Challenges in Machine Translation of Interactive Multimodal Exercises

Lucie Poláková, Miroslav Hrabal, Věra Kloudová,
Michal Novák, Mariia Anisimova and Martin Popel
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
{surname, mnovak}@ufal.mff.cuni.cz

Abstract

This paper describes linguistic and technological challenges encountered within an applied project aimed at expanding a large e-learning portal from its original Czech to three other languages: Ukrainian, English and German. Although there seems to be a general belief that machine translation is a solved task in 2026, we show that translating educational content, which in our case is highly terminological, multimodal, interactive and encoded in XML, brings along many challenges of different types, some easily solvable and some not. We also compare our results from the early phase of the project (Transformer-based machine translation) with those after the switch to the LLM-based translation methods. We show that both MT methods are prone to different types of errors, some of which are quite new (such as the undesired correction of counterfactual statements) and require new ways of handling them. The resulting four-language edition of the educational web portal will be freely available to educators, students and researchers by the end of 2026.

1 Introduction

Online learning and hybrid learning, which integrate traditional textbooks with online practice, are becoming increasingly widespread. In addition to many other products and services, major textbook publishers nowadays routinely link their new print textbooks to supplementary material and exercises available online, often via QR codes (e.g. Nadziro et al., 2023, Kempe and Grönlund, 2019). In the Czech educational ecosystem, this gradual shift is occurring alongside a notable increase in non-native speakers within the school system, particularly following the recent arrival of Ukrainian pupils and students (Popel et al., 2024). Ensuring that every learner can access education in a language they understand, at least before they are capable of education in the language of the host country, is a pressing concern and a key technological challenge. Machine translation (MT) offers a promising way to extend the reach of educational content, though it still faces practical limitations.

In this paper, we explore a case study based on an product development project where multiple translation methods were applied to interactive web exercises, examining the advantages and constraints of these approaches in a real-world educational setting. In addition to Ukrainian as a target language, we focus on English and German as the two most frequent second languages in Czechia.

The paper has two goals: (i) to report and analyze the results of the project that have not yet been published, which in particular include the **switch** from traditional MT methods **to methods based on large language models (LLMs)**; to compare these results on Ukrainian; to report and analyze the results on English and German; and (ii) to describe and analyze the challenges of the chosen task, the key feature of which is **the translation of XML that encodes interactive web exercises with multimodal content.** We identify and classify the linguistic/translatological challenges as well as the technical ones, determine their origins and suggest possible solutions. The description of our experiments could provide valuable insights to the EdTech community about the recent pros and cons of using open source LLMs in domain-specific translation tasks.

The paper is structured as follows: §2 reviews recent developments in the field of MT. §3 describes the data and §4 presents the methodology we employ. In §5, we assess translation quality using automatic metrics, followed by human evaluation. §6 forms the core of the paper, providing a classification and a discussion of the challenges related to our task, along with examples. §7 concludes the paper. Appendix A presents additional examples of mistranslations organized by type; Appendix B offers details for the in-image translation task.

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 141–152

2 Machine Translation

The MT field has undergone a fundamental architectural shift, transitioning from the long-standing dominance of encoder-decoder Transformers (Vaswani et al., 2017) to the adoption of decoder-only Large Language Models (LLMs). Historically, specialized encoder-decoder systems, such as the No Language Left Behind (NLLB) model (NLLB Team et al., 2022), were preferred for their ability to be trained on specific, non-English parallel corpora. For language pairs like Czech-Ukrainian, these dedicated systems achieved superior results by performing direct translation, thereby avoiding the translationese and loss of gender and politeness information associated with traditional pivoting through English (Popel et al., 2024). This architectural specificity allowed for high-fidelity mappings in low-resource or linguistically distant directions and domains that general-purpose models often struggled to capture.

However, recent findings from the WMT25 General Machine Translation Shared Task (Kocmi et al., 2025) indicate that the performance gap between dedicated NMT systems and LLMs has largely closed. Modern LLMs now consistently place in the winning clusters across a vast majority of language directions.[1] The historical advantage of encoder-decoder systems is rapidly vanishing as modern LLMs become natively multilingual, encoding diverse linguistic structures within a unified semantic space during pre-training.

Current research now focuses on refining these LLM-based systems through advanced alignment techniques that go beyond standard Supervised Fine-Tuning (SFT). A notable development is Contrastive Pairwise Optimization (CPO), which trains models to distinguish between gold-standard references and translations that are adequate but imperfect (Xu et al., 2024). By leveraging preference data rather than simple maximum likelihood estimation, CPO enables moderate-scale LLMs to match or exceed the performance of both proprietary models and traditional competition winners.

Translating structured educational content necessitates the high-fidelity preservation of formatting markup, such as HTML, XML or DOCX. This task is sometimes called *label projection*, especially when used for the cross-lingual transfer of span-annotated NLP datasets with relatively simple labels (i.e. formatting markup) from a high-resource to a low-resource language. There are two main approaches to the task. The first approach involves the joint translation of text and labels, a method that has gained significant traction with the advent of LLMs. By embedding labels, such as XML tags (Thennal et al., 2026) or square brackets (Chen et al., 2023), directly into the input sequence, models learn to generate formatted output in a single pass. While effective for simple entity markers, these joint systems often struggle with the structural integrity required for complex markup, such as nested online exercises in XML formats. Conversely, the second approach, which we use, utilizes a modular two-step pipeline where the plain text is translated first, followed by a separate projection of the labels using word alignment or forced decoding (Parekh et al., 2024). This decoupled strategy remains robust for contemporary localization workflows for structured documents (often using the XLIFF format and Translation Management Systems) as it allows for arbitrary pre- and post-processing of the text and labels independently, ensuring that the structural validity of the markup is maintained regardless of the translation model's complexity (Zenkel et al., 2021).

