TransLoco: AI-Driven Real-Time Transcription, Translation, and Summarization

A self-hosted free-software conference solution

Simon Haller-Seeber[1] **and Samuel Frontull**[1]

[1] Leopold-Franzens-Universität Innsbruck

Abstract

Advanced AI models, such as Large Language Models (LLMs), require significant computational resources and are thus typically hosted on centralized cloud infrastructure. This implies data protection risks and dependencies on major platforms, such as those of tech monopolies. This paper presents TransLoco, a self-hosted, free-software framework that enables near real-time transcription and translation of audio streams locally on one's own devices. We discuss practical free-software tools and models, along with their hardware requirement specifications, and demonstrate how advanced AI applications can be deployed without any dependency on external resources. This allows for greater security, control, and full digital sovereignty. Moreover, as part of our effort to support linguistic diversity, we include the low-resource language Ladin and release an Italian-Ladin translation model optimized for fast inference. The framework was demonstrated and tested live at the Medien-Wissen-Bildung 2025 conference, confirming its setup and practical applicability in real-world scenarios. We made the complete source code publicly available at https://git.uibk.ac.at/informatik/iis/iis-projects/TransLoco.

Haller-Seeber, Simon, & Frontull, Samuel. (2026). TransLoco: AI-Driven Real-Time Transcription, Translation, and Summarization. A self-hosted free-software conference solution. In Klaus Rummler, Günther Pallaver, Petra Missomelius, Valentin Dander, & Oliver Leistert (Eds.), *Streifzüge an den Nahtstellen von Medien, Bildung und Philosophie* (2. erw. Aufl., Medien – Wissen – Bildung 2025, S. 397–418). Zürich: OAPublishing Collective. https://doi.org/10.21240/978-3-03978-164-5_21

KI-gestützte Transkription, Übersetzung und Zusammenfassung in Echtzeit. Eine selbstgehostete Konferenzlösung auf Basis freier Software

Zusammenfassung

Fortgeschrittene KI-Modelle, wie Large Language Models (LLMs), sind sehr rechenintensiv und werden daher in der Regel über zentralisierte Cloud-Dienste genutzt. Dies birgt jedoch Datenschutzrisiken und die Gefahr der Abhängigkeit von großen Plattformen. In diesem Beitrag wird TransLoco vorgestellt: ein freies Software-Framework, das die Transkription und Übersetzung von Audio-Streams in nahezu Echtzeit auf eigenen Geräten ermöglicht. Wir präsentieren freie Software-Tools und -Modelle sowie deren Hardware-Anforderungen und zeigen auf, wie sich fortschrittliche KI-Anwendungen ohne Abhängigkeit von externen Ressourcen betreiben lassen. Dies ermöglicht eine höhere Sicherheit, mehr Kontrolle und volle digitale Souveränität. Um die sprachliche Vielfalt zu unterstützen, haben wir die ressourcenschonende Sprache Ladinisch einbezogen und ein für eine schnelle Inferenz optimiertes Übersetzungsmodell für Italienisch-Ladinisch veröffentlicht. Das Framework wurde auf der Konferenz Medien-Wissen-Bildung 2025 vorgeführt und seine praktische Anwendbarkeit unter realen Bedingungen getestet. Der vollständige Quellcode wurde unter https://git.uibk.ac.at/informatik/iis/iis-projects/TransLoco öffentlich zugänglich gemacht.

1. Introduction

Modern Natural Language Processing (NLP) applications – such as machine translation, intelligent writing assistants, speech recognition, and speech synthesis – are integral to communication, education, and professional environments. AI systems excel at (i) identifying patterns, (ii) compressing large volumes of information, and (iii) performing repetitive tasks efficiently. These capabilities make them particularly well-suited for applications designed to support humans in processing data if strategically combined (Marcus & Davis, 2019). In our increasingly globalised and digital world, such tools are becoming indispensable in professional environments and everyday communication, education, and public services. Due to their computational demands, however, users are often forced to rely on external infrastructures or commercial proprietary platforms, which carry

risks associated with surveillance, data storage, and vendor lock-in. Commercial platforms routinely collect, store, and exploit user data as a core part of their business model, turning personal behavior into predictive and monetizable assets (Zuboff, 2018). In educational contexts, such dependencies can lead to platform dependencies, limiting institutional autonomy and adaptability (Haller, 2018), while also subjecting students and educators to extensive data collection and surveillance practices (Schneier, 2015). In addition, only selected languages are supported in these systems, and so-called low-resource languages are often overlooked. Due to the ever-increasing integration of these technologies in our everyday lives, smaller languages are being used less and less (Zaugg et al., 2022), which poses the risk of them being regarded as "second-class" languages. However, many of the tasks handled by commercial tools can also be performed locally using open-source models.

