Evaluating Large Language Models for Text-to-Gloss Translation in Kazakh-Russian Sign Language: A Pilot Study

Zhanibek Kozhirbayev, Alfarabi Imashev
National Laboratory Astana, Nazarbayev University
{zhanibek.kozhirbayev, alfarabi.imashev}@nu.edu.kz

Abstract

Conceptual glossing involves a systematic linguistic transformation in which the models must preserve meaning, grammatical integrity, and punctuation while turning the real language into a more structured structure. The purpose of this study is to assess the accuracy and dependability of glosses produced by these models by juxtaposing them with human-annotated standards, investigating whether the models maintain essential linguistic characteristics. By identifying the strengths and weaknesses of each model, we want to determine which architectures are most suitable for organized language tasks, such as glossing. This may reduce the manual labor required for linguistic annotation by experts while maintaining superior quality outcomes. And help deaf signers with weak reading skills interpret written paragraphs into glosses, making them more comprehensible and naturally looking to them. Text-to-gloss translation converts written or spoken language into sign language glosses, enhancing accessibility for the Deaf and Hard of Hearing (DHH) community. This pilot study evaluates four large language models (LLMs): GPT-4-turbo, Grok 3, Deepseek-V3, and Gemini 20 Flash to generate conceptual glosses in Kazakh-Russian Sign Language (K-RSL), still an under-resourced sign language. Using a dataset of 250 Russian sentences with expert-annotated K-RSL glosses, we assess performance across METEOR, BLEU, BERTScore, and WER. Results show Deepseek-V3 excels on complex texts (METEOR: 0.426 for K-RSL word order, 0.377 for fairytale paragraphs), while Gemini 20 Flash performs strongly on short sentences (METEOR: 0.602). These findings demonstrate LLMs' potential to automate gloss production, reducing manual annotation and aiding DHH individuals with reading comprehension. Challenges include K-RSL's unique grammar and limited datasets. This is the first study to apply LLMs to K-RSL glossing and examine the potential efficacy of autonomous gloss production.

Keywords: LLMs, sign language generation, text to gloss

1. Introduction

Text-to-gloss (Text2Gloss) translation converts written or spoken language into sign language glosses, textual representations of signs, to improve accessibility for the Deaf and Hard of Hearing community. It has been proven that deaf students experience a higher degree of limitations and, unfortunately, in the end, they may usually have a lower knowledge or insufficient reading and writing skills (Morere, 2012; Sarchet et al., 2014; Hrastinski and Wilbur, 2016; Takahashi et al., 2017; Alqraini, 2018; González-Cuenca et al., 2024). According to the observations of the interpreters with whom we cooperate, even deaf signers with higher reading and writing skills may usually experience cognitive burdens when reading long sentences in Russian or Kazakh and have to divide concepts into glosses in mind, and prefer to express their communication using concepts sequences which are more similar to gloss annotations, usually paying less attention to proper word order of Kazakh or Russian sentences. In Kazakhstan, deaf signers may generally face difficulties reading Russian or Kazakh and require additional elaborations due to fewer educational opportunities (Imashev et al., 2024a), thus may prefer gloss-like sequences to grasp the idea. Although transformer-based models have advanced Text2Gloss for languages like American Sign Language (Stoll et al., 2018), K-RSL remains understudied and large language models (LLMs) have not been applied to this task.

This pilot study evaluates four LLMs: GPT-4-turbo (Shafik, 2024), Grok 3 (Carvalho et al., 2025), Deepseek-V3 (Lu et al., 2024), and Gemini 20 Flash (Imran and Almusharraf, 2024) for generating K-RSL glosses from Russian sentences. We address the gap in applying LLMs to low-resource sign languages, focusing on K-RSL's unique features. Our contributions are:

- The first evaluation of LLMs for Text2Gloss in K-RSL, using 250 expert-annotated sentences.
- Comprehensive analysis across semantic and structural metrics (e.g., METEOR, BLEU).
- Insights into model suitability for short and long texts, aiding DHH accessibility.
- Publicly available dataset and code for reproducibility.

The paper is organized as follows: Section 2 reviews related work, Section 3 describes our methodology, Section 4 presents results, Section 5 discusses findings, and Section 6 concludes.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9964–9972
11–16 May 2026.

2. Related work

Machine translation (MT), the automated conversion of text from a source language to a target natural language, has undergone significant advancements in recent decades. Neural Machine Translation (NMT) has emerged as a state-of-the-art technology to address the shortcomings of previous translation methodologies. In contrast to such methodologies, the NMT approach aims to delineate and train a neural network capable of accommodating broader textual context windows in a flexible manner (Bahdanau et al., 2014).

Unfortunately, Sign Language Machine Translation (SLMT) cannot directly employ machine translation methodologies designed for written language translation. Early systems relied on Statistical Machine Translation (SMT) to generate glosses for sign languages like Spanish Sign Language (San-Segundo et al., 2012), often using glosses as intermediaries for 3D avatar rendering. These approaches struggled with limited datasets and complex sign language grammar. Recent advances leverage Neural Machine Translation, with Recurrent Neural Networks (RNNs) (Stoll et al., 2020) and transformer architectures (Saunders et al., 2020; Zhu et al., 2023) achieving higher accuracy. For instance, Saunders et al. (2022) reported state-of-the-art BLEU scores for Deutsche Gebärdensprache (DGS) using transformers on the RWTH-PHOENIX14T dataset (De Martino and Christinele, 2024). Enhancements like syntax-aware transformers, which incorporate dependency parsing into embeddings (Gómez et al., 2021), and transfer learning with pre-trained models like mBART (Egea Gómez et al., 2022) have further improved glossing quality for high-resource sign languages such as American Sign Language (ASL).

