Journal of Computational Systems and Applications

https://jcsa.cultechpub.com/jcsa

Cultech Publishing

Review

Trustworthy Cross-Lingual Knowledge Distillation for Clinical NLP: A Conceptual Review of Evidence Preservation, Explainability, Calibration, and Compound Risk

Hamza Qureshi, Hamid Ghous*

Department of Computer Science and Information Technology, University of Southern Punjab, Pakistan

*Corresponding author: Hamid Ghous, hamidghous@usp.edu.pk

Abstract

Under low-resource conditions, knowledge distillation is a feasible approach to reduce the size of the clinical language models. But similarity between teacher and student task scores is not enough to ensure clinical deployment. The accuracy and F1 may not evolve but the evidence links, contraindications, reasoning boundaries, explanation quality and uncertainty can change during compression. This problem is complicated by the fact that there is limited variation of the terminology, context, population shift, and target language validation as part of the cross lingual transfer. This conceptual review connects the dots on research in clinical NLP, multilingual transfer, compression, factual grounding, explainable AI, calibration, fairness and deployment safety. It is the first to create evidence drift, wherein the link between a recommendation or advice and clinical evidence weakened. Evidence-based distillation, clinicians and patients dual explanations and explanation-based calibration with a taxonomy of 12 categories of single, paired and compounded errors are used for analysis. The loss of evidence and translation error, the incomplete explanation, the overconfidence and the hidden stratification can combine to form a clinical-risk cascade in a chronic disease scenario. Evaluation occurs in the following nodes, Task performance, Evidence fidelity, Explanation fidelity, Calibration, Stakeholder utility and Cascade safety, organized in the format of an operational framework. Their classification and structure assist each other in future empirical studies of the trustworthy multilingual compression. Evaluation should be connected to predictive performance and should be auditable, explainable in a role/aware and culture/aware manner and be calibrated to the target population, include explicit abstention, and include intersectional subgroup monitoring.

Keywords

Knowledge distillation, Cross-lingual transfer, Clinical NLP, Evidence grounding, Explainable AI, Calibration, Trustworthiness

Article History

Received: 20 July 2026
Revised: 03 September 2026
Accepted: 14 October 2026
Available Online: 05 November 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0

1. Introduction

The use of clinical NLP is increasingly common, and it is likely that it will be used in a wide variety of settings ranging from different languages and different types of documentation to different prevalence of diseases, different hardware, different access to specialists, and so on. While large language models demonstrate broad biomedical capabilities [1-3], they are also costly and complex to deploy [4,5]. Smaller models, including distilled and compressed models, have been used to address these challenges [6,7]. DistilBERT and subsequent clinical lightweight-transformer studies show that efficiency gains are possible without sacrificing task performance [8,9]. The issue of prediction of outcomes and multilingual medical summarization in Clinical KD has been addressed by a number of studies [10,11]. The results here provide feasibility data, but not clinical data on the equivalence.

This review is a commentary on the main thrust that accuracy retention is not clinical knowledge retention. A student may reproduce a teacher's label while losing the evidence chain, contraindication, guideline qualification, and uncertainty that made the output defensible. Indeed, the study of faithfulness has shown that a clinically relevant info loss can be missed in the presence of a lexical overlap and even in the case of a specialized automatic metric [12-15]. With crosslingual deployment, more opportunities for semantic drift may be created as well as opportunities for terminology drift. Even there may be differences between various morphological systems of how to write the same token. [16,17], [18-20].

But quality of the translation is not the only the risk. Compression, evidence loss, explanation design, confidence error, and population shift can feed each other into a vicious cycle. Therefore, it might be more reliable, particularly in areas with less detailed clinical evidence, that a good indication of the level of target language fluency should be provided. The interaction, not some distinct checklists for accuracy, explainability and fairness, helps to create a cascade perspective.

This paper addresses this question by investigating trustworthiness as a connected clinical requirement instead of an independent properties of the models. Introduces an idea of "evidence drift" which describes a possible divergence between substance of a recommendation and its supporting evidence. The idea is extended and further developed in the context of evidence based distillation, temporal brittleness, language and clinical transfer, dual explainability and explanation grounded calibration. The analysis is then brings these considerations and suggests a taxonomy of 12 interacting failures and an architecture for enabling the deployment of the application in a more safe manner across languages. In other words, the taxonomy main areas reviewed are reflected in the relationship given in Figure 1. The proposed framework is intended to guide evaluation and future empirical research.

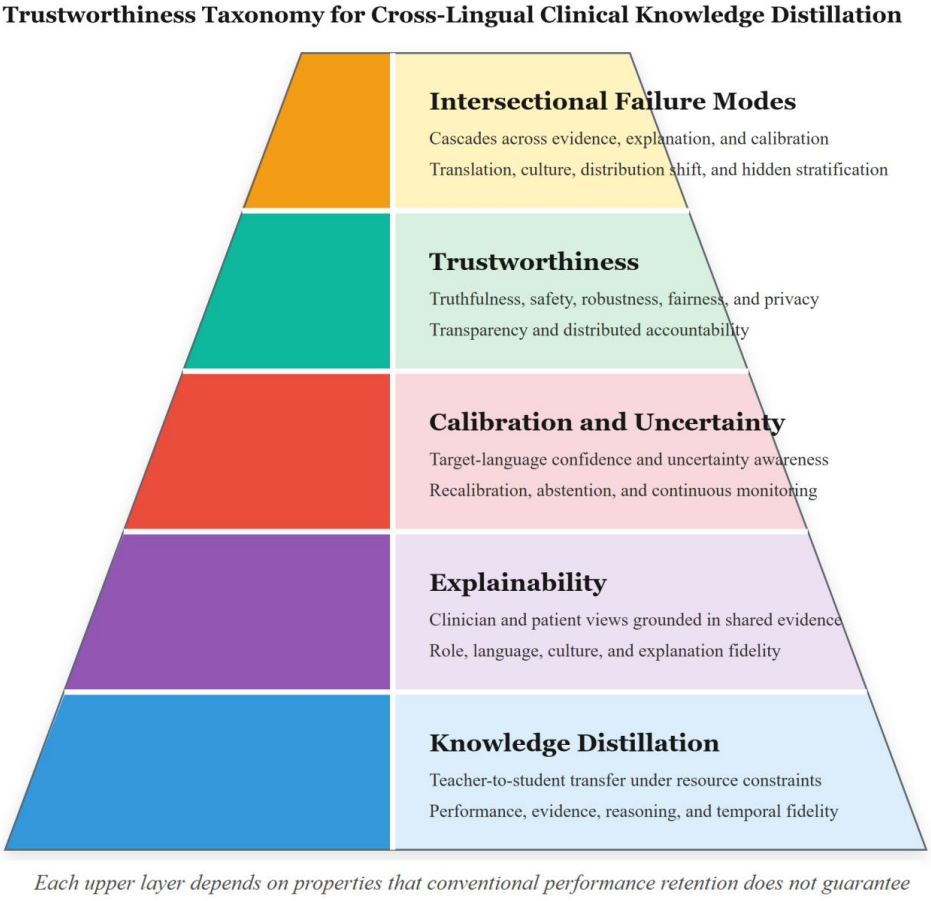


Figure 1. Author-created conceptual taxonomy of challenges in trustworthy cross-lingual clinical KD. The relationships are an author-developed synthesis and do not represent measured effect sizes.

2. Scope and Literature Approach

This conceptual/critical review integrates research on clinical knowledge distillation, cross-lingual transfer, evidence preservation, explainability, calibration, and trustworthy medical AI. The literature was selected for its relevance to the relationships among these areas, with emphasis on peer-reviewed journal and conference studies. Recent preprints were considered where published evidence remained limited. The review does not claim exhaustive coverage. Its purpose is to connect findings from related fields, identify unresolved clinical risks, and develop a framework for future empirical evaluation.

The areas of evidence gathered encompass KD surveys and objectives [21,22], general model evaluation [23,24]. The review also draws on clinical model evaluation [25-29]. Further sources address grounding and faithfulness [30-33]. The review also considers healthcare governance and risk [34-37]. Numerical claims were drawn from full-text sources. This is because it is not a pooling of percentages across studies, there are differences in datasets, endpoints, languages and language validation designs.

The four questions are used to organize the synthesis. First, when can a distilled student explain the knowledge with predictive power, losing the aspects of clinical knowledge? Second, what occurs in the cross-lingual transfer in terms of evidence, explanation and confidence? Third, what interactions are there between these dimensions that suggest that there is cause for concern beyond what each measure alone might suggest? Fourth, what kinds of architectural and evaluation practices would be able to show them before and after they were deployed? Each source is used only for the capability it can support, an NER study could be applied to the capacity of linguistic task transfer but not to medication safety, a calibration study to the capacity of confidence reliability but not to explanation comprehension. This interpretation, based on the capability-bound approach, is required due to the fact that there is currently no common clinical-KD benchmark that covers the entire framework.

3. Knowledge Distillation and Cross-Lingual Clinical Transfer

One of the common strategies used in KD is to minimize a combination of task loss and divergence between teacher and student outputs. In simplified form,

$$L_{KD} = \alpha L_{task} + (1 - \alpha)T^2 D_{KL}(p_T^{(T)} \mathbin{/\!/} p_S^{(T)}) \quad (1)$$

This objective targets the transfer of behaviour at the output, but does not require keeping of source evidence or reasoning process. Alternative divergence objectives can be used to provide an alternate way to address the clinical omission that can enhance optimization [22]. Pretraining with synthetic data and data annotation can improve the amount of supervision, or can be used to add generator error or distribution bias [38,39].

Useful task gains have been realized in the cross-lingual transfer context. Papaioannou et al. explored clinical phenotyping across multiple languages, Gupta et al. investigated cross-lingual knowledge distillation for answer sentence selection, and Nguyen et al. explored cross-lingual summarization via optimal transport [40-42]. According to Basir et al. (2025) the Sindhi XLM-R score has increased from 0.302 to 0.681 with the use of combined pretraining of Urdu and Sindhi. This would be an absolute increase of 0.379, and a relative increase of 125.5% [43]. The final F1 of 68.1% reflects a substantial improvement from a weak baseline while also showing the remaining difficulty of the task. They should not be directly compared because of differences in the amount of space used on a label, sampling and test split, clinicality and quality of the annotations. Table 1 summarizes examples of task-specific evidence relevant to cross-lingual clinical transfer.

Table 1. Examples of task-specific evidence for aspects relevant to cross-lingual clinical transfer.

Data/sample	Task/languages	Evaluation	Relevance to clinical transfer	Ref.
Machine-labelled Urdu and Sindhi corpora, with held-out evaluation reported by the authors	NER in Urdu and Sindhi	Model-wise precision, recall, and F1	Sindhi XLM-R increased from 0.302 to 0.681 F1, showing a large gain under scarce-data conditions.	[43]
Clinical phenotyping corpora across source and target languages	Multilingual phenotyping	Cross-lingual predictive evaluation	Demonstrates multilingual label transfer and motivates further assessment of calibration and explanation quality.	[40]
100 MIMICIV discharge-summary cases	English patient-centred discharge instructions	Clinician quality and safety review	Potentially harmful safety issues occurred in 18%, including hallucinations in 6% and new medications in 3%. Fluency alone was insufficient.	[44]

The studies provide evidence for separating medical meaning, evidence and safety walls and calibrated uncertainty from useful explanations in the context of clinical transfer and successful prediction across languages in the context of linguistic transfer. Thus, clinical readiness is not just a general multilingual indicator, but also a measure of the performance of NER, which must be assessed [45-47].

4. Evidence Preservation and Evidence Drift

Evidence drift occurs when there is a shift/change, degradation or disappearance of the connection between a recommendation and clinical evidence in the compression or cross-lingual transfer. If the teacher does not provide evidence E and reasoning R for x → y with respect to a standard KD, the student is allowed to store x → y without the evidence and reasoning. The answer could be correct on an in-distribution test, but would be incorrect if found in a contraindication or conflicting source, and/or changes in the guideline. This mechanism is also shown in Figure 2 with hypothetical values.

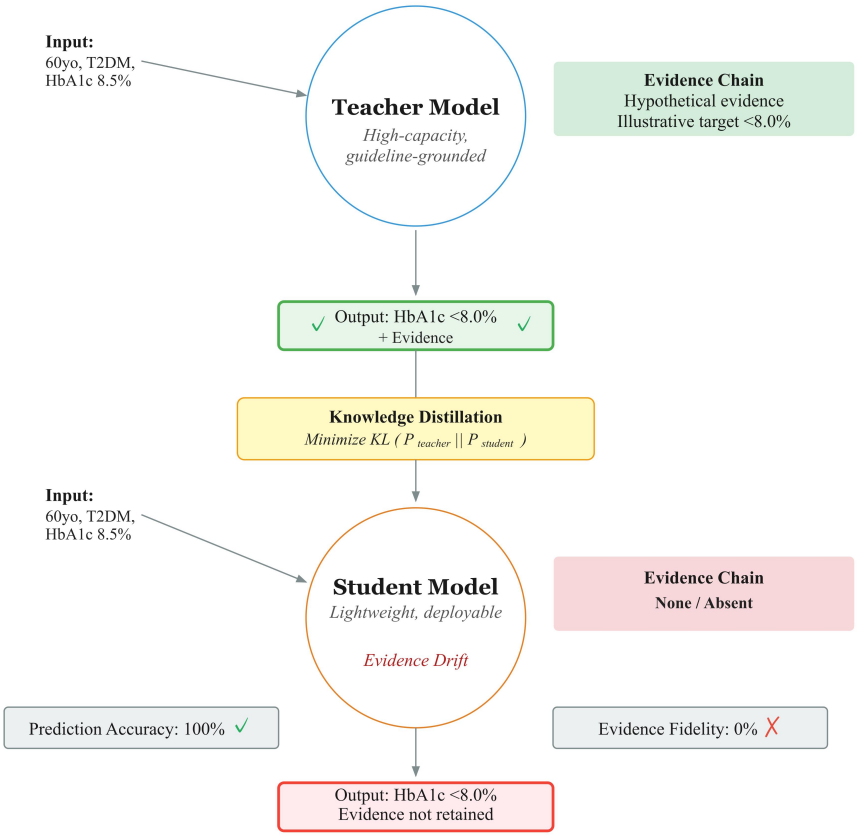


Figure 2. Evidence drift: Conceptual illustration created for this review. The values shown are only hypothetical and are not indicative to show any connection between the accuracy and the fidelity of prediction.

There are four associated risks to evidence drift. Firstly because of the conflict between compression and fidelity it is possible for benchmark retention to cover up the lack of clinical qualifiers. Secondly, temporal brittleness in students occurs when they save a suggestion, but do not have a means to identify a changed guideline. Third, there are or can be diagnoses, medicines, dosages or negation in the diagnosis, medicine, dosage, etc. that leads to semantic drift. Fourth, evaluation myopia rewards accuracy, F1, AUROC, BLEU, or ROUGE without directly testing evidence correctness, guideline conformity, subgroup harm, or deployment safety. There are partially solutions to these failures based on faithful summarization, reference revision, and fact-guided reranking [13,48-50]. The evidence retrieved to support a claim could be irrelevant or misapplied, but retrieval and structured knowledge can lead to more auditable claims [51,52].