To address this digital marginalization and privacy concerns, we present TransLoco, a free, open, and self-hosted conference tool developed to support conference participants in understanding, following, and documenting presentations, with a particular focus on including Ladin, a low-resource language. By combining (i) a speech recognition system, (ii) a summarization model, and (iii) machine translation engines, the tool can provide real-time transcriptions of speech with live translations in multiple languages, as well as a history and concise summaries of the main topics. In this framework, the data is processed directly on local devices, eliminating reliance on external servers and/or cloud providers. Therefore, we make the following main contributions:

- documentation of how audio streams can be processed locally without external dependencies;
- demonstration that real-time speech processing can be performed locally, offering a viable alternative to cloud-based solutions in specific scenarios;
- inclusion of Ladin, a low-resource language, and release of a machine translation system optimized for fast inference in this language;
- public release of the source code.

By doing that, we showcase how individuals, institutions, and organizations can take back control of their data. Moreover, we encourage the adoption of decentralized alternatives that prioritize privacy, autonomy, and contribute to linguistic diversity and cultural preservation. This aligns with a broader vision of digital sovereignty, promoting the principles of data protection and free software. We hope to inspire others to disrupt the digital status quo of centralized, corporate-driven AI.

Background

The initial framework was developed for scientific outreach efforts at the Long Night of Research 2024[1], where it was designed to transcribe and translate a panel discussion into German and English. Lessons learned from the Media, Inclusion & AI Space within the INNALP Education Hub (Bouvier et al., 2025) informed our approach. While INNALP primarily focuses on educational tools for schools, the Long Night of Research emphasized public engagement. More broadly, self-hosted AI and free software tools offer valuable applications beyond classrooms, benefiting researchers, educators, and institutions seeking digital autonomy. Presented as a live demo at Medien-Wissen-Bildung 2025 conference[2], this work fits in with the conference's broader discussions on the intersections of media, education, and philosophy. It challenges the assumption that advanced AI solutions must be outsourced to major technology providers.

2. Related Work

The idea of combining multiple NLP technologies into a digital assistant with real-time transcriptions and summaries, as the one presented in this paper, is not entirely new. An early attempt to automatically generate meeting notes from audio using speech recognition and summarization models was presented in 2018 with ProMETheus (Liu et al., 2018). This work focused on supporting multiple speakers. However, it predates the emergence of LLMs, and language technologies have since advanced significantly. More recent efforts, such as

1 https://www.uibk.ac.at/de/mip/events/lange-nacht-der-forschung/.
2 https://www.uibk.ac.at/de/medien/veranstaltungen/tagungen/medien-wissen-bildung-2025/.

Khonde, et al. (2024), Wyawahare, et al. (2024), and Akshatha et al. (2024), have explored similar ideas for generating meeting minutes, but their architecture relies on proprietary services, including Amazon Transcribe for speech-to-text conversion and ChatGPT for summarization.

In contrast, TransLoco is designed to address these privacy and dependency concerns by leveraging open-source components and providing a complete infrastructure to run everything locally. In this regard, it shares similar goals with the open-source system proposed by Haz et al. (2024), which also prioritizes transparency and user control. Their system focuses on post-hoc analysis of recorded video presentations, integrating open-source components for speech recognition, summarization, and slide segmentation. While their approach is modular and effective in offline scenarios, our work targets real-time support with a lightweight and privacy-respecting architecture.

Our framework incorporates Ladin (Val Badia), one of many low-resource languages often overlooked in mainstream machine translation systems. Previous work has focused on developing and evaluating methods that enable the use of state-of-the-art neural models for machine translation for Ladin, more specifically, the Val Badia variant (Frontull & Moser, 2024a,b). Building on this foundation, our framework incorporates a newly trained model into the real-time translation pipeline to support Ladin.

3. Core NLP Tasks

In this section, we describe the state-of-the-art approach for the individual tasks in our framework.

3.1 Automatic Speech Recognition (ASR)

The process of ASR involves the interpretation of a time-varying acoustic signal using a stochastic model (Dighe et al., 2020). The first scientific approaches in this field can be traced back to the 1930s, when phonetic elements were first systematically described (Juang & Rabiner, 2005). With advances in computing power and the availability of larger amounts of data, these systematic approaches have been replaced by end-to-end methods (Graves & Jaitly, 2014),

where, again, the transformer architecture has become the dominant paradigm (Pratap et al., 2024). While these require less human effort, they require vast amounts of data (Prabhavalkar et al., 2023) and need dedicated infrastructure to operate.

3.2 Automatic Text Summarization (ATS)

ATS is the task of summarizing a text while retaining the most important information (Rennard et al., 2023). The literature distinguishes between *extractive* summarization, where the task is to select the most important sentences from a source text and concatenate them into a summary, *abstractive* summarization, which aims to capture the essential information of the original text and to generate new sentences that summarize it, and *hybrid* summarization that combines both approaches (Rennard et al., 2023). In our work, we focus on abstractive summarization, as it helps to smooth out transcription errors. As with many other NLP tasks, transformer-based architectures and generative models have become the foundation for abstractive summarization (Zhang et al., 2025a,b).

