Set of Guidelines for the Development of a Web-Based Assistant Focused on the Creation of Dubbing Scripts

Renzo Manuel Ramos Ramirez[1], Orlando Arturo Roca Huapaya[1] y Eduardo Díaz[1]

[1] Universidad Peruana de Ciencias Aplicadas, Prolongación Primavera 2390, Lima 15023 – Perú
U202113745@upc.edu.pe, U201919742@upc.edu.pe, pcsijord@upc.edu.pe

Abstract. Dubbing studios in Latin America face challenges in creating translated scripts due to linguistic inconsistencies, limited cultural adaptation, and extensive manual processes. This article proposes ten guidelines for designing a web application for script preparation in dubbing projects. The guidelines are based on 31 scientific articles and are organized into two categories: (i) Guidelines for interaction design and user experience, and (ii) Guidelines for the architecture and technologies of the web application. To provide further clarity, an illustrative example of the proposed guidelines is presented. Additionally, an experiment was conducted with 20 participants to measure satisfaction with the web application designed according to the proposed guidelines, yielding positive results. This research may be of interest to software designers, dubbing studios, and audiovisual localization researchers seeking to improve productivity and narrative quality in dubbing workflows.

Keywords: Web application, generative artificial intelligence, translated scripts, automatic transcription, machine translation, web assistant.

1 Introduction

Dubbing script preparation faces recurrent bottlenecks in linguistic adaptation, cultural adjustments, lip-sync, and narrative coherence, which extend timelines even when translation quality is maintained. Illustrative cases include Avatar (three months to resolve jargon and colloquialisms) [1] and Maya the Bee: The Honey Games (better synchronization and cohesion in Peruvian Spanish than in European Spanish, with longer production) [2]. A further factor is the lack of uniform standards in some markets, for example the English-speaking one, which complicates lip-sync, prosody, and fluency [48]. Current AI tools are insufficient for these workflows: Heygen speeds multilingual video with lip-sync but limits video length and lacks cultural adaptation [3], while Vink uses Whisper for diarization and transcription but still struggles with synchronization and adaptation [4].

This work proposes ten guidelines derived from 31 studies. Interaction/UX: text-based conversational interfaces, color/contrast accessibility, mixed filtering and search, rich-text editing, adjustable-speed video playback. Architecture/technology: LLM integration, speech-to-text, diarization, Transformer-based MT, AI-assisted translation.

Unlike general HCI guidance for editing or subtitling, the proposal targets dubbing needs such as character-based timecodes, lip-sync, and culturally appropriate adaptation for Latin America, making it applicable to studios in emerging markets. The paper includes an illustrative implementation and is organized as follows: Section 2 (state of the art), Section 3 (guidelines), Section 4 (example), Section 5 (early validation), and Section 6 (conclusions and future work).

2 State of Art

This section reviews prior work on the technologies needed to build a web assistant that transcribes and generates translations for dubbing scripts. We conducted a Targeted Literature Review (TLR), non-systematic and informative, to select the most relevant studies and reduce selection bias. Searches were run in Scopus using the query: ("speaker diarization" OR "speech segmentation" OR "speaker identification") AND ("language model" OR "large language model" OR "LLM" OR "specialized domain"). Sixty articles were initially retrieved. After applying inclusion criteria (use of speaker diarization for voice separation; application of LLMs in specialized domains) and exclusion criteria (no focus on accuracy, efficiency, or applicability; unrelated to transcription, dubbing, or specialized text generation), ten articles remained for detailed analysis. The final set was classified into two groups:

2.1 Speaker diarization technologies

Speaker diarization. It segments an audio stream by speaker to answer who spoke when without prior identities, and it supports transcription, turn-taking analysis, and discourse studies [46]. Recent work shows consistent gains across settings: Pan et al. introduce ATC-SD Net for noisy aeronautical radiotelephony, reporting 6.3% DER on ATCSPEECH and outperforming VBx [5]; Khoma et al. present a PyAnnote-based supervised system where clustering beats segment classification (15.52% vs. 20.14% DER) [6]; Lyu et al. build a multilingual real-time pipeline on Whisper with WDER of 2.68% for two speakers and 6.96% for three on political interviews [7]; Aung et al. enhance EEND with a multiscale decoder, reaching 4.37% DER and surpassing i-/x-vector baselines [8]; Xylogiannis et al. fuse acoustic and spatial cues, reducing DER by 2–3% on AVLAB and SpeaD-M3C [9]. Overall, the literature emphasizes neural architectures [5][8], multilingual robustness [7], and multisensor fusion [9]. Our project adopts these insights to design a computationally efficient diarization module suitable for web-based workflows and varied conversational contexts.

