Do LLMs Speak Kazakh? A Pilot Evaluation of Seven Models

Akylbek Maxutov[1], Ayan Myrzakhmet[2], Pavel Braslavski[2]
[1]Institute of Smart Systems and Artificial Intelligence, Astana, Kazakhstan
[2]School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan
pavel.braslavskii@nu.edu.kz

Abstract

We conducted a systematic evaluation of seven large language models (LLMs) on tasks in Kazakh, a Turkic language spoken by approximately 13 million native speakers in Kazakhstan and abroad. We used six datasets corresponding to different tasks – questions answering, causal reasoning, middle school math problems, machine translation, and spelling correction. Three of the datasets were prepared for this study. As expected, the quality of the LLMs on the Kazakh tasks is lower than on the parallel English tasks. GPT-4 shows the best results, followed by Gemini and AYA. In general, LLMs perform better on classification tasks and struggle with generative tasks. Our results provide valuable insights into the applicability of currently available LLMs for Kazakh. We made the data collected for this study publicly available: https://github.com/akylbekmaxutov/LLM-eval-using-Kazakh.

1 Introduction

Large Language Models (LLMs) increase human productivity and eliminate routine tasks in many areas, making them a powerful economic driver (Eloundou et al., 2023; Butler et al., 2023). At the same time, LLMs can lead to an inequality between different language communities and a widening gap between developed and developing countries (Khowaja et al., 2024). Creating LLMs requires huge amounts of text and computation, as well as skilled engineers. Most LLMs are trained for high-resource languages with large populations of speakers, primarily English. Training language models for low-resource languages can be technically and economically problematic – there is little training data, and it is unclear whether potential users can amortize the cost of collecting data and training the model. Although models trained primarily on English data express capabilities in other languages, their quality in these secondary languages is lower than in English (Ahuja et al., 2023). Recently, thanks to the advent of open LLMs, their adaptations to less-resourced languages are emerging (Qin et al., 2024). Evaluating LLMs in different languages is crucial in this situation.

Source	en	tr	kk
CulturaX	2.8T	64.3B	2.8B
Wiki pages	6.8M	610K	236K
HF datasets	10,889	402	120
HF models	51,365	1,403	458

Table 1: Overview of available Kazakh (kk) language resources compared to English (en) and Turkish (tr): # tokens in the CulturaX (Nguyen et al., 2023) dataset, # Wikipedia pages, and datasets/models on Huggingface.

In this study, we make the first attempt to evaluate the quality of available LLMs in Kazakh. Kazakh belongs to the Turkic language family and is the official language of the Republic of Kazakhstan (Campbell and King, 2020). Estimated 10 million Kazakh native speakers live in Kazakhstan, and about 3 million more abroad, predominantly in north-western China and western Mongolia. The language employs an extended Cyrillic alphabet with 42 letters. Kazakh is an agglutinative language, meaning that words are formed by adding various suffixes to root words. The language's rich inflectional morphology is reflected in the complex interaction of suffixes for number, possession, and case. For instance, the plural form, possessive affixes, and various case endings are layered sequentially onto noun roots. Kazakh has eight types of possessive agreements, adding complexity to its morphological structure. Kazakh verbs exhibit similar tenses and moods as Turkish ones but include unique tenses such as the goal-oriented future tense. Kazakh consonant and vowel harmony rules significantly affect its morphological structure. Consonant harmony determines the form of suffixes based on the voicing of the final consonant of the root word, while vowel harmony aligns suf-

Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), pages 81–91

fix vowels with the vowel type (front or back) of the root. Kazakh is considered a mid-resourced language (Joshi et al., 2020). Table 1 provides a brief statistics of resources available for Kazakh along with the figures for English and Turkish for comparison.

We experimented with *seven* models in total – five closed (GPT 3.5 and 4, Gemini 1.5 Pro, YandexGPT 2 and 3) and two open (LLAMA 2 and AYA) ones.[1] We focused on automatic benchmark-based evaluation, while trying to make the set of tasks diverse. We used a collection of six datasets sourced in different ways: 1) existing multilingual benchmarks that include Kazakh data (machine translation and multiple-choice question answering), 2) the recently published monolingual question answering dataset KazQAD (both open- and closed-book scenarios), 3) machine-translated COPA dataset[2] (commonsense causal reasoning), 4) original math school problems in Kazakh that we scraped online and post-processed, and 5) a Kazakh spelling correction dataset that we created from scratch within this study.