3 Data

The educational web portal we work with, *Škola s nadhledem*,[2] currently contains more than 9,000 interactive exercises designed for practicing curriculum material from the first years of primary school through to the final high-school year. They cover a wide range of school subjects, including foreign languages and computer science. In terms of interactivity, the exercises include many different types, such as drag-and-drop tasks, yes/no quizzes, multiple-choice questions, gap-filling activities, memory games, crosswords, mind maps, timelines, etc. They also work with various media, so in addition to text, they may contain images, audio and other elements. In the past three years, the portal has attracted over 1.5 million users, with more than 100,000 students practicing on it each month. It is also part of the hybrid approach to educational materials, combining printed textbooks or workbooks with interactive exercises that provide

[1] In WMT25, LLMs *were the predominant approach, used by all but one external submission* (Kocmi et al., 2025). *Overall, the best-performing model in the human evaluation is Gemini 2.5 Pro [...]. It places in the top cluster for 14 of the 16 evaluated language pairs and is on par with or surpasses human translation in 10 of those pairs.* (Kocmi et al., 2025).

[2] https://skolasnadhledem.cz

immediate feedback. Users, regardless of whether they use the company's printed textbooks at school, can freely practice topics of their choice.

Some of the exercises are not suitable for translation. For example, reading exercises for the youngest learners, such as building words out of letters or syllables, anagram-based games or crosswords. The selection of content for translation was therefore based on the general suitability of each exercise for translation across different subjects and grade levels. So far, we have worked with portal content covering approximately 1,000 preselected exercises from subjects like history, geography, math, physics, biology, chemistry, civics, etc., and we are gradually expanding this scope by adding other suitable thematic areas.

From the translation perspective, the web portal data is highly specific: exercises are typically very short, rarely provide multi-sentence context, include many domain-specific terms, and are formatted in very complex HTML, with images or other elements. These factors make the task challenging even for state-of-the-art MT systems.

Standard MT evaluation requires reference data, i.e., human translations for comparison. We therefore created an evaluation dataset by manually translating 396 selected exercises from biology, chemistry and geography, split into a development set (190 exercises) and a test set (206 exercises). Both sets were translated into Ukrainian, English, and German by professional translators.

4 Methodology

4.1 Collecting data for domain adaptation

In order to adapt our system for translation, we need to obtain parallel bilingual and monolingual in-domain data in the target language. Parallel data can be used directly during fine-tuning, while monolingual data must be back-translated into the source language first.

For Czech-English, we used two data sources: publicly available parallel corpus CzEng 2.0 (Kocmi et al., 2020) and back-translated excerpts from English and Czech Wikipedia.

For Czech-Ukrainian, we used the dataset collected in our previous work (Poláková et al., 2025).

For Czech-German, we identified several online sources that could contain texts from the educational domain (e.g. www.goethe.de, www.oegp.cz, www.rkfpraha.cz). Using the Bitextor tool (Esplà-Gomis, 2009), we downloaded publicly available pages from these websites, extracted the texts, aligned them based on dictionary and metadata, and then segmented, paired, and filtered them at the sentence level. The result of this process was a dataset containing 57,000 sentence pairs. The monolingual data were extracted from www.oebv.at using custom scripts tailored to this source (targeted textbook collection, filtering out foreign-language textbooks), obtaining 1.2 million German sentences from the educational domain. We further augmented the data with back-translated excerpts from German and Czech Wikipedia.

4.2 Model adaptation

We used a locally installed EuroLLM-9B-Instruct LLM (with 9 billion parameters) for translation from Czech into all three target languages. As full fine-tuning of such a large model would be computationally very demanding, we used the *Low-Rank Adaptation* (LoRA) method for fine-tuning, in which most of the model's weights remain frozen and smaller adapters, consisting of an order of magnitude fewer parameters, are trained for individual linear layers of the model.

We used three different methods for fine-tuning the models: Supervised Fine-Tuning (SFT), Contrastive Preference Optimization (CPO, Xu et al., 2024) and Simple Preference Optimization (SimPO, Meng et al., 2024). For SFT, we simply used the collected parallel data. For CPO and SimPO, parallel data alone are insufficient; a preference dataset consisting of triplets (source text, preferred translation, non-preferred translation) is required. We therefore also translated the parallel data using the model after the SFT phase. We selected the preferred translation using a combination of publicly available *quality estimation* models (MetricX, Juraska et al., 2024) and publicly available LLMs (Gemma-3-27B-it). Preliminary experiments confirmed that CPO and SimPO bring the best results when applied on top of SFT (we coin these systems SFT+CPO and SFT+SimPO, respectively).