2.2 Large Language Models

Large Language Models (LLMs) are Transformer-based neural networks trained for next-token prediction and applied to generation, translation, and summarization, often outperforming traditional approaches [47]. Evidence from machine translation research shows that LLMs can approach offline quality in simultaneous settings through adaptive read–write training and supervised fine-tuning data design [10]. Idiom handling

improves when aligning figurative meaning via Semantic Idiom Alignment and LLM-based Idiom Alignment, with human evaluations favoring semantic alignment [11]. Chinese idiom translation remains challenging, and even top systems still make substantial errors while common metrics correlate poorly with human judgments [12]. For low-resource Indian languages, evaluations using sentiment and semantic analyses compare GPT-4o, Gemini, and Google Translate against expert references to assess quality [13]. Iterative Bilingual Understanding further refines translations by generating and improving bilingual context, yielding measurable COMET gains across domains [14] . Taken together, these results indicate strong applicability of LLMs to translation, including real-time and idiomatic use cases, while highlighting the need for robust evaluation and control strategies [10–14].

3 Definition of the Proposed Guidelines

This section presents a set of 10 guidelines derived from the literature review for designing a web application for the preparation of dubbing scripts. The guidelines are based on the analysis of 31 scientific articles. Each guideline is supported by scientific studies that describe the necessary considerations for designing a web application for dubbing script preparation. The analyzed studies address aspects such as interface design, the use of services like diarization, artificial intelligence, Transformers, and Large Language Models (LLMs). These guidelines are grouped into two main categories: (i) Guidelines for interaction design and user experience, and (ii) Guidelines for the architecture and technologies of the web application. Each guideline is identified with the prefix “G” followed by a number (for example, G1, G2, G3, etc.).

3.1 Guidelines for Interaction Design and User Experience

G1: Conversational Interfaces. Studies on text-based conversational interfaces for web assistants indicate measurable UX gains on complex tasks when designs provide clear structure, visual feedback, and accessible controls. Kuang et al. compare voice and text assistants in UX evaluations and report that text systems are perceived as more efficient, easier to use, and less intrusive in analytical contexts, supported by visual clarity and the option to review asynchronously [31]. Silva and Canedo, in a systematic review, identify dialog-style message structuring, visible interaction history, and upfront transparency about limitations as core factors [32]. Srinivas et al. examine proactive LLM-based assistants for assessment tasks in complex cognitive processes and find that proactivity, user control over interaction flow, and adjustable assistant behavior improve perceived usefulness, trust, and control [33]. In conclusion, use a persistent, text-based interface embedded in the workflow with clear visual feedback and simple turn-taking. Provide a visible, revisitable history, adjustable proactivity, and upfront transparency about limits. This enables reviewing, repeating, or modifying steps improving efficiency, trust, and perceived control.

G2: Colors and Text. Research on color schemes and visual accessibility in subtitling and dubbing contexts indicates that careful theme and contrast choices improve readability, reduce eye strain, and support inclusive, long-duration work. Fan et al.

report that under negative polarity (dark mode) yellow on black yields the lowest visual fatigue, while red and some green hues increase discomfort (eye-tracking, blink rate) [34]. Szarkowska and Boczkowska find that color-coding subtitles by language does not harm comprehension and can increase immersion for certain viewers [35]. Cheng, Vahdat, and Lin introduce the Primary Color Difference metric and show that high-contrast text–background pairs, such as green on black or white, improve recognition efficiency and reduce cognitive load on LCDs, especially at higher contrast levels [36]. In conclusion, it is recommended that the web assistant for dubbing script creation implement a dark mode with text in high-contrast colors such as yellow, green, or white, avoiding colors like red that increase visual fatigue.

