al. 2022, 146). Modern neural networks usually have many connected hidden layers, as networks with more layers are generally capable of solving more complex tasks (Pérez-Ortiz et al. 2022, 147). The final layer in the network is called the **output layer** (Pérez-Ortiz et al. 2022, 146). The activation states of the neurons in this layer represent the solution of the problem that was introduced to the neural network at the start – in the case of NMT systems, this is the target text (Pérez-Ortiz et al. 2022, 146-8). The layered structure is displayed in Figure 1, taken from IBM (n. d. b., sic.).

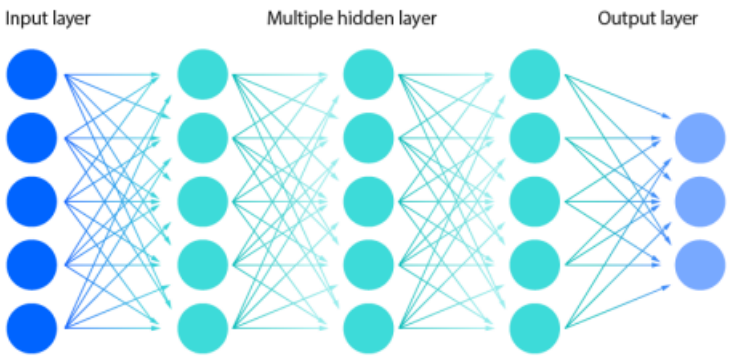

Figure 1. Architecture of a neural network with multiple hidden layers (IBM., n.d.b).

In order for artificial neural networks to perform their tasks successfully, they need to be "trained" using large, labeled corpora (Pérez-Ortiz et al. 2022, 148). The training process involves, among other things, adjusting the value of the aforementioned weights (ibid.). The weight-training process starts with random

weights or weights taken from a neural network previously trained for similar tasks (ibid.). When the neural network produces an output, a training algorithm compares that output against the training examples from the corpus and slightly updates the weights (ibid.). This process is repeated until the difference between the neural network's output and the example data is small enough (ibid.). Some models can also be trained on unlabeled corpora, which is referred to as unsupervised learning (IBM n. d. d., Pérez-Ortiz et al. 2022, 162).

The words that go through these neuron layers are represented numerically, more precisely by a series of numbers called **word embeddings** (Pérez-Ortiz et al. 2022, 150). These numbers capture semantic information about individual words (Pérez-Ortiz et al. 2022, 152). We can imagine the numbers as coordinates in a coordinate system: words with similar or related meanings are placed closer to each other in that coordinate system, as depicted below in Figure 2, taken from Pérez-Ortiz (2022, 153).

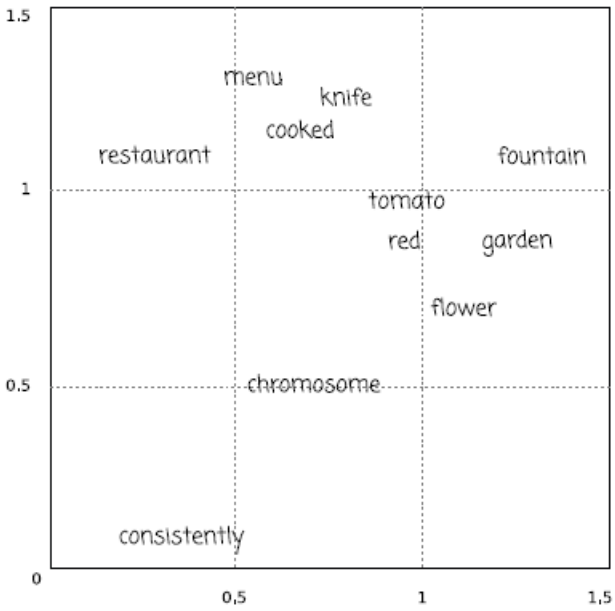

Figure 2. Word embeddings imagined as coordinates in a two-dimensional grid (Pérez-Ortiz 2022, 150-3).

In an NMT system, however, the word embeddings do not only have two coordinates, but potentially up to hundreds (Pérez-Ortiz 2022, 151). In this way,

word embeddings can capture many different shades of meaning, as words can be placed closer together or further away from each other in different dimensions for different reasons (Pérez-Ortiz 2022, 152). The word embeddings can, similarly to weights, be trained (Pérez-Ortiz 2022, 151).

One of the problems is that an embedding such as described can only capture one meaning of a word, independent of the context (Pérez-Ortiz et al. 2022, 153). This is why the kinds of embeddings described above are called **non-contextual word embeddings** (ibid.). To solve this problem, there needs to be a word embedding that also takes into account the context in which the word appears (ibid.). These are called **contextual word embeddings** and can be produced by applying a special mechanism called **attention** to the non-contextual embeddings (ibid.).

Neural machine translation systems quickly showed themselves to offer many quality improvements over their statistical predecessors: studies conducted as early as 2016 showed improved fluency, i. e. NMT systems produced results that read better and lacked overt mistakes, regardless of translation accuracy (Forcada 2017, 13). Automatic evaluation also judged the overall quality of NMT to be better than SMT for certain language pairs (ibid.). Human evaluators concluded that post-editing effort, i. e. the effort it takes to bring MT output to a publishable quality was reduced with NMT systems for many language pairs (ibid.).

