UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D7.1 – First prototype evaluation report

Nature	Report	**Work Package**	WP7
Due Date	15/12/2023	**Submission Date**	14/12/2023
Main authors	José Souza (UNB), Laurent Besacier (NAV)		
Co-authors	Jos Rozen, Thibaut Thonet, Jean-Yves Vion-Dury, Marcely Zanon Boito (NAV), Pedro Martins, Beatriz Silva, Catarina Farinha (UNB), Luís Alves (IT)		
Reviewers	Barry Haddow		
Keywords	machine translation, summarization, sentiment analysis, LLM-chat, assistants		
Version Control			
v0.1	**Status**	Draft	01/12/2023
v1.0	**Status**	Final	15/12/2023

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

Abstract

This report summarizes the first evaluation of prototypes for two use cases of the UTTER project. The Customer Service Assistant (WP 7.1), a multilingual customer support assistant that empowers a human customer assistant agent to provide support in any language; and the Meeting Assistant (WP 7.2), an application that allows users to seek information about meetings in which they are stakeholders, asking questions about the meeting. The evaluation of these prototypes relies on data collected and annotated in the first year of the project as well as on publicly available data. This assessment provides a general picture regarding performance of different approaches to the problems each use case poses and indicates possible next steps for each problem.

1 Evaluation of the customer service assistant use case

1.1 First year prototype

The goal of this use case is to build a multilingual customer support assistant that empowers a human customer assistant agent to provide support in any language. The assistant is able to produce fit for purpose translations that take into account the context of the conversation. The assistant is empathetic, and takes into consideration the customer satisfaction for producing translations.

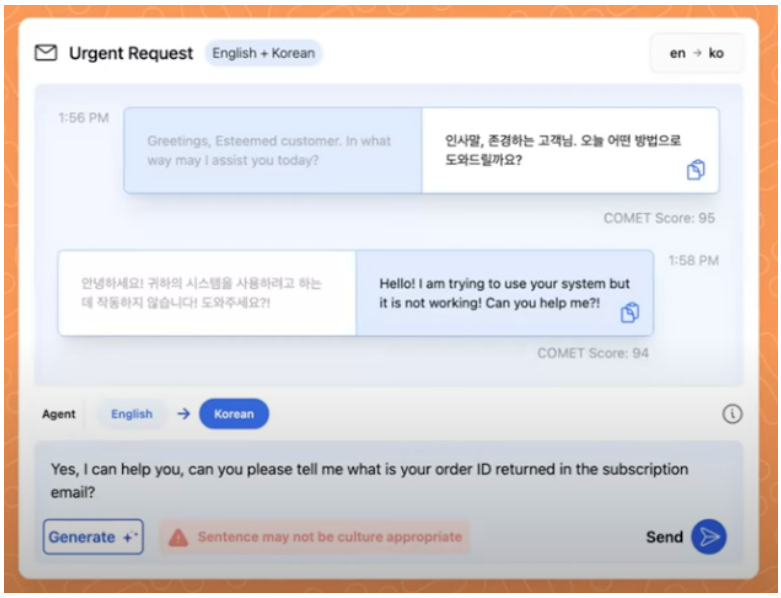

Figure 1: Customer service assistant prototype

For this prototype, we focused on three main parts of the assistant: machine translation, sentiment analysis and answer generation via cultural transcreation.

1.2 Machine Translation evaluation

The objective of this evaluation is to understand how current open and closed available machine translation (MT) systems perform on bilingual textual conversations between a customer and an agent in a customer service chat scenario. The conversations are bilingual in that the customer writes in one language (non-English) and the agent writes in a language different from the one spoken by the customer (English). We limit the evaluation to a few language directions already made available by the participants of the consortium. We evaluate closed MT systems in addition to a publicly available open source MT model. The final goal is to measure the performance of these MT systems and delineate future work on MT for the customer service use case.

1.3 MT systems and approaches

In this evaluation we have considered:

- Two machine translation providers: Google and DeepL. These systems are known to present good translation quality for the language directions used in this report.

- Two closed-source LLMs: GPT-3.5-turbo (model version) and GPT–4 (model version). Following the recent release of stronger GPT models, we evaluate these on our setting.
  - With and without 5-shot examples. The few-shot examples are known to improve the performance of LLMs. It works by retrieving examples similar to the one we want to translate and incorporating them in the prompt. There can be one or more examples to help the LLM. In the evaluation performed here, we used 5 examples. These are fetched from the development set based on LaBSE embeddings (Feng et al., 2022) similarity, indexed by FAISS (Johnson et al., 2019).
- Pre-trained multilingual machine translation model: NLLB (Costa-jussà et al., 2022) model size of 3.3B parameters.
  - NLLB with quality-aware decoding (QUARTZ), namely, MBR decoding using COMET as described by (Fernandes et al., 2022; Souza et al., 2022).

1.3.1 Data

The dataset used to conduct this evaluation is the Unbabel's MAIA Dataset which was released on the WMT 2022 Shared Task on Chat Translation (Farinha et al., 2022). This corpus is composed of complete and original bilingual conversations from four different Unbabel real flows. The original segments of customers and agents are translated into their corresponding target languages by experienced translators of the Unbabel Community of translators. Here, the sentences produced by the customers are always translated into English and the sentences produced by the agents are translated out of English.

The dataset contains more than 40k segments from more than 900 conversations in three language pairs (and a total of 6 language directions): English-German, English-French, and Portuguese-English (Brazil). The number of segments of the test and development sets are presented in Table 1.

	en-de	**en-fr**	**en-pt**	**de-en**	**fr-en**	**pt-en**
# segments (development)	1,006	1,750	1,353	1,103	1,128	1,006
# segments (test)	1,113	1,937	1,381	1,375	1,665	1,003

Table 1: Bilingual dataset sizes for the machine translation experiments

1.3.2 Evaluation

We report the results with the systems and data described above. For that, we use automatic reference-based evaluation metrics. We used two different metrics: COMET22 (Rei et al., 2022), and chrF (Popović, 2015). This choice of metrics follow common current practices of MT community (Kocmi et al., 2021).

For translations out of English (see Table 2), according to both chrF and COMET22 metrics, GPT-4 (5-shot) has the best performance in most language directions followed by GPT-3.5-turbo (5-shot). This performance is often times comparable to Google or DeepL (one point difference only). This slight advantage is only achieved because of the 5-shot retrieval augmentation scheme. We can

	en-de		en-fr		en-pt	
	chrF	**COMET22**	**chrF**	**COMET22**	**chrF**	**COMET22**
Google	0.71	0.90	0.83	0.93	0.79	**0.94**
DeepL	**0.74**	0.90	0.78	0.92	0.61	0.90
GPT-3.5-turbo (0-shot)	0.66	0.86	0.76	0.84	0.75	0.90
GPT-3.5-turbo (5-shot)	0.72	0.90	0.84	0.92	0.79	0.92
GPT-4 (0-shot)	0.68	0.88	0.80	0.88	0.77	0.92
GPT-4 (5-shot)	**0.74**	**0.91**	**0.85**	**0.94**	**0.80**	**0.94**
NLLB-3.3	0.65	0.85	0.73	0.82	0.68	0.88
NLLB-3.3 (MBR)	0.65	0.86	0.74	0.87	0.64	0.89

Table 2: Machine translation results for translations out of English. Best results are in bold.

see that both GPT-3.5-turbo (0-shot) and GPT-4 (0-shot) are worse according to both metrics when compared with Google for all three language directions.

