From General LLM to Translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning

Denis Elshin **Nikolay Karpachev** **Alexander Antonov** **Anton Chekashev**

Alexander Chernyshev **Kirill Denisov** **Ekaterina Enikeeva** **Vera Frantsuzova**

Ilya Golovanov **Boris Gruzdev** **Georgy Ivanov** **Ekaterina Latypova**

Vladimir Layner **Vladislav Negodin** **Dmitry Popov** **Nickolay Skachkov**

Abstract

This paper describes Yandex submission to the WMT2024 General Translation Task. More specifically, we present a novel pipeline designed to build a strong paragraph-level translation engine with an emphasis on video subtitles domain. In particular, we apply a multi-stage adaptaion pipeline on top of LLM pretraining to align the model for translation task and subsequently to the video subtitles format. Our submission ranks 3rd on the preliminary general translation leaderboard.

1 Introduction

In this paper, we present unconstrained system submitted by the Yandex LLC NLP team to the WMT 2024 General MT Translation track, focusing on English-to-Russian translation. Our approach involves training a YandexGPT[1] LLM-based model for translation tasks using a multi-stage process to ensure high-quality and contextually accurate translations.

We are not capable of revealing all the details of the model due to NDA reasons, however, we can say that it is a Yandex GPT-like model, specifically trained for the translation task.

Our multi-stage approach, which combines extensive pre-training, targeted fine-tuning, advanced prompt-tuning, and structure-preserving techniques, ensures that our model delivers high-quality, fluent, and structurally consistent translations and performs well both in competitive benchmarks and real-world applications.

[1] https://yandex.cloud/en/services/yandexgpt

2 System Overview

2.1 Pretraining

The foundation of our approach is a robust pretraining phase involving a Large Language Model (LLM) trained on a vast corpus of clean texts in multiple languages, with a predominant focus on Russian and English. The quality of this pretrained model is evaluated using a comprehensive suite of benchmarks, including both automated metrics and human evaluation.

This initial phase ensures that the model captures a wide range of linguistic features and nuances across different languages, thereby establishing a strong base for subsequent fine-tuning.

2.2 Incorporating Parallel Data

Following the pretraining phase, we enhance the model by incorporating parallel data, where English and Russian texts are concatenated using a delimiter. This step is crucial for aligning the model's understanding of both languages in a translation context. We use a proprietary CommonCrawl-like parallel corpus of pages crawled from the Web. The data is meticulously curated to ensure high quality using Bicleaner-likeRamírez-Sánchez et al. (2020)

Proceedings of the Ninth Conference on Machine Translation, pages 247–252

pipeline:

- Texts are selected using automated parallelism filters.
- Duplicates are removed to maintain a clean dataset.

This concatenation strategy enables the model to establish connections between two languages and to learn direct mappings from English to Russian and vice versa.

2.3 Sentence-level vs. Paragraph-level Translation

Our initial translation model primarily focuses on sentence-level translation. However, through extensive experimentation, we have observed that paragraph-level translation benefits significantly more from clean, coherent paragraph-level data. Unlike isolated sentences, paragraphs provide a broader context, which is essential for maintaining the flow and coherence in translations.

To leverage this, we gather texts that are inherently structured in paragraphs. These texts are preprocessed to ensure they meet our quality standards:

- Automated filters are employed to assess text parallelism and quality.
- Rigorous deduplication processes are applied to eliminate any repeated content, ensuring that the data fed into the model is both diverse and representative.

2.4 Structured content translation

Although the document-level translation system we have obtained using the pipeline above has high translation quality on generic textual data, it is incapable of consistently translating data in structured format, e.g. data in HTML format. Particularly, when presented texts with tags or other strict markup, model is prone to dropping or altering the markup and thus generating an invalid HTML page.

To handle this problem, we have designed a data augmentation strategy aimed at guiding the model towards HTML domain and such an augmentation have been incorporated into our document-level alignment stage.

2.5 Fine-Tuning LLM for Subtitle Translation

Building on a pre-trained LLM proficient in translating tagged web pages, we developed a method to train the model for subtitle translation. The key idea of this approach involves enclosing each speaker and dialogue in brackets, ensuring accurate parsing into individual dialogues.

This adaptation enhances the LLM's ability to meet the specific challenges of subtitle translation, ensuring contextually accurate outputs with proper segmentation by speaker and timing.

In the subsequent sections we further describe the main stages of our pipeline.

3 Supervised Fune-Tuning (SFT)

Firstly, we align the pretrained language model to the machine translation task. We conduct supervised fine-tuning (SFT) on an in-house dataset of parallel books fragments of up to 1000 tokens length.

We use multilayer prompt-tuning as in Liu et al. (2021) with each p-tuning block size of 100.

Overall LLM input consists of an English source text surrounded by two p-tuning blocks:

Figure 1: PTune blocks layout.

4 Human Feedback Alignment

Following the Supervised Fine-Tuning stage, we further improve core translation capabilities of the model using our internal Human Preferences dataset.

4.1 Data

We collect the training data using Side-By-Side human evaluation of paragraph-level translations, where an expert has to choose which of the two translations is better. The annotated data is presented in triplets (source, winner, loser), where 'winner' and 'loser' correspond to the compared translations. The source segments are sampled from various domains including books of different genres, web pages etc.
Our training dataset consists of the following parts:

Sentence-level data
Sentence part of the corpus consists of side-by-side comparisons between different model generations, in total 100.000 sentence triplets.

Document-level data
Document part of the corpus contains two primary sources of human feedback annotations.

Firstly, similarly to the sentence-level alignment data, we collect several thousands of document-level side-by-side comparisons between different versions of our model.

Secondly, we collect an additional contrastive triplet corpus aimed specifically at improving translation fluency.

Total document-level corpora size is several tens of thousands triplets.

4.2 Modeling

We fine-tune the model obtained at SFT stage using contrastive learning objective.
The model is trained using Contrastive Preference Optimization (CPO) loss function as in Xu et al. (2024).

$$\mathcal{L}(\pi_\theta; U) = \min_\theta \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{\text{prefer}}} \underbrace{-E_{(x,y_w)\sim\mathcal{D}}[\log \pi_\theta(y_w|x)]}_{\mathcal{L}_{\text{NLL}}}. \tag{1}$$

where

$$\mathcal{L}_{\text{prefer}}(\pi_\theta, U) = -E_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Big]. \tag{2}$$

We train with batch size of 64, 1 epoch and triangular learning rate schedule (warmup length of 0.1 epochs, peak learning rate 1e-6).
It is worth mentioning that, due to the dataset imbalance between sentences and documents, training on a uniform mixture yields results almost equal to only sentence-wise training. To handle this discrepancy between sources, we employ a variation of curriculum learning (Bengio et al. (2009)).
In particular, we implement an easy-to-hard schedule, where we start with training only on sentence-level data and shift towards longer documents to the end of the training. This enables more effective leveraging of low-resource document-level corpora.

5 Structured content translation

In this section, we explore the methodology developed to improve the translation of pages with structured data (e.g. web page or video subtitles data) by Large Language Models (LLMs). Traditional LLMs, when tasked with translating structured content, often exhibit significant hallucination level. This manifests as omission of tags, partial tag loss, or incorrect translation of tags. Our goal is to achieve a more robust and accurate translation of such content by ensuring the correct transfer of tags.

5.1 Current Challenge: Tag Hallucination

During free-form translation, LLMs struggle to maintain the integrity of HTML tags. This issue is critical as tags are essential for preserving the structure and formatting of HTML documents. A common problem observed is the complete omission of tags or their partial loss, which leads to a significant decrease in the quality of the translated document. An initial assessment showed a low percentage of correctly transferred tags. Tags are preserved only in 36% for CPO model that proves the need of a more reliable approach to tag preservation.

Test data: To test the accuracy of tag preservation we used a corpus of HTML-fragments. We collected innerHTML of block HTML tags from 10 Wikipedia pages.

Proposed Solution: Bracket Substitution and Model Adaptation
To address the issue of hallucination and improve tag preservation, we propose the following approach:

5.1.1 Tag Substitution with Brackets

Paired HTML tags are replaced with paired brackets (e.g., becomes {, and becomes }) to simplify the text structure for the model. Unpaired tags are also converted to a bracket format: every unpaired tag becomes a pair {}. This increases the proportion of sentences with retained tags to 76%.

```
a. I saw a cat.
b. <span><a>I</a> saw a <span>cat</span>.</span>
c. _< |span|>< |a|> |I| </ |a|> |_saw|_a|_< |span| > |cat| </ |span|>.</ |span|>
d. {[I] saw a [cat].}
```

Figure 2: a. Plain sentence. b. Sentence with html tags. c. Sentence with tags displayed as subword tokens processed by LLM. d. Sentence with tags replaced with braces.

5.1.2 Adaptation Using Parallel Corpus

We utilize a parallel corpus of HTML texts sourced from open repositories. This corpus serves as a foundation for generating synthetic data necessary for model fine-tuning.

5.1.3 Training Dual-Network System

Train data: We used the same parallel corpus as for SFT training but with tags aligned from original HTML documents. Sentence pairs with non-matching HTML tags were filtered out.
First Network: This network is trained to insert brackets and line breaks correctly into the text in the original language. This step helps to maintain the structural consistency of the text.
Second Network: Given a source text with tags and its translation without tags, this network learns to accurately re-insert the tags into the translated text. This network ensures that the translated content preserves the necessary HTML tags.

5.2 Synthetic Data Generation

By leveraging the dual-network system, we generate a substantial amount of synthetic data. This data includes the original text with brackets and line breaks, and the corresponding translated text with correctly inserted tags. Specifically, for the Contrastive Preference Optimization (CPO), we use:

1. The output of the first network as the source sentence in English.
2. The output of the second network on a good translation as the positive example.
3. The output of the second network on a poor translation as the negative example.

The good/poor translation pairs were obtained using human annotation as described above.

5.3 Results

Our experimental results demonstrate that the proposed methodology effectively increases the percentage of sentences with correctly transferred tags to 99%.
This substantial improvement underscores the effectiveness of our approach in reducing tag hallucination and ensuring a more stable and accurate translation of HTML content.
By substituting HTML tags with brackets, adapting the model using a parallel HTML corpus, and incorporating a dual-network system for synthetic data generation, we have developed a robust method to enhance HTML translation. This approach not only mitigates the problem of tag hallucination but also ensures the structural integrity of translated HTML documents. The success of this methodology paves the way for more reliable and efficient translation of structured data formats, significantly benefiting applications in web content translation and beyond.

6 Fine-Tuning LLM for Subtitle Translation

Building upon a pre-trained model that has demonstrated proficiency in translating tagged web pages, we have adapted the following approach to train a subtitles translation system. Its core idea is straightforward: we enclose each speaker and their corresponding dialogue in brackets, as shown in figure below.

```
[
{man 1:} {
    Hey, guys, Kevin here from snowboard pro camp, in this video I'm
    going to give you a list of the first ten tricks to learn on your
    snowboard.
}
{woman 1:} {
    These tricks are in order and each trick will teach you a skill that
    you'll use in the next trick on the list.
}
]
```

Figure 3: Subtitles input format.

This ensures that the translation preserves these brackets, allowing the entire text to be parsed into individual speaker dialogues.
The production version of the algorithm is somewhat more sophisticated, as it must align the translations of longer dialogues with their corresponding timestamps. However, for the purposes of this discussion, a more detailed description is unnecessary and is therefore omitted from this paper.
We fine-tune the model using publicly available subtitle corpora, which we preprocess to fit the above mentioned format. This additional training step has led to noticeable improvements in our human evaluation scores, particularly within the domain of movies and YouTube video subtitle translation. The reason for employing this model is that part of the competition data is presented in audio format, making effective subtitle translation a critical component of our approach.

