CITS4012: Natural Language Processing
Project 2
Semester 2, 2025

Mahit Gupta *(23690265)*
James Wigfield *(23334375)*
Mitchell Otley *(23475725)*

Abstract

This study evaluates three neural architectures for Natural Language Inference: a bidirectional GRU, a Variational Siamese Autoencoder (VSAE), and a self-attention transformer. The bidirectional GRU achieved the highest test accuracy (70.0%) but lowest F1-score (0.650), indicating majority class bias. The VSAE demonstrated superior class balance with the highest F1-score (0.678) despite lower accuracy (68.1%). The transformer achieved intermediate performance (69.1% accuracy, 0.654 F1). Ablation studies showed self-attention substantially outperformed cross-attention (69.1% vs 62.8%), while moderate reconstruction weight ($\gamma = 0.3$) improved VSAE validation performance but reduced test generalization, revealing trade-offs between latent semantic structure and discriminative accuracy in imbalanced NLI tasks.

1 Introduction

Humans have the inherent ability to draw conclusions from information that may not be explicitly stated, harnessing our gathered knowledge about the world around us to extract further meaning that may not be initially apparent. This ability has many applications to machines, improving capabilities for question answering, sentiment analysis, semantic role labelling, deduction, and more (Storks et al., 2020). Rather than relying on static inference rules to map the relation between pieces of information, Natural Language Inference (NLI) is a subset of deep learning tasks that seek to mimic this trait by tasking models to classify the entailment of a hypothesis h, given the initial premise p. MacCartney (2009) highlights the challenges associated with NLI, including variability in lexical choice, linguistic expression, and semantics, that make logical inference difficult. To be able to accurately classify the entailment, models need to comprehend the underlying knowledge of the input (Li and Yuan, 2025).

This investigation assesses the performance of three NLI models designed to classify premise-hypothesis pairs as *entails* (i.e. h can be inferred from p), or *neutral* (i.e. h cannot be inferred nor contradicted from p). The implemented models are a bidirectional Gated Recurrent Unit (GRU) model, a variational siamese autoencoder (VSAE) model, and a self-attention transformer model. Further ablation studies are conducted to compare the performance of self-attention with an alternative cross-attention implementation on an identical transformer architecture, as well as to investigate the effect of the reconstruction weight, γ, in the VSAE model. The γ ablation assesses how varying the contribution of the reconstruction loss influences latent semantic structure and classification balance, providing insight into the trade-off between discriminative and generative regularisation within the VSAE framework.

2 Methods

2.1 Bidirectional GRU (BiGRU)

The GRU structure is an implementation of a Recurrent Neural Network (RNN), that uses an update gate, based on the current input and hidden state of the network, to retain important information and discard less valuable information from the sequence. In this implementation, the outputs for each token in the final hidden layer are used in the rest of the network, rather than the hidden layer outputs for the final token. This allows for the token-level representation to be passed through the network, rather than just the combined state of the final token.

Figure 2 illustrates the bidirectional GRU model. The embedded premise/hypothesis pairs are passed through individual bidirectional GRU layers. Layer normalisation is used to normalise values across the feature-dimension, without requiring any reshaping of the GRU output. Unlike batch normalisation,

this is independent of the other samples that pass through the network and is not impacted by the padding of the sequences. The premise and hypothesis pairs are then concatenated together before being passed through another bidirectional GRU. This secondary GRU learns the interaction between the outputs of the individual GRUs. After layer normalisation, the output undergoes maximum pooling along the sequence dimension to reduce the dimensionality of the values. Maximum pooling preserves only the strongest features across each token, whilst mean pooling averages the features for each token. Maximum pooling was selected, as it offers stronger invariance to noise than mean pooling, and is quicker to compute. A dense linear network with a ReLU activation layer is used, to learn the features of the pooled GRU output, before the network classifies the values as either neutral or entails. Dropout layers were introduced as regularisation techniques during training, used before and after the dense linear layer, as they were found to improve model performance.

Hyperparameter tuning was used to find the best parameters for the network. After tuning, the individual GRU layers were set to 256 units, whilst the combined GRU was set to 128 units. As each GRU is bidirectional, each subsequent layer depth is multiplied by two, to account for the forward and backward GRUs. Each bidirectional GRU is stacked to five layers each. The linear layer has 64 units, and each dropout layer was set to a dropout rate of 0.4. The model is trained for 10 epochs, with a learning rate of 10^{-5} using an ADAM optimiser.

2.2 Variational Siamese Autoencoder (VSAE)

The Variational Siamese Autoencoder (VSAE) combines the generative strengths of variational autoencoders (VAEs) with the pairwise reasoning structure of Siamese neural networks, tailored for the Natural Language Inference (NLI) task. The model classifies premise–hypothesis pairs (p, h) as either *entails* or *neutral*, depending on whether the hypothesis can be logically inferred from the premise.