We therefore propose evidence-grounded KD: the student learns both the outputs and the connections to versioned evidence, clinically relevant concepts, contraindications, or triggers for abstention. Evidence and reasoning words could be added to a suitable objective,

$$L = L_{KD} + \lambda_E L_{evidence} + \lambda_R L_{reason} + \lambda_C L_{calibration} \quad (2)$$

This objective treats evidence accuracy separately from prose fluency. This formulation provides an avenue to test the rationale of future model development and comparison.

There are different types of evidence preservation. Source fidelity does the evidence have a cited guideline or record and does the recommendation survive compression without a qualifier or negator. Entailment fidelity does the evidence version survive compression without the loss of an important fact, like a contraindication or some exceptions. Coverage fidelity asks whether decisive facts, including contraindications and exceptions, survive compression. Temporal fidelity is it a current evidence version. Finally, decision fidelity whether or not the students will make corresponding recommendations if there is a change in evidence or context for the patient. These levels are different from an auditable evidence chain and an authoritative looking citation that was created.

An evidence-grounded student also needs explicit failure behavior. If retrieval returns conflicting guidance, the model should expose the conflict rather than compress it into one unqualified answer. If a decisive patient variable is absent, the student should request it or abstain. If a source-language concept lacks a validated target-language equivalent, the output should retain the original term, identify the translation uncertainty, and escalate when the ambiguity affects treatment. Such behavior makes uncertainty actionable and limits the conversion of missing knowledge into fluent certainty.

5. Explainability for Clinical Stakeholders

Medical XAI provides explanations of attention and attribution, local surrogate explanations, counterfactual explanations, examples, and generated narratives as outlined below [53-57]. The different approaches have a focus on different aspects of the explanation problem. Visualization in Attention is intuitive, but does not necessarily follow the causal behaviour of the model. SHAP and LIME can be technically complicated open up the possibility of feature sensitivity. Actionable reasoning is supported by counterfactuals, under medical constraint. Materials are easily accessible, reasons and citations are verified with narration. Established rules indicate that explanations, to be clinically useful, should have four characteristics: They are actionable, They are contrastive, They are credible or consistent with medical evidence, They are adapted to the stakeholder.

The explanation evaluation gap is also large. Technical proxy scores and/or convenience sample assessments of explainability are not substitutes for explainability assessments by physicians, checking medication information by pharmacists, and assessing comprehension and actionability by patients/caregivers for the assessment of explainability [58-61]. In such cases, the explanations that are created using natural languages can seem credible even if the evidence provided is minimal [62,63]. Thus, explanation fidelity should encompass attribution alignment, counterfactual consistency, as well as fact verification within the narration and the accuracy of evidence citations and conformability to guidelines [64-66].

The dual-explainability framework provides two kinds of explanation of the same prediction. The clinical perspective is on evidence, quantitative risk, uncertainty, adherence to guidelines, alternatives and contraindications. Patient view emphasizes simple language, steps to take, alerts, and culturally appropriate illustrations. There should be two views, the first role aware, the second language aware and culture aware. There are several ways of adapting dietary examples, measurement conventions, family decision making, health beliefs, reading level and access constraints which cannot be done by simply translating the dietary example into another language [17,67,68]. Separation, with a common core, is illustrated in Figure 3.

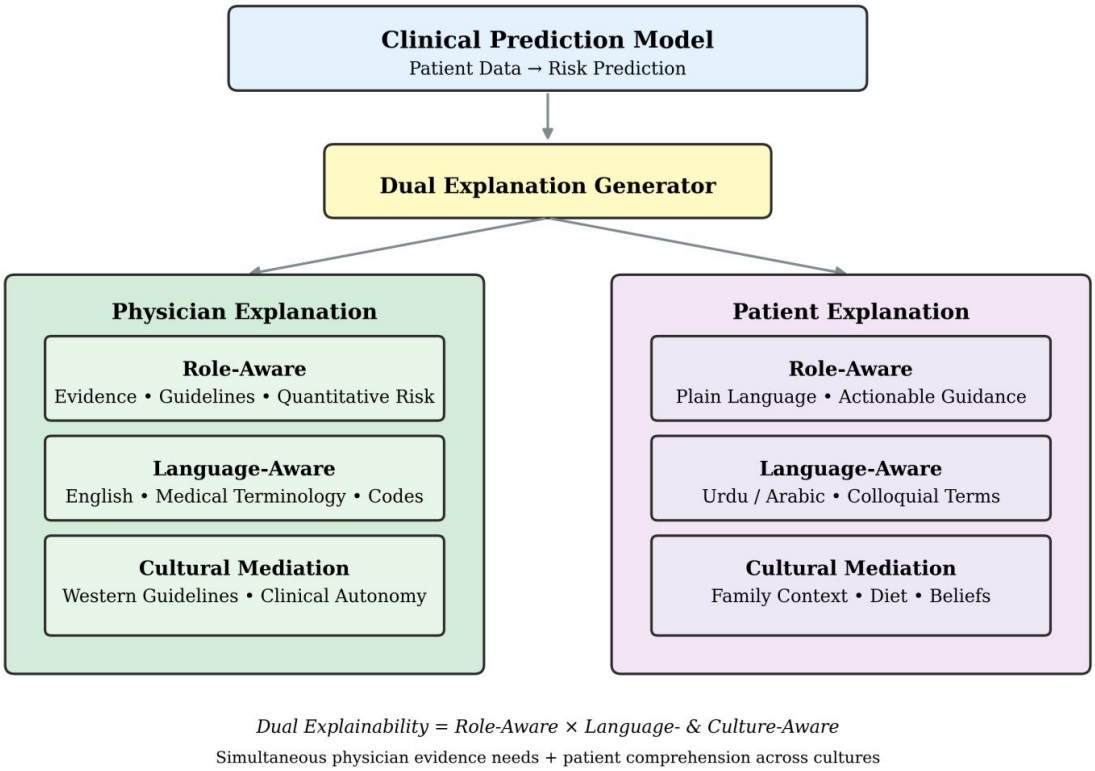


Figure 3. Conceptual dual-explainability framework developed for this review. Both stakeholder perspectives are based on the same prediction, evidence record and uncertainty estimate.

Dual explainability does not mean generating two unrelated narratives. The two views should be based on one unified explanation object that includes the prediction, the supporting and opposing evidence, the important patient variables, uncertainty, limitations and escalation condition. Clinician layer can show options and guidance versions. The patient

layer can express the elements they wish to be understood as actions or questions to the professional. This is an audit signal when there is a difference between the two views. A patient explanation should not, for example, cloud an explanation contraindicated to the clinician.

The following should be 5 questions for judging faithfulness of the explanation: Does the explanation follow model behavior? Is it factually consistent with the record and evidence? Does it serve the purpose for the stated clinical position? Does it make sense in target language? Is it feasible and appropriate in the target culture and care setting? The other 4 answers all need to make use of readability. The decisions when determining the stakeholders should be based on realism and not preference. The clinician is aware of when they would validate and/or challenge a recommendation. The patient is able to explain these next steps, and what the red flags are and what he or she was confused about.

