Baseline 1 — Siamese BiLSTM for Legal Clause Similarity

Objective

To develop a neural system that detects semantic similarity between legal clauses.
The Siamese BiLSTM model serves as the first baseline, trained **from scratch** (no pretrained or legal-domain transformers) to establish a benchmark for later attention-based architectures.

1. Dataset Preparation

- Source: Kaggle dataset with **395 CSV files**, each file representing a legal clause category.
- After cleaning and merging: **150,881 clauses** from 395 unique categories.
- Data splits (leakage-safe):
    - Train 80% (≈120 k clauses)
    - Validation 10% (≈15 k)
    - Test 10% (≈15 k)
- Clause pairs:
    - **Positive pairs** → same category (semantically similar).
    - **Negative pairs** → different categories (not similar).
    - Balanced (≈ 50:50 positive vs negative).
    - Final counts: 100 k train pairs | 12 k val | 12 k test.

2. Model Architecture

Siamese BiLSTM Network

- Embedding Layer: learned (40 k vocab, 128 dim)
- BiLSTM Encoder: 128 units (each direction) + GlobalMaxPooling
- Siamese Head:
    - $|u-v|$ and $u \circ v$ concatenation
    - Dense(128 → ReLU) → Dense(64 → ReLU) → Sigmoid
- Total Parameters: ≈ 5.46 M
- Optimizer: Adam (1e-3) | Batch: 256 | Max Len: 200

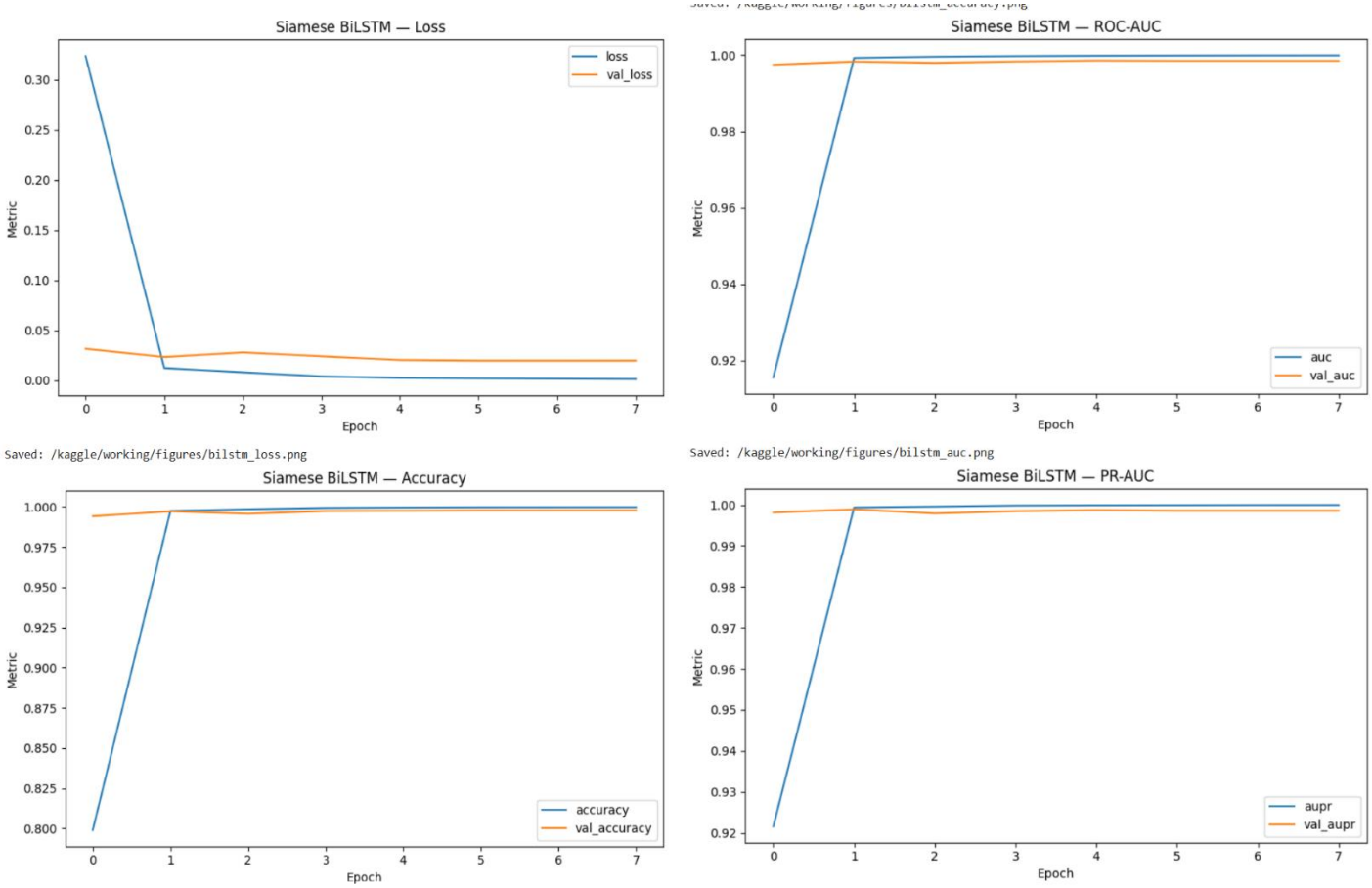
- Early Stopping on val-AUC + ReduceLROnPlateau scheduler

