Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

ADAPTING AI CHATBOTS TO THE UZBEK LANGUAGE FOR EDUCATIONAL PURPOSES

Khakimova Gullola Nasirovna
Master Student, Sharda University Uzbekistan
+998 99 8555715 hakimovagullola491@gmail.com

Abstract
Large language model chatbots are entering school education, yet most of them were built around English and work noticeably worse in Uzbek. This article examines what has to be changed in such a chatbot before it can serve Uzbek-speaking pupils reliably. The linguistic and technical obstacles are analyzed first: agglutinative morphology, the parallel use of Latin and Cyrillic scripts, inconsistent apostrophe characters in the letters o‘ and g‘, code-switching with Russian and the shortage of educational data. An original experiment then measures how four widely used tokenizers process 30 parallel school-domain sentences in English and Uzbek. Uzbek in Latin script required 1.63–2.04 times more tokens than English, Cyrillic Uzbek up to 5.24 times more, and the choice of apostrophe character alone changed token counts by up to 5.7%. On this basis a five-layer adaptation framework is proposed, covering text normalization, safety, curriculum-based retrieval, an Uzbek-adapted language model and a pedagogical response policy, together with criteria for its evaluation.

Keywords: Uzbek language, educational chatbot, large language models, low-resource languages, tokenization, transliteration, retrieval-augmented generation, natural language processing, school education, artificial intelligence.

Introduction

АДАПТАЦИЯ ЧАТ-БОТОВ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА К УЗБЕКСКОМУ ЯЗЫКУ В ОБРАЗОВАТЕЛЬНЫХ ЦЕЛЯХ

Хакимова Гуллола Насировна
Магистрант, Университет Шарда в Узбекистане
+998 99 8555715 hakimovagullola491@gmail.com

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

Аннотация

Чат-боты на основе больших языковых моделей приходят в школьное образование, однако большинство из них создано с ориентацией на английский язык и заметно хуже работает на узбекском. В статье рассматривается, что необходимо изменить в таком чат-боте, чтобы он надёжно обслуживал узбекоязычных учащихся. Сначала анализируются лингвистические и технические препятствия: агглютинативная морфология, параллельное использование латиницы и кириллицы, несогласованность символов апострофа в буквах o‘ и g‘, смешение с русским языком и нехватка образовательных данных. Затем в оригинальном эксперименте измеряется, как четыре широко используемых токенизатора обрабатывают 30 параллельных предложений школьной тематики на английском и узбекском языках. Узбекский текст на латинице потребовал в 1,63–2,04 раза больше токенов, чем английский, на кириллице – до 5,24 раза больше, а один лишь выбор символа апострофа менял число токенов до 5,7%. На этой основе предложена пятиуровневая модель адаптации, включающая нормализацию текста, безопасность, поиск по учебной программе, адаптированную к узбекскому языку модель и педагогическую политику ответов, а также критерии её оценки.

Ключевые слова. Узбекский язык, образовательный чат-бот, большие языковые модели, малоресурсные языки, токенизация, транслитерация, генерация с дополненным поиском, обработка естественного языка, школьное образование, искусственный интеллект.

Conversational systems built on large language models (LLMs) have moved from research laboratories into everyday school life within a very short time. Pupils use them to check homework, to ask for an explanation of a difficult topic or to practise a foreign language, and teachers use them to prepare tasks and lesson materials. Reviews of the educational use of such systems point both to their potential for individual support and to risks such as uncritical reliance on generated answers and factual errors. For a school system the attractive feature of a chatbot is simple: it is available at any time and can repeat an explanation as many times as a pupil needs.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