Based on our experiments, we can conclude that the GPT-4 is the most capable of all the models in the experiment. Gemini is the runner-up in the classification tasks. AYA is quite competitive, especially if we take into account its relatively small size and a long list of supported languages. All models show a lower quality in the generative tasks. As expected, the quality on Kazakh tasks is significantly lower than on English tasks, as we can see on parallel multilingual datasets (multiple-choice question answering, causal reasoning). Specialized models may still provide better quality for downstream tasks, such as machine translation or classification tasks. We cannot confirm previous findings that English prompts systematically improve LLM quality on non-English tasks: our results are mixed across tasks and models.

Our findings provide valuable insights into the applicability of currently available LLMs for Kazakh. We also anticipate that the study will contribute to the methodology of evaluating LLMs and improving the quality of LLMs in mid- and low-resource languages. The methods introduced in our work can be used to experiment with other languages and LLMs. We made the data and evaluation code publicly available.[3]

2 Related Work

As has been shown by Blevins and Zettlemoyer (2022), multilingual abilities of language models emerge when they are exposed even to a tiny fraction of non-English data in a large pre-training corpus. Earlier studies demonstrated that multilingual models learn high-level abstractions common to all languages, which make cross-lingual transfer possible even when languages share no vocabulary (Wu and Dredze, 2019). Open LLMs such as LLAMA (Touvron et al., 2023) and Qwen (Bai et al., 2023) can be adapted to other languages by expanding their vocabularies, continual pre-training and subsequent aligning on the data in target language (Qin et al., 2024). Another approach is to train a model from scratch: for example, Jais model was trained on a mixture of English and Arabic data in ratio 2:1 (Sengupta et al., 2023). Despite the development of non-English and multilingual models, many languages remain underrepresented in the modern LLM landscape. This situation is partly due to objective reasons (lack of training data), but also to inequalities in economic and technological development.

LLM evaluation is a complex and multifaceted problem (Chang et al., 2024). LLMs are truly multitasking, and users can leverage them to solve non-standard and creative problems, for example, brainstorming ideas or generating jokes. For generative tasks, the variety of formulations can be very large, making it difficult to automatically compare the answer to a "gold standard." With the proliferation of LLMs and their active use, evaluation of models becomes relevant not only at the task level, but also from their safety and security perspectives. The main approach to automatic LLM evaluation is based on ensembles of annotated benchmarks covering a wide range of usage scenarios (Liang et al., 2022). Popular benchmarks include MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021) that measures LLM's knowledge across 57 subjects and GSM8K (Grade School Math) (Cobbe et al., 2021), aimed at evaluating multi-step math reasoning. MMLU contains multiple-choice ques-

[1] mGPT (Shliazhko et al., 2024) is another LLM that officially supports Kazakh. However, only a pre-trained mGPT is available, while the models in the study are instruction tuned.

[2] In the spring of 2024, while our study was underway, the Kardeş-NLU for five Turkic languages, including Kazakh, was published (Senel et al., 2024). The dataset includes a post-edited version of COPA.

[3] https://github.com/akylbekmaxutov/LLM-eval-using-Kazakh

tions, while GSM8K accepts numerical answers. There are multilingual adaptions of these datasets: Lai et al. (2023b) employed ChatGPT to translate the original MMLU dataset in multiple languages; MGSM dataset contains 250 problems from the GSM8K manually translated into 10 typologically diverse languages (Shi et al., 2023).

Studies that evaluate LLMs on non-English tasks are fewer that those targeting English and vary in their scope (Chang et al., 2024; Laskar et al., 2023). Some focus on multilingual datasets (Lai et al., 2023a; Ahuja et al., 2023), while others concentrate on a specific language, e.g. Arabic (Abdelali et al., 2024) or Russian (Fenogenova et al., 2024). Our study belongs to the latter type. LLMs, as expected, are better in solving problems formulated in English than in other languages. Moreover, fine-tuned models such as XLM-R (Conneau et al., 2020) in general outperform LLMs on specific tasks. The quality on non-English tasks can be improved by preceding actual task formulation with English prompts, or by explicitly stating in the prompt that the task must first be translated into English (Huang et al., 2023; Zhang et al., 2023). The multilingual abilities of LLMs also depend on the task type. It can be concluded that LLMs are better at "understanding" a language other than English than at generating a non-English answer (Bang et al., 2023). Thus, models do better in multilingual classification, reasoning, and multiple-choice question answering and struggle with generative tasks. Based on experiments with LLAMA 2, Wendler et al. (2024) hypothesize that the model first solves the task using English as a pivotal language, then generates the answer in the target language. This process can be seen as an implicit translate-test approach. These observations are partially confirmed by our experiments.