5 Results

The experiments were evaluated intrinsically using automated metrics (§5.1) and human evaluation (§5.2), based on manually translated test sets. The system selected for publication was then evaluated extrinsically within the web portal for all three languages (§5.3). The project is primarily an

applied product-development effort rather than a purely research-oriented study, hence the evaluation aim was to develop a functional educational product and identify exercises suitable for real deployment in the target platform. From this perspective, the evaluations described below thus focused on practical criteria such as pedagogical usefulness and production readiness, serving product-selection and quality-assurance rather than establishing annotator-independent ground truth.

5.1 Automatic Metrics

Table 1 shows the automatic evaluation on the development sets for each language pair using two automatic metrics: chrF2 (Popović, 2015) and MetricX24 (Juraska et al., 2024).[3] System B using SFT+CPO was evaluated as the best one in all three language pairs according to both automatic metrics.

5.2 Intrinsic Manual Evaluation

In the intrinsic manual evaluation, the annotators saw side by side the source Czech text, translations from three (or four, in the case of Ukrainian) MT systems (anonymized and presented in random order) and the human reference translation. The evaluation focused primarily on selected, more difficult or longer passages – i.e., those that were linguistically complex. Experts in the target language (native speakers for Ukrainian) evaluated these segments on a 0–10 scale (0 being the worst, 10 being the best). Since the goal of this evaluation phase was to select the best-performing system, the evaluators skipped lines where they would assign the highest scores (10) to all three systems. As a result, evaluation results cannot be used to determine the absolute performance of individual systems (nor to compare across languages); rather, they represent a relative comparison of the systems against one another.

For Ukrainian (Table 2), we compared a standard encoder-decoder NMT system (the pre-LLM Charles Translator, Popel et al., 2024) with the following three systems: SFT (A), SFT+CPO (B), and EuroLLM-9B-Instruct (C) on 1800 segments. The three LLM-based systems are substantially better than the pre-LLM system; SFT+CPO is significantly the best system ($p<0.05$, paired t-test).

For German and English (Table 3), all translations were of very high quality, and the results of the individual systems did not differ significantly from one another in either language. Approximately 500 segments were evaluated for each language. For German, we used systems A, B and C (we have not trained a Pre-LLM system). For English, we replaced A (which had the lowest score in German) with a newly trained SFT+SimPO system (D). Systems B and C were the best, with no significant difference. We have selected system B (SFT+CPO) for deployment for both languages.

5.3 Extrinsic Manual Evaluation

The extrinsic human evaluation phase consisted of testing the functionality of 540 translated exercises within the web portal. The procedure was quite straightforward: the evaluator "played the game" and tested several aspects of the exercise at once: translation quality and functionality within the context of the given exercise/task, consistency and technical aspects such as switching among the four languages at any step, correct visualization, correct exercise scoring and display of the correct solution (and also correct solution explanation, where present). There was an easy three-way division: *good* (the exercise has no issues, it works well and can be directly published), *okay* (there are some minor issues that do not affect the functionality of the exercise and mostly can be fixed easily,[4] see §6), *bad* (the exercise has serious issues and cannot be published in its current form)[5]. Additionally, there was a *not applicable* option for exercises where it was determined that their translation does not make sense, e.g., completing morphological suffixes in Czech, disambiguating homonyms, etc.

In this way, typical patterns of dysfunction were revealed, categorized and they are gradually solved. After the functionality of the exercises is refined, a subset that is already working well is deployed to the production portal accessible to the public. This process is iterative.

The results of the extrinsic human evaluation in Table 4 show that 90% of the translated exercises were evaluated as either *good* or *okay* for Ukrainian, while for English and German, the rates were 75% and 69% respectively. However, the results are not directly comparable across languages, as the translations were evaluated by different evaluators.

[3] Specifically, we used the QE version of https://huggingface.co/google/metricx-24-hybrid-xl-v2p6, without the "reference" parameter.

[4] such as non-standard use of upper- and lowercase letters, punctuation or special signs, some minor inflection errors

[5] incorrect translations, missing translations or other technical issues (see above) that would prevent the user from successfully completing the exercise

System	German		English		Ukrainian	
	chrF2↑	MetricX24↓	chrF2↑	MetricX24↓	chrF2↑	MetricX24↓
Pre-LLM Charles Translator					63.1 ± 0.8	3.73 ± 0.11
A: SFT	65.6 ± 0.7	3.30 ± 0.10	68.5 ± 0.7	5.04 ± 0.10	62.1 ± 0.8	3.59 ± 0.08
B: SFT+CPO	**67.1 ± 0.7**	**2.90 ± 0.09**	**73.1 ± 1.0**	**4.68 ± 0.10**	**64.6 ± 0.7**	**3.33 ± 0.08**
C: EuroLLM-9B-Instruct	65.6 ± 0.8	3.15 ± 0.10	71.6 ± 0.7	4.80 ± 0.10	63.5 ± 1.0	3.59 ± 0.09
D: SFT+SimPO	–	–	72.9 ± 0.7	4.78 ± 0.10	–	–

Table 1: Translation quality measured with automatic MT metrics chrF2 (↑ = higher is better, with 95% CI intervals calculated using 1000 bootstrap samples) and MetricX24 (↓ = lower is better). The best score is in **bold**.

System	Ukrainian↑
Pre-LLM Charles Translator	8.20 ± 2.6
A: SFT	8.75 ± 2.2
B: SFT+CPO	**8.81 ± 2.1**
C: EuroLLM-9B Instruct	8.68 ± 2.2

Table 2: Human evaluation results for Czech to Ukrainian, average on the scale 0–10 and standard deviation. The significantly ($p<0.05$) best score is in **bold**.

