International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Automatic Speech Recognition in Healthcare in the Post-LLM Era: A Scoping Review Protocol

INPLASY202640033
doi: 10.37766/inplasy2026.4.0033
Received: 9 April 2026
Published: 10 April 2026

Alabbad, M; Alhoshan, W.

Corresponding author:
Waad Alhoshan

wmaboud@imamu.edu.sa

Author Affiliation:
Imam Mohammad Ibn Saud Islamic University (IMSIU).

ADMINISTRATIVE INFORMATION

Support - No external funding received.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202640033

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 10 April 2026 and was last updated on 10 April 2026.

INTRODUCTION

Review question / Objective Objective:
To systematically map and characterize the emerging landscape of Automatic Speech Recognition systems integrated with Large Language Models in healthcare contexts, examining their applications, technical architectures, evaluation practices, and reported implementation insights and challenges.

Review Questions:
RQ1 (Applications): In which healthcare application contexts and settings have LLM-based ASR systems been applied and evaluated?
RQ2 (Technical Architectures): What models (including LLM and PLM), training datasets, and model adaptation techniques have been utilized in target healthcare applications?
RQ3 (Evaluation Methods): What evaluation environment and methods, including performance metrics, are employed to assess LLM-based ASR systems in healthcare contexts?
RQ4 (Reported Insights & Challenges): What are the benefits, implementation challenges, and ethical considerations that have been reported in the included studies?

PCC Elements:
Population (P): Healthcare and health-related contexts where ASR-LLM systems are deployed or evaluated.
Concept (C): Integration of Automatic Speech Recognition (ASR) systems with Large Language Models (LLMs), including: applications (clinical documentation, diagnostic assessment, therapeutic interventions, patient communication, medical education, emergency services, administration); technical architectures (ASR models, LLM models, datasets, adaptation techniques); evaluation methodologies and metrics; and implementation considerations (privacy, equity, reproducibility, clinical outcomes).
Context (Cx): Healthcare settings (hospitals, clinics, emergency departments, ambulances, home healthcare, telehealth, therapeutic environments, medical education institutions,

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/

administrative contexts); studies published January 2022 through December 2025; open-access English-language original research articles.

Background Automatic Speech Recognition (ASR) technology converts spoken language into written text, serving as a critical interface between clinicians and healthcare information systems. The field has undergone four major evolutionary phases. Hidden Markov Models combined with Gaussian Mixture Models dominated from the 1980s through early 2010s, requiring manual feature engineering and separate acoustic, language, and pronunciation models. Deep Neural Networks replaced statistical approaches around 2012, dramatically improving acoustic modeling through learned representations. End-to-end architectures emerged subsequently, with Connectionist Temporal Classification and attention-based encoder-decoder models eliminating the need for frame-level alignments. The current era is defined by Transformer-based architectures—including Conformer, wav2vec 2.0, and Whisper—that leverage self-supervised learning on massive unlabeled datasets to achieve near-human transcription accuracy on benchmark tasks.

In healthcare contexts, ASR applications have historically focused on three areas: clinical documentation (radiology reports, physician dictation, encounter notes), diagnostic assessment (speech intelligibility evaluation, cognitive screening), and accessibility (live captioning, voice-controlled medical devices). Performance evaluations consistently reported challenges with medical terminology, speaker variability, acoustic noise, and cross-population disparities. Benchmark Word Error Rates demonstrated significant performance gaps across demographic groups, raising equity concerns about clinical deployment.

Large Language Models represent a distinct technological lineage emerging from natural language processing research. Unlike earlier language models that captured statistical patterns, contemporary LLMs—exemplified by GPT-4 and LLaMA—and pretrained language models (PLMs) such as BERT-based architectures demonstrate emergent capabilities including complex reasoning, instruction following, few-shot learning, and sophisticated text generation. Healthcare applications of LLMs span clinical decision support, medical question answering, patient communication, documentation generation, and medical education. However, systematic evaluations have documented critical limitations including hallucinations (confident generation of factually incorrect information), inconsistent performance across medical specialties, and challenges with medical knowledge integration.

The convergence of ASR and LLM technologies represents a fundamental architectural shift. Traditional ASR-only pipelines produced verbatim transcripts requiring human interpretation and structuring. Contemporary ASR-LLM systems integrate speech recognition with language understanding to generate structured clinical notes, perform speaker diarization, extract diagnostic entities, suggest clinical codes, and correct transcription errors through contextual reasoning. This integration transforms speech technology from a transcription utility into a clinical intelligence tool capable of semantic understanding and clinical workflow integration.

