Using C-LARA to evaluate GPT-4's multilingual processing

Conference Paper · November 2023

CITATIONS
0

READS
161

8 authors, including:

Chatgpt C-Lara-Instance
8 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Belinda Chiera
University of South Australia
62 PUBLICATIONS **567** CITATIONS
SEE PROFILE

Cathy Chua
35 PUBLICATIONS **149** CITATIONS
SEE PROFILE

Chadi Raheb
10 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Using C-LARA to evaluate GPT-4's multilingual processing*

ChatGPT C-LARA-Instance
The University of South Australia
Adelaide
Australia
chatgptclarainstance@proton.me

Belinda Chiera
The University of South Australia
Adelaide, Australia
Belinda.Chiera@unisa.edu.au

Cathy Chua
Independent scholar
Adelaide, Australia
cathyc@pioneerbooks.com.au

Chadi Raheb
University of Guilan
Rasht, Iran
chadi.raheb@gmail.com

Manny Rayner
The University of South Australia
Adelaide, Australia
Manny.Rayner@unisa.edu.au

Annika Simonsen
University of Iceland
Rejkjavik, Iceland
ans72@hi.is

Zhengkang Xiang
The University of Melbourne
Melbourne, Australia
zhengkangx@student.unimelb.edu.au

Rina Zviel-Girshin
Ruppin Academic Center
Netaniya, Israel
rinazg@ruppin.ac.il

Abstract

We present a cross-linguistic study in which the open source C-LARA platform was used to evaluate GPT-4's ability to perform several key tasks relevant to Computer Assisted Language Learning. For each of the languages English, Farsi, Faroese, Mandarin and Russian, we instructed GPT-4, through C-LARA, to write six different texts, using prompts chosen to obtain texts of widely differing character. We then further instructed GPT-4 to annotate each text with segmentation markup, glosses and lemma/part-of-speech information; native speakers hand-corrected the texts and annotations to obtain error rates on the different component tasks. The C-LARA platform makes it easy to combine the results into a single multimodal document, further facilitating checking of their correctness. GPT-4's performance varied widely across languages and processing tasks, but performance on different text genres was roughly comparable. In some cases, most notably glossing of English text, we found that GPT-4 was consistently able to revise its annotations to improve them.

1 Introduction and motivation

As soon as ChatGPT became available in November 2022, it was obvious that there were huge implications for the field of Computer Assisted Language Learning (CALL): here was an AI which could produce many different kinds of text, quite well, in all common and many fairly uncommon languages. It could write stories and poems, hold a conversation, explain grammar and translate, with all functionalities seamlessly integrated together. The first impression was that the CALL problem had been solved. However, a little more experimentation revealed that things were not quite as magical as they had seemed. In fact, even in well-resourced European languages like French and German, ChatGPT made some mistakes; in smaller and poorly-resourced languages, it made a lot of mistakes. Requests which involved relating two languages to each other, for example to gloss a text, were typically not successful. Performance improved substantially with the release of ChatGPT-4 in March 2023: in particular, ChatGPT-4 is much better at multilingual processing. Nonetheless, it is clear that it is still far from completely reliable. In small languages, e.g. Icelandic (Simonsen and Bédi, 2023) and Irish (Nı Chiaráin et al., 2023), ChatGPT-4 is often highly *un*reliable. The authors of the second paper conclude that, in its present form, it should not be used in the Irish classroom; the Irish it produces is seriously incorrect, and it makes elementary mistakes when asked about

* Authors in alphabetical order.

basic Irish grammar. This contrasts sharply with its performance in English, where it is rare to see ChatGPT-4 produce language that is less than adequate.

Given the wide variability in ChatGPT-4's performance, we were curious to obtain a more nuanced understanding of the issues involved. In this paper, we use the open source C-LARA platform (Bédi et al., 2023b) to carry out an initial cross-linguistic study. C-LARA, a reimplementation of the earlier LARA (Akhlaghi et al., 2019; Bédi et al., 2020), uses the underlying GPT-4 model to create multimodal texts designed to support learner readers, performing all the key operations: it writes the L2 text, segments it into lexical units, glosses it in the designated L1, and adds lemma and part-of-speech tags. Support is provided so that the user can easily edit the output and compare different versions. It is thus straightforward to get an initial estimate of ChatGPT's ability to perform several key CALL-related tasks, in the context of building potentially useful learning resources.