System	German↑	English↑
A: SFT	**7.81 ± 3.2**	no eval
B: SFT+CPO	**7.96 ± 3.1**	**8.16 ± 3.1**
C: EuroLLM-9B-Instruct	**7.95 ± 3.1**	**8.19 ± 2.9**
D: SFT+SimPO	not used	8.00 ± 3.1

Table 3: Human evaluation results for Czech to German and Czech to English, average on the scale 0-10, higher is better ↑. The best score and other scores that are not significantly ($p<0.05$) worse are in **bold**.

6 Analysis of the Results/Challenges

The results of the manual evaluations (§5.2 and §5.3) show that the translation systems perform well in less contextualized translation scenarios. In the educational domain, these are usually topics related to society, family, culture, and traditions. Another aspect is text complexity: as expected, translation quality was higher for less advanced exercises. In fact, it was very high and these exercises were ready to use (no further editing) even for more specialized topics such as mathematics, history, geography, etc. The relatively low proportion of well functioning exercises in German (see Table 4) is most likely attributed to its typological specificity in connection with a special type of exercise (filling gaps), see §6.3.

Quality	DE (%)	EN (%)	UK (%)
good	43	46	54
okay	26	29	36
bad	25	17	4
not applicable	7	7	6

Table 4: Percentage distribution in human evaluation of the demo portal for German (DE), English (EN), and Ukrainian (UK). Evaluated exercises in total: 522 (DE), 523 (EN), 540 (UK).

6.1 Typical errors in pre-LLM and LLM-based MT systems

Based on our extensive experience in evaluating traditional MT systems,we have a good understanding of the typical errors made by machine translators before the LLM era. Also, many studies have already examined machine translation outputs, focusing on the typology of errors or the typology of changes made by post-editors to the output of machine translation, e.g. Kloudová et al. (2021).

Typical errors in standard MT include failing to distinguish between homonyms in the given context, inconsistency in using gender, spelling of foreign words adopted from the original language, improper collocations, preserving the number of nouns (plural vs singular) and word-to-word translations that do not reflect the meaning of the original, to name just the most common ones.

According to the evaluations performed, new types of errors were identified that differ in certain aspects from those made by standard MT. The following subsections provide a comprehensive, data-driven classification of these errors.

6.2 Lexical challenges

6.2.1 Translation of domain-specific terms

A quite significant type of error identified in translations to German and English, similarly to errors de-

tected earlier in the Ukrainian translation (Poláková et al., 2025), is the translation of domain-specific terms. See Example 1 or Appendix A. This problem occurred primarily in biology. These terms are often unfamiliar not only to the average native Czech speaker but also to professional translators without specific knowledge of the given scientific field.[6]

(1) **SRC$_{cs}$:** *ortorula*
MT$_{en}$: orthorhombic
HT$_{en}$: orthogneiss
MT$_{de}$: Orthorula
HT$_{de}$: der Orthogneis

Currently, our system cannot fully handle these cases. We have been able to replicate these terminology errors in commercial machine translation systems and large language models such as Google Translate, DeepL and ChatGPT up to its free version 5 mini. In order to achieve a more successful translation of domain-specific terminology, we are currently experimenting with different approaches.

We are extending the system with optional terminology constraints via external glossaries. Glossaries allow us to curate in-domain terms independently and quickly introduce missing entries identified during human evaluation, without the need to retrain the underlying model. Building these glossaries at scale remains an open engineering task, and we are exploring sources such as bilingual titles from WikiData and term pairs mined from our parallel training data.

6.2.2 Homonymy

Our predominantly sentence-level (or, due to the nature of our data, sometimes phrase-level) segmentation can reduce the ability to disambiguate homonymous expressions. This is a broad issue with impact on the whole vocabulary we translate, including terminology. Solving lexical homonymy without a sufficient context is known to be an impossible task in theory. At the same time, lexical errors are particularly serious because they substantially change meaning and impede understanding. Very often, because of a single lexical error, the whole exercise is not functional, e.g. in Example 2:

(2) **SRC$_{cs}$:** *Má svíčku.*
MT$_{en}$: She has a candle.
HT$_{en}$: It (the engine) has a spark plug.

[6]In the following examples, we use **SRC$_{cs}$** for the source Czech text, **MT** for machine translation by the published SFT+CPO system and **HT** for the reference human translation.

Dealing with homonymy each time anew also sometimes leads to translation inconsistencies within one exercise. For instance, the Czech word *předmět* has multiple meanings (mostly *object*), but there is a high probability it should keep the same meaning within each exercise. Nevertheless, several different translations of this word were documented in the English version of one exercise: *object*, *subject*, *item*, *product*. Similarly, in the German version: *Schulprojekt*, *Objekt*, *Gegenstand*.

To mitigate these issues, we are planning to modify the translation process to operate on larger chunks of exercise segments jointly. We expect this to enable the model to select the appropriate semantic variant of a word based on context. According to our preliminary small-scale experiments on several such examples, this indeed does help; however, we need to conduct additional testing to ensure that this does not degrade the quality of the translations in different ways.