Despite these advances, most Text2Gloss research focuses on well-resourced sign languages, leaving low-resource languages like Kazakh-Russian Sign Language (K-RSL) underexplored. K-RSL, used by deaf communities in Kazakhstan, features unique grammatical structures, the usual omission of prepositions, and less attention to proper written Kazakh or Russian word order grammar rules, which may challenge standard translation models (Imashev et al., 2024b). Existing K-RSL studies usually provide annotated corpora but lack automated glossing solutions, relying on manual expert annotation.

Large Language Models (LLMs) have transformed NLP tasks, excelling in text generation and multilingual processing. Models like GPT-4, Deepseek-V3, and Gemini 20 Flash leverage vast pre-training datasets to handle complex linguistic transformations, yet their application to Text2Gloss remains limited. Recent work (De Martino and Christinele, 2024) explored LLMs for glossing in DGS, but no studies have addressed K-RSL. Our pilot study is the first to evaluate LLMs (GPT-4-turbo, Grok 3, Deepseek-V3, Gemini 20 Flash) for K-RSL Text2Gloss, addressing this critical gap and paving the way for accessible NLP solutions in low-resource sign languages.

3. Experimental setup

We evaluated four large language models (LLMs): GPT-4-turbo, Grok 3, Deepseek-V3, and Gemini 20 Flash — for generating K-RSL glosses from Russian sentences. These models were selected for their complementary strengths: GPT-4-turbo and Gemini 20 Flash excel in multilingual NLP, handling Russian's complex morphology; Deepseek-V3 is optimized for structured reasoning, suitable for glossing's formalized transformations; and Grok 3 offers robust general language understanding, providing a baseline for comparison.

The dataset comprises 250 short Russian sentences, each paired with expert-annotated K-RSL glosses, sourced from datasets used in linguistic studies of Kuznetsova et al. (2022); Imashev et al. (2020). A professional K-RSL interpreter curated the glosses, ensuring fidelity to K-RSL conventions. The dataset is divided into three subsets to capture K-RSL's linguistic diversity.

The first set of written translations was created in a manner that most deaf signers would write in a text message (see Figure 1).

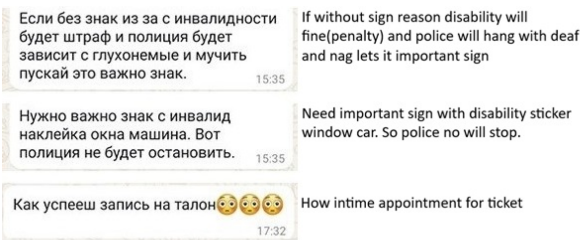

Figure 1: Examples of WhatsApp messages interpreters usually receive from deaf signers who probably have higher writing skills: not skipping prepositions, make less mistakes and use proper cases and numbers of nouns, verb forms, and tenses quite often.

The second set (see Figure 2) is similar to the previous one, but skipping prepositions or merging them with the nearest gloss if there is a sign that has that meaning such as UMENYA: U (a possessive preposition) + MENYA (an object pronoun) = UMENYA for the sign meaning "I Have" or even just "I"(with meaning HAVE); or NOT+NEED= NONEED (for the sign that means "there is no need to ..."), and others. The third set included glosses with K-RSL word order only - without any prepositions.

Additionally, we included paragraphs from Alice in Wonderland, glossed in K-RSL, to test performance on longer, narrative texts. This dataset spans diverse grammatical structures, semantic nuances, and punctuation patterns, challenging models to preserve meaning and structure.

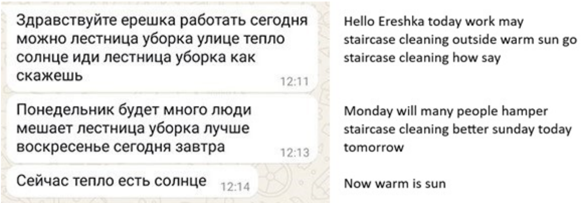

Figure 2: Examples of WhatsApp messages interpreters receive from deaf signers who skip prepositions and use words in a manner closer to glosses.

Copilot	Gemini 20 Flash
Как твоя работа?	Твоя работа как? (YOUR WORK HOW?)
У меня все отлично	Я всё отлично (I ALL GREAT)
У меня все очень плохо	Я всё очень плохо (I ALL VERY BAD)
Вы что делаете?	Вы делать что? (YOU DO WHAT?)
Я отдыхаю	Я отдыхать (I REST)
Я работаю	Я работать (I WORK)
Я смотрю телевизор	Я телевизор смотреть (I TV WATCH)
Как твое самочувствие?	Твоё чувствовать как? (YOUR FEEL HOW?)
Что случилось?	Случиться что? (TOHAPPEN WHAT?)

Figure 3: Some similar outputs of Copilot and Gemini 20 Flash

Copilot	Gemini 20 Flash	English GLOSS
Мое прозвище высокий	Моё прозвище высокий	MY NICKNAME TALL
Давно живете в Киеве?	Давно живёте в Киеве?	LONGTIME LIVE KIYV?
Поздравляю, с днем рождения	Поздравляю, с днём рождения	CONGRATULATION WITH BIRTHDAY

Figure 4: One letter difference

Each model received the prompt: “Convert each Russian sentence into a conceptual K-RSL gloss, preserving all punctuation”. The examples of selected LLMs outputs are presented in Figures 5 - 7.