G3: Dynamic Filters and Classification Systems. Evidence shows that navigation improves when interfaces combine free-text search, hierarchical categories, and tags with clear interactions, immediate feedback, and flexible user control. Niu et al. report that faceted filters should expose relevant attributes, keep active selections visible at all times, and prevent combinations that yield empty results, users prefer filters that can be applied and removed intuitively without interrupting the workflow [37]. Melenhorst et al. find that user-generated tags support flexible, personalized classification when they are visible in the UI, allow multiple combinations, and align with natural language, which suits scene, emotion, or pending-edit tagging in script work [38]. Mahdi et al. review dynamic filter design and support real-time application, per-option result previews, and the pairing of hierarchies with free-form tags to accommodate different profiles and large editable datasets [39]. In conclusion, it is recommended that the web application for dubbing script creation implement a mixed filtering system composed of textual search, hierarchical categories, and customizable tags, allowing filters to be applied dynamically, displaying active filters at all times, and enabling easy removal.

G4: Text Editing. Studies on rich-text editors in collaborative settings show efficiency gains in writing and review when features such as anchored comments, version history, and change tracking are implemented with visual clarity, authorship control, and accessibility. Shulgina et al. report that comments attached to specific fragments significantly improve writing performance, whereas using tracked changes in isolation is less effective [40]. Das et al. design accessible collaboration techniques for visually impaired users, adding auditory representations of comments and suggested edits; in tests with 48 screen-reader users these functions improved understanding of changes, increased perceived participation, and reduced barriers [41]. Birnholtz and Ibara analyze the social dynamics of collaboration, finding that visible edits influence interpersonal relations, many contributors prefer to justify modifications through comments, and exposing authorship with controlled accept or reject actions supports coordination [42]. In conclusion, it is recommended that the web application for dubbing script creation implement a rich-text editor that allows the user to apply standard formatting (bold, italics, underline), insert comments anchored to specific script fragments, conduct reviews with visible change tracking, and access a version history with author identification.

G5: Video Player. Studies on web video players show efficiency gains in editing and dubbing when interfaces include speed control, subtitles, and tight script synchronization, implemented with intuitive and accessible controls. Christel et al. develop a player with synchronized transcription that lets users navigate via line-linked text; tests

show faster retrieval of relevant segments and better comprehension, with direct jumps from script to scene reducing cognitive load [43]. Valor Miró et al. find that simple, functional subtitle editors help users adjust auto-generated captions without technical overhead, cutting editing time by up to 70% compared to manual transcription [44]. Pantula compares accessible web players under WAI guidelines and reports that recent systems often integrate speed control, editable subtitles, and assistive-tech synchronization, improving experiences for users with visual or hearing impairments [45]. In conclusion, it is recommended that the web application for dubbing script creation implement a video player with adjustable playback speed, synchronized subtitle display, direct editing capabilities for the text, and script-based navigation. These functions should be accessible, intuitive, and provide clear feedback, allowing users to quickly move between scenes, review content with precision, and maintain an efficient dubbing-focused workflow.

3.2 Guidelines for the Architecture and Technologies of the Web Application

G6: LLM Integration. Studies on controlled, reliable LLM use in specialized applications point to three consistent practices. Woollaston et al. present TAMMY, an LLM-powered chatbot for EFL translation tasks, showing effectiveness when interaction flows are constrained, and the model operates at low temperature to limit variability [15]. Raunak et al. propose GPT-4 for automatic post-editing, demonstrating significant improvements over baseline MT while stressing the need for human oversight and careful prompt design to avoid hallucinations or unnecessary edits [16]. Lee et al. introduce a Modular Prompted Chatbot (MPC) that maintains long-term consistency by decomposing services into modules and applying techniques such as chain-of-thought and specialized prompting, removing the need for costly fine-tuning in many cases [17]. In conclusion, integrate LLMs through a modular, asynchronous architecture with the model as a decoupled service (API or queue) in a staged pipeline that includes reasoning, contextual prompting, and post-editing to enrich inputs and maintain coherence without fine-tuning. Use a middleware flow manager to orchestrate calls, adjust creativity and variability, and route domain-specific prompts.