6.2.3 Non-existent words

Words that do not exist in the target language (e.g. non-existent compounds, non-standard spellings, see Example 3 or Appendix A) have been encountered in standard MT. Their impact on readers' comprehension of the text was examined (Macken et al., 2019). Based on our analysis of exercises translated by the new LLM-based systems, we identified a relatively large number of newly created words in a comparatively small sample of texts.

(3) **SRC$_{cs}$:** *valcha*
MT$_{en}$: valcharium
HT$_{en}$: washboard

However, a quantitative study confirming a higher incidence of these inventions, or their different structure (see the Latin suffix in Example 3), in LLM-based translation has yet to be conducted.

6.2.4 Mistranslations based on form similarity

Another recurring phenomenon are mistranslations that apparently arose due to the similarity in form between unrelated words (Examples 4–5).

(4) **SRC$_{cs}$:** *kopí*
MT$_{en}$: kopy
HT$_{en}$: spear
MT$_{de}$: kopieren
HT$_{de}$: Speer

(5) **SRC$_{cs}$:** *žně*
MT$_{en}$: gently
HT$_{en}$: harvest
MT$_{de}$: herzlich
HT$_{de}$: Ernte

In these cases, the incorrect forms appear to have been triggered by superficial similarity: Example 4 by confusion with *kopírovat* (*to copy*), Example 5

by interference from the Czech *něžně* (*gently*). For more such cases, see Appendix A.

6.2.5 Untranslated words

In some cases, the translation failed for very specific groups of words, which remained untranslated. It is words that are in the source language: (i) very short (*var – boiling point, sup – vulture*), (ii) across-language homographs (*city – feelings, rosa – dew, pink* in German), or they are (iii) very rare, again almost terminological concepts (*otava – aftermath* in agriculture, *plebejský – plebeian*, *řemdih – flail*, a medieval weapon, *prvohory – paleozoic*).

6.3 Syntactic challenges

Czech allows flexible word order, whereas English and, to some extent, German follow more fixed patterns. This makes the translation of some exercises, especially gap-filling, challenging. If the words in the gap are omitted in the source sentence, a correct translation may be impossible even for human translators in some cases. Luckily, our data contain the correct option and several incorrect ones for each gap. However, we need to preserve the (complex) XML markup for each gap as well, which is also challenging, as shown below.

6.3.1 Word order in gap-filling

A naive approach that translates each XML element independently often fails because it cannot produce the correct target-language word order:

(6) **SRC$_{cs}$:** *Z [vodní páry] vznikají oblaky.*
MT$_{en}$: From [water vapor] clouds are formed.
HT$_{en}$: Clouds are formed from [water vapor].

6.3.2 Compound alignment in gap-filling

Even if we translate the whole sentence without tags and then apply label projection using word alignment (see §2), the task is challenging because the alignment is not always one-to-one.

(7) **SRC$_{cs}$:** *Severní Ameriku od Asie odděluje [Beringův] průliv.*
HT$_{de}$: Nordamerika ist durch die [Bering]straße von Asien getrennt.

The two words *Beringův průliv* (Bering Strait) should be translated as a single German compound word, *Beringstraße*. The web portal may not be ready to present gaps as part of a word (in our case, it renders a drop-down menu followed by a space). An alternative is to translate the word in the gap with a final hyphen: *Bering-*, but this decision is based on the final presentation capabilities.

6.3.3 Non-monotonic alignment in gap-filling

In German, parts of verbs (such as infinitives in modal constructions or participles) tend to appear in sentence-final position (e.g., *odděluje* (*separates*) is translated as *ist ... getrennt* in Example 7). This can result in long-distance reordering and possibly in a different order of the gaps (e.g., if *odděluje* would be a gap). The XML has ID attributes attached to each gap, but the web portal may not be adapted to accommodate different orders of the gaps in each language.

6.3.4 Distractors in gap-filling

Originally, we applied a preprocessing method of gap-filling XML markup where the correct option is translated within the sentence, while the incorrect options (distractors) are translated separately afterwards. This improved the translation quality, both in terms of the form (inflection in Ukrainian and German) of the correct option as well as the rest of the sentence. However, the incorrect options were still often translated incorrectly due to the lack of context. In preliminary experiments, we tried to translate the incorrect options within the context of the whole sentence as well, but we noticed an increased number of mis-corrected counterfactual statements (see §6.4.1 below). Additionally, we have noticed that the rest of the sentence may be translated differently for each option and using the correct-option context may provide hints that make the exercise too easy, e.g. by revealing the grammatical gender of the correct option.

6.4 Other challenges

The following issues are either closely related to the translation method, like those in previous sections, or they are connected to the purpose of the translation (interaction with images in exercises).

6.4.1 Counterfactual statements

There are exercises with yes/no questions that contain counterfactual (i.e. false) statements. However, the LLM-based systems tend to correct false statements during translation:

(8) **SRC$_{cs}$:** *Prvoka trypanozómu, který vyvolává spavou nemoc, přenáší* ***octomilka obecná****.*
MT$_{en}$: The trypanosome protozoan that causes sleeping sickness is transmitted by the ***tsetse fly***.

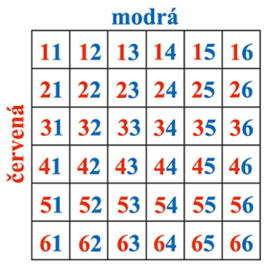

Figure 1: Example of embedded Czech text requiring in-image translation (*We roll a red die and a blue die. Determine the probability that the sum is at least 10.*)

HT_{en}: ... by the ***fruit fly***.