3.3 Machine Translation (MT)

Current state-of-the-art machine translation systems typically use transformer-based architectures, which provide significant gains compared to older statistical approaches (Koehn, 2020). In specific settings, e.g., document-level translation, LLMs have proven to be even more effective and can provide contextually more appropriate translations (Wang et al., 2023). However, LLMs are too computationally expensive when speed and responsiveness are critical. In our framework, we thus opted for (smaller) neural machine translation models. Neural MT models are trained on parallel data, i.e., source texts and their corresponding translations. From this data, the models learn input–output relationships as well as patterns in the target language. For real-time text-to-text translation from English to German and Italian, we used Argos Translate (Finlay, P.J. and Argos Translate, Contributors, 2025), and for translations from Italian to Ladin we trained and integrated a quantized model optimized for fast inference that we have publicly released on Hugging Face[3].

3 https://huggingface.co/sfrontull/transloco-ita-lld

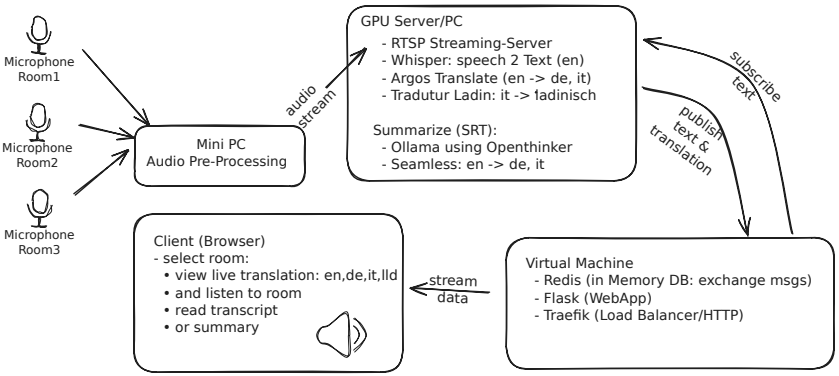

Figure 1: High-level overview of TransLoco (Haller-Seeber and Frontull, 2025).

4. Technical Implementation

This section outlines the technical implementation of TransLoco. The system architecture combines multiple components to ensure seamless processing of speech input, near real-time transcription, and translation. The integrated tools consist of free and open-source software released under an open license.

Figure 1 presents a high-level overview of the TransLoco framework with three main components: (a) audio client, (b) GPU Server, (c) web-client. To simplify the deployment and management of all services, the system uses containerization with Docker. For efficient communication between components, we use Redis[4] as an in-memory database. In the following, we describe the distinct components in more detail. The source code is available at: https://git.uibk.ac.at/informatik/iis/iis-projects/TransLoco

4 https://github.com/redis/redis.

4.1 Audio Processing

We use PulseAudio[5] and FFMPEG[6] for capturing audio from microphones and performing necessary pre-processing tasks and stream the audio to the streaming server using the real-time streaming protocol (RTSP). Mediamtx[7], a simple RTSP server, handles live audio streams.

When a client connects to the GPU machine via an RTSP audio stream, a dedicated docker container begins by performing Voice Activity Detection (VAD) (SileroTeam, 2021), which identifies the presence of human speech in the incoming audio stream. This step is crucial for minimizing errors and hallucinations in the transcription process by ensuring that the transcription model processes only relevant speech. Once speech is detected, the audio is transcribed into text. The transcribed text is then translated into the target languages (German, Italian, and Ladin). If a stream stops or is not available, the respective listening container restarts and waits for a new audio stream connection.

4.2 Transcription, Translation and Summarization

At the core of the system's capabilities, several AI models are integrated for transcription, translation, and summarization. The transcription of live speech is handled by Whisper Live (Radford et al., 2022), a near real-time transcription tool built upon OpenAI's Whisper model[8]. This model is highly efficient in transcribing both live audio input from microphones and pre-recorded audio files, adjusting the transcription output based on the most probable words and sentence structures.

For multilingual translation, TransLoco integrates different models based on the use case. For translating summaries and the transcription history (stored as srt-file) into German and Italian, it employs

5 https://www.freedesktop.org/wiki/Software/PulseAudio.
6 https://github.com/FFmpeg/FFmpeg.
7 https://github.com/bluenviron/mediamtx.
8 This model could handle translations as well, but we wanted to evaluate different models in one go.

SeamlessM4T (Gong & Veluri, 2024), a large-scale multilingual machine translation model known for its broad language coverage and high translation quality. In contrast, for real-time translation of live transcriptions, TransLoco uses smaller and faster models from argos-translate[9] , which allow low-latency inference. For all translations into Ladin (Val Badia variety), the system uses the transloco-lld-ita[10] model optimised specifically for fast inference.