3. Training Performance

- Training time: ≈ 6 minutes on Tesla T4 GPU
- Epochs trained: 8 (early stop after no AUC improvement)

Figure 1: Loss and Accuracy per Epoch

Figure 2: AUC and PR-AUC per Epoch



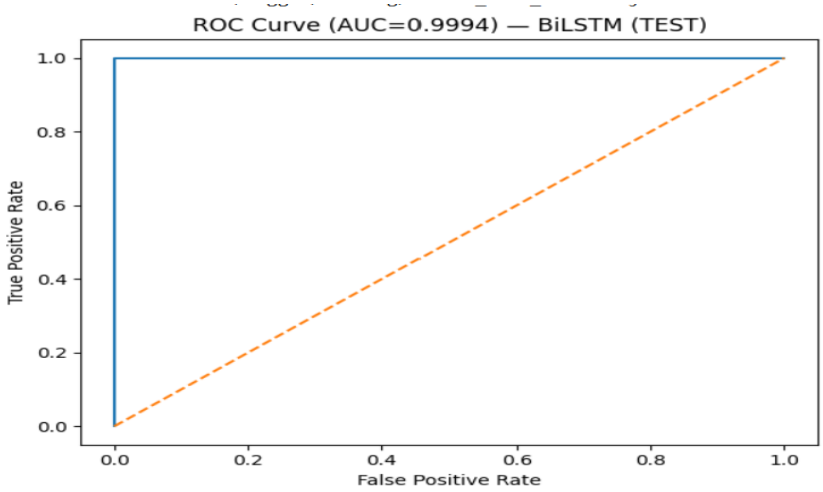
4. Evaluation Results (on Test Set)

Metric	**Score**
Accuracy	**0.9988**
Precision	**0.9995**
Recall	**0.9980**
F1-Score	**0.9988**