We noticed that Copilot and Gemini 20 Flash produced almost the same outputs in most cases, which can be explained by the fact that Copilot utilizes Gemini under the hood. We were also interested in testing such LLMs with longer textual paragraphs. So, several fairytale passages have been used without testing Copilot (see Figure 9).

Рахмет - Рақмет

Figure 5: An example of one letter difference case keeping the same meaning in Kazakh

Outputs were compared against expert glosses using a suite of metrics: METEOR for semantic similarity, BLEU-1 to BLEU-4 for n-gram precision, BERTScore for contextual similarity, and Word Error Rate (WER) for word-level errors. These metrics evaluate both meaning preservation and structural accuracy, critical for glossing’s dual demands.

This setup enables a robust comparison of model performance across short and long texts, addressing K-RSL’s unique grammar while providing insights into LLM suitability for low-resource sign language tasks.

4. Results

Initially, our goal was to provide a comparative evaluation of the outputs of five models: we evaluated four large language models (LLMs): GPT-4-turbo, Grok 3, Deepseek-V3, Copilot, and Gemini 20 Flash.

For several cases from the first 20 sentences, Copilot and Gemini 20 Flash provided very similar output, but Gemini 20 Flash outputs looked much closer to the real GLOSS manner (see Figure 3). In only several cases, the difference was just about one letter, which in modern written Russian is allowed to be replaced if it does not change the meaning (see Figure 4); similar cases may happen for the Kazakh language as well, such as *ra**kh**met* and *ra**q**met* (meaning "thanks") (see Figure 5). The other longer sentences were almost the same as the given sentences.

Thus, it has been decided to compare four large language models GPT-4-turbo, Grok 3, Deepseek-V3, and Gemini 20 Flash on generating K-RSL glosses from 250 Russian sentences and Alice in Wonderland paragraphs, comparing outputs to expert-annotated glosses. Performance was assessed using METEOR (semantic similarity), BLEU-4 (n-gram precision), BERTScore (contextual similarity) and WER (word-level errors), capturing meaning preservation and structural accuracy. Table 1 summarizes the results in three sentence subsets (Set 1: gloss manner but closer to normal text, Set 2: gloss manner with missed or merged prepositions, and Set 3: K-RSL word order) and fairytale paragraphs. To provide a detailed comparison of the performance of the LLMs evaluated across the four experimental sets, we compiled a comprehensive table of evaluation metrics. This table includes METEOR, CHRF, BERTScore, BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, WER,

Given sentences	Ground-truth	GPT-4-turbo	Grok 3	Deepseek-V3	Copilot	Gemini 20 Flash
У меня вчера было день рождения // It was my birthday yesterday	У МЕНЯ ВЧЕРА БЫЛО ДЕНЬ УШИ ТЯНУТЬ // 'U' ME YESTERDAY WAS DAY EARS PULL	Вчера мой день рождение	Я день рождения вчера	У меня вчера быть день рождение	У меня вчера было день рождения	У меня вчера было день рождения
Я работаю плотником в мастерской по металлу // I work as a carpenter in a metal workshop	Я РАБОТА ПЛОТНИК МЕСТО МАСТЕРСКОЙ О ЖЕЛЕЗО // I WORK CARPENTER PLACE WORKSHOP ABOUT METAL	Я работать плотник металл мастерская	Я работа плотник в мастерской по металлу	Я работать плотник мастерская металл	Я работаю плотником в мастерской по металлу	Я работаю плотником в мастерской по металлу
Дорогой дом построили, а автомобиль продали // They built an expensive house, but sold the car	ДОРОГОЙ ДОМ СТРОЙКА ГОТОВА, АВТОМОБИЛЬ ПРОДАЖА // EXPENSIVE HOUSE BUILDING READY, CAR SELL	Дом дорогой построить, авто продать	Дорогой дом постройка, а автомобиль продажа	Дорогой дом построить, а автомобиль продать	Дорогой дом построили, а автомобиль продали	Дорогой дом построили, а автомобиль продали
Нет, им сейчас не нужен адвокат // No, they don't need an advocate now	НЕТ, ОНИ СЕЙЧАС АДВОКАТ НЕНУЖНО // NO, THEY NOW ADVOCATE NONEED	Нет, им адвокат не нужен	Нет, им сейчас адвокат нет	Нет, они сейчас не нужный адвокат	Нет, им сейчас не нужен адвокат	Нет, им сейчас не нужен адвокат
Военный едет домой в Астану на своей машине // A serviceman drives his car home to Astana	ВОЕННЫЙ СВОЙ РУЛЬ НАПРАВЛЕНИЕ ВПЕРЕД МЕСТО ДОМ АСТАНА // SERVICEMAN OWN STEERINGWHEEL DIRECTION FORWARD PLACE HOME ASTANA	Военный домой Астана ехать на машина	Военный дом в Астану на своей машине	Военный ехать дом Астана свой машина	Военный едет домой в Астану на своей машине	Военный едет домой в Астану на своей машине