However, NMT is not perfect. When translating longer stretches of text, these systems struggle with issues that Castilho (2022, 3018) calls "context-related issues": since they cannot take into account the entire text simultaneously, they may, for instance, misidentify the referent of a pronoun if the referent was named much earlier in the text, or mistranslate a polysemous word because the context necessary to correctly interpret the word was provided earlier in the text. NMT systems are also difficult to train, as they require very large corpora, and the training process requires much computational power (Forcada 2017, 5; Pérez-Ortiz 2022, 148).

2.2.2 Transformer architecture, large language models and ChatGPT

The transformer is a variant of the neural network introduced in 2017 by a group of scientists at Google (Vaswani et al. 2017). This new model was cheaper and easier to train than older neural network models, such as recurrent neural networks (IBM, n.d. c.). In addition to NMT, transformer architecture was used to develop **large language models** (LLMs), systems that can not only translate texts into many languages, but also perform various language tasks, such as summarization or even creative writing (IBM, n. d. a). In 2022, the company OpenAI launched an LLM-based chatbot named **ChatGPT** (short for "generative pre-trained transformer") (OpenAI 2022). It quickly became a sensation, garnering the interest of companies and the general public due to its remarkable capabilities (Weise et al. 2023). The user can communicate with the chatbot in simple, conversational language and ask it to perform various language tasks in multiple languages (OpenAI 2022). The 3.5 version of the chatbot has a context window of 8000 tokens, meaning it can take into account contextual information from conversations as long as 8000 words (OpenAI, n.d.). One might hypothesize that this would make ChatGPT less prone to context-related errors than more conventional NMT systems.

Some research suggests the quality of ChatGPT's translations is close to or on a par with more conventional NMT systems, at least when dealing with high-resource languages, but that the quality of ChatGPT's translations is lower for lower-resource languages (Peng et al. 2023, 1). As for ChatGPT's performance in English-Croatian translation, research has been relatively sparse thus far. Using the automatic BLEU metric, Omazić and Šoštarić (2023, 78-80) found that, when translating legal texts from English into Croatian, ChatGPT performed slightly worse than other commercial (neural) MT systems such as Google Translate and Microsoft Translate.

2.3 Neural machine translation at the European Central Bank

As previously mentioned, neural networks became the dominant paradigm in machine translation in the mid-2010s. It was during this time that the European Commission also decided to replace its proprietary statistical machine translation system MT@EC with a new neural system called **eTranslation** (European Commission, n.d.). Various European institutions currently use this MT system,

including the European Central Bank (Villani 2021, 11). eTranslation has multiple specialized engines adapted to different domains: EU Formal Language, General Text, Finance etc. (European Commission, n.d.) The finance engine is trained on various texts from the economic domain, including texts translated by the ECB's Croatian unit.

At the Croatian unit in the European Central Bank, eTranslation's finance engine is regularly used in the environment of the computer-assisted translation tool Trados Studio, which uses ECB-specific translation memories. These tools help with productivity, though substantial post-editing of MT output is usually necessary for producing translations of publishable quality. The Croatian translation unit translates various texts from finance and banking, both for an expert audience and a more general audience. Many translations are published on the ECB's official website (European Central Bank, n.d. a). To my knowledge, no LLM was used at the ECB while this research was conducted.

3. Aims and hypotheses

The main aim of this study is to evaluate the usefulness of ChatGPT, a general-purpose large language model, for machine translation in a practical setting, and compare it to a more conventional, but specialized neural machine translation system, in this case eTranslation. An additional aim is to explore if ChatGPT makes fewer context-related errors than more conventional NMT systems.

Two research questions were formulated:

- Is ChatGPT more useful for translators at the European Central Bank than eTranslation, i. e. does it produce a better output, possibly reducing post-editing time and effort?
- Does ChatGPT fare better than eTranslation in terms of context-related issues?

I hypothesized that eTranslation's finance engine might deliver a translation more in line with ECB-specific style and domain-specific terminology, as it is trained on financial documents from EU institutions, including the ECB. On the other hand,

I hypothesized that ChatGPT might perform better when it comes to context-related issues such as reference, number, gender, etc., due to its more advanced architecture and its nature as an LLM chatbot.

4. Methodology

4.1 Choice of text and production of machine translations

For the source text, I chose four paragraphs from the September 2023 ECB staff macroeconomic projections for the euro area published on the website of the European Central Bank (European Central Bank 2023). The ECB's macroeconomic projections contain various terminology relating to inflation, economic growth, wages, etc., and are published four times per year (European Central Bank, n.d., c). In accordance with the language policy of the ECB's website, the projections are translated into as many of the 24 official languages of the EU as possible, including Croatian (European Central Bank, n.d. b). The projections are therefore fitting for the purposes of the present research, as they are a text from the relevant domain, of sufficient length and complexity, as well as a type of text that the Croatian unit regularly translates. It should be noted that the official Croatian translation of the projections was published only after the machine translated excerpts used in this study were revised and evaluated by the human revisers. The chosen excerpt deals with inflation as measured by the Harmonized Index of Consumer Prices (HICP). Considering the four chosen paragraphs make a cohesive whole and occasionally reference each other, this text also offers me the opportunity to test how both machine translation systems deal with context-related issues. The ChatGPT and eTranslation output, as well as the source text, can be found in the Appendix.

The version of ChatGPT used was ChatGPT 3.5 which, at the time of research, was the freely available version of the system (OpenAI 2022). Another relevant aspect for ChatGPT in particular is the choice of prompt. Whereas eTranslation is a straightforward translation tool to which a source text can be uploaded and automatically translated, ChatGPT is a chatbot that can be used for various purposes, meaning it needs to be asked to produce a translation before being given

the source text. Additional information can also be provided in the prompt, i. e. the type of text, the domain or special considerations, such as not using certain words in the translation. There has been some research suggesting that even relatively small changes in prompt phrasing can influence the quality of the resulting translation. For example, Jiao et al. (2023, 2-3) found that the prompt "Please provide the [target language] translation for these sentences" performed slightly better than other similar prompts, such as "What do these sentences mean in [target language]". Furthermore, Peng et al. (2023, 1) found that including task and domain information in the prompt also improves the resulting translations. After taking all this information into account and running several preliminary experiments involving paragraphs from the macroeconomic projections from June 2023, I decided to use the following prompt:

> The text below is part of the ECB staff macroeconomic projections for the euro area published by the European Central Bank. Please provide a Croatian language translation of the text below using the terminology and style appropriate to the financial domain.

It should be noted that prompts are a potentially vast research topic, as there are many variations that could be experimented with. However, the focus on prompts is beyond the scope of the present study.

As for eTranslation, I decided to use the Finance engine, as this is the engine fine-tuned to the financial and economic domains, and is the one regularly used by the Croatian translation unit.

4.2 Evaluation of the machine translations

Since both automatic evaluation and human evaluation have their strengths and weaknesses, I decided to combine these two kinds of evaluations to achieve the most accurate possible evaluation of the general quality of the machine translations.

4.2.1 Automatic evaluation

In the first phase, the unrevised machine translations were uploaded to MATEO (Machine Translation Evaluation Online), an online tool developed at the University of Ghent (Vanroy 2023 b), to be evaluated automatically. Six different metrics were used in the automatic evaluation: BERTScore, BLEURT, COMET, BLEU, ChrF, and TER. Details on the individual metrics can be found in the "Background" section of the MATEO website (Vanroy 2023 a). Some of these algorithms require a reference translation to evaluate the quality of the machine translation. For this, I used the Croatian translation published on the ECB's official website (European Central Bank, n. d. c). The reference translation can also be found in the Appendix. All the above metrics rate the quality of the translation on a scale from 0 to 100, with 0 being the lowest and 100 being the highest rating. The exception is the TER score, which denotes the number of interventions needed to bring the machine translation in line with the reference translation: This means a low TER score signifies a text more like the reference translation and thus better in quality, while a high TER score signifies that the machine translation needs to be rewritten more extensively (Snover et al. 2006, 228).

4.2.2 Human evaluation

In the next phase, machine translations were given to two revisers for bilingual revision and evaluation. Both revisers had extensive experience translating similar ECB texts from the financial domain from English into Croatian. They were thus familiar with financial terminology, highly skilled in both languages, and aware of the ECB's stylistic, terminological and other conventions.

The aim of the human evaluation phase was ultimately to evaluate the usefulness of the machine translation systems in a real-world setting. This is why care was given to simulate the conditions in which the Croatian unit at the ECB usually works with machine translated texts: the revisers were asked to revise the two translations until they were ready to be published on the official ECB website. This meant correcting any grammatical or factual errors, but also ensuring that the target text is in line with the ECB's stylistic, terminological and other conventions. They were also allowed to use any tools they usually employ in their work, including

ECB-specific term bases and translation memories, but also publicly available online and physical resources (dictionaries, parallel corpora, encyclopedias etc.).

The revisers were asked to measure the time required for completing their revisions. To counter the order effect, they were given the translations in the reversed order – i. e. Reviser A was first given the eTranslation output and only then asked to revise the ChatGPT output, whereas Reviser B was first given ChatGPT's translation and only then the target text produced by eTranslation. Neither reviser knew which system produced which output. After completing the revisions, the revisers were asked to rate both translations on a scale from 1 to 10, with 1 signifying a translation which is unusable, meaning it would take significantly longer to revise the translation than to translate the source text from scratch. A 10 would mean a translation that could be published immediately on the ECB website, without any revision. The revisers were also given the opportunity to comment on the quality of the translations and the differences between them, though this was not obligatory.

There also remains the question of inter-reviser agreement, i. e. if the revisers corrected the same errors. To take this into account, I first counted all the unique errors made by each system, i.e. I counted the changes that either of the two revisers made to ChatGPT's and eTranslation's output respectively. After that, I counted those errors corrected by both revisers. Finally, I compared the two numbers.