Each sentence is embedded using pretrained GloVe-Twitter-100 vectors and encoded with a bidirectional LSTM (BiLSTM) network, which produces contextual hidden states. Masked mean pooling is applied to obtain sentence-level embeddings $\mathbf{h}_1$ and $\mathbf{h}_2$ for the premise and hypothesis respectively. A variational layer then projects each embedding into the parameters of a Gaussian posterior:

$$q(\mathbf{z}_i \mid \mathbf{h}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)), \tag{1}$$

and samples latent representations via the reparameterisation trick:

$$\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

Following the Siamese architecture, the latent representations are combined as

$$\mathbf{v} = [\,|\mathbf{z}_1 - \mathbf{z}_2|;\, \mathbf{z}_1 \odot \mathbf{z}_2\,], \tag{3}$$

where $\odot$ denotes elementwise multiplication, and the concatenated vector $\mathbf{v}$ is passed to a multilayer perceptron (MLP) classifier that outputs the entailment label.

To promote semantically meaningful latent spaces, the model includes a reconstruction decoder that predicts the masked-mean embedding of the hypothesis from its latent vector $\mathbf{z}_2$. The total training objective combines discriminative and generative components:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta\, \mathcal{L}_{\text{KL}} + \gamma\, \mathcal{L}_{\text{rec}}, \tag{4}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy classification loss, $\mathcal{L}_{\text{KL}}$ is the Kullback–Leibler divergence regulariser encouraging latent normality, and $\mathcal{L}_{\text{rec}}$ is a mean-squared reconstruction loss. The hyperparameters β and γ control the regularisation and reconstruction strength, respectively.

A schematic overview of the VSAE architecture is presented in Figure 3, illustrating the dual BiLSTM encoders, variational heads with reparameterization, feature combination, and reconstruction decoder. Similar architectures that integrate VAEs with Siamese networks have demonstrated the ability to capture structured latent representations in multimodal and sequence learning tasks (Arad et al., 2023). This model follows the same principle, balancing generative and discriminative learning for improved semantic generalisation in NLI.

2.3 Transformer

The Transformer architecture leverages self-attention mechanisms to capture long-range dependencies without sequential processing constraints. Unlike RNNs and LSTMs, transformers process entire sequences in parallel, enabling efficient training while modeling complex linguistic relationships. This section describes the main NLITransformer

model architecture employing self-attention with stable pooling. An ablation variant using cross-attention is discussed in Section 2.5.

Figure 4 illustrates the model architecture. The architecture begins with pretrained GloVe embeddings scaled by $\sqrt{d_{\text{model}}}$, followed by sinusoidal positional encodings:

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \\ PE_{(pos,2i+1)} &= \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \end{aligned} \tag{5}$$

where pos denotes token position and i the dimension index. Layer normalization is applied before passing to the transformer encoder.

The encoder consists of stacked multi-head self-attention layers with feed-forward networks and pre-layer normalization for improved training stability. Multi-head attention computes:

$$\text{MH}(Q, K, V) = \text{Concat}(h_1, \ldots, h_n)W^O, \tag{6}$$

where each head computes scaled dot-product attention:

$$h_i = \text{softmax}\left(\frac{QW_i^Q K^T W_i^{K^T}}{\sqrt{d_k}}\right) VW_i^V. \tag{7}$$

The premise and hypothesis are concatenated before encoder processing, with padding masks preventing attention over padding tokens.

A learned attention-weighted pooling mechanism aggregates encoder outputs. For sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$, attention scores are:

$$\alpha_i = \frac{\exp(w^T \mathbf{h}_i)}{\sum_{j=1}^{n} \exp(w^T \mathbf{h}_j)}, \tag{8}$$

where w is learned. The final representation is $\mathbf{s} = \sum_{i=1}^{n} \alpha_i \mathbf{h}_i$, allowing dynamic focus on informative tokens with numerical stability via max-score subtraction.

The model employs 4 attention heads, 2 encoder layers, feed-forward dimension of 512, and dropout of 0.2. Training uses AdamW optimization with learning rate 8×10^{-5}, weight decay 0.01, and gradient clipping (max norm = 0.5).

2.4 Dataset Preprocessing

The textual entailment dataset consists of 226,518 pairs of premise and hypothesis statements. The hypotheses are derived from multi-choice science exams, where the question and correct answer are converted into an assertive statement, whilst the premise is derived from a large text corpus of web sentences. 23,088 samples are used for model training, whilst 1,304 labels are reserved for the validation set, and 2,126 labels are reserved for the test set. Non-alphanumeric characters, such as punctuation, are removed, and sentences are normalised to lower case. Enforcing all words to lowercase guarantees consