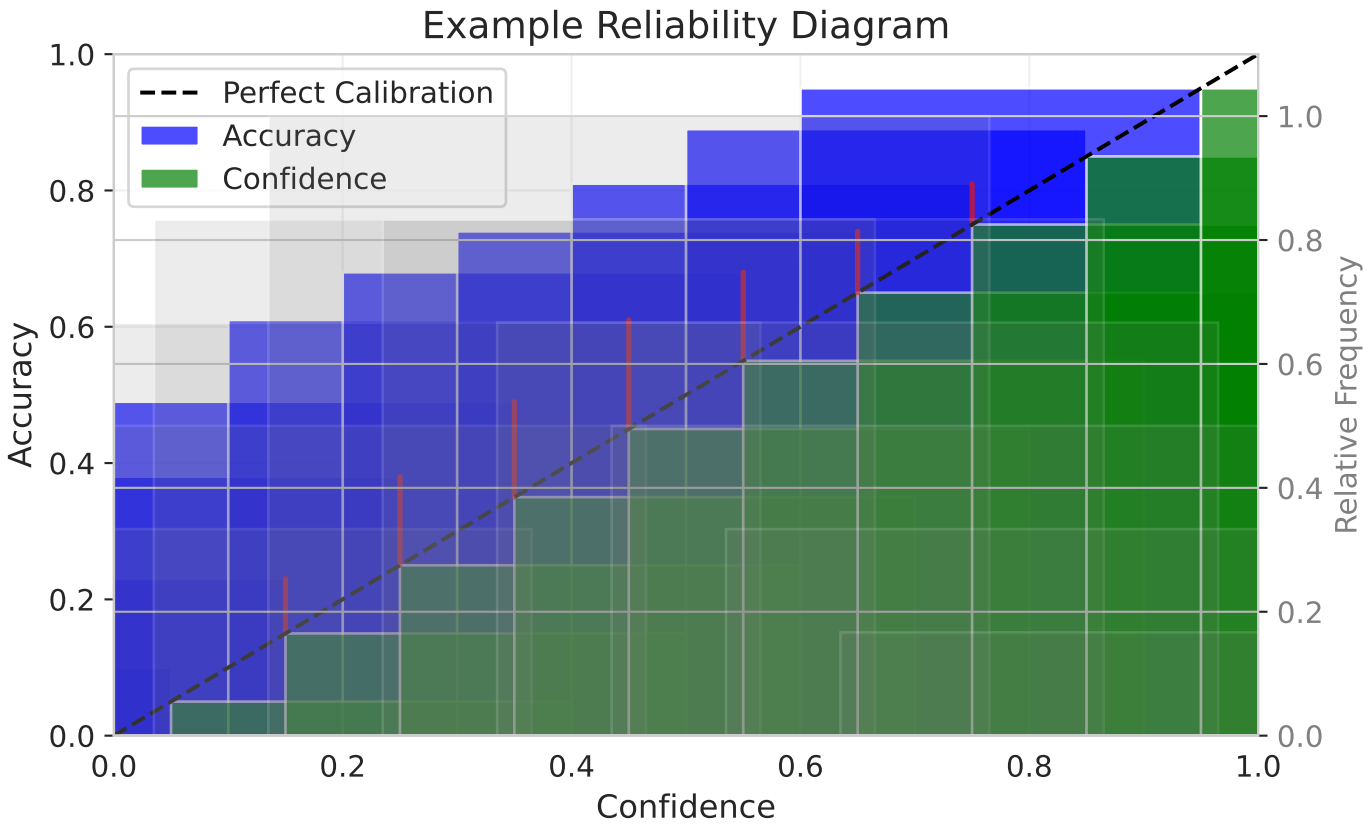
Understanding Uncertainty Metrics

This report explains the uncertainty metrics and visualizations
for 5 images analyzed with VAE-UNet.