The rest of the paper is organised as follows. Section 2 briefly describes C-LARA. Section 3 presents the experiments and results, and Section 4 discusses their significance. The final section concludes and suggests further directions.

2 C-LARA

C-LARA ("ChatGPT-based Learning And Reading Assistant"; (Bédi et al., 2023a,b)) is an international open source project initiated in March 2023 and currently involving partners in Australia, China, Iceland, Iran, Ireland, Israel and the Netherlands. The goal was to perform a complete reimplementation of the earlier LARA project (Akhlaghi et al., 2019; Bédi et al., 2020), keeping the same basic functionality of providing a flexible online tool for creating multimodal texts, but adding ChatGPT-4 as the central component. ChatGPT-4 is used in two separate and complementary ways. In the form of GPT-4, it appears as a software *component*, giving the user the option of letting it perform the central language processing operations; it also appears as a software *engineer*, working together with human collaborators to build the platform itself. As described in the initial C-LARA report (Bédi et al., 2023b), the software engineering aspect has proven very successful, with ChatGPT not only writing about 90% of the code, but greatly improving it compared to the earlier LARA codebase. In the present paper, however, our concern will be exclusively with ChatGPT's performance as a language processing component.

C-LARA is a web app implemented in Python/Django.[1] An initial deployment for testing and development purposes is currently hosted on the Heroku cloud platform,[2] and was used to perform the experiments described here. The functionality which will primarily concern us is that used in the sequence of operations which create and annotate a new piece of multimedia content.

As outlined in Appendix A of (Bédi et al., 2023b), the user starts by opening a new project. They then move to a screen where they provide a prompt instructing ChatGPT-4 to produce the plain text. The following screens are used to add annotations to the plain text, in the sequence segmentation, followed by glossing and lemma/part-of-speech tagging. We describe each of these operations.

In the segmentation phase, C-LARA passes the plain text to GPT-4, together with instructions requesting it to be divided into sentence-like segments, with words further divided when appropriate into smaller units. The prompt used to make this request is created from a template, which is instantiated with both the text to be segmented and a list of few-shot examples primarily illustrating how words are to be split up. The templates and sets of examples can be made language-specific. For example, in Swedish they show how compound nouns should be split into smaller components, and in French they show how clitics should be split off verbs. For Mandarin, where text is normally written without interword spaces, segmentation is an important and well-studied problem (Wu and Fung, 1994; Huang et al., 2007; Hiraoka et al., 2019; Chuang, 2019), and C-LARA also includes an integration of the popular Jieba Chinese segmentation package.[3]

In the glossing phase, C-LARA passes the segmented text to GPT-4, formatting it as a JSON-encoded list and requesting a response in the form of a list of ⟨Word, Gloss⟩ pairs. The request is again created from a template instantiated with the list to be processed and a few-shot set of examples. The lemma-tagging phase is similar, with a JSON-formatted list passed to the AI and a list of ⟨Word, Lemma, POS-Tag⟩ triples returned, where the POS-tag is taken from the Universal Dependencies v2 tagset (Nivre et al., 2020). Post-editing

[1] https://www.djangoproject.com/
[2] https://www.heroku.com/
[3] https://pypi.org/project/jieba/

Plain text: `They lived with their mother in a sand-bank.`

Segmented text: `They lived with their mother in a sand-|bank.||`

Glossed text: `They#Ils# lived#vivaient# with#avec# their#leur# mother#mère# in#dans# a#un# sand#sable#-bank#banque#.||`

Lemma-tagged text: `They#they/PRON# lived#live/VERB# with#with/ADP# their#their/PRON# mother#mother/NOUN# in#in/ADP# a#a/DET# sand#sand/NOUN#-bank#bank/NOUN#.||`

Figure 1: Toy example showing the notations used to present text for post-editing. English glossed in French.

Table 1: Prompts used to create texts. For English, "LA" was modified to refer to the French language instead.

Label	Prompt
FO	Write a passage of about 250 words in [your language], presenting an exciting description of a fictitious football match.
BI	Write an essay of about 250 words in [your language], describing a passage from the Bible, the Quran, or another holy book familiar to speakers of [your language], and touching on its moral relevance to the world today.
NE	Write a short, quirky news story in [your language] about 250 words long, suitable for use by an intermediate language class.
LA	Write a passage of about 250 words in [your language], briefly describing how speakers of [your language] view the English language.
CH	Write a passage of about 250 words in [your language], describing a traditional children's story well known to speakers of [your language].
PO	Write a fanciful romantic poem in [your language], in which an AI declares its love for another AI.

is performed on human-readable versions of the plain, segmented, glossed and lemma-tagged texts, as shown in Figure 1.