G7: Automatic Speech-to-Text Transcription. Prior work supports a two-stage pipeline for spoken content: first transcribe, then translate. Zhang formalizes this separation, noting that structured text simplifies semantic, grammatical, and syntactic handling compared with operating directly on audio [18]. Siddique Latif et al. show that Transformer-based ASR outperforms RNN and CNN approaches, better managing variability, accents, and complex linguistic patterns required for high-quality transcripts [19]. Radford et al. present Whisper, a multilingual model trained on 680,000 hours that transcribes multiple languages without extra training and outperforms alternatives such as wav2vec 2.0 across diverse recordings, with near-human results in tests [20]. This evidence motivates adopting ASR as an essential first step in multilingual workflows, followed by MT on the transcribed text. In conclusion, automatic Speech-to-Text transcription is recommended as an essential step for capturing dialogues in multilingual contexts. Furthermore, it is suggested that the web assistant for dubbing script preparation implement a two-stage workflow: (i) transcription of audio using advanced models such as Whisper, and (ii) translation of the transcribed text.

G8: Integration of Diarization Services. Studies on integrating external diarization into web transcription systems point to joint or modular pipelines that handle multiple speakers and languages. Cheng et al. present a simultaneous ASR–diarization approach using Conformer, with information sharing across stages and Seq2Seq-TSVAD for parallel target-speaker activity detection, avoiding multiple passes [21]. Vachhani et al. use Whisper for transcription and language ID, adding ECAPA-TDNN accent detection, multi-window analysis, and DOVER-Lap fusion to improve multilingual accuracy [22]. Pappala et al. combine fine-tuned Whisper for Hindi/Marathi with Pyannote.audio for speaker segmentation in a reproducible, service-oriented pipeline that also accommodates sentiment analysis [23]. Collectively, these works support modular, decoupled integration of transcription, language detection, and diarization services exposed via APIs to scale across complex, multilingual audio. In conclusion, it is recommended to integrate automatic speaker diarization services in a modular and decoupled manner, employing existing tools such as Whisper (for transcription and language detection) and Pyannote.audio (for speaker segmentation). These services should be invoked through REST APIs or other asynchronous mechanisms such as message queues, facilitating their integration into a scalable architecture. This strategy ensures that the system can produce scripts correctly segmented by speaker, even in complex contexts with overlapping voices or language alternation.

G9: Transformer-Based Machine Translation. Reviews and evaluations highlight the centrality of Transformer models in MT and their advantages over statistical and RNN approaches, with better semantic context capture and fluency [24]. For dubbing, these properties help preserve rhythm, intent, and narrative cohesion in dialogue. Comparative studies show GPT-4 achieves accuracy comparable to junior human translators, though it can miss cultural and expressive nuances that still require human refinement [25]. Additional analyses report GPT-4 is competitive with commercial MT systems, especially when prompting is carefully designed [26]. In conclusion, it is recommended that the web assistant for dubbing script preparation implement a transformer-based machine translation system, considering the use of GPT-4 to generate high-quality initial translations and applying assisted post-editing to ensure semantic fidelity.

G10: Human-in-the-Loop Translation. Studies on AI-assisted MT with human intervention for professional, confidentiality-sensitive settings converge on hybrid workflows. Yuksel et al. combine LLMs with post-editing: the model proposes multiple hypotheses, scores them with COMETQE and GEMBA, and routes only critical segments to human editors; accepted fixes are reused via active learning [27]. Chatzitheodorou et al. design a secure pipeline with pseudo-anonymization before MT, human review on anonymized text, and reinsertion of original entities after validation, which is pertinent for unpublished scripts [28]. Han introduces HilMeMe, a human-centric MT metric focused on idioms and multi-word expressions, classifying errors by type and severity to guide fluent and adequate revisions [29]. Yang et al. present an LLM pipeline that generates a draft, then refines it with examples and feedback using in-context learning; past corrections are stored for reuse, and the system can select between draft and revision through model self-evaluation [30]. Collectively, these works support modular MT pipelines where the model drafts, humans correct targeted segments, and the system learns from edits while preserving data security. In conclusion, it is recommended

that the web application implement a Human-in-the-Loop-based machine translation workflow, where the AI generates a first draft of the script and this is refined through human intervention. To achieve this, it is suggested that the components responsible for translation, review, correction storage, and improvement suggestions be developed as decoupled microservices, invoked via REST APIs.

Table 1 summarizes the ten recommendations proposed for designing a web application for creating dubbing scripts.