The integration of Ollama[11], a framework for managing LLMs, further enhances the system with access to OpenThinker, which is based on the Qwen2.5 family of models, to provide summaries for the transcribed talks (OpenThoughtsTeam, 2025). We generate a rolling summary every two minutes. Each new summary builds on the previous one, incorporating the most recent 15 minutes of transcription. This iterative process allows the model to continuously refine its understanding of the ongoing session. However, it should be noted that this may result in older information gradually disappearing from the summary.

4.3 Web Application

The final transcriptions, translations, and summaries are displayed on a dynamic webpage, which provides users with live updates of the transcription, translation, and summarization process. This interface is powered by Flask[12], a lightweight web framework that ensures smooth user interaction and dynamic updates. Additionally, Traefik[13] is used as the entry point for accessing the system from the web.

9 https://github.com/argosopentech/argos-translate.
10 https://huggingface.co/sfrontull/transloco-ita-lld.
11 https://ollama.com.
12 https://flask.palletsprojects.com/en/stable/.
13 https://github.com/traefik/traefik.

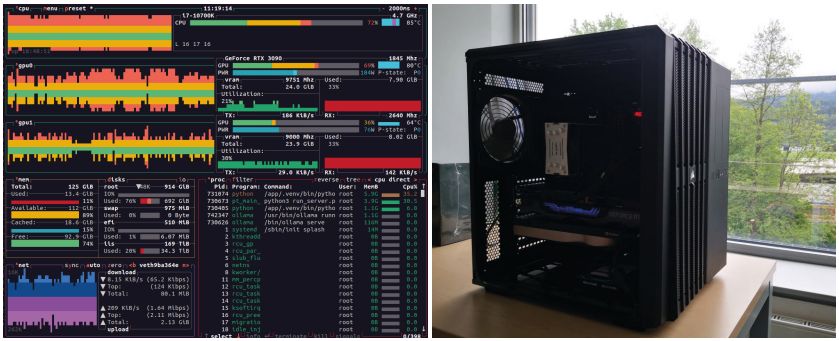

Figure 2: Left: GPU, CPU, Memory and Network utilization; Right: Workstation with Dual GPU Setup.

4.4 Hardware

To support live transcription, translation, and summarization, we deployed a high-performance workstation optimized for running LLMs in near real time. The workstation is built around an Intel Core i7-10700K processor running at up to 4.99 GHz across its 8 cores and 16 threads, and is equipped with 128 GB of RAM. It features two consumer-grade NVIDIA GPUs: a GeForce RTX 3090 and an RTX 4500 Ada Generation, each with 24 GB of VRAM, making the system well-suited for running medium-sized LLMs. Storage is provided by a high-performance 1 TB Samsung NVMe SSD. Network connectivity is supported by multiple Ethernet interfaces, including a high-speed Aquantia AQC107 controller. The system operates under moderate thermal and CPU loads, as shown on the left of Figure 2 via btop monitoring. The CPU is at approximately 72 % utilization and 85 °C, while the GPUs exhibit moderate usage levels, indicating available headroom for additional LLM requests.

5. Evaluation and Results

To assess the effectiveness of the proposed transcription framework, we conducted both manual and automated evaluations and monitored its live performance during the Medien-Wissen-Bildung 2025 conference. This combination of methods enabled us to evaluate the system's reliability, accuracy, and behavior under real-world conditions, using formal performance metrics/indicators and insights from user experience and observed user interactions.

5.1 Evaluation Metrics

Below, we describe the quantitative and qualitative metrics used in the automated evaluation.

Quantitative Metrics

Word Error Rate (WER) is the most common metric for ASR, quantifying transcription accuracy by counting substitutions, deletions, and insertions relative to the reference transcript (Morris, et al., 2004):

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is deletions, I is insertions, and N is the total number of words in the reference (Morris et al., 2004). The alignment was performed based on content similarity, not temporal alignment. This was implemented using the Python script jiwer[14] and custom alignment heuristics.

Sentence Error Rate (SER) measures the percentage of sentences with at least one error, providing a complementary perspective on transcription quality at the sentence level. We implement this using Python's difflib with case-insensitive comparisons.

Precision and Recall for Key Terms In domain-specific ASR tasks, accurately capturing critical terminology is crucial. We measure precision and recall over predefined keyword sets using Python set operations to evaluate this aspect.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

BLEU is a standard for automated MT evaluation, despite known limitations. It measures n-gram overlap between candidate and reference translations, adjusted by a brevity penalty to account for length differences (Papineni et al., 2002).

14 https://github.com/jitsi/jiwer.

Qualitative Metrics

Semantic Fidelity Evaluates whether the transcribed text preserves the intended meaning of the original. Semantic similarity is measured using sentence embeddings and cosine similarity across the entire corpus. For the computation of the embeddings, we used 'sentence-transformers' and 'scikit-learn' for cosine similarity.

Fluency and Grammar Assesses the grammaticality and readability of the transcription through LanguageTool an automated grammar-checking tool. The number and severity of errors are compared to the original reference text to quantify degradation.

