UTTER

Unified Transcription and Translation for Extended Reality
(UTTER)

Horizon Europe Research and Innovation Action
Number: 101070631
D7.3 – Third prototype evaluation report

Nature	Report	**Work Package**	WP7
Due Date	30/09/2025	**Submission Date**	30/09/2025
Main authors	Amin Farajian (UNB), Laurent Besacier (NAV)		
Co-authors	Thibaut Thonet (NAV)		
Reviewers	Barry Haddow		
Keywords	machine translation, grammatical error correction, cultural adaptation, LLM-chat		
Version Control			
v0.1	**Status**	Draft	18/09/2025
v1.0	**Status**	Final	31/09/2025

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 document summarizes the third evaluation of the prototypes for two use cases of the UTTER project, the customer service assistant and the meeting assistant. In this report we focus on the field trials that are a step beyond the main prototyping cycle of development and testing. Having all the features implemented in the prototypes, we measure the impact of the tools on the larger workflow and operation.

The first part of the report focuses on the evaluation of the customer service assistant, and the impact of the prototype on the conversations between the customer and the agent in a bilingual setting. The second part of the report focuses on the evaluation of the meeting assistant.

1 Evaluation of the customer service assistant use case

1.1 Third year evaluation

The goal of this use case is to develop a multilingual customer support assistant that enables human agents to provide support in any language. The assistant delivers suitable translations that consider both the conversational context and the cultural norms of the customer's language.

In the third year, we fully integrated the Tower+ models (Rei et al., 2025)—the latest generation of TowerLLM models developed within UTTER—into the demonstration system. This integration enabled a broader range of functionalities, including machine translation, grammatical error correction, cultural appropriateness detection and adaptation, and emotion recognition. Compared to the second prototype version, where Tower was applied solely for machine translation and grammatical error correction, Tower+ significantly extends the system's capabilities. In addition, we incorporated xCOMET (Guerreiro et al., 2024) to evaluate the quality of translations generated by Tower+. A video with a demonstration of the prototype is available.[1]

