NLI with BiLSTM Cross-Attention, ESIM-Style BiGRU, and a Lightweight Transformer

Shuo Ma (23914891) CITS4012 Group 35 23914891@student.uwa.edu.au

Kunhong Zou (24257885) CITS4012 Group 35 24257885@student.uwa.edu.au

Mohaimen Rashid (24117314) CITS4012 Group 35 24117314@student.uwa.edu.au

Abstract

We study Natural Language Inference (NLI) on the science-domain dataset released for CITS4012 by training three from-scratch architectures that emphasise interpretability, alignment reasoning, and efficiency. Model A couples static 200-dimensional word2vec embeddings with bilinear cross-attention over BiLSTM encoders; Model B implements an ESIM-style BiGRU pipeline with inference composition; and Model C is a lightweight Transformer cross-encoder with learned token and segment embeddings. All models are trained solely on the provided train/validation/test JSON splits using a consistent preprocessing pipeline. Model B delivers the strongest held-out performance (73.0% validation / 72.7% test accuracy; macro-F1 0.708), while Model C provides the best test-only accuracy among the lightweight setups (71.4% test, macro-F1 0.692) and Model A offers interpretable attention maps despite slightly lower scores. Ablations show bilinear cross-attention maintains a small but consistent validation edge over dot-product and attention-free variants. We analyse quantitative and qualitative behaviour and discuss limitations and future refinements.

1 Introduction

Natural Language Inference (NLI) requires determining whether a hypothesis is entailed by, or neutral with respect to, a premise. The science-domain dataset released for CITS4012 contains 23,088 training pairs and stresses models with technical vocabulary and long sentences, making it a useful test-bed for attention mechanisms and lightweight Transformers under compute constraints.

We study three NLI architectures that target complementary trade-offs between interpretability, alignment reasoning, and efficiency. Our contributions are:

- a unified preprocessing and training pipeline that trains all parameters from scratch on the provided JSON splits while ensuring reproducibility through deterministic seeding;
- an empirical comparison of a bilinear cross-attention BiLSTM, an ESIM-style BiGRU, and a shallow Transformer cross-encoder, evaluated with accuracy, macro-F1, and confusion matrices;
- quantitative ablations and qualitative alignment analyses that expose how attention design choices affect the science-domain NLI task.

We release consolidated notebooks and logging artefacts to document each experiment.

2 Methods

2.1 Model A: BiLSTM Cross-Attention

Model A operates on a shared vocabulary built from tokens that occur at least twice in the training split; sequences are truncated to 64 tokens. We train 200-dimensional skip-gram word2vec embeddings (window=5, epochs=10) on the cleaned premises and hypotheses, initialise the embedding layer with this matrix, and fine-tune during supervised learning. Premise and hypothesis are encoded with separate bidirectional LSTMs with 128 hidden units per direction. A bilinear cross-attention module produces aligned representations that are combined with element-wise difference and product features. We apply global max- and average-pooling, concatenate premise and hypothesis summaries, and feed them to a 256-unit ReLU layer with 0.3 dropout before the final softmax classifier. The model has 5.1M trainable parameters and is optimised with Adam (learning rate 2×10^{-3}, batch size 64) for five epochs.

2.2 Model B: ESIM-Style BiGRU with Inference Composition

Model B follows the ESIM template and reuses the cleaned vocabulary with dynamic padding per batch. We initialise 200-dimensional embeddings randomly and train them jointly with the network. A shared bidirectional GRU with 128 hidden units encodes the premise and hypothesis; soft alignment attention yields context-aware pairs that are enhanced by concatenation, element-wise difference, and product. A second bidirectional GRU composes local inference features, after which we apply average and max pooling, concatenate the pooled vectors, and pass them through a 256-unit multilayer perceptron with ReLU activation and 0.3 dropout. We optimise with AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-5}), gradient clipping at 5.0, batch size 32, and early stopping with patience three. The configuration contains 3.36M trainable parameters.

2.3 Model C: Lightweight Transformer Cross-Encoder

Model C constructs a vocabulary from the training split (minimum frequency two) and inserts special tokens for `[CLS]`, `[SEP]`, and padding. Each example is tokenised as `[CLS]` + premise + `[SEP]` + hypothesis, paired with segment identifiers for the three regions, and truncated to 256 tokens. We learn 256-dimensional token, segment, and position embeddings that are summed and layer-normalised before entering a three-layer Transformer encoder with four attention heads and 512-dimensional feed-forward blocks (dropout 0.1, GELU activations). The pooled `[CLS]` representation feeds a classifier consisting of a linear layer, GELU, dropout, and a final linear layer over the two labels. Training uses AdamW (learning rate 2×10^{-4}, weight decay 1×10^{-4}), a 10% warm-up, gradient clipping at 1.0, batch size 32, and eight epochs with the best checkpoint selected by validation accuracy.

3 Experiment Setup

3.1 Dataset

We normalise each field with Unicode NFKC, collapse whitespace, remove unsupported symbols, and lowercase tokens before tokenisation. The released splits contain 23,088 training, 1,304 validation, and 2,126 test pairs. The training set is moderately imbalanced (63.3% neutral vs. 36.7% entails), whereas the validation split is balanced and the test split contains 60.4% neutral labels. Premises average 21.1 tokens (95th percentile 38) and hypotheses 13.2 tokens (95th percentile 22), motivating the 64-token truncation used by the recurrent models, while Model C retains up to 256 tokens to accommodate the longest sequences.