This way, the originally false statement (*trypanosome is transmitted by the fruit fly*) is converted into a true statement (*trypanosome is transmitted by the tsetse fly*), which makes the whole translated yes/no exercise harmful for the students. Our plan is to instruct the LLMs in the system prompt that false statements should not be corrected and to provide the whole exercise for translation (including instructions such as "Are the following statements true?").

6.4.2 Hallucinations

We encountered cases where the LLM-based translation completes an unfinished statement, which spoils the gap-filling exercise:

(9) **SRC_{cs}:** *Nad psaným textem většinou najdeme titulek neboli...*
MT_{de}: Über dem geschriebenen Text finden wir meist einen Titel oder eine Überschrift. *(lit. 'Above a written text, we usually find a title or* ***heading****.')*
HT_{de}: Über dem geschriebenen Text finden wir meist einen Titel oder...

This issue, similarly to the non-existent words (§6.2.3) and counterfactuals (§6.4.1), may be mitigated by adapting the LLM system prompt. However, we noticed that our SFT (and CPO/SimPO) fine-tuning decreases the model's capability to follow general instructions.

6.4.3 In-Image Translation

Exercises are often accompanied by images. If an image is purely decorative, it can usually be ignored during translation. However, some images, such as in Figure 1, contain embedded text (e.g., labels, captions, or instructions) that is necessary to solve the exercise. In such cases, the pipeline should apply in-image machine translation (MT) so that learners see the original visual content with the embedded text rendered in the target language.

In preliminary experiments, we examined whether low-parameter open-source Vision Language Models (VLMs) can detect images whose embedded text should be translated. On 1,000 images randomly sampled from the exercise image store, the best-performing model reached 84% recall (F1 = 0.64). Given that fewer than 7% of the sample actually required translation, this level of recall suggests the detector can surface most relevant images while filtering out the large majority that do not warrant resource-intensive in-image MT. See Appendix B for details.

For in-image MT, we plan to use more capable VLMs to extract the embedded text, translate it, and render the translation back into the image. Alternatively, one could generate a translated image in a single step, but this risks modifying non-textual content.

7 Conclusion

We presented a project that employs various MT methods for translating educational content – interactive web-based exercises – from Czech to Ukrainian, English, and German. We described several experiments to achieve the best translation quality using open source LLMs (EuroLLM-9B-Instruct), in particular the process of fine-tuning and further experiments on the given domain. We selected the three most promising systems using automatic MT metrics. Two rounds of manual evaluations helped to select the best system (SFT+CPO) and observe its performance in a real-world scenario within the web portal. We were able to significantly improve our original Pre-LLM system for Czech-Ukrainian in terms of both automatic (+1.5 chrF2, -0.4 MetricX24) and human (+0.61 score on a 0–10 scale) evaluation, thus confirming the superiority of LLMs with SFT+CPO adaptation for the education domain.

The second goal of the paper was to conduct an in-depth analysis of the translation errors, along with a description of the advantages and limitations of the approach chosen for our use case. Translating such a web portal required a substantial amount of engineering work as well as experimentation,

particularly in the area of terminology, word order and syntax in gap-filling exercises, and in-image translation. These issues have not yet been fully resolved, but we have outlined the next steps to address them. We have classified the translation errors linguistically: some types of lexical errors are the most significant issue and can occasionally impair the functionality of the entire exercise. At the same time, we tried to distinguish between traditional, well-documented MT errors and newly emerging ones, which are more typical of LLM-based translation methods. These errors include hallucinations, such as unintended corrections of false statements, or completing unfinished statements.

The translation systems were published in 2025 in the LINDAT/CLARIAH repository, the release consists of two items: a package of three machine translation models adapted for educational use: Czech-Ukrainian, Czech-English, and Czech-German (Hrabal et al., 2025)[7] and a second package with an improved version of the translation software required for the newly released models (Popel et al., 2025)[8]. This package includes three tools: a web interface (Charles-translator-web-frontend) for machine translation with phonetic transcription of Ukrainian suitable for Czech speakers, an API server, and a tool for translating tagged documents including html, docx, odt, pptx, pdf, and odp. We will release updated system versions during 2026 in the same way. Furthermore, by the end of 2026, translated exercises will be made freely available to the public via Fraus-Klett educational web portal *Škola s nadhledem*.

Limitations

An obvious limitation is the disparity in capabilities between open-source and the newest commercial models. It can certainly be expected that all types of models will continue to improve quickly in tasks such as machine translation, and that current versions of large commercial models will perform better. On the other hand, data confidentiality is an issue. In our approach, all data remains on our servers and those of our partner company. Second, as far as we could tell, some translation issues remain unsolved in large commercial models, mainly concerning translations to or from small languages and highly specialized terminology (see 6.2.1).

As stated in §5, the evaluation served mainly as a quality-assurance and decision-making mechanism within the product development cycle, rather than as an attempt to establish annotator-independent ground truth or to measure reproducibility of theoretical judgments. For this reason, inter-annotator agreement was not treated as a central methodological requirement. In this applied evaluation context, practical effectiveness and deployment readiness took precedence over annotation-consistency metrics typical of hypothesis-driven research.