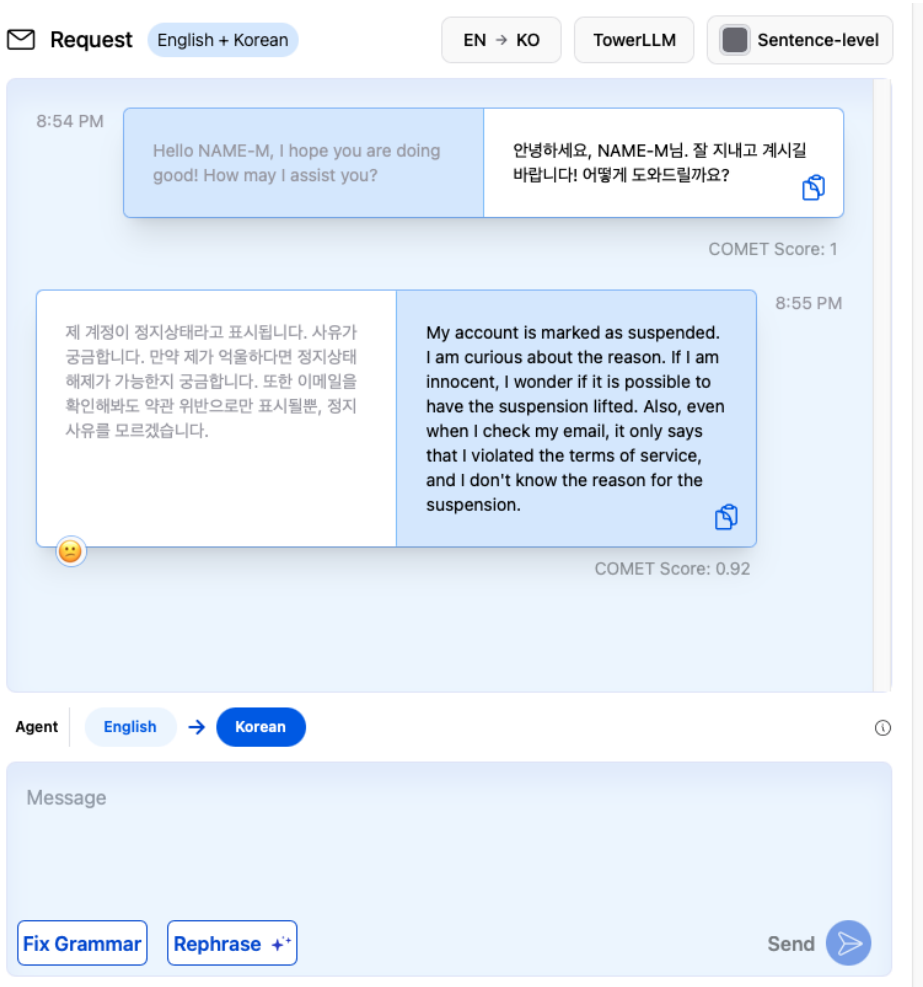

Figure 1: Customer service assistant 3rd year prototype.

[1] https://www.youtube.com/watch?v=UK_9BOgKLn4&t=215s

For evaluation, we report the results of the field tests in two language pairs: English-Korean and English-Portuguese. For each pair, we selected approximately 300 sentence pairs from the test sets of the MAIA corpus—introduced and used in the Chat Translation Shared Task (Mohammed et al., 2024)—and asked an in-house bilingual professional linguist to act as a customer service human agent in our prototype. In one round, the agent used the original English sentences provided in the dataset, while in another round she applied the prototype's cultural adaptation module (for English–Korean) or Grammar Error Correction module (for English–Portuguese) to enhance the source sentences before passing them to the translation step and to the customer. The following sections provide a more detailed analysis of the field test results.

1.2 Cultural Adaptation

The prototype is designed so that once the agent finishes writing a sentence, it automatically checks its appropriateness for the target language. If any issues are detected, the system displays a red warning, prompting the agent to rephrase the sentence before passing it to the translation step. Figure 2 shows a screenshot of the prototype in this scenario.

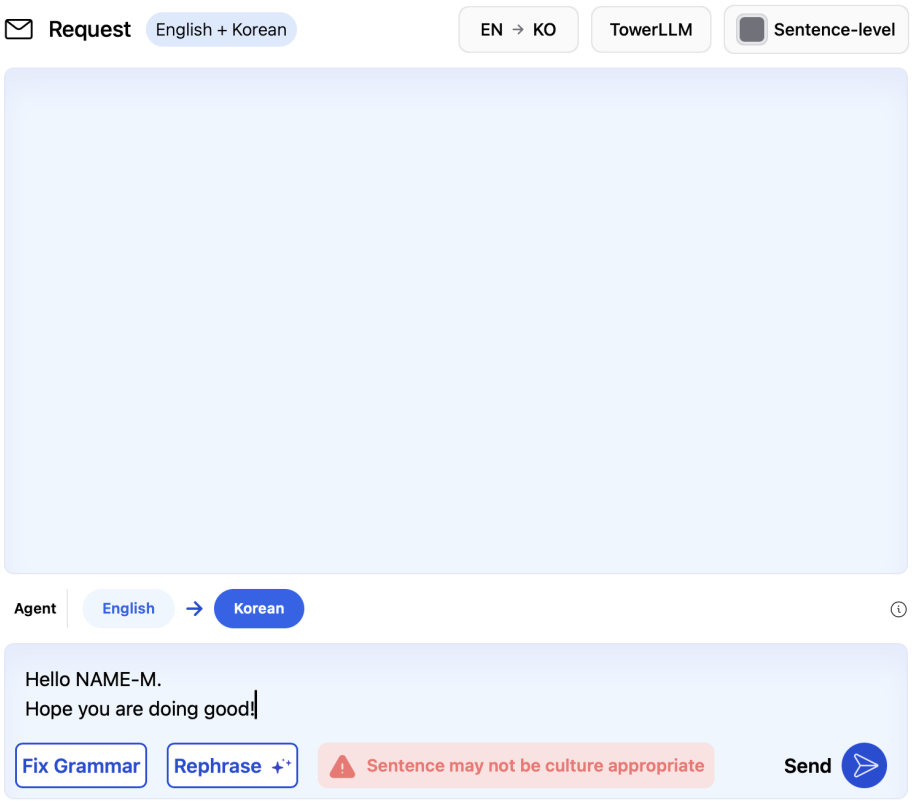

Figure 2: Prototype warning for culturally inappropriate sentences before translation.