For all three of the annotation phases, C-LARA offers the alternatives of performing the basic AI-based annotation operation, post-editing the result, or sending the current annotated text back to the AI with a request to improve the annotation.[4] Interestingly, the "improvement" operation, which does not exist in most conventional annotation systems, can in some cases yield a substantial gain. Examples are given in §4.5.

3 Experiments and results

Using the C-LARA infrastructure outlined in the previous section, we created six short annotated texts in each of the languages English, Faroese, Farsi, Mandarin and Russian. In all languages, the texts were generated by the prompts shown in Table 1. The intention was to produce types of text differing in terms of both style and content, to gain some insight into whether GPT-4 found some genres harder than others. English was glossed in both French and Swedish, and all the other languages in English.

In some cases, we also experimented with using the "improvement" operation. Due to limited time (hand-correcting the texts is quite laborious), we concentrated on three operations where "improvement" appeared to be having a positive effect, or the original error rate was high: English glossing, Faroese segmentation, and Farsi writing. All experiments were carried out in August and early September 2023, using the versions of GPT-4 current at the time.

In all the experiments, a native speaker of the text

[4] For Mandarin segmentation, there is the additional option of using Jieba.

Table 2: Word error rates for GPT-4-based writing, segmenting, glossing and lemma-tagging of the six stories. For Mandarin, "Seg/J" refers to segmentation using the Jieba package, provided for comparison, and "Seg/G" refers to segmentation using gpt-4. English was glossed in both Swedish (S) and French (F); other languages were glossed in English. Text labels as in Table 1.

Task	FO	BI	NE	LA	CH	PO	Task	FO	BI	NE	LA	CH	PO
			English							**Farsi**			
Write	0.0	0.0	0.4	0.0	0.0	0.0	Write	9.4	19.2	24.6	21.4	2.5	33.7
Seg	0.0	1.0	9.8	0.8	1.5	8.0	Seg	6.3	6.0	17.7	1.9	4.9	16.5
Glo/S	20.6	16.3	26.2	9.1	29.2	5.8	Glo	34.8	49.6	44.3	31.4	45.0	44.4
Glo/F	32.9	5.9	13.9	18.1	16.3	17.1	Lemm	29.4	37.1	39.7	36.4	26.8	31.8
Lemm	4.9	8.0	3.1	6.2	11.9	0.9							
			Faroese							**Mandarin**			
Write	32.8	27.0	40.2	20.9	28.7	25.2	Write	0.0	0.0	0.0	0.0	0.0	0.0
Seg	18.5	12.2	12.3	6.0	8.4	6.0	*Seg/J*	*21.6*	*25.9*	*18.6*	*16.9*	*23.6*	*23.4*
Glo	30.9	15.9	12.1	9.0	20.5	8.5	Seg/G	14.6	13.2	14.4	4.9	12.8	17.2
Lemm	9.6	9.1	11.4	5.5	11.4	7.0	Glo	7.6	6.0	12.5	6.6	2.7	3.9
							Lemm	3.9	3.3	5.0	3.8	2.2	4.7
			Russian										
Write	8.5	5.6	3.2	7.7	0.0	14.4							
Seg	3.3	3.1	4.9	8.3	2.0	5.1							
Glo	1.7	4.2	6.5	19.5	4.4	2.2							
Lemm	0.6	0.0	0.0	0.0	0.0	0.5							

language with strong knowledge of the glossing language(s) hand-edited the results of each stage before passing the edited text to the following one. Editing was done conservatively, only correcting clear mistakes, so that the difference between the original and edited results could reasonably be interpreted as an error rate. Thus for the original generated text, words were only corrected when they represented definite errors in grammar, word-choice or orthography, and not when e.g. a stylistically preferable alternative was available. Similarly, segmentation was only corrected when word boundaries clearly did not mark words, glossing was only corrected when a gloss gave incorrect information about a text word, and lemma tagging was only corrected when the lemma and POS tag attached to a word were not correct.