Model	Val Acc	Test Acc	Macr
Model A (BiLSTM + Cross-Attn)	0.709	0.694	0.6
Model B (ESIM / BiGRU)	0.730	0.727	0.7
Model C (Transformer)	0.698	0.714	0.6

Table 1: Main results for the three models. Metrics are averaged over the single provided validation and test splits.

3.2 Training Details

Experiments were executed on a Google Colab runtime with a single NVIDIA T4 GPU, TensorFlow 2.19, and PyTorch 2.8 (CUDA 12.6). Deterministic seeds (42 for TensorFlow, 2025 for PyTorch) are set across NumPy, Python, and the frameworks. Model A is trained for five epochs with early stopping on validation accuracy and maintains the word2vec embeddings trainable; Word2Vec is trained for ten epochs with window size five and minimum count two. Model B trains for up to ten epochs with AdamW, gradient clipping, and patience-three early stopping; we log the full training history and reload the best checkpoint before evaluation. Model C trains for eight epochs with AdamW and a linear warm-up scheduler, persisting the per-epoch metrics to `modelC_history.json`. Inference batches for Models B and C use dynamic padding to minimise unnecessary computation.

3.3 Evaluation Metrics

We report accuracy and macro-F1 using `sklearn.metrics` for every split with labels. Confusion matrices on validation and test data highlight class-specific behaviour, and per-label accuracies are recorded for Model C. For qualitative analysis we extract attention weights and alignment scores from the trained models. Test labels remain unseen until the final evaluation stage to prevent leakage.

4 Results

4.1 Main Results

The ESIM-style Model B achieves the strongest overall performance, topping both validation

Ablation	Val Acc	Test Acc	Test Macro F1
A: Bilinear Attention (baseline)	0.709	0.694	0.661
A: Dot-Product Attention	0.707	0.708	0.682
A: No Attention	0.704	0.686	0.677
A: Bilinear (LSTM hidden=64)	0.705	0.697	0.677

Table 2: Model A ablations highlighting attention design and encoder capacity.

(73.0%) and test (72.7%) accuracy while delivering the highest macro-F1 (0.708). The lightweight Transformer (Model C) remains competitive with 71.4% test accuracy and strong neutral recall, offering a favourable efficiency–performance trade-off. Model A trails on macro-F1 because its bilinear attention slightly overfits the majority neutral class despite comparable validation accuracy and remains valuable for interpretability through attention visualisations.

4.2 Ablations

Removing attention causes Model A to lose roughly half a validation point relative to the bilinear baseline, confirming that cross-attention remains helpful for the science-domain vocabulary. Dot-product attention recovers most of the test accuracy (70.8%) and improves macro-F1, but it still trails the bilinear variant on validation score. Shrinking the LSTM hidden size to 64 dimensions slightly reduces both validation accuracy and macro-F1, indicating that the richer interaction features benefit from higher-capacity encoders.

5 Qualitative Results

Qualitative inspection of Model A attention heatmaps shows the bilinear module focusing on pivot tokens such as "rotates" and quantitative phrases when resolving entailment, while neutral predictions spread mass across unrelated nouns. For Model B we examine the alignment weights of the correctly classified validation example in which the hypothesis "replace another in a molecule happens to atoms during a substitution reaction" is predicted neutral. The ESIM attention links "replace" and "another" to "molecule" and "chemical", highlighting that the model grounds decisions in chemically relevant tokens even when asserting neutrality. Errors for all models often arise from hypotheses that paraphrase premises via scientific synonyms absent in the training data, indicating that lexical coverage remains a key limitation.

6 Conclusion

We compared three from-scratch NLI architectures on the CITS4012 science-domain dataset. The bilinear cross-attention BiLSTM provides interpretable alignments and competitive accuracy, the ESIM-style BiGRU balances performance with transparent alignment features and achieves the highest validation and test metrics, and the lightweight Transformer cross-encoder offers a strong efficiency–performance compromise. Ablations confirm that attention design and encoder capacity materially influence results. Future work includes expanding the Transformer ablations (head count and feed-forward width), exploring class reweighting to mitigate label imbalance, and integrating error-driven data augmentation to improve entailment recall.

Limitations

The dataset is binary (neutral vs. entails) and moderately imbalanced, limiting conclusions about contradiction handling. Normalisation removes domain-specific symbols (e.g., chemical notation), which may discard informative cues. Compute constraints cap the Transformer depth, so we do not assess larger cross-encoders that could further improve validation accuracy. Our Model C ablations currently cover only the baseline configuration; sweeping head counts and feed-forward widths remains as future work. Finally, qualitative analyses rely on single examples, and broader human evaluation would strengthen interpretability claims.

Team Contribution

- Shuo Ma (23914891): Designed and implemented Model A, trained word2vec embeddings, and produced the cross-attention ablations and visualisations.
- Kunhong Zou (24257885): Built the ESIM-style Model B pipeline, conducted dataset diagnostics, and generated the alignment analyses and confusion matrices.
- Mohaimen Rashid (24117314): Developed the Transformer cross-encoder (Model C), executed training runs and logging, and integrated the consolidated notebook and report.