Mean Expected Calibration Error (ECE): 0.0054
Mean Dice Score: 0.6566
Mean Brier Score: 0.0059
Mean Sparsification Error: 0.9507

Understanding Reliability Diagrams



RELIABILITY DIAGRAM EXPLANATION:

A reliability diagram shows how well a model's predicted probabilities match actual outcomes.

- Blue bars: The actual frequency of positive pixels in each confidence bin
- Green bars: The mean predicted probability (confidence) for each bin
- Gray histogram: Distribution of predictions across confidence levels
- Red lines: Highlight gaps between confidence and actual frequency
- Diagonal line: Perfect calibration (confidence = actual frequency)

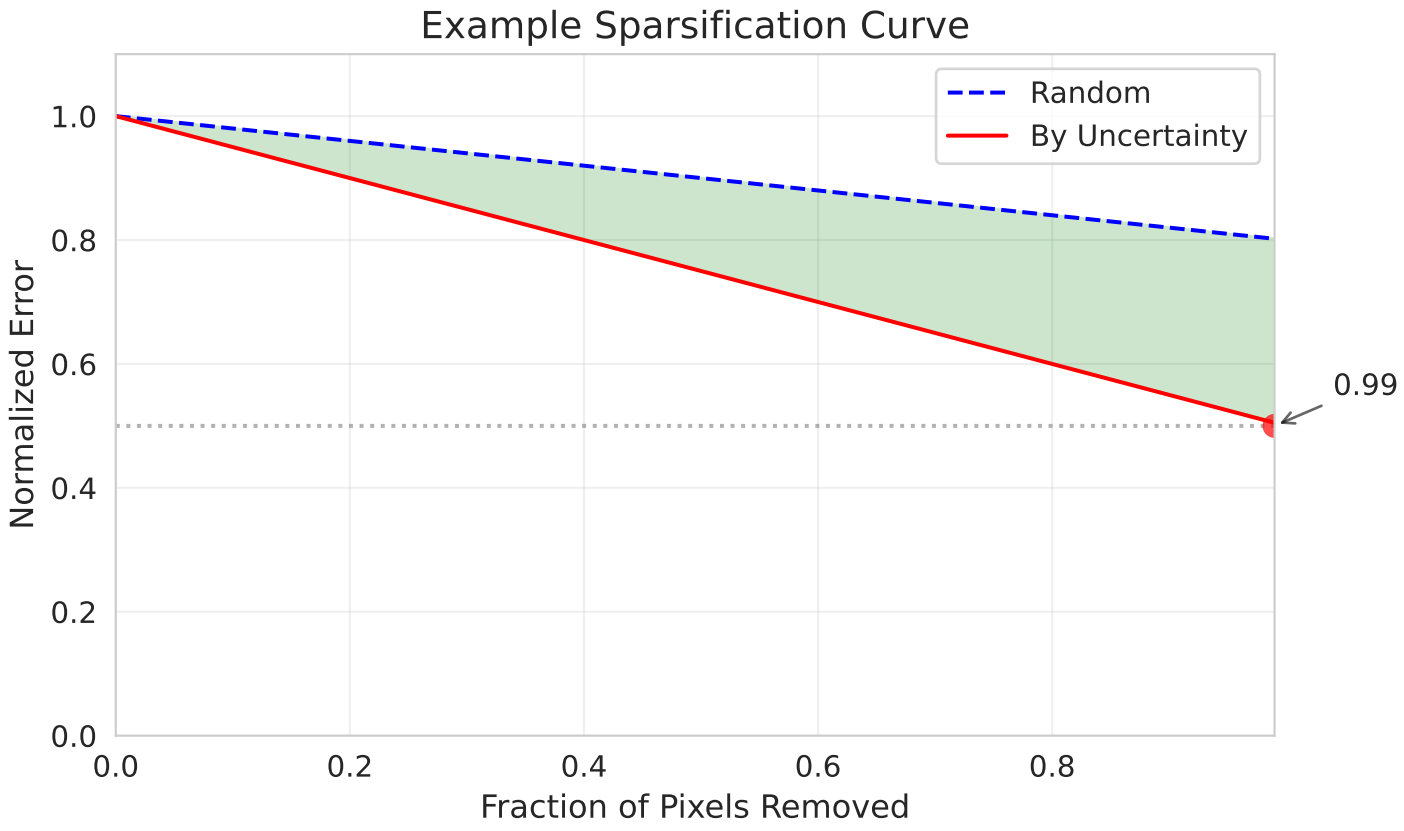
INTERPRETATION:

- When blue bars are higher than green bars: Model is underconfident
- When green bars are higher than blue bars: Model is overconfident
- Expected Calibration Error (ECE): Weighted average of gaps between bars
- Lower ECE values (closer to 0) indicate better calibration

ECE Values Interpretation:
- < 0.01: Excellent calibration
- 0.01-0.05: Good calibration
- 0.05-0.15: Fair calibration
- > 0.15: Poor calibration

WHY IT MATTERS:
Good calibration means the confidence values from your model are reliable. For medical segmentation, this helps clinicians know when to trust the model's predictions.

Understanding Sparsification Curves



SPARSIFICATION CURVE EXPLANATION:

A sparsification curve shows whether your model's uncertainty estimates correlate with actual errors.

• Blue dashed line: Error when removing pixels randomly
• Red solid line: Error when removing pixels with highest uncertainty first
• Green/Red fill: Area between curves (Sparsification Error)
• Red dot: Fraction of pixels that must be removed to halve the error

INTERPRETATION:

• If red line is below blue line (green area): Good uncertainty estimates!
This means removing high-uncertainty pixels reduces error faster than random removal.

• If red line is above blue line (red area): Poor uncertainty estimates.
Your model's uncertainty doesn't correlate well with actual errors.

• Sparsification Error (SE): Area between the curves
- Positive SE: Good uncertainty estimates
- Negative SE: Poor uncertainty estimates
- Larger positive values indicate better uncertainty quality

WHY IT MATTERS:
Good uncertainty estimates help identify which predictions might be wrong and where the model needs human verification in clinical applications.

Understanding the Visualization Plots

CORRELATION MATRIX EXPLANATION:

The correlation matrix shows how different metrics relate to each other:
• Values close to 1: Strong positive correlation (one increases, the other increases)
• Values close to -1: Strong negative correlation (one increases, the other decreases)
• Values close to 0: Little or no correlation

Key relationships to look for:
• Dice Score vs. Uncertainty Metrics: Does better performance correlate with better calibration?
• ECE vs. Sparsification Error: Do different uncertainty metrics agree with each other?

CALIBRATION ANALYSIS PLOT EXPLANATION:

This plot helps understand the pattern of calibration errors:
• X-axis: Maximum Calibration Error (MCE) - the largest calibration error in any bin
• Y-axis: Mean Absolute Calibration Error (MACE) - the average calibration error
• Color: Expected Calibration Error (ECE) - weighted average of calibration errors
• Size: Dice Score - larger points indicate better segmentation performance

Interpretation by location:
• Points near the diagonal: Errors are consistent across all confidence levels
• Points below diagonal: Errors concentrated in specific confidence bins
• Bottom-left corner: Best calibration overall (low errors)
• Larger points in bottom-left: Ideal models (good performance, good calibration)

PAIRPLOT EXPLANATION:

The pairplot shows the relationships between all pairs of metrics:
• Diagonal: Distribution of each individual metric
• Off-diagonal: Scatter plots showing relationship between pairs of metrics

TEMPERATURE SCALING:

If your model has an ECE of 0.005, that's excellent calibration!
With temperature=2.0 giving better results, this suggests your model was slightly overconfident at the default temperature (T=1.0). Higher temperatures 'soften' predictions, making very confident predictions less extreme.