Clinical Text Classification: Robustness, Calibration, and Rationale Consistency Analysis

Team Members: Peter Ozo-ogueji (po3783a@american.edu)
Project Type: Custom Project
Mentor: Prof. Mousavi

Abstract

Clinical text classifiers must maintain consistent predictions under paraphrases and section-order changes while producing calibrated confidence scores that clinicians can trust. We evaluate two rationale-aware models—a BiLSTM with attention-based rationale extraction and a lightweight Transformer with gradient-derived rationales. Our comprehensive reliability assessment includes: (1) a clinical paraphrase probe testing label and rationale stability across synonym, abbreviation, and negation variations; (2) section-order stress testing via shuffle and ablation of clinical note sections (History/Exam/Assessment/Plan); and (3) dual confidence calibration using temperature scaling for predictions and power scaling for rationales. On a chest-pain note dataset, BiLSTM exhibits superior paraphrase consistency (90.6% label agreement vs 74.2% for Transformer) but requires aggressive calibration (59% NLL reduction). Transformer models are better pre-calibrated but less robust to textual variations. Dual calibration reduces Expected Calibration Error (ECE) by 15-26% across models. Both architectures show concerning over-reliance on History sections (≈0.92 probability drop on ablation), suggesting insufficient weighting of objective clinical findings. We release artifacts including reliability diagrams, section ablation reports, paraphrase consistency analyses, and calibrated embeddings for reproducibility.

1. Introduction

Clinical NLP systems operate in high-stakes environments where reliability is as critical as accuracy. Beyond performance metrics, clinicians need systems that: (i) behave consistently under benign textual variations (paraphrases of equivalent clinical content), (ii) remain robust to incomplete or reordered documentation, and (iii) provide well-calibrated confidence estimates aligned with empirical outcome frequencies. Model explanations (rationales) must also remain stable and clinically meaningful across input variations to support auditing and error analysis.

Traditional NLP evaluation focuses on accuracy, precision, and recall, often neglecting critical reliability dimensions that determine safe clinical deployment. A model might achieve 90% accuracy yet be dangerously overconfident, assigning 98% probability to predictions that are wrong 15% of the time. Similarly, models may flip predictions when "denies chest pain" becomes "reports no chest discomfort," despite identical clinical meaning. Such brittleness undermines clinical trust and creates liability risks.

We build on previous work developing compact clinical text encoders (Week 2) and rationale-aware architectures (Week 3) to conduct comprehensive reliability testing. Our evaluation framework addresses three critical questions: (1) Do models maintain consistent decisions across paraphrased clinical content? (2) Are predictions robust to document structure variations common in clinical practice? (3) Can we calibrate both prediction confidence and rationale confidence to reflect true reliability?

Contributions:

- A domain-aware paraphrase generation system for clinical text with medical synonyms, abbreviation expansions, and negation variations
- Section-order stress testing framework quantifying model sensitivity to document structure
- Dual calibration methodology jointly optimizing prediction and rationale confidence
- Comprehensive artifacts (reliability diagrams, JSON reports, calibrated embeddings) supporting reproducibility

2. Background and Key Concepts

2.1 Clinical Reliability Challenges

Clinical text classification faces unique reliability requirements beyond general NLP. We examine critical dimensions with medical discharge note examples:

Model Robustness measures whether models maintain consistent predictions under clinically equivalent input variations. Consider these equivalent statements:

```
Version A: "Patient denies chest pain and shortness of breath"
Version B: "Individual reports no chest discomfort or dyspnea"
```

A fragile model might classify Version A as low cardiac risk but Version B as high risk, despite identical clinical meaning. This brittleness arises from over-reliance on specific lexical patterns rather than semantic understanding.

Paraphrase Consistency evaluates prediction stability across equivalent clinical expressions. Medical documentation exhibits substantial lexical variation:

```
Expression 1: "Patient complains of severe chest pain radiating to left arm"
Expression 2: "Individual reports intense chest discomfort extending to left upper extremity"
Expression 3: "Subject describes significant chest pain spreading to left limb"
```

A consistent model should assign similar acute cardiac event probabilities to all three descriptions. Inconsistent models may vary by 20+ percentage points, creating documentation-dependent diagnoses.

Section-Order Invariance requires stable predictions regardless of clinical note organization. Medical content doesn't change with section ordering:

```
Standard order:
  History: Patient reports chest pain for 2 hours
  Physical Exam: Diaphoretic, tachycardic
  Assessment: Possible MI
  Plan: ECG, troponins

Shuffled order:
  Assessment: Possible MI
  History: Patient reports chest pain for 2 hours
  Plan: ECG, troponins
  Physical Exam: Diaphoretic, tachycardic
```

Both should yield identical predictions. Section-order sensitivity creates unreliable behavior in emergency settings where documentation may be incomplete or non-standard.

Calibration Quality aligns predicted probabilities with actual outcome frequencies. Well-calibrated models satisfy: when predicting X% confidence, outcomes occur X% of the time. Poor calibration manifests as:

```
Poorly calibrated:
  100 patients with 90% MI prediction → 60 actually have MI (30% error)

Well calibrated:
  100 patients with 90% MI prediction → 90 actually have MI (aligned)
```

Miscalibration leads to inappropriate clinical decisions—overconfident predictions may skip necessary testing, while underconfident predictions trigger excessive workups.

Overconfident Predictions occur when models assign extreme probabilities (near 0% or 100%) more frequently than warranted by accuracy. Clinical example:

Overconfident model:
Note: "Acute chest pain with radiation to left arm, diaphoretic"
Prediction: 98% MI probability
Reality: Wrong 15% of the time on such confident predictions

Realistic model:
Same note
Prediction: 75% MI probability
Reality: Wrong 25% of the time (calibrated)

Overconfidence is particularly dangerous in clinical settings where uncertain cases require careful human review and additional testing.

Stress Testing systematically challenges models to reveal failure modes. Section ablation quantifies feature dependencies:

Full note: "History: Chest pain 2 hours. Exam: Diaphoretic, tachycardic. Assessment: Rule out MI. Plan: ECG"
Prediction: 85% MI risk

Remove History: "Exam: Diaphoretic, tachycardic. Assessment: Rule out MI. Plan: ECG"
Prediction: 45% MI risk (40 percentage point drop)

Remove Exam: "History: Chest pain 2 hours. Assessment: Rule out MI. Plan: ECG"
Prediction: 75% MI risk (10 percentage point drop)

Large History drops reveal concerning over-reliance on subjective patient reports versus objective examination findings. In practice, incomplete patient histories are common in emergency scenarios.

Paraphrase Generation Systems create domain-aware text variations for systematic robustness testing. Medical-specific transformations include:

Synonym substitution:
"Patient denies chest pain" → "Individual reports no chest discomfort"
"Acute onset" → "Sudden onset" → "Sharp onset"

Abbreviation expansion:
"EKG shows ST elevation" → "ECG shows ST elevation" → "Electrocardiogram shows ST elevation"
"Patient is SOB" → "Patient has shortness of breath" → "Patient experiences dyspnea"

Negation variation:
"No chest pain" → "Without chest pain" → "Absence of chest pain" → "Denies chest pain"

These transformations preserve clinical semantics while testing lexical robustness.

Post-hoc Calibration Techniques improve probability reliability without retraining. Temperature scaling adjusts raw model outputs:

```
Uncalibrated:
  Raw logit: 2.3
  Probability: sigmoid(2.3) = 90.9%
  Actual accuracy at this confidence: 75%

Temperature calibrated (T=1.8):
  Scaled logit: 2.3/1.8 = 1.28
  Probability: sigmoid(1.28) = 78.2%
  Now matches actual accuracy
```

Temperature T>1 reduces overconfidence by "cooling" probability distributions, pushing extreme values toward 0.5.

Rationale Reliability ensures model explanations remain consistent and clinically meaningful. Consider rationale shifts across paraphrases:

```
Original: "Patient complains of severe chest pain, positive for cardiac risk factors"
Model rationales: ["severe" (0.23), "chest" (0.19), "pain" (0.17), "cardiac" (0.15), "risk" (0.12)]

Paraphrase: "Individual reports intense chest discomfort, has cardiac risk factors"
Model rationales: ["reports" (0.31), "chest" (0.24), "factors" (0.19), "cardiac" (0.15)]
```

The focus shift from clinically critical "severe/intense" to generic verb "reports" indicates unreliable explanation quality. Such instability undermines clinical trust in model reasoning.

2.2 Calibration Metrics

Three complementary metrics quantify calibration quality:

Expected Calibration Error (ECE) measures alignment between confidence and accuracy across probability bins.

Computation process:

1. Partition predictions into M bins by confidence: $B_1 = [0, 0.1)$, $B_2 = [0.1, 0.2)$, ..., $B_{10} = [0.9, 1.0]$
2. For each bin B_m, calculate:
   - Bin accuracy: $\text{acc}(B_m) = (\#\text{ correct predictions in } B_m) / |B_m|$

- Bin confidence: conf(B_m) = average predicted probability in B_m

3. Compute weighted error:

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

Medical Example:

Bin [80%, 90%]: 10 MI predictions
  Average confidence: 85%
  Actual MI cases: 6/10 = 60% accuracy
  Bin error: |85% - 60%| = 25%

Bin [90%, 100%]: 5 MI predictions
  Average confidence: 95%
  Actual MI cases: 5/5 = 100% accuracy
  Bin error: |95% - 100%| = 5%

ECE = (10/15 × 0.25) + (5/15 × 0.05) = 0.183

Perfect calibration (ECE = 0) means accuracy equals confidence in every bin.

Negative Log-Likelihood (NLL) heavily penalizes confident incorrect predictions.

Formula:

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Medical Example:

Patient 1: True MI (y=1), Model predicts 90% (p=0.90)
  Contribution: -log(0.90) = 0.105 (low penalty for correct confidence)

Patient 2: True MI (y=1), Model predicts 40% (p=0.40)
  Contribution: -log(0.40) = 0.916 (high penalty for underconfidence)

Patient 3: No MI (y=0), Model predicts 85% (p=0.85)
  Contribution: -log(1-0.85) = -log(0.15) = 1.897 (very high penalty!)