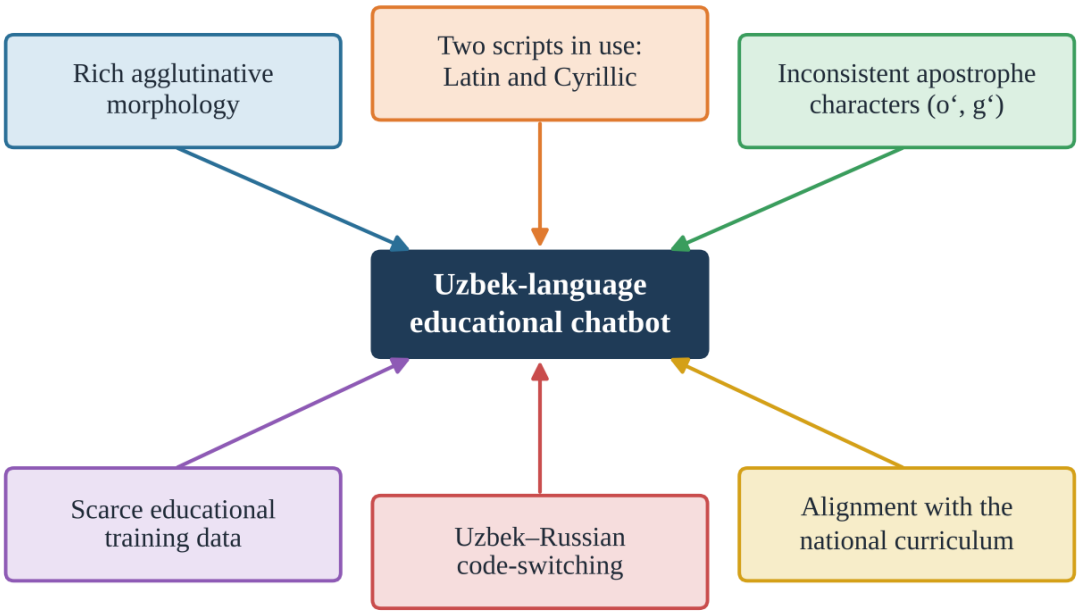
The situation is less favourable for pupils who study in Uzbek. Uzbek is the language of instruction for the majority of schoolchildren in Uzbekistan, and the global Uzbek population exceeds 34 million people, yet from the point of view of language technology it remains a low-resource language. Research on the Central Asian Turkic languages repeatedly reports the same obstacles: limited data, few annotated resources and slow technology development. General-purpose chatbots can usually answer in Uzbek, but their answers often contain unnatural word forms, Russian or English insertions and terminology that does not match the textbooks used in Uzbek schools.

The quality gap begins before the model generates a single word. Modern language models read text as a sequence of tokens produced by a tokenizer, and tokenizers trained mainly on English split texts in other languages into many more pieces. Petrov et al. showed that the same content translated into different languages can differ in tokenized length by up to 15 times. Longer token sequences mean higher cost for paid services, slower answers and less room in the context window for textbook material, which is exactly what an educational chatbot needs most.

The aim of this article is to determine which components of an AI chatbot must be adapted to make it suitable for school education in Uzbek and to propose a practical framework for such adaptation. The work makes three contributions. First, it systematizes the linguistic and technical features of Uzbek that affect chatbot quality. Second, it presents an original measurement of how four widely used tokenizers process parallel school-domain texts in English, Latin-script Uzbek and Cyrillic-script Uzbek, including the effect of different apostrophe characters. Third, it proposes a five-layer adaptation framework, a pedagogical dialogue policy and a set of evaluation criteria for Uzbek educational chatbots.

Features of the Uzbek language that affect chatbots. The difficulties that an Uzbek educational chatbot faces can be grouped into six interrelated problems (Picture 1). Some of them are linguistic and stem from the structure of the language; others are social or technical and stem from the way Uzbek is written and used today.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.



Picture 1. Main challenges in adapting an AI chatbot to Uzbek for school education.

Agglutinative morphology. Uzbek belongs to the Turkic family, whose members are generally described as morphologically rich. Grammatical meaning is expressed by chains of suffixes attached to a stem. The word o‘quvchilarimizga (“to our pupils”) consists of the stem o‘quv (study), the suffix -chi (agent), the plural -lar, the possessive -imiz (our) and the dative -ga. A single noun can therefore appear in dozens of surface forms, and a model that has seen too few of them tends either to produce non-existent forms or to choose the wrong case ending. For a chatbot this matters in two places: in understanding a pupil's question and in searching educational materials, where a query word and a textbook word may share a stem but differ in every suffix.

Two scripts. Uzbek replaced Cyrillic with a Latin-based alphabet in 1993, and the alphabet was reformed in 1995; nevertheless both scripts remain popular in law, books, the web and the media [4]. A pupil may therefore type a question in Latin, copy a passage from an older Cyrillic book, or mix the two. For a language model the Latin and Cyrillic versions of the same sentence are two different strings with different token sequences, so the knowledge learned from one script is only partly available in the other.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

Apostrophe characters. The Latin alphabet contains the letters o‘ and g‘ and the separating mark ’. The official forms use special modifier characters, but on ordinary keyboards people type the ASCII apostrophe ('), a typographic quotation mark (‘) or even a grave accent (`). For a human reader all of these look almost the same; for a tokenizer they are different symbols. The developers of one Uzbek-adapted model explicitly advise users to replace apostrophes with a special placeholder sequence before passing text to the model [6], which shows that the problem is practical rather than theoretical.

Code-switching and dialects. In everyday speech and writing, especially in cities, Uzbek is often mixed with Russian words and phrases, and school terminology in physics, chemistry and informatics still contains many Russian and international borrowings. In addition, Southern Uzbek, spoken by around 5 million people in Afghanistan, differs significantly from the northern variety in phonology, lexicon and orthography. A chatbot must understand such mixed input but answer in standard literary Uzbek.

Data scarcity. Open Uzbek resources are still limited in size and genre. Many corpora are built from news websites, while educational dialogues, worked solutions of school problems and teacher explanations are almost absent. A model trained mainly on news learns the vocabulary of politics and economics better than the vocabulary of a seventh-grade geometry lesson.

Curriculum alignment. Finally, an educational chatbot must follow the national curriculum: the same concept should be called by the same term as in the pupil's textbook, examples should match the grade level, and the order of topics should not run ahead of the programme. This requirement cannot be solved by language modelling alone; it needs access to the actual educational content.

Existing resources for Uzbek language technology. Over the last five years the Uzbek research community has produced a number of open resources that can serve as building blocks for an educational chatbot. The most relevant of them are summarized in Table 1.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

Table 1. Selected open resources for Uzbek language technology

Resource	Type	Key characteristics
UzBERT [1]	Encoder model	First public monolingual BERT for Uzbek; trained on a news corpus of about 142 million words in Cyrillic script
BERTbek [2]	Encoder models	BERT models for Uzbek; outperform multilingual BERT in sentiment analysis, topic classification and named entity recognition
BBPOS [7]	Tagging model and dataset	Part-of-speech tagging with Uzbek BERT models, about 91% average accuracy; first public UPOS benchmark for Uzbek
USC [3]	Speech corpus	105 hours of transcribed speech from 958 speakers; 17.4% word error rate on the test set in initial experiments
UzTransliterator [4]	Software tool	Transliteration between Cyrillic, Latin and the newly announced Latin alphabet; rule-based and fine-tuned methods
Llama-3.1-8B-Instruct-Uz [5]	Generative LLM	Llama 3.1 instruction-tuned on Uzbek and English data; English→Uzbek BLEU raised from 6.72 to 11.58 on FLORES+
alloma-8B-Instruct [6]	Generative LLM	Llama-based model with a custom Uzbek tokenizer: about 1.7 tokens per Uzbek word versus about 3.5 for the original Llama

Two trends can be seen in the table. The first generation of resources consisted of encoder models such as UzBERT and BERTbek, which are well suited for classification and search but cannot conduct a dialogue. The second, more recent generation consists of generative models obtained by adapting open multilingual LLMs to Uzbek. The Llama-3.1-8B-Instruct-Uz model, for example, nearly doubled the BLEU score of its base model for English-to-Uzbek translation and raised the accuracy of Uzbek sentiment analysis from 68.96% to 82.42% [5]. The alloma models go further and replace part of the tokenizer vocabulary, which roughly halves the number of tokens per Uzbek word [6].

At the same time, none of these resources was designed specifically for school education. They are evaluated on translation, sentiment and news classification, not on whether a pupil receives a correct, curriculum-consistent and pedagogically sound explanation. This gap motivates the present work.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

Materials and methods. To quantify one of the problems described in Section 2 — the cost of Uzbek text for existing tokenizers — a small controlled experiment was carried out.

Corpus. A parallel corpus of 30 sentences typical of school communication was written for this study. The sentences cover mathematics, physics, biology, geography, history, literature and native language, and include teacher instructions, factual statements and typical pupil questions (for example, “I did not understand this topic, could you give me one more example?”). Each sentence was written in English and in literary Uzbek in the Latin script by the author. The Cyrillic version was produced automatically with the open UzTransliterator tool [4] and checked manually. Examples are given in Table 2. In total the English part contains 287 words and the Uzbek Latin part 208 words; the Uzbek version is 27.5% shorter in words because suffixes replace separate English prepositions, articles and pronouns.

Table 2. Examples of sentences from the parallel corpus

English	Uzbek (Latin)	Uzbek (Cyrillic)
The teacher explained the new topic to the students.	O'qituvchi o'quvchilarga yangi mavzuni tushuntirdi.	Ўқитувчи ўқувчиларга янги мавзуни тушунтирди.
Please explain step by step how to solve this equation.	Iltimos, bu tenglamani qanday yechishni bosqichma-bosqich tushuntirib bering.	Илтимос, бу тенгламани қандай ечишни босқичма-босқич тушунтириб беринг.
Cells are the smallest structural units of living organisms.	Hujayralar tirik organizmlarning eng kichik tuzilish birliklaridir.	Ҳужайралар тирик организмларнинг энг кичик тузилиш бирликларидир.

Apostrophe variants. Four versions of the Latin text were prepared, identical except for the character used in o‘, g‘ and the separating mark: the ASCII apostrophe (U+0027), the modifier letter turned comma used in the official orthography (U+02BB), the left single quotation mark (U+2018) and the grave accent (U+0060). These are the variants most often found in real Uzbek texts typed on different devices.

Tokenizers. Four byte-level BPE tokenizers were compared: GPT-2 (an early English-centred tokenizer used as a reference point), cl100k_base (used by GPT-4 and GPT-3.5), o200k_base (used by GPT-4o) and the Llama 3 tokenizer (the

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

base of many open models, including the Uzbek models in Table 1). The OpenAI tokenizers were run with the open-source gpt-tokenizer library, version 4.1.2, and the Llama 3 tokenizer with llama3-tokenizer-js, version 1.2.0; special tokens were not added.

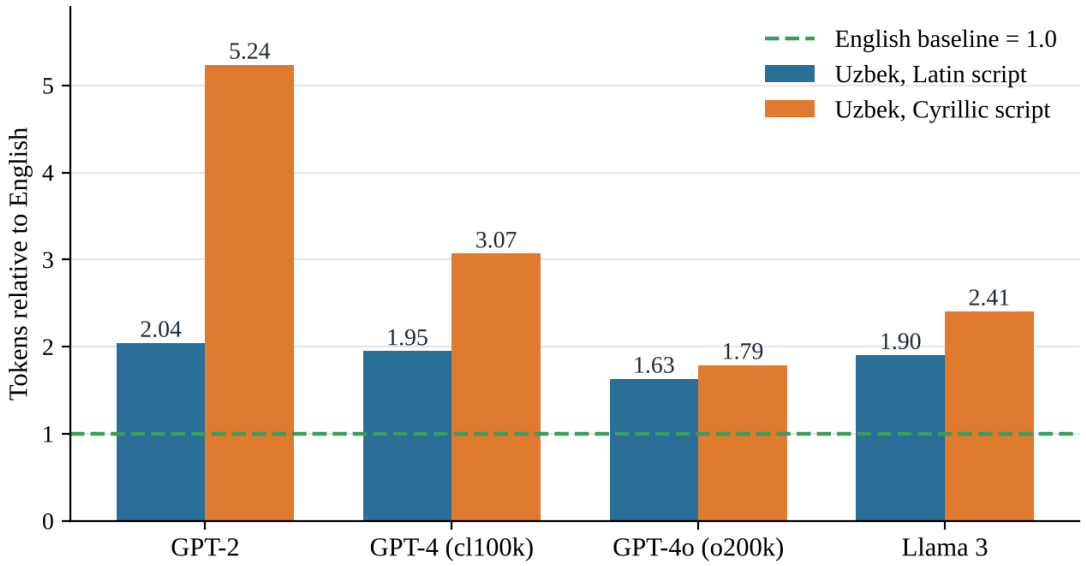
Measures. Two standard measures were calculated. Fertility is the average number of tokens per word; a value close to 1 means that most words are stored in the vocabulary as whole units. The tokenization premium is the ratio between the number of tokens of the Uzbek text and of its English equivalent, following the notion of tokenizer parity introduced by Petrov et al.; a premium of 2.0 means that the same content costs twice as many tokens in Uzbek. All scripts and the corpus are kept by the author and can be shared for reproduction.

Results. Tokenization premium. The results are summarized in Table 3 and Picture 2. All four tokenizers process English almost identically, at about 1.2 tokens per word. For Uzbek in Latin script the cost is substantially higher: from 1.63 times the English length for o200k_base to 2.04 times for GPT-2. The Llama 3 tokenizer needed 3.22 tokens per Uzbek word, which agrees well with the figure of about 3.5 reported by the developers of the alloma models [6].

Table 3. Tokenization of 30 parallel sentences (author's measurements)

Tokenizer	English, tokens	Uzbek Latin, tokens (premium)	Uzbek Cyrillic, tokens (premium)	Tokens per word: EN / Lat / Cyr
GPT-2	351	717 (2.04)	1838 (5.24)	1.22 / 3.45 / 8.79
GPT-4 (cl100k)	352	686 (1.95)	1080 (3.07)	1.23 / 3.30 / 5.17
GPT-4o (o200k)	347	566 (1.63)	621 (1.79)	1.21 / 2.72 / 2.97
Llama 3	352	670 (1.90)	847 (2.41)	1.23 / 3.22 / 4.05

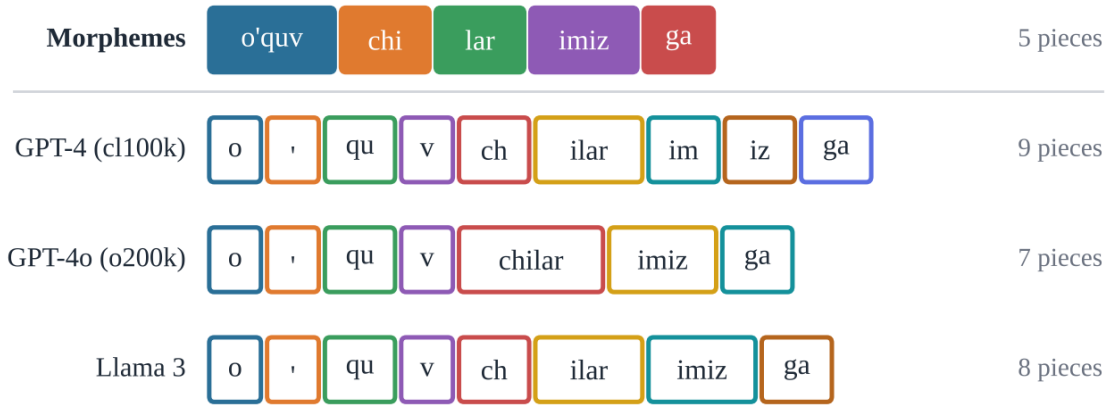
Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.



Picture 2. Tokenization premium of Uzbek relative to English for four tokenizers (author's measurements).

The script has an even larger effect than the language. The Cyrillic version of the same Uzbek text was 5.24 times longer than English for GPT-2 and 3.07 times longer for cl100k_base. The reason is that Cyrillic letters occupy two bytes in UTF-8 and, unless the tokenizer has learned frequent Cyrillic sequences, they are encoded byte by byte. The newest tokenizer, o200k_base, which has a much larger vocabulary, almost closes the gap between the two scripts (1.63 versus 1.79), but even for it Uzbek remains considerably more expensive than English.
Segmentation quality. Token counts do not tell the whole story; it also matters where the boundaries fall. Picture 3 compares the morphological structure of the word o'quvchilarimizga with the pieces produced by three tokenizers.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.



Picture 3. Morphemes of the word o'quvchilarimizga ("to our pupils") and its segmentation by three tokenizers (author's analysis).

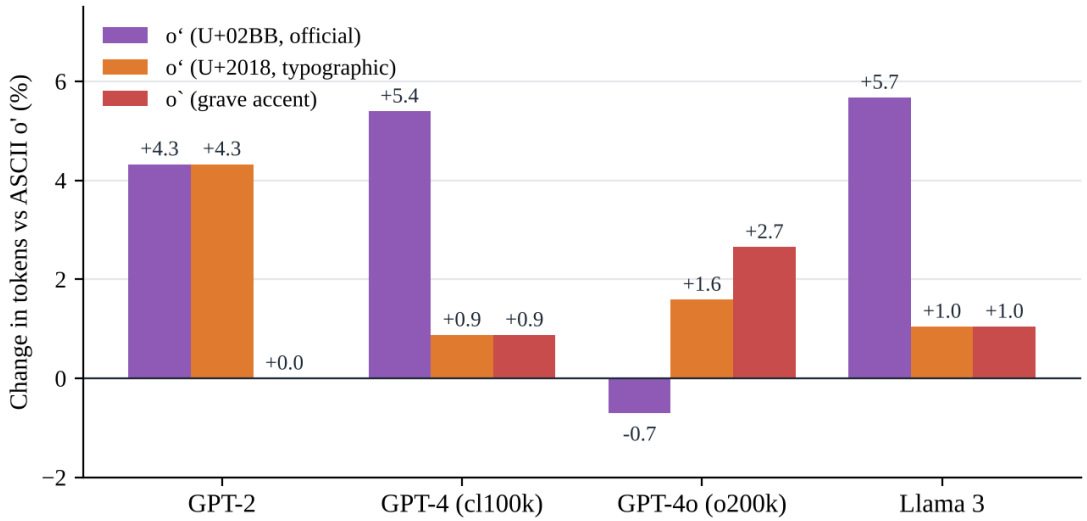
None of the tokenizers respects the morpheme boundaries completely. All of them separate the apostrophe from the letter o and break the stem o'quv into three meaningless fragments. The cl100k_base tokenizer also splits the plural and possessive suffixes into parts. The o200k_base and Llama 3 tokenizers keep the possessive -imiz intact, which is a sign that they have seen more Uzbek text, but the stem is still fragmented. When the stem of a word is cut differently in different forms, the model has to learn the connection between related forms from data, and in a low-resource setting such data are scarce.

Effect of the apostrophe character. Table 4 and Picture 4 show how the number of tokens changes when only the apostrophe character is replaced.

Table 4. Number of tokens for the Uzbek Latin text with different apostrophe characters (author's measurements)

Tokenizer	ASCII o'	o‘ (U+02BB)	o‘ (U+2018)	o` (grave)
GPT-2	717	748	748	717
GPT-4 (cl100k)	686	723	692	692
GPT-4o (o200k)	566	562	575	581
Llama 3	670	708	677	677

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.



Picture 4. Change in the number of tokens caused by the apostrophe character, relative to the ASCII apostrophe (author's measurements).

The officially recommended character U+02BB increased the token count by 5.4% for cl100k_base and by 5.7% for Llama 3, because these tokenizers do not have it in their vocabulary and encode it as two separate bytes. For o200k_base, on the contrary, the official form was the cheapest variant, since this tokenizer has learned combinations such as “o‘” as single tokens. In other words, there is no apostrophe variant that is best for all models, and the same word written by two pupils on two different phones may be represented to the model in two different ways.