Figure 6: Examples of LLMs outputs for set #1

Given sentences	Ground-truth	GPT-4-turbo	Grok 3	Deepseek-V3	Copilot	Gemini 20 Flash
Помогите, голова болит третий день // Help, my head has been hurting for three days	ПОМОГИТЕ, ГОЛОВАБОЛИТ ТРЕТИЙДЕНЬ ПРОШЛО //HELP HEADPAIN THIRDDAY PASS	Помогите, голова болить 3 день!	Помощь, голова боль третий день	Помогать, голова болеть третий день	Помогите, голова болит третий день	Помогите, голова болит третий день
Нужно лекарство от простуды, температура высокая // I need cold medicine, I have a high temperature	ТЕМПЕРАТУРА ВЫСОКО НУЖНО ЛЕКАРСТВО ПРОТИВ СОПЛИ ГОРЛО // TEMP. HIGH NEED MEDICINE AGAINST SNOT THROAT	Нужно лекарство, температура высокая	Лекарство от простуды, температура высокая	Нужно лекарство от простуда, температура высокий	Нужно лекарство от простуды, температура высокая	Нужно лекарство от простуды, температура высокая
Нет, им сейчас не нужен адвокат (защита чел.) // No, they don't need a lawyer (defence person) now	НЕТ, ИМ СЕЙЧАС НЕНУЖНО АДВОКАТ (ЗАЩИТА ЧЕЛ) // NO THEY NOW NONEED ADVOCATE (DEVENCE PERSON)	Нет, им адвокат не нужен	Нет, им сейчас адвокат (защита чел) нет	Нет, они сейчас не нужный адвокат (защита человек)	Нет, им сейчас не нужен адвокат (защита чел)	Нет, им сейчас не нужен адвокат
Военный едет домой в Астану на своей машине // A serviceman drives his car home to Astana	ВОЕННЫЙ СВОЕЙ РУЛЬНАПРАВЛЕНИЕВПЕРЕД МЕСТО ДОМ АСТАНА // SERVICEMAN OWN STEERINGWHEELDIRECTIONFORWARD PLACE HOME ASTANA	Военный домой Астана ехать на машина	Военный дом в Астану на своей машине	Военный ехать дом Астана свой машина	Военный едет домой в Астану на своей машине	Военный едет домой в Астану на своей машине

Figure 7: Examples of LLMs outputs for set #2

Given sentences	Ground-truth	GPT-4-turbo	Grok 3	Deepseek-V3	Copilot	Gemini 20 Flash
Где можно сделать операцию бесплатно? // Where can surgery be done for free?	ГДЕ МЕСТО ОПЕРАЦИЯ БЕСПЛАТНО ДЕЛАТЬ МОЖНО? // WHERE PLACE SURGERY FREE DO MAY?	Где операция бесплатно сделать?	Операция бесплатно где?	Где можно сделать операция бесплатно?	Где можно сделать операцию бесплатно?	Где можно сделать операцию бесплатно?
Папе выдали бесплатные лекарства от аллергии // Dad was given free allergy medication	ЛЕКАРСТВО АЛЛЕРГИЯ БОРЬБА ПАПА ВЗЯТЬ БЕСПЛАТНО ВСЕ //MEDICINE ALLERGY FIGHT FATHER TAKE FREE ALL	Папа лекарство бесплатно получить	Папа бесплатные лекарства от аллергии	Папа выдать бесплатный лекарство от аллергия	Папе выдали бесплатные лекарства от аллергии	Папе выдали бесплатные лекарства от аллергии
Режим приема лекарств поменяли // The medication regimen has been changed	МНЕ ТАБЛЕТКА ПИТЬ РАСПИСАНИЕ МЕНЯТЬ ВСЕ //TOME PILL DRINK SCHEDULE CHANGE ALL	Режим лекарство прием изменить	Режим приём лекарств смена	Режим прием лекарство поменять	Режим приема лекарств поменяли	Режим приёма лекарств поменяли
Сестре выдали бесплатную путевку // My sister was given a free trip.	СЕСТРА ОТДЫХ БУМАГА ПОЛУЧИТЬ БЕСПЛАТНО РАНЬШЕ // SISTER VACATION PAPER GET FREE EARLY	Сестра путевка бесплатно получить	Сестра бесплатная путёвка	Сестра выдать бесплатный путёвка	Сестре выдали бесплатную путевку	Сестре выдали бесплатную путёвку