6. Calibration and Uncertainty Across Languages

Discrimination and calibration are two separate issues. AUROC measures discrimination, and F1 summarizes classification performance at a chosen threshold. Calibration assesses agreement between predicted probabilities and observed outcome frequencies. Noisy observations (aleatoric uncertainty) is different from missing knowledge of the model (epistemic uncertainty) in the reviews of calibration and uncertainty [69-71]. Epistemic uncertainty may exist in clinical NLP when the disease under study is uncommon, the drugs or combination of drugs is uncommon, the dialect is not the one that the clinician has been trained in, or the institution that the clinician was trained at is not present.

Post-hoc temperature, Platt, or isotonic scaling learned from English data may lose reliability in a target-language setting. Errors and confidence can change due to changes in language, demographic, institutional and prevalence. Conference signals validation is also important, as stated in studies carried out in the medical field directly [72-74]. This is called cross-lingual calibration degradation and is explained in this paper. The pathway is qualitatively shown in Figure 4.

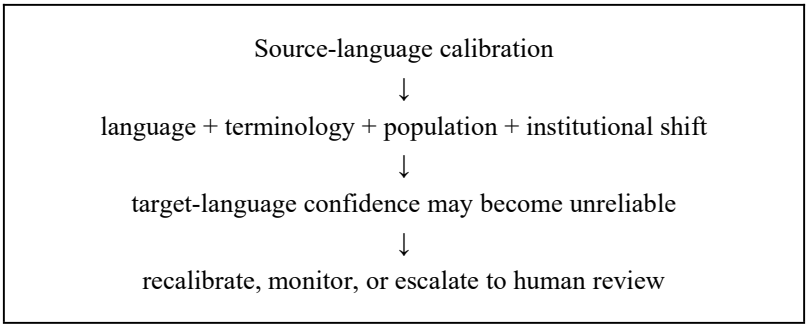


Figure 4. Non-quantitative model developed for this review. It traces cross-lingual calibration degradation from source-language calibration to target-setting monitoring and review.

Explanation-grounded calibration links confidence to the quality of the available evidence/explanation that supports a prediction. Confidence should be lowered, or human review should be initiated, when evidence retrieval is unsuccessful, or when there is low level of agreement with guidelines, conflicting explanations, or uncertainty regarding terminology in the target language. Evaluation during deployment should include the following: clinically selected escalation thresholds, evaluation of reliability diagrams, expected calibration error or Brier score, evaluation of subgroup analysis, evaluation of shift monitoring. This direction is related to abstention-enabled systems [75]. Confidence should always be used for decision-making, but never as a replacement for clinical judgment.

A calibration assessment should be conducted at the level where a decision is made. The frequency curve may be okay but the confidence may be low, in one language, in one hospital, at one disease stage or in one intersectional subgroup. Conversely, a small subgroup estimate might be unstable and not foundation for a stable recalibration. Reporting should therefore include the size of each calibration stratum and the uncertainty associated with each stratum, the prevalence and decision threshold, and the relative cost of false reassurance and of needed escalation. When the number of labels is small, the uncertainty with regard to calibration should be clearly expressed, not hidden in a large number of single estimates.

Testing should be done on an ongoing basis as calibration may be deteriorating without a corresponding decrease in discrimination. Change in confidence can be attributed to changes in coding practice, translation workflow, population mix, version of guidelines/teacher model. Any changes identified during monitoring should be matched with a pre-defined action to investigate, re-calibrate, restrict use of use case or stop automated advice. Regression test and regression tests for evidence retrieval/explanation agreement/abstention and subgroup calibration should be run for every model update.

7. Multidimensional Trust and Compound Failure

There are key concepts for creating trustworthy clinical AI which include truthfulness, safety, robustness, privacy, transparency and accountability [34,35]. Fairness is also a key consideration [76-78]. The review covers six general failure modes: hallucination, miscalibration, terminology drift, subgroup bias, privacy exposure, and accountability gaps.

Involved in distillation are the roles of the teacher developers, distillation engineers, translators, institutions and the clinical users, which are distributed. This disjointed chain can be smoothed with a fluent interface.

These failures interact. Lost evidence can be concealed by a persuasive explanation. A miscalibrated score can suppress appropriate skepticism. Terminology drift can cause problems when finding evidence, as well as when explaining. A concentration of subgroup harm may be obscured in an overall assessment. Beware of many seemingly good averages that may have clinically important failure modes, as demonstrated by hidden stratification [79,80]. Real-world primary-care decision support studies for tasks such as population specificity, site specificity, or prediction of patient outcomes in the context of medical conditions like hypertension, diabetes, kidney disease, and others have shown to highlight the importance of population and site context [81-84]. Studies of diabetes, kidney disease, and clinical outcomes also illustrate why findings depend on the population and clinical task [85-88]. Other studies examine these issues in outcome prediction, treatment effects, and phenotyping [89-91].

7.1 A 12-Category Intersectional Taxonomy

Table 2 presents the proposed taxonomy. The three letters are E, X and C, for evidence, explanation and calibration. The categories are ‘analytical lenses’ that overlap and need empirical support in different languages and contexts.

Table 2. Intersectional failure taxonomy for cross-lingual clinical KD.

Failure mode	E	X	C	Principal risk
Evidence-only drift	×	√	√	Fluent, confident recommendation lacks a defensible source.
Explanation-only loss	√	×	√	Evidence exists but cannot be communicated or interrogated.
Calibration-only failure	√	√	×	Correct evidence and explanation receive unreliable confidence.
Evidence-explanation gap	×	×	√	Unsupported output appears statistically controlled.
Evidence-calibration gap	×	√	×	Persuasive explanation and overconfidence conceal weak evidence.
Explanation-calibration gap	√	×	×	Evidence is present but opacity and confidence error obstruct use.
Triple cascade	×	×	×	Evidence, intelligibility, and uncertainty fail together.
Translation-induced drift	×	√	√	A decisive term, qualifier, negation, or dose is corrupted.
Cultural explanation gap	√	×	√	Literal translation is unusable in the patient's context.
Distribution-shift miscalibration	√	√	×	Source-language confidence is transferred to a different population.
Compounded cross-lingual failure	×	×	×	Linguistic loss amplifies all three dimensions.
Hidden-stratification cascade	×	×	×	Harm concentrates in dialect, demographic, presentation, or site subgroups.

7.2 Worked Clinical Illustration

Imagine the case of a hypothetical fully grown-up person whose native language is Urdu, who suffers from hypertension, diabetes, poor kidney function, and who has not gone to doctor visits without bringing a list of medications. A compressed model suggests that treatment intensity should be increased and has high confidence. The scenario represents multiple interrelated nodes in the taxonomy. If a renal contraindication or guideline condition is not specified during distillation, this will result in drift based on the evidence only. A change in meaning of a medicine or dose word in translation introduces translation-induced drift. If a clinician facing explanation of the risk factors excludes the contraindication, then there is an evidence-explanation gap. If the explanation is being offered to the patient at face level without paying attention to the units, or using culturally irrelevant dietary information, there is a cultural explanation gap. Distribution-shift miscalibration from confidence-misplaced from an English source cohort. The cascade of hidden stratification occurs when these errors are clustered in a regional dialect or in a single comorbidity subpopulation, even though the overall F1 remains unchanged. There are some features that may use risk-score or observational studies, and there are still features that rely on full clinical evaluation that will be used in individual recommendations [82,86,87].

8. Integrated Architecture and Design Principles

The proposed architecture is the assembly of four components. They are both retrieval-augmented and versioned evidence preservation, a simplified prediction model with explicit evidence supervision, and dual clinician-patient explanations, as well as target-population calibration using a system of monitoring of subgroups. It can be integrated with multilingual medical models and retrieval systems [26-28,51]. Figure 5 illustrates the conceptual cascade. It also acknowledges the constraints of privacy, security, and governance [37,92-95].