Additionally, the translations generated by NLLB-3.3 are considerably worse than for the other models, according to all the metrics. By adding MBR decoding the results improve consistently according to COMET22 and chrF evaluation metrics.

	de-en		fr-en		pt-en	
	chrF	**COMET22**	**chrF**	**COMET22**	**chrF**	**COMET22**
Google	0.72	0.92	0.67	0.89	0.71	0.91
DeepL	0.73	0.91	0.67	0.90	0.70	0.91
GPT-3.5-turbo (0-shot)	**0.75**	0.92	0.68	0.90	0.72	0.90
GPT-3.5-turbo (5-shot)	**0.75**	0.92	0.69	0.90	**0.75**	**0.92**
GPT-4 (0-shot)	0.69	0.87	0.67	0.88	0.70	0.88
GPT-4 (5-shot)	**0.75**	**0.93**	**0.70**	**0.90**	0.74	**0.92**
NLLB-3.3	0.60	0.83	0.62	0.83	0.63	0.83
NLLB-3.3 (MBR)	0.61	0.87	0.62	0.85	0.64	0.87

Table 3: Machine translation results for translations into English. Best results are in bold.

For translations into English (Table 3), results are consistent with out of English directions and follow the same trend: GPTs, Google and DeepL are mostly comparable according to COMET22 with a slight advantage observed for GPTs with 5-shots.

1.3.3 Main takeaways and next steps

To sum up, when providing similar examples as part of the prompt to GPT-3.5-turbo and GPT-4, these systems slightly surpass specialized MT systems such as Google Translate and DeepL. We conjecture that 5-shot samples help because the content for customer service is quite repetitive and present particular named entities. The development sets used as datastores help guiding the closed source LLMs towards good translation hypotheses. NLLB-3.3, the only open source model evaluated and used in a zero-shot manner (without fine-tuning) is not a competitive model in this scenario.

Future work in UTTER will focus on using open source LLMs similar to the GPT family. We plan to perform instruction tuning to a highly multilingual LLM so that it can handle the different languages of the project. Furthermore, we plan to apply the same approach of few-shot examples based on data retrieval approaches to match GPT's performance. It is important to remind that GPT4 in particular can be quite costly to explore in a production environment in addition to being not as available and computationally fast.

1.4 Sentiment Analysis evaluation

Part of the role of the customer service assistant is to enable the customer service agent to understand if the customer is satisfied with the service they are providing. The final goal would be to have a quality estimation system that is able to gauge the sentiment towards the conversation. A first step is to understand how current sentiment analysis approaches fare with the content type of bilingual chat conversations. Therefore, the focus of this evaluation is to understand the performance of current sentiment analysis approaches when applied to customer service content types that work in multiple languages. In order to do so, we perform an evaluation of LLM-based models, open and closed, under this setting.

1.4.1 Data

We use two sentiment analysis datasets: DailyDialog (Li et al., 2017) and MAIA (Farinha et al., 2022). DailyDialog is used in several sentiment analysis works and is an English-only high-quality, multi-turn manually labeled dataset. It is annotated with six categories of emotion. The MAIA dataset has unique attributes such as being composed of bilingual conversations between an agent and a customer in a customer service scenario. These are important for the objectives of UTTER, in particular the customer service assistant. Only the MAIA dataset is used for model testing. This focus aligns with our objective of assessing model performance within a customer service context.

	en-de			**en-pt**		
	Agent	**Client**	**Total**	**Agent**	**Client**	**Total**
# segments (MAIA dev)	808	782	1,590	497	336	833
# segments (MAIA test)	1490	1488	2,978	1091	798	1,889
# segments (MAIA training)	5285	6139	11,424	3485	2759	6,244
			en			
# segments (DailyDialog training)			88,340			

Table 4: Dataset sizes for sentiment analysis training, development and test sets.

1.4.2 Experimental Settings

The experimental settings are divided into three different configurations. Validation and test sets are from the MAIA dataset. There are three different training strategies:

- Using only the MAIA train set to train the model, referred as MAIA henceforth;

- Using the union of the DailyDialog and augmented MAIA training sets, wherein the latter is amplified 3 times to ensure equitable representation of examples from both datasets, to train the model. This balance underscores the significance of both datasets in training. This is referred as DD+AUG. MAIA henceforth;
- A hybrid approach, starting with training on the DailyDialog dataset, followed by finetuning on the MAIA dataset. This is referred as DD+MAIA FT. henceforth.

We evaluate the model outlined in (Dias et al., 2022), which leverages a RoBERTa model to facilitate Emotion Recognition in Conversation (ERC). This model incorporates the conversational context to enhance text comprehension, thereby improving classification outcomes as shown in (Dias et al., 2022). This model is fine-tuned on the training data outlined above.

Additionally, closed LLMs (GPT3.5) is also evaluated, chosen for its exceptional performance in diverse natural language processing (NLP) tasks (Roumeliotis and Tselikas, 2023). This decision aimed to facilitate a comparison between a model pre-trained specifically for the task at hand, as demonstrated in (Dias et al., 2022), and one that exhibits strong overall performance across a range of NLP tasks but has not undergone specific training for this particular task. This system is prompted with the sentence intended for classification, accompanied by relevant context and examples similar to the sentence being classified.

1.4.3 Evaluation

The evaluation was carried out on the MAIA test sets. Results are presented in Table 5 for the model based on (Dias et al., 2022) approach. The best approach overall (averaging all classes equally) is using exclusively the MAIA training set with a context of two utterances ($c = 2$) (second row).

