NPFL087 Project Report
Analysis of WMT2023 Results

Dasha Hryhoryeva

Intro

This project focuses on analyzing the results of the WMT23 conference, specifically evaluating GPT4's performance in translating from English to Russian. The goals were to analyze the translations provided by GPT4 for this language pair, identify errors present in the outputs, try to generalize these errors, and develop a test suite to search for similar errors in the future.

Project setup

Data

Source datasets, including references, outputs, and automatic evaluations, are available on the WMT23 Metrics Task website [1, 2]. Manual evaluations were provided by Toloka [3].

Combining scores into one file

I created a script to parse the source, reference, and output data, and align them with automatic scores from GPT4, DeepL, and Google Translate systems. The final version of the script is available on GitHub [4]. I consolidated all the data into a single table that contains structured source, reference, and output data, aligned with both automatic and manual evaluations, along with my comments [5].

Perplexity calculation

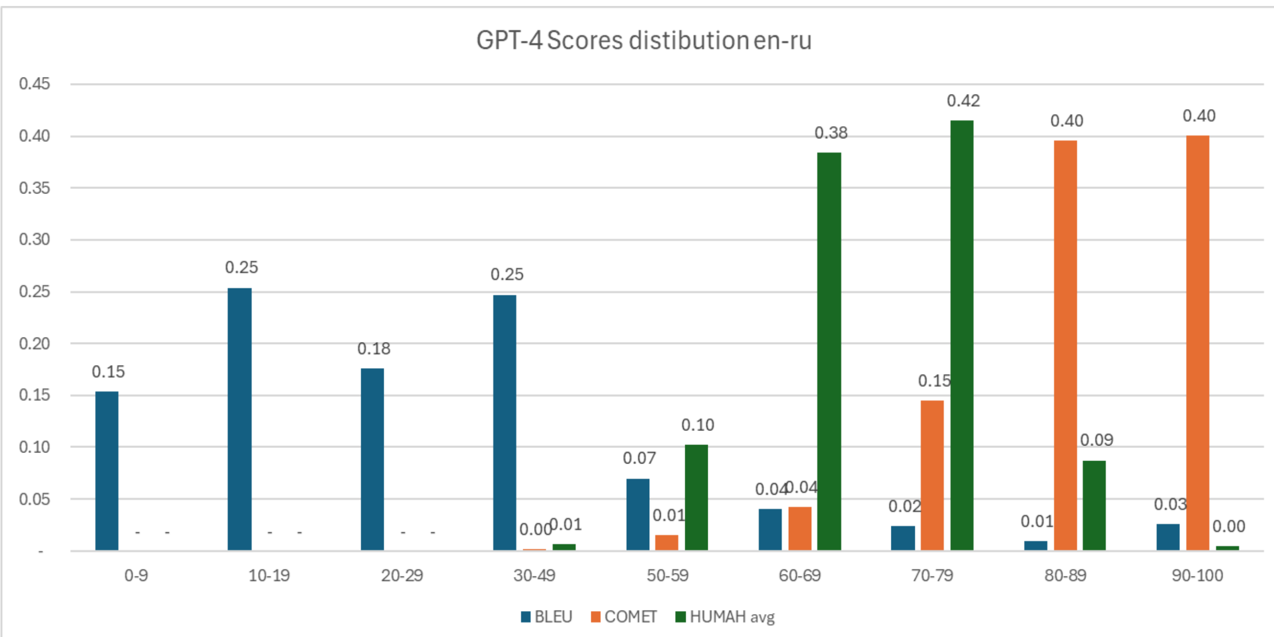
I used perplexity to check how unique a specific phrase is within the general dataset, which contains a large enough sample to assume that GPT4's data diversity is similar to this dataset's diversity. Based on this assumption, I used perplexity calculated on the general dataset as an approximate measure of a phrase's uniqueness in GPT4. Initially, I attempted to use the KenLM tool [8, 9] and the English Common Crawl Monolingual Training Data available on the WMT23 website. However, I was unable to get it to work due to the compressed format of the data. The main issue was that the unix pipe approach in bash cannot unpack compressed TGZ files piece by piece, and for large volumes of data, there is not enough RAM. As an alternative solution, I used a perplexity metric by HuggingFace [6], which allows the use of different models as a base dataset for perplexity calculation. For my experiments, I used the GPT2 model, assuming that models from this family are partially based on the same datasets.