Patient 4: No MI (y=0), Model predicts 10% (p=0.10)
  Contribution: -log(0.90) = 0.105 (low penalty for correct confidence)

NLL = (0.105 + 0.916 + 1.897 + 0.105) / 4 = 0.756

Lower NLL indicates better calibration. Confident wrong predictions (Patient 3) receive extreme penalties due to logarithmic scaling.

Brier Score computes mean squared error between predicted probabilities and binary outcomes.

Formula:

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$$

Medical Example:

Patient 1: True MI (y=1), Model predicts 90% (p=0.90)
  Squared error: (0.90 - 1)² = 0.01

Patient 2: True MI (y=1), Model predicts 40% (p=0.40)
  Squared error: (0.40 - 1)² = 0.36

Patient 3: No MI (y=0), Model predicts 85% (p=0.85)
  Squared error: (0.85 - 0)² = 0.7225

Patient 4: No MI (y=0), Model predicts 10% (p=0.10)
  Squared error: (0.10 - 0)² = 0.01

Brier = (0.01 + 0.36 + 0.7225 + 0.01) / 4 = 0.276

Brier score range: 0 (perfect) to 1 (worst). Score of 0.25 equals random guessing when base rate is 50%.

The Brier score decomposes into:

$$\text{Brier} = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$$

where Reliability measures calibration error, Resolution measures discrimination ability, and Uncertainty reflects dataset base rates.

2.3 Related Work

Clinical text classification: BiLSTM and Transformer architectures have been extensively applied to EHR note classification. However, most work emphasizes accuracy over reliability, with limited attention to calibration and robustness in deployment settings.

Model calibration: Temperature scaling is a standard post-hoc calibration technique in computer vision. Clinical text applications remain underexplored, particularly for rationale-based models where both prediction and explanation confidence require calibration.

Robustness evaluation: NLP robustness studies primarily address general domains. Clinical text introduces unique challenges including specialized terminology, abbreviation conventions, and templated documentation structures requiring domain-specific evaluation frameworks.

Rationales and faithfulness: Attention-based and gradient-based explanation methods are widely used for interpretability. However, consistency under input variation and faithfulness to model decision-making remain active research areas, with limited work in clinical settings.

3. Methodology

3.1 Rationale-Aware Architectures

We implement two models with distinct rationale extraction mechanisms:

BiLSTM with Attention-Based Rationales

Architecture components:

Embedding layer:

$$\mathbf{e}_t = \mathbf{E}[w_t] \in \mathbb{R}^{d_e}$$

where $\mathbf{E} \in \mathbb{R}^{V \times d_e}$ is the trainable embedding matrix, V is vocabulary size, and d_e is embedding dimension.

Bidirectional LSTM:

Forward pass:

$$\mathbf{h}_t^{(\text{forward})} = \text{LSTM}(\mathbf{e}_t, \mathbf{h}_{t-1}^{(\text{forward})})$$

Backward pass:

$$\mathbf{h}_t^{(\text{backward})} = \text{LSTM}(\mathbf{e}_t, \mathbf{h}_{t+1}^{(\text{backward})})$$

Concatenated hidden state:

$$\mathbf{h}_t = [\mathbf{h}_t^{(\text{forward})}; \mathbf{h}_t^{(\text{backward})}] \in \mathbb{R}^{2H}$$

where H is hidden dimension.

Rationale head:

Compute attention scores:

$$a_t = \mathbf{W}_r \mathbf{h}_t + b_r \in \mathbb{R}$$

Normalize via softmax:

$$r_t = \frac{\exp(a_t)}{\sum_{j=1}^{L} \exp(a_j)}$$

where $\mathbf{W}_r \in \mathbb{R}^{1 \times 2H}$ and $b_r \in \mathbb{R}$ are learned parameters. The softmax ensures $\sum_{t=1}^{L} r_t = 1$, creating a probability distribution over tokens.

Document representation:

$$\mathbf{z} = \sum_{t=1}^{L} r_t \mathbf{h}_t \in \mathbb{R}^{2H}$$

This weighted sum emphasizes tokens with high rationale weights.

Classification:

$$\hat{y} = \sigma(\mathbf{W}_c \mathbf{z} + b_c)$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, $\mathbf{W}_c \in \mathbb{R}^{1 \times 2H}$, and $b_c \in \mathbb{R}$.

Rationale tokens are the top-k by score a_t or normalized weight r_t.

Transformer with Gradient-Based Rationales

Architecture components:

Token and position embeddings:

$$\mathbf{x}_t = \mathbf{E}[w_t] + \mathbf{P}[t] \in \mathbb{R}^{d_e}$$

where $\mathbf{P} \in \mathbb{R}^{L \times d_e}$ provides learned positional information, L is maximum sequence length.

Multi-head self-attention:

For each head $i \in \{1, ..., h\}$:

Query projection:

$$\mathbf{Q}^{(i)} = \mathbf{X}\mathbf{W}_Q^{(i)}$$

Key projection:

$$\mathbf{K}^{(i)} = \mathbf{X}\mathbf{W}_K^{(i)}$$

Value projection:

$$\mathbf{V}^{(i)} = \mathbf{X}\mathbf{W}_V^{(i)}$$

Attention computation:

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{Q}^{(i)}{\mathbf{K}^{(i)}}^T}{\sqrt{d_k}}\right)\mathbf{V}^{(i)}$$

Multi-head output:

$$\text{MultiHead}(\mathbf{X}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}_O$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d_e \times d_k}$ and $\mathbf{W}_O \in \mathbb{R}^{hd_k \times d_e}$.