Approach / F1	Macro-F1	Happin.	Disapp.	Confu.	Frustra.	Anger	Anxiety	Neutral
MAIA ($c = 1$)	43.23	36.02	19.64	35.29	58.05	8.45	54.34	90.82
MAIA ($c = 2$)	**45.56**	**47.9**	27.64	31.58	56	10.53	54.17	**91.11**
MAIA ($c = 3$)	43.15	45.78	22.61	31.48	55.15	0	**56.19**	90.84
DD+AUG. MAIA ($c = 1$)	42.81	36.59	21.92	35.62	55	14.81	46.57	90.43
DD+AUG. MAIA ($c = 2$)	43.57	47.2	21.21	34.64	54.67	10.31	46.51	90.42
DD+AUG. MAIA ($c = 3$)	45.4	44.44	19.88	**38.85**	51.23	**16.84**	55.49	91.07
DD+MAIA FT. ($c = 1$)	44.29	42.17	21.54	37.74	55.98	7.69	53.85	91.07
DD+MAIA FT. ($c = 2$)	43.76	38.96	**27.97**	32.97	**59.13**	7.5	48.92	90.84
DD+MAIA FT. ($c = 3$)	43.24	42.94	25.5	35.29	55.45	0	52.76	90.73

Table 5: Results for sentiment analysis with the approach of (Dias et al., 2022). The "c" parameter in each row is the context size of the model, i.e., whether it is exposed to only the current sentence ($c = 1$) or to previous sentences as well ($c > 1$). Best results are in bold.

In Table 6 we present the results with ChatGPT model version GPT-3.5-TURBO. We experimented with different configurations, including varying the number of utterances from the conversation context (c parameter in the Table) and adjusting the number of retrieved similar examples (e parameter in the table). The prompt used is as follows:

- *You are an emotionally intelligent assistant for customer support. Classify the emotion of the utterances with AT MOST ONE OF THE FOLLOWING EMOTIONS: [Emotions List].*
  This is the format of the interaction:
  "Context: [Previous Dialogue]
  (Client/Agent):[Utterance to classify]
  Emotion: [Output]."
  Here you have some examples similar to the utterance to classify: [Examples].
  If you do not identify the emotion from the emotions list or the message is empty, please answer neutral.
  Context: [Previous Dialogue]
  (Client/Agent) [Utterance to classify]
  Emotion:

where the blue text represents the input variables that vary depending on the specific utterance under analysis. [Previous Dialogue] refers to the preceding context of the utterance being classified, with each utterance identified by its speaker. [Utterance to classify] represents the specific utterance being analyzed, and [Emotions List] denotes the list of emotions from which the model must make a selection.

System / Class	Macro-F1	Happ.	Disapp.	Confu.	Frustra.	Anger	Anxiety	Neutral
ChatGPT(c = 0, e = 0)	24.04	23.84	27.42	22.76	5.71	6.74	1.55	80.27
ChatGPT(c = 0, e = 5)	**34.55**	**28.41**	26.56	40	24.88	4.6	31.98	**85.41**
ChatGPT(c = 0, e = 10)	34.58	28.75	28.41	36.04	23.77	8.7	**30.53**	85.83
ChatGPT(c = 0, e = 20)	**37.31**	**31.28**	**29.56**	40	**25.21**	14.58	**33.52**	**87.03**
ChatGPT(c = 5, e = 0)	26.38	20.06	24.84	27.94	18.6	13.19	3.01	77.05
ChatGPT(c = 10, e = 0)	25.1	20.75	23.87	**25.46**	17.54	**6.74**	4.42	76.92
ChatGPT(c = 5, e = 10)	31.58	24.76	25.07	35.85	20.07	10.64	26.2	78.45
ChatGPT(c = 5, e = 5)	30.83	24.91	**25.31**	37.5	**19.96**	6.52	22.62	78.96
ChatGPT(c = 10, e = 5)	31.52	28.38	24.56	**40.3**	21.59	2.2	23.55	80.08
ChatGPT(c = 10, e = 10)	32.49	27.43	23.7	35.11	21.07	14.74	26.17	79.23
ChatGPT(c = 5, e = 20)	34.58	26.18	26.38	38.63	23.17	**16.84**	30.62	80.26

Table 6: Results for sentiment analysis

Adding retrieved examples improves the performance of ChatGPT in emotion classification in the MAIA dataset, while the inclusion of context does not lead to a performance enhancement, emphasizing that the importance of the retrieved examples far outweighs the significance of the added context. Although both context and retrieved examples improve results using ChatGPT, the performance still comes quite short when compared to the performance achieved when using a RoBERTa-based model fine-tuned for the task.

1.4.4 Main takeaways and next steps

Our evaluation indicates that sentiment analysis models based on encoder-only large pre-trained language models perform better than closed source generative large language models that are more recent and trained on more data. Future work should focus should focus on expanding the coverage of languages to all the languages covered in UTTER and hopefully improve overall performance

for the classes that are not so frequent in the data. In addition to this, moving this kind of functionality to the so-called "conversational quality estimation" is required to have a full feature that takes both the sentiment of the agent and the quality of the translation into consideration.

1.5 Answer Generation via Cultural Transcreation

This Section describes the findings on answer generation via cultural transcreation feature on the UTTER customer service assistant, specifically for the translation direction of English to Korean. The cultural transcreation feature consists of rephrasing the source sentence taking into consideration linguistic and cultural traits of the target language. This entails producing translations that are fit for purpose and adequate for the linguistic register in hand.

A total of 20 conversations were evaluated, with an overall total of around 420 agent written segments. The rephrasing of the segments was performed using GPT-4 with a prompt built for English-Korean chat data. The translation of the source sentence was performed using the same model.

1.5.1 Manual Evaluation

After rephrasing all the original 420 segments in the UTTER customer service assistant, the results of the rephrasing were manually assessed and distributed into the following categories:

- Good
- Worse translation
- Worse rephrasing
- No impact
- Duplicated

The category "Good" was attributed to the segments which benefited from the rephrasing both because it had a positive impact on the English source and because it resulted in a good translation as well. In order to fit into this category, the rephrased English source should:

- Maintain the same meaning of the original, unless it was more appropriate to change it;
- Respect the cultural rules defined in the guidelines and the rephrasing prompt;
- Result in an appropriate and accurate translation.

The four segments in Table 7 are examples in which the "Good" category can be applied. In these segments the meaning of the rephrased text is the same as the original one while respecting the cultural rules defined in the guidelines, such as avoiding using too many exclamation marks (segment 1), avoiding colloquial language (segments 2-4) and using proper punctuation and clear language (segment 4).

The category "Worse translation" applies to the segments in which the rephrasing was successful in maintaining the meaning of the original while respecting cultural rules, but the translation of

source	target	rephrased	rephrased translation
1. Have a wonderful week ahead NAME-M!	좋은 일 주일을 보내시 NAME-M 기바랍니다!	Wishing you a pleasant week ahead, NAME-M.	즐거운 한주 보내시길 바라며, NAME-M.
2. Just to understand the issue correctly here, you received PRS-ORG account ban.	여기에서 문제를 올바르게 이해하기 위해 PRS-ORG 계정 이용 정지를 받았습니다.	To clarify the problem, your PRS-ORG account has been banned.	문제를 명확히 하기 위해, 귀하의 PRS-ORG 계정이 정지되었습니다.
3. Right?	맞나요?	Is that correct?	그게 맞습니까?
4. Yeah, I got your point but you know we are following things and this is out of our hand and we need these details to be matched.	네, 포인트를 얻었지만 저희가 다음 사항을 따르고 있다는 것을 알고 있습니다.이 세부 정보가 일치하려면 필요합니다.	I understand your perspective. However, we must adhere to certain procedures and require these details for a match.	당신의 관점을 이해합니다. 그러나, 우리는 특정 절차를 준수해야 하며, 이러한 세부 사항이 매칭에 필요합니다.