Several reviews have examined components of this landscape. Previous scoping reviews examined digital scribes in clinical practice but predated the emergence of LLMs. Systematic reviews assessed ASR performance for clinical documentation using traditional metrics. Surveys examined intelligent speech technologies across multiple healthcare applications. Reviews have focused specifically on speech emotion recognition for mental health. Multiple reviews have examined LLM applications in healthcare broadly but without addressing speech technology integration. No existing review has comprehensively synthesized evidence on the convergence of ASR and LLM technologies across healthcare applications.

Rationale Expanding Clinical Landscape. LLM-enhanced ASR systems are being rapidly deployed across diverse healthcare contexts: diagnostic assessment (analyzing speech for cognitive decline, mental health conditions, disease detection), therapeutic interventions (psychological counseling, speech therapy), patient communication (multilingual translation, accessibility tools), clinical documentation, medical education, emergency services, and healthcare administration. Each domain presents distinct technical requirements, evaluation challenges, and implementation considerations, creating urgent need for comprehensive evidence mapping.

The Evidence Gap. Despite accelerating development across multiple healthcare domains, the field lacks consolidated knowledge essential for evidence-based decision-making. Stakeholders currently operate without clear understanding of: which clinical applications and settings have been rigorously evaluated versus underexplored; what technical architectures and LLM-ASR combinations demonstrate effectiveness in different contexts; how performance should be

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/

measured beyond traditional transcription accuracy; what implementation challenges emerge across diverse settings; and what ethical risks—privacy violations, algorithmic bias, safety concerns—require domain-specific mitigation. This evidence vacuum creates risks of premature deployment, duplicated research, wasted investments, and potential patient harm.

The Methodological Heterogeneity Challenge. Published studies employ diverse LLM architectures (GPT-4, Claude, LLaMA, Whisper, domain-adapted models), varied ASR components, inconsistent evaluation metrics (from Word Error Rate to clinical outcomes), different clinical contexts (controlled laboratories to noisy emergency departments), and varied populations. This heterogeneity—combined with rapid technological evolution—makes traditional literature synthesis impossible. Systematic scoping review is necessary to map this complex landscape, identify patterns, and establish structured knowledge.

The Equity and Safety Imperative. Pre-LLM research documented significant ASR performance disparities across demographic groups. LLM integration introduces new risks varying by application: hallucinated information in medical records (documentation), misdiagnosis based on flawed speech analysis (diagnostic applications), culturally inappropriate responses (counseling systems), mistranslation of critical information (patient communication), and exacerbation of healthcare disparities. Each application presents unique safety and equity challenges. Without systematic evidence synthesis across the full spectrum of healthcare applications, deployment may introduce new harms or worsen existing inequities.

The Critical Timing Window. Commercial ASR-LLM products are being introduced into clinical workflows while systematic evidence remains limited. The field is in its early stages, yet deployment is already occurring. A scoping review at this formative stage can help establish evaluation standards, identify promising applications, and highlight safety considerations before practices become entrenched.

The Unique Contribution. This review addresses gaps no existing review has filled. Rather than examining ASR or LLMs separately, it focuses on their integration. Rather than limiting scope to single applications, it comprehensively maps the full spectrum: documentation, diagnosis, therapy, patient communication, education, emergency services, and administration. Rather than focusing solely on technical performance, it examines applications, architectures, evaluation methods, clinical outcomes, implementation challenges, and ethical considerations. Rather than traditional systematic review of mature fields, it employs scoping methodology for emerging, heterogeneous domains requiring exploratory synthesis.

Methodological Justification. Scoping review is appropriate because: the field is emerging and rapidly evolving across multiple domains; methodological and application heterogeneity requires exploratory mapping rather than quantitative synthesis; the objective is identifying knowledge gaps and research priorities rather than answering specific narrow questions; diverse stakeholders require different insights; and the goal is providing foundational evidence to inform future systematic reviews and evidence-based guidelines across healthcare contexts where LLM-enhanced ASR is deployed.

METHODS

Strategy of data synthesis Data will be synthesized using narrative synthesis methods appropriate for scoping reviews. Extracted data will be organized according to the four research questions. Descriptive statistics will summarize study characteristics (publication year, country, clinical setting, target population). Thematic analysis will identify patterns in applications, technical approaches, and evaluation methods. Results will be presented through frequency tables, taxonomies categorizing LLM and ASR models, visual representations (e.g., charts), and narrative summaries highlighting trends, gaps, and implementation considerations. The synthesis will map the landscape of LLM-enhanced ASR in healthcare rather than provide quantitative meta-analysis, consistent with scoping review methodology.