Feed-forward network:

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$$

Transformer block with residuals:

First sublayer:

$$\mathbf{z}_1 = \text{LayerNorm}(\mathbf{X} + \text{MultiHead}(\mathbf{X}))$$

Second sublayer:

$$\mathbf{z}_2 = \text{LayerNorm}(\mathbf{z}_1 + \text{FFN}(\mathbf{z}_1))$$

Document pooling: Max or mean pooling over non-padding tokens.

Gradient-based rationale extraction:

Compute gradient norm with respect to input embeddings:

$$I_t = \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{x}_t} \right\|_2 = \sqrt{\sum_{d=1}^{d_e} \left(\frac{\partial \mathcal{L}}{\partial x_{t,d}} \right)^2}$$

where $\mathcal{L}$ is classification loss. High gradient magnitudes indicate tokens whose perturbation significantly affects predictions.

3.2 Domain-Aware Paraphrase Generation

We develop a clinical paraphrase generator with four transformation strategies:

Medical synonym substitution: $$\text{synonyms} = \begin{cases} \text{patient} \rightarrow {\text{individual, subject, person, case}} \\ \text{denies} \rightarrow {\text{reports no, states no, refutes, negates}} \\ \text{acute} \rightarrow {\text{sudden, severe, sharp, intense}} \\ \text{chronic} \rightarrow {\text{persistent, ongoing, long-term, prolonged}} \end{cases}$$

Abbreviation expansion: $$\text{abbreviations} = \begin{cases} \text{EKG} \rightarrow {\text{ECG, electrocardiogram}} \ \text{SOB} \rightarrow {\text{shortness of breath, dyspnea}} \ \text{MI} \rightarrow {\text{myocardial infarction, heart attack}} \end{cases}$$

Negation variation: $$\text{negations} = \begin{cases} \text{denies} \rightarrow {\text{reports absence of, states no, negative for}} \ \text{no} \rightarrow {\text{absence of, without, lacking}} \ \text{without} \rightarrow {\text{in absence of, lacking, devoid of}} \end{cases}$$

Multi-token replacement: Bounded edits preserving clinical semantics.

Generation example:

> Original:
> "Patient denies chest pain and reports no shortness of breath during examination"
>
> Generated paraphrases:
> 1. "Individual reports no chest discomfort and states no shortness of respiration during assessment"
> 2. "Subject refutes chest pain and denies difficulty with breathing during evaluation"
> 3. "Person reports absence of chest discomfort and without dyspnea during exam"
> 4. "Case denies chest discomfort and negative for SOB during physical examination"

Each paraphrase maintains clinical equivalence while varying lexical surface forms.

Consistency metrics:

Label consistency rate:

$$\text{Consistency} = \frac{1}{N} \sum_{i=1}^{N} 1[\text{label}(x_i) = \text{label}(\tilde{x}_i)]$$

where $\tilde{x}_i$ is a paraphrase of x_i.

Confidence gap:

$$\text{Gap} = \mathbb{E}[|p(\tilde{x}) - p(x)|]$$

Rationale overlap:

$$\text{Overlap} = \frac{|\text{top-k}(x) \cap \text{top-k}(\tilde{x})|}{|\text{top-k}(x)|}$$

3.3 Section-Order Stress Testing

Clinical notes typically follow structured formats. We implement heuristic section detection:

Section identification: Pattern matching for keywords:

- History: "history of present illness", "HPI", "chief complaint"
- Exam: "physical exam", "PE:", "examination", "vital signs"
- Assessment: "assessment", "impression", "diagnosis"
- Plan: "plan:", "recommendations", "follow-up"

Stress testing procedures:

Shuffle test: Randomly permute detected sections and measure:

$$\Delta_{\text{shuffle}} = |p(\text{shuffled}) - p(\text{original})|$$

Ablation test: Systematically remove each section s and compute:

$$\Delta_{\text{ablate}}(s) = p(\text{original}) - p(\text{without } s)$$

Large drops indicate high dependency on specific sections.

Example analysis:

Original structure:
History: Patient reports substernal chest pain worsening with exertion, relieved by rest
Physical Exam: No acute findings, patient appears comfortable
Assessment: Atypical chest pain in young female
Plan: Anxiety-related symptoms noted, recommend follow-up

Shuffled:
Assessment: Atypical chest pain in young female
History: Patient reports substernal chest pain worsening with exertion, relieved by rest
Plan: Anxiety-related symptoms noted, recommend follow-up
Physical Exam: No acute findings, patient appears comfortable

Ablated (remove History):
Physical Exam: No acute findings, patient appears comfortable
Assessment: Atypical chest pain in young female
Plan: Anxiety-related symptoms noted, recommend follow-up

3.4 Dual Confidence Calibration

We calibrate both prediction and rationale confidence using temperature-based scaling.