4.2.3 Error analysis

Based on the revisions made by the two revisers, I performed a classification and analysis of the errors each system made. I decided to use the "MQM FULL error typology" (The MQM Council, n.d.), as it offers a robust classification system, with 7 high-level error types (terminology, accuracy, linguistic conventions, style, locale convention, audience appropriateness, design and markup) and many subtypes.

5. Results

5.1 Results of the automatic evaluation

All automated evaluation metrics considered eTranslation better than ChatGPT (Figure 3, Table 1).

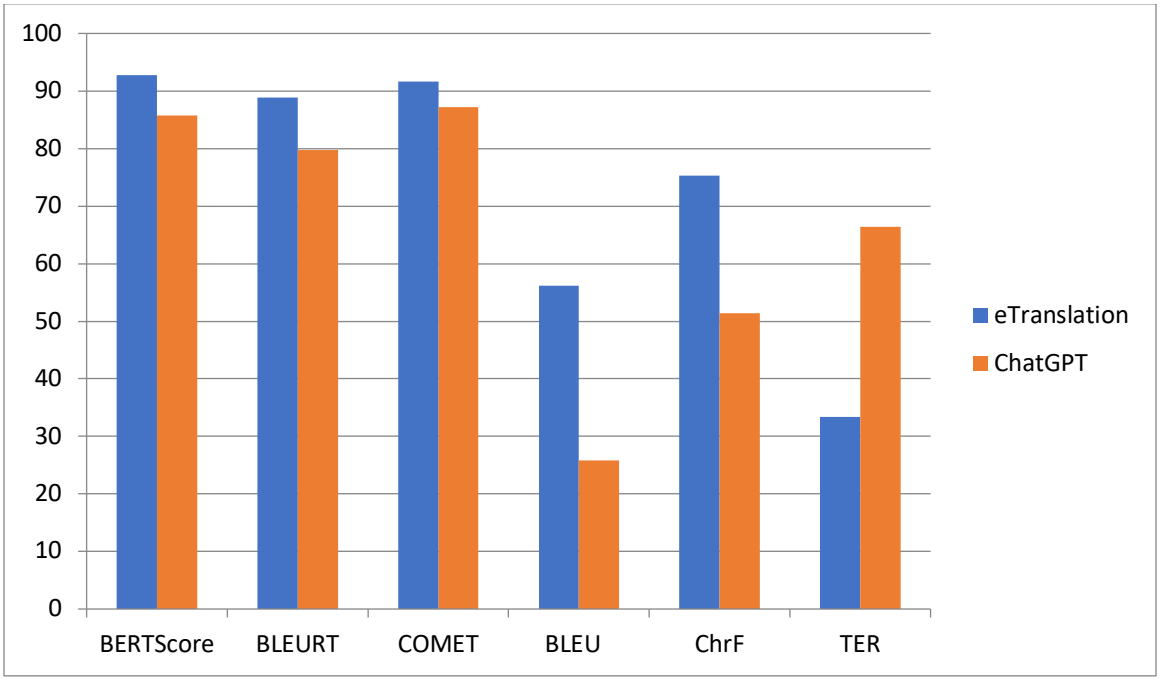

Figure 3. Automatic evaluation results (graph)

Table 1. Automatic evaluation results (table)

	BERTScore	BLEURT	COMET	BLEU	ChrF	TER
eTranslation	92.8	88.9	91.6	56.2	75.3	33.4
ChatGPT	85.7	79.7	87.2	25.8	51.4	66.4

5. 2 Results of the human evaluation

Table 2. Results of the human evaluation

	eTranslation	ChatGPT
Reviser A	6/10, 70 minutes	2/10, 50 minutes
Reviser B	5/10, 40 minutes	1/10, 90 minutes

Reviser A was first given eTranslation's translation and then ChatGPT's translation. The eTranslation output took approximately 70 minutes to revise and was given the

score of 6 on the scale of 1 to 10. The reviser commented that the translation contained certain factual errors, omissions, terminological inconsistencies and incorrect interpretations of temporal relations, but that the terminology was properly used in many cases, and that there were significantly fewer errors than in ChatGPT's translation. The revision of ChatGPT's translation took approximately 50 minutes, and the translation was given a score of 2. Reviser A noted that the text was "brimming" with semantic, terminological, grammatical, orthographic and stylistic errors, that even the most basic terms, such as the acronym "ECB" and the different types of inflation, were inaccurately translated, and that some solutions offered by the system were entirely inappropriate. According to the reviser, it would be significantly easier to translate the text from scratch than to revise the translation to a publishable quality.

Reviser B was first given the ChatGPT output and then the eTranslation output. The eTranslation output required about 40 minutes of revision and was given a score of 5. ChatGPT's output took 90 minutes to revise and was given a score of 1. The reviser noted that, in a real-world setting, it would be better to translate from scratch than to use ChatGPT's output as a starting point, not only because of the time the revision would take, but also because using a low-quality machine translation as a starting point lowers the quality of the final translation.

5.3 Results of error analysis

Table 3 shows all the unique changes made to the MT outputs, i. e. changes made by either of the two revisers. Table 4 shows changes made by both revisers.