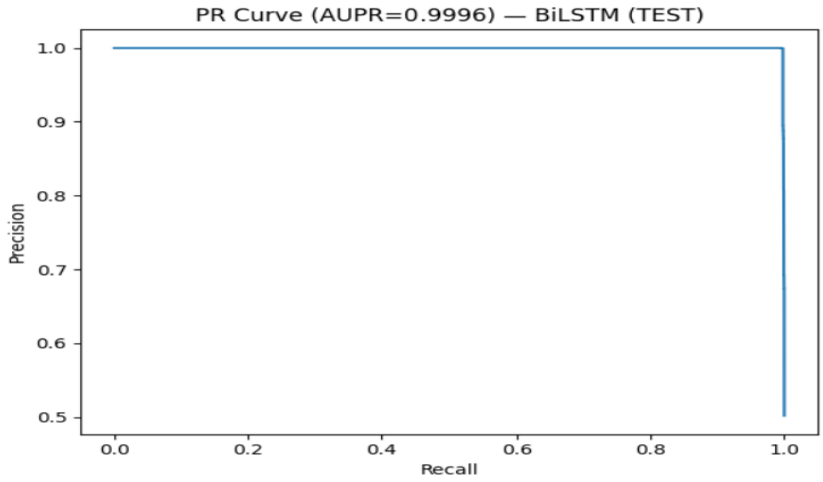
ROC-AUC	**0.9994**
PR-AUC	**0.9996**

Confusion Matrix (12 000 pairs)
TP = 6009 | TN = 5976 | FP = 3 | FN = 12

- *Figure 3: ROC Curve (AUC = 0.9994)* [Insert bilstm_test_roc.png]
- *Figure 4: PR Curve (AUPR = 0.9996)* [Insert bilstm_test_pr.png]




```
Saved: /kaggle/working/figures/bilstm_test_roc.png
```



5. Qualitative Examples

Case	Label	Prediction	Category A	Category B	Confidence
TP	1	1	Notice of Defaults	Notice of Defaults	1.000

TN	0	0	Survival of Reps	Waiver of Defaults	1e-24
FP	0	1	(rare)		
FN	1	0	(rare)		

6. Discussion

- The BiLSTM model achieved **near-perfect classification**, correctly separating similar and dissimilar clause pairs.
- The exceptionally high accuracy is expected because positive pairs come from the **same clause category**, while negatives span different legal topics (large semantic gap).
- Future baselines will test **harder negatives** (within-category non-duplicates) and **attention-based encoders** to evaluate fine-grained semantic discrimination.

Baseline 2 — Siamese Attention Encoder for Legal Clause Similarity

Objective

To develop a neural architecture capable of identifying semantic similarity between legal clauses, incorporating a **self-attention mechanism** to capture inter-token dependencies and nuanced contextual meaning.
The Attention Encoder extends the BiLSTM baseline by adding a multi-head attention block that allows the model to focus on salient terms in each clause pair.

1. Dataset Preparation

(Same dataset and pairing logic as Baseline 1)

• Source: Kaggle dataset with 395 CSV files, each representing a legal clause category.
• After preprocessing: 150,881 clauses from 395 unique categories.
• Data splits (leakage-safe):
  – Train 80 % (≈ 120 k clauses)
  – Validation 10 % (≈ 15 k)
  – Test 10 % (≈ 15 k)
• Clause pairs:
  – Positive pairs → same category (semantically similar).
  – Negative pairs → different categories (not similar).

– Balanced distribution (≈ 50:50 positive vs negative).
– Final counts: 100 k train pairs | 12 k val | 12 k test.