Figure 8: Examples of LLMs outputs for set #3

Given sentences	In english	Ground-truth	GPT-4-turbo	Grok 3	Deepseek-V3	Gemini 20 Flash
ОНА СИДЕЛА И РАЗМЫШЛЯЛА, НЕ ВСТАТЬ ЛИ ЕЙ И НЕ НАРВАТЬ ЛИ ЦВЕТОВ ДЛЯ ВЕНКА; МЫСЛИ ЕЕ ТЕКЛИ МЕДЛЕННО И НЕСВЯЗНО - ОТ ЖАРЫ ЕЕ КЛОНИЛО В СОН. КОНЕЧНО, СПЛЕСТИ ВЕНОК БЫЛО БЫ ОЧЕНЬ ПРИЯТНО, НО СТОИТ ЛИ РАДИ ЭТОГО ПОДЫМАТЬСЯ?	SHE SAT AND CONSIDERED WHETHER SHE SHOULD GET UP AND PICK SOME FLOWERS FOR A WREATH; HER THOUGHTS WERE SLOW AND DISORDERED - THE HEAT MADE HER DROZEN. OF COURSE, WEAVING A WREATH WOULD BE VERY PLEASANT, BUT WAS IT WORTH GETTING UP?	ОНА СИДЕТЬ, ДУМАТЬ ВСТАТЬ ЛИ? ЦВЕТЫ РВАТЬ ЛИ? ВЕНОК ДЕЛАТЬ ЛИ? МЫСЛИ МЕДЛЕННО ИДТИ, ЖАРА, ОНА СПАТЬ ХОТЕТЬ.ВЕНОК ДЕЛАТЬ - ПРИЯТНО. НО ВСТАТЬ НАДО ЛИ?	ОНА СИДЕТЬ, ДУМАТЬ: ВСТАТЬ? ЦВЕТЫ СОБРАТЬ? ВЕНОК СДЕЛАТЬ? МЫСЛИ МЕДЛЕННО, БЕСПОРЯДОК. ЖАРКО — ХОЧЕТСЯ СПАТЬ. КОНЕЧНО, ВЕНОК ДЕЛАТЬ — ПРИЯТНО. НО — СТОИТ ЛИ ВСТАВАТЬ?	ОНА СИДЕТЬ ДУМАТЬ, ВСТАТЬ-МОЖЕТ ЦВЕТЫ СОБРАТЬ ВЕНОК ДЛЯ; МЫСЛИ МЕДЛЕННО НЕСВЯЗНО ТЕЧЬ - ЖАРА СОН КЛОНИТЬ. КОНЕЧНО ВЕНОК ПЛЕСТИ ПРИЯТНО ОЧЕНЬ, НО СТОИТ-ЛИ ВСТАТЬ ДЛЯ ЭТО?	ОНА СИДЕТЬ, РАЗМЫШЛЯТЬ — ВСТАТЬ? СОБРАТЬ ЦВЕТЫ ВЕНОК? МЫСЛИ ТЕЧЬ МЕДЛЕННО, НЕСВЯЗНО — ЖАРА. ВЕНОК СПЛЕСТИ — ПРИЯТНО , НО : ВСТАВАТЬ ?	ОНА СИДЕТЬ И ДУМАТЬ , НЕ ВСТАТЬ ЛИ ОНА И НЕ НАРВАТЬ ЛИ ЦВЕТЫ ДЛЯ ВЕНОК ; МЫСЛИ ЕЕ ТЕЧЬ МЕДЛЕННО И НЕСВЯЗНО - ОТ ЖАРА ЕЕ КЛОНИТЬ В СОН . КОНЕЧНО , СПЛЕСТИ ВЕНОК БЫТЬ ОЧЕНЬ ПРИЯТНО , НО СТОИТ ЛИ РАДИ ЭТОГО ПОДЫМАТЬСЯ ?
ВДРУГ МИМО ПРОБЕЖАЛ БЕЛЫЙ КРОЛИК С КРАСНЫМИ ГЛАЗАМИ. КОНЕЧНО, НИЧЕГО УДИВИТЕЛЬНОГО В ЭТОМ НЕ БЫЛО. ПРАВДА, КРОЛИК НА БЕГУ ГОВОРИЛ: - АХ, БОЖЕ МОЙ, БОЖЕ МОЙ! Я ОПАЗДЫВАЮ.	SUDDENLY A WHITE RABBIT WITH RED EYES RAN PAST. OF COURSE, THERE WAS NOTHING SURPRISING ABOUT THIS. HOWEVER, AS THE RABBIT RAN, HE SAID: - OH, MY GOD, OH MY GOD! I'M LATE.	ВДРУГ БЕЛЫЙ КРОЛИК БЕГАТЬ МИМО, ГЛАЗА КРАСНЫЙ. НИЧЕГО УДИВИТЬ НЕТУ. НО КРОЛИК БЕГАТЬ И ГОВОРИТЬ: "АЙ-АЙ! ОПАЗДЫВАТЬ!"	ВНЕЗАПНО — БЕЛЫЙ КРОЛИК БЕЖАТЬ МИМО. ГЛАЗЫ КРАСНЫЕ. КОНЕЧНО, УДИВЛЕНИЕ — НЕТ. НО КРОЛИК БЕЖАТЬ, ГОВОРИТ: "БОЖЕ МОЙ! БОЖЕ МОЙ! Я ОПАЗДЫВАТЬ!"	ВДРУГ КРОЛИК БЕЛЫЙ ГЛАЗА-КРАСНЫЕ МИМО БЕЖАТЬ; КОНЕЧНО НИЧЕГО УДИВИТЕЛЬНО НЕТ. ПРАВДА КРОЛИК БЕЖАТЬ ГОВОРИТЬ: - АХ БОЖЕ-МОЙ БОЖЕ-МОЙ! Я ОПАЗДЫВАТЬ.	ВДРУГ БЕЛЫЙ КРОЛИК ПРОБЕЖАТЬ, КРАСНЫЕ ГЛАЗА. КОНЕЧНО, НИЧЕГО УДИВИТЕЛЬНОГО. НО: КРОЛИК ГОВОРИТЬ — «АХ, БОЖЕ МОЙ, БОЖЕ МОЙ! Я ОПАЗДЫВАЮ!».	ВДРУГ ПРОБЕЖАТЬ БЕЛЫЙ КРОЛИК КРАСНЫЕ ГЛАЗА . КОНЕЧНО , НИЧЕГО УДИВИТЕЛЬНОГО В ЭТОМ НЕ БЫЛО . ПРАВДА , КРОЛИК НА БЕГУ ГОВОРИТЬ : - АХ , БОЖЕ МОЙ , БОЖЕ МОЙ ! Я ОПАЗДЫВАТЬ .

Figure 9: Examples of LLMs outputs for the fairytale paragraphs

Exact Match, and TER, capturing semantic fidelity, structural accuracy, and error rates for each model (see Table 2 with consolidated performance metrics for all evaluated datasets (Set #1, Set #2, Set #3, and "Fairytale" paragraphs.

Gemini 20 Flash excels on Sets 1 and 2, achieving the highest METEOR (0.602, 0.592), BLEU-4 (0.265, 0.253), and BERTScore (0.895, 0.890), and lowest WER (0.541, 0.575). This suggests strong semantic and structural fidelity for short glosses, likely due to its multilingual NLP capabilities. Deepseek-V3 outperforms others on Set 3 (METEOR: 0.426, BLEU-4: 0.135) and fairytale paragraphs (METEOR: 0.377, BLEU-4: 0.073), indicating robustness for complex K-RSL grammar, attributed to its structured reasoning design. GPT-4-turbo shows moderate performance (e.g. METEOR: 0.394, WER: 0.675 for Set 3), while Grok 3 consistently lags (e.g., METEOR: 0.377, WER: 0.745 on Set 2), possibly due to less optimization for K-RSL's linguistic nuances.