For the English-Korean language pair, we selected 6 conversations containing 260 sentences, 160 of which were from the agent's side. Among these, Tower+ flagged 74 sentences as culturally inappropriate for a Korean audience. The analysis of our in-house professional linguist confirmed that in 81% of cases (60 out of 74) the rephrased versions were more culturally appropriate, while in 9% (7 out of 74) the original sentences were still better, and in 1% (1 out of 74) both versions were judged equally inappropriate.

When comparing translations of the original and rephrased source sentences, we observed that in 87% of the cases (52 out of the 60 rephrased cases, the translations were of equal or higher quality than those of the original sentences. Only 13% of cases (8 out of 60) showed a negative impact on translation quality.

The analysis conducted by our in-house professional linguist confirms the positive impact of cultural adaptation on the translation quality provided by Tower+. In these tests, customer sentences were kept unchanged and only the effect of cultural adaptation on the agent side was evaluated. This approach reflects real-world multilingual customer support settings, where the primary goal is to maximize customer satisfaction.

Given the sensitivity of the Korean audience to cultural nuances and the high proportion of culturally inappropriate agent sentences (46%) in this small sample, the importance of the cultural adaptation module in the pipeline becomes evident. By identifying and correcting such sentences before they are translated into Korean and reach the customer, the module helps ensure that communications are culturally appropriate, reducing the risk of misunderstandings and enhancing overall customer satisfaction in multilingual support interactions. It is worth noting that the original conversations of the MAIA corpus were sampled from real bilingual interactions between customers and agents in an actual customer support environment, highlighting that these issues are not just theoretical but occur in real exchanges between English-speaking agents and Korean customers.

1.3 Grammar Error Correction

To evaluate the impact of the GEC module, we conducted a field test using 333 sentences from 12 conversations sampled from the English–Portuguese portion of the MAIA corpus (Mohammed et al., 2024). Our in-house professional bilingual linguist acted as the agent in two settings: in the first, she used the original agent sentences from the MAIA corpus without any correction, while in the second, the sentences were processed through the GEC module[2] before being passed to the MT step. This setup allowed us to directly compare the quality of translations with and without grammatical corrections, providing a clear measure of the effectiveness of the GEC module. By simulating real-world customer support interactions, we were able to assess not only improvements in grammatical accuracy, but also the resulting impact on overall translation fluency, readability, and naturalness of communication between English-speaking agents and Portuguese-speaking customers. It is worth noting that, unlike the cultural appropriateness detection, which is applied automatically to all messages, grammar correction is performed on demand when the agent clicks the "*Fix Grammar*" button.

Out of 333 sentences, 168 were from the agent's side. Because grammar error detection is not automated in our prototype, we asked our in-house professional linguist to manually apply the GEC module to all agent sentences. This process resulted in 90 sentences being modified by Tower+, with changes varying in magnitude: some involved minor edits, such as adding or removing punctuation, while others addressed more substantial grammatical issues, including corrections to verb tense, sentence structure, and overall syntactic clarity. These modifications were aimed at improving the readability and fluency of the sentences before they were passed to the MT step, ensuring higher-quality translations for the end customer.

Of the 90 sentences modified by Tower+, 78 (i.e. 87%) were assessed as having improved quality compared to the original source sentences, demonstrating clear benefits from the GEC module. In

[2] The button that activates this module in the prototype's user interface is labeled "*Fix Grammar*".

9% of cases (8 out of 90), the modifications neither improved nor degraded the sentence quality. Only in 4% of instances (4 out of 90) the original source sentences were judged to be better than the modified versions.

Among the 78 sentences where the modified version was considered an improvement over the original, 41% (i.e. 32 out of 78) produced translations that were judged to be of higher quality than those of the original sentences, reflecting a clear positive effect of the GEC module on translation accuracy and fluency. In 42% of cases (33 out of 78), the translations of both the original and modified sentences were deemed equally good, indicating that the modifications did not negatively or positively affect the translation outcome. Interestingly, in 17% of cases (13 out of 78), the translation of the modified source was rated lower than that of the original sentence, suggesting that certain grammatical changes, while improving the source sentence, could occasionally lead to less optimal translations due to interactions with the MT system. This highlights the complexity of balancing source sentence correction with downstream translation quality.