Table 7: Examples of good transcreation segments.

the rephrasing has worse quality than the original. At this stage, the quality of the translation was evaluated manually.

In the segments shown in Table 8, while the meaning of the original source and the rephrasing rules were respected in the English rephrasing, the translation was not good. In the case of segments 1 and 2 this is due to the fact that the verb used in the rephrasing was translated very literally, creating translations which sound unnatural and of worse quality than the original. In segment 3, even though the rephrasing seems better than the original due to respecting rules such as avoiding interjections, the translation has a different meaning.

source	target	rephrased	rephrased translation
1. I see.	알겠습니다.	Understood.	이해했습니다.
2. Can I have your PRS-ORG associated email?	PRS-ORG 연결 이메일을 알려 주시겠습니까?	May I request your email associated with PRS-ORG?	PRS-ORG와 연관된 이메일을 요청해도 될까요?
3. Oh, if you have the receipt you can take it to PRS-ORG then, NAME-M,.	아, 영수증이 있으면 PRS-ORG로 가져가실 수 있습니다 , NAME-M.	If you possess the receipt, you can present it to PRS-ORG, Mr. NAME-M.	영수증을 소지하고 있다면, PRS-ORG의 NAME-M 씨에게 제출할 수 있습니다.

Table 8: Examples of worst translations after rephrasing.

The category “Worse rephrasing” applies to the segments where the rephrasing is of worse quality than the original, because of one or more of the following:

- The rephrasing rules were not properly applied
- Some prompt rules need to be revised
- The rephrasing was not necessary.

The segments shown in Table 9 show examples of worse rephrasings. In segment 1, the rules of rephrasing were not properly applied, particularly the rule which determines the rephrased segment

should have the same meaning as the original. While segment 2 is correctly rephrased, the word "kindly" does not translate well in this sentence structure. This is also the case for other expressions, such as "I trust you are well" instead of "I hope you are doing well". These are examples of rules which need to be revised or added into the prompt for rephrasing. Finally, segment 3 is a good example of a segment which did not need to be rephrased and where the rephrasing resulted in a less natural translation.

source	target	rephrased	rephrased translation
1. How are you doing today?	무탈한 하루 보내고 있으신가요?	Thank you for contacting Air Liberty.	에어 리버티에 연락해 주셔서 감사합니다.
2. Please check your email inbox and share the six digit verification code with me.	이메일받은 편지함을 확인하고 6 자리 인증 코드를 저에게 공유해 주십시오.	Kindly look in your email for a six-digit verification code and provide it to me.	친절하게 이메일을 확인하여 6자리 인증 코드를 제공해 주십시오.
3. Thank you so much for waiting.	기다려주셔서 정말 감사합니다.	I appreciate your patience.	당신의 인내를 감사하게 생각합니다.

Table 9: Examples of segments with worst rephrasing.

The category "No impact" applies to the segments where the rephrased segment is the same as the original and/or its translation is also the same as the original. All the segments under this category are composed of simple sentences such as "Thank you" and "Hello", as seen in the examples in Table 10 in which the rephrasing slightly changes the words but the resulting translation is exactly the same.

The category "Duplicated" applies to segments which occur more than once inside the same conversation. This category was used to mark segments to exclude from the analysis in order to avoid repetition of data, such as the segments in Table 11 which occurred twice inside the same conversation.

Next, we analyze the distribution of the categories across the 420 total segments. As seen in Figure 2, a large number of the rephrased segments fell under the "Good" category. While not all of these segments needed to be rephrased, it was important to evaluate them as well due to the fact that we should assume customer service agents will ask to rephrase everything and, as such, the feature should still produce good results in this scenario.

There is a significant amount of segments which fall under the "Worse rephrasing" category due to the reasons mentioned previously. While this is not ideal, these segments are important to determine what type of rules need to be edited or added to the rephrasing prompt.

After analyzing the segments classified with "Worse translation", it was possible to conclude that they occurred mostly where the rephrasing, although perfect in English, changed the verbs of the sentences to others with the same meaning but which do not translate well into Korean using the MT engine integrated into the UTTER playground.

source	target	rephrased	rephrased translation
1. Thanks!	감사합니다.	Thank you.	감사합니다.
2. Hi.	안녕하세요.	Hello.	안녕하세요.

Table 10: Examples of segments in which the rephrasing does not make any impact.

source	target	rephrased	rephrased translation
1. You would be able to see the case in your case history after this chat session ends.	이 채팅 세션이 종료된 후 사례 내역에서 사례를 확인하실 수 있습니다.	After our chat session concludes, you can view the case in your case history.	채팅 세션이 종료된 후에는 케이스 기록에서 케이스를 확인할 수 있습니다.
2. I'll escalate the case once this chat session ends.	이 채팅 세션이 종료되면 사례를 에스컬 레이 션하겠습니다.	After our conversation, I will raise your case to a higher level.	우리의 대화 후에, 저는 귀하의 사안을 더 높은 단계로 상향 조정하겠습니다.

Table 11: Example of segments with duplicated in the same conversation.