High BERTScore values (0.790–0.895) across models indicate robust conceptual equivalence, despite low BLEU-4 scores (0.008–0.265), reflecting glossing's tolerance for functional variation (e.g., "I GO SHOP" vs. "I SHOP GO"). WER varies widely (0.541–1.706), with Gemini and Deepseek-V3 minimizing errors in shorter texts. Fairytale results show lower performance overall, likely due to narrative complexity, with Deepseek-V3 maintaining the lead. These findings highlight the suitability of Gemini 20 Flash for concise glosses and the strength of Deepseek-V3 for longer, grammatically intricate texts, informing model selection for K-RSL accessibility applications.

5. Discussion

Our evaluation demonstrates that large language models (LLMs) can effectively generate K-RSL glosses, paving the way for automated accessibility tools for Kazakhstan's Deaf and Hard of Hearing community. Gemini 20 Flash excels on short, text-like glosses (METEOR: 0.602 for Set 1, 0.592 for Set 2), adeptly capturing K-RSL's concise, preposition-free syntax, likely due to its robust multilingual NLP capabilities. Deepseek-V3 outperforms on complex texts (METEOR: 0.426 for Set 3, 0.377 for fairytale paragraphs), navigating K-RSL's non-linear word order and gestural polysemy through structured reasoning. High BERTScore values (0.790–0.895) across models indicate strong semantic equivalence, while moderate WER (0.541–1.706) reflects K-RSL's tolerance for functional variations, such as "I GO SHOP" versus "I SHOP GO". These results highlight LLMs' ability to address K-RSL's linguistic challenges, enabling scalable gloss production for applications like educational platforms, sign language subtitles, and literacy aids.

Some studies already utilize computer-based text simplification to enhance the clarity of inquiries; methodologies for text simplification in Kazakh (Toleu et al., 2025) and Russian (Dmitrieva et al., 2021) have already been established. However, in our case, there are concerns that text simplification may not adequately address the challenges posed by sign polysemy and the requirement to use glosses from the selected K-RSL sign language, thereby ensuring that Deaf and Hard of Hearing (DHH) respondents can fully understand the texts. The comprehension of concepts with closely related translations in sign language can be particularly confusing for deaf participants.

Ethically, LLMs also risk encoding biases or oversimplifying K-RSL's culturally nuanced signs, potentially misrepresenting context-dependent gestures and erode trust between deaf signers. Co-design with K-RSL communities is critical to validate outputs and ensure cultural fidelity. By reducing the

Model	**METEOR**	**BLEU-4**	**BERTScore**	**WER**
		Set 1: Text-like Glosses		
GPT-4-turbo	0.392	0.121	0.846	0.690
Grok 3	0.383	0.107	0.830	0.722
Deepseek-V3	0.523	0.225	0.885	0.545
Gemini 20 Flash	**0.602**	**0.265**	**0.895**	**0.541**
		Set 2: Merged Prepositions		
GPT-4-turbo	0.388	0.119	0.843	0.705
Grok 3	0.377	0.105	0.827	0.745
Deepseek-V3	0.517	0.220	0.880	0.573
Gemini 20 Flash	**0.592**	**0.253**	**0.890**	**0.575**
		Set 3: K-RSL Word Order		
GPT-4-turbo	0.394	0.122	0.850	**0.675**
Grok 3	0.361	0.107	0.833	0.715
Deepseek-V3	**0.426**	**0.135**	**0.852**	0.695
Gemini 20 Flash	0.309	0.089	0.829	0.811
		Fairytale Paragraphs		
GPT-4-turbo	0.343	0.016	0.821	0.980
Grok 3	0.277	0.020	0.813	1.006
Deepseek-V3	**0.377**	**0.073**	**0.839**	**0.912**
Gemini 20 Flash	0.259	0.008	0.790	1.706

Table 1: Performance of LLMs on K-RSL Text2Gloss. Best scores are **bolded**.

