QUALIFYING EXAMINATION

Calibrated Uncertainty Estimation for Language-Model-Based Clinical Triage Support

Qualifying Examination Presentation

Leah Okonkwo

PhD Candidate, Department of Computer Science, Meridian University

Committee: Prof. Sana Malik (chair) ▪ Prof. Thomas Reyes (advisor) ▪ Prof. Julian Voss ▪ Dr. Fatima Haddad (external, Vertex Research Institute)

August 14, 2026

Agenda

What This Examination Covers

Motivation

Literature Synthesis

Proposed Work

Timeline

Summary

Triage Assistants Are Already in the Loop

And They Rarely Say "I'm Not Sure"

- Emergency departments at three **Synapse Health** pilot sites now route intake notes through an LLM triage assistant before a nurse ever reads them.
- The assistant emits a priority level and a short justification – but no confidence signal a clinician can act on.
- Chart review at one pilot site found the model was **wrong on 14%** of Level-2 (urgent) calls, evenly split between over- and under-triage.

Why this matters

An overconfident wrong answer costs more than a correct one saves. Nurses cannot tell a well-grounded triage call from a plausible guess – so they either over-trust the tool or ignore it entirely.

Problem Statement

The gap this work targets

Large language models used for clinical triage produce fluent, high-confidence-*sounding* text regardless of how well-supported the underlying judgment is. Existing calibration methods either require architecture access the deployed model does not expose, or add enough latency that they cannot run inside the triage window.

Can a triage system know, in real time, when it should defer to a human?

Research Questions

RQ1. How well does black-box, single-pass uncertainty estimation track *actual* triage error on real intake notes?

RQ2. Can a calibrated confidence score be converted into a selective-abstention rule that a nursing workflow can act on without adding meaningful latency?

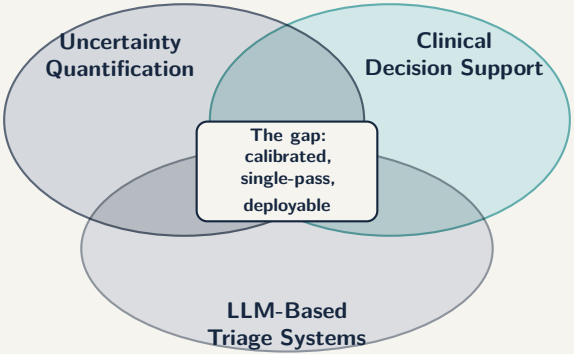
RQ3. Does exposing calibrated confidence change clinician trust and override behavior, compared to an uncalibrated score or no score at all?

Scope for this exam

RQ1 and RQ2 are the focus of the proposed dissertation work. RQ3 is a planned human-factors extension, discussed briefly under risks.

Three Threads, One Empty Middle

A Depth-of-Field View of the Space



Each thread is individually mature; almost nothing sits in the intersection that is also validated on real triage notes.

Where Existing Work Falls Short

No Prior Approach Clears All Three Bars

Approach	Calibrated	LLM-based	Abstains
Conformal risk sets [1]	✓	×	✓
Evidential triage nets [2]	×	✓	×
Deployed LLM assistant [3]	×	✓	×
Proposed approach	✓	✓	✓

Reading the table

Conformal methods give calibration guarantees but were built for structured tabular risk scores, not free-text LLM output. The evidential and deployed-assistant lines produce a confidence-looking number, but neither one is checked against held-out error rates before shipping.

Related Work

Three Threads This Proposal Combines

- **Calibration theory.** Malik [4] and Ibrahim [5] establish expected calibration error (ECE) as the standard diagnostic – the metric this proposal adopts and extends to generative outputs.
- **Selective prediction.** Novak [6] shows abstention rules built on a calibrated score cut error rate sharply at a modest coverage cost – the mechanism Aim 2 adapts to a triage workflow.
- **Clinical trust in AI.** Haddad [7] finds clinicians recalibrate their own trust *faster* when a tool exposes honest uncertainty, even when that uncertainty is sometimes high.