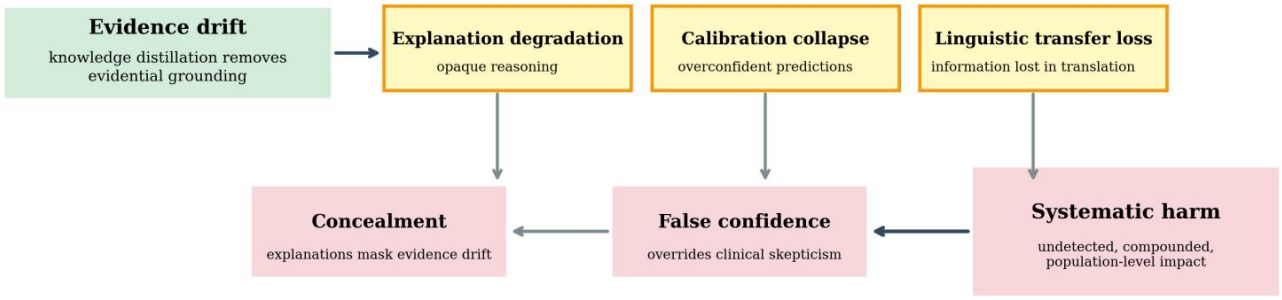


Figure 5. Conceptual cascade architecture developed for this review. It links evidence preservation, dual explanation, calibration, translation, and subgroup monitoring. Arrows indicate possible failure propagation rather than estimated causal effects.

There are 5 Design Principles. First, make sure to record chains of evidence, not outputs. Second, assess the quality of evidence and explanations. Third, perform cross-lingual Trust audits, in each target setting. Fourth, monitor sub-group performance instead of overall performance. Last but not least, test end-to-end cascaded impacts vs. individual failures. Evaluation reform must then make a shift from a focus on accuracy retention to a focus on reasoning and evidence retention, from generic fluency to stakeholder comprehension, and from aggregate performance to subgroup equity.

9. Operational Evaluation Framework

The conceptual framework turns into operational when tested separately and together its dimensions. The five design principles are converted to an evaluation plan in Table 3. When you have the incorrect evidence behind a good report, it is not possible to obtain strong readability, similarly, when there is serious subgroup error without a good average report, the average report can't be used to gain strong calibration.

Table 3. To read about how the trustworthy cross-lingual clinical KD can be assessed.

Dimension	Unit	Indicative tests	Stratification	Failure response
Task performance	Patient, note, entity, or question	Accuracy, F1, or AUROC with uncertainty and comparison against teacher and non-distilled baselines	Language, site, condition, documentation quality	Restrict use or retrain
Evidence fidelity	Claim and source pair or recommendation and guideline pair	Citation validity, entailment, decisive-fact coverage, contraindication retention, update sensitivity	Language, guideline version, specialty	Retrieve again, abstain, verify
Explanation fidelity	Prediction and explanation pair	Attribution stability, constrained counterfactuals, narrative fact checks, clinician consistency review	Role, language, literacy, culture	Suppress or escalate
Calibration	Confidence and outcome pair	Reliability diagrams, Brier score, ECE sensitivity, selective risk and coverage	Language, site, prevalence, demographic intersections	Recalibrate or change threshold
Stakeholder utility	Decision episode	Comprehension, action selection, override behavior, time, and workload	Clinical role, patient group, caregiver context	Redesign interface or workflow
Cascade safety	Complete scenario	Joint perturbation of evidence, translation, explanation, confidence, and subgroup shift	Intersections of language, disease, site, and demography	Suspend deployment pathway

Baseline for teacher, student and task will be measured on identical splits and should be used to evaluate. In the second stage, the interpretations of evidence for fidelity loss in the score retention are compared across the same cases and evidence is further interpreted across the cases for fidelity loss in the score retention. The third stage introduces controlled inference in the other language: terminology differences, negation, units, code switching, dialect and culturally-specific expressions. Under population shift and institutional shift a fourth tests calibration and abstention. The last stage is perturbations to reveal cascade behavior. This set does not imply any fault, it just points to the place of a failure in the system–rather than putting all failures on the general charge of translation or the model.

Also negative controls are required. A system should be tested against reasonable, but false, evidence, dates, text of guidelines, conflicting retrieved documents, and explanations that are fluent without providing support for the prediction. High numbers for these outputs would suggest that there is no direct relationship between the metric and the property evaluated. The levels of expertise and target language proficiency, adjudication, and disagreement are all situation-specific and should be catalogued in human-evaluation scripts as well as explanation and safety judgments.

Deployment and accountability require ongoing oversight because pre-deployment validation cannot anticipate every interaction. It is important that the deployed system retains the evidence version, model version, translation component, confidence transformation, explanation view and escalation outcome (assuming there is appropriate privacy protection). Collates a chain of information in case of otherwise distributed responsibility and aids in looking into a change if a change in responsibilities (often a teacher update) and/or change in term or local workflow brings an unexpected change.

Governance decisions must be made regarding addition of a new language, changing the decision thresholds, approving Medical Terminology, as well as approving the system. Successful transfer does NOT mean clinical and linguistic evaluation. The output should be deterministic and clearly have an owner for high-risk components and should not place ownership on the patient for patient facing text that communicates the uncertainty of the system and its role. Privacy controls must cover intermediate translations, retrieved records, explanation logs, and monitoring data, not only the final prediction.

10. Discussion

There are two ways that multilingual compression can be motivated: efficiency and fairness. Contrary to this, the settings the most likely to require small models are also likely to have the least evidence, calibration data and post deployment monitoring in the target language. The clinical readiness is the outcome of an evaluation of a number of capabilities, not one single capability. All of these tasks focus on a different aspect of this evaluation, including the case of large-model reasoning [96], multilingual QA [29,97]. Other work addresses answer sentence selection [41], radiology summarization [49], imaging, and risk prediction (as explored by [88]).

Stanceski et al. evaluated 100 discharge instructions: 18% had safety issues that were potentially harmful, and 6% included hallucination occurrences, and 3% new medications [44]. The study focuses on the instructions in English for discharge, but results demonstrate the need for clinical verification even for instructions that were patient oriented and readable, notwithstanding.

Other compressions are also planned for the future, which are to be combined with KD, pruning, quantization and matrix factorization. Failure modes they might have in their interactions may not be evident if they are examined individually. Moreover, novel systems for proving, tracing and accountability for the provenance and sources of agents and synthetic notes emerge [98,99]. Here are some hypotheses provided by the framework. Compressed evidence-aware students should maintain guideline links. Dual explanations should enhance fidelity and offer specific understanding for the stakeholders. Target-language calibration should diminish the overconfidence following transfer. Subgroup audits should be cascade aware and should be expected to reveal in-depth what is lacking in the aggregate measures.

11. Limitations and Future Work

This review has not been conducted systematically and therefore literature is not comprehensive. No original model experiment or meta analysis, or patient study conducted. There is a wide variety of tasks, data, metrics and languages contained in the evidence making direct comparison difficult. Little research is conducted in languages except English and the few language combinations, and there is little evidence from under-researched dialects and in clinical settings. Preprints are cited in some recent sources which should be taken with a grain of salt. These 12 categories are a taxonomy need to be validated empirically, while the concepts of dual-explainability, explanation-grounded calibration and integrated architecture must be supported empirically.