Prediction calibration via temperature scaling:

Raw model logits are scaled before sigmoid:

$$p_{\text{cal}} = \sigma\left(\frac{\text{logit}(p)}{T_{\text{pred}}}\right) = \frac{1}{1 + \exp(-\text{logit}(p)/T_{\text{pred}})}$$

Temperature effect:

- $T_{\text{pred}} > 1$: Reduces overconfidence by flattening probability distribution
- $T_{\text{pred}} < 1$: Increases confidence sharpness
- $T_{\text{pred}} = 1$: No scaling (original model)

Optimal temperature minimizes Brier score on calibration split:

$$T^*_{\text{pred}} = \arg\min_{T} \frac{1}{N} \sum_{i=1}^{N} \left(\sigma\left(\frac{\text{logit}(p_i)}{T}\right) - y_i\right)^2$$

Rationale calibration via power scaling:

Rationale confidence aggregates attention statistics:

$$r_{\text{conf}} = \frac{\max(\mathbf{w}) \times (1 + \text{mean}(\mathbf{w}))}{1 + \text{std}(\mathbf{w})}$$

Components:

- $\max(\mathbf{w})$: Peak attention weight (strongest focus)
- $\text{mean}(\mathbf{w})$: Average attention level (overall engagement)
- $\text{std}(\mathbf{w})$: Attention spread (focus vs. diffusion)

Numerator increases when both peak and average attention are high. Denominator decreases confidence when attention is scattered.

Power scaling:

$$r_{\text{cal}} = \text{clip}(r^{1/T_{\text{rat}}}, 0, 1)$$

Temperature effect:

- $T_{\text{rat}} > 1$: Compresses high confidences toward moderate values
- $T_{\text{rat}} < 1$: Amplifies confidence differences
- $T_{\text{rat}} = 1$: No scaling

For rationale calibration, we use a proxy target correlating confidence with prediction extremeness:
$$\text{target}_i = 2 \times |y_i - 0.5| = \begin{cases}
2y_i & \text{if } y_i \geq 0.5 \\
2(1 - y_i) & \text{if } y_i < 0.5
\end{cases}$$

This assigns high target confidence (near 1) to clear cases (probabilities near 0 or 1) and low confidence (near 0) to uncertain cases (probabilities near 0.5).

Optimal rationale temperature:

$$T_{\text{rat}}^* = \arg\min_T \frac{1}{N} \sum_{i=1}^{N} \left(\text{clip}(r_i^{1/T}, 0, 1) - \text{target}_i \right)^2$$

Grid search optimization:

Search space: $T_{\text{pred}}, T_{\text{rat}} \in [0.1, 5.0]$ with 50 equally-spaced points (2,500 combinations).

For each temperature pair:

1. Apply scaling to calibration set
2. Compute Brier score
3. Track optimal parameters

4. Experiments

4.1 Dataset

We evaluate on a curated chest-pain clinical note dataset containing 12 notes covering diverse scenarios:

Note types:

- Emergency Department (ED): Acute presentations

- Progress: Inpatient monitoring
- Cardiology: Specialist consultations
- Psychiatry: Psychosomatic presentations
- Musculoskeletal (MSK): Non-cardiac chest pain
- Pulmonary: Respiratory causes
- Gastroenterology (GI): Reflux-related symptoms

Clinical examples:

> Negative case (Label: 0):
> "Patient denies chest pain and reports no shortness of breath during examination"
> Note Type: ED
>
> Positive case (Label: 1):
> "Acute chest pain with radiation to left arm, patient appears distressed and diaphoretic"
> Note Type: ED
>
> Complex case (Label: 1):
> "Patient complains of severe chest pain, positive for cardiac risk factors, EKG pending"
> Note Type: Cardiology
>
> Challenging case (Label: 0):
> "Chest tightness with anxiety symptoms, patient reports recent stressors"
> Note Type: Psychiatry

Binary classification target: Acute cardiac event risk (positive vs. negative).

4.2 Experimental Configuration

BiLSTM configuration:

- Embedding dimension: 128
- Hidden units: 128 × 2 (bidirectional)
- Dropout: 0.3 (higher regularization)
- Max sequence length: 256 tokens
- Pooling: Rationale-weighted attention or max/mean
- Rationale extraction: Attention-based

Transformer configuration:

- Embedding dimension: 128
- Architecture: 1 block, 4 attention heads
- Feed-forward dimension: 4× embedding dimension (512)
- Dropout: 0.1 (lower dropout)
- Max sequence length: 256 tokens
- Pooling: Max or mean over non-padding tokens
- Rationale extraction: Gradient-based

Training setup:

- Optimizer: Adam with default parameters
- Tokenization: Standard Week-2 preprocessing
  - Number normalization: all digits → <num> token
  - De-identification: remove protected health information markers
  - Lowercasing and punctuation handling

Calibration setup:

- Data split: 50% calibration, 50% test (6 notes each)
- Temperature search: Grid over [0.1, 5.0] with 50 points
- Optimization metric: Brier score minimization

Evaluation metrics:

- Classification: Accuracy, precision, recall, F1-score
- Calibration: ECE (10 bins), NLL, Brier (pre/post calibration)
- Consistency: Label agreement rate, confidence gap
- Robustness: Shuffle deltas, section ablation drops
- Rationales: Top-k overlap, weight correlation