Model	**METEOR**	CHRF	**BERTScore**	BLEU-1	BLEU-2	BLEU-3	**BLEU-4**	ROUGE-L	**WER**	Exact Match	TER
					Set 1: Text-like Glosses						
GPT-4-turbo	0.392	50.20	0.846	0.373	0.232	0.158	0.121	0.392	0.690	0.073	65.26
Grok 3	0.383	46.83	0.830	0.341	0.205	0.143	0.107	0.369	0.722	0.053	68.38
Deepseek-V3	0.523	62.16	0.885	0.544	0.388	0.290	0.225	0.555	0.545	0.186	52.08
Gemini 20 Flash	0.602	65.21	0.895	0.584	0.433	0.343	0.265	0.582	0.541	0.235	48.54
					Set 2: Merged Prepositions						
GPT-4-turbo	0.388	49.97	0.843	0.367	0.226	0.155	0.119	0.386	0.705	0.069	66.63
Grok 3	0.377	46.57	0.827	0.332	0.197	0.139	0.105	0.361	0.745	0.053	70.26
Deepseek-V3	0.517	61.73	0.880	0.530	0.376	0.280	0.220	0.541	0.573	0.178	54.26
Gemini 20 Flash	0.592	64.58	0.890	0.567	0.412	0.324	0.253	0.565	0.575	0.223	51.24
					Set 3: K-RSL Word Order						
GPT-4-turbo	0.394	48.56	0.850	0.408	0.222	0.160	0.122	0.418	**0.675**	0.048	**63.25**
Grok 3	0.361	42.83	0.833	0.350	0.197	0.144	0.107	0.378	0.715	0.057	68.56
Deepseek-V3	**0.426**	**53.80**	**0.852**	**0.439**	**0.249**	**0.176**	**0.135**	**0.443**	0.695	**0.091**	65.16
Gemini 20 Flash	0.309	43.89	0.829	0.317	0.148	0.107	0.089	0.328	0.811	0.029	78.15
					Fairytale Paragraphs						
GPT-4-turbo	0.343	41.08	0.821	0.237	0.100	0.030	0.016	0.298	0.980	0.000	95.93
Grok 3	0.277	41.31	0.813	0.252	0.073	0.033	0.020	0.270	1.006	0.000	98.99
Deepseek-V3	**0.377**	**44.92**	**0.839**	**0.291**	**0.143**	**0.103**	**0.073**	**0.317**	**0.912**	0.000	**90.16**
Gemini 20 Flash	0.259	37.96	0.790	0.149	0.033	0.013	0.008	0.225	1.706	0.000	170.57

Table 2: Consolidated results of LLMs on K-RSL Text2Gloss. Best scores are **bolded**.

reliance on manual annotation, LLMs can democratize access to K-RSL resources, but their deployment must prioritize community-driven evaluation to align with deaf signers' needs and linguistic diversity.

6. Limitations

The dataset's modest size (250 sentences) and a single annotator may limit generalizability across K-RSL's varied registers, such as informal versus formal signing. The fairytale subset, while useful for testing narrative complexity, is less representative of everyday K-RSL communication, constraining practical insights for real-world applications. Additionally, the evaluation focused on gloss accuracy without assessing usability by deaf signers, a critical gap for accessibility tools. Future studies should incorporate larger, multi-annotator datasets and participatory testing with K-RSL communities to enhance robustness and relevance.

7. Conclusions and Future Work

This pilot study is the first to evaluate LLMs for K-RSL Text2Gloss, demonstrating Gemini 20 Flash's efficacy for short glosses and Deepseek-V3's strength for complex texts. These findings highlight LLMs' potential to automate gloss production, improving accessibility for deaf signers in Kazakhstan. Future work includes: (1) expanding datasets with diverse K-RSL texts, (2) engaging multiple annotators for robust validation, (3) conducting usability studies with deaf signers, and (4) fine-tuning LLMs to capture K-RSL's gestural and cultural nuances.

To this end, there is still room for dataset improvement (greater datasets are needed for K-RSL), nuances of each sign language should be taken into account for the future step of proper Gloss2Text interpetation: for example, glosses related to professions are usually presented by glosses of two signs (DOCTOR = HEALTHCARE + PERSON), or concepts that may also be represented by several glosses (FRIDGE = FREEZE + BOX). Other types of gloss-sign mismatches. For example, ME and I are manifested by the same sign. Sign languages have unique grammatical structures that differ from spoken languages. Phenomena that occur in the local sign language, such as sign variability, sign polysemy, or phonological minimal pairs (Imashev et al., 2024a) should also be taken into account for future linguistically correct Gloss2Text interpretation. For example, there is a sign that has three different meanings: WEATHER, NATURE, and AIR, which should be stored like WNA-1, WNA-2, and WNA-3 in the corpus, but appear proper words (meanings) for user-end translation.

Neural Machine Translation (NMT) has arisen as a cutting-edge technology to address the deficiencies of earlier translation methods, especially in low-resource contexts (Ranathunga et al., 2023; Haque et al., 2021), and has already been implemented in the Kazakhstani setting (Yeshpanov et al., 2024; Kozhirbayev, 2024; Kozhirbayev and Islamgozhayev, 2023; Yessenbayev et al., 2020).

The sign language used in Kazakhstan is based on the signing system developed in the USSR, so deaf signers usually use the same signs for the same concepts, but with different mouthings related to the spoken language (Kazakh or Russian), except for local concepts (traditional dishes, famous figures, and landmarks). Nevertheless, testing the gloss generation from Kazakh and code-switching written sentences will be quite interesting, considering that nowadays people in Kazakhstan often utilize components of both languages in informal text communication.

Conducting a user study juxtaposing LLMs from the perspective of potential end users (deaf signers) to determine which model may be more convenient and desirable for their everyday use is also advisable.

Despite the promising performance of several LLMs tested, LLM outputs should be evaluated by potential end-user stakeholders: native deaf signers. Let all outputs be evaluated by deaf people of different age, education, occupation, and skills. Perhaps a separate methodology should be established on the evaluation of gloss annotations by deaf signers in a questionnaire manner, similar to (Mich, 2009; Imashev et al., 2022; Fuentes-Cortázar and Rojano-Cáceres, 2024).

In our case, the fairytale ground truth gloss interpretation was performed by a single professional interpreter; nevertheless, it is preferable for each ground truth gloss interpretation to be performed collaboratively by several proficient sign language interpreters. A written fairytale interpretation into glosses seems not as challenging as interpretation of news or scientific/encyclopedic information. So, such interpretations should be tested as well.

More accurate interpretation into glosses can significantly assist deaf students and adults in comprehending information more effectively and faster, while also alleviating cognitive burdens. This might become a useful additional tool for the education of deaf signers.

Acknowledgments

This work was supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23489529).