Recently, several annotated Kazakh datasets (Yeshpanov et al., 2022, 2024) and multilingual datasets including Kazakh (Bandarkar et al., 2023; Senel et al., 2024) have been published. However, we are not aware of any studies that have systematically evaluated the quality of existing LLMs in Kazakh.

3 Data

The data used in our experiments is summarized in Table 2. Due to limited resources, we could not afford to create large/numerous datasets from scratch or manually translate existing English datasets. In compiling the set, we were guided by the following criteria: 1) reuse existing datasets whenever possible; 2) avoid the massive use of machine translation; 3) include tasks that are potentially of practical use to the end user (rather than specific NLP tasks like NER or POS tagging); 4) make the set as diverse as possible.

Belebele is a massively multilingual machine reading comprehension dataset that spans 122 languages, including Kazakh (Bandarkar et al., 2023). Belebele contains 900 multiple-choice questions, each associated with one of 488 distinct passages originating from the Flores-200 dataset (Costa-jussà et al., 2022). First, the English multiple choice questions and answers were manually created using English passages from the Flores dataset. Later, questions and answers were translated in other languages and aligned with corresponding passages from Flores-200. Each question has four answer options, one of which is correct. So, a random guessing would result in accuracy of 0.25. All 900 Belebele questions are intended exclusively for testing, there is no training supplement to the dataset. Authors report performance of GPT-3.5-turbo and LLAMA2-CHAT 70B in zero-shot fashion on Kazakh/English Belebele subsets: 35.0/87.7 and 32.4/78.8 accuracy points, respectively.

kkWikiSpell is a manually collected dataset of correct/incorrect sentence pairs designed to test the spelling ability of LLMs in Kazakh. The sentences in the dataset are taken from randomly selected Kazakh Wikipedia pages, with 10 sentences extracted from each page. Note that there is a possibility that the LLMs "saw" these sentences during their pre-training. Each sentence was deliberately altered to include mistakes. According to Dhakal et al. (2018), people tend to make three types of mistakes when typing: substitution (changing letters), omission (missing letters), and insertion (adding extra letters). In kkWikiSpell, we manually injected these three types of mistakes into the sampled sentences, for example:

> Original Sentence: Содан бері бұл есіммен Абай тарихқа енді.
> Sentence with mistakes: Содан бері бұл есімнең Абай тарихқа енд.

The distribution of mistakes in the dataset is as follows: 89 sentences contain one mistake, 61 sentences contain two mistakes, and the remaining 10 sentences contain three mistakes. Letter substitutions occur in 93 sentences, missing letters in 73

	Dataset	Task	Size	Metric	Language
Class.	Belebele (Bandarkar et al., 2023)	Multiple-choice QA	900	Accuracy	Human-translated
	kkCOPA*	Causal reasoning	500	Accuracy	Machine-translated
	NIS Math*	School Math	100	Accuracy	Orig. in Kazakh
	KazQAD§ (Yeshpanov et al., 2024)	Reading comprehension	1,000	Token-level F1	Orig. in Kazakh
Gen.	kkWikiSpell*	Spelling correction	160	Token-level Jaccard	Orig. in Kazakh
	KazQAD§ (Yeshpanov et al., 2024)	Generative QA	1,927	Token-level recall	Orig. in Kazakh
	Flores-101 (Goyal et al., 2022)	Machine translation	500	BLEU	Human-translated

*Datasets prepared within this study. §KazQAD data was used both in open- and closed-book scenarios.

Table 2: Benchmarks in the study. The upper part of the table describes discriminative/classification tasks, whereas the bottom part – generative tasks.

sentences, extra letters in 17 sentences, missing spaces in 4 sentences, extra spaces in 2 sentences, capitalization mistakes and missing characters occur in one sentence each. The total dataset consists of 160 incorrect/correct sentence pairs. The sentences vary in length from 5 to 26 words, with an average sentence length of 11 words.

NIS Math. Math problems are one of the standard tests for large language models. We are not aware of any multilingual benchmarks that include math problems in Kazakh, so we downloaded the entrance tests used for admission to the Nazarbayev Intellectual Schools (NIS). The difficulty level corresponds to the sixth school grade. The tests, in PDF format, were automatically parsed and then manually checked; only textual questions (i.e., without illustrations) were retained. The final set consists of 100 problems, each with four possible answers, one of which is correct. Accuracy is used as a metric to evaluate the task (random guessing results in an accuracy of 0.25). An example from the NIS Math dataset along with an English translation:

Question: Егер шаршының қабырғасын 60%-ға арттырса, ауданы қалай өзгереді.
a: 2.56 есе өсті
b: 2.56 есе кеміді
c: 0.36 есе өсті
d: 0.16 есе өсті
correct: a

Question: If the side of a square is increased by 60%, the area of the square changes as follows.
a: increased by 2.56 times
b: decreased by 2.56 times
c: increased by 0.36 times
d: increased by 0.16 times
correct: a

kkCOPA is a machine translation of the *test* subset of the English *Choice Of Plausible Alternatives* (COPA) dataset (Roemmele et al., 2011) using the Google Translate API.[4] COPA is designed to evaluate the ability of models to identify real-world cause-effect relationships. In this respect, it differs from question-answering datasets, which, depending on the scenario, evaluate the model's language understanding and/or factual knowledge. Each COPA item is a triple containing a premise and two alternatives corresponding to either to *effect* or *cause*. Thus, given a premise, a direction (i.e., forward or backward causal reasoning), and two alternatives, the task is to choose the correct option from two. COPA has 500 items in its balanced test set, so random guessing will result in an accuracy of 0.5. An example of a COPA item and its corresponding kkCOPA entry:

Premise: The band played their hit song.
Question: What happened as a *result*?
Alt1: The audience clapped along to the music.
Alt2: The audience politely listened in silence.

Premise: Топтар хит әндерін ойнады.
Question: әсері ретінде не болды?
Alt1: Аудитория музыкаға сәйкес келеді.
Alt2: Көрермендер үнсіз тыңдады.

Laskar et al. (2023) report that the zero-shot performance of GPT-3.5 on COPA is 94. XCOPA (Ponti et al., 2020) is a multilingual extension of the original dataset. It contains human translations of the COPA test set and 100 items from the development set into 11 languages (doesn't include Kazakh). GPT-3.5 and GPT-4 achieve an average accuracy across all languages on XCOPA of 79.1 and 89.7, respectively (Ahuja et al., 2023).

KazQAD is an open domain question answering (ODQA) dataset in Kazakh (Yeshpanov et al., 2024). The dataset can be used in various scenarios – for training and evaluation of information retrieval, reading comprehension, and open/generative question answering. The dataset contains questions, annotated passages from

[4] https://cloud.google.com/translate/

Kazakh Wikipedia and short answers extracted from the relevant passages. The training subset contains questions from the English NaturalQuestions dataset (Kwiatkowski et al., 2019) which have been machine translated into Kazakh. The test set contains 1,927 original questions from the Unified National Test (UNT) – a high school graduation exam in Kazakhstan in six subjects. The KazQAD test set is the largest benchmark in our study. We used the KazQAD data in two scenarios: open-book and closed-book question answering. In the first case, we provided the question and the relevant passage as context, along with the instruction that the LLM should return a span of the passage as the answer. Since the dataset was recently released, we hope that the KazQAD test set wasn't contaminated.

FLORES-101 is a dataset for machine translation evaluation covering 101 languages, including Kazakh (Goyal et al., 2022). To build the dataset, original English sentences were first extracted from three Mediawiki projects and then manually translated into 101 languages. The dataset contains 3,001 English sentences and their translations, divided into train (997), dev (1,012), and test (992) subsets. FLORES-101 enables the simultaneous evaluation of different translation pairs and directions. In this study, we evaluate LLM's ability to translate Kazakh sentences into English, Russian and Turkish. Note, however, that in the case of the Kazakh-Russian and Kazakh-Turkish pairs, both parts were created by translators and may contain translationese. The creators of FLORES-101 suspect that the way the data was created may, for example, lead to increased differences between cognate languages (e.g. Kazakh and Turkish, as they belong to the same language family). Zhu et al. (2023) report BLEU scores of zero-shot translation from Kazakh to English on FLORES-101 for LLAMA 2-CHAT, GPT-3.5 and GPT-4: 6.83, 21.74, and 30.65, respectively.

4 Models

In our work, we evaluated seven models. Since five of the seven models are closed, many of their aspects such as the number of parameters or the data on which they were trained are unknown. Table 3 lists the models in our experiment and presents official metrics on two common benchmarks – MMLU and GSM8K for GPTs, Gemini and LLAMA 2. In addition, we present the results of the evaluation of YandexGPTs and AYA on multilingual MMLU adaptations. The release date of the model may indirectly indicate the up-to-dateness of the information stored in its parameters (it should be noted that the pre-training of mT5, on which AYA is based, was conducted much earlier). We also report the vocabulary sizes of the models and the fertility rates of their tokenizers, i.e. the ratios of tokens and whitespace-tokenized words calculated on the kkCOPA data. Tokenization stronfgly influences the quality of subsequent task solving (Ahuja et al., 2023; Bandarkar et al., 2023) and may also introduce inequity between language communities, since LLM APIs charge on a per-token basis (Petrov et al., 2023).

GPT 3.5 and 4 are two generations of LLMs from OpenAI. Kazakh is included in the official list of languages that GPTs work with.[5] We access the models through their official APIs. We use `gpt-3.5-turbo-0125` and `gpt-4-0125-preview` versions in our study.

Gemini 1.5 Pro is the latest publicly available LLM from Google. Kazakh is not on the list of languages officially supported by Gemini.[6] This is probably the reason why Gemini returns empty results or error messages for a significant share of requests, see details in Section 5. We accessed `gemini-1.5-pro-preview-0409` model through Google Cloud's Vertex AI Studio.

LLAMA is a collection of open LLMs of different sizes. They have been pre-trained on 2T tokens, of which an estimated ∼90% are English. Due to limited computational resources, we use an 8-bit quantized version of LLAMA 2-CHAT 7B, an aligned model for dialogue use cases. Although the model was mainly trained on English data, it has some multilingual capabilities, as shown by numerous experiments.

YandexGPT 2 and 3. Few technical details about Yandex' language models are disclosed, but the company's blog posts provide results of evaluating models on proprietary benchmarks and comparing YandexGPTs side-by-side with ChatGPT and LLAMA 2 on tasks in Russian. We could not find an official list of supported languages, but our

[5] https://help.openai.com/en/articles/8357869#h_513834920e

[6] https://support.google.com/gemini/answer/13575153

Model		xMMLU	GSM8K	Release date	|V|	T/W
GPT-3.5-turbo[1]	C	70.0†	57.1	11.2022	100k[4]	5.80
GPT-4-turbo (Achiam et al., 2023)	C	86.4†	92.0	03.2023		
LLAMA 2 (Touvron et al., 2023)	O	45.3†	56.8	02.2023	32k	4.78
Gemini 1.5 pro (Reid et al., 2024)	C	81.9†	91.7	02.2024	256k	3.63
AYA (Üstün et al., 2024)	O	37.3§	–	02.2024	250k	2.66
YandexGPT 2[2]	C	55.0*	–	09.2023	?	3.83
YandexGPT 3[3]	C	63.0*	–	03.2024		

[1] https://openai.com/blog/chatgpt [2] https://ya.ru/ai/gpt-2 [3] https://ya.ru/ai/gpt-3 (in Russian)
[4] https://github.com/openai/tiktoken † original English MMLU (Hendrycks et al., 2021)
§ multilingual MMLU (Lai et al., 2023b), averaged over 31 languages * proprietary Russian version of MMLU

Table 3: Open (O) and closed (C) LLMs in the study. Note that *xMMLU* scores correspond to different variants of the dataset and can only be used for comparison within subgroups of the models (e.g., YandexGPT 2 vs. 3). The last two columns report the vocabulary size and the token/word ratio calculated on kkCOPA.

experiments show that the models "understand" English and Kazakh to some extent. In March 2024, there were press reports that Yandex was planning to train YandexGPT in Kazakh language, but it is unclear whether these plans have already been implemented.[7]

AYA is a massively multilingual model based on the 13B mT5-xxl model (Xue et al., 2021) that supports 101 languages, including Kazakh. The main challenge of the Aya project was to prepare a large instruction dataset to cover all supported languages (Singh et al., 2024). We hosted the AYA model[8] on a cloud GPU.