Future research should create a multilingual measure of trustworthiness that would be based on evidence anchoring, explanation fidelity, guideline adherence, uncertainty calibration, language/cultural appropriateness and subgroup effects to show evidence's trustworthiness. There should be a multi-lingual representation of assessment participants: clinicians, pharmacists, patients, caregivers and community health workers from all disciplines and institutions. Future research should validate updates of the guidelines in the temporal dimension, changes to the terminology, changes to the distribution, the abstention thresholds, and pipelines including a combination of KD, pruning and quantization.

12. Conclusion

Increased access to clinical NLP can benefit cross-lingual KD, but retained accuracy does not provide a foundation for retained clinical knowledge. Evidence, explanation, calibration, language, culture and safety for subgroups are a system of interdependent components. Five design principles and a 12-category taxonomy, evidence drift, dual explainability, explanation-grounded calibration are a conceptual vocabulary and architecture for examining that system that this review presents. Now these proposals must be put to the test on the ground with the stakeholders in various languages. The extent to which the model is "safe" to compress should be determined not only by whether or not the model's score corresponds to the teacher's score, but whether or not: the model is evidence auditor, clearly understood and valid, calibrated in its target population and can amplify uncertainty.

Acknowledgements

The authors have no additional acknowledgements.

Funding

The authors received no specific funding or grant support for this work.

Ethics Statement

This article is a conceptual review of previously published literature. It did not involve original research with human participants, collection of identifiable patient data, or interventions involving human subjects. Ethics approval was therefore not required.

Data Availability Statement

No new datasets were generated or analysed for this conceptual review. The literature informing the review is cited in the manuscript. MIMIC resources discussed as possible datasets for future validation are hosted by PhysioNet (https://physionet.org/) and are subject to the applicable credentialing and data-use requirements.

Author Contributions

Hamza Qureshi contributed to conceptualization, literature review, development of the framework, visualization, and preparation of the manuscript. Hamid Ghous contributed to conceptual development, interpretation, thematic organization, manuscript review, and coordination of the submission. Both authors read and approved the manuscript.

Conflict of Interest

The authors do not have any conflict of interest.

Generative AI Statement

Generative AI was used only to search for and identify relevant academic papers. No AI-generated text, analysis, or interpretations are included in this manuscript.

References

[1] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature, 2023, 620, 172-180. DOI: 10.1038/s41586-023-06291-2

[2] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375, 2023. DOI: 10.48550/arXiv.2303.13375

[3] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023. DOI: 10.48550/arXiv.2303.08774

[4] Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, et al. Large language models: A survey. arXiv preprint arXiv:2402.06196, 2024. DOI: 10.48550/arXiv.2402.06196

[5] Yenduri G, Ramalingam M, Chemmalar Selvi GC, Supriya Y, Srivastava G, Maddikunta PKR, et al. GPT (Generative Pre-Trained Transformer)–A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, 2024, 12, 54608-54649. DOI: 10.1109/ACCESS.2024.3389497

[6] Jung KH. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations. Healthcare Informatics Research, 2025, 31(2), 114-124. DOI: 10.4258/hir.2025.31.2.114

[7] Yu E, Chu X, Zhang W, Meng X, Yang Y, Ji X, et al. Large language models in medicine: Applications, challenges, and future directions. International Journal of Medical Sciences, 2025, 22(11), 2792-2801. DOI: 10.7150/ijms.111780

[8] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019. DOI: 10.48550/arXiv.1910.01108

[9] Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Nooralahzadeh F, Clifton L, et al. Lightweight transformers for clinical natural language processing. Natural Language Engineering, 2024, 30(5), 887-914. DOI: 10.1017/S1351324923000542

[10] Ibrahim ZM, Bean D, Searle T, Qian L, Wu H, Shek A, et al. A knowledge distillation ensemble framework for predicting short- and long-term hospitalization outcomes from electronic health records data. IEEE Journal of Biomedical and Health Informatics, 2022, 26(1), 423-435. DOI: 10.1109/JBHI.2021.3089287

[11] Lee C, Kumar S, Vogt KA, Munshi M, Tallapudi P, Vogt A, et al. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. Scientific Reports, 2025, 15(1), 27619. DOI: 10.1038/s41598-025-10451-x

[12] Adams G, Zucker J, Elhadad N. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. Proceedings of Machine Learning Research, 2023, 219, 2-30.

[13] Zhang N, Zhang Y, Guo W, Mitra P, Zhang R. FaMeSumm: Investigating and improving faithfulness of medical summarization. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, 10915-10931. DOI: 10.18653/v1/2023.emnlp-main.673
[14] Li W, Wu W, Chen M, Liu J, Xiao X, Wu H. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. arXiv preprint arXiv:2203.05227, 2022. DOI: 10.48550/arXiv.2203.05227
[15] Trienes J, Joseph S, Schlötterer J, Seifert C, Lo K, Xu W, et al. InfoLossQA: Characterizing and recovering information loss in text simplification. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2024, 4263-4294. DOI:10.18653/v1/2024.acl-long.234
[16] Shaitarova A, Zaghir J, Lavelli A, Krauthammer M, Rinaldi F. Exploring the latest highlights in medical natural language processing across multiple languages: A Survey. Yearbook of Medical Informatics, 2023, 32(1), 230-243. DOI: 10.1055/s-0043-1768726
[17] Hershcovich D, Frank S, Lent H, de Lhoneux M, Abdou M, Brandl S, et al. Challenges and strategies in cross-cultural NLP. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, 6997-7013. DOI:10.18653/v1/2022.acl-long.482
[18] García Subies G, Barbero Jiménez Á, Martínez Fernández P. A survey of spanish clinical language models. arXiv preprint arXiv:2308.02199, 2023. DOI: 10.48550/arXiv.2308.02199
[19] Yada S, Nakamura Y, Wakamiya S, Aramaki E. Cross-lingual natural language processing on limited annotated case/radiology reports in English and Japanese: Insights from the real-MedNLP workshop. Methods of Information in Medicine, 2024, 63(5-06), 145-163. DOI: 10.1055/a-2405-2489
[20] Asgari E, El Kheir Y, Sadraei Javaheri MA. MorphBPE: A morpho-aware tokenizer bridging linguistic complexity for efficient LLM training across morphologies. arXiv preprint arXiv:2502.00894, 2025. DOI: 10.48550/arXiv.2502.00894
[21] Xu X, Li M, Tao C, Shen T, Cheng R, Li J, et al. A survey on knowledge distillation of large language models. arXiv preprint arXiv:2402.13116, 2024. DOI: 10.48550/arXiv.2402.13116
[22] Wu T, Tao C, Wang J, Yang R, Zhao Z, Wong N. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In: Proceedings of the 31st International Conference on Computational Linguistics. Association for Computational Linguistics, 2025, 5737-5755. Available from: https://aclanthology.org/2025.coling-main.383/ (accessed on 4 November 2026)
[23] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3), 39, 1-45. DOI: 10.1145/3641289
[24] Qin L, Chen Q, Feng X, Wu Y, Zhang Y, Li Y, et al. Large language models meet NLP: A survey. Frontiers of Computer Science, 2026, 20(11), 2011361. DOI: 10.1007/s11704-025-50472-3
[25] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research, 2023. Available from: https://discovery.ucl.ac.uk/id/eprint/10216468/ (accessed on 4 November 2026)
[26] Wang X, Chen N, Chen J, Wang Y, Zhen G, Zhang C, et al. Apollo: A lightweight multilingual medical LLM towards democratizing medical AI to 6B people. arXiv preprint arXiv:2403.03640, 2024. DOI: 10.48550/arXiv.2403.03640
[27] Gangavarapu A. Introducing L2M3, a multilingual medical large language model to advance health equity in low-resource regions. arXiv preprint arXiv:2404.08705, 2024. DOI: 10.48550/arXiv.2404.08705
[28] Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: A collection of open-source pretrained large language models for medical domains. In: Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, 5848-5864. DOI: 10.18653/v1/2024.findings-acl.348
[29] Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, 675-718. DOI: 10.18653/v1/2023.ijcnlp-main.45
[30] Zhang Y, Ives Z, Roth D. Evidence-based trustworthiness. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019, 413-423. DOI: 10.18653/v1/P19-1040
[31] Chandu KR, Bisk Y, Black AW. Grounding grounding in NLP. In: Findings of the Association For Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, 4283-4305. DOI: 10.18653/v1/2021.findings-acl.375
[32] Kenthapadi K, Sameki M, Taly A. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2024, 6523-6533. DOI: 10.1145/3637528.3671467
[33] Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in medicine: A systematic review with large language models and beyond. medRxiv [Preprint]. 2023, 2023.04.18.23288752. DOI: 10.1101/2023.04.18.23288752
[34] Aljohani M, Hou J, Kommu S, Wang X. A comprehensive survey on the trustworthiness of large language models in healthcare. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics, 2025, 6720-6748. DOI: 10.18653/v1/2025.findings-emnlp.356
[35] Huang Y, Sun L, Wang H, Wu S, Zhang Q, Li Y, et al. TrustLLLM: Trustworthiness in large language models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, 2024, 235, 20166-20270.
[36] He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. Information Fusion, 2025, 118, 102963. DOI: 10.1016/j.inffus.2025.102963
[37] Chen Y, Esmaeilzadeh P. Generative AI in medical practice: In-depth exploration of privacy and security challenges. Journal of Medical Internet Research, 2024, 26, e53008. DOI: 10.2196/53008
[38] Alcoba Inciarte A, Kwon SY, Nagoudi EMB, Abdul-Mageed M. On the utility of pretraining language models on synthetic data. In: Proceedings of The Second Arabic Natural Language Processing Conference. Association for Computational Linguistics, 2024, 265-282. DOI: 10.18653/v1/2024.arabicnlp-1.23
[39] Tan Z, Li D, Wang S, Beigi A, Jiang B, Bhattacharjee A, et al. Large language models for data annotation and synthesis: A survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, 930-957. DOI: 10.18653/v1/2024.emnlp-main.54