While we observed positive impacts from applying GEC to the source sentences, its effect on the final translation was smaller compared to the cultural adaptation module. For this reason, we decided not to apply grammar error correction automatically to source sentences, leaving the decision to use GEC entirely to the agent. One possible explanation for this lower impact on translation quality is the robustness of LLMs to noisy inputs (Thonet et al., 2024; Peters and Martins, 2024). Even when the source sentences contain grammatical errors, the models are often capable of producing reasonable translations, which in turn reduces the added value of applying GEC before the translation step.

1.4 Conclusion

In this simulated field test, we evaluated the impact of the cultural adaptation module and the grammar error correction (GEC) module on the quality of translations delivered in multilingual customer support scenarios. The results highlight that cultural adaptation plays a particularly important role in ensuring that agent messages are appropriate and resonate with customers, thereby directly contributing to customer satisfaction. Grammar correction, while useful for improving readability and grammatical accuracy, showed a more limited effect on translation quality due to the robustness of modern MT systems to minor errors. Overall, the findings underscore the importance of the developed modules in enhancing the quality of customer support interactions across languages.

2 Evaluation of the meeting assistant use case

2.1 Meeting assistant prototype: a staged vision

We have developed the third prototype of our UTTER meeting assistant, also known as Trust Mediator *TM*. This prototype is showcased in greater detail in a YouTube presentation.[3]

While in the first year we focused on building a general-purpose, LLM-powered assistant designed to answer meeting-related questions in a friendly and informal context, this prototype was not yet ready for deployment. Real-world applications pose significant challenges, such as handling

[3] https://www.youtube.com/watch?v=W2GFtgXpIDc

questions that fall beyond the assistant's intended scope, that may be inappropriate, or that are phrased ambiguously. We have therefore defined the following staged vision for our meeting assistant for the remainder of the project:

Year 1. A general-purpose, LLM-powered assistant, customized to answer meeting questions in a friendly, cooperative setting.

Year 2. A production-grade meeting assistant for the real, less-amicable world: robust to noisy inputs, equipped with input filters, and able to verify compliance with stated "principles".

This year. A trustworthy-by-design assistant ready to go public— based on our *Trust Mediator (TM) framework*.

This section reports on two types of evaluation made during the third year of the project:

(a) We continued to evaluate the accuracy of recent LLMs on our meeting assistant benchmark. For this we organized a shared task on this topic (see Section 2.2, *Benchmarking LLMs for the Meeting Assistant Use Case*).

(b) In parallel, we conducted a user study to evaluate our *TrustMediator* framework, designed for developing a chatbot that is trustworthy by design (see Section 2.3, *TrustMediator Evaluation*).

2.2 Benchmarking LLMs for the Meeting Assistant Use Case

We co-organized the third edition of AutoMin, a shared task on automatic meeting summarization into minutes. The 2025 edition featured the main task of *minuting*—producing structured meeting minutes in English and Czech for both project meetings and European Parliament sessions—and a new task introduced by UTTER: question answering (QA) based on meeting transcripts. The QA task focused on project meetings and was offered in two settings: monolingual QA in English, and cross-lingual QA, where questions in Czech were answered from English meetings.

Participation in 2025 was more limited than in previous years (one team for minuting, two teams for QA). To ensure robust evaluation, we provided multiple baseline systems, enabling assessment of current large language models (LLMs) across both tasks. In this report, we describe only the task B (meeting QA), as it most directly aligns with the meeting assistant use case.

2.2.1 Task B: Meeting QA

Task B is new and consists of answering questions based on meeting transcripts.

The evaluation builds on the work of Thonet et al. (2024) who recently created ELITR-Bench, a collection of 271 manually crafted questions in English along with their corresponding ground-truth answers related to a subset of English meeting minutes from the ELITR Minuting Corpus. Subsequently, ELITR-Bench (i.e. the set of questions and reference answers, not the underlying English meeting transcripts) was manually translated to Czech.