Named Entity Recognition (NER) Accuracy Measures the correct identification and transcription of named entities and domain-specific expressions (e.g., "Big Tech", "LLMs") using SpaCy.

ASR Evaluation	**FLORES+**
Quantitative	
Word-Error-Rate	0.121
Sentence-Error-Rate	0.065
Precision	0.913
Recall	0.910
Qualitative	
Cosine Similarity	0.822
Grammar Errors (avg/sent.)	89/0.07
NER	0.661
MT Evaluation	**BLEU/TER**
English to German	31.88 / 51.16
English to Italian	26.07 / 45.79
Italian to Ladin	19.77 / 62.68

Table 1: Evaluation Results from microphone audio stream to English Transcription.

5.2 Automated Tests

To test the microphone setup and the transcription model, we conducted a trial in which selected English, German and Italian texts were synthesized into speech, recorded using the TransLoco client, and transcribed to English, replicating conditions similar to those of a live conference.

We use the English devtest split of the FLORES+ (NLLB Team, 2024) dataset comprising 1,012 sentences as a text source. For these texts, we first generated an MP3 audio file using eSpeak NG, a lightweight text-to-speech (TTS) engine (that does not rely on large language models). This audio was then played through a standard speaker and recorded using Hollyland Lark Max microphones. The setup was intended to mimic the acoustic conditions typical of a conference environment. To facilitate sentence-level segmentation during transcription, a 3-second pause was inserted between sentences. The resulting audio recordings were streamed into the TransLoco framework. Regardless of the input language, the system was configured to output English transcriptions.

The complete pipeline, from speech input to transcribed output, was evaluated by comparing the generated transcriptions with the original reference texts, using a range of quantitative and qualitative metrics to assess the accuracy and linguistic quality of the system. Moreover, to provide a general picture of the translation quality, we also evaluated the machine translation systems used to translate the English transcriptions into German and Italian, and from Italian into Ladin. All the results are reported in Table 1.

5.3 TransLoco at Medien – Wissen – Bildung 2025

Figure 3 is a screenshot of the system taken during the Medien-Wissen-Bildung 2025 conference, showing live transcription of the German talk given by Heribert Insam with the title “Medien, Mikroben und Kommunikation” (c.f. Insam, 2026) in Ladin. The right sidebar shows a summary. The system operated smoothly throughout the event, which confirms that the hardware setup was well-configured. However, the quality of the transcriptions and generated summaries remains far from perfect. For instance, the translation of the summary

into Ladin shown in Figure 3 is generally readable; however, it is not fluent. The German phrase “wesentlich weniger stringent“ was neither recognized nor translated into English and, consequently, was not translated correctly in any of the other target languages – a somewhat fitting accident. On a more positive note, the (rolling) summary displayed on the right-hand side of the screenshot performs reasonably well in this instance, capturing the key points of the talk so far.

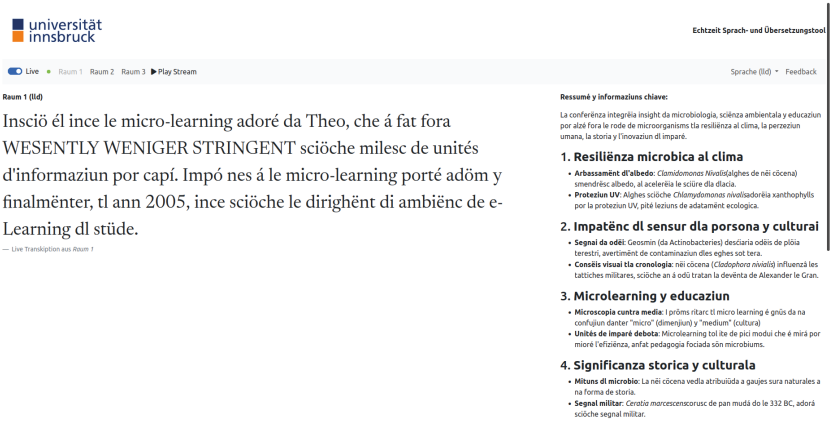

Figure 3: Screenshot of the webpage of TransLoco showing the live transcription and a summary in Ladin during the talk given by Heribert Insam.

6. Discussion

Based on the results of our evaluations and the experience gained during the live demonstration at Medien-Wissen-Bildung 2025 conference, we discuss the most important findings and observations, as well as areas where we see considerable potential for improvement in future work.

Real-time multilingual speech processing is feasible on local infrastructure: Our demonstration shows that it is feasible to run a real-time conferencing tool with live transcription, translation, and summarization entirely on local infrastructure. While we do not provide a direct benchmark against commercial platforms such as Microsoft Teams or Zoom, the system’s performance in tested scenarios indicates that locally-run solutions can offer a viable alternative for certain use cases, particularly where data privacy or offline capability is

a concern. Specifically, we demonstrated that the entire pipeline can operate with a configuration of just two modern GPUs. This configuration is sufficient and necessary to ensure that all models run at the required speed and responsiveness. During real-time use, the GPUs were continuously busy with transcription and translation, requiring resource allocation to support translation and summarization tasks in parallel.

We estimate the cost of setting up the required infrastructure for such a deployment to be around €5,000. This configuration provides a scalable, reusable solution for future events. The measured cost of operating the system during the conference was approximately €1 illustrating the potential ecological and economic benefits of a local deployment in this specific context. Comparisons with production-grade infrastructure or cloud-based services are more complex. In real-world scenarios, factors such as scalability, reliability, maintenance overhead, and long-term operational costs must be taken into account. Our system ran in a development environment, which, while sufficient for the event, does not fully reflect the demands or architecture of a production-ready, continuously operating (24/7/365) solution. Moreover, development infrastructure often needs to be maintained regardless of the chosen deployment model, which can blur the overall cost calculation. A feasible deployment approach depends on context-specific priorities, including privacy, performance, environmental impact, and operational control.

Sensitivity to Audio Conditions: The quality and reliability of the transcription, translations, and the generated summaries vary considerably and appear to be highly dependent on audio conditions, such as microphone connectivity and clarity of speech. During the first session of the conference, we observed that the setup of the audio stream server had a significant impact: two microphones were streaming from relatively distant positions, which resulted in a more scrambled audio signal. This led to a low transcription quality, which in turn affected the downstream translation and summarization. Additionally, we observed that during recognition and transcription,

the system occasionally alters individual word tokens, resulting in incomplete or fragmented sentences. Again, a problem that propagates through the translations.

Translation Accuracy is Domain-Dependent: While modern neural translation models can generalize to a certain extent, the domain of the training data plays a crucial role (Shen, et al., 2021). Consequently, one cannot always expect such models to produce accurate translations. Domain-specific terminology was occasionally misidentified; for example, OpenSource was incorrectly transcribed and summarized as OpenSync. This highlights limitations in the system's ability to accurately recognize named entities, as also reflected by a relatively low NER accuracy score of 0.661. This limitation has been explored in Walter et al. (2025), where the strengths and weaknesses of AI translation were examined, highlighting the need for human oversight. AI tend to simplify language for broader accessibility, thus it may also reduce textual complexity and depth. Maintaining high-quality translations requires careful attention to context, tone, and linguistic intricacies. A hybrid approach, combining AI efficiency with human expertise, ensures accuracy and preserves linguistic richness.

AI Summaries Sometimes Introduce Hallucinations: For AI summaries, we saw cases where the system tended to normalize – rather than capturing specific and nuanced content, the system sometimes produced generalized or overly abstract statements. In some instances, the system generated statements that were not part of the original talk, suggesting instances of content hallucination. These issues may be mitigated through more targeted prompt engineering, instruction tuning, and system prompt design. Notably, the AI summarizer sometimes went beyond summarization, providing a form of evaluative commentary or critical analysis. In a few instances, it generated remarks akin to constructive feedback. These findings underscore the need for critical oversight when interpreting automatically generated summaries, particularly in academic or professional contexts. In our case, the summaries were generated on a rolling basis. As a result, the summaries evolved dynamically over time. While not always precise, they provided an interesting glimpse into the ongoing

talks, including for those attending different sessions or unable to be physically present at the conference. Examples of AI-generated summaries produced by our framework at the Medien-Wissen-Bildung 2025 conference are available at https://hug-web.at/mwb2025/.

The Value of AI Translation Increases with the User's Linguistic Awareness: AI-powered translation tools such as ChatGPT and DeepL are becoming increasingly prevalent, offering speed and convenience for tasks like subtitle generation and product descriptions. However, Walter et al. (2025) from the Linguistics Department at the University of Innsbruck cautions that these tools are not flawless. While some errors are obvious – such as social media subtitle mistakes – others are more subtle, particularly in cultural nuances, legal terminology, or humor. Professional translators emphasize that AI-generated translations can lack contextual understanding, potentially altering meaning and intent. Users with strong linguistic awareness can recognize these subtle issues and correct them, thereby increasing the practical value of AI translation tools.

Based on the above observations, we identify several promising approaches for improving the overall performance of the system. One important possibility lies in the development of domain-specific machine translation systems that could be trained when the topic or context is known in advance, thereby improving translation (e.g., NER) accuracy. In addition, refining prompting strategies for summary generation could lead to more coherent results. In particular, experimenting with alternative approaches to generating rolling summaries could help ensure that important information is not lost over time.

7. Conclusion

In this paper, we presented TransLoco, a free-software conference tool that leverages advanced AI technologies running on local infrastructure. Our framework demonstrates that self-hosted AI solutions can enhance translation accuracy while preserving user control over data. By integrating multilingual and minority language support, we further promote inclusivity. This approach challenges the narrative

that AI translation must rely on centralized platforms, advocating instead for localized, free, and open alternatives that prioritize digital sovereignty and linguistic diversity.