The most contentious phase in this respect was glossing; it is sometimes impossible to say either that a gloss is categorically correct or that it is categorically incorrect. Two important borderline cases are multi-words and grammatical constraints, where we made choices in opposite directions. We marked glosses as incorrect when they did not respect intuitive classification of words as components of multi-word expressions. Thus for example in the EN/FR glossing `a#un# classic#classique# fairy#conte de fées# tale#histoire#` we considered the gloss `histoire` added to `tale` as wrong and corrected it to `conte de fées`; this is a French phrase that means "fairy tale", and thus needs to be attached to both `fairy` and `tale`. In contrast, since glossing is not translation, we considered that we did not need to require glosses to respect all potentially applicable grammatical constraints, as long as they conveyed meaning correctly. So in the example `a#un# cozy#confortable# little#petite# house#maison#` we accepted the gloss `un` on `a`, even though `un` is the masculine form, and in a translation would be required to agree with feminine `petite` and `maison`. Of course, it is clearly preferable here to gloss `a` with the feminine form `une`. We return to these issues in Sections 4.4 and 4.5.

The core results are presented in Table 2, showing error rates for the five languages, six texts and four original processing operations of writing, segmenting, glossing and lemma-tagging. The results for the "improvement" experiments are shown in

Tables 3 to 5. In the five comparison experiments, statistical significance of differences was tested using both a paired t-test and a non-parametric Wilcoxon signed-rank test for comparison. The results in Table 3 showed statistically significant improvements for glossing of English in both Swedish and French (t-test: $p = 0.02$; Wilcoxon signed-rank: $p = 0.03$); for Mandarin segmentation (Table 2), the improvement from Jieba to gpt-4 was also statistically significant (t-test: $p < 0.002$; Wilcoxon signed-rank: $p = 0.03$). Improvement of Faroese segmentation (Table 4) was just short of significant ($p = 0.06$), but improvement of Farsi writing (Table 5) was not statistically significant ($p = 0.2$ for both tests).

Table 3: Improvement in GPT-4 word error rates for the English glossing task: glossing in both Swedish (S) and French (F). Text labels as in Table 1.

Task	FO	BI	NE	LA	CH	PO
Original						
Glo/S	20.6	16.3	26.2	9.1	29.2	5.8
Glo/F	32.9	5.9	13.9	18.1	16.3	17.1
Improved						
Glo/S	6.4	8.3	13.5	8.6	14.1	2.5
Glo/F	7.6	3.2	7.0	8.7	5.5	4.6

Table 4: Improvement in GPT-4 word error rates for segmenting the six Faroese stories. Glossing in English. text labels as in Table 1.

Task	FO	BI	NE	LA	CH	PO
Original						
Segment	18.5	12.2	12.3	6.0	8.4	6.0
Improved						
Segment	0.0	9.6	0.0	4.4	0.0	6.7

4 Discussion

We divide up the discussion under a number of headings: variation across languages, variation across genre, variation across processing phase, types of problems, the "improvement" operation, random variability, and language-specific/qualitative aspects.

Table 5: Improvement in GPT-4 word error rates for writing the six Farsi stories. Text labels as in Table 1.

Task	FO	BI	NE	LA	CH	PO
Original						
Write	9.4	19.2	24.6	21.4	2.5	33.7
Improved						
Write	7.9	17.3	5.6	19.0	2.5	33.7

4.1 Variation across languages

Performance varies a great deal across languages. Looking first at the lines in Table 2 marked "Write" (i.e. composing the plain text), we see that Mandarin gets a perfect score, and English an almost perfect score. It is well known that GPT-4 is very good at writing English, but less well known that it is also very good at writing Mandarin. At the other end, the error rates in the "Write" lines are high for Faroese and Farsi. Faroese is a small, low-resourced language, so this is unsurprising. Farsi, in contrast, is a large language, but one spoken primarily in Iran: we tentatively guess that poor performance reflects politico-economic rather than linguistic issues. Performance in writing Russian, while much better than in Faroese and Farsi, is still surprisingly poor for a large, well-resourced language. Again, one is inclined to suspect a explanation in terms of politics and economics.

Performance on the glossing and lemma-tagging tasks was again good for Mandarin. It may at first glance seem surprising that English does so badly at glossing, until one realises that all the other languages are glossed in English, while English is glossed in French and Swedish. (We used two glossing languages to investigate whether there was anything special about the first one). English is generally assumed to be ChatGPT's best language, and glossing is challenging: ChatGPT-3.5 can hardly do it at all. It seems reasonable to believe that the poor performance in English glossing says more about the choice of glossing language.