[40] Papaioannou JM, Grundmann P, van Aken B, Samaras A, Kyparissidis I, Giannakoulas G, et al. Cross-lingual knowledge transfer for clinical phenotyping. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, 2022, 900-909. Available from: https://aclanthology.org/2022.lrec-1.95/ (accessed on 4 November 2026)
[41] Gupta S, Matsubara Y, Chadha A, Moschitti A. Cross-lingual knowledge distillation for answer sentence selection in low-resource languages. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, 14078-14092. DOI: 10.18653/v1/2023.findings-acl.885
[42] Nguyen TT, Luu AT. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10), 11103-11111. DOI: 10.1609/aaai.v36i10.21359
[43] Basir N, Hakro DN, Khoumbati KUR, Bhatti Z. Leveraging machine-labeled data and cross-lingual transfer for NER in Urdu and Sindhi. Journal of Information & Communication Technology, 2025, 19(1), 1-8.
[44] Stanceski K, Zhong S, Zhang X, Khadra S, Tracy M, Koria L, et al. The quality and safety of using generative AI to produce patient-centred discharge instructions. NPJ Digital Medicine, 2024, 7(1), 329. DOI: 10.1038/s41746-024-01336-w
[45] Tan S, Joty S, Baxter K, Taeihagh A, Bennett GA, Kan MY. Reliability testing for natural language processing systems. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, 4153-4169. DOI: 10.18653/v1/2021.acl-long.321
[46] Litschko R, Müller-Eberstein M, van der Goot R, Weber-Genzel L, Plank B. Establishing trustworthiness: Rethinking tasks and model evaluation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, 193-203. DOI: 10.18653/v1/2023.emnlp-main.14
[47] Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: Systematic review. JMIR Mental Health, 2024, 11, e57400. DOI: 10.2196/57400
[48] Adams G, Shing HC, Sun Q, Winestock C, McKeown K, Elhadad N. Learning to revise references for faithful summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, 4009-4027. DOI: 10.18653/v1/2022.findings-emnlp.296
[49] Xie Q, Zhou J, Peng Y, Wang F. FactReranker: Fact-guided reranker for faithful radiology report summarization. arXiv preprint arXiv:2303.08335, 2023. DOI: 10.48550/arXiv.2303.08335
[50] Piya FL, Beheshti R. Agenticsum: An agentic inference-time framework for faithful clinical text summarization. arXiv preprint arXiv:2602.20040, 2026. DOI: 10.48550/arXiv.2602.20040
[51] Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac: Retrieval-augmented language models for clinical medicine. NEJM AI, 2024, 1(2), AIoa2300068. DOI: 10.1056/AIoa2300068
[52] Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, et al. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. JMIR AI, 2025, 4, e58670. DOI: 10.2196/58670
[53] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(11), 4793-4813. DOI: 10.1109/TNNLS.2020.3027314
[54] Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of explainable artificial intelligence in healthcare. Computers and Electrical Engineering, 2024, 118, 109370. DOI: 10.1016/j.compeleceng.2024.109370
[55] Bharati S, Mondal MRH, Podder P. A review on explainable artificial intelligence for healthcare: Why, how, and when? IEEE Transactions on Artificial Intelligence, 2024, 5(4), 1429-1442, 2024. DOI: 10.1109/TAI.2023.3266418
[56] Mumuni F, Mumuni A. Explainable artificial intelligence (XAI): From inherent explainability to large language models. Array, 2026, 31, 101133. DOI: 10.1016/j.array.2026.101133
[57] Bilal A, Ebert D, Lin B. LLMs for explainable AI: A comprehensive survey. arXiv preprint arXiv:2504.00125, 2025. DOI: 10.48550/arXiv.2504.00125
[58] Donoso-Guzmán I, Kacafírková KS, Szymanski M, Jacobs A, Parra D, Verbert K. A systematic review of user-centred evaluation of explainable AI in healthcare. arXiv preprint arXiv:2506.13904, 2025. DOI: 10.48550/arXiv.2506.13904
[59] Pierce RL, Van Biesen W, Van Cauwenberge D, Decruyenaere J, Sterckx S. Explainability in medicine in an era of AI-based clinical decision support systems. Frontiers in Genetics, 2022, 13, 903600. DOI: 10.3389/fgene.2022.903600
[60] Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery, 2020, 10(5), e1379. DOI: 10.1002/widm.1379
[61] Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Frontiers in Digital Health, 2021, 3, 645232. DOI: 10.3389/fdgth.2021.645232
[62] Cabrero-Daniel B, Sanagustín Cabrero A. Perceived trustworthiness of natural language generators. In: Proceedings of the First International Symposium on Trustworthy Autonomous Systems. ACM, 2023, 1-9. DOI: 10.1145/3597512.3599715
[63] Shailya K, Rajpal S, Krishnan GS, Ravindran B. LExT: Towards evaluating trustworthiness of natural language explanations. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. ACM, 2025, 1565-1587. DOI: 10.1145/3715275.3732104
[64] Ras G, Xie N, van Gerven M, Doran D. Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research, 2022, 73, 329-397. DOI: 10.1613/jair.1.13200
[65] Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access, 2020, 8, 42200-42216. DOI: 10.1109/ACCESS.2020.2976199
[66] Edin J, Maistro M, Maaløe L, Borgholt L, Havtorn JD, Ruotsalo T. An unsupervised approach to achieve supervised-level explainability in healthcare records. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, 4869-4890. DOI: 10.18653/v1/2024.emnlp-main.280
[67] Ding K, Zhou M, Chaudhari A, Zhang S, Metaxas DN. Aligning large language models with healthcare stakeholders: A pathway to trustworthy AI integration. arXiv preprint arXiv:2505.02848, 2025. DOI: 10.48550/arXiv.2505.02848
[68] Dubey A, Yang Z, Hattab G. AI readiness in healthcare through storytelling XAI. In: Proceedings of the First Workshop on Explainable Artificial Intelligence for the Medical Domain (EXPLIMED 2024). CEUR Workshop Proceedings, 2024, 3831, paper 12. Available from: https://ceur-ws.org/Vol-3831/ (accessed on 4 November 2026)
[69] Wang C. Calibration in deep learning: A survey of the state-of-the-art. arXiv preprint arXiv:2308.01222, 2023. DOI: 10.48550/arXiv.2308.01222