Terms and electronic databases included: Search Terms: The search strategy combined three components using Boolean operators:
- ASR terminology: "automatic speech recognition" OR "speech recognition" OR "speech-to-text" OR "voice recognition"
- Healthcare domain: "healthcare" OR "medicine" OR "clinical" OR "patient" OR "hospital" OR "medical"
- LLM identifiers: "Large Language Model" OR "LLMs" OR "GPT" OR "Whisper"

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/

Databases: PubMed, Scopus, IEEE Xplore, Web of Science (searched January 2022 through December 2025).

Eligibility criteria Population: Healthcare applications and healthcare-related contexts where LLM-enhanced ASR systems are deployed or evaluated.

Concept: Integration of Large Language Models (LLMs) with Automatic Speech Recognition (ASR) systems creating speech-to-clinical-intelligence pipelines. This includes: applications (clinical documentation, diagnostic assessment, therapeutic interventions, patient communication, medical education, emergency services, healthcare administration); technical architectures (LLM models such as GPT-4, LLaMA, Whisper; ASR components; training datasets; adaptation techniques including fine-tuning and prompting); evaluation methodologies and performance metrics; implementation outcomes (efficiency, accuracy, user satisfaction); and ethical considerations (privacy, algorithmic bias, equity across populations, regulatory compliance).

Context: Healthcare and health-related settings including hospitals, clinics, emergency departments, ambulances, home healthcare, telehealth platforms, therapeutic environments, medical education institutions, and administrative contexts. Studies published between January 1, 2022 and December 31, 2025. Open-access, English-language original research articles.

Eligibility Criteria
Studies were selected based on predefined inclusion and exclusion criteria in accordance with PRISMA-ScR guidelines. To be considered eligible, studies had to meet the following criteria:
- Published between January 1, 2022 and December 31, 2025
- Original research articles (empirical, experimental, or solution-based studies)
- Full text available in English and open access
- Focus on ASR in healthcare or health-related contexts (e.g., clinical documentation, diagnosis, therapy, patient communication, medical education, accessibility, administration)
- Investigation of integrated ASR–LLM pipelines for downstream clinical tasks

Studies were excluded if they: (1) lacked an ASR component, (2) did not integrate LLM capabilities, (3) focused on non-clinical or non-health applications, or (4) were secondary research such as reviews or surveys.

Source of evidence screening and selection The study selection process followed PRISMA-ScR guidelines across three phases: identification, screening, and inclusion.

Identification phase: Database searches were conducted in PubMed, Scopus, IEEE Xplore, and Web of Science using the predefined search strategy. All retrieved records were imported into reference management software. Duplicate records were removed using automated deduplication followed by manual verification.

Screening phase: Two independent reviewers screened titles and abstracts of unique records against eligibility criteria. Disagreements were documented and resolved through discussion. Records clearly not meeting inclusion criteria were excluded. Potentially eligible studies proceeded to full-text assessment.

Inclusion phase: Full texts of all potentially eligible studies were retrieved. Two independent reviewers assessed each full text against eligibility criteria. Reasons for exclusion at this stage were systematically documented (no ASR component, no LLM integration, non-clinical application, secondary research type). Disagreements during full-text assessment were resolved through consensus discussion between the two reviewers.

A PRISMA flow diagram documents the complete selection process including number of records identified from each database, duplicates removed, records screened, full texts assessed, studies excluded with reasons, and final number of studies included in the review.

Data management A structured data extraction form was developed based on the research questions, aligned with established data extraction schemes for scoping reviews, and piloted on three studies before refinement. Two reviewers independently extracted data from all included studies using a standardized Excel spreadsheet.

The extraction form captured seven domains: (1) Bibliographic Metadata: authors, first author country, publication year, database source, paper format, paper type; (2) Study Contexts: study motivation, objective, application context, application setting with reasoning, target users, supported languages; (3) Technology and Methods - ASR: system used, role (primary/secondary), adaptation approach, training dataset, dataset availability; (4) Technology and Methods - LLM: generative models used, role, adaptation, training dataset, dataset availability, prompting technique

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/

with reasoning, tasks performed; (5) Technology and Methods - PLM: supporting pre-trained models, adaptation, training dataset, availability, tasks; (6) Evaluation Methods: ASR evaluation metrics, ASR human-in-the-loop, LLM evaluation metrics, LLM human-in-the-loop, PLM evaluation, PLM human-in-the-loop, external validation approaches; (7) Ethics and Implementation: privacy and data governance measures, equity considerations, replication package availability.

Discrepancies between reviewers were identified through comparison and resolved through discussion until consensus was reached. Final extracted data were organized in tabular format facilitating cross-study analysis and synthesis.