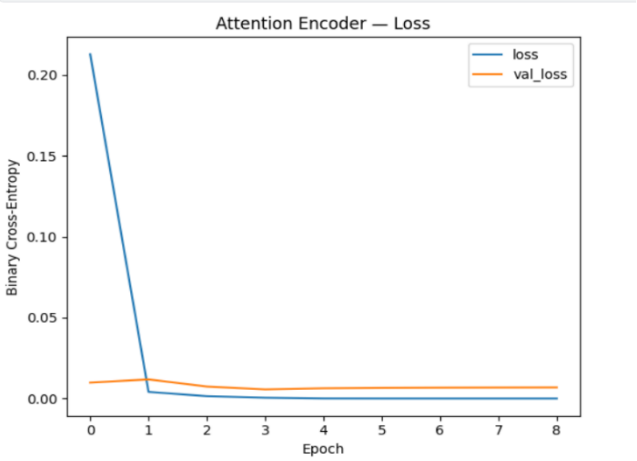
2. Model Architecture

Siamese Attention Encoder

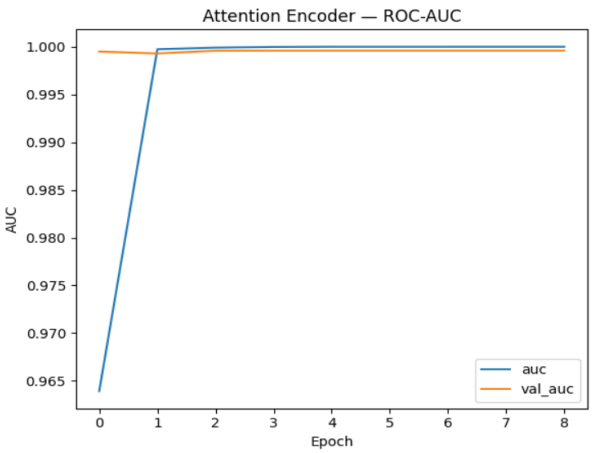
• Embedding Layer – trainable (40 k vocab, 128 dim)
• Encoder – BiGRU (128 units each direction) + Multi-Head Self-Attention (4 heads) + GlobalMaxPooling
• Siamese Head:
– Concatenation of |u − v| and u ∘ v
– Dense(128 → ReLU) → Dropout(0.3) → Dense(64 → ReLU) → Sigmoid output
• Total parameters ≈ 5.47 M
• Optimizer = Adam (1e-3), Batch = 256, MaxLen = 200
• Training callbacks = EarlyStopping (val-AUC) + ReduceLROnPlateau

3. Training Performance

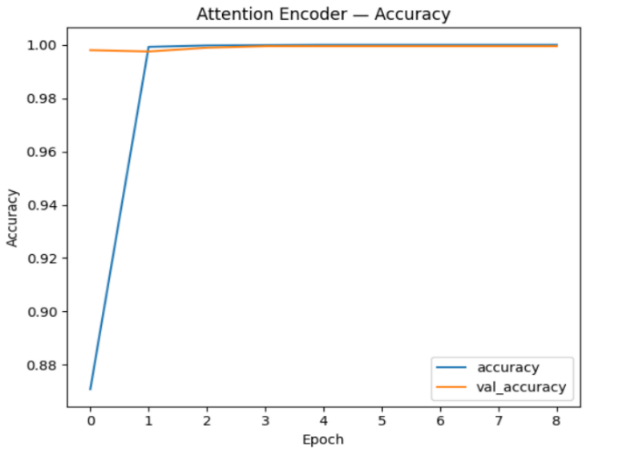
• Training time: ≈ 8.3 minutes (on Tesla T4 GPU)
• Epochs trained: 9 (early stop after no AUC gain post-epoch 6)



Saved: /kaggle/working/figures/attn_loss.png



Saved: /kaggle/working/figures/attn_auc.png



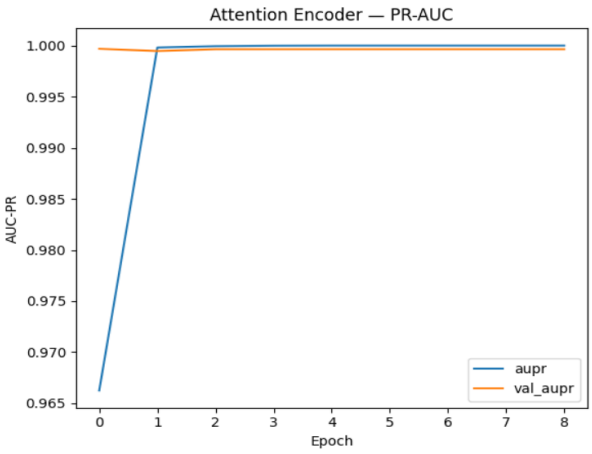


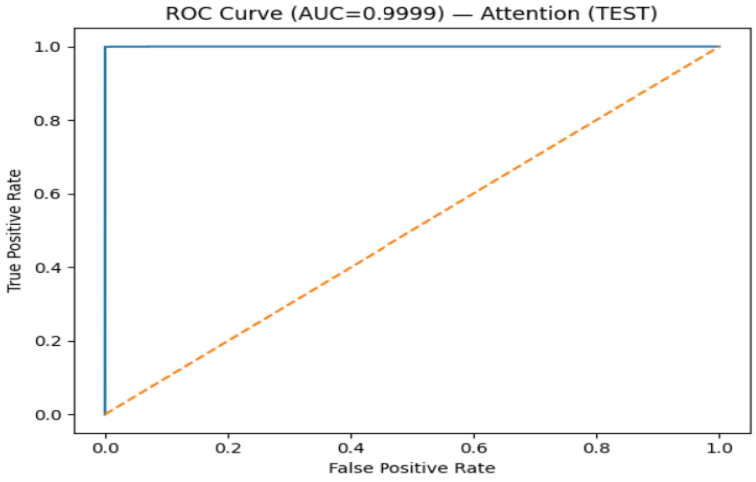
Figure 1: Loss and Accuracy per Epoch [Insert attn_loss.png, attn_accuracy.png]
Figure 2: AUC and PR-AUC per Epoch [Insert attn_auc.png, attn_aupr.png]