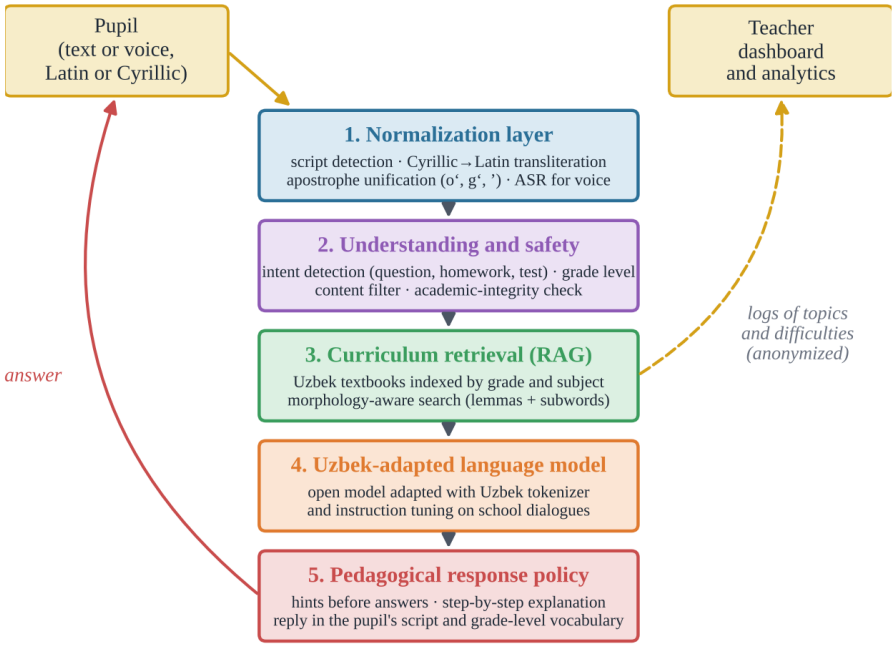
Taken together, the results lead to three practical conclusions. First, an educational chatbot that relies on a general-purpose model pays roughly twice as much in tokens for Uzbek content as for English, which shortens the amount of textbook material that can be placed into the context. Second, input in Cyrillic should be converted to Latin before it reaches the model if the model was trained mostly on Latin-script Uzbek. Third, apostrophe characters must be unified to the variant that a particular model processes best. These conclusions are built into the framework proposed below.

Proposed framework for adapting a chatbot to Uzbek. Based on the analysis above, a five-layer framework for an Uzbek educational chatbot is proposed (Picture 5). Each layer addresses one or more of the challenges shown in Picture

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

1, and the layers can be implemented step by step, starting from those that require the least data and computing resources.

Layer 1: normalization. Every message first passes through a normalization module. It detects the script of the text and, if necessary, transliterates Cyrillic into Latin with a tool such as UzTransliterator [4]; unifies all apostrophe-like characters into the single form that the chosen model handles best, as measured in Section 5; and, for voice input, converts speech to text using a recognizer trained on Uzbek speech, for example on the USC corpus [3]. The script of the original message is remembered so that the final answer can be returned in the same script. This layer requires no training data and gives an immediate improvement, which makes it the natural first step.



Picture 5. Five-layer framework of an Uzbek-language educational chatbot (proposed by the author).

Layer 2: understanding and safety. The second layer classifies the request: is it a question about a topic, a request to check homework, a request to solve a test, or something unrelated to learning? It also estimates the pupil's grade and filters unsafe content. Encoder models such as BERTbek, which already perform well on Uzbek classification tasks [2], are well suited for this layer because they are small and fast. An academic-integrity rule is applied here: if the request looks like

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

a current test or examination question, the system switches to hint mode instead of giving a ready answer.

Layer 3: curriculum retrieval. The third layer implements retrieval-augmented generation, in which a language model receives relevant documents found by a search component together with the question. Here the documents are Uzbek school textbooks and teacher materials, divided into short passages and indexed by subject and grade. Because of agglutination, the search must not rely only on exact word matching: the index should store lemmas obtained with a morphological analyzer as well as subword representations, so that a question containing tenglamalarni ("the equations", accusative) finds a passage that uses tenglama ("equation"). Retrieval keeps terminology consistent with the textbook and reduces the risk of factual errors.

Layer 4: Uzbek-adapted language model. The core generator should be an open model already adapted to Uzbek, such as the models in Table 1, which can then be further instruction-tuned on a relatively small set of school dialogues. Using an open model allows the school or the ministry to run it on its own servers and to keep pupils' data inside the country. The choice of tokenizer is important: a model with an Uzbek-optimized vocabulary processes the same textbook passage in about half as many tokens [6], which directly increases the amount of curriculum context that fits into one request.