Table 1. Summary of Recommendations

Category	Guidelines	Details
Guidelines for Interaction Design and User Experience	G1	Conversational interfaces
	G2	Colors and text
	G3	Dynamic filters and classification systems
	G4	Text editing
	G5	Video player
Guidelines for the Architecture and Technologies of the Web Application	G6	LLM integration (e.g., GPT-4)
	G7	Automatic speech-to-text transcription
	G8	Integration of diarization services
	G9	Transformer-based machine translation
	G10	Translation with human-in-the-loop

4 Illustrative Example of the Proposed Guidelines

This section presents an illustrative example of seven guidelines for designing a web application for dubbing script preparation, demonstrating how these guidelines enhance the process of creating dubbing scripts.

Fig 1. Assistant Conversational Interface with AI-Assisted Cultural Adaptation.

Figure 1 shows a chatbot responding to a request to translate a script fragment into Spanish. It illustrates G6 through the use of an LLM to handle user queries, G10 by involving the user in reviewing and refining the draft rather than accepting an automatic result, and G1 in the UI/UX through a text-based interface with preserved message history.

Fig 2. Interface Showing the Results of Transcription, Translation, and Diarization of Dubbing Scripts Produced by the Web Assistant.

Figure 2 presents two tables: the transcription (G7) and the translated script (G9). Both list "Timecode," "Character," and "Dialogue." The separation of entries is enabled by diarization (G8), which segments the audio, identifies speakers, and provides the corresponding durations.

5 Early validation

This section describes an early validation performed to verify the applicability of the proposed guidelines for the development of a dubbing web assistant. This prototype was created following key recommendations such as G1 (Conversational interfaces), G2 (Colors and text), G6 (Integration of LLMs), G7 (Automatic speech-to-text transcription), G8 (Integration of diarization services), G9 (Machine translation based on transformers), and G10 (Translation with human-in-the-loop). This first version implemented main functionalities like transcription, speaker diarization and translation. The validation involved twenty participants with professional experience in dubbing workflows, ranging from junior practitioners to professionals with moderate industry expertise, who performed predefined tasks using the web assistant. Each subject interacted with the prototype by uploading an English-language or French-language MP4 file and completing the process to obtain a translated and segmented script in Spanish. The system allowed participants to visualize and revise speaker-specific segments, edit translations, and export the script. When subjects finished solving the experimental cases, they completed a satisfaction survey. The satisfaction was measured online using a 5-point Likert scale questionnaire based on the framework developed by Moody's [49], which defined a framework (based on the work of Lindland et al. [50]) to evaluate satisfaction in terms of Perceived Ease to Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU). This framework has been previously validated and is widely used. The possible answers for each statement in the PEOU, PU, and ITU questionnaire are: Totally disagree, Fairly disagree, Neutral, Fairly agree, and Totally agree. A numerical value is provided to each statement from 1 (Totally disagree) to 5 (Totally agree). Six questions were defined to measure PEOU; the metric was calculated by adding the numerical values of the answers and classifying the result into a rank of five possible values: Rank 1–6: Totally disagree, Rank 7–12: Fairly disagree, Rank 13–18: Neutral, Rank 19–24: Fairly agree, Rank 25–30: Totally agree. For example, if a subject

answers 5 questions with Totally agree and 1 question with Neutral in PU, the result of this metric will be 28 (Totally agree). Six questions were defined to measure PU; the metric was calculated by adding the numerical values of the answers that each subject filled in through the six questions, and the result of this addition is classified into a rank with the five possible options: Rank 1–8: Totally disagree, Rank 9–16: Fairly disagree, Rank 17–24: Neutral, Rank 25–32: Fairly agree, Rank 33–40: Totally agree. Six questions were defined to measure ITU; the metric was calculated by adding the numerical values of the answers and classifying the result into a rank of five possible values: Rank 1–2: Totally disagree, Rank 3–4: Fairly disagree, Rank 5–6: Neutral, Rank 7–8: Fairly agree, Rank 9–10: Totally agree. Figure 4 shows the satisfaction results.