Ethical Considerations

This research was conducted in accordance with the ACM Code of Ethics and Professional Conduct. The work does not involve harmful applications, and human annotation was performed with informed consent and fair compensation. We aim to ensure transparency, reproducibility, and responsible release of resources.

Acknowledgments

This work was supported by the project TQ01000458 (EdUKate) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Programme. It has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2023062).

We would like to thank our colleagues at Fraus-Klett Publishing: Adam Jelínek, Barbora Bartošová, Matěj Sutr, Barbora Tvarohová and Kateřina Berková for our ongoing cooperation.

References

Nuaman Chen, Li Zhou, and Zilong Wang. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4500–4515, Singapore. Association for Computational Linguistics.

Miquel Esplà-Gomis. 2009. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.

Miroslav Hrabal, Martin Popel, Lucie Poláková, Michal Novák, Věra Kloudová, and Mariia Anisimova. 2025. EdUKate translation models 2025.

[7] http://hdl.handle.net/11234/1-6032
[8] http://hdl.handle.net/11234/1-6033

LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Workshop on Machine Translation (WMT 2024)*.

Anna-Lena Kempe and Åke Grönlund. 2019. Collaborative digital textbooks–a comparison of five different designs shaping teaching and learning. *Education and Information Technologies*, 24(5):2909–2941.

Věra Kloudová, Ondřej Bojar, and Martin Popel. 2021. Detecting post-edited references and their effect on human evaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 114–119, Online. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondřej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *Preprint*, arXiv:2007.03006.

Lieve Macken, Laura Van Brussel, and Joke Daems. 2019. NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. *Computational Linguistics in the Netherlands Journal*, 9:67–80.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Maslahatun Nadziro, Aufia Aisa, and Rina Dian Rahmawati. 2023. The development of QR code-based Arabic textbooks for non-arabic education students. *Scaffolding: Jurnal Pendidikan Islam dan Multikulturalisme*, 5(2):960–981.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Lucie Poláková, Martin Popel, Věra Kloudová, Michal Novák, Mariia Anisimova, and Jiří Balhar. 2025. Mitigating language barriers in education: Developing multilingual digital learning materials with machine translation. In *EDULEARN25 Proceedings*, pages 8754–8760, Valencia, Spain. IATED.

Martin Popel, Michal Novák, Jiří Balhar, Ondřej Košarko, Jiří Mayer, Lucie Poláková, Věra Kloudová, and Mariia Anisimova. 2025. EdUKate translation software 2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Popel, Lucie Poláková, Michal Novák, Jindřich Helcl, Jindřich Libovický, Pavel Straňák, Tomáš Krabač, Jaroslava Hlaváčová, Mariia Anisimova, and Tereza Chlaňová. 2024. Charles Translator: A Machine Translation System between Ukrainian and Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3038–3045, Torino, Italy. European Language Resources Association.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

D. K. Thennal, Chris Biemann, and Hans Ole Hatzel. 2026. Just Use XML: Revisiting Joint Translation and Label Projection. *Preprint*, arXiv:2603.12021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

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-based machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2021. Automatic bilingual markup transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3524–3533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Appendices

A Additional Examples of Mistranslations by Type

A.1 Lexical challenges:

Terminology (6.2.1):

(10) **SRC$_{cs}$:** *svor*
MT$_{en}$: svor
HT$_{en}$: mica schist
MT$_{de}$: Schlüsselwort
HT$_{de}$: der Glimmerschiefer

(11) **SRC$_{cs}$:** *motýlice obecná*
MT$_{en}$: common hawker
HT$_{en}$: Beautiful demoiselle
MT$_{de}$: Großer Blaupfeil
HT$_{de}$: Blauflügel-Prachtlibelle

(12) **SRC$_{cs}$:** *candát obecný*
MT$_{en}$: European pike
HT$_{en}$: zander
MT$_{de}$: der Europäische Wels
HT$_{de}$: der Zander

Homonymous expressions (6.2.2):

(13) **SRC$_{cs}$:** *sloupek*
MT$_{de}$: die Säule *(lit. 'pillar')*
HT$_{de}$: die Kolumne *(lit. 'journalistic column')*

(14) **SRC$_{cs}$:** *kůže*
MT$_{de}$: die Haut *(lit. 'skin')*
HT$_{de}$: das Leder *(lit. 'leather')*

(15) **SRC$_{cs}$:** *pěvci*
MT$_{de}$: Sänger und Sängerinnen *(lit. 'singers')*
HT$_{de}$: Singvögel *(lit. 'songbirds')*

Non-existent words (6.2.3):

(16) **SRC$_{cs}$:** *obkládat / dlaždičkovat*
MT$_{de}$: flieschen
HT$_{de}$: fliesen / mit Fliesen belegen *(lit. 'to tile')*

(17) **SRC$_{cs}$:** *směsný komunální odpad*
MT$_{de}$: Mischmüllgemisch
HT$_{de}$: Restmüll *(lit. 'residual waste')*

(18) **SRC$_{cs}$:** *Zakavkazsko*
MT$_{de}$: Zakavkazien
HT$_{de}$: Südkaukasus *(lit. 'South Caucasus')*

(19) **SRC$_{cs}$:** *traktor*
MT$_{de}$: Tractur
HT$_{de}$: der Traktor *(lit. 'tractor')*

Mistranslations based on form similarity (6.2.4):
In Example 21, the incorrect translation appears to have been triggered by confusion with *opakování* (*repetition*), Example 22 perhaps by interference from Slovak *len* (*only*), Example 23 by similarity to *míchat* (*to mix, to stir*), and Example 24 perhaps by the Czech abbreviations for *sobota* and *neděle* (Samstag and Sonntag).