Layer 5: pedagogical response policy. The last layer controls how the answer is given rather than what it contains. It is described in detail in Section 7. In addition, anonymized records of which topics pupils ask about and where they make mistakes are passed to a teacher dashboard, so that the teacher can see which parts of the programme need more attention in class.

Table 5 compares the main technical strategies that can be used to adapt the language model itself. They differ greatly in the amount of data and computing power they require, which is a decisive factor for educational institutions with limited budgets.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

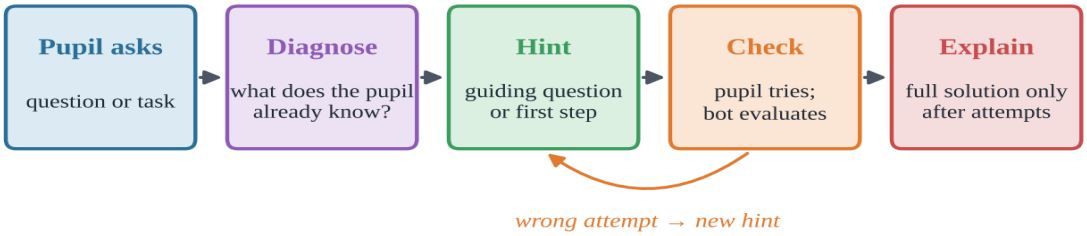
Table 5. Strategies for adapting a language model to Uzbek (compiled by the author)

Strategy	Data needed	Cost	Expected effect
Prompting and text normalization	None	Very low	Fewer errors caused by script and apostrophe variation; no change in the model's knowledge of Uzbek
Retrieval-augmented generation	Textbooks and teacher materials	Low	Correct terminology and facts from the curriculum; quality still limited by the model's Uzbek fluency
Instruction fine-tuning	Thousands of Uzbek school dialogues	Medium	More natural Uzbek and the desired teaching style
Tokenizer adaptation and continual pretraining	Billions of tokens of Uzbek text	High	Shorter token sequences and better morphology; requires GPU resources
Training a model from scratch	Very large Uzbek corpus	Very high	Full control, but unrealistic for a single institution

For most schools and universities in Uzbekistan the realistic path is to combine the first three strategies on top of an existing open Uzbek-adapted model, leaving tokenizer adaptation and large-scale pretraining to national research centres and industry.

Pedagogical design of the dialogue. A chatbot that simply gives the correct answer to every question is convenient for a pupil but harmful for learning, because it removes the effort through which knowledge is formed. For this reason the response policy follows the logic of guided questioning (Picture 6). The bot first finds out what the pupil already knows, then offers a hint or the first step, lets the pupil try, evaluates the attempt and only after several attempts presents the complete solution.

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.



Picture 6. Pedagogical dialogue policy of the chatbot (proposed by the author).

An example of a dialogue designed according to this policy is shown in Table 6. It is an illustrative scenario written by the author to show the intended behaviour, not output recorded from a working system.

Table 6. Illustrative dialogue for a seventh-grade algebra task

Step	Uzbek	English translation
Pupil	3x + 5 = 20 tenglamani yechib bering.	Solve the equation 3x + 5 = 20 for me.
Diagnose	Keling, birga yechamiz. Chap tomondagi 5 sonidan qanday qutulish mumkin, deb o'ylaysiz?	Let's solve it together. How do you think we can get rid of the 5 on the left side?
Pupil	5 ni qo'shamizmi?	Do we add 5?
Hint	Yo'nalish to'g'ri, lekin 5 ni yo'qotish uchun uni ikkala tomondan ayirish kerak. Qanday tenglama hosil bo'ladi?	The direction is right, but to remove the 5 we must subtract it from both sides. What equation do we get?
Pupil	3x = 15	3x = 15
Check	Ajoyib! Endi x ni topish uchun nima qilamiz?	Excellent! Now what do we do to find x?
Pupil	15 ni 3 ga bo'lamiz, x = 5.	We divide 15 by 3, x = 5.
Explain	To'g'ri! Tekshiramiz: 3 · 5 + 5 = 20. Javob: x = 5.	Correct! Let's check: 3 · 5 + 5 = 20. Answer: x = 5.

Several language-specific details are built into this policy. The bot addresses the pupil with the polite plural form (siz) or the inclusive first person plural (keling, yechamiz – “let's solve”), which is natural for teacher speech in Uzbek. Mathematical terminology follows the Uzbek textbook (tenglama for “equation”,

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

ayirish for “subtraction”) rather than Russian borrowings. Explanations are kept short, because long answers in Uzbek also mean long token sequences.

Evaluation framework. Standard benchmarks for Uzbek models measure translation quality, sentiment or news classification [5, 6]. These measures are necessary but not sufficient for an educational chatbot. Table 7 proposes evaluation criteria that reflect the actual purpose of the system.