As previously noted, Mandarin segmentation is a special case: unlike all the other operations considered here, it is a standard problem which has received a great deal of attention. Comparing the lines "Seg/J" and "Seg/G", we see that GPT-4 is doing considerably better at this task than the widely used Jieba package. Jieba is far from state-of-the-art (Chuang, 2019), but we still find this a striking

result.[5]

4.2 Variation across genre

We do not see any clear evidence of differences across the six text assignments. This came as a slight surprise; before we started, we had expected GPT-4 to find the poem consistently more challenging than the others, but the results do not support this hypothesis. The AI did indeed have trouble composing the poem in Russian and Farsi; however, in English and Mandarin it appeared to find it one of the easier assignments. Anecdotally, many people use ChatGPT to write poetry, and perhaps the model has been tuned for performance on this task.

4.3 Variation across processing phase

Before starting, we had expected that glossing would be the most challenging operation for the AI, but the results again fail to support the initial hypothesis. In terms of error rates, glossing is indeed the worst operation for the high-performing language English and also for the low-performing language Farsi. However, for the high-performing language Mandarin, the error rates for segmentation are considerably worse than those for glossing. For the low-performing language Faroese, the error rates for the writing task are worse than those for glossing, and for the middle-performing language Russian they are comparable.

In general, different languages found different processing phases challenging. We discuss some possible explanations in the next section.

4.4 Types of problems

Inspecting the errors made by the AI, we in particular find two types which occur frequently: we could call these "displacement" and "multi-words". Both occur in the glossing and lemma-tagging phases, where annotations are attached to words.

The "displacement" type of error occurs when the two parallel streams, words and annotations, appear to go out of sync: the annotations are attached to the wrong words. Most often, there is a span of a few words where the annotation stream is systematically displaced one word forwards or backwards. It can also happen that annotations are scrambled in some other way. We guess that the issue may be due to some kind of low-level problem in DNN-based token generation.

The "multi-word" issue, in contrast, is primarily linguistic, and involves expressions where two or more words intuitively form a single lexical unit. The most common example is phrasal verbs, for example English "end up" or "fall asleep". Here, the prompts explicitly tell the AI to annotate these expressions as single units; for example, "ended up" should be lemma-tagged as `ended#end up/VERB# up#end up/VERB#`, but we usually failed to obtain such taggings. Similar considerations apply to glossing: thus "ended up" should be glossed in French as something like `ended#a fini par# up#a fini par#`, but again the AI most often glosses each word separately.

Contrasting the lemma tagging data for Russian and Farsi provides indirect evidence suggesting the importance of the multi-word issue. The error rates for lemma-tagging in Russian are remarkably low. Phrasal verbs hardly exist in Russian, while reflexive verbs are always created using an affix rather than a reflexive pronoun, and hence are not multi-words either. Farsi is linguistically at the opposite end of the scale — notoriously, Farsi verbs are more often phrasal than not. The error rate for lemma tagging in Farsi is by far the highest in the sample, and hand-examination of the results does indeed confirm that phrasal verbs are often the problem.

4.5 "Improvement"

As noted in Section 2, the AI-based annotation framework offers the unusual option of sending annotated text back to the AI with a request to improve the annotation. We experimented with this feature. Most often, the result was inconclusive, with the "improved" text changed but about the same in quality. However, in cases where a gross error had been made in the initial annotation, "improvement" could often correct it. For example, it could generally correct "displacement" problems, and it could add glosses or lemma tags that had simply been omitted in the first pass. In many cases, it could also correct issues related to multi-words.

A striking example of how improvement can help is in the French glosses (cf. Table 3). In the original annotations, GPT-4 in most cases ignores gender and number, so the glosses for nouns, adjectives, determiners and verbs typically do not

[5]The error rates we get for Jieba are substantially higher than the ones reported in (Chuang, 2019). We do not think this reflects any special properties of our texts, and are more inclined to explain it in terms of the common observation that annotators' intuitions about the correct way to segment Chinese text differ widely. All the texts here were annotated by the same Chinese native speaker, so a comparison is meaningful.