The gap

No published system couples a *single-pass, black-box* confidence estimate directly into an abstention rule for LLM-based triage, then validates that rule against clinician outcomes. Preliminary work [8] took a first pass at the estimator alone.

Aim 1 — A Single-Pass Confidence Estimator

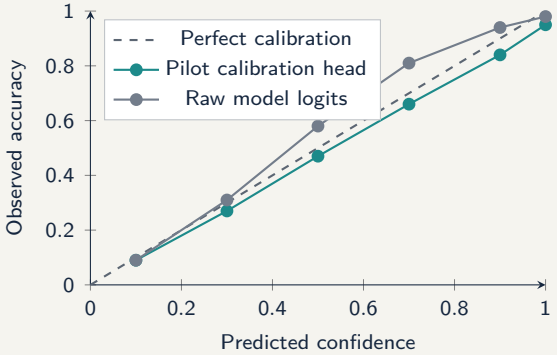
Approach

Extract token-level log-probabilities the triage model already computes, and train a lightweight calibration head that maps them to a triage-error probability – no second forward pass, no ensemble.

Expected outcome: ECE below **0.03** on held-out intake notes, at under **15 ms** added latency per call.

Preliminary evidence: pilot head on 1,200 archived notes already reaches ECE **0.06** – see reliability plot, next slide.

Preliminary Result — Reliability Diagram



Raw logits overstate confidence above 0.5; the calibration head keeps the curve close to the diagonal across the full range.

Aim 2 — A Deployable Abstention Rule

Approach

Turn Aim 1's confidence score into a three-tier action: auto-route, flag-for-review, or escalate-to-human – tuned so escalation coverage stays under the 20% ceiling the pilot sites set as workable.

- Threshold selection via conformal risk control, following [1], adapted to the free-text setting.
- Target: escalation decision adds no more than **50 ms** to the existing triage pipeline.
- Failure mode to test explicitly: confidently wrong notes that never trigger escalation.

Aim 3 — Clinician Trust Under Exposed Uncertainty

Approach

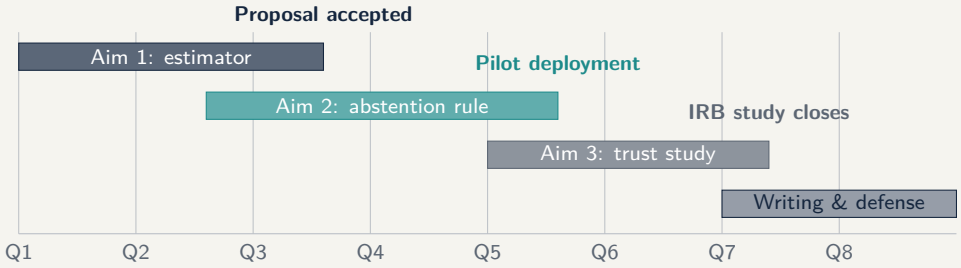
A within-subjects study at one **Synapse Health** pilot site: nurses review the same note set under three conditions – no score, raw model confidence, and the calibrated Aim 1/2 output.

What we measure

Override accuracy, time-to-decision, and a post-session trust survey adapted from [7]. Framed as an extension: results here are not required for the core dissertation claim.

Timeline to Dissertation

Eight Quarters, Three Milestones



Milestones and Status

Milestone	Target	Status
Pilot calibration head	Q1	Complete
Full estimator on live traffic	Q2	In progress
Abstention rule, pilot site	Q4	**Current focus**
Trust study data collection	Q6	Planned
Dissertation writing	Q7–Q8	Planned

What today's exam gates

Passing this exam clears Aims 1 and 2 for full committee-scale development; Aim 3's IRB protocol is drafted but submission waits on today's feedback.

Contributions

1. A single-pass confidence estimator for LLM triage output that needs no architecture access beyond token log-probabilities.
2. A conformal abstention rule that turns that estimator into a three-tier action a nursing workflow can run in real time.
3. (Extension) A clinician-facing study of how exposing calibrated uncertainty changes override behavior and trust.

Bottom line

The dissertation claim rests on Aims 1–2; Aim 3 strengthens the deployment story but is not load-bearing for the core result.