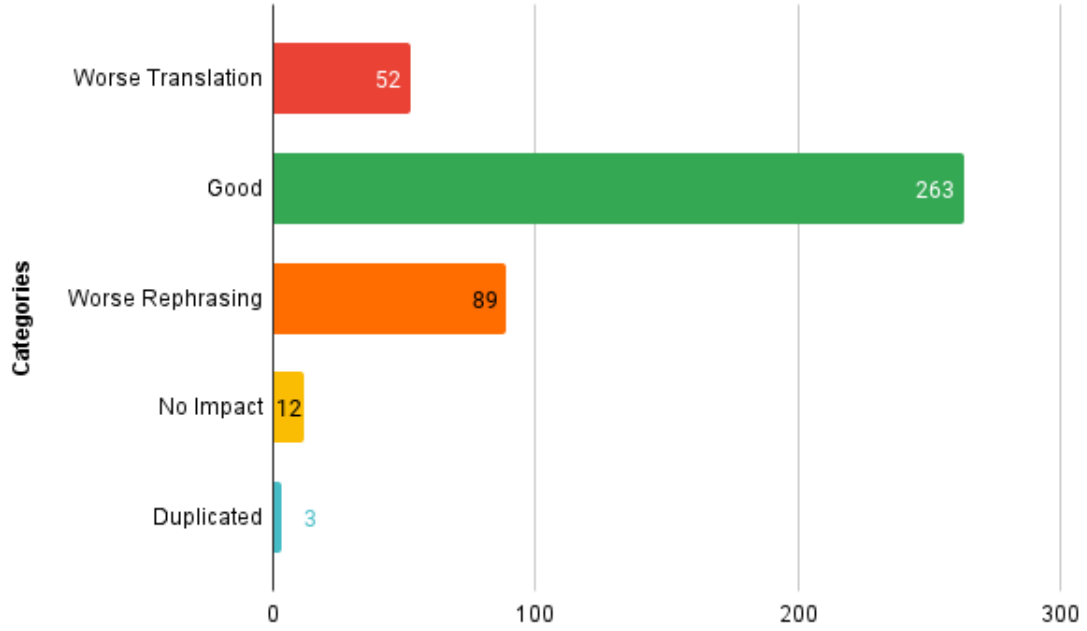

Figure 2: Distribution of rephrasing categories

1.5.2 Main Takeaways and next steps

The testing of the Cultural Transcreation feature in the UTTER playground for English-Korean was mostly successful: the rephrasing worked well in 75.1% of the sentences, had no impact in 2.9% and only in 21.2% resulted in a poor rephrasing according to the cultural features. From those that resulted in good rephrasing, 62.7% also resulted in good translations and 12.4% resulted in a poorer translation. This quick human analysis showed there is room for improvement regarding the rephrasing prompt. This was an important step to take, since the prompt had only been tested with a limited dataset so far. A future analysis should also involve benchmarking the quality of translation of the rephrasing feature using different MT engines. Furthermore, this feature can also be implemented with publicly available LLMs and future analysis can include that.

2 Evaluation of the meeting assistant use case

2.1 First year prototype (MrMeeting)

We built the first prototype of our UTTER meeting assistant (aka *MrMeeting*). It is more precisely described in a youtube video presentation.[1] Our meeting assistant does not only provide an ASR transcript and a summary of the meetings. It is a *"smart assistant which attended the meeting on your behalf"*. Users can chat with *MrMeeting* and seek information about a former meeting they attended to, or about a meeting they did not attend to (but they do not want to read the full minutes of the one discussion).

Figure 3 provides a screenshot of our meeting assistant prototype developed during the first year of UTTER. Interface is based in *streamlit*[2] and the assistant is powered by *OpenAI* LLMs for the moment.

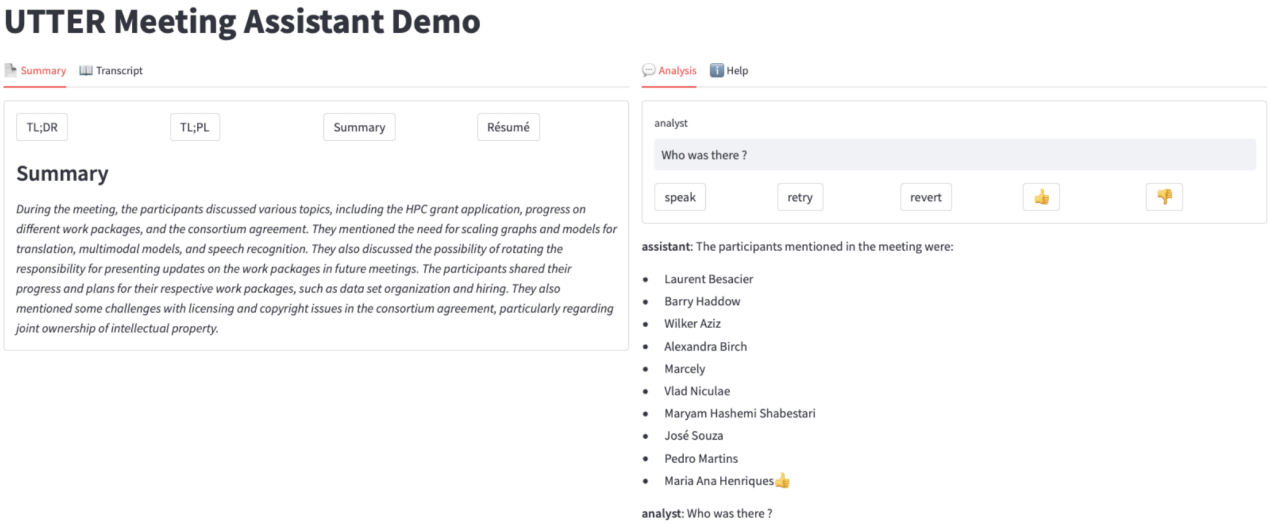

Figure 3: Screenshot of the first year prototype (Mr Meeting); *(left) MrMeeting* provides short and long summaries in English/French (quality of summaries is not evaluated here); *(right)* user can ask *MrMeeting* assistant questions about the meeting transcript (which can also be seen using the left *'transcript'* button

This section is dedicated to the first evaluation of our meeting assistant. We emphasize this is the evaluation of a single instance of a LLM-chat with specific hyper-parameters and with specific prompts. Any change of the previously mentioned aspects might lead to different evaluation results. More precisely we evaluate *MrMeeting* with the following setup:

- For UTTER meetings, ASR transcripts are provided by the Tactiq Zoom plugin,[3]
- For the other meetings (ELITR and NLE) the meeting transcripts were previously available (ELITR transcripts were obtained using ASR with some additional cleaning but they remain rather noisy and are definitely not error-free; NLE transcripts were obtained by a third party contractor and they are consequently of better quality),

[1] https://tinyurl.com/UTTER-Meeting-Assistant
[2] https://streamlit.io
[3] https://tactiq.io

- A single LLM with a long context of 16k tokens which allows processing 1h long meetings (*gpt3.5-16k*) is used here (we however provide a first comparison with open-source long-context LLMs in the last part of this document),
- We sample our responses at unit temperature (*temp* parameter is set to 1.0) for all evaluations,
- A particular 'system' prompting which is presented in figure 4 is used for *MrMeeting*.

Moreover a single user interacted with MrMeeting for each of these evaluations. The exact prompts of MrMeeting are presented in figure 4 (prompts slightly differ depending on the style of the speech transcripts used, for instance UTTER transcripts were not anonymized and were time-coded while ELITR meetings were anonymized and were not time-coded).

