Artificial intelligence language technologies in multilingual healthcare: Grand challenges ahead

Vicent Briva-Iglesias
School of Applied Languages and Intercultural Studies (SALIS)
CTTS, ADAPT Centre
Dublin City University
vicent.brivaiglesias@dcu.ie

Abstract

AI language technologies (AILTs), increasingly enabled by large language models (LLMs), are becoming embedded in multilingual healthcare workflows for translation, rewriting, documentation, interpreting, and messaging in language-discordant settings. Yet fluent output is not the same as clinically safe or equitable communication: performance varies across languages, accents, tasks, and workflows, and efficiency gains can hide errors, reduce traceability, and shift responsibility across clinicians, translators, interpreters, and health systems. This narrative review synthesises recent peer-reviewed evidence across written communication, spoken communication, and emerging agentic workflows. Using the Human-Centered AI Language Technology (HCAILT) lens, it examines capabilities, evaluation practices, implementation patterns, and recurrent errors through reliability, safety culture, and trustworthiness. We identify key convergences and contradictions in the literature and propose seven grand challenges for the next phase of research and deployment. Progress, we argue, requires not only better models but also accountable sociotechnical design, calibrated human oversight, and stronger collaboration across MT/NLP, translation studies, HCI, clinical practice, implementation science, and policy.

1 Introduction

Multilingual healthcare is one of the clearest high-stakes environments in which AI language technologies can produce both meaningful benefit and meaningful harm. When clinicians and patients do not share a language, the consequences extend beyond inconvenience to comprehension, adherence, care continuity, and negative outcomes (Rawal et al., 2019; Lion et al., 2024). A recent systematic review and meta-analysis found that adult patients in language-discordant settings had higher odds of readmission and emergency department revisits, whereas access to verified interpretation attenuated those differences (Chu et al., 2024). Woods et al. (2022) likewise concluded that limited English proficiency is associated with poorer outcomes after hospital-based care, while van Lent et al. (2025) found that shared language and professional interpreters continue to outperform informal interpreting and most digital tools in complex care situations, as supported by extensive literature (Dew et al., 2018; Genovese et al., 2024; Valdez et al., 2025).

At the same time, pressure to adopt AI-powered language technologies is increasing. Recent generative AI (genAI) systems can translate text, simplify complex documents, transcribe and summarise conversations, draft responses, and increasingly operate inside broader workflows linked to patient portals, documentation systems, and electronic health records (Brown et al., 2020). However, apparent fluency is an unreliable proxy for safety. A system may produce highly readable discharge instructions while mishandling dosage information, perform well in a dominant language while degrading sharply in a minor one, or reduce documentation time while introducing inac-

curacies that are difficult to detect (Koenecke et al., 2024; Leung et al., 2025). For this reason, the central question is no longer whether these systems 'work' in a general sense, but for whom, in which language, for what task, under what workflow conditions, and with what consequences.

This paper addresses that question through a narrative review of recent peer-reviewed literature on AILTs in multilingual healthcare. A narrative review is appropriate because the evidence base is heterogeneous across modalities, user groups, tasks, settings, and outcomes, and because the field is evolving quickly enough that conceptual synthesis is as important as point-by-point benchmarking (Sukhera, 2022). The paper is intentionally user-centred in two senses. First, it focuses on the real users who encounter these systems: patients, clinicians, translators, interpreters, administrative staff, and healthcare organisations. Second, it examines the broader sociotechnical conditions that shape use, including interfaces, workflow design, escalation pathways, institutional accountability, and language inequity.

The paper has two linked aims. The first is descriptive: to synthesise the state of the art in multilingual healthcare AILTs across written communication, spoken communication, and emerging agentic workflows. The second is programmatic: to use that synthesis to articulate the major grand challenges that remain open for the next phase of research and deployment. In this sense, the paper is positioned not only as a healthcare AILTs review, but also as a contribution to the Translators and Users agenda in MT research: it asks how multilingual communication is being reconfigured by AI systems, what kinds of users are expected to rely on them, and what kinds of oversight, competencies, and institutional safeguards are required if these tools are to support rather than undermine safe care.

The review is selective rather than exhaustive. It focuses on recent peer-reviewed studies that examine AILTs used for multilingual healthcare communication across text, speech, and workflow orchestration, with priority given to work reporting empirical evaluation, implementation experience, or clinically relevant governance implications. The aim is not to catalogue every healthcare AILT paper involving language, but to identify the strongest current evidence, the main points of convergence and contradiction, and the unresolved problems that are most consequential for future research and deployment.

2 Conceptual framing

The grand-challenges tradition in HCI offers a useful way of structuring an interdisciplinary agenda for a field undergoing rapid technological change. Rather than listing isolated problems, grand-challenge papers identify cross-cutting tensions that span methods, stakeholders, and application domains. Stephanidis et al. (2019) framed seven HCI grand challenges around issues such as accessibility, ethics, privacy, security, and health. Their second revisit argued that these challenges had not receded, but had instead been intensified by AI, particularly in relation to transparency, value alignment, explainability, user control, and accountability (Stephanidis et al., 2025). Their most recent update reaches a similar conclusion: genAI does not replace earlier human-centred concerns, but intensifies them around human autonomy, operational safety under non-deterministic outputs, accountability, governance-by-design, and alignment with human cognitive processes (Winslow et al., 2026). That framing is especially relevant to multilingual healthcare because questions of trust, human control, and explainability become concrete in language-discordant clinical encounters. They concern who understands a diagnosis, who detects a mistranslation, who reviews a rewritten discharge instruction, and who takes responsibility when an AI-mediated communication failure affects care.

A grand-challenge framing is useful here for three reasons. First, the literature on multilingual healthcare AILTs is fragmented across MT/NLP, translation studies, health communication, implementation science, and HCI. Second, the empirical evidence is uneven: some applications already have promising deployment studies (Tu et al., 2025; Gallifant et al., 2025), while others remain early, speculative, or weakly evaluated in naturalistic tasks (Mellinger et al., 2025). Third, healthcare adoption depends on much more than model quality. It also depends on workflow design, training, monitoring, reporting, procurement, and institutional legitimacy. A grand-challenge framing helps keep those dimensions visible at the same time.