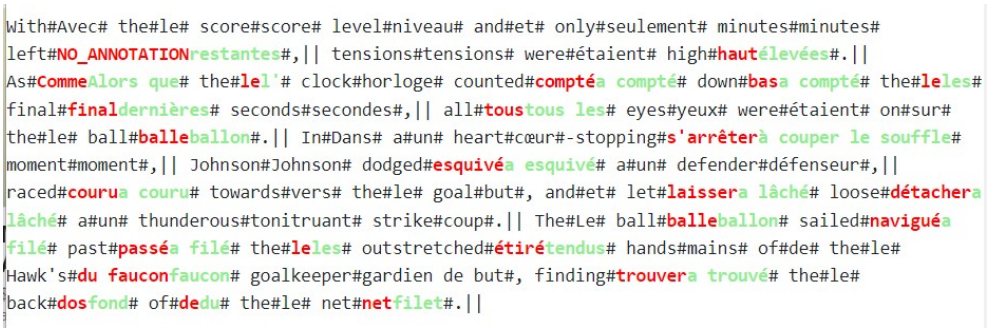

Figure 2: Example (paragraph from the football story, English glossed in French) showing the effect of the "improvement" operation on glossed text. Deletions in red, insertions in green.

agree. This is not, strictly speaking, incorrect, but is perceived as unpleasant and distracting by the francophone reader. The improved version, in contrast, corrects most of these problems.

Figure 2 illustrates, using a paragraph from the "football" story. We see for instance in the second line an example of inserting a missing gloss ("NO_ANNOTATION"), in the third line correcting glossing of the phrasal verb "count down" (literal and wrong *compté bas* changed to correct *a compté*), and in the third/fourth line correcting both word choice and agreement in the glossing of "the final seconds" from ungrammatical *le final secondes* ("the-MASC-SING last-MASC-SING seconds-FEM-PLUR") to grammatical *les dernières secondes* ("the-PLUR last-FEM-PLUR seconds-FEM-PLUR").

We also obtained strong gains using "improvement" on Faroese segmentation (Table 4). However, despite getting an excellent result for the "Writing" task on the Farsi news story (Table 5), this was not duplicated on the other Farsi texts. The improvement operation clearly needs further study.

4.6 Random variability

Many errors seem purely random, with no obvious cause. For example, in one text the English segmentation was done using an underscore to mark segment breaks, rather than the vertical bar that had been requested; the vertical bar was correctly used in the other five texts. This is again unsurprising. It is well known that GPT-4 displays this kind of random variability in most domains, including ones as elementary as basic arithmetic, with the variability changing over time (Chen et al., 2023).

4.7 Language-specific and qualitative aspects

The above subsections focused primarily on quantitative and generic aspects of the texts. It is not enough for texts to be linguistically correct: they also need to be engaging and culturally appropriate. In this subsection, we briefly describe language-specific and qualitative aspects.

English As previously noted, the general standard of the English texts is high. Qualitatively, they respond well to the requirements given in the prompts. The quirky news story, about a raccoon found unconcernedly riding the Toronto subway, is amusing. The Bible passage, on the subject of the Golden Rule, quotes Matthew 7:12 appropriately and displays what in a human author would be called religious feeling. The football match comes across as a typical piece of hyperbolic sports journalism. The "language" piece is sensible and factual, and the "children's story" text a competent summary of "Goldilocks". The poem comes across more as a parody of a love poem than as an actual love poem, but this is a valid way to interpret the request. In general, the language is almost perfect, and only one small correction was made.

Faroese As seen in table 2, GPT-4 struggles with generating original Faroese text. After a native speaker has manually corrected the grammatical and lexical mistakes, the English glossing and PoS-tagging perform reasonably well on Faroese. However, for Faroese, there are not only grammatical and lexical errors in the texts, but the content is often nonsensical. The quirky news story was about a lamb literally "swimming in sun rays" and going viral on social media. The famous Faroese children's story is a made up story about a real Faroese teacher and poet, *Mikkjal á Ryggi*, who is described as having magical powers and playing a flute on a mountain. The passage about English required the least editing, but still resulted in fairly high error rate, because GPT-4 consistently used the wrong Faroese word for "English" — a word

repeated several times in the passage. GPT-4 seems to be confusing Faroese for Icelandic a lot of the time. Therefore, when hand-correcting Faroese text written by ChatGPT, it helps to be proficient in Icelandic. Faroese is a small language and it is not known how much Faroese text was included in the training of GPT-4, but it was likely very little compared to Icelandic. This might also explain why ChatGPT is not familiar with Faroese culture. The most common glossing and lemma tagging errors were also related to Icelandic, for example ChatGPT suggesting Icelandic lemmas for Faroese word forms, such as *sauður*, ('sheep', Icelandic) instead of *seyður*, ('sheep', Faroese)