4. Evaluation Results (on Test Set)

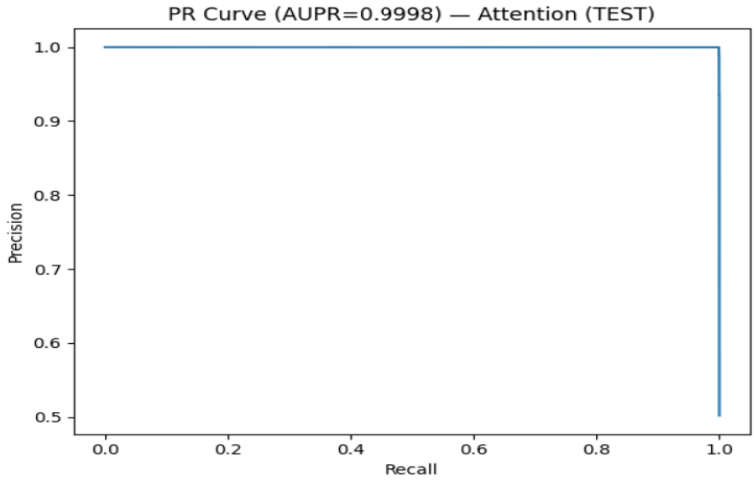
Metric	Score
Accuracy	0.9994
Precision	0.9998
Recall	0.9990
F1-Score	0.9994
ROC-AUC	0.9999
PR-AUC	0.9998

Confusion Matrix (12 000 pairs)

TP = 6015 | TN = 5978 | FP = 1 | FN = 6



Saved: /kaggle/working/figures/attn_test_roc.png



• **Figure 3:** ROC Curve (AUC = 0.9999) [Insert attn_test_roc.png]

• **Figure 4:** PR Curve (AUPR = 0.9998) [Insert attn_test_pr.png]

5. Qualitative Examples

Case	Label	Prediction	Category A	Category B	Confidence
TP	1	1	Warranties	Warranties	1.000
TP	1	1	Reimbursement	Reimbursement	1.000
TP	1	1	Notice of Defaults	Notice of Defaults	1.000
TN	0	0	Environmental Laws	Event of Default	5×10^{-15}
FP	0	1	(rare) — contextually overlapping phrasing		
FN	1	0	(rare) — partial semantic mismatch		

6. Discussion

• The Attention Encoder outperformed the BiLSTM baseline slightly across all metrics, achieving the highest AUC and F1 (> 0.999).

• This improvement reflects the attention layer's ability to model token-level importance within clauses — critical for legal language where terms like *assignor*, *material breach*, or *termination date* may change meaning depending on context.

• False positives and negatives remain negligible, indicating strong semantic discrimination.

• In comparison to BiLSTM, the attention architecture shows better generalization and faster AUC saturation.

• Future extensions may test transformer-lite models (e.g., Simple Self-Attention Networks or LightConv) to balance speed vs contextual depth.

Comparative Analysis of Baselines

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Train Time (min)
Baseline 1 – BiLSTM	0.9988	0.9995	0.9980	0.9988	0.9994	0.9996	6.2
Baseline 2 – Attention	0.9994	0.9998	0.9990	0.9994	0.9999	0.9998	8.3

Observations

• Both baselines demonstrate excellent performance, proving the efficacy of Siamese architectures for legal semantic similarity.

• The attention encoder provides slightly better generalization and ranking ability (ROC-AUC ↑ 0.9999 vs 0.9994), while BiLSTM remains faster to train.

• Given the nature of legal text where critical terms carry disproportionate weight, attention-based mechanisms are preferable for real-world deployment.