If grand challenges provide the paper's structure, the Human-Centered AI Language Technology (HCAILT) framework provides its analytical

lens (Briva-Iglesias and O'Brien, 2026). HCAILT adapts broader human-centred AI thinking to multilingual language technologies and foregrounds three linked pillars: reliability, safety culture, and trustworthiness. Reliability refers to whether a system performs consistently across languages, tasks, and conditions, and whether it is fit for a particular communicative purpose. Safety culture refers to the organisational practices that anticipate failure rather than assuming success, including monitoring, auditing, literacy, incident reporting, role clarity, and explicit escalation. Trustworthiness refers not to how much users happen to trust a system, but to whether that trust is warranted because the system's behaviour can be inspected, governed, challenged, and corrected. Recent work on medical AI trustworthiness reinforces this distinction, arguing that trustworthy systems require not only technical performance but institutional infrastructures for oversight and contestability (Goisauf et al., 2025; Shneiderman, 2020; van Leersum and Maathuis, 2025).

In this paper, AILTs in multilingual healthcare are defined broadly as "AI systems that process or generate language across text, speech, or multimodal inputs and outputs in ways that affect multilingual healthcare communication". The emphasis falls on three practical domains: written communication, spoken communication, and agentic workflows. The central concern is therefore not simply what has been built, but what kinds of communicative problems remain unsolved when these systems are examined through the combined lenses of UX, clinical risk, and multilingual inequity.

3 State of the art in AI language technologies for multilingual healthcare

This section reviews the current state of the art across three domains of use: written communication, spoken communication, and agentic workflows. Across all three domains, the literature points to the same broad pattern. Technical capability has improved rapidly, but performance remains highly task- and language-dependent, and the most consequential questions now concern workflow design, evaluation, governance, and human oversight rather than generation quality alone.

3.1 Written communication

Written communication remains the most mature and empirically developed area of AILT use in multilingual healthcare (Dew et al., 2018; Genovese et al., 2024). This is partly because text-based tasks map onto existing organisational bottlenecks: discharge instructions must be delivered quickly, educational materials need broad language coverage, and document workflows are easier to evaluate than live spoken encounters. At the same time, written communication is not a single task. It includes direct translation of clinical, specialised documents, translation of generic patient education and public-health materials, patient-facing information that needs to be understood, and hybrid workflows that combine automation with expert review.

3.1.1 Machine translation of patient-specific clinical documents

Earlier work already suggested that MT could support healthcare communication, while also stressing the limitations of general-purpose systems. Dew et al. (2018) identified promise for MT in health communication but called for stronger domain adaptation and more meaningful evaluation. Zeng-Treitler et al. (2010) similarly showed that access to translated medical content is not enough on its own if the resulting text remains difficult to understand. Herrera-Espejel and Rach (2023), reviewing public-health and epidemiological communication, argued that MT is becoming increasingly useful for outreach, but only when its limitations are understood and human validation remains central in higher-risk contexts.

Recent work on patient-specific discharge materials is more ambitious and more mixed. Ray et al. (2025) evaluated GPT-4o translations of personalised paediatric patient instructions into Spanish and found performance comparable to professional translation under an MQM-style framework. That is an important result, because it suggests that for a high-resource language and a constrained document type, contemporary LLM-based translation can approach professional-quality output. However, other studies complicate any simple parity narrative. Martos et al. (2025) compared AI-generated and professionally translated discharge instructions across Spanish, Chinese, Vietnamese, and Somali, and found that AI was non-inferior only for Spanish adequacy and error severity. Brewster et al. (2025) likewise reported substantial variation across Arabic, Armenian, Bengali, Chinese, Somali, and Spanish, with weaker performance in digitally underrepresented

languages.

Taken together, these studies suggest that the relevant question is not whether LLM translation works in healthcare in the abstract, but in which language pairs, for which document types, under what review conditions, and with what acceptable level of risk (Pym, 2025). They also point to an important methodological shift: evaluation is moving away from generic fluency scoring and toward clinically meaningful error categorisation. That shift matters because a minor stylistic awkwardness and a dosage omission do not have the same consequences. In multilingual healthcare, surface quality is therefore an insufficient basis for deployment decisions.

3.1.2 Post-editing translation workflows

A second major lesson from the written literature is that the most useful comparison is often not AI versus humans, but one workflow configuration versus another. Brewster et al. (2025) showed that what they call "human-in-the-loop" (hereafter, post-editing or "PE", since we think the HITL concept dehumanises the user) strategies can produce translations comparable to, or better than, professional translation alone while being substantially faster than a fully manual process. This is nothing new within the MT and Translation Studies communities (see, for example, Terribile (2024)), but may have not been known in the medical informatics field. These findings are especially relevant for healthcare implementation because they suggest that the central design question is not whether AI should replace professional language workers, but how automation should be positioned within a reviewed and accountable pipeline, and that the crucial stakeholders - MT/TS communities and medical informatics - are not aware of what the others are doing.

Lopez et al. (2025) make a related argument from an implementation perspective. Rather than focusing only on benchmark quality, they emphasise the organisational work required for safe deployment: integration with documentation and translation workflows, terminology control, auditability, clear lines of responsibility, and evaluation of patient comprehension rather than translation quality alone. This helps reconcile apparent contradictions across the literature. Studies such as Ray et al. (2025) show that raw model performance may already be strong enough to reduce manual effort in some high-resource settings. Studies such as Martos et al. (2025) and Brewster et al. (2025), however, show that residual risk remains substantial across the broader multilingual spectrum. Once differences in language set, task, and workflow are taken seriously, the broader conclusion becomes clearer: AI can already support safer and faster multilingual document production, but not yet in ways that justify removing expert oversight.

3.1.3 Plain language and rewriting

Plain-language transformation and rewriting are central to multilingual accessibility because accurate translation alone may still leave patients with text that is technically correct but difficult to understand (Hansen-Schirra and Maaß, 2020). Recent evidence suggests that generative AI can be useful here, although again under bounded conditions. Zaretsky et al. (2024) found that genAI could substantially improve readability and comprehensibility of discharge content, while also emphasising the need for better accuracy, completeness, and clinician review. Briva-Iglesias and Peñuelas-Gil (2025) also shared similar results in informed consent forms via automatic readability metrics, and Rust et al. (2025) reported similar improvements in cardiology discharge-summary simplification, though with limitations in personalisation.

Findings are also encouraging, though conditional, for broader informational materials. Ugas et al. (2025) found that human translation still outperformed MT across languages, even when MT often produced acceptable quality. Chen et al. (2025a), evaluating critical-care educational content in Mandarin, Spanish, and Ukrainian, also showed meaningful access gains alongside substantial platform- and language-dependent variation. McMinn et al. (2025) reported that a bespoke AI process could produce more readable first drafts of scientific plain-language summaries than medical-writer workflows alone.