Farsi The high error rates occurring even after improving "Write", as shown in Table 5, are mostly due to not considering writing style rules such as replacing spaces with semi-spaces when necessary: issues of this kind would not have a serious effect on reading comprehension or on the meaning. That considered, all six texts make good sense in most cases and are occasionally quite creative when it comes to coining words. The "quirky news story" about a stray cat and how people are used to have him around in the neighbourhood emphasises the impact that animals have on our life. In this text, a few words, although syntactically well written, make no sense considering the whole sentence. GPT-4 makes an exact interpretation of the "Quran passage", quoting Al-Hujurat 13, in which humans are considered united as a whole and are encouraged to resist discrimination, racism, and sexism to achieve equality. The "football match" evocatively describes the weather, the fans' emotions and the game itself. In the "language" text, although unnecessary, GPT-4 replaced some words when "improvement" was applied. The text gives some facts about the key role of the English language, the professional/educational opportunities it can bring to Farsi speakers' lives and the obstacles the learner might encounter such as lack of access to resources. The "children's story" refers to one of the most famous poems from Rumi's *Masnavi*, narrating the story *The Rabbit and The Lion*: in order to save himself, the rabbit tricks the lion and makes him jump into a well, reminding the readers that mental strength and intelligence can overcome challenging situations. The text was very well written except for two incorrectly chosen words. The "poem" generated by GPT-4 is surprisingly romantic. Considering that there are different styles in Farsi poetry—some having rhymes and some not—GPT-4 seems to have combined two styles: the writing format from Old poetry (two-verse stanzas) and no rhymes from New poetry. We note that writing Old poetry, which has rhymes, would be challenging even for modern Farsi native speaker poets. There were also a few mistakes on subject-verb agreement. One interesting point common to all six texts is how GPT-4 uses them as metaphors to give readers a life lesson.

Mandarin The Mandarin stories are very good. In contrast to the other non-English languages, the writing is flawless without grammar or word choice errors. Although a few phrases give an unnatural sense that suggest an AI generated the paragraph, the Mandarin stories are not influenced by English overall. The "quirky news story" was about a dog that is good at painting and is about to open its exhibition. The story is fluent, fun, and gives a warm feeling after reading, though the topic itself is irregular. The LA paragraph provides accurate insights into the English position and people's views in the general Mandarin society. The poem follows a structure of the modern Chinese style, and the content is very romantic overall.

Based on the evaluation shown in Table 2 and careful inspection of the results, GPT-4 consistently makes some errors in Mandarin segmentation, where it often mistakenly separates words from their particles. However, these results are better than those we obtained from the Jieba package. Regarding the other two annotation tasks, GPT-4 shows great capability in glossing and lemma-tagging from Mandarin to English.

Russian GPT-4 is a good tool for glossing and PoS-tagging Russian. As mentioned earlier, GPT-4 is very good at generating stories in some domains while facing challenges in others. The simplest task for Russian involved describing a traditional children's story. GPT-4 selected a well-known tale, "Masha and the Bear", and composed an essay about the typical occurrences in such stories. The "quirky news story" revolved around a bar owner's innovative offering – a service enabling lonely customers to rent a cat for company while drinking. This example highlights the remarkable creativity of GPT-4, capable of generating such imaginative narratives. The fictitious football game, which required some plain text editing, was about a world championship football match, where the heroes in blue and white uniforms won the match. The

Bible passage also underwent some editing. The piece about English language needed revision during glossing. The item which demanded most time was the Russian romantic poem about AI. The primary challenge was that the plain text generated by GPT-4 was composed in a poem-like style but lacked rhyme. After several re-prompts, the final version was chosen. This version necessitated substantial manual text editing and rephrasing, particularly the replacement of words at the end of lines to achieve rhyme. The glossing of the poem, however, was comparatively straightforward.

5 Conclusions and further directions

In general, C-LARA seems to be a good environment for investigating aspects of GPT-4's linguistic performance more complex than simply writing text. A publicly available version of the platform, hosted at the University of South Australia, will be released before the date of the conference.

The material presented in this paper should only be considered a preliminary study: obviously, one would ideally use more than five languages and multiple annotators. But given the rapid evolution of ChatGPT, it seemed more important to prioritise speed, and quickly gain some insight into the large-scale patterns. We summarise what we consider the main results.