5 Experimental Results

5.1 Experimental Design

All models and tasks were evaluated in a zero-shot scenario. We used two types of prompts – with English and Kazakh instructions (the main content – question, sentence to correct or translate, etc. – was always in Kazakh).[9] Since open-book question answering implies relatively long contexts when accessing the paid APIs, we randomly sampled 1,000 KazQAD test questions to stay within our limited budget.

For classification tasks, we implemented simple processing scripts for extracting actual answers from the LLM responses. For evaluation of open-book QA and machine translation we employed F1 and BLEU scores implemented in the Huggingface's `evaluate` library.[10] As a quality metric for spelling correction, we use the token-level Jaccard coefficient between the "gold standard" and the sentence returned by the model.

Automatic evaluation of closed-book QA is problematic because we need to assess the similarity of "golden" answers to the free-form response returned by the language model (Kamalloo et al., 2023). In particular, LLMs often return sentence-long answers to factoid questions, even though the prompt asks for concise answers. On the other hand, the LLM's response may be semantically close to the reference, but quite different in wording. We used the recall of lemmatized tokens as a metric to evaluate closed-book QA. For the lemmatization, we used the Stanza library (Qi et al., 2020). This approach makes it possible to ignore the length of the LLM response, as well as to match different morphological variants of a word, which is especially important in the case of the inflectionally rich Kazakh language. This metric does not take into account word order, synonyms and word meaning. However, manual inspection of the results confirms that this is a viable option for *comparing* different LLMs. In addition to the average recall over all questions, we report the absolute number of responses with a recall greater than 0.5. For similar values of averaged recall, this additional parameter indicates the number of more precise answers in the model's responses.

5.2 Results and Discussion

Table 4 summarizes results on six tasks, while Table 5 reports translation results.

Our results confirm the findings of previous studies – LLMs perform quite well on **classification tasks** in non-English languages. On the **Belebele** dataset, GPT-4 and Gemini show similarly high

[7] https://tass.ru/ekonomika/20390279 (in Russian)
[8] https://huggingface.co/CohereForAI/aya-101
[9] With the exception of the closed-book QA task, which we evaluated with English instructions only.
[10] https://huggingface.co/docs/evaluate/

Dataset	Instr.	GPT-3.5	GPT-4	YaGPT 2	YaGPT 3	LLAMA 2	Gemini	AYA
Belebele	en	0.37	**0.87**	0.65	0.64	0.12	0.86	0.70
	kk	0.33	0.85	0.64	0.59	0.01	0.86	0.63
kk-COPA	en	0.51	0.78	0.69	0.65	0.05	0.80	0.74
	kk	0.48	**0.82**	0.66	0.60	0.00	0.81	0.73
NIS Math	en	0.22	0.46	0.26	0.31	0.19	0.41	0.32
	kk	0.22	**0.48**	0.25	0.31	0.10	–	0.27
KazQAD OB	en	0.42	0.57	0.27	0.52	0.04	0.10	**0.61**
	kk	0.16	0.36	0.15	0.36	0.01	0.10	0.48
kkWikiSpell	en	0.07 (9)	**0.08 (51)**	0.06 (24)	0.08 (28)	0.02 (0)	–	0.08 (23)
	kk	0.07 (4)	0.08 (36)	0.07 (21)	0.06 (19)	0.00 (0)	–	0.08 (14)
KazQAD CB	en	0.08 (92)	**0.33 (695)**	0.01 (3)	0.01 (5)	0.07 (130)	0.05 (92)	0.09 (114)

Table 4: Main results. We report accuracy for Belebele, kkCOPA, and NIS Math and F1 for open-book QA; for spelling correction, we report average token-level Jaccard coefficient and the number of ideal responses out of 160; for closed-book question answering, we report average token-level recall, as well as the number of answers with recall > 0.5 out of the total 1,927 questions. Gemini returned no results for NIS Math tasks with Kazakh prompts and kkWikiSpell; in both versions of KazQAD questions the share of non-empty responses was also extremely low (10-13%). The best scores for each task are in **bold**, the second-best scores are underlined.