Discussion

Data statistics

The dataset consists of 2074 segments coming from 192 documents concatenated into a sequence. I worked with BLEU [10] and COMET [11] automatic evaluations, which cover all segments, and manual evaluations, which cover 53% of the segments and have at least 5 entries per covered segment.
I also used Google Translate (online-B system in WMT23 data) and DeepL (online-W system in WMT data) results as additional reference translations.

The distribution of automatic and human scores on a scale from 0 to 100 is illustrated in the following bar chart. As the bar chart suggests, the majority of COMET scores exceed 80 points, whereas human evaluations mostly fall within 60 to 80 point range, with a small number of segments being estimated above 80 points. In my analysis, I concentrated on segments with COMET scores ranging from 60 to 80 points taking COMET score as the main indicator.



I also attempted to analyze the BLEU score. The first column of the table below shows the human average scores split by ranges. The second and third columns display the average COMET and BLEU scores for segments that fall into the specific ranges of human scores. As shown in the table, the BLEU score does not appear to correlate with human scores. For example, in the 60-69 and 70-79 ranges, both human and COMET scores increase, while the BLEU score decreases. Therefore, I will disregard the BLEU score in further analysis.

	Human avg	COMET avg	BLEU avg
0-29			
30-49	**47.14**	83.15	25.32
50-59	**56.23**	85.22	31.47
60-69	**65.13**	85.90	30.35
70-79	**73.82**	86.47	29.34
80-89	**82.83**	85.80	28.02
90-100	**91.20**	94.74	60.93

The table below displays a comparison of average BLEU, COMET, and human scores. Human average and minimum scores are available for the GPT4 system only. It was also interesting to examine the average scores for different systems and note that, on average, GPT4, Google Translate, and DeepL are very close based on the COMET score. However, for segments with COMET scores below 80 points, Google Translate and DeepL performed better according to COMET score, i.e. GPT-4 COMET scores under 80 are visibly lower compared to the two competitors. My observations as a human evaluator confirmed this finding, i.e. that GPT-4, when departing more from the reference as seen by COMET, indeed does translate worse than standard Transformer systems. Consequently, I used Google Translate and DeepL as secondary and tertiary references when I disagreed with the initial reference.

Average values

	BLEU	**COMET**	**COMET <80**	**Human**	**Human min**
GPT4	29.9	86.1	72.1	69.4	41.8
Google Translate	39.2	86.2	75.3		
DeepL	30.9	86.6	76.2		

Observations

- The main reference translation used for all automatic score evaluations in WMT23 for English to Russian tends to be quite artistic. By this, I mean that for better sentence construction and comprehension for native speakers, the reference translation may omit certain parts of the sentence (e.g., introductory words) or rephrase clauses in a way that is semantically close but far from a literal translation. If a segment translation makes good sense and reflects the source meaning correctly but receives a low score due to differences from the reference, I do not try to improve such translations. This decision is based on the idea that machine translation is not meant to be artistic, so a lower score reflecting this is appropriate.
- However, my most common observation was that GPT4 produces "too literal" translations. By this, I mean a lack of fluency in the translation—instances where the sense of the segment is clear, but there are better ways to convey the same meaning.
- The dataset contains a significant amount of slang words and expressions, most often found in direct speech or informal contexts, where even human evaluators may struggle to grasp the segment's meaning. Unsurprisingly, GPT4 also struggled with such cases. In my analysis, I marked these instances but did not include errors of this kind in the test suite. As a non-native

speaker, I found it challenging to be definitive about slang translation or to find resources with similar examples.

Main cases

In this part, I want to describe the main cases I analyzed and found interesting for the project test suite.

Possessive nouns and possessive pronouns.

Example 1 shows that GPT4 struggles with possessive nouns of the type "dog's", the expected translation to Russian is the Adjective "собачий", while the proposed solution is "для собак" (gloss: for dogs). Although the proposed translation makes sense, it is not the best choice in this context.
Example 2 shows a similar issue with possessive pronouns of type "someone's mom", the expected translation to Russian is the Adjective + Noun "чья-то мама". The translation GPT4 suggests is "мама кого-то" (gloss: mom of someone), which doesn't sound natural in Russian.