Table 3. Error analysis results: changes made by either reviser

	Termin.	Accur.	Ling. conv.	Style	Locale con.	Aud. approp.	Design and markup	Total
eTranslation	12	13	6	49	0	0	0	80
ChatGPT	29	31	26	73	0	0	0	159

Table 4. Error analysis results: errors corrected by both revisers

	Termin.	Accur.	Ling. conv.	Style	Locale conv.	Aud. approp.	Design and markup	Total
eTranslation	6	10	4	17	0	0	0	37
ChatGPT	28	27	24	42	0	0	0	121

5.3.1 eTranslation error analysis

The revisers made 80 unique changes to the eTranslation output. Most of them were due to eTranslation employing style that the revisers thought unidiomatic, awkward or simply not in line with the Croatian unit's stylistic conventions (a total of 49). The system also made 12 terminological errors, using terms that were not consistent with entries in the Croatian unit's termbase (e.g. "komponenta" instead of "sastavnica", "prehrambeni proizvod" instead of "hrana"). Six errors related to linguistic conventions, such as grammar, punctuation and collocations. For example, eTranslation did not insert a period after the number "4" in the chapter heading, which is necessary in Croatian, and misspelled the verb "porasti" as "porast".

As for accuracy errors, 13 of them were found. Some of them were related to the ambiguity in the source sentence. For instance, in the first sentence of the second paragraph, eTranslation uses the past tense, although the source text refers to projected future events:

> Following **an uptick in 2024**, related to the unwinding of fiscal support measures, energy inflation is expected to add only marginally to headline inflation in 2025.
>
> Očekuje se da će inflacija cijena energije, **koja je u 2024. porasla** zbog postupnog ukidanja mjera fiskalne potpore, u 2025. tek neznatno povećati ukupnu inflaciju.

A similar error is repeated later in the same paragraph:

> This turnaround is seen to reflect renewed increases in energy commodity prices following declines over the past year, base effects **changing the sign** from the fourth quarter of 2023 onwards, and the withdrawal of energy and inflation compensatory fiscal measures.

> Taj preokret posljedica je ponovnog povećanja cijena energetskih sirovina nakon smanjenja u posljednjih godinu dana, baznih učinaka koji **su promijenili predznak** iz četvrtog tromjesečja 2023. nadalje te povlačenja energetskih i inflacijskih fiskalnih mjera.

It could be argued that these errors stem from eTranslation's inability to take sufficient context into account. A human translator could easily understand that the English sentence refers to predicted future events by reading previous sentences and paragraphs.

Another similar error was found in the final sentence of the third paragraph, in which eTranslation misidentified the subject of "remain elevated":

> Over the medium term, food inflation is expected to decline more gradually, averaging 2.3% in 2025, as upward price pressures from food commodities ease and profit margins normalise, although **it will remain somewhat elevated** owing to dynamic growth in labour costs.
>
> Očekuje se da će inflacija cijena prehrambenih proizvoda u srednjoročnom razdoblju postupno padati i iznositi prosječn 2,3 % u 2025. jer će se pritisci na rast cijena prehrambenih sirovina smanjiti, a profitne marže normalizirati, **premda će ostati donekle povišene** zbog dinamičnog rasta troškova rada.

It is clear that the original text says inflation will remain elevated, but the translation says the profit margins will remain elevated. Considering the length and complexity of the sentence, it could be hypothesized that this error is also due to eTranslation being unable to take sufficient context into account.

Finally, at several points, eTranslation omitted or disregarded certain information from the source text, which directly led to inaccurate translations, e. g. translating "HICPX inflation" as "inflacija mjerena HIPC-om" instead of "inflacija mjerena HIPC-om bez hrane i energije" or translating "pandemic re-opening effects" as "učinci pandemije".

When it comes to agreement between the revisers, it should be noted that the total number of changes that both revisers made was significantly lower than the number of unique changes (37 as opposed to 80). The error distribution remains

similar, however, with the highest number of errors in the style category, followed by accuracy, terminology and linguistic conventions.

5.3.2 ChatGPT error analysis

The revisers made a total of 159 unique changes to the ChatGPT output. Many of them – 73 – were stylistic. Many terminological errors were made (29 in total), including erroneous translations of basic terms such as "ECB", "headline inflation", "euro area" and the acronym "HICP". ChatGPT made almost as many errors relating to Croatian linguistic conventions (26 in total) – it omitted blank spaces between numbers and percentage signs, omitted periods after ordinal numbers, placed commas in the wrong places and occasionally used an insufficiently formal register.

The system also made many accuracy errors, a total of 31. Some words and phrases were translated in an overly literal manner (e. g. "projected" as "projektiran" when referring to a forecasted future event), some syntactic relations were also misinterpreted (e.g. switching the object and the subject of a verb). At several points, ChatGPT used words or phrases which were entirely unfit for the context:

> The profile is also affected initially by strong base effects in the energy and food components and, throughout the horizon, by fiscal policy measures and **commodity** price assumptions [...]
>
> Profil također prvotno utječe na snažne bazne efekte u energetskim i prehrambenim komponentama, a tijekom cijelog razdoblja utječu i fiskalne mjere te pretpostavke o cijenama **komoditeta** [...]