In this challenge, the participants were asked to answer these questions either monolingually (English questions on English transcripts) or cross-lingually (Czech questions on English transcripts), using ASR-generated meeting transcripts.

Q: Who were the participants of the meeting?
A: [PERSON14], [PERSON10], [PERSON5], [PERSON9], [PERSON1], [PERSON11]

Q: What was the main purpose of this meeting?
A: Discuss and finalize the technical setup for a demo

Q (Conv version): How many scenarios were discussed?
Q (QA version): How many scenarios were discussed for the demo setup?
A: 3 (plans A, B and C)

Q (Conv version): Which scenario was chosen eventually?
Q (QA version): Which scenario was chosen eventually for the demo setup?
A: Plan C

Figure 3: An illustration of variants of ELITR-Bench questions for the QA and Conv versions. No mention of the version indicates that the same question is used for both versions. Differences underlined for presentation purposes.

Given that meetings tend to be lengthy (with an average one-hour meeting generating around 20,000 tokens), solving this task effectively requires large language models capable of handling long contexts.

2.2.2 Evaluation

The evaluation of question answering on meeting transcripts adheres to the methodology introduced in Thonet et al. (2024). The approach is based on large language models as automated judges to assess the quality of responses. Specifically, these models compare the system-generated answers to the human-crafted gold reference answer for each given query. Following Thonet et al. (2024); Kim et al. (2024), our LLM-as-a-judge is provided with a 10-point score rubric that details the expected quality criteria for each grade level in order to guide the LLM towards the relevant numeric score. The LLM adopted for the evaluation is GPT-4o.[4] This choice is motivated by the fact that GPT-4 was shown to be highly correlated with human evaluators in Thonet et al. (2024).

2.2.3 Dataset: ELITR-Bench

Task B leverages ELITR-Bench[5] – a benchmark for the evaluation of long-context LLMs on meeting transcripts. The meeting data used in this benchmark originally come from the ELITR Minuting Corpus.

For the cross-lingual variant, we used ELITR-Bench questions which we had professionally translated from English into Czech. This variant of the test set allows to evaluate the following setting: asking questions in Czech, locating answers in English meeting minutes, reporting answers in Czech and comparing them to the golden-truth Czech answer.

ELITR-Bench is available in two settings. In ELITR-Bench-QA, we designed for each meeting a set of stand-alone questions (along with their answers) that can be addressed solely based on the

[4] We specifically used the gpt-4o-2024-11-20 model.
[5] https://github.com/utter-project/ELITR-Bench

meeting transcript, without additional context. We also designed a modified ELITR-Bench-Conv version where questions are to be asked in sequence, in a pre-defined order within a conversation. In this setting, some of the questions contain pronominal references or ellipses, for which previous conversational context (i.e., previous questions and answers) must be used to answer properly (see figure 3). The two settings overlap greatly: of the total of 271 questions only 33 differ between the -QA and -Conv variant.

Note that for AutoMin 2025, only the -QA setting was used.

2.2.4 Participants' Submissions

The submissions for Task B consist of two systems which are detailed below. In addition to these systems, we include a comparison against baselines that simply involve prompting an LLM with the meeting transcript in its context, following the same approach as described in Thonet et al. (2024). The LLMs considered in these baselines are the following: GPT-4o,[6] LLaMA-3.2-3B-Instruct,[7] LLaMA-3.1-8B-Instruct,[8] Phi-4-Mini-Instruct,[9] Phi-3-Small-128k-Instruct.[10]

GETALP Team GETALP's system (Gonzalez-Saez et al., 2025) investigates the combination of Retrieval Augmented Generation (RAG) and Abstract Meaning Representation (AMR) techniques. The former consists in retrieving relevant passages from the meeting transcript while the latter builds a graph representation of the semantic components of the retrieved sentences. This results in three systems: RAG-only (the context only contains retrieved passages), AMR-only (the context only contains the sentences from the graph representations) and RAG+AMR (the context contains both retrieved passages and graph-based sentences). These approaches are extractive, requiring the prompted LLM – a LLaMA-3.1-8B-Instruct model – to provide the 1-2 most relevant sentences from the context. The different systems are validated on the monolingual (English-only) and cross-lingual (English-Czech) tasks.