Example 1

Source:
It was a one dog night and that doggo was not getting off the bed not even for "Dog's Breakfast - Breakfast for Dog's!"
GPT4:
Это была ночь с одной собакой, и этот пёс не собирался слезать с кровати даже ради "Завтрака для собак - Завтрак для собак!"
Gloss:
It was a night with one dog and that dog was not getting off the bed, not even for "Breakfast for dogs - Breakfast for dogs!"
Expected translation for the highlighted green words:
Собачий завтрак - Завтрак для собак

Example 2

Source:
I'd never seen anyone's mom make it or had it shoved at me at a party or seen it appear in conjunction with anyone's saucemaking escapades or baking days
GPT4:
Я никогда не видел, чтобы мама кого-то готовила это, или мне подавали это на вечеринке, или видел, как это появляется в связи с кулинарными приключениями или днями выпечки кого-либо.
Gloss:
I've never seen mom of someone make this, or had it served to me at a party, or seen it appear in connection with cooking adventures or baking days of someone.
Expected translation for the highlighted green words:
чья-то мама, чьими-то приключениями по приготовлению соусов или днями выпечки

Time vs age prepositional phrase

Example 1 shows the phrase "...rule, from last August" in the meaning "rule that was issued in last August", which is translated by GPT4 as [gloss] "rule, since last August". The expected translation in Russian would be "правило от прошлого августа" or similar.
In Example 2, the error is similar.

Example 1

Source:
And we can also do the very same thing with the ATF receiver/frame rule, from last August.
GPT4:
И мы также можем сделать то же самое с правилом приемника/рамки ATF, начиная с прошлого августа.
Gloss:
And we can also do the same with the ATF receiver/frame rule, since last August.
Expected translation for the highlighted green words:
от прошлого августа

Example 2

Source:
A recycled Twitter joke from two months ago.
GPT4:
Переработанная шутка из Твиттера два месяца назад.
Gloss:
A recycled Twitter joke two months ago.
Expected translation for the highlighted green words:
двухмесячной давности = "two-month old"

Parentheses and brackets

There was no clear guide in the data about translating the words in parentheses, and I based my analysis on the assumption that all such words should be translated. It is simple to imagine the situation with playwriting with characters' names in parentheses or recording transcripts with speakers' names in parentheses, and I would expect that names in such texts should be translated. Conversely, in the text containing, for example, markup language snippets the translation of words in brackets is undesirable, and I would assume the system to distinguish such cases.
Example 1 demonstrates that GPT4 does not translate words in square brackets. Although in this specific domain, not translating is the preferable option, which makes the chosen reference somewhat unfortunate.

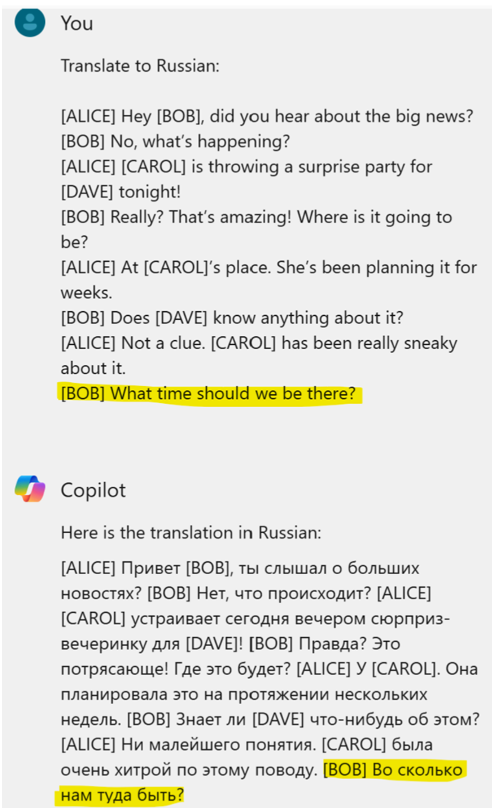
Example 1

Source	Reference	GPT4 Output
And if this debugging can be done by the author of that component, here, in this case [PERSON4], it would be most efficient for all of us.	И наиболее эффективно для всех нас будет, если эта отладка может быть выполнена самим автором компонента, вот в данном случае [ЧЕЛОВЕК4].	И если эту отладку может выполнить автор этого компонента, в данном случае [PERSON4], это было бы наиболее эффективно для всех нас.
So, [PERSON8], please confirm that you agree with this idea of like do-it-yourself integration.	[ЧЕЛОВЕК8], пожалуйста, подтвердите, что вы согласны с идеей самостоятельной интеграции.	Итак, [PERSON8], пожалуйста, подтвердите, что вы согласны с этой идеей самостоятельной интеграции.