> Food inflation is projected to continue to decline, given significant base effects, easing **pipeline pressures** and an assumed decline in euro area food commodity prices.
>
> Prognozira se da će inflacija hrane nastaviti padati, uz značajne bazne efekte, smanjenje **pritisaka u cjevovodima** i pretpostavljeno smanjenje cijena hrane u eurozoni.

The word ChatGPT used in the first sentence, "komoditet", means "comfort" rather than "commodity". In the second sentence, ChatGPT translated the phrase

"pipeline pressures" literally despite the fact Croatian does not use a pipeline metaphor to describe that economic phenomenon. It could be hypothesized that these errors are due to domain mismatch: the words in question were translated without regard to the fact that they were used in an economic text, and thus ChatGPT produced overly literal translations.

Finally, ChatGPT arguably made three context-related errors. In the second sentence of the final paragraph, the pronoun "its", which refers back to the word "inflation" in the previous sentence, was mistranslated as the masculine pronoun "njegovih" rather than the feminine "njezinih":

> After standing at a rate of 5.3% in August 2023, HICPX **inflation** is envisaged to recede to 2.9% in 2024. Regarding **its** two main components, non-energy industrial goods inflation is foreseen to decline more than services inflation [...]
>
> Nakon što je iznosila 5,3% u kolovozu 2023., očekuje se da će se HICPX **inflacija** smanjiti na 2,9% u 2024. godini. Što se tiče **njegovih** dviju glavnih komponenti, inflacija netoenergetskih industrijskih dobara predviđa se da će se smanjiti više od inflacije usluga [...]

ChatGPT also twice translated the adjective "this" as "ovaj" when referring back to something previously mentioned in the text:

> After a sharp decline in the course of 2023, headline inflation is expected to move broadly sideways in the first half of 2024, before falling gradually further in 2025 [...]. **This** decline in headline inflation over the projection horizon reflects decreases in the annual rates of change of all the main components [...]
>
> Nakon oštrog pada tijekom 2023., očekuje se da će glavna inflacija kretati širom u prvoj polovini 2024. godine prije daljnjeg postupnog pada u 2025. [...]. **Ovaj** pad glavne inflacije tijekom projekcijskog razdoblja odražava smanjenje godišnjih stopa promjene svih glavnih komponenti [...]

> Energy inflation [...] will remain negative until the last quarter of 2023 before turning positive and rising to 7.4% in the second quarter of 2024. **This** turnaround is seen to reflect renewed increases in energy commodity prices following declines over the past year [...]

> Energetska inflacija trebala bi [...] ostati negativna do posljednjeg tromjesečja 2023. prije nego što postane pozitivna i poraste na 7,4% u drugom tromjesečju 2024. godine. **Ovaj** obrat očekuje se zbog ponovnog rasta cijena energetskih komoditeta nakon padova tijekom protekle godine [...]

The use of "ovaj" in the examples above could be interpreted in multiple ways. On the one hand, some Croatian grammarians believe that only the pronoun "taj" should be used for anaphoric reference, i. e. when referring to something previously mentioned in the text, whereas the pronoun "ovaj" should only be used for cataphoric reference, i. e. when referring to words or phrases used later in the text (Ham 2012, 63). From this point of view, one could argue that ChatGPT made two context-related errors in the examples above, since it misidentified the referent of the pronoun "this". On the other hand, this rule is not followed by the majority of Croatian speakers, especially not in more casual speech or writing. Therefore, the use of "ovaj" in ChatGPT's output could also be considered a stylistic error or not an error at all. Finally, even if we do consider the use of "ovaj" to be a grammatical error, it is unclear if the error was made because ChatGPT is unable to retrieve sufficient context from the preceding sentence or because most of the texts in ChatGPT's training material do not differentiate between the two demonstrative pronouns.

The revisers seemed to be in agreement when it comes to errors made by ChatGPT – if one compares the changes made by either translator with the changes made by both translators in terms of terminology, accuracy and linguistic convention, one will discover their numbers are similar (29 – 28, 31 – 27, 26 – 24). It is only in the style category that there is a more significant difference between the two numbers (159 – 121).

6. Discussion

The human and automatic evaluations both suggest that the translation of the source text produced by eTranslation is of higher quality, as does the error analysis.

The revisers clearly did not consider ChatGPT useful in a practical setting, as attested by the low scores they gave it (1/10, 2/10) and their accompanying

comments. By contrast, it seems reasonable to assume that eTranslation could be of use in certain settings, as it received satisfactory scores (5/10 and 6/10), although much revision was necessary to bring the MT output to a publishable quality. The fact that Reviser A needed more time for the revision of eTranslation's output than for the revision of ChatGPT's output (approximately 70 minutes, compared to approximately 50 minutes) at first glance does not seem to support the conclusion that eTranslation's output is of higher quality. However, this apparent inconsistency can easily be explained by the fact that Reviser A was given eTranslation's text first. This likely gave them the opportunity to familiarize themselves with the source text and already think of possible solutions to translation problems, which would make the revision of ChatGPT's translation significantly faster. Reviser B, who was given the translations in the reversed order, spent far less time revising eTranslation's output (approximatey 40 minutes) than on revising ChatGPT's output (approximately 90 minutes).