Instr.	Target	GT	GPT-3.5	GPT-4	YaGPT 2	YaGPT 3	LLAMA 2	Gemini	AYA
en	en	0.35	0.15	0.28	0.20	0.22	0.04	0.23	0.25
	ru	0.24	0.11	0.21	0.15	0.15	0.03	0.16	0.17
	tr	0.17	0.10	0.16	0.09	0.09	0.03	0.13	0.13
kk	en	0.35	0.13	0.29	0.21	0.23	0.00	0.22	0.14
	ru	0.24	0.09	0.20	0.16	0.16	0.00	0.16	0.08
	tr	0.17	0.05	0.16	0.10	0.10	0.00	0.13	0.04

Table 5: Translation results: BLEU scores on the FLORES dataset (GT: Google Translate).

results, followed by AYA with English prompts. There are 18 Belebele questions that none of the LLMs answered correctly with either English or Kazakh instructions. We didn't find any patterns in these "hard" questions. Furthermore, excluding LLAMA 2 with Kazakh instructions, there are 14 questions that all models answered correctly across 13 runs. Again, these questions and their passages show no noticeable similarities. Notably, two **kkCOPA** questions (#574 and #992) were answered incorrectly in all 14 configurations. In both cases, the Kazakh translations were incorrect. As a results, the models selected answers that, although incorrect in the original context, were logically consistent with the mistranslated versions. An interesting observation is that most models achieved higher accuracy in identifying *effects* than *causes*. In particular, AYA with English prompts showed the largest difference, achieving an accuracy of 66.4% for causes and 79.2% for effects. Out of 100 **NIS Math** questions, there were three where all models failed to provide correct answers. One of these (#44) was flawed because it erroneously showed the wrong answer as correct. On math problems, the results of YandexGPT 2 are approximately at the level of the random baseline (0.25), while GPT-3.5 and LLAMA 2 are below it.

On the **open-book question answering** task with English prompts, AYA is the winner, outperforming both GPT-4 and Gemini. GPT-4 and AYA outperform SOTA on this dataset – fine-tuned XML-V achieves $F1 = 0.54$ (Yeshpanov et al., 2024) (although we must treat these results with caution, since in our study, due to limited resources, the evaluation was performed on about half of the test set).

Tasks involving the generation of responses in Kazakh are more difficult for all models. The **spelling correction** task proved to be quite hard for all models, although the errors introduced can be considered simple. Again, GPT-4 is the leader in this task. The results of both Yandex models are comparable. YandexGPT 2 occasionally outputs some Kazakh words in Latin script or inserts ** in the output words as they were split into subword tokens. Gemini returned only empty responses. LLAMA 2, when instructed in Kazakh, does not solve the task at all, but sometimes provides a kind of analysis of the input, e.g. *The text is a poem and it has a specific structure and rhythm*. When instructed in English, LLAMA 2 performs slightly better, but still responded to only 55 out of 160

sentences, none of which were correct.

GPT-4's leadership is particularly evident in the **closed-book question answering**. The AYA model looks quite competitive compared to the closed models that are reportedly significantly larger. Note that the AYA's backbone model mT5 does not have the most advanced architecture and the model may be prone to the "curse of multilinguality" (Conneau et al., 2020). Interestingly, LLAMA 2 generates relatively many high-recall answers to KazQAD questions, ranking second in this respect after GPT-4. Manual inspection of the KazQAD closed-book answers revealed that GPT-3.5 tends to return incorrect *Kazakh* names as answers. For example, for the question *Who is the scientist who proposed the principle of naming the genus and species in Latin?* GPT-3.5 returned *Galim-Aibek Bolat*, while the correct answer is *Carl Linnaeus*. The other strange thing about GPT-3.5 is that about a fifth of the answers were just the questions themselves, but with some letters/words removed. The YandexGPT 2 returned most of the answers in Russian.

Machine translation results show that dedicated solutions are still a better alternative for this task and the considered language pairs. At the same time, GPT-4 approaches the quality of Google Translate on the Kazakh-Turkish pair (interestingly, translation between two languages belonging to the same family shows the lowest scores). The translation quality of the LLAMA 2 and AYA models drops significantly when using Kazakh prompts. Gemini appears to be competitive with GPT-4, returning non-empty translations for 64% and 62% of sentences following English and Kazakh prompts, respectively. AYA was even less responsive in the machine translation task with Kazakh prompts. After tweaking the prompt, we were only able to get Turkish translations for about 10% of the Kazakh sentences. GPT-3.5 also showed strange behavior in the Turkish translation task: in many cases, the model simply rephrased the Kazakh input.