(PERSON17) Yeah, so for the multi-accent English.	(ЧЕЛОВЕК17) Да, для английской речи с разными акцентами.	(PERSON17) Да, так что для английского с множественными акцентами.
So, we are now [PERSON14] putting that together to just one technical solution.	Итак, теперь [ЧЕЛОВЕК14] объединяет все это в одно техническое решение.	Итак, мы сейчас [PERSON14] объединяем это в одно техническое решение.

I conducted a few experiments to check what are the conditions when GPT4 translates or does not translate words with square and round brackets.
Firstly, I created a simple dialog between a few people where all names are written in capital letters and wrapped in square brackets (Picture 1 in the table below). In GPT4's translation names are kept in English. Then the square brackets were replaced with round brackets (Picture 2 in the table below), GPT4's translation again contains names in English. The next step was to write names conventionally with first letter capital (Picture 2 in the table below), the result of translation was the same as in two previous requests. Finally, I changed the English names to the ones that are popular in the Russian language (Picture 4 in the table below) and this made GPT4 translate the names into Russian.

1



You

Translate to Russian:

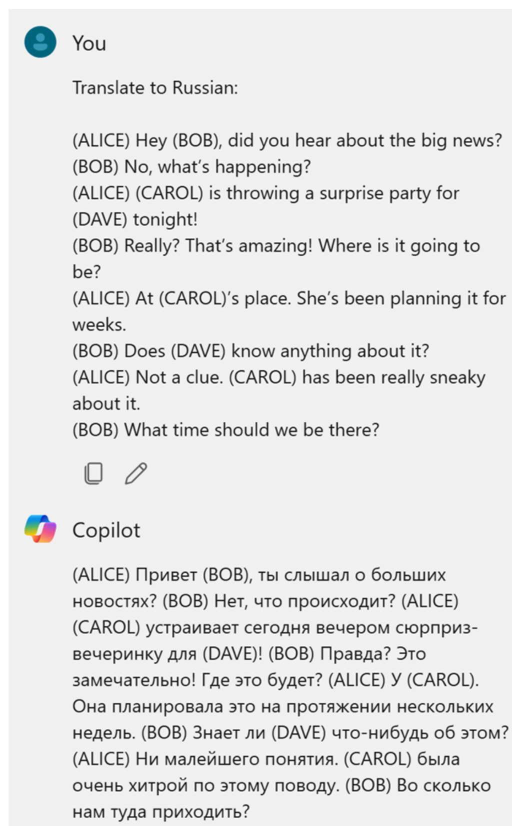
[ALICE] Hey [BOB], did you hear about the big news?
[BOB] No, what's happening?
[ALICE] [CAROL] is throwing a surprise party for [DAVE] tonight!
[BOB] Really? That's amazing! Where is it going to be?
[ALICE] At [CAROL]'s place. She's been planning it for weeks.
[BOB] Does [DAVE] know anything about it?
[ALICE] Not a clue. [CAROL] has been really sneaky about it.
[BOB] What time should we be there?

Copilot

Here is the translation in Russian:

[ALICE] Привет [BOB], ты слышал о больших новостях? [BOB] Нет, что происходит? [ALICE] [CAROL] устраивает сегодня вечером сюрприз-вечеринку для [DAVE]! [BOB] Правда? Это потрясающе! Где это будет? [ALICE] У [CAROL]. Она планировала это на протяжении нескольких недель. [BOB] Знает ли [DAVE] что-нибудь об этом? [ALICE] Ни малейшего понятия. [CAROL] была очень хитрой по этому поводу. [BOB] Во сколько нам туда быть?

2



You

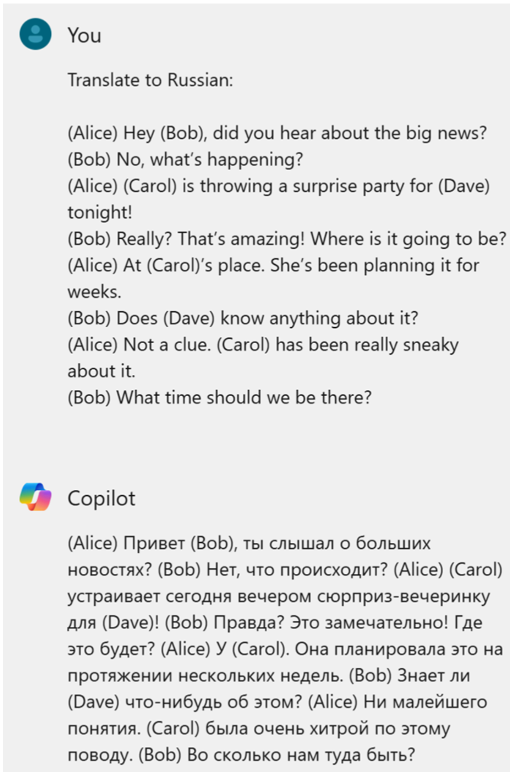
Translate to Russian:

(ALICE) Hey (BOB), did you hear about the big news?
(BOB) No, what's happening?
(ALICE) (CAROL) is throwing a surprise party for (DAVE) tonight!
(BOB) Really? That's amazing! Where is it going to be?
(ALICE) At (CAROL)'s place. She's been planning it for weeks.
(BOB) Does (DAVE) know anything about it?
(ALICE) Not a clue. (CAROL) has been really sneaky about it.
(BOB) What time should we be there?

Copilot

(ALICE) Привет (BOB), ты слышал о больших новостях? (BOB) Нет, что происходит? (ALICE) (CAROL) устраивает сегодня вечером сюрприз-вечеринку для (DAVE)! (BOB) Правда? Это замечательно! Где это будет? (ALICE) У (CAROL). Она планировала это на протяжении нескольких недель. (BOB) Знает ли (DAVE) что-нибудь об этом? (ALICE) Ни малейшего понятия. (CAROL) была очень хитрой по этому поводу. (BOB) Во сколько нам туда приходить?

3



You

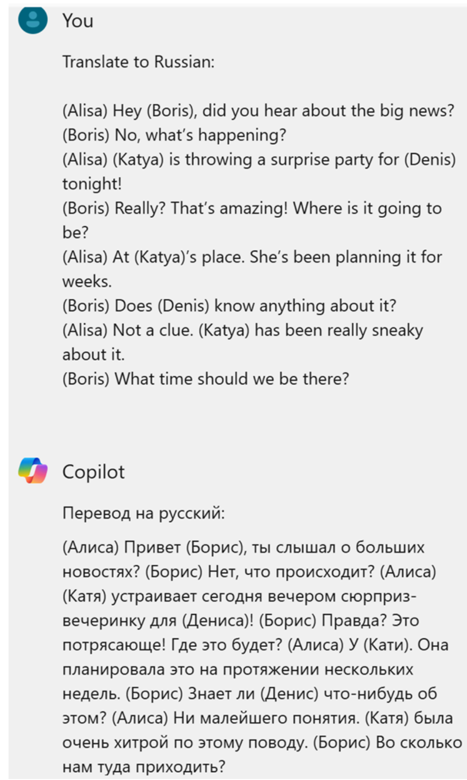
Translate to Russian:

(Alice) Hey (Bob), did you hear about the big news?
(Bob) No, what's happening?
(Alice) (Carol) is throwing a surprise party for (Dave) tonight!
(Bob) Really? That's amazing! Where is it going to be?
(Alice) At (Carol)'s place. She's been planning it for weeks.
(Bob) Does (Dave) know anything about it?
(Alice) Not a clue. (Carol) has been really sneaky about it.
(Bob) What time should we be there?

Copilot

(Alice) Привет (Bob), ты слышал о больших новостях? (Bob) Нет, что происходит? (Alice) (Carol) устраивает сегодня вечером сюрприз-вечеринку для (Dave)! (Bob) Правда? Это замечательно! Где это будет? (Alice) У (Carol). Она планировала это на протяжении нескольких недель. (Bob) Знает ли (Dave) что-нибудь об этом? (Alice) Ни малейшего понятия. (Carol) была очень хитрой по этому поводу. (Bob) Во сколько нам туда быть?

4



You

Translate to Russian:

(Alisa) Hey (Boris), did you hear about the big news?
(Boris) No, what's happening?
(Alisa) (Katya) is throwing a surprise party for (Denis) tonight!
(Boris) Really? That's amazing! Where is it going to be?
(Alisa) At (Katya)'s place. She's been planning it for weeks.
(Boris) Does (Denis) know anything about it?
(Alisa) Not a clue. (Katya) has been really sneaky about it.
(Boris) What time should we be there?

Copilot

Перевод на русский:

(Алиса) Привет (Борис), ты слышал о больших новостях? (Борис) Нет, что происходит? (Алиса) (Катя) устраивает сегодня вечером сюрприз-вечеринку для (Дениса)! (Борис) Правда? Это потрясающе! Где это будет? (Алиса) У (Кати). Она планировала это на протяжении нескольких недель. (Борис) Знает ли (Денис) что-нибудь об этом? (Алиса) Ни малейшего понятия. (Катя) была очень хитрой по этому поводу. (Борис) Во сколько нам туда приходить?