By adapting the LLM in this manner, we enhance its ability to handle the unique challenges posed by subtitle translation, ensuring that the final outputs are both contextually accurate and properly segmented according to speaker and timing, which is crucial for maintaining the integrity of the original content in the translated version.

7 Evaluation Metrics and Results

Ablation

In order to estimate the effect of each stage of the pipeline, we compare our models using BLEURT-20 (Sellam et al. (2020)) and COMET (Rei et al. (2020)) automatic metrics, as well as BLEU. We rely primarily on neural metrics results as suggested in Freitag et al. (2022). Table 1 shows the scores on WMT-22 English to Russian testset.

Model Ablation (wmt-22 fwd)			
Model Stage	BLEURT	COMET	BLEU
PTune	0.76	0.836	31.3
cpo-sents	**0.789**	**0.860**	**31.52**
cpo-curriculum-base	0.787	0.855	24.8
cpo-curriculum-tags	0.784	0.855	27.1

Table 1: Metrics by stage (sentence-level).

Model Ablation (wmt-22 fwd news)			
Model Stage	BLEURT	COMET	BLEU
PTune	0.728	0.835	**27.78**
cpo-sents	0.733	0.847	25.55
cpo-curriculum-base	**0.743**	**0.850**	19.61

Table 2: Metrics by stage (document-level).

Firstly, the model trained only on parallel data (PTune) is already capable of generating decent quality translations. However, it exhibits bias towards literal translations and poor fluency.
During the alignment stage (cpo-curriculum-base) the model is exposed to a variety of high-quality translations (including contrastive triplets aimed specifically at improving fluency) and, hence, the model after initial CPO training is much more fluent, but prone to tags omission and format inconsistency.
Augmented CPO training solves the problem with format and tags without sacrificing the model's target language fluency capabilities.

Overall, the metrics ablation highlights the following:
1) BLEU correlates poorly with model quality, especially on document-level benchmarks due to high preference for literal translations.
2) On sentence-level evaluation contrastive learning model trained only on sentence data yields superior results both on neural and n-gram based metrics.
3) Tag-focused augmentation does not lead to quality degeneration on primary benchmarks whilist increasing model stability (see tag accuracy evaluations).
4) Contrastive learning phase with curriculum learning training improves the quality on document-level inputs, but only on neural metrics. We hypothesize that curriculum learning model increases fluency of the translations and introduces more complicated paraphrases that BLEU fails to score adequately.

WMT'24 Results

The quality of our system is assessed by the organizers using the following metrics:
MetricX-23-XL (Juraska et al. (2023)) – a reference-based metric built on top of the mT5 model. CometKiwi-DA-XL (Rei et al. (2023)) – a quality estimation metric built on the XLM-R XL model. Both metrics are among the top-performing metrics in the field (Freitag et al. (2023)). According to these metrics, our system currently ranks third on the leaderboard, with a MetricX score of 2.9 and a CometKiwi-DA-XL score of 0.705. The final leaderboard will be determined based on human evaluation results.

Ethics Statement

Our system was trained on the publicly available data. This unrestricted access to data allowed us to leverage a vast and diverse set of examples, enabling the model to learn from a wide array of linguistic patterns, contexts, and domains.
The absence of data limitations contributed to the development of a robust and versatile model, capable of generalizing well across various tasks and applications. By incorporating extensive datasets

from different sources, our system gained the ability to handle complex and varied scenarios, enhancing its overall performance and adaptability.
This approach ensured that the model could effectively capture and respond to the nuances of different data types, ultimately leading to more accurate and reliable outputs in real-world applications.

Acknowledgements

We are grateful to the organizers for providing a challenging dataset that allowed us to apply and evaluate our subtitle model.

References

Y. Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. volume 60, page 6.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu - neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv*, abs/2110.07602.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. pages 2685–2702.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.