Table 7. Criteria for evaluating an Uzbek educational chatbot (proposed by the author)

Criterion	What is assessed	How it can be measured
Linguistic correctness	Grammar, word forms, absence of Russian or English insertions	Rating by Uzbek language teachers; share of answers without errors
Script consistency	Answer in the same script as the question; correct o‘ and g‘	Automatic check of characters
Curriculum fidelity	Terms and facts match the textbook for the given grade	Comparison with a set of reference questions and answers per subject
Pedagogical quality	Hints before answers; explanations suited to the grade	Rubric-based assessment by subject teachers
Safety and integrity	Refusal of harmful content; no ready answers to test questions	Set of adversarial test prompts
Efficiency	Tokens per answer, response time, cost per pupil	Automatic logging
Learning effect	Improvement in pupils' results	Pre-test and post-test with control and experimental groups

The last criterion is the most important and the most expensive: it requires a pedagogical experiment in which one group of pupils studies a topic with the chatbot and a control group studies without it, with knowledge measured before and after. Such an experiment is planned as the next stage of the author's research.

Discussion. The results show that the weakness of general-purpose chatbots in Uzbek is not only a question of how much Uzbek text the model has seen. Part of the problem lies in the way text is converted into tokens, and part in the variation of writing practice: two scripts and several apostrophe characters. These two

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

sources of difficulty are cheap to address. Normalization and transliteration require neither training data nor graphics processors, yet they remove variation that otherwise forces the model to treat the same word as several different words. The findings are consistent with international research on tokenization unfairness and with the experience of the Uzbek developers who built a custom tokenizer for their models [6]. The present study adds two observations specific to Uzbek: the very large penalty for Cyrillic text in older tokenizers and the measurable effect of the apostrophe character, which changes direction from one tokenizer to another. The latter means that normalization rules cannot be fixed once and for all; they should be chosen together with the model.

The study has limitations. The corpus is small (30 sentence pairs) and written by one author, so the exact numbers may change on a larger and more varied set of texts; however, the direction of the effects is clear and agrees with published results. Only tokenization was measured experimentally; the proposed framework and the dialogue policy have not yet been tested with pupils. Finally, the field changes quickly, and newer models may handle Uzbek better than those examined here. These limitations define the directions of future work: building a larger open corpus of Uzbek school dialogues, implementing a prototype of the framework and testing it in real classrooms.

Conclusion. Adapting an AI chatbot to Uzbek for school education requires more than translating its interface. The article has shown that the main obstacles are agglutinative morphology, the coexistence of Latin and Cyrillic scripts, inconsistent apostrophe characters, code-switching, the shortage of educational data and the need to follow the national curriculum.

The experiment with four tokenizers demonstrated that the same school content costs 1.63–2.04 times more tokens in Latin-script Uzbek than in English and up to 5.24 times more in Cyrillic, and that the choice of apostrophe character alone changes the token count by up to 5.7%. These findings justify placing a normalization layer at the entrance of any Uzbek chatbot.

The proposed five-layer framework — normalization, understanding and safety, curriculum retrieval, an Uzbek-adapted language model and a pedagogical response policy — combines existing open Uzbek resources into a system that can be built step by step even with limited resources. Together with the proposed evaluation criteria, it can serve as a basis for developing chatbots that help Uzbek-

Educa Journal of Studies
ISSN (E): 3054-3355
Volume 01, Issue 11, November, 2026
Website: https://researchiapress.com
Licensed under a Creative Commons Attribution 4.0 International License.

speaking pupils learn in their native language rather than simply supplying them with answers.

REFERENCES

1. Mansurov B., Mansurov A. UzBERT: pretraining a BERT model for Uzbek. – arXiv:2108.09814, 2021.
2. Kuriyozov E., Vilares D., Gómez-Rodríguez C. BERTbek: A Pretrained Language Model for Uzbek // Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL) @ LREC-COLING 2024. – Torino: ELRA and ICCL, 2024. – P. 33–44.
1. Musaev M., Mussakhojayeva S., Khujayorov I., Khassanov Y., Ochilov M., Varol H.A. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments // Speech and Computer: SPECOM 2021. – Springer, 2021. – P. 437–447.
2. Salaev U., Kuriyozov E., Gómez-Rodríguez C. A machine transliteration tool between Uzbek alphabets // CEUR Workshop Proceedings. – 2022. – Vol. 3315. – arXiv:2205.09578.
3. Fozilov E., Urinov A., Juraev K. Llama-3.1-8B-Instruct-Uz: model card. – Hugging Face, 2024. – URL: https://huggingface.co/behbudiy/Llama-3.1-8B-Instuct-Uz
4. alloma-8B-Instruct: Uzbek-optimized Llama model, model card. – Hugging Face. – URL: https://huggingface.co/uzlm/alloma-8B-Instruct
5. Bobojonova L., Akhundjanova A., Ostheimer P., Fellenz S. BBPOS: BERT-based Part-of-Speech Tagging for Uzbek. – arXiv:2501.10107, 2025.