These studies show that rewriting is a central component of equitable health communication. They also blur the boundary between translation and authorship, because these systems do not merely transfer content across languages, they also reshape it for particular users and contexts (Montalt-Resurrecció et al., 2024). In practice, this means that translation and plain-language rewriting should be treated as linked communicative tasks, both of which require context-aware review when clinical nuance or risk is involved.

3.1.4 Main learnings and pitfalls in written communication

Across the written literature, five lessons stand out. First, performance can already be very strong in high-resource languages and relatively constrained document types. Second, quality remains uneven across languages, especially for digitally underrepresented ones. Third, the strongest implementation evidence increasingly favours reviewed PE workflows over either raw automation or fully manual extremes. Fourth, readability and translation quality must be treated as distinct outcomes. Fifth, many of the hardest problems are organisational rather than purely technical, including terminology governance, workflow integration, quality assurance, and role allocation. This is where HCI methods should come into play.

The main pitfalls are equally consistent: clinically meaningful omissions, mistranslations hidden by fluent output, and evaluations that focus on text quality while neglecting patient comprehension and correct action. There is also a clear equity risk. If AI translation transforms service delivery in dominant languages while remaining unreliable in minor ones, multilingual healthcare may become more efficient and more unequal at the same time.

3.2 Spoken communication

If written communication is currently the most mature domain of AILT deployment, spoken communication is the most revealing one. Speech exposes the complexity of real-world multilingual healthcare: accent variation, code-switching, overlapping talk, noise, incomplete utterances, and emotionally charged exchanges. For that reason, the spoken literature is especially valuable for understanding end-to-end system risk.

3.2.1 Machine interpreting

Professional interpreting remains the benchmark for complex spoken multilingual care. van Lent et al. (2025) make this clear: shared language and professional interpreters generally outperform informal interpreters and most digital tools in situations involving complexity, nuance, or clinical risk. That does not mean digital tools are irrelevant, but it does suggest that their role is currently best understood as bounded and task-contingent rather than universally substitutive.

Recent deployment studies support this interpretation. Olsavszky et al. (2025) piloted a digital translation platform designed to support consultations via multilingual mediation and found the platform feasible, but also identified personnel availability as a major bottleneck. Kothari et al. (2025), by contrast, described a system-wide digital medical interpretation framework focused on infrastructure, EHR integration, hardware, and operational scale. Together, these studies show that multilingual spoken communication is not only a model-performance problem. It is also a systems-design, procurement, and workflow problem.

The literature on direct speech translation in healthcare remains thinner and methodologically harder to interpret than the literature on written translation. Today, most research on spoken AILTs has focused on computer-assisted interpreting (CAI) and not machine interpreting (Lu and Fantinuoli, 2025). Even so, Iranzo-Sanchez et al. (2025) showed in multilingual medical education that domain-adapted ASR and speech translation pipelines can substantially outperform general systems. This is encouraging evidence for the value of domain adaptation, but it still comes from settings that are more controlled than typical bedside care. Spoken multilingual healthcare therefore remains an area where technical progress is real, but deployment claims should remain cautious.

3.2.2 Ambient scribes

Ambient scribes are currently the most visible application of spoken healthcare AILTs. They extend the earlier digital scribe concept (Coiera et al., 2018) by combining ASR, speaker diarisation, and LLM-based note generation. Recent studies suggest clear benefits, but also show why these systems should not be evaluated on perceived usefulness alone.

Balloch et al. (2024) found that an ambient AI documentation tool improved documentation quality scores, shortened consultations, and reduced task load in simulated encounters. Stults et al. (2025) reported improved clinician satisfaction and reduced time spent on note-writing after deployment, while Olson et al. (2025) found reductions in burnout, cognitive load, and after-hours documentation. Shah et al. (2025) likewise reported positive clinician perceptions regarding workload and patient engagement.

However, the picture is less reassuring when the outcome shifts from experience to fidelity. Lukac et al. (2025) found only modest reductions in documentation time and reported persistent accuracy

concerns, including occasional clinically significant inaccuracies. Wang et al. (2025) proposed a formal evaluation framework showing that fluent notes can coexist with weaknesses in transcription, diarisation, factual accuracy, and medication capture. These findings are especially important for multilingual healthcare because each additional processing step introduces another opportunity for error.

From a multilingual perspective, ambient scribes raise a further concern. Although many are evaluated in predominantly monolingual documentation settings, their architecture is readily repurposed for multilingual encounters. Once that happens, the system may be transcribing accented English, patient speech in another language, interpreted speech, or machine-translated speech. The resulting error stack is cumulative. The literature therefore supports a cautious conclusion: ambient tools may already reduce administrative burden, but there is much weaker evidence that they are ready for multilingual clinical communication.

3.2.3 Main learnings and pitfalls in spoken communication

Spoken-language AI is now infrastructural in healthcare, because ASR underpins dictation, ambient documentation, conversational agents, and speech-to-speech systems. Across the literature, three points are clear. First, speech tools can improve usability and reduce burden, especially for documentation-related tasks. Second, domain adaptation matters. Third, performance remains highly context-dependent, and human review is still necessary in high-stakes settings (Ng et al., 2025).

The major risks are both technical and equity-related. Koenecke et al. (2020) showed racial disparities in ASR outside healthcare, while Zolnoori et al. (2024) reported analogous disparities in patient-nurse communication. For multilingual care, this expands the equity problem beyond named languages to include accent, dialect, conversational style, and racialised speech. The strongest implication is methodological: spoken systems should be evaluated as end-to-end pipelines rather than as isolated components, because compounding errors across recognition, translation, summarisation, and note generation can distort the clinical record even when workflow efficiency appears to improve. This is something that most reviewed papers currently lack.

3.3 Agentic workflows

The third domain, agentic workflows, is empirically the least mature but strategically the most consequential. Here the focus shifts from single-task systems to orchestrated pipelines that combine language processing with retrieval, planning, routing, documentation, and interaction with clinical systems (Briva-Iglesias, 2025). In multilingual healthcare, this may include tools that receive patient messages, detect language, translate content, retrieve relevant context, draft a response, and route the case onward, or systems that process spoken encounters into notes and structured records (Heydari et al., 2025).

This area matters because multilingual care is rarely experienced as a series of isolated language tasks. Patients need symptom intake, appointment preparation, portal communication, after-visit summaries, navigation guidance, and follow-up messaging. Clinicians need support with note drafting, inbox management, referral text, and multilingual communication. Agentic workflows promise to connect these tasks, but they also risk blurring task boundaries and obscuring where accountability lies (Ojewale et al., 2025).