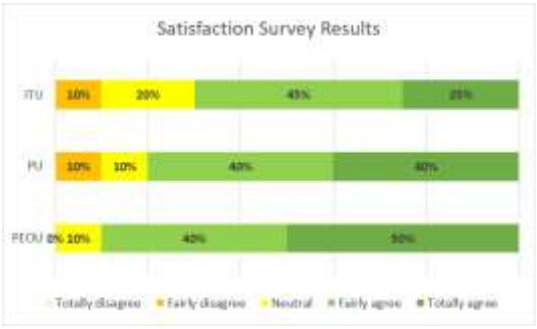

Fig. 4 Results of satisfaction

Figure 4 shows the distribution of subjects responses across of PEOU, PU, and ITU. A large majority of respondents selected “Fairly agree” or “Totally agree” in each category Specifically, 90% of subjects agreed that the system was easy to use (PEOU), 80% considered it useful for dubbing workflows (PU), and 70% expressed their intention to use it in future tasks (ITU). These results demonstrate encouraging early adoption potential and support the relevance of the proposed design guidelines.

6 Conclusions and future work

This work proposes ten guidelines for designing a web application for dubbing script preparation, derived from 31 studies. G1 covers conversational interfaces; G2 colors and text; G3 dynamic filters and classification; G4 rich-text editing; G5 the video player; G6 LLM integration; G7 speech-to-text transcription; G8 speaker diarization; G9 Transformer-based machine translation; and G10 AI-assisted translation with human oversight. We present an illustrative implementation that integrates G1, G2, G6, G7, G8, G9, and G10, showing how these guidelines improve the workflow. An early study with 20 participants reports positive satisfaction results. Future work includes adding guidelines grounded in new evidence and running broader empirical evaluations of effort and satisfaction in dubbing scenarios.

References

1. Cutando Plumed, A. (2023). Traducción de guiones y tiempos de producción en la industria audiovisual: un análisis del caso de la saga Avatar. https://hdl.handle.net/10234/203484
2. Guevara, L. & Quiroz, M. (2022). Comparación de versiones dobladas en distintos mercados: Un análisis de "Maya the Bee: The Honey Games" en español peruano y español peninsular. https://hdl.handle.net/20.500.12692/129174
3. Miquel-Vergés, J. (2024). Heygen: Soluciones de inteligencia artificial para la creación rápida de vídeos multilingües con sincronización labial. https://doi.org/10.31637/epsir-2024-358
4. Tolle, H., Castro, M. D. M., Wachinger, J., Putri, A. Z., Kempf, D., Denkinger, C. M., & McMahon, S. A. (2024). From voice to ink (Vink): development and assessment of an automated, free-of-charge transcription tool. BMC research notes, 17(1), 95. https://doi.org/10.1186/s13104-024-06749-0
5. Pan, W., Wang, Y., Zhang, Y., & Han, B. (2024). ATC-SD Net: Radiotelephone Communications Speaker Diarization Network. Aerospace, 11(7), 599. https://doi.org/10.3390/aerospace11070599
6. Khoma, V., Khoma, Y., Brydinskyi, V., & Konovalov, A. (2023). Development of supervised speaker diarization system based on the PyAnnote audio processing library. Sensors. https://www.mdpi.com/1424-8220/23/4/2082
7. Lyu, K.-M., Lyu, R., & Chang, H.-T. (2024). Real-time multilingual speech recognition and speaker diarization system based on Whisper segmentation. PeerJ Computer Science, 10, e1973. https://doi.org/10.7717/peerj-cs.1973
8. Myat Aye Aye Aung, Win Pa Pa, & Hay Mar Soe Naing. (2024). End-to-End Neural Diarization for Unknown Number of Speakers with Multi Scale Decoder. International Journal of Intelligent Engineering and Systems, 17(5), 870–881. https://www.doi.org/10.22266/ijies2024.1031.66
9. Xylogiannis, P., Vryzas, N., Vrysis, L., & Dimoulas, C. (2024). Multisensory Fusion for Unsupervised Spatiotemporal Speaker Diarization. Sensors, 24(13), 4229. https://doi.org/10.3390/s24134229
10. Fu, B., Liao, M., Fan, K., Li, C., Zhang, L., Chen, Y., & Shi, X. (2025). LLMs Can Achieve High-quality Simultaneous Machine Translation as Efficiently as Offline. https://doi.org/10.18653/v1/2025.findings-acl.1045
11. Donthi, S., Spencer, M., Patel, O., Doh, J. Y., Rodan, E., Zhu, K., & O'Brien, S. (2025). Improving LLM Abilities in Idiomatic Translation https://aclanthology.org/2025.loreslm-1.13/
12. Yang, C., Dou, Y., Heineman, D., Wu, X., & Xu, W. (2025). Evaluating LLMs on Chinese Idiom Translation. https://doi.org/10.48550/arXiv.2508.10421
13. Chandra, R., Chaudhari, A., & Rayavarapu, Y. (2025). An Evaluation of LLMs and Google Translate for Translation of Selected Indian Languages via Sentiment and Semantic Analyses. https://www.doi.org/10.1109/ACCESS.2025.3585629
14. Chen, A., Chen, K., Xiang, Y., Bai, X., Yang, M., Feng, Y., Zhao, T., & Zhang, M. (2025). LLM-based Translation Inference with Iterative Bilingual Understanding. https://doi.org/10.18653/v1/2025.findings-acl.867
15. Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024). TAMMY: Supporting EFL Translation Practice with an LLM-Powered Chatbot. International Conference on Computers in Education. https://doi.org/10.58459/icce.2024.4901
16. Raunak, V., Sharaf, A., Wang, Y., Awadallah, H. H., & Menezes, A. (2023). Leveraging GPT-4 for Automatic Translation Post-Editing. https://doi.org/https://doi.org/10.48550/arXiv.2305.14878

17. Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., & Lee, K. (2023). Prompted LLMs as Chatbot Modules for Long Open-domain Conversation. Findings of the Association for Computational Linguistics: ACL 2023, 4536–4554. https://doi.org/10.18653/v1/2023.findings-acl.277
18. Zhang, J. (2022). Bridging speech recognition and machine translation: Two-stage approaches for spoken language translation. IEEE Access, 10, 98765–98778. https://doi.org/10.1109/ACCESS.2022.9876543
19. Latif, S., Rana, R., Qadir, J., & Epps, J. (2020). Survey of deep learning techniques for automatic speech recognition. IEEE Access, 7, 19143–19165. https://doi.org/10.1109/ACCESS.2020.2968524
20. Radford, A., Jain, S., Mielke, J., Ramesh, A., Kim, J. W., & Sutskever, I. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. OpenAI. https://cdn.openai.com/papers/whisper.pdf
21. Cheng, M., Wang, W., Zhang, Y., Qin, X., & Li, M. (2022). Target-Speaker Voice Activity Detection via Sequence-to-Sequence Prediction. https://doi.org/10.1109/ICASSP49357.2023.10094752
22. Vachhani, B., Singh, D., & Lawyer, R. (2023). Multi-resolution Approach to Identification of Spoken Languages and To Improve Overall Language Diarization System Using Whisper Model. INTERSPEECH 2023, 1993–1997. https://doi.org/10.21437/Interspeech.2023-1354
23. Papala, G., Ransing, A., & Jain, P. (2023). Sentiment Analysis and Speaker Diarization in Hindi and Marathi Using using Finetuned Whisper. Scalable Computing: Practice and Experience, 24(4), 835–846. https://doi.org/10.12694/scpe.v24i4.2248
24. Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001
25. Yan, J., Yan, P., Chen, Y., Li, J., Zhu, X., & Zhang, Y. (2024). Benchmarking GPT-4 against Human Translators: A Comprehensive Evaluation Across Languages, Domains, and Expertise Levels. https://doi.org/10.48550/arXiv.2411.13775
26. Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. https://doi.org/10.48550/arXiv.2301.08745
27. Yuksel, K. A., Gunduz, A., Anees, A. B., & Sawaf, H. (2025). Efficient Machine Translation Corpus Generation: Integrating Human-in-the-Loop Post-Editing with Large Language Models. https://doi.org/10.48550/arXiv.2502.12755
28. Chatzitheodorou, K., Escrivá, M. Á. G., & Lacal, C. G. (2023, June). Machine translation of anonymized documents with human-in-the-loop. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation (pp. 429-436). https://aclanthology.org/2023.eamt-1.41/
29. Han, L. (2022). HilMeMe: A Human-in-the-Loop Machine Translation Evaluation Metric Looking into Multi-Word Expressions. https://doi.org/10.48550/arXiv.2211.05201
30. Yang, X., Zhan, R., Wong, D. F., Wu, J., & Chao, L. S. (2023). Human-in-the-loop Machine Translation with Large Language Model. https://doi.org/10.48550/arXiv.2310.08908
31. Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., & Shinohara, K. (2023, April). Collaboration with conversational AI assistants for UX evaluation: Questions and how to ask them (voice vs. text). In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-15). https://doi.org/10.1145/3544548.3581247
32. Silva, G. R. S., & Canedo, E. D. (2024). Towards user-centric guidelines for chatbot conversational design. International Journal of Human–Computer Interaction, 40(2), 98-120. https://doi.org/10.1080/10447318.2022.2118244
33. Srinivas, V., Xu, X., Liu, X., Ayush, K., Galatzer-Levy, I., Patel, S., ... & Althoff, T. (2025). Substance over Style: Evaluating Proactive Conversational Coaching Agents. https://doi.org/10.48550/arXiv.2503.19328