The automatic metrics all gave eTranslation a better rating than ChatGPT. However, it should be noted that the gap between the systems seems somewhat smaller than in human evaluation.

When comparing the error types made by the two systems, it can be seen that ChatGPT made more errors and that the errors it made were more severe compared to those made by eTranslation. ChatGPT made more terminological errors, and some of its inaccurate translations seem to be due to domain mismatch, which is in line with my initial expectations. It is unclear whether ChatGPT performed better in terms of context-related issues, as the arguably erroneous translation of the relative pronoun "this" could also be due to the underlying training material rather than due to ChatGPT's inability to take sufficient context into account.

One of the reasons for ChatGPT's underwhelming performance compared to eTranslation is likely that eTranslation's finance engine is specifically designed for financial translation of English texts into Croatian and trained on texts regularly translated at the ECB. It should, however, be noted that the version of ChatGPT I used for the test was 3.5, the most widely used and accessible free version of the

system. The newer version of ChatGPT, 4.0, reportedly boasts significantly greater capabilities (OpenAI 2023), which is also the case for its machine translation capabilities (Jiao 2023, 1). It might be the case that this more advanced version of the system would have yielded better results in this test as well.

7. Conclusion

The results of this study suggest that large language models such as ChatGPT are currently not as useful for machine translation in a real-world setting as more conventional NMT systems fine-tuned to a specific purpose and trained on domain-specific data. It is unclear if ChatGPT fares better when dealing with context-related issues, since its arguably non-standard use of Croatian relative pronouns can be interpreted in multiple ways. It should be noted that the analysis was conducted on ChatGPT-3.5 instead of the more advanced ChatGPT-4 model, which could partially explain the underwhelming results. It should also be noted that the study was conducted on a small text sample and that it involved only two human revisers, which are additional limitations. The use of ChatGPT for English-Croatian machine translation thus remains underexplored. Future research could, among other potential questions, further explore variations in prompting.

References

Castilho, Sheila. 2022. "How Much Context Span is Enough? Examining Context-Related Issues for Document-level MT." *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp. 3017–3025.* European Language Resources Association (ELRA).

Dale, David, Elena Voita, Loïc Barrault and Marta R. Costa-jussà. 2022. "Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better." *arXiv.Org.* Accessed March 21, 2024. https://doi.org/10.48550/arXiv.2212.08597.

DeepL. 2021. "How does DeepL work?" Accessed March 21, 2024. https://www.deepl.com/en/blog/how-does-deepl-work

DuPont, Quinn. "The Cryptological Origins of Machine Translation." In *Amodern 8*. Accessed March 6, 2024. amodern.net/article/cryptological-origins-machine-translation/.

European Central Bank. 2023. "ECB staff macroeconomic projections for the euro area, September 2023." Accessed March 6, 2024. www.ecb.europa.eu/pub/projections/html/ecb.projections202309_ecbstaff~4eb3c5960e.en.html.

European Central Bank, n.d. a. "Home." Accessed March 22, 2024. https://www.ecb.europa.eu/home/html/index.hr.html

European Central Bank. n. d. b. "Language Policy of This Website." Accessed March 6, 2024. www.ecb.europa.eu/services/using-our-site/language-policy/html/index.en.html.

European Central Bank. n. d. c. "Macroeconomic Projections." Accessed March 6, 2024. www.ecb.europa.eu/pub/projections/html/index.en.html.

European Commission. n. d. "eTranslation." Accessed March 6, 2023. https://commission.europa.eu/resources-partners/etranslation_en

Forcada, Mikel. 2017. "Making sense of neural machine translation." In *Translation Spaces* 6:2, 291–309. John Benjamins Publishing Company.

Ham, Sanda. 2012. *Školska gramatika hrvatskoga jezika*. *4. izdanje*. Školska knjiga.

Hughes, Alex. 2023. "ChatGPT: Everything you need to know about OpenAI's GPT-4 tool." In *BBC Science Focus*. Accessed March 21, 2024. https://www.sciencefocus.com/future-technology/gpt-3

Hutchins, John. 2004. "Two precursors of machine translation: Artsrouni and Trojanskij." In *International Journal of Translation* 16, 11-31.

Hutchins, John. 2006. "Machine Translation: History." In *Encyclopedia of Language and Linguistics, 2nd ed*., edited by Keith Brown, 375–383. Elsevier.

IBM, n. d., a. "What Are Large Language Models?" Accessed March 21, 2024. https://www.ibm.com/topics/large-language-models

IBM, n. d., b. "What Is a Neural Network?" Accessed March 21, 2024. https://www.ibm.com/topics/neural-networks

IBM, n. d., c. "What Is a Transformer Model?" Accessed March 21, 2024. https://www.ibm.com/topics/transformer-model

IBM, n. d. d. "What Is Unsupervised Learning?" Accessed July 8, 2024. https://www.ibm.com/topics/unsupervised-learning

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi and Zhaopeng Tu. 2023. "Is Chatgpt a Good Translator? Yes with GPT-4 as the Engine." *arXiv.Org*. Accessed March 21, 2024. arxiv.org/abs/2301.08745.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge University Press.