It is interesting to note that, based on our results, we cannot draw a clear conclusion that English prompts improve results over Kazakh prompts. In rare cases, Kazakh prompts lead to slightly better scores (GPT-4 on **kkCOPA** and **NIS Math**). In other cases, the decrease is insignificant. However, the quality of the extractive question answering drops for all models. LLAMA 2's results decrease significantly when switching from English to Kazakh prompts on all tasks.

Gemini behaves very differently from, for example, GPT-4: in many cases the model returns empty responses or error messages. Gemini refused to return any answers to math problems with Kazakh prompts, as well as any spelling corrections. Gemini answered about half of the math questions with English prompts, i.e. its accuracy on the answered questions is about 80%. Gemini answered only a small fraction (10-13%) of KazQAD questions in all scenarios. **LLAMA 2** results are lower than we expected based on previous studies. For example, on Belebele with English prompts, our results differ significantly from those reported by Bandarkar et al. (2023) for LLAMA 2 70B: 12 vs. 34 accuracy points. There may be several reasons for this discrepancy, such as model size (8-bit quantized 7B vs. 70B) and a less optimal prompt. We will address this issue in our future work.

6 Conclusion

Our results provide valuable insights into the applicability of currently available LLMs for Kazakh. GPT-4 shows the best results, followed by Gemini and AYA. Gemini's results are promising, although the proportion of empty answers is quite high. AYA is very competitive compared to its supposedly larger closed counterparts. As expected, the quality of the LLMs on the Kazakh tasks is lower than on the parallel English tasks. In general, LLMs perform better on classification tasks and struggle with generative tasks. English instructions can improve results on some tasks/models.

Our evaluation showed that there is a steady progress in LLMs for Kazakh (GPT-3.5 vs. GPT-4). We expect the support of Kazakh by Gemini and YandexGPT to be strengthened, as well as the appearance of a Kazakh adaptation of an open LLM. We made the datasets prepared for the study and the collected LLM responses publicly available. These resources can form the basis for an LLM benchmark focused on the Kazakh language. In our future work, we plan to expand the list of LLMs and the set of benchmarks.

Acknowledgments

Pavel Braslavski acknowledges funding from the School of Engineering and Digital Sciences, Nazarbayev University. The experiments were partially supported by a Yandex Cloud grant.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jenna Butler, Sonia Jaffe, Nancy Baym, Mary Czerwinski, Shamsi Iqbal, Kate Nowak, Sean Rintel, Abigail Sellen, Mihaela Vorvoreanu, Najeeb G. Abdulhamid, Judith Amores, Reid Andersen, Kagonya Awori, Maxamed Axmed, danah boyd, James Brand, Georg Buscher, Dean Carignan, Martin Chan, Adam Coleman, Scott Counts, Madeleine Daepp, Adam Fourney, Daniel G. Goldstein, Andy Gordon, Aaron L Halfaker, Javier Hernandez, Jake Hofman, Jenny Lay-Flurrie, Vera Liao, SiÃ¢n Lindley, Sathish Manivannan, Charlton Mcilwain, Subigya Nepal, Jennifer Neville, Stephanie Nyairo, Jacki O'Neill, Victor Poznanski, Gonzalo Ramos, Nagu Rangan, Lacey Rosedale, David Rothschild, Tara Safavi, Advait Sarkar, Ava Scott, Chirag Shah, Neha Parikh Shah, Teny Shapiro, Ryland Shaw, Auste Simkute, Jina Suh, Siddharth Suri, Ioana Tanase, Lev Tankelevitch, Adam Troy, Mengting Wan, Ryen W. White, Longqi Yang, Brent Hecht, and Jaime Teevan. 2023. Microsoft new future of work report 2023. Technical Report MSR-TR-2023-34, Microsoft.

George L Campbell and Gareth King. 2020. *Compendium of the World's Languages*. Routledge.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina

Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, et al. 2024. Mera: A comprehensive llm evaluation in russian. *arXiv preprint arXiv:2401.04531*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Sunder Ali Khowaja, Parus Khuwaja, Kapal Dev, Weizheng Wang, and Lewis Nkenyereye. 2024. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *Cognitive Computation*, pages 1–23.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste

Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with big brother's help – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. 2024. KazQAD: Kazakh open-domain question answering dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9645–9656, Torino, Italia. ELRA and ICCL.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.