(20) **SRC$_{cs}$:** *city*
MT$_{en}$: city
HT$_{en}$: feelings
MT$_{de}$: Stadt
HT$_{de}$: Gefühle

(21) **SRC$_{cs}$:** *opak*
MT$_{de}$: Wiederholung *(lit. 'repetition')*
HT$_{de}$: Gegenteil *(lit. 'opposite')*

(22) **SRC$_{cs}$:** *len*
MT$_{de}$: nur *(lit. 'only')*
HT$_{de}$: Flachs / Leinen *(lit. 'flax / linen')*

(23) **SRC$_{cs}$:** *mícha*
MT$_{de}$: mix *(lit. 'mix')*
HT$_{de}$: Rückenmark *(lit. 'spinal cord')*

(24) **SRC$_{cs}$:** *sob polární*
MT$_{de}$: Sonntag polare *(lit. 'polar Sunday')*
HT$_{de}$: Rentier *(lit. 'reindeer')*

A.2 Syntactic challenges (6.3):

(25) **SRC$_{cs}$:** *Znečištění vzduchu [může] pocházet z lidské činnosti.*
MT$_{en}$: Air pollution [can] originating from human activity.
MT$_{de}$: Luftverschmutzung [kann] entstammen menschlicher Aktivität.

(26) **SRC$_{cs}$:** *Společnými znaky strunatců je [páteř/struna] hřbetní, [trubicovitá/rozptýlená] nervová soustava a [otevřená/uzavřená] cévní soustava.*
MT$_{en}$: The common characteristics of chordates are [string/spine] dorsal, [scattered/tubular] nervous system and [open/closed] vascular system.
MT$_{de}$: Gemeinsame Merkmale der Chordatiere sind: [Wirbelsäule/die Saite] rückseitig, [röhrenförmig/zerstreut] das Nervensystem und [offen/geschlossen] das Gefäßsystem.

B Detecting Images for Translation

Some images used in our online exercises contain embedded text that should be translated. To avoid unnecessary computation for in-image MT, it is desirable to automatically detect and filter such images. In principle, this decision may depend on the surrounding exercise context; here, we adopt a simplified setting and classify each image in isolation, based solely on the type of text it contains.

The decision is also target-language dependent. For English, German, and Ukrainian, numbers, mathematical symbols, and variables typically do not require translation. However, because Ukrainian uses the Cyrillic script, measurement units written in Latin script should be translated.

We constructed a dataset for this experiment from our image store associated with the web exercises, which contains over 31k images. We randomly sampled 1,000 images and manually annotated (i) whether the image contains text and (ii) whether the text should be translated. Following the simplified setting, annotation was performed without access to the surrounding exercise context. We therefore labeled text as "requires translation" if it would be unintelligible to a target-language reader without translation, which is a superset of the text that is strictly necessary to solve the exercise. In our sample, 292 images contain text, and 68 contain text that requires translation.

We experimented with three quantized, relatively small open-source multimodal models: `gemma3-12b-qat`,[9] `qwen2.5-omni-7b-gptq`,[10] and `llama4-scout-17b-16e-iq4`.[11] The first two use dense feed-forward layers, whereas the latter uses a mixture-of-experts architecture, with 109B total parameters but 17B active parameters. We use the prompt shown in Figure 2 in a zero-shot manner.

Results in Table 5 show that the F1 score for detecting images with text that requires translation falls in a narrow range of 0.62–0.64 across all models. However, because this is a filtering step, recall is more important than precision. Therefore, among systems with similar F1 scores, we prefer the one with the highest recall, which is `gemma3-12b-qat` in our case. Precision remains reasonably high; that is, the detector filters out most images that either contain no text or contain embedded text that does not require translation.

System prompt:

```
You are a classification assistant. Analyze the provided image for text and determine if translation is required based on these rules:
1. No text: The image contains no readable characters.
2. No text to translate: Text is present, but it is NOT in Czech, or it consists solely of numbers, mathematical symbols, or variables.
3. Text to translate: Text is present in Czech, including units of measurement.
Constraint: You must respond with exactly one of these three labels: ["No text", "No text to translate", "Text to translate"]. Do not provide explanations or preamble.
```

User prompt:

```
Analyze this image.
<IMAGE>
```

Figure 2: Prompt template used for identifying images whose embedded text requires translation.

Model	P	R	F1
`gemma3-12b-qat`	0.52	**0.84**	**0.64**
`qwen2.5-omni-7b-gptq`	0.50	0.79	0.62
`llama4-scout-17b-16e-iq4`	0.56	0.74	0.63

Table 5: The quality of identification images containing a text that requires translation (support = 68).

[9] https://huggingface.co/google/gemma-3-12b-it-qat-q4_0-gguf

[10] https://huggingface.co/Qwen/Qwen2.5-Omni-7B-GPTQ-Int4

[11] https://huggingface.co/unsloth/Llama-4-Scout-17B-16E-Instruct-GGUF