We would like to extend our gratitude to Gulmira Baizhanova, an interpreter of the local sign language, for her involvement and support, as well as the native signers who provided the examples of their text messages.

8. Bibliographical References

Faisl M Alqraini. 2018. Identifying similarities and differences on how deaf and hard of hearing students learn new vocabulary knowledge. *International Journal of Instruction*, 11(4):61–74.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Breno W Carvalho, Artur S Garcez, Luís C Lamb, and Emílio Vital Brazil. 2025. Grokking explained: A statistical phenomenon. *arXiv preprint arXiv:2502.01774*.

José Mario De Martino and Dener Stassun Christinele. 2024. Exploring pre-trained transformers for translating portuguese text to brazilian sign language. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 67–75.

Anna Dmitrieva, Jörg Tiedemann, et al. 2021. Creating an aligned russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*. ACL Anthology.

Santiago Egea Gómez, Luis Chiruzzo, Euan McGill, and Horacio Saggion. 2022. Linguistically enhanced text to sign gloss machine translation. In *International Conference on Applications of Natural Language to Information Systems*, pages 172–183. Springer.

Andrés Eduardo Fuentes-Cortázar and José Rafael Rojano-Cáceres. 2024. Methods for evaluating the user experience of applications for deaf users: A mapping review. In *2024 12th International Conference in Software Engineering Research and Innovation (CONISOFT)*, pages 287–294. IEEE.

Santiago Egea Gómez, Euan McGill, and Horacio Saggion. 2021. Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th workshop on building and using comparable corpora (BUCC 2021)*, pages 18–27.

Antonia González-Cuenca, Marina González-Jerez, María José Linero, and Rocío Lavigne. 2024. Differences in the type of vocabulary understood by deaf and hearing students: Results to guide interventions. *Journal of Communication Disorders*, 111:106458.

Rejwanul Haque, Chao-Hong Liu, and Andy Way. 2021. Recent advances of low-resource neural machine translation. *Machine Translation*, 35(4):451–474.

Iva Hrastinski and Ronnie B Wilbur. 2016. Academic achievement of deaf and hard-of-hearing students in an asl/english bilingual program. *Journal of deaf studies and deaf education*, 21(2):156–170.

Alfarabi Imashev, Aigerim Kydyrbekova, Medet Mukushev, Anara Sandygulova, Shynggys Islam, Khassan Israilov, Aibek Makazhanov, and Zhandos Yessenbayev. 2024a. Retrospective of kazakh-russian sign language corpus formation. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 111–122.

Alfarabi Imashev, Aigerim Kydyrbekova, Nurziya Oralbayeva, Azamat Kenzhekhan, and Anara Sandygulova. 2024b. Learning sign language with mixed reality applications-the exploratory case study with deaf students. *Education and information technologies*, 29(13):17261–17292.

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. K-rsl: a corpus for linguistic understanding, visual evaluation, and recognition of sign languages. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Alfarabi Imashev, Nurziya Oralbayeva, Vadim Kimmelman, and Anara Sandygulova. 2022. A user-centered evaluation of the data-driven sign language avatar system: A pilot study. In *Proceedings of the 10th international conference on human-agent interaction*, pages 194–202.

Muhammad Imran and Norah Almusharraf. 2024. Google gemini as a next generation ai educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22.

Zhanibek Kozhirbayev. 2024. Enhancing neural machine translation with fine-tuned mbart50 pre-trained model: An examination with low-resource translation pairs. *Ingenierie des Systemes d'Information*, 29(3):831.

Zhanibek Kozhirbayev and Talgat Islamgozhayev. 2023. Cascade speech translation for the kazakh language. *Applied Sciences*, 13(15):8900.

Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman. 2022. Functional data analysis of non-manual marking of questions in kazakh-russian sign language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*. European Language Resources Association (ELRA).

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Ornella Mich. 2009. Evaluation of software tools with deaf children. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 235–236.

Donna A Morere. 2012. Measures of writing, math, and general academic knowledge. In *Assessing literacy in deaf individuals: Neurocognitive measurement and predictors*, pages 127–137. Springer.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Rubén San-Segundo, Juan Manuel Montero, Ricardo Cordoba, Valentin Sama, F Fernández, Luis Fernando D'Haro, Verónica López-Ludeña, D Sánchez, and Antonio García. 2012. Design, development and field evaluation of a spanish into sign language translation system. *Pattern Analysis and Applications*, 15:203–224.

Thomastine Sarchet, Marc Marschark, Georgianna Borgna, Carol Convertino, Patricia Sapere, and Richard Dirmyer. 2014. Vocabulary knowledge of deaf and hearing postsecondary students. *Journal of postsecondary education and disability*, 27(2):161.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wasswa Shafik. 2024. Introduction to chatgpt. In *Advanced applications of generative AI and natural language processing models*, pages 1–25. IGI Global.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Noboru Takahashi, Yukio Isaka, Toshikazu Yamamoto, and Tomoyasu Nakamura. 2017. Vocabulary and grammar differences between deaf and hearing students. *The Journal of Deaf Studies and Deaf Education*, 22(1):88–104.

Alymzhan Toleu, Gulmira Tolegen, and Irina Ualiyeva. 2025. Fine-tuning large language models for kazakh text simplification. *Applied Sciences*, 15(15).

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. KazParC: Kazakh parallel corpus for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9633–9644, Torino, Italia. ELRA and ICCL.

Zhandos Yessenbayev, Zhanibek Kozhirbayev, and Aibek Makazhanov. 2020. Kaznlp: A pipeline for automated processing of texts written in kazakh language. In *Speech and Computer*, pages 657–666, Cham. Springer International Publishing.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541.