4.3 Results and Analysis

4.3.1 Paraphrase Consistency

BiLSTM performance:

- Label agreement: 90.6%
- Average confidence gap: 0.089 (8.9 percentage points)

- Rationale overlap: ~67%

Transformer performance:

- Label agreement: 74.2%
- Average confidence gap: 0.132 (13.2 percentage points)
- Rationale overlap: ~45%

Clinical interpretation:

BiLSTM exhibits superior stability across paraphrases. Example:

```
Original: "Patient denies chest pain"
  BiLSTM prediction: 23% MI risk
  Rationales: ["denies" (0.34), "chest" (0.28), "pain" (0.21)]

Paraphrase: "Individual reports no chest discomfort"
  BiLSTM prediction: 19% MI risk (Δ = 4%, consistent)
  Rationales: ["reports" (0.31), "no" (0.27), "chest" (0.24)]
```

The 4 percentage point shift is clinically acceptable. Rationales maintain focus on chest-related terms (67% overlap).

Transformer shows greater sensitivity:

```
Original: "Patient denies chest pain"
  Transformer prediction: 31% MI risk
  Rationales: ["denies" (0.38), "chest" (0.31), "pain" (0.22)]

Paraphrase: "Individual reports no chest discomfort"
  Transformer prediction: 45% MI risk (Δ = 14%, inconsistent!)
  Rationales: ["individual" (0.42), "chest" (0.28), "discomfort" (0.19)]
```

The 14 percentage point jump for identical clinical content is concerning for deployment. The rationale shift from "denies" to "individual" (a generic noun) suggests unstable attention patterns.

Clinical implications: BiLSTM's sequential processing preserves local context better, maintaining stability across synonym substitutions. Transformer's global attention is disrupted by lexical changes, causing prediction instability that could lead to documentation-dependent diagnoses.

4.3.2 Calibration Improvements

BiLSTM calibration results:

- ECE: 0.625 → 0.518 (−17% improvement)
- NLL: 2.405 → 0.977 (−59% improvement)
- Brier: 0.588 → 0.368 (−37% improvement)
- Optimal T_pred: 3.200
- Optimal T_rat: 5.000

Transformer calibration results:

- ECE: 0.205 → 0.174 (−15% improvement)
- NLL: 0.707 → 0.695 (−2% improvement)
- Brier: 0.257 → 0.251 (−2% improvement)
- Optimal T_pred: 5.000
- Optimal T_rat: 5.000

Analysis of BiLSTM calibration:

The uncalibrated BiLSTM exhibited severe overconfidence (NLL = 2.405). Example cases:

> Case 1: "Acute chest pain with radiation to left arm, patient appears distressed and diaphoretic"
> Actual outcome: Positive MI (y = 1)
> Raw prediction: 98% MI probability (overconfident)
> Calibrated prediction: 73% MI probability (realistic)
>
> Case 2: "Patient denies chest pain and reports no shortness of breath"
> Actual outcome: No MI (y = 0)
> Raw prediction: 2% MI probability (overconfident low)
> Calibrated prediction: 18% MI probability (acknowledges uncertainty)

The 59% NLL reduction indicates the raw model assigned extreme probabilities (near 0 or 1) too frequently. Temperature scaling with T=3.2 "cools" these extreme predictions toward more realistic middle-range probabilities.

Analysis of Transformer calibration:

The Transformer showed much better initial calibration (ECE = 0.205, NLL = 0.707). Minimal improvements post-calibration suggest the architecture naturally produces well-calibrated probabilities.

Case: "Chest discomfort described as pressure-like, associated with nausea and diaphoresis"
Actual outcome: Positive MI (y = 1)
Raw prediction: 78% MI probability
Calibrated prediction: 75% MI probability (minimal adjustment)

The architecture's self-attention mechanism and global context modeling may contribute to better initial probability estimation.

Rationale calibration:

Both models showed rationale-ECE improvements after power scaling. The high T_rat values (5.0 for both) indicate rationales needed significant compression to match the correctness proxy target.

Key finding: Architecture choice creates a fundamental trade-off between initial calibration quality and paraphrase robustness. BiLSTM requires aggressive post-hoc calibration but maintains consistency; Transformer is well-calibrated initially but sacrifices robustness.

4.3.3 Section Robustness

BiLSTM ablation results:

- History ablation: 0.92 probability point drop (largest)
- Full text removal: 0.97 drop
- Exam ablation: 0.50 drop

Transformer ablation results:

- History ablation: 0.44 probability point drop
- Full text removal: 0.52 drop
- Exam ablation: 0.53 drop

Detailed example (BiLSTM):

Full note (Cardiology):
"History: Chest discomfort described as pressure-like, associated with nausea and diaphoresis.
Physical Exam: Diaphoretic, tachycardic, blood pressure 145/95.
Assessment: Acute coronary syndrome suspected.
Plan: Troponins, ECG, cardiology consult."
Prediction: 78% MI risk

Remove History:
"Physical Exam: Diaphoretic, tachycardic, blood pressure 145/95.
Assessment: Acute coronary syndrome suspected.
Plan: Troponins, ECG, cardiology consult."
Prediction: 12% MI risk (dropped 66 percentage points!)