Reporting results / Analysis of the evidence Results will be reported according to PRISMA-ScR guidelines and organized by the four research questions:

RQ1 (Applications): Frequency distributions of application contexts (diagnosis, therapy, documentation, education, emergency, administration, communication); clinical settings (hospital, telehealth, ambulance, home); target populations (clinicians, patients, specific demographic groups); and supported languages. Patterns in application coverage and underexplored areas will be identified.

RQ2 (Technical Architectures): Taxonomies of ASR systems (Whisper, commercial APIs, custom models) and generative LLMs (GPT-4, Claude, LLaMA variants, domain-adapted models); categorization of adaptation techniques (fine-tuning, prompting, none); characterization of datasets (nature, size, availability); analysis of system roles (primary vs. secondary); examination of prompting techniques; documentation of supporting PLMs. Technical diversity and prevailing approaches will be characterized.

RQ3 (Evaluation Methods): Summary of evaluation metrics for ASR components (WER, accuracy), LLM components (clinical utility, semantic accuracy), and PLM components; documentation of human-in-the-loop mechanisms for each component; characterization of external validation approaches. Gaps in evaluation standards will be identified.

RQ4 (Insights & Challenges): Synthesis of privacy/governance measures, equity considerations (multilingual support, accent sensitivity, accessibility), reproducibility through replication package availability. Implementation challenges and reported benefits will be thematically analyzed. Methodological heterogeneity will be characterized rather than statistically pooled. Knowledge gaps and future research priorities will be explicitly identified.

Presentation of the results Results will be presented through multiple complementary formats organized by research question:
Study Overview:
- PRISMA flow diagram documenting study selection with numbers at each stage and exclusion reasons
- Temporal and methodological distribution showing publication trends and study types (solution-based vs. empirical)
- Geographic distribution by first-author country affiliation
- Language landscape showing supported languages across studies

RQ1 - Applications:
- Application context taxonomy table categorizing identified domains
- Heatmap visualizing application contexts mapped against clinical settings to identify clustering patterns
- Clinical setting taxonomy table documenting intended use environments
- Language group analysis showing distribution across application contexts and clinical settings to identify patterns in monolingual vs. multilingual deployment

RQ2 - Technical Architectures:
- Distribution charts of ASR systems and generative LLM families employed
- LLM prompting techniques distribution documenting prevalence of different strategies
- Task taxonomy tables for generative LLMs and supporting PLMs
- Cross-tabulation matrix of ASR and LLM adaptation strategies revealing dominant patterns (frozen vs. fine-tuned components)

RQ3 - Evaluation Methods:
- Evaluation metrics taxonomy table grouped by metric family and pipeline component (ASR, LLM, PLM) with frequency counts
- Human-in-the-loop involvement comparison across ASR and LLM evaluation showing differential patterns
- External validation methods table documenting approaches employed

RQ4 - Implementation and Ethics:

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/

- Privacy and data governance measures distribution showing reported approaches
- Equity considerations distribution documenting addressed concerns
- Reproducibility assessment through replication package availability

Narrative Synthesis:
- Each research question will be accompanied by narrative synthesis highlighting key patterns, exemplar studies, knowledge gaps, and evidence-based insights for future research and implementation.
- All visualizations will include precise study counts and percentages, with explicit acknowledgment when categories are non-mutually exclusive (i.e., when counts exceed N due to studies employing multiple approaches).

Language restriction English only. Studies published in English were included to ensure consistency in data extraction.

Country(ies) involved Saudi Arabia.

Other relevant information Protocol registered after review completion but before publication. The scoping review is complete and manuscript is under preparation for journal submission. Extracted data will be publicly available upon publication.

Keywords Automatic Speech Recognition; Large Language Models; Healthcare; Clinical 21Documentation; Scoping Review.

Dissemination plans Results will be disseminated through publication in a peer-reviewed journal. The protocol is registered in INPLASY for transparency. The goal is to inform researchers, clinicians, policymakers, and technology developers.

Contributions of each author

Author 1 - Maram Alabbad - M.A.: Contributed to conceptualization and methodological design; conducted the literature search and investigation; managed data curation and extraction procedures; drafted the original protocol manuscript; and prepared visualizations and structural elements.
The author (M.A.) read and approved the protocol registration.
Email: 446012265@sm.imamu.edu.sa
Author 2 - Waad Alhoshan - Led the development of the review protocol, including conceptualization of the questions and objectives; designed the methodological framework; supervised all stages of protocol preparation; contributed to data curation planning and validation of eligibility criteria; critically supervised, reviewed and edited the manuscript. The author (W.A.) read and approved the protocol registration.
Email: wmaboud@imamu.edu.sa

Alabbad et al. INPLASY protocol 202640033. doi:10.37766/inplasy2026.4.0033 Downloaded from https://inplasy.com/inplasy-2026-4-0033/