Liu, Qun and Xiaojun Zhang. 2015. "Machine Translation: General." In *The Routledge Encyclopedia of Translation Technology*, edited by Chan Sin-wai, 105-119. Routledge.

Microsoft, n.d. "Neural Machine Translation." Accessed March 21, 2024. https://www.microsoft.com/en-us/research/project/neural-machine-translation/

MQM Council, n. d. "The MQM FULL Typology." Accessed March 6, 2024. https://themqm.org/the-mqm-full-typology/

Omazić, Marija and Blaženka Šoštarić. 2023. "New Resources and Methods in Translating Legal Texts: Machine Translation and Post-Editing of Machine-Translated Legal Texts." In *Language(s) and Law*, edited by Ljubica Kordić, 71-84. Faculty of Law Osijek.

OpenAI. 2022. "Introducing ChatGPT." Accessed March 21, 2024. https://openai.com/blog/chatgpt

OpenAI. 2023. "GPT-4." Accessed March 6, 2024. https://openai.com/research/gpt-4.

OpenAI, n.d. "Pricing." Accessed March 21, 2024. https://openai.com/chatgpt/pricing

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang and Dacheng Tao. 2023. "Towards Making the Most of ChatGPT for Machine Translation." *arXiv.Org*, https://doi.org/10.48550/arXiv.2303.13780.

Pérez-Ortiz, Juan Antonio, Mikel Forcada and Felipe Sánchez-Martínez. 2022. "How neural machine translation works." In *Machine translation for everyone: Empowering users in the age of artificial intelligence*, edited by Dorothy Kenny, 141-164. Language Sciences Press.

Snover, Matthew, Bonnie Dorr, Richard Schwarz, Linnea Micciulla and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation." In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* 223-231. Association for Machine Translation in the Americas.

Stein, Daniel. 2016. "Machine translation: Past, present and future." In *Language technologies for a multilingual Europ*e, edited by Georg Rehm, Felix Sasaki, Daniel Stein, Andreas Witt, TC3 III, 5–17. Language Science Press.

Vanroy, Bram. 2023. a. "Background." Accessed March 21, 2024. https://mateo.ivdnt.org/Background

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. b. "MATEO : MAchine Translation Evaluation Online." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, edited by Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, et al., 499–500. European Association for Machine Translation (EAMT).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. "Attention Is All You Need." *arXiv.Org*. Accessed March 6, 2024. arxiv.org/abs/1706.03762.

Villani, Rossana. 2021. "The changing profile of the translator profession at the European Central Bank." In *Proceedings of Translation and Interpreting Technology Online Conference, June 2021,* edited by Ruslan Mitkov, Vilelmini Sosoni, Julie Christine Giguere, Elena Murgolo and Elizabeth Deysel, 7-14. Incoma Ltd.

Weise, Karen, Cade Metz, Nico Grant and Mike Isaac. "Inside the A.I. Arms Race That Changed Silicon Valley Forever." *New York Times*, December 5, 2023. Accessed March 6, 2024. https://www.nytimes.com/2023/12/05/technology/ai-chatgpt-google-meta.html

Wu, Yonhgui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey and Maxim Krikun et al. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arXiv.org.* Accessed March 6, 2024. https://doi.org/10.48550/arXiv.1609.08144

Sažetak

Najnovija inovacija na području obrade prirodnog jezika veliki su jezični model kao ChatGPT, koji mogu obavljati mnoge jezične zadaće, uključujući strojno prevođenje. U ovoj studiji istražuju se sposobnosti prevođenja jezičnog modela ChatGPT s engleskog na hrvatski u stvarnom profesionalnom okružju. Odlomak iz dokumenta Europske središnje banke (ESB) preveden je s pomoću modela ChatGPT i sustava eTranslation, konvencionalnijeg neuronskog strojnog prevoditelja koji rabe

prevoditelji u ESB-u. Provedene su ljudska procjena, automatska procjena i analiza pogrešaka kako bi se utvrdilo bi li ChatGPT hrvatskim prevoditeljima u ESB-u bio korisniji od sustava eTranslation te je li ChatGPT uspješniji u rješavanju kontekstnih problema. Rezultati istraživanja upućuju na to da ChatGPT hrvatskim prevoditeljima u ESB-u trenutačno ne bi bio korisniji od sustava eTranslation. Nije jasno je li ChatGPT uspješniji u rješavanju kontekstnih problema. S obzirom na to da je upotreba modela ChatGPT za strojno prevođenje s engleskog na hrvatski i dalje velikim dijelom neistražena, buduća istraživanja mogla bi dodatno istražiti tu temu.

Ključne riječi: neuronsko strojno prevođenje, veliki jezični model, ChatGPT, eTranslation, prijevod s engleskog na hrvatski, problemi povezani s kontekstom