Acknowledgments

We would like to thank Georg Moser and Justus Piater for their support throughout this project. On a more personal note, I (Simon) would like to express my deep gratitude to Theo, who has profoundly shaped my life. His support and encouragement have been instrumental in my development as a critical thinker. Theo, thank you for raising me with the values of independence and curiosity. I know I can always turn to you – whether for guidance or simply to share a moment of reflection – and I will, especially when I need it most.

References

Akshatha, P. S., Akash, B., Harshith, V., Sumukh, C., Gowardhan Reddy, V., & Shetty, Shravya. (2024). NLP-Driven Video Transcription: A Comprehensive Survey and Innovative Office Meeting Automation. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7. https://doi.org/10.1109/ICCCNT61001.2024.10724409

Bouvier, Fabian, Gleirscher, Lena, Haller-Seeber, Simon, Hug, Theo, Kaiserer, Madeleine, & Sonntag, Miriam. (2025). Innovative Lehr- und Lernansätze in Lernlaboren: Einblicke in den Media, Inclusion & AI Space des INNALP Education Hub. *Medienimpulse*, *63*. https://doi.org/10.21243/mi-01-25-24

Dighe, Pranay, Asaei, Afsaneh, & Bourlard, Hervé. (2020). On quantifying the quality of acoustic models in hybrid DNN-HMM ASR. *Speech Communication*, *119*, 24–35. https://doi.org/10.1016/j.specom.2020.03.001

El-Kassas, Wafaa S., Salama, Cherif R., Rafea, Ahmed A., & Mohamed, Hoda K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/10.1016/j.eswa.2020.113679

Finlay, P.J. and Argos Translate, Contributors. (2025). Argos Translate. In *Github Repository*. https://github.com/argosopentech/argos-translate

Frontull, Samuel, & Moser, Georg. (2024a). Rule-Based, Neural, and LLM Back-Translation: Comparative Insights from a Variant of Ladin. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, & Xiaobing Zhao (Eds), *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)* (pp. 128–138). Association for Computational Linguistics. https://aclanthology.org/2024.loresmt-1.13

Frontull, Samuel, & Moser, Georg. (2024b). Traduzione automatica "neurale" per il ladino della Val Badia. *Ladinia XLVIII (2024)*, 119–144. https://micura.it/upload-ladinia/files/700.pdf

Gong, Hongyu, & Veluri, Bandhav. (2024). *SeamlessExpressiveLM: Speech Language Model for Expressive Speech-to-Speech translation with Chain-of-Thought*. https://arxiv.org/abs/2405.20410

Graves, Alex, & Jaitly, Navdeep. (2014). Towards End-To-End Speech Recognition with Recurrent Neural Networks. In Eric P. Xing & Tony Jebara (Eds), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32, Issue 2, pp. 1764–1772). PMLR. https://proceedings.mlr.press/v32/graves14.html

Haller, Simon. (2018). Über "Bildung 4.0", "Schule 4.0" und andere Dinge, die keine Versionierung brauchen. *Medienimpulse*, *26*(1/2018). http://www.medienimpulse.at/articles/view/1187

Haller-Seeber, Simon, & Frontull, Samuel. (2025). TransLoco: AI-Driven Real-Time Transcription, Translation, and Summarization: SourceCode. In *Gitlab Repository*. University of Innsbruck. https://git.uibk.ac.at/informatik/iis/iis-projects/TransLoco

Haz, Amma Liesvarastranta, Panduman, Yohanes Yohanie Fridelin, Funabiki, Nobuo, Fajrianti, Evianita Dewi, & Sukaridhoto, Sritrusta. (2024). Fully Open-Source Meeting Minutes Generation Tool. *Future Internet*, *16*(11). https://doi.org/10.3390/fi16110429

Insam, Heribert. (2026). Medien, Mikroben und Kommunikation. In Klaus Rummler, Günther Pallaver, Petra Missomelius, Valentin Dander, & Oliver Leistert (Eds.), *Streifzüge an den Nahtstellen von Medien, Bildung und Philosophie* (2. erw. Aufl., Medien – Wissen – Bildung 2025, S. 303–316). Zürich: OAPublishing Collective. https://doi.org/10.21240/978-3-03978-164-5_16.

Juang, Biing-Hwang, & Rabiner, Lawrence R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*(67), 1. https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf

Kachhoria, Renu, Daga, Netal, Ramteke, Himanshu, Akotkar, Yash, & Ghule, Samarth. (2024). Minutes of Meeting Generation for Online Meetings Using NLP & ML Techniques. *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 1–6. https://doi.org/10.1109/ESCI59607.2024.10497256

Khonde, Kaushal Rajendra, Shah, Jaimeel, & Patel, Pratik. (2024). EchoSense AI Transcrib Using DevOps. *2024 Parul International Conference on Engineering and Technology (PICET)*, 1–5. https://doi.org/10.1109/PICET60765.2024.10716195

Koehn, Philipp. (2020). *Neural Machine Translation*. Cambridge University Press.