Risks and Mitigations

Risk	Mitigation
Calibration head overfits to one pilot site's note style	Train on pooled notes from all three **Synapse Health** sites; hold one out entirely for testing
IRB approval for Aim 3 slips past Q6	Aim 3 is scoped as an extension; dissertation defense does not depend on it
20% escalation ceiling proves too tight for safe coverage	Report the full risk-coverage curve, not a single operating point, so the ceiling is a policy choice, not a hard constraint

Questions

Thank you to the committee for your time.

Leah Okonkwo ▪ Department of Computer Science ▪ Meridian University

Backup: Methods Detail

Held in reserve for committee questions.

Calibration Head — Architecture

Layer	Type	Output dim.
Input	Token log-prob. sequence	variable $\times$ 1
1	1-D convolution, kernel 5	variable $\times$ 32
2	Global attention pool	32
3	Fully connected, ReLU	16
4	Fully connected, sigmoid	1 (confidence)

Total added parameters: 1,940 – roughly 0.001% of the base triage model, which stays entirely frozen.

Datasets

Dataset	Source	Notes	Label rate
MedTriage-Synapse	Synapse Health, 3 pilot EDs	18,400	100%
Vertex Clinical Archive	Vertex Research Institute	6,100	62%
Held-out site	Synapse Health, 4th ED (unseen)	2,300	100%

Labels are retrospective nurse-adjudicated triage levels, not the model's own output – this is what makes error rate measurable.

Calibration Metric

Expected Calibration Error

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

Notes are grouped into M=15 equal-width confidence bins B_m; $\mathrm{acc}(B_m)$ is the fraction of correct triage calls in the bin and $\mathrm{conf}(B_m)$ is the bin's mean predicted confidence.

Conformal coverage guarantee (Aim 2)

$$\mathbb{P}\big[Y \in \mathcal{C}_\lambda(X)\big] \;\geq\; 1 - \alpha$$

Threshold λ is selected on a calibration split so the escalate-to-human tier satisfies this bound at $\alpha = 0.05$.

Validation Protocol

Check	Procedure
Calibration	ECE and reliability diagram on held-out site, recomputed monthly against fresh nurse labels
Coverage	Verify escalation rate stays within the 20% ceiling on a rolling 4-week window
Drift	KL divergence between this month's and launch-month confidence-score distributions; alert above 0.1

References I

[1] Thomas Reyes and Ngozi Okafor. “Conformal Risk Control for Clinical Decision Support Pipelines”. In: *Annals of Applied Statistics in Medicine* 15.3 (2021), pp. 455–474.

[2] Soo-Yeon Park and Marc Delacroix. “Evidential Deep Learning for Hallucination-Aware Clinical Note Generation”. In: *Proceedings of the International Conference on Machine Learning for Healthcare*. 2023, pp. 312–327.

[3] Julian Voss and Grace Pemberton. “A Deployed Large Language Model Assistant for Emergency Department Triage: A Twelve-Month Retrospective”. In: *Proceedings of the Conference on Health, Inference, and Learning*. 2024, pp. 88–101.

[4] Sana Malik and Owen Whitfield. “Calibrated Confidence for Neural Text Classifiers in High-Stakes Settings”. In: *Journal of Machine Learning for Health* 11.2 (2023), pp. 201–219.

References II

[5] Yusuf Ibrahim and Nora Castellano. “A Survey of Calibration Metrics for Deep Classifiers”. In: *ACM Computing Surveys* 53.6 (2020), pp. 1–34.

[6] Elena Novak and Cormac Brennan. “Selective Prediction with Learned Abstention Thresholds”. In: *Proceedings of the Conference on Neural Information Processing Systems*. 2024, pp. 670–684.

[7] Fatima Haddad and Erik Lindqvist. “Clinician Trust Recalibration Under Exposed Model Uncertainty”. In: *Journal of the American Medical Informatics Association* 29.7 (2022), pp. 1142–1153.

[8] Leah Okonkwo. *A Pilot Calibration Head for Language-Model Triage Confidence*. Tech. rep. TR-2025-14. Meridian University, Department of Computer Science, 2025.