Remove Exam:
"History: Chest discomfort described as pressure-like, associated with nausea and diaphoresis.
Assessment: Acute coronary syndrome suspected.
Plan: Troponins, ECG, cardiology consult."
Prediction: 42% MI risk (dropped 36 percentage points)

Remove Assessment:
"History: Chest discomfort described as pressure-like, associated with nausea and diaphoresis.
Physical Exam: Diaphoretic, tachycardic, blood pressure 145/95.
Plan: Troponins, ECG, cardiology consult."
Prediction: 71% MI risk (dropped 7 percentage points)

Critical findings:

1. **Over-reliance on History:** The 66-point drop when removing patient-reported symptoms reveals dangerous dependency on subjective information. In emergency settings, patients may be unable to provide complete histories (altered mental status, language barriers, critical condition).
2. **Under-weighting of Exam:** Objective physical exam findings (tachycardia, diaphoresis, hypertension) should be highly diagnostic but contribute less to predictions. This suggests insufficient learning of objective clinical reasoning.
3. **Assessment section paradox:** Despite containing the explicit diagnosis ("acute coronary syndrome suspected"), removing Assessment causes minimal prediction changes. Models may be learning pattern matching rather than clinical reasoning.

Shuffle testing:

Section reordering produced non-zero but moderate delta values, suggesting some order sensitivity but less than ablation effects. This indicates models primarily rely on content presence rather than sequential structure.

4.3.4 Error Analysis by Note Type

Performance by clinical context:

ED (Emergency Department) notes:
  BiLSTM accuracy: 100%
  Transformer accuracy: 100%
  Example: "Acute chest pain with radiation to left arm" (clear presentation)

Cardiology specialist notes:
  BiLSTM accuracy: 100%
  Transformer accuracy: 100%
  Example: "Severe chest pain, positive cardiac risk factors, EKG pending"

Psychiatry notes:
  BiLSTM accuracy: 67%
  Transformer accuracy: 50%
  Example: "Chest tightness with anxiety symptoms, patient reports recent stressors"
  Challenge: Distinguishing psychosomatic from cardiac presentations

MSK (Musculoskeletal) notes:
  BiLSTM accuracy: 100%
  Transformer accuracy: 100%
  Example: "Chest wall tenderness on palpation, musculoskeletal origin suspected"

Pulmonary notes:
  BiLSTM accuracy: 100%
  Transformer accuracy: 100%
  Example: "Sharp stabbing chest pain, pleuritic in nature, worsens with deep inspiration"

GI (Gastroenterology) notes:
  BiLSTM accuracy: 100%
  Transformer accuracy: 100%
  Example: "Burning chest sensation after meals, likely gastroesophageal reflux disease"

Error pattern analysis:

Psychiatric notes pose the greatest challenge. Consider this case:

Note: "Chest tightness with anxiety symptoms, patient reports recent stressors"
True label: Negative (no cardiac event)
BiLSTM prediction: Positive (67% confidence) - WRONG
Transformer prediction: Positive (85% confidence) - WRONG

Key challenge: "Chest tightness" triggers cardiac risk patterns, but context ("anxiety symptoms", "recent stressors") should modulate interpretation.

Both models struggle with context-dependent differential diagnosis, suggesting they've learned lexical correlations ("chest" + "tightness" → cardiac) without deeper semantic understanding of psychological vs. physiological presentations.

Clear-cut cases (acute cardiac presentations, musculoskeletal pain, reflux symptoms) achieve perfect accuracy, indicating models perform well on prototypical examples but struggle with ambiguous presentations requiring nuanced clinical reasoning.

5. Discussion and Analysis

5.1 Architecture Trade-offs

Our results reveal fundamental architecture-dependent trade-offs:

BiLSTM strengths:

- Superior paraphrase consistency (90.6% vs 74.2%)
- Stable rationale extraction across input variations
- Sequential processing preserves local context

BiLSTM weaknesses:

- Severe initial miscalibration (ECE = 0.625, NLL = 2.405)
- Requires aggressive temperature scaling (T = 3.2)
- Over-reliance on sequential patterns may cause overconfidence

Transformer strengths:

- Excellent initial calibration (ECE = 0.205, NLL = 0.707)
- Requires minimal post-hoc adjustment
- Global attention captures long-range dependencies

Transformer weaknesses:

- Poor paraphrase robustness (26% disagreement rate)
- Large confidence swings (13.2 point average gap)
- Unstable rationale focus across lexical variations

5.2 Clinical Deployment Implications

Reliability requirements for clinical systems:

1. **Consistency imperative:** Documentation style shouldn't affect diagnosis. BiLSTM's 90.6% consistency better meets this requirement, though 9.4% disagreement rate remains concerning for high-stakes decisions.
2. **Calibration necessity:** Clinicians need accurate confidence estimates to guide testing and treatment decisions. Transformer's native calibration advantage is significant, but BiLSTM achieves similar post-calibration performance.
3. **Robustness to incomplete information:** Both models' heavy reliance on History sections (0.92 drop for BiLSTM) creates vulnerability in emergency scenarios with incomplete documentation.
4. **Explainability stability:** Rationale consistency across paraphrases is essential for clinical trust. BiLSTM's 67% rationale overlap exceeds Transformer's 45%, supporting more reliable explanation-based auditing.

Recommended deployment strategy:

Given these trade-offs, clinical deployment should:

- Use BiLSTM with mandatory post-hoc calibration for consistency-critical applications
- Consider ensemble approaches combining both architectures
- Implement paraphrase-consistency checks as quality control
- Flag cases with high History-dependency for manual review
- Develop section-aware architectures addressing objective finding under-weighting

5.3 Limitations

Dataset scale: Our 12-note evaluation set provides controlled testing but may not capture full clinical documentation diversity. Larger-scale validation on real EHR data is essential.

Paraphrase coverage: Domain-aware generation covers synonym, abbreviation, and negation variations but doesn't exhaust possible linguistic transformations. More sophisticated paraphrasing (syntactic restructuring, semantic preservation with lexical diversity) warrants exploration.

Rationale evaluation: Token-level overlap measures are simplistic proxies for semantic faithfulness. More sophisticated metrics (e.g., embedding-based semantic similarity, clinical concept alignment) would better assess rationale quality.

Binary classification: Real clinical decision-making involves multi-class risk stratification and continuous risk scores. Extension to more complex prediction tasks is necessary.

Synthetic note types: While note type annotations enable subgroup analysis, our synthetic labels may not reflect true clinical documentation categories.

5.4 Future Directions

Consistency-aware training: Augment training with paraphrase pairs and consistency regularization terms: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathbb{E}_{x,\tilde{x}}[D(f(x), f(\tilde{x}))]$where D measures prediction divergence between original and paraphrased inputs.

Section-aware architectures: Develop models with explicit section encoders:

- Separate embedding spaces for History, Exam, Assessment, Plan
- Learned section importance weights
- Cross-section attention mechanisms

Semantic rationale evaluation: Replace token overlap with semantic similarity:

- Embed rationale tokens using clinical concept embeddings
- Measure semantic coherence across paraphrases
- Align rationales with medical ontology concepts (SNOMED-CT, UMLS)

Larger-scale evaluation: Validate on MIMIC-III discharge summaries (thousands of notes) with real outcome labels.

Pretrained clinical LMs: Evaluate ClinicalBERT, BioBERT, and other domain-adapted transformers for inherent robustness and calibration properties.

Multi-task learning: Joint training on related clinical tasks (ICD coding, entity extraction, outcome prediction) may improve robustness through shared representations.

6. Conclusion

We present a comprehensive reliability evaluation framework for clinical text classifiers, addressing paraphrase consistency, section-order robustness, and dual confidence calibration. Our analysis reveals critical trade-offs: BiLSTM architectures provide superior paraphrase stability (90.6% agreement) but require aggressive calibration (59% NLL reduction); Transformer models exhibit better initial calibration but suffer from paraphrase brittleness (74.2% agreement, 13.2 point confidence gaps).

Both architectures show concerning over-reliance on subjective History sections (≈0.92 probability drop on ablation), suggesting insufficient weighting of objective clinical findings. This creates vulnerability in

incomplete documentation scenarios common in emergency medicine.

Our dual calibration methodology successfully improves both prediction confidence (ECE reductions of 15-26%) and rationale confidence without affecting classification decisions. Temperature scaling with T=3.2 for BiLSTM and T=5.0 for Transformer corrects overconfidence patterns, enabling more reliable probability estimates for clinical decision support.

The artifacts released with this work—reliability diagrams, section ablation reports, paraphrase consistency analyses, and calibrated embeddings—support reproducibility and downstream integration into multimodal clinical systems.

Key contributions:

1. Domain-aware paraphrase generation system for systematic robustness testing
2. Section-order stress framework revealing structure dependencies
3. Dual calibration approach for predictions and rationales
4. Comprehensive evaluation methodology for clinical text reliability

Critical findings:

1. Architecture choice determines calibration-consistency trade-off
2. Sequential models (BiLSTM) favor consistency over calibration
3. Attention models (Transformer) favor calibration over consistency
4. Both architectures under-weight objective clinical findings
5. Post-hoc calibration is essential for safe clinical deployment

Future work should focus on consistency-aware training objectives, section-aware architectures, and large-scale validation on real clinical data. Integration of medical knowledge graphs and clinical concept embeddings may improve both robustness and interpretability.

The reliability challenges identified in this work—paraphrase brittleness, structural dependencies, miscalibration—represent critical gaps between research metrics and clinical deployment requirements. Addressing these challenges is essential for developing trustworthy AI systems that can safely augment clinical decision-making.

References

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321-1330.

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *AAAI Conference on Artificial Intelligence*, 2901-2907.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 1-9.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 4171-4186.

Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *Clinical NLP Workshop*, 72-78.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *NAACL-HLT*, 3543-3556.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *EMNLP-IJCNLP*, 11-20.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *ACM SIGKDD*, 1135-1144.

Acknowledgments

This work builds on Week 2 encoder development and Week 3 rationale-aware architectures. We thank Prof. Mousavi for guidance and feedback throughout this project.