HallucinationIndexes The HallucinationIndexes system (Katwe et al., 2025) adopts a reinforcement learning approach to reduce hallucinations in summarization. It introduces the Entity Hallucination Index (EHI) as a reward signal and is model-centric, domain-agnostic, and fine-tuned on BART to improve factual consistency. Importantly, it does not rely on meeting-specific preprocessing, such as speaker segmentation or dialogue structuring. In Task B, the base model employed was Flan-T5, which is constrained by a context window of 1024 tokens and may therefore struggle to process the lengthy ELITR-Bench transcripts.

2.2.5 Task B Results

The results of Task B are presented in Tables 1 and 2 for the monolingual and cross-lingual question-answering subtasks respectively. The scores are based on evaluations by an LLM-judge using a scoring rubric from 1 to 10 (with higher scores indicating better responses) following the

[6] We used the gpt-4o-2024-11-20 model.
[7] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[9] https://huggingface.co/microsoft/Phi-4-mini-instruct
[10] https://huggingface.co/microsoft/Phi-3-small-128k-instruct

Approach	Mean Score
baseline_gpt-4o-2024-11-20	7.74
baseline_llama-3.1-8B-instruct	7.08
baseline_phi-4-mini-instruct	6.84
baseline_phi-3-small-128k-instruct	6.65
baseline_llama-3.2-3B-instruct	6.33
GETALP@AutoMin	4.55
GETALP@AutoMin_amr	4.34
GETALP@AutoMin_amr_only	2.73
HallucinationIndexes@AutoMin	2.28

Table 1: Monolingual subtask (English-only) results for Task B. The reported performance corresponds to the average of scores within a 1–10 range (higher is better).

methodology described earlier. They represent the average over the scores obtained on the 130 questions of the ELITR-Bench test set.

For the monolingual subtask, our results confirm those of Thonet et al. (2024), who showed that LLMs such as GPT-4o and LLaMA-3.1/3.2 perform strongly on the meeting QA long-context task, with closed models still outperforming open ones. We also observe that the baseline based on Phi-4-Mini-Instruct obtained a highly competitive score, suggesting it to be a strong model in the 3-4B parameter class.

In contrast, the participants' submissions (*GETALP* and *HallucinationIndexes*) received significantly lower scores than our baselines, although *GETALP* outperformed *HallucinationIndexes*. We hypothesize that this is primarily due to the extractive nature of both systems, which may be disadvantaged by LLM-based evaluation methods. Moreover, extractive approaches are generally less effective than abstractive models when it comes to answering complex questions that require capturing nuanced information, understanding context, or performing reasoning. In the specific case of *HallucinationIndexes*, the lower performance is likely also due to limited handling of long-context inputs: the system is based on Flan-T5, which has a maximum context window of just 1,024 tokens, while meeting transcripts frequently exceed 10,000 tokens. If this limitation was not properly addressed (e.g., via chunking, summarization cascades, or retrieval mechanisms), it raises concerns about the reliability of the reported results, as the model would only process a small portion of the transcript, potentially overlooking key information necessary for accurate answers or summaries.

For the cross-lingual task, we observe that GPT-4o remains quite robust, achieving a score very close to that of the monolingual setting. In contrast, open models such as LLaMA and Phi experience a performance drop of more than one point in this more challenging cross-lingual scenario. Among the participants, only *GETALP* submitted a system for the cross-lingual task, and we similarly observe a notable degradation in performance compared to their monolingual results.

Approach	Mean Score
baseline_gpt-4o-2024-11-20	7.69
baseline_llama-3.1-8B-instruct	6.21
baseline_phi-4-mini-instruct	5.41
baseline_llama-3.2-3B-instruct	5.11
baseline_phi-3-small-128k-instruct	4.77
GETALP@AutoMin_amr	3.11
GETALP@AutoMin	2.85
GETALP@AutoMin_amr_only	2.15

Table 2: Cross-lingual subtask (English-Czech) results for Task B. The reported performance corresponds to the average of scores within a 0–10 range (higher is better).