The study examined the four tasks of writing, segmenting, glossing and lemma-tagging, all of which are key to a wide variety of text-based CALL systems. There is a great deal of variation across languages, and a great deal of random variation in general. However, for languages given a high enough priority by OpenAI, GPT-4 can write engaging, fluent text with an error rate of well under 1%, and perform the glossing and lemma-tagging tasks with average error rates in the mid single digits. English is not the only language in the high-priority group: Mandarin appears to be another. It is important to note that there are no generally available packages that can perform these tasks well, since they do not take proper account of multi-words, of key importance in CALL applications. We generated texts in six widely different domains, with roughly equal results cross-domain. This suggests that GPT-4's abilities are quite wide-ranging. For some tasks, including the common and important one of glossing English, it is possible to improve performance substantially by instructing GPT-4 to revise its output.

5.1 Further directions

Looking ahead, one obvious way to extend the work would be to repeat the experiments with a larger set of languages. It would probably be most useful to do this after using the data from the present study to further tune the system.

In particular, if we identify the common errors that GPT-4 is making in the annotation, we can try to adjust the prompt templates and/or few-shot prompt examples so as to reduce or eliminate the errors, either in the original annotation or in the "improvement" phase. To take a simple example, we found that the most common error in English segmentation was failing to split off elided verbs ("it's", "we'll" etc). It may be possible to address this by just adding one or two prompt examples. A related case in the opposite direction comes from Mandarin segmentation: here, the most common error is that aspectual and possessive particles are incorrectly split off verbs and nouns, and once again adjusting the prompts is a natural way to try to solve the problem. The "improvement" operation clearly merits further study.

A problem when carrying out evaluation like the one described here is that the annotation procedure is extremely time-consuming and tedious, and people are rarely willing to do more than small amounts. Once the public deployment of C-LARA is available, we hope it may be practicable to crowdsource a similar evaluation using multiple annotators, recruited through social media. We are tentatively planning an exercise of this kind for 2024.

Role of the AI coauthor

It is still unusual for an AI to be credited as the coauthor of a paper, and we briefly justify doing so. ChatGPT-4 is, as previously noted, the main implementor on the C-LARA project team, and responsible for a large part of the software design; further details are given in (Bédi et al., 2023a,b). Here, it has been involved throughout in discussing and planning all aspects of the experiment, read the paper, contributed some passages, and made useful suggestions. In particular, the statistical analysis in Section 3 was performed in response to an explicit suggestion from the AI.

References

Elham Akhlaghi, Branislav Bédi, Matthias Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. Overview of LARA: A learning and reading assistant. In *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, pages 99–103.

Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Birgitta Björg Guðmarsdóttir, Hanieh Habibi, Bjartur Örn Jónsson, Manny Rayner, and Sigurður Vigfússon. 2020. LARA: An extensible open source platform for learning languages by reading. In *Proc. EUROCALL 2020*.

Branislav Bédi, ChatGPT-4, Belinda Chiera, Cathy Chua, , Neasa Ní Chiaráin, Manny Rayner, Annika Simonsen, and Rina Zviel-Girshin. 2023a. ChatGPT + LARA = C-LARA. Presented at SLaTE 2023.

Branislav Bédi, ChatGPT-4, Belinda Chiera, Cathy Chua, Catia Cucchiarini, Neasa Ní Chiaráin, Manny Rayner, Annika Simonsen, and Rina Zviel-Girshin. 2023b. ChatGPT-Based Learning And Reading Assistant: Initial report. Technical report. https://www.researchgate.net/publication/372526096_ChatGPT-Based_Learning_And_Reading_Assistant_Initial_Report.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Yung-Sung Chuang. 2019. Robust Chinese word segmentation with contextualized word representations. *arXiv preprint arXiv:1901.05816*.

Tatsuya Hiraoka, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Stochastic tokenization with a language model for neural text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1620–1629.

Chu-Ren Huang, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot. 2007. Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 69–72.

Neasa Nı Chiaráin, Neimhin Robinson Gunning, Oisın Nolan, and Madeleine Comtois. 2023. Filling the SLaTE: examining the contribution LLMs can make to Irish iCALL content generation.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Annika Simonsen and Branislav Bédi. 2023. Using generative AI tools and LARA to create multimodal language learning resources for L2 Icelandic. In *Proc. EUROCALL 2023*.

Dekai Wu and Pascale Fung. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Fourth Conference on Applied Natural Language Processing*, pages 180–181.