3.3.1 Evidence base and current maturity

The empirical literature on real-world clinical LLM workflows remains relatively thin. Artsi et al. (2025), in a systematic review of real-world clinical workflows, found surprisingly few peer-reviewed empirical studies despite the high level of public attention surrounding LLM and AI agents deployment. Reported applications included message drafting, outpatient communication, mental health support, and information extraction, and some studies reported gains in efficiency and user satisfaction. At the same time, the review highlighted limited generalisability, regulatory delays, and a lack of robust post-deployment monitoring.

Chen et al. (2025b) similarly argue that LLMs and agents in healthcare require richer evaluation frameworks than traditional task-based systems. In multilingual healthcare this matters especially because, as previously discussed, a workflow may combine translation, retrieval, summarisation, and action. A system may perform reasonably at each subtask in isolation while still failing as a workflow if it loses negation during translation, retrieves outdated policy information, or generates an overconfident patient-facing summary.

The current evidence therefore supports a cautious position. Agentic systems may offer operational value, but the literature is presently stronger on potential than on mature multilingual clinical deployment. For this reason, the most defensible stance is neither dismissal nor exuberance, but tightly governed experimentation.

3.3.2 Why agentic workflows are especially important for multilingual healthcare

Multilingual healthcare is especially likely to benefit from agentic designs because language-discordant care often requires chains of actions rather than isolated outputs (Heydari et al., 2025). A patient portal message, for example, may require language identification, translation, urgency recognition, retrieval of relevant medication information, drafting of a plain-language response, and routing to the appropriate team. Similar logic applies to discharge preparation, appointment reminders, consent support, and navigation tasks.

At the same time, multilingual healthcare is one of the least forgiving environments in which to assume that agentic flexibility is automatically beneficial. Language tasks may sit close to legal, ethical, and clinical thresholds. A mistranslated symptom, an over-smoothed explanation, or an inferred but undocumented detail can produce downstream harm. Agentic systems also raise the risk of silent task expansion: a tool perceived as a translation assistant may also begin summarising, simplifying, or prioritising without the user fully noticing that the communicative task has changed.

The HCAILT lens is useful here because it brings these risks into a single frame (Briva-Iglesias and O'Brien, 2026). Reliability requires that each step in the chain be fit for purpose and that end-to-end behaviour be evaluated rather than inferred from component quality. Safety culture requires explicit scope boundaries and escalation pathways. Trustworthiness requires that users understand what the system has done, what sources it has used, and where uncertainty remains.

3.3.3 Governance, non-determinism, and bounded agency

The governance literature suggests that agentic clinical systems may require new regulatory categories and stronger operational safeguards. Tan et al. (2026) argue that general-purpose, non-deterministic, increasingly agentic software fits poorly within traditional medical-device paradigms. This concern is reinforced by Winslow et al. (2026), and is especially relevant to multilingual language agents, which are often flexible, prompt-sensitive, and repurposable rather than narrowly locked systems.

In practical terms, this means that a multilingual agent linked to a patient portal or EHR should not treat translation, simplification, summarisation, triage framing, and explanation as interchangeable tasks. These activities have different risk profiles and should be bounded accordingly. See, for example, the EU AI Act's tier risk (EU, 2024). Useful safeguards may include constrained retrieval, task-specific thresholds, confidence-aware escalation, editable intermediate outputs, provenance displays, and explicit human fallback. The critical question is not whether "humans remain symbolically in the loop", according to some, but whether they are genuinely positioned to detect and correct consequential failure.

3.3.4 Main learnings and pitfalls in agentic workflows

The main lesson from the current agentic literature is that the field is advancing conceptually faster than empirically. The strongest evidence still comes less from mature multilingual patient-care deployments than from workflow reviews, governance papers, and early implementation studies in adjacent clinical uses (Liu et al., 2025; Karunanayake, 2025). This does not reduce the importance of the area. It suggests that now is the right time to shape expectations before unsafe assumptions solidify.

The main pitfalls are opaque task boundaries, weak post-deployment monitoring, and a tendency to overinterpret workflow automation as a purely technical gain. In multilingual healthcare, agentic systems are attractive because they can connect communicative tasks that currently fragment care, but their danger lies in connecting those tasks without sufficient transparency, constraint, and accountability.

4 Grand challenges ahead

The literature reviewed does not point to a single bottleneck. It reveals a cluster of interlocking tensions that cannot be solved by model improvement alone. These are best understood as grand challenges because they span modalities, disciplines, and institutional layers. Table 1 summarises the seven challenges proposed here.

Grand challenge	What is at stake	Primary modalities	HCAILT linkage
1. Clinically valid, risk-sensitive evaluation	Preventing fluent-but-harmful output; demonstrating comprehension and correct action.	Text, speech-to-text, speech-to-speech	Reliability + Trustworthiness
2. End-to-end multilingual fidelity	Preventing cumulative errors across recognition, translation, rewriting, and summarisation.	Speech, text, multimodal pipelines	Reliability
3. Bounded agency and safe failure	Ensuring that flexible systems stay within scope and escalate appropriately.	Agentic workflows	Safety culture
4. Redesign of human roles and competencies	Defining who uses, reviews, governs, and takes responsibility for AI-mediated language work.	Text, speech, workflows	Safety culture + Trustworthiness
5. Equity for minor languages, dialects, and accents	Preventing a two-tier communication infrastructure.	All	Reliability + Safety culture
6. Governance, regulation, and reporting	Creating institutional and regulatory mechanisms for non-deterministic systems.	All	Safety culture
7. Trust-oriented UX and overreliance prevention	Designing interfaces that support calibrated use rather than blind acceptance.	All	Trustworthiness

Table 1: Proposed grand challenges for AI language technologies in multilingual healthcare

4.1 Clinically valid, risk-sensitive evaluation

The first grand challenge is to redesign evaluation so that it reflects clinical reality rather than linguistic convenience. Current translation and speech studies increasingly incorporate severity judgements, but multilingual healthcare still relies too heavily on metrics and protocols that are insufficient for deployment decisions. Ray et al. (2025) and Martos et al. (2025) illustrate why: a system can appear excellent under one evaluation setup and far more fragile under another, especially once multiple languages are included.

Risk-sensitive evaluation therefore requires at least four shifts. Errors must be weighted by potential clinical consequence. Outcomes must include comprehension, actionability, and correct follow-through rather than expert text comparison alone. Reporting must be stratified by language, dialect, accent, and task. And evaluation must become continuous rather than purely predeployment. In this sense, multilingual healthcare needs not just better benchmarks, but monitoring and auditing infrastructures aligned with real use (Mellinger et al., 2025). Are we doing this in MT?

4.2 End-to-end multilingual fidelity

The second challenge is end-to-end multilingual fidelity. Much of the literature still evaluates ASR, MT, simplification, summarisation, or note generation separately. In real healthcare workflows, however, these components increasingly operate in sequence. Spoken encounters may pass through recognition, translation, diarisation, summarisation, and note insertion, while written workflows may combine retrieval, translation, rewriting, and messaging. Each stage creates new opportunities for distortion.

The strongest implication of the spoken-language literature is that acceptable performance at one stage does not guarantee faithful end-to-end outcomes. The field therefore needs to move beyond component benchmarking and towards auditing how information changes as it passes through complete multilingual pipelines. And how errors stack after each step.

4.3 Bounded agency and safe failure

The third grand challenge is to ensure that flexible systems fail safely. Agentic systems are attractive because they can handle multiple subtasks, but that same flexibility makes them risky in multilingual care. A single system may translate, simplify, summarise, prioritise, and route information. In high-stakes settings, that is too much responsibility for an unbounded tool.

A strong research agenda on bounded agency would ask which tasks can be safely combined, where explicit hand-offs are required, how uncertainty should trigger abstention or escalation, and how intermediate outputs can remain visible for review, for example via quality estimation (Sindhujan et al., 2025). In multilingual healthcare, safe

failure should be treated as a first-class design objective rather than as an afterthought.

4.4 Redesign of human roles and competencies

The fourth challenge is the redesign of human roles and competencies. As AILTs enter healthcare, translators, interpreters, clinicians, informaticians, and administrative staff are not disappearing from multilingual communication workflows, but their roles are changing. The literature on PE workflows suggests that language professionals remain central, but increasingly as reviewers, terminology stewards, quality controllers, and escalation points (Briva-Iglesias and O'Brien, 2022; Ehrensberger-Dow et al., 2023). At the same time, clinicians are becoming more direct users of AI-mediated communication tools, often without systematic training (Marwaha et al., 2025).

This creates a dual need. People require practical AI literacy (Long and Magerko, 2020): an understanding of likely failure modes, appropriate reliance, and the limits of automation. They also require stronger integration into workflow design and governance. Health systems, in turn, need explicit pathways specifying when professional interpreters, translators, or clinical reviewers must be involved. This challenge links translation studies, HCI, workforce redesign, and implementation science.

4.5 Equity for minor languages, dialects, and accents

The fifth challenge is equity. Across the literature, the strongest gains are least secure precisely where communication vulnerability is often greatest: minor languages, accent-diverse speech, and sociolinguistically marginalised communities. Current systems do not fail evenly, and this is not a minor technical inconvenience. It is a structural risk fulfilling the inverse care law, which states that the availability of good medical care tends to vary inversely with its need (Hart, 1971).

If healthcare organisations adopt AILTs because they work well for dominant languages, they may reduce cost and waiting times for some patients while entrenching lower-quality service for others. Equity therefore cannot be treated as a later optimisation problem. It must be built into evaluation, procurement, deployment, and reporting from the outset through stratified quality reporting, curated data strategies, community involvement, and explicit thresholds below which digital support should not displace professional services.

4.6 Governance, regulation, and reporting

The sixth challenge is governance. If multilingual communication is treated as clinical infrastructure, then failures in language mediation must become governable in ways analogous to other patient-safety issues. That requires institutional policies on approved use cases, audit trails, terminology management, version control, privacy safeguards, and incident reporting (Shneiderman, 2020). It also requires regulatory thinking capable of addressing systems whose behaviour varies with prompts, retrieval sources, context windows, and workflow integrations.

Broader AI-in-healthcare frameworks such as SPIRIT-AI and CONSORT-AI provide useful foundations for clearer reporting (Cruz Rivera et al., 2020; Liu et al., 2020), and Tan et al. (2026) push further by arguing that agentic systems require a different regulatory lens. What remains underdeveloped is a multilingual health communication perspective within these frameworks. The field still lacks strong norms for reporting language-specific variability, communication outcomes, and the boundary between linguistic and clinical responsibility.

4.7 Trust-oriented UX and overreliance prevention

The seventh challenge concerns trust and interface design. The literature repeatedly suggests that fluent output invites overreliance (Robinette et al., 2016). What multilingual healthcare needs, however, is not maximal trust, but calibrated trust (Zerilli et al., 2022). Users should be able to see what the system has done, what remains uncertain, and when review or escalation is necessary (see "seams" within HCI literature (Ehsan et al., 2024)).

For multilingual healthcare, this has direct interface implications. Systems should surface provenance when retrieval is involved, highlight unresolved terminology, flag uncertainty around clinically critical entities such as medication, dosage, negation, and time, and make editing and error reporting straightforward. Poor interface design can make even a technically strong model unsafe by encouraging autopilot behaviour. Conversely, well-designed interfaces can make imperfect systems safer by supporting reflection, verification,

and contestation.

Taken together, these seven challenges suggest that the next phase of research should focus less on isolated tools and more on accountable multilingual communication infrastructures. The future of the field will depend not on whether AI can generate language, but on whether it can be embedded into systems that remain reliable, safe, and trustworthy under real-world conditions (Reiter, 2025).

5 Conclusions and research agenda

Recent literature on AI language technologies in multilingual healthcare justifies both optimism and restraint. The optimism is real. For some written tasks and some language pairs, LLM-based translation is already approaching professional quality (Ray et al., 2025). Reviewed PE workflows can accelerate multilingual document production without abandoning expert oversight (Brewster et al., 2025). Plain-language rewriting can make difficult discharge materials more understandable (Rust et al., 2025; Zaretsky et al., 2024). Ambient documentation tools can reduce at least some dimensions of documentation burden and burnout (Olson et al., 2025; Stults et al., 2025). Domain-adapted speech pipelines can also improve performance in specialist settings (Iranzo-Sanchez et al., 2025). These are meaningful advances for access, efficiency, and UX in multilingual healthcare.

The restraint is equally important. Gains are not distributed evenly across languages, accents, and settings. Spoken systems remain particularly fragile in real-world conversational conditions. Minor languages continue to face poorer performance. Agentic workflows advance faster rhetorically than empirically. And many of the hardest problems are institutional and not only technical: procurement, escalation pathways, role allocation, AI literacy, monitoring, and governance.

Consequently, the future of multilingual healthcare cannot be reduced to a competition between models and BLEU scores. It is a question of how communication systems are designed, how users are expected to rely on them, and how accountability is preserved when language mediation becomes increasingly automated. This is also why the paper belongs squarely within a Translators and Users perspective. Translators, interpreters, clinicians, and patients are not peripheral to these technologies. They are the actors through whom reliability, safety, and trustworthiness are either realised or undermined. AILT research in healthcare therefore needs to examine not only outputs, but also user roles, revised workflows, human oversight, and the conditions under which trust becomes warranted.

If HCAILT is used as a lens (Briva-Iglesias and O'Brien, 2026), a practical research agenda for the next phase can be organised around three linked pillars. First, a reliability pillar should develop risk-sensitive multilingual benchmarks, end-to-end pipeline evaluations, and comprehension-oriented assessments grounded in real healthcare tasks. Second, a safety-culture pillar should focus on implementation: role design, AI literacy, escalation, terminology governance, and multilingual incident reporting. Third, a trust-and-governance pillar should develop interface patterns for calibrated use, reporting norms for multilingual AI interventions, and regulatory approaches that acknowledge the realities of non-deterministic and increasingly agentic systems.

The main implication of this review is therefore transdisciplinary. MT/NLP researchers contribute modelling and evaluation expertise. Translation studies contributes long-standing knowledge about mediation, revision, function, and language inequity. Healthcare researchers and practitioners ground the field in clinical reality. HCI contributes user-centred design and trust calibration. And policymakers shape the institutional conditions under which these systems can be used responsibly. No single field can solve these problems alone.

This review has limitations. It is a narrative synthesis rather than a formal systematic review, and it reflects a rapid-moving field, especially for agentic workflows, where discourse still outpaces strong deployment evidence. Even so, the present moment is decisive. Multilingual healthcare is one of the first domains in which AI language technologies are becoming infrastructure. The choices made now about evaluation, workflow design, governance, and equity will determine whether that infrastructure becomes merely efficient, or genuinely safe, just, and trustworthy.

References

Artsi, Y., V. Sorin, B. S. Glicksberg, P. Korfiatis, G. N. Nadkarni, and E. Klang. 2025. Large language models in real-world clinical workflows: A systematic review of applications and implementation. Frontiers in Digital Health, 7, 1659134.

Balloch, J., S. Sridharan, G. Oldham, F. Morehouse,

H. Brant, and A. V. Varadarajan. 2024. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. Future Healthcare Journal, 11(3), 100157.

Brewster, R. C. L., A. S. Desai, D. Lall, C. Do, S. Xiong, A. Hernandez, et al. 2025. Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: A multidisciplinary analysis. npj Digital Medicine, 8(1), 629.

Briva-Iglesias, Vicent and Sharon O'Brien. 2022. The Language Engineer: A Transversal, Emerging Role for the Automation Age. *Quaderns de Filologia - Estudis Lingüístics*, 27(0):17–48, December.

Briva-Iglesias, V. and S. O'Brien. 2026. Human-Centered AI Language Technology (HCAILT): An empathetic design framework for reliable, safe and trustworthy multilingual communication. International Journal of Human-Computer Interaction.

Briva-Iglesias, Vicent and Isabel Peñuelas Gil. 2025. Simplifying healthcare communication: Evaluating AI-driven plain language editing of informed consent forms. In Ginel, María Isabel Rivas, Patrick Cadwell, Paolo Canavese, Silvia Hansen-Schirra, Martin Kappus, Anna Matamala, and Will Noonan, editors, *Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL)*, pages 55–65, Geneva, Switzerland, June. European Association for Machine Translation.

Briva-Iglesias, Vicent. 2025. Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 365–377, Geneva, Switzerland, June. European Association for Machine Translation.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners, July.

Chen, C. L., J. Kim, S. Surani, et al. 2025a. A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally. BMC Medical Education, 25(1), 1022.

Chen, X., J. Xiang, S. Lu, Y. Liu, M. He, and D. Shi. 2025b. Evaluating large language models and agents in healthcare: Key challenges in clinical applications. Intelligent Medicine, 5(3).

Chu, J. N., J. Wong, N. S. Bardach, I. E. Allen, J. Barr-Walker, M. Sierra, U. Sarkar, and E. C. Khoong. 2024. Association between language discordance and unplanned hospital readmissions or emergency department revisits: A systematic review and meta-analysis. BMJ Quality & Safety, 33(7), 456-469.

Coiera, E., A. B. Kocaballi, J. Halamka, and L. Laranjo. 2018. The digital scribe. npj Digital Medicine, 1, 58.

Cruz Rivera, S., X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nature Medicine, 26, 1351-1363.

Dew, K. N., A. M. Turner, Y. K. Choi, A. Bosold, and K. Kirchhoff. 2018. Development of machine translation technology for assisting health communication: A systematic review. Journal of Biomedical Informatics, 85, 56-67.

Ehrensberger-Dow, Maureen, Alice Delorme Benites, and Caroline Lehr. 2023. A new role for translators and trainers: MT literacy consultants. *The Interpreter and Translator Trainer*, 17(3):393–411, July.

Ehsan, Upol, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daume. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–29, April.

EU. 2024. European Union AI Act.

Gallifant, Jack, Katherine C. Kellogg, Matt Butler, Amanda Centi, Shan Chen, Patrick F. Doyle, Sayon Dutta, Joyce Guo, Matthew J. Hadfield, Esther H. Kim, David E. Kozono, Hugo JWL Aerts, Adam B. Landman, Raymond H. Mak, Rebecca G. Mishuris, Tanna L. Nelson, Guergana K. Savova, Elad Sharon, Benjamin C. Silverman, Umit Topaloglu, Jeremy L. Warner, and Danielle S. Bitterman. 2025. A Field Guide to Deploying AI Agents in Clinical Practice, December.

Genovese, Ariana, Sahar Borna, Cesar A. Gomez-Cabello, Syed Ali Haider, Srinivasagam Prabha, Antonio J. Forte, and Benjamin R. Veenstra. 2024. Artificial intelligence in clinical settings: A systematic review of its role in language translation and interpretation. *Annals of Translational Medicine*, 12(6):117, December.

Goisauf, M., M. Cano Abadia, A. Fiske, E. Vayena, and S. Hurst. 2025. Trust, trustworthiness, and the future of medical AI: Outcomes of an interdisciplinary expert workshop. Journal of Medical Internet Research, 27, e71236.