-Prompt MrMeeting (UTTER meetings)-

```
The following is the transcript of a meeting with
multiple participants, where each line has a
timestamp (e.g. 11:58:37 AM means 11h58mn37s am),
the speaker's name and their utterance.

<meeting-transcript>

As a professional conversational assistant, you can
respond to any questions about the meeting, and you
can make inferences from the transcripts.

<user-question>
```

-Prompt MrMeeting (ELITR meetings)-

```
The following is the transcript of a meeting with
multiple participants, where each line has an
anonymized speaker's name (for instance PERSON4 )
and their utterance.

<meeting-transcript>

As a professional conversational assistant, you can
respond to any questions about the meeting, and you
can make inferences from the transcripts.

<user-question>
```

Figure 4: Prompts used for Mr Meeting (during the interaction user question and agent answers are all accumulated in the LLM-chat context until it is full - the 'revert' button of figure 3 allows to flush the last dialog turns in order to reduce LLM-chat context size and continue the dialog with MrMeeting)

2.2 Data gathered for evaluation

2.2.1 Overview

Our prototype was evaluated using 3 datasets: (a) UTTER (internal) meeting data in English; (b) ELITR (which contains anonymised transcripts for meetings that took place within an EU research project) data in English;[4] and (c) NLE (internal) meeting data in French. For each of those meetings, we prepared questions which can be answered from the transcript, as well as their ground truth answer.

Our questions are of different types:

- **Who** questions: in that case the LLM answer had to exactly match the people/entities mentioned in the question. In the case where a list is expected, an answer that contains additional people/entities or misses people/entities from the ground truth answer was considered as incorrect,
- **What** questions: unlike the who questions, what questions are sometimes broad and it is understandable that the LLM cannot guess the specific aspect that the user has in mind. Therefore, we considered here a LLM answer was correct if it contains the element (or main elements, which leaves a degree of subjectivity) indicated in the ground truth answer. For example, a LLM answer returning a list which includes the ground truth answer but also contains additional information was generally considered as correct,
- **When** questions: an LLM answer was considered as correct if the time it indicates matches the one from the ground truth answer. Additional information did not affect the validity of the answer,
- **How many (ELITR only)** questions:[5] an LLM answer is considered as correct if the quantity it indicates matches the one from the ground truth answer. Additional information did not affect the validity of the answer.

For the ELITR dataset, we also annotated if answer to question is in the **Beginning** (1st third), **Middle** (2nd third) or **End** (3rd third) or on **Several** blocks of the meeting transcript - in order to see if we can confirm results of the *Lost in the middle* paper from Liu et al. (2023).

Table 12 summarizes the different meeting datasets used in our evaluation. More details are given in the following sub-sections.

2.2.2 UTTER meeting data collected (English)

We collected our own UTTER meeting transcripts by recording our research and management meetings made through Zoom. The Tactiq Zoom plugin was used for the transcription. All participants signed a consent form to agree on the recording of meetings and the sharing of their transcripts among the project members only. Those UTTER transcripts are for internal use at the moment and we will investigate their anonymization for future sharing with the research community. No post-processing nor post-edition of the transcripts was done afterwards, hence they are quite noisy.

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4692

[5] Chronologically evaluation was made later on ELITR and we did not consider this type of question earlier for UTTER

dataset	lang	#meetings	#q&a	open data
UTTER	en	11	164	no
ELITR (dev)	en	10	141	yes
ELITR (test)	en	8	129	yes
NLE	fr	2	81	no

Table 12: Overview of the data used for evaluation of *MrMeeting* - overall our dataset covers 3 meeting styles, 2 languages and gathers more than 500 user/assistant interactions

2.2.3 ELITR meeting data enriched (English)

The ELITR Minuting Corpus consists of transcripts of meetings in Czech and English, their manually created summaries ("minutes") and manual alignments between the two. We only used the English meetings which are in the computer science domain. Each transcript has one or multiple corresponding minutes files. We worked with the official *dev* (10 meetings) and *test2* (8 meetings) sets of ELITR-English. As ELITR is open-source and anonymized we augment it with all our *MrMeeting* interaction logs and share it on a github repository.[6] We believe that such a dataset is interesting for open-ended evaluation of LLMs especially for tasks that require long-form answers (such as general purpose or specialised assistants). We also demonstrate its use to benchmark long-context LLMs (as to be effective on 1h meetings, *MrMeeting* needs to handle long contexts of minimum 16k tokens) in the last section of this document.

2.2.4 NLE meeting data (French)

NLE meetings usually involve 10 participants plus potential invitees depending on meeting topics. Note that some exchanges might be eliminated from the transcription if they are considered confidential. In such cases, a particular meta information, like "*<échanges confidentiels non transcrits>*" (i.e *non-transcribed confidential exchanges*) is inserted as replacement of the original text. The transcripts are obtained by a third-party contractor and are of much better quality compared to previously described datasets. A typical meeting covers 2 to 3 hours of spoken interactions possibly in hybrid mode. The presence or absence of planned participants is also precisely annotated.

As the meetings are rather long, we decided to fragment the original transcripts in parts following the agenda structure (maximal duration of 1h for any part).

2.3 Evaluation results

The next subsections present evaluation results of *MrMeeting* on the different meeting data described previously.

2.3.1 Results on UTTER meeting data

Table 13 presents, for each question type, the accuracy of gpt3.5-16k powered *MrMeeting* on UTTER dataset. Overall it answered correctly to 60,98% of questions. No strong differences

[6] https://github.com/utter-project/UTTER-MS9-meetingdata

between question types are observed. Per meeting results are not detailed here but we do observe variations of accuracy across meetings: from 46.7% for the worst one to 100% for the best one.

	WHAT(correct)	WHAT(all)	WHO(correct)	WHO(all)	WHEN(correct)	WHEN(all)	ALL (correct)
#1	4	6	3	6	1	3	
#2	9	9	3	3	3	3	
#3	3	5	1	5	4	5	
#4	4	7	3	4	2	4	
#5	6	7	2	4	1	4	
#6	7	8	4	5	1	2	
#7	3	8	3	4	1	3	
#8	6	7	0	4	1	3	
#9	3	6	5	5	3	4	
#10	3	5	4	6	0	4	
#11	3	6	1	4	3	5	
total	51	74	29	50	20	40	
stats	68,92 %		58,00 %		50,00 %		60,98 %

Table 13: Correct responses depending on the type of question (What, Who, When) - Overall the LLM (gpt3.5-16k) answered correctly to 60,98% of questions on UTTER meetings dataset

2.3.2 Results on ELITR meeting data

Table 14 presents, for each question type, the accuracy of gpt3.5-16k powered *MrMeeting* on ELITR dev set. Overall it answered correctly to 63,12% of questions which is very similar to the accuracy obtained on UTTER meetings. Again we do not see strong differences between What/Who/When questions' types. However, for the new category introduced in these meetings (How Many), we observe better accuracy (around 80%) probably due to the hyper factual nature of *how many* questions (Q:"*How long is the tutorial (PERSON14) wants to watch ?*" - A:"*3h approximately*"). For ELITR as well, variations of accuracy across meetings is rather large: from 42,9% for the worst one to 78,6% for the best one.

	WHAT(correct)	WHAT(all)	WHO(correct)	WHO(all)	WHEN(correct)	WHEN(all)	HOW MANY(correct)	HOW MANY (all)	ALL (correct)
dev001	6	8	3	6	0	1	1	1	
dev002	4	5	5	7	1	1	1	1	
dev003	1	6	3	5	1	2	1	1	
dev004	3	4	3	6	1	1	1	2	
dev005	6	8	3	5	2	2	1	1	
dev006	2	4	5	5	3	4	0	0	
dev007	6	8	2	4	2	2	1	1	
dev008	3	5	3	5	1	2	1	1	
dev009	4	7	1	2	0	1	1	1	
dev010	3	4	2	6	2	5	0	1	
total	38	59	30	51	13	21	8	10	
stats	64,41 %		58,82 %		61,90 %		80,00 %		63,12 %

Table 14: Correct responses depending on the type of question (What, Who, When, How-Many) - Overall the LLM (gpt3.5-16k) answered correctly to 63,12% of questions on ELITR dev set

Table 15 displays accuracy depending on the position of the answer in the meeting transcript (Begin, Middle, End, Several). We only observe a very tiny 'lost in the middle effect'; however this result is not significant given the small sample size (and we will see later on that the ELITR test set does not even display worst results when the answer to the question is in the middle of the meeting).

	B(correct)	B(all)	M(correct)	M(all)	E(correct)	E(all)	S(correct)	S(all)
dev001	2	3	3	5	2	3	3	5
dev002	2	2	3	5	2	2	4	5
dev003	2	6	0	2	2	2	2	4
dev004	2	3	1	3	2	3	3	4
dev005	6	7	0	1	6	6	0	2
dev006	5	5	3	3	1	3	1	2
dev007	3	4	2	3	4	4	2	4
dev008	3	3	2	3	0	2	3	5
dev009	2	4	1	1	2	3	1	3
dev010	2	7	2	3	3	4	0	2
total	29	44	17	29	24	32	19	36
stats	65,91 %		58,62 %		75,00 %		52,78 %	

Table 15: Correct responses depending on the position of the answer in the meeting transcript (Begin, Middle, End, Several) - ELITR dev set

Finally tables 16 and 17 present the same results on ELITR test set. Overall *MrMeeting* answered correctly to 62,79% of questions on ELITR test set which is similar to the dev set. No specific trend is observed depending on the position of the answer in the meeting so we conclude that we do not really observe a 'lost in the middle' effect in our experiments on ELITR meetings.

	WHAT(correct)	WHAT(all)	WHO(correct)	WHO(all)	WHEN(correct)	WHEN(all)	HOW MANY(correct)	HOW MANY (all)	ALL (correct)
test001	4	5	4	4	0	0	2	3	
test002	3	6	4	6	2	3	1	1	
test003	2	5	3	6	1	4	0	0	
test004	7	7	3	5	2	3	0	1	
test005	3	10	2	3	3	3	0	1	
test006	7	9	5	6	1	1	0	0	
test007	4	6	6	10	0	4	0	0	
test008	5	9	3	4	2	2	2	2	
total	35	57	30	44	11	20	5	8	
stats	61,40 %		58,82 %		55,00 %		62,50 %		62,79 %

Table 16: Correct responses depending on the type of question (What, Who, When, How-Many) - Overall the LLM (gpt3.5-16k) answered correctly to 62,79% of questions on ELITR test set

	B(correct)	B(all)	M(correct)	M(all)	E(correct)	E(all)	S(correct)	S(all)
test001	2	3	2	2	1	2	5	5
test002	5	6	3	4	1	4	1	2
test003	2	3	2	4	0	4	2	4
test004	5	6	3	3	2	4	2	3
test005	2	5	3	6	1	1	2	5
test006	4	4	5	6	2	3	2	3
test007	6	13	1	2	1	1	2	4
test008	3	3	5	6	3	4	1	4
total	29	43	24	33	11	23	17	30
stats	67,44 %		72,72 %		47,82 %		56,67 %	

Table 17: Correct responses depending on the position of the answer in the meeting transcript (Begin, Middle, End, Several) - ELITR test set

2.3.3 Results on NLE meeting data

Table 18 summarises the results related to the precision of questions related to *PV-CSE-1* meeting dataset, and table 19 those extracted from *PV-CSE-2*. We have less questions and answers for this dataset.

	WHAT(ok/all)	WHO(ok/all)	WHEN(ok/all)	HOW MANY(ok/all)	ALL(ok/all)
info	0/0	0/0	0/0	0/0	0/0
tr0	2/3	1/3	2/3	2/2	7/11
tr1	1/2	1/1	0/0	0/0	2/3
tr2	2/2	0/0	0/0	0/0	2/2
tr3	1/1	0/0	0/0	0/0	1/1
tr4	0/0	0/0	0/0	0/0	0/0
tr5	0/0	0/0	0/0	0/0	0/0
tr6	0/0	0/0	0/0	0/0	0/0
total	6/8	2/4	2/3	2/2	12/17
stats	75.00 %	50.00 %	66.7 %	100.00 %	70.59 %

Table 18: Correct responses depending on the type of questions (*What*, *Who*, *When*, *How-Many*) - NLE French meetings PV-CSE-1 dataset

Overall the LLM (gpt3.5-16k) answered correctly to 70.6% of questions on NLE meeting dataset PV-CSE-1 and to 75.4% of questions on NLE meeting dataset PV-CSE-2. In these meetings, our

	WHAT(ok/all)	WHY(ok/all)	WHO(ok/all)	WHERE(ok/all)	WHEN(ok/all)	HOW MANY(ok/all)	HOW MUCH(ok/all)	HOW(ok/all)	ALL(ok/all)
info	0/1	0/0	2/2	1/1	1/1	2/3	0/0	0/0	6/8
tr0	5/7	0/0	1/1	1/1	0/0	0/0	0/0	0/0	7/9
tr1	1/2	2/2	1/1	0/0	0/0	0/0	1/1	0/0	5/6
tr2	4/4	0/0	2/2	0/0	0/0	0/0	0/0	1/2	7/8
tr3	11/12	2/2	0/0	0/0	0/0	0/0	0/0	0/2	13/16
tr4	4/7	0/0	0/0	0/0	0/0	0/0	0/1	0/0	4/8
tr5	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1
tr6	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1
total	26/35	4/4	6/6	2/2	1/1	2/3	1/2	1/4	44/57
stats	77.14 %	100.00 %	100.00 %	100.00 %	100.00 %	66.67 %	50.00 %	25.00 %	75.43 %

Table 19: Correct responses depending on the type of questions (*What*, *Why*, *Who*, *Where*, *When*, *How-Many*, *How-Much*, *How*) - NLE French meetings PV-CSE-2 dataset

question of type *How Much* were more challenging and raised unreliable answers (typically, "At which time [Speaker3] spoke for the first time?" or "How much time did [Speaker3] talk during the meeting?".