Neogism

The next example is a failed attempt to create a neologism from a slang word. The English word "dad-ification" is translated by GPT4 as "отечественность", which I would describe as a mix of "fatherland" and suffix -ility. That is interesting that GPT4 tries to create a new word, but the result does not convey the meaning.

Source:
Dad-ification usually posits that the solution to a situation where a parent figure must care for or protect a child or child stand-in is overwhelming strength, and using that strength to force a very specifically shaped power dynamic.
GPT4:
Отечественность обычно предполагает, что решение ситуации, когда родительский авторитет должен заботиться или защищать ребенка или заменителя ребенка, заключается в огромной силе и использовании этой силы для создания очень специфической формы властных отношений.
Gloss:
[Neologism: mix of fatherland and -ility suffix] generally assumes that the solution to a situation where parental authority must care for or protect a child or child surrogate is enormous power and the use of that power to create a very specific form of power relationship.
Expected translation for the highlighted green words:
Папификация (neologism)

Change of the meaning

The phrase "we're reluctant to buy" in this example suggests that the purchase didn't happen since the buyers have their doubts about it, they are reluctant to buy. The translation to Russian in its turn means [gloss] "we are reluctantly buying", which means the purchase took place. This case is particularly tricky because subtle differences in meaning determine whether "reluctant to do something" should be translated as "do not want to do something" or "do something, but reluctantly", although the former translation "do not want" is more probable. In this particular example, the meaning of the expression was changed from "did not buy" to "buy".

Source:
Having been radicalized here, we're reluctant to buy a new FAANG machine*.
GPT4:
Подвергшись радикализации здесь, мы неохотно покупаем новую машину FAANG*.
Gloss:
Having been radicalized here, we are reluctantly buying a new FAANG* car.
Expected translation for the highlighted green words:
Мы не хотим покупать = We don't want to buy

Wrong disambiguation

A flyer has two major definitions:

1. A person or thing that flies
2. A small handbill advertising an event or product

In the below example, the disambiguation of flyer meaning happens incorrectly.

Source:
You decide that establishing contact again would be dangerous and return to your flyer, but the experience has cleared your mind which had been befuddled ever since you jumped into the future.
GPT4:
Вы решаете, что возобновление контакта будет опасным и возвращаетесь к своему листовке, но это опыт прояснил ваш разум, который был смущен с тех пор, как вы прыгнули в будущее.
Gloss:
You decide that resuming contact would be dangerous and return to your flyer, but the experience has cleared your mind, which had been embarrassed since you jumped into the future.
Expected translation for the highlighted green words:
Флаер [flyer] = paper with some information

Conclusion

- I analyzed the WMT23 results for the language pair English-Russian and created a mini test suite [7] based on the cases described in this report.
- The most problematic part for me was to generalize the found examples. The observed segments with low evaluations rarely showed the same (or even similar) linguistic phenomenon. It was relatively simple to re-create the error for the same phrase in different contexts, whilst attempting to find similar examples by nature without using the exact phrase wasn't successful.
- This observation confirms that the translation provided by large language models is good enough and comparable with other leading systems like DeepL. Although this is true for English-Russian translation, it might not apply as well to other languages. For example, the English-Chinese (Mandarin) translation analysis we discussed in class did not show the same positive results as my English-Russian analysis.

- I also checked many of the analyzed examples in the newly released GPT4o model and found out that the majority of them are translated correctly.

Links

1. https://wmt-metrics-task.github.io/
2. https://drive.google.com/file/d/1oRrZI9nrM3739tdq0PDYaYOr94X2PIgQ/view
3. https://drive.google.com/drive/folders/1HNusvFK_ZQvh6eHyW2bV2wqgOL4QtTVI
4. https://github.com/DashaII/MT_Project
5. https://github.com/DashaII/MT_Project/blob/main/en_ru_scores_table.xlsx
6. https://huggingface.co/spaces/evaluate-metric/perplexity
7. https://github.com/DashaII/MT_Project/blob/main/test-suite.csv
8. https://kheafield.com/papers/edinburgh/estimate_paper.pdf
9. https://kheafield.com/code/kenlm/
10. https://aclanthology.org/P02-1040.pdf
11. https://aclanthology.org/2020.emnlp-main.213.pdf