Lin, Chin-Yew. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013/

Liu, Hui, Wang, Xin, Wei, Yuheng, Shao, Wei, Liono, Jonathan, Salim, Flora D., Deng, Bo, & Du, Junzhao. (2018). ProMETheus: An Intelligent Mobile Voice Meeting Minutes System. *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 392–401. https://doi.org/10.1145/3286978.3286995

Marcus, Gary, & Davis, Ernest. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

Morris, Andrew Cameron, Maier, Viktoria, & Green, Phil. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. *Interspeech 2004*, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Muppidi, Satish, Kandi, Jayanthi, Kondaka, Bhavani Sankar, Kethireddy, Chakradhar, & Kandregula, Sai Eswar. (2023). Automatic meeting minutes generation using Natural Language processing. *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, 1–7. https://doi.org/10.1109/EASCT59475.2023.10393102

NLLB Team. (2024). Scaling neural machine translation to 200 languages. *Nature*, *630*(8018), 841–846. https://doi.org/10.1038/s41586-024-07335-x

OllamaTeam. (2024). *Ollama: A Framework for Running Large Language Models Locally*. https://github.com/ollama/ollama

OpenThoughtsTeam. (2025, January). *Open Thoughts*. https://openthoughts.ai

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Prabhavalkar, Rohit, Hori, Takaaki, Sainath, Tara N., Schlüter, Ralf, & Watanabe, Shinji. (2023). End-to-End Speech Recognition: A Survey. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, *32*, 325–351. https://doi.org/10.1109/TASLP.2023.3328283

Pratap, Vineel, Tjandra, Andros, Shi, Bowen, Tomasello, Paden, Babu, Arun, Kundu, Sayani, Elkahky, Ali, Ni, Zhaoheng, Vyas, Apoorv, Fazel-Zarandi, Maryam, Baevski, Alexei, Adi, Yossi, Zhang, Xiaohui, Hsu, Wei-Ning, Conneau, Alexis, & Auli, Michael. (2024). Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, *25*(1).

Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavey, Christine, & Sutskever, Ilya. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv Preprint arXiv:2212.04356*. https://arxiv.org/abs/2212.04356

Rennard, Virgile, Shang, Guokan, Hunter, Julie, & Vazirgiannis, Michalis. (2023). Abstractive Meeting Summarization: A Survey. *Transactions of the Association for Computational Linguistics*, *11*, 861–884. https://doi.org/10.1162/tacl_a_00578

Schneier, Bruce. (2015). *Data and Goliath: The Hidden Battles to Capture Your Data and Control Your World* (1st edn). W. W. Norton & Company.

Shen, Jiajun, Chen, Peng-Jen, Le, Matthew, He, Junxian, Gu, Jiatao, Ott, Myle, Auli, Michael, & Ranzato, Marc'Aurelio. (2021). The Source-Target Domain Mismatch Problem in Machine Translation. In Paola Merlo, Jorg Tiedemann, & Reut Tsarfaty (Eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1519–1533). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.130

SileroTeam. (2021). Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. In *GitHub repository*. GitHub.

Walter, Katharina, Mayer, Martina, Prandi, Bianca, & Agnetta, Marco. (2025, March). *Doku: Übersetzen im Zeitalter von KI - droht sprachliche Verarmung?* University of Innsbruck, Department of Translation Studies. https://www.youtube.com/watch?v=k5_uu1a1hgI

Wang, Longyue, Lyu, Chenyang, Ji, Tianbo, Zhang, Zhirui, Yu, Dian, Shi, Shuming, & Tu, Zhaopeng. (2023). Document-Level Machine Translation with Large Language Models. In Houda Bouamor, Juan Pino, & Kalika Bali (Eds), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 16646–16661). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.1036

Woodard, J. P., & Nelson, J. T. (1982). An information theoretic measure of speech recognition performance. *Workshop on Standardisation for Speech I/O Technology, Naval Air Development Center, Warminster, PA*.

Wyawahare, Medha, Shelke, Madhuri, Bhorge, Siddharth, & Agrawal, Rohit. (2024). AI Powered Multilingual Meeting Summarization. *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 86–91. https://doi.org/10.1109/Confluence60223.2024.10463307

Zaugg, Isabelle A., Hossain, Anushah, & Molloy, Brendan. (2022). Digitally-disadvantaged languages. *Internet Policy Review*, *11*(2), 1–11. https://doi.org/10.14763/2022.2.1654

Zhang, Haopeng, Yu, Philip S., & Zhang, Jiawei. (2025). A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. *ACM Comput. Surv.* https://doi.org/10.1145/3731445

Zhang, Yang, Jin, Hanlei, Meng, Dan, Wang, Jun, & Tan, Jinghua. (2025). *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. https://arxiv.org/abs/2403.02901

Zuboff, Shoshana. (2018). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st edn). PublicAffairs.