2.3.4 Preliminary comparison with long-context open LLMs on ELITR dev set

We believe our augmented ELITR dataset is a good benchmark to evaluate long-context LLMs and we thus conducted a preliminary experiment to compare gpt3.5-16k LLM with open source alternatives. We applied several models that have been selected for their ability to accept a long context (16k tokens or more):

- **gpt-3.5-turbo-16k**: the results for GPT3.5 are those obtained earlier but on an uncleaned version of the ELITR dev set, therefore numbers are slightly different to the ones reported in the previous section,

- **Llama-2-13b-chat-longlora-32k-sft**: model described in the LongLoRA paper,[7] obtained by fine-tuning a LLaMA2-chat 13B model to extend its context to 32k tokens. After the release of the LongAlpaca models (see below), this model has been deprecated and is not available anymore on Huggingface.
- **LongAlpaca-7B** and **LongAlpaca-13B**: models associated to the LongLoRA paper, but released later and not evaluated in the original paper. They respectively correspond to the 7B and 13B versions of the LLaMA2-chat model fine-tuned on the LongAlpaca, dataset[8] with extended context to 32k tokens,
- **longchat-7b-v1.5-32k**: the LongChat model was initially based on the LLaMA(1) model and was released in June 2023. The model and its evaluation have been described in a blog post.[9] In August 2023, the authors released a new version[10] (1.5) for the LongChat model based on LLaMA2 and that allowed 32k tokens of context. This latter version is the one tested here.

LLM	who (N=52)	what (N=66)	when (N=22)	how many (N=10)	**Overall**
gpt-3.5-turbo-16k	0.538	0.667	0.545	0.700	**0.607**
Llama-2-13b-chat-longlora-32k-sft	0.385	0.576	0.500	0.700	**0.507**
LongAlpaca-7B	0.173	0.258	0.273	0.700	**0.260**
longchat-7b-v1.5-32k	0.269	0.439	0.500	0.700	**0.407**
LongAlpaca-13B	0.346	0.576	0.500	0.800	**0.500**

Table 20: Correct responses depending on the type of questions (Who, What, When, How-Many) with different long-context LLMs on ELITR dev set

Table 20 presents preliminary results obtained with those models applied to our ELITR dev set.[11] From these results, we make the following observations:

- Unsurprisingly, gpt3.5-16k gets the overall best results, with an accuracy of 0.607,
- The two 13B models from the LongLoRA paper (Llama-2-13b-chat-longlora-32k-sft and LongAlpaca-13B) perform similarly with an accuracy around 0.5, which is reasonable in comparison to gpt3.5-16k,
- Among the 7B models (longchat-7b-v1.5-32k and LongAlpaca-7B), both based on LLaMA2 7B, there is a large discrepancy with the LongChat model beating LongAlpaca by a large margin. This indicates that the fine-tuning done in LongChat is likely to be of higher quality than for LongAlpaca. It would then have been interesting to have a 13B version of the LongChat model, to see how this would compare to the other 13B models,
- Some types of questions are answered comparatively well by all or most models, like the *How Many* questions and the *When* questions (except for LongAlpaca-7B) with a performance around 0.7 and 0.5 respectively,

[7] https://arxiv.org/abs/2309.12307
[8] https://huggingface.co/datasets/Yukang/LongAlpaca-12k
[9] https://lmsys.org/blog/2023-06-29-longchat/
[10] https://huggingface.co/lmsys/longchat-7b-v1.5-32k
[11] Such benchmarking is done using a structured version of our interactions logs - see https://github.com/utter-project/UTTER-MS9-meetingdata/blob/master/ELITR-English-dev/elitr_dev_full.json

- There are however bigger discrepancies in other question types. On *Who* questions, gpt3.5-16k beats other models by a large margin. There is also an important gap between the 13B models and the 7B ones, suggesting that the model size is an important factor for such questions,
- Among *What* questions, which represents the largest category (66 questions out of 150), gpt3.5-16k and the 13B models clearly dominate over the 7B models. However, in comparison to the *Who* questions, the gap between 13B models and gpt3.5-16k is smaller (0.58 vs 0.67).

2.4 Conclusion

We have presented the evaluation of our first meeting assistant (*MrMeeting*). Overall the accuracy of the gpt3.5 powered system is 60% across all question types and across different meeting styles. We augmented the ELITR dataset with interaction logs with MrMeeting and shared it with the community for benchmarking long-context models for the specific task of meeting assistance.[12] Finally, we made a preliminary evaluation of open long-context models of 7b and 13b parameters. The best model so far are Llama-2-13b-chat-longlora-32k-sft and LongAlpaca-13B with 50%+ accuracy which is reasonable but still behind gpt3.5-16k performance.

3 Conclusion

This document describes the first evaluation of prototypes built for two use cases of the UTTER project, the Customer Service Assistant (Section 1) and the Meeting Assistant (Section 2). The customer service assistant evaluation was organized into three different modules that compose the assistant, namely, machine translation, sentiment analysis and answer generation via cultural transcreation. Results for each module vary but overall for machine translation specialized open source models such as NLLB are behind closed solutions using Google, DeepL and OpenAI's GPT models for the bilingual chat data used as a benchmark. For sentiment analysis, approaches based on open source LLMs such as XLMR perform better than closed OpenAI GPT models. For answer generation via transcreation, only one closed model was leveraged (GPT4) and the results are quite positive. Results for MrMeeting indicate that the best approach to incorporate in the use case leverage OpenAI's GPT models and that open source LLMs such as LLaMA2 are close behind. Future work for all the streams points to the direction of exploring open source LLMs with the required modifications and improvements to achieve comparable or better performance than closed API solutions.

[12] https://github.com/utter-project/UTTER-MS9-meetingdata

References

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

Isabel Dias, Ricardo Rei, Patrícia Pereira, and Luisa Coheur. Towards a sentiment-aware conversational agent. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392488. doi: 10.1145/3514197.3549692. URL https://doi.org/10.1145/3514197.3549692.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.70.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, 2022.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.100. URL https://aclanthology.org/2022.naacl-main.100.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.57.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-1099.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

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

September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Konstantinos I. Roumeliotis and Nikolaos D. Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6), 2023. ISSN 1999-5903. doi: 10.3390/fi15060192. URL https://www.mdpi.com/1999-5903/15/6/192.

José G.C. de Souza, Ricardo Rei, Ana C. Farinha, Helena Moniz, and André F. T. Martins. QUARTZ: Quality-aware machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 315–316, Ghent, Belgium, June 2022. European Association for Machine Translation. URL https://aclanthology.org/2022.eamt-1.47.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D7.1 First prototype evaluation report