Hansen-Schirra, Silvia and Christiane Maaß. 2020. Easy Language - Plain Language - Easy Language Plus: Perspectives on Comprehensibility and Stigmatisation. pages 17–38. September.

Hart, Julian Tudor. 1971. THE INVERSE CARE LAW. *The Lancet*, 297(7696):405–412, February.

Herrera-Espejel, P. S. and S. Rach. 2023. The use of machine translation for outreach and health communication in epidemiology and public health: Scoping review. JMIR Public Health and Surveillance, 9, e50814.

Heydari, A. Ali, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, Ahmed A. Metwally, Brent Winslow, Yubin Kim, Kumar Ayush, Yuzhe Yang, Girish Narayanswamy, Maxwell A. Xu, Jake Garrison, Amy Armento Lee, Jenny Vafeiadou, Ben Graef, Isaac R. Galatzer-Levy, Erik Schenck, Andrew Barakat, Javier Perez, Jacqueline Shreibati, John Hernandez, Anthony Z. Faranesh, Javier L. Prieto, Connor Heneghan, Yun Liu, Jiening Zhan, Mark Malhotra, Shwetak Patel, Tim Althoff, Xin Liu, Daniel McDuff, and Xuhai "Orson" Xu. 2025. The Anatomy of a Personal Health Agent, September.

Iranzo-Sanchez, J., M. Roldan, T. Rossetto, A. Cattelan, E. Salesky, C. Escolano, et al. 2025. Speech translation for multilingual medical education leveraged by large language models. Artificial Intelligence in Medicine, 166, 103147.

Karunanayake, Nalan. 2025. Next-generation agentic AI for transforming healthcare. *Informatics and Health*, 2(2):73–83, September.

Koenecke, A., A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. 2020. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

Koenecke, Allison, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1672–1681, New York, NY, USA, June. Association for Computing Machinery.

Kothari, U., A. Squires, J. Austrian, A. Feldman, I. Syed, and S. Jones. 2025. Implementing systemwide digital medical interpretation: A framework for healthcare organizations. JAMIA Open, 8(6), ooaf100.

Leung, Tiffany I., Andrew J. Coristine, and Arriel Benis. 2025. AI Scribes in Health Care: Balancing Transformative Potential With Responsible Integration. *JMIR Medical Informatics*, 13(1):e80898, August.

Lion, K. Casey, Yu-Hsiang Lin, and Theresa Kim. 2024. Artificial Intelligence for Language Translation: The Equity Is in the Details. *JAMA*, 332(17):1427–1428, November.

Liu, X., S. Cruz Rivera, D. Moher, M. J. Calvert, and A. K. Denniston. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nature Medicine, 26, 1364-1374.

Liu, Xianyuan, Jiayang Zhang, Shuo Zhou, Thijs L. van der Plas, Avish Vijayaraghavan, Anastasiia Grishina, Mengdie Zhuang, Daniel Schofield, Christopher Tomlinson, Yuhan Wang, Ruizhe Li, Louisa van Zeeland, Sina Tabakhi, Cyndie Demeocq, Xiang Li, Arunav Das, Orlando Timmerman, Thomas Baldwin-McDonald, Jinge Wu, Peizhen Bai, Zahraa Al Sahili, Omnia Alwazzan, Thao N. Do, Mohammod N. I. Suvon, Angeline Wang, Lucia Cipolina-Kun, Luigi A. Moretti, Lucas Farndale, Nitisha Jain, Natalia Efremova, Yan Ge, Marta Varela, Hak-Keung Lam, Oya Celiktutan, Ben R. Evans, Alejandro Coca-Castro, Honghan Wu, Zahraa S. Abdallah, Chen Chen, Valentin Danchev, Nataliya Tkachenko, Lei Lu, Tingting Zhu, Gregory G. Slabaugh, Roger K. Moore, William K. Cheung, Peter H. Charlton, and Haiping Lu. 2025. Towards deployment-centric multimodal AI beyond vision and language, September.

Long, Duri and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–16, New York, NY, USA, April. Association for Computing Machinery.

Lopez, I., D. E. Velasquez, J. H. Chen, and J. A. Rodriguez. 2025. Operationalizing machine-assisted translation in healthcare. npj Digital Medicine, 8(1), 584.

Lu, Xinchao and Claudio Fantinuoli. 2025. Machine and Computer-assisted Interpreting: Innovations in and Implications for Interpreting Practice, Pedagogy and Research. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 24, December.

Lukac, P. J., P. Mishra, E. Polychronopoulou, et al. 2025. Ambient AI scribes in clinical practice: A randomized trial. NEJM AI, 2(12).

Martos, M., B. Fields, S. G. Finlayson, et al. 2025. Accuracy of artificial intelligence vs professionally translated discharge instructions. JAMA Network Open, 8(9), e2532312.

Marwaha, Jayson S., William Yuan, Mukund Poddar, Pansy Elsamadisi, and Gabriel A. Brat. 2025. The algorithmic consultant: A new era of clinical AI calls for a new workforce of physician-algorithm specialists. *npj Digital Medicine*, 8(1):552, August.

McMinn, D., T. Grant, L. DeFord-Watts, V. Porkess, M. Lens, C. Rapier, W. Q. Joe, T. A. Becker, and W. Bender. 2025. Using artificial intelligence to expedite and enhance plain language summary abstract writing of scientific content. JAMIA Open, 8(2), ooaf023.

Mellinger, Christopher D., Nicoletta Spinolo, Maureen Ehrensberger-Dow, and Sharon O'Brien. 2025. Designing studies with naturalistic tasks. In *Research Methods in Cognitive Translation and Interpreting Studies*, pages 49–68. John Benjamins, April.

Montalt-Resurrecció, Vicent, Isabel García-Izquierdo, and Ana Muñoz-Miquel. 2024. *Patient-Centred Translation and Communication*. Taylor & Francis, December.

Ng, J. J. W., E. Wang, X. Zhou, K. X. Zhou, C. X. L. Goh, G. Z. N. Sim, H. K. Tan, S. S. N. Goh, and Q. X. Ng. 2025. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: A systematic review. BMC Medical Informatics and Decision Making, 25(1), 236.

Ojewale, Victor, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–29, New York, NY, USA, April. Association for Computing Machinery.

Olsavszky, V., M. Bazari, T. Ben Dai, A. Olsavszky, F. Finkelmeier, M. Friedrich-Rust, S. Zeuzem, E. Herrmann, J. Leipe, F. A. Michael, H. von Westernhagen, and O. Ballo. 2025. Digital translation platform (Translatly) to overcome communication barriers in clinical care: Pilot study. JMIR Formative Research, 9, e63095.