34. Fan, Q., Xie, J., Dong, Z., & Wang, Y. (2024). The Effect of Ambient Illumination and Text Color on Visual Fatigue under Negative Polarity. Sensors, 24(11), 3516. https://doi.org/10.3390/s24113516
35. Szarkowska, A., & Boczkowska, J. (2022). Colour coding subtitles in multilingual films – a reception study. Perspectives, 30(3), 520–536. https://doi.org/10.1080/0907676X.2020.1853186
36. Cheng, Z., Vahdat, V., & Lin, Y. (2018). A novel approach to study the effect of font and background color combinations on text recognition efficiency on LCDs. arXiv. https://doi.org/10.48550/arXiv.1812.08842
37. Niu, X., Fan, X., & Zhang, T. (2019). Understanding Faceted Search from Data Science and Human Factor Perspectives. ACM Transactions on Information Systems, 37(2), 14:1–14:27. https://doi.org/10.1145/3284101
38. Melenhorst, M., Hoekstra, R., & Koolen, M. (2018). Exploring User Tagging for Supporting Search in Streaming Video Services. Journal of the Association for Information Science and Technology, 69(4), 505–517. https://doi.org/10.1002/asi.23981
39. Mahdi, M. N., Ahmad, A. R., Natiq, H., Subhi, M. A., & Qassim, Q. S. (2021). Comprehensive Review and Future Research Directions on Dynamic Faceted Search. Applied Sciences, 11(17), 8113. https://doi.org/10.3390/app11178113
40. Shulgina, G., Costley, J., Shcheglova, I., & Zhang, H. (2024). Online peer editing: the influence of comments, tracked changes and perception of participation on students' writing performance. Smart Learning Environments, 11(1), 30. https://doi.org/10.1186/s40561-024-00315-8
41. Das, M., Piper, A. M., & Gergle, D. (2022). Design and Evaluation of Accessible Collaborative Writing Techniques for People with Vision Impairments. ACM Transactions on Computer-Human Interaction (TOCHI), 29(2), 1–42. https://doi.org/10.1145/3480169
42. Birnholtz, J. P., & Ibara, S. (2012). Tracking changes in collaborative writing: Edits, visibility and group maintenance. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW), 809–818. https://doi.org/10.1145/2145204.2145325
43. Christel, M. G., Stevens, S. M., Maher, B. S., & Richardson, J. (2010, October). Enhanced exploration of oral history archives through processed video and synchronized text transcripts. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1333-1342). https://doi.org/10.1145/1873951.1874215
44. Valor Miró, J., Silvestre-Cerdà, J. A., & Ferrer, M. A. (2015). Post-editing of automatic subtitles: Quality and usability evaluation. Speech Communication, 74, 61–71. https://doi.org/10.1016/j.specom.2015.09.006
45. Pantula, M. (2023). A study on media players from an accessibility perspective. International Journal of Computer Aided Engineering and Technology (IJCAET), Vol. 18, No. 4, 2023. https://doi.org/10.1504/IJCAET.2023.131920
46. Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, 101317. https://doi.org/10.1016/j.csl.2021.101317
47. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2). https://doi.org/10.48550/arXiv.2303.18223
48. Miggiani, G. S. (2021). English-language dubbing: challenges and quality standards of an emerging localisation trend. The Journal of Specialised Translation, (36), 2-25. https://doi.org/10.26034/cm.jostrans.2021.054
49. Moody, D. L. (2003). The method evaluation model: a theoretical model for validating information systems design methods. https://aisel.aisnet.org/ecis2003/79
50. Lindland OI, Sindre G, Solvberg A (1994) Understanding quality in conceptual modeling. IEEE Softw 11(2):42–49. https://doi.org/10.1109/52.268955