[70] He W, Jiang Z, Xiao T, Xu Z, Li Y. A survey on uncertainty quantification methods for deep learning. ACM Computing Surveys, 2026, 58(7), 1-35. DOI: 10.1145/3786319
[71] Gao R, Li T, Tang Y, Xu Z, Kammer M, Antic SL, et al. A comparative study of confidence calibration in deep learning: From computer vision to medical imaging. arXiv preprint arXiv:2206.08833, 2022. DOI: 10.48550/arXiv.2206.08833
[72] Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the confidence of large language models in answering clinical questions: Cross-sectional evaluation study. JMIR Medical Informatics, 2025, 13, e66917. DOI: 10.2196/66917
[73] Yu Y, Gomez-Cabello CA, Haider SA, Genovese A, Prabha S, Trabilsy M, et al. Enhancing clinician trust in AI diagnostics: A dynamic framework for confidence calibration and transparency. Diagnostics, 2025, 15(17), 2204. DOI: 10.3390/diagnostics15172204
[74] Popat R, Ive J. Embracing the uncertainty in human-machine collaboration to support clinical decision-making for mental health conditions. Frontiers in Digital Health, 2023, 5, 1188338. DOI: 10.3389/fdgth.2023.1188338
[75] Islam MR. An explainable agentic AI framework for uncertainty-aware and abstention-enabled acute ischemic stroke imaging decisions. arXiv preprint arXiv:2601.01008, 2026. DOI: 10.48550/arXiv.2601.01008
[76] Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 2023, 7(6), 719-742. DOI: 10.1038/s41551-023-01056-8
[77] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys, 2021, 54(6), 1-35. DOI: 10.1145/3457607
[78] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2023, 2(6), e0000278. DOI: 10.1371/journal.pdig.0000278
[79] Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. ACM, 2020, 2020, 151-159. DOI: 10.1145/3368555.3384468
[80] Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E. A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery, 2022, 12(3), e1452. DOI: 10.1002/widm.1452
[81] Korom R, Kiptinness S, Adan N, Said K, Ithuli C, Rotich O, et al. AI-based clinical decision support for primary care: A real-world study. arXiv preprint arXiv:2507.16947, 2025. DOI: 10.48550/arXiv.2507.16947
[82] Mroz T, Griffin M, Cartabuke R, Laffin L, Russo-Alvarez G, Thomas G, et al. Predicting hypertension control using machine learning. PLoS One, 2024, 19(3), e0299932. DOI: 10.1371/journal.pone.0299932
[83] Araujo-Moura K, Souza L, de Oliveira TA, Rocha MS, De Moraes ACF, Chiavegatto Filho A. Prediction of hypertension in the pediatric population using machine learning and transfer learning: A multicentric analysis of the SAYCARE study. International Journal of Public Health, 2025, 70, 1607944. DOI: 10.3389/ijph.2025.1607944
[84] Bada AB, Garko AB, Gabi D. Hypertension prediction using deep learning with transfer learning techniques. FUDMA Journal of Sciences, 2024, 8(6), 257-263, 2024. DOI: 10.33003/fjs-2024-0806-2855
[85] Liu Y, Mo W, Wang H, Shao Z, Zeng Y, Bi J. Feature selection and risk prediction for diabetic patients with ketoacidosis based on MIMIC-IV. Frontiers in Endocrinology, 2024, 15, 1344277. DOI: 10.3389/fendo.2024.1344277
[86] Zhang W, Yang Z. Association between the triglyceride glucose index and the risk of acute kidney injury in critically ill patients with hypertension: Analysis of the MIMIC-IV database. Frontiers in Endocrinology, 2024, 15, 1437709. DOI: 10.3389/fendo.2024.1437709
[87] SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration. SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. European Heart Journal, 2023, 44(28), 2544-2556. DOI: 10.1093/eurheartj/ehad260
[88] Betzler BK, Chee EYL, He F, Lim CC, Ho J, Hamzah H, et al. Deep learning algorithms to detect diabetic kidney disease from retinal photographs in multiethnic populations with diabetes. Journal of the American Medical Informatics Association, 2023, 30(12), 1904-1914. DOI: 10.1093/jamia/ocad179
[89] Röhr T, Figueroa A, Papaioannou JM, Fallon C, Bressem K, Nejdl W, et al. Revisiting clinical outcome prediction for MIMIC-IV. In: Proceedings of the 6th Clinical Natural Language Processing Workshop. Association for Computational Linguistics, 2024, 208-217. DOI: 10.18653/v1/2024.clinicalnlp-1.18
[90] Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nature Medicine, 2024, 30(4), 958-968. DOI: 10.1038/s41591-024-02902-1
[91] Harrigian K, Tang T, Gonzales A, Cai CX, Dredze M. An eye on Clinical BERT: Investigating language model generalization for diabetic eye disease phenotyping. arXiv preprint arXiv:2311.08687, 2023. DOI: 10.48550/arXiv.2311.08687
[92] Wang F, Preininger A. AI in Health: State of the art, challenges, and future directions. Yearbook of Medical Informatics, 2019, 28(1), 16-26. DOI: 10.1055/s-0039-1677908
[93] Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJPC. Generative AI for transformative healthcare: A comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access, 2024, 12, 31078-31106. DOI: 10.1109/ACCESS.2024.3367715
[94] Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digital Medicine, 2023, 6(1), 210. DOI: 10.1038/s41746-023-00958-w
[95] Sharaf S, Anoop VS. An analysis on large language models in healthcare: A case study of BioBERT. arXiv preprint arXiv:2310.07282, 2023. DOI: 10.48550/arXiv.2310.07282
[96] Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions. Patterns, 2024, 5(3), 100943. DOI: 10.1016/j.patter.2024.100943
[97] Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239, 2022. DOI: 10.48550/arXiv.2201.08239
[98] Tong Z, Liu Y, Fan W, Qin J, Han Z, Gong H, et al. Agentic AI in medicine: Architectures, applications, evaluation, and challenges for clinical translation. arXiv preprint arXiv:2607.25489, 2026. DOI: 10.48550/arXiv.2607.25489
[99] Poulett W, Waterhouse A, Wallace B, Kynoch S, Imaz Blanco A, Spence M, et al. A pipeline for generating longitudinal synthetic clinical notes using large language models. arXiv preprint arXiv:2606.26879, 2026. DOI: 10.48550/arXiv.2606.26879