Olson, K. D., D. Meeker, M. Troup, et al. 2025. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Network Open, 8(10), e2534976.

Pym, Anthony. 2025. Risk Management in Translation. *Elements in Translation and Interpreting*, February.

Rawal, Shail, Jeevitha Srighanthan, Arthi Vasantharoopan, Hanxian Hu, George Tomlinson, and Angela M. Cheung. 2019. Association Between Limited English Proficiency and Revisits and Readmissions After Hospitalization for Patients With Acute and Chronic Conditions in Toronto, Ontario, Canada. *JAMA*, 322(16):1605–1607, October.

Ray, M., D. J. Kats, J. Moorkens, et al. 2025. Evaluating a large language model in translating patient instructions to Spanish using a standardized framework. JAMA Pediatrics, 179(9), 1026-1033.

Reiter, Ehud. 2025. We Should Evaluate Real-World Impact. *Computational Linguistics*, pages 1–13, September.

Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–108, Christchurch, New Zealand, March. IEEE.

Rust, P., J. Frings, S. Meister, and L. Fehring. 2025. Evaluation of a large language model to simplify discharge summaries and provide cardiological lifestyle recommendations. Communications Medicine, 5(1), 208.

Shah, S. J., T. Crowell, Y. Jeong, et al. 2025. Physician perspectives on ambient AI scribes. JAMA Network Open, 8(3), e251904.

Shneiderman, B. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504.

Sindhujan, Archchana, Diptesh Kanojia, and Constantin Orăsan. 2025. Reference-Less Evaluation of Machine Translation: Navigating Through the Resource-Scarce Scenarios. *Information*, 16(10):916.

Stephanidis, C., G. Salvendy, M. Antona, J. Y. C. Chen, J. Dong, V. G. Duffy, X. Fang, C. Fidopiastis, G. Fragomeni, L. P. Fu, Y. Guo, D. Harris, A. Ioannou, K. A. Jeong, S. Konomi, H. Kromker, M. Kurosu, J. R. Lewis, A. Marcus, et al. 2019. Seven HCI grand challenges. International Journal of Human-Computer Interaction, 35(14), 1229-1269.

Stephanidis, C., G. Salvendy, M. Antona, V. G. Duffy, Q. Gao, W. Karwowski, S. Konomi, F. Nah, S. Ntoa, P. L. P. Rau, K. Siau, and J. Zhou. 2025. Seven HCI grand challenges revisited: Five-year progress. International Journal of Human-Computer Interaction, 41(19), 11947-11995.

Stults, C. D., S. Deng, M. C. Martinez, et al. 2025. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Network Open, 8(5), e258614.

Sukhera, J. 2022. Narrative reviews: Flexible, rigorous, and practical. Journal of Graduate Medical Education, 14(4), 414-417.

Tan, C., D. V. Gunasekeran, R. Agarwal, M. I. Bittner, K. V. Carson, et al. 2026. Regulation of clinical artificial intelligence in the age of agents: Unconfined non-deterministic clinical software (UNDCS) systems for healthcare. npj Digital Medicine, 9(1).

Terribile, Silvia. 2024. Is post-editing really faster than human translation? *Translation Spaces*, 13(2):171–199, December.

Tu, Tao, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe

Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, June.

Ugas, M., M. A. Calamia, J. Tan, B. Umakanthan, C. Hill, K. Tse, A. Cashell, Z. Muraj, M. Giuliani, and J. Papadakos. 2025. Evaluating the feasibility and utility of machine translation for patient education materials written in plain language to increase accessibility for populations with limited English proficiency. Patient Education and Counseling, 131, 108560.

Valdez, Susana, Floor van Heeswijk, and Noa Warren. 2025. Machine Translation at the Hospital: Healthcare Professionals' Perspectives on Use, Appropriateness, and Policy. *Tradumàtica tecnologies de la traducció*, (23):244–265, December.

van Leersum, C. M. and C. Maathuis. 2025. Human centred explainable AI decision-making in healthcare. Journal of Responsible Technology, 21, 100108.

van Lent, L. G. G., N. G. Yilmaz, S. Goosen, J. Burgers, S. Giani, B. C. Schouten, and M. W. Langendam. 2025. Effectiveness of interpreters and other strategies for mitigating language barriers in healthcare: A systematic review. Patient Education and Counseling, 136, 108767.

Wang, H., R. Yang, M. Alwakeel, A. Kayastha, A. Chowdhury, J. M. Biro, A. D. Sorrentino, J. L. Handley, S. Hantzmon, S. Bessias, N. J. Economou-Zavlanos, A. Bedoya, M. Agrawal, R. M. Ratwani, E. G. Poon, M. J. Pencina, K. I. Pollak, and C. Hong. 2025. An evaluation framework for ambient digital scribing tools in clinical applications. npj Digital Medicine, 8(1), 358.

Winslow, Brent, Ozlem Ozmen Garibay, Tesh Goyal, Sean Koon, George Margetis, Gavriel Salvendy, Ben Shneiderman, Aida Tayebi, and Laura Vardoulakis. 2026. Revisiting the Six Human-Centered Artificial Intelligence Grand Challenges in the Age of Generative AI. *International Journal of Human–Computer Interaction*, 42(7):4697–4738, April.

Woods, A. P., B. S. Davis, A. Xie, S. Campbell, L. Tieu, J. K. Hoang, and J. A. Rodriguez. 2022. Limited English proficiency and clinical outcomes after hospital-based care in English-speaking countries: A systematic review. Journal of General Internal Medicine, 37(8), 2050-2061.

Zaretsky, J., K. Rector, M. Aiken, et al. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Network Open, 7(3), e240357.

Zeng-Treitler, Q., H. Kim, G. Rosemblat, and A. Keselman. 2010. Can multilingual machine translation help make medical record content more comprehensible to patients? Studies in Health Technology and Informatics, 160(Pt 1), 73-77.

Zerilli, John, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4), April.

Zolnoori, M., S. Vergez, Z. Xu, E. Esmaeili, A. Zolnour, K. A. Briggs, J. K. Scroggins, S. F. Hosseini Ebrahimabad, J. M. Noble, M. Topaz, S. Bakken, K. H. Bowles, I. Spens, N. Onorato, S. Sridharan, and M. V. McDonald. 2024. Decoding disparities: Evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. JAMIA Open, 7(4), ooae130.