2.3 TrustMediator Evaluation

The Trust Mediator is a principle-centered workbench designed to help service owners configure and govern LLM-powered chatbots. It enables the exploration of chatbot behavior, the elicitation and refinement of governing principles, and the detection of conflicts or overlaps among them. Recall that Trust Mediator is presented in more details in this YouTube presentation.[11] A key feature is persona-driven simulation, allowing chatbot responses to be tested across diverse user profiles and situational contexts. By making implicit norms explicit and linking abstract trust requirements to concrete interaction rules, the Trust Mediator bridges organizational expectations with operational chatbot behavior, thus supporting transparency, accountability, and trust calibration.

We conducted an exploratory between-subjects study ($n = 12$) to evaluate how different configurations of the Trust Mediator support the authoring and testing of chatbot governance principles. Participants were randomly assigned to one of two conditions: (A) a complete system with all LLM-assisted features (persona-based query generation, feedback-to-principle elicitation, principles conflict detection, and evaluation support), or (B) a simplified manual system without such assistance. The sessions lasted 60–75 minutes and involved reflective exploration, principle creation (3–5 principles), and interactive testing with the chatbot. Data collection combined self-reports, quantitative metrics of principle quality (specificity, coherence, coverage), and qualitative observations. Overall, both groups endorsed principles as a useful mechanism for evaluating chatbot behavior. Assisted participants produced more principles on average (4.8 vs. 3.8), with broader coverage across cognitive, emotional, and organizational dimensions, while manual participants achieved higher scores in specificity and coherence. Subjective ratings were favorable across conditions, although usefulness was rated higher in assisted mode (e.g., 4.5—-4.8 vs. 3.2–4.0). Interestingly, manual authoring showed a trend towards greater improvements in chatbot responses (mean gain 1.13 vs 0.64), reflecting the focus of participants on operational precision. Qualitative analyses reinforced this complementarity: assisted authoring fostered breadth and inclusivity, whereas manual workflows encouraged depth, ownership, and refinement. Taken together, the findings suggest that hybrid workflows—combining manual reflection with targeted assistance—may offer the strongest path toward robust and trustworthy chatbot governance. *This experiment is currently documented in*

[11] https://www.youtube.com/watch?v=W2GFtgXpIDc

a full paper that will be submitted to CHI 2026 ('Surfacing Governing Principles for Chatbots: A Workbench and Comparative Study').

3 Conclusion

This document describes the third prototypes evaluation built for the two use cases of the UTTER project, the Customer Service Assistant (Section 1) and the Meeting Assistant (Section 2). For the customer service assistant use case, the main finding is that cultural adaptation has a stronger impact than grammar correction on the quality of multilingual customer support translations. This prototype is showcased in greater detail in a YouTube presentation.[12]

For the meeting assistant use case, we organized a shared task (Automin 2025) to further evaluate the performance of current LLMs on meeting QA. In addition, we assessed our trustworthy-by-design assistant—built on the Trust Mediator (TM) framework.

[12] https://www.youtube.com/watch?v=UK_9BOgKLn4

References

Gabriela Gonzalez-Saez, Felix Herron, Diandra Fabre, Jeongwoo Kang, Markarit Vartampetian, and Yongxin Zhou. Getalp@automin 2025: Leveraging rag to answer questions based on meeting transcripts, 2025. URL https://arxiv.org/abs/2508.00476.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Praveenkumar Katwe, Rakesh Chandra Balabantaray, and Naman Kabadi. Reducing hallucinations in summarization via reinforcement learning with entity hallucination index, 2025. URL https://arxiv.org/abs/2507.22744.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing evaluation capability in language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Wafaa Mohammed, Sweta Agrawal, M. Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C. Farinha, and José G. C. de Souza. Findings of the wmt 2024 shared task on chat translation, 2024. URL https://arxiv.org/abs/2410.11624.

Ben Peters and André F. T. Martins. Did translation models get more robust without anyone even noticing?, 2024. URL https://arxiv.org/abs/2403.03923.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. Tower+: Bridging generality and translation specialization in multilingual llms, 2025. URL https://arxiv.org/abs/2506.17080.

Thibaut Thonet, Jos Rozen, and Laurent Besacier. Elitr-bench: A meeting assistant benchmark for long-context language models, 2024. URL https://arxiv.org/abs/2403.20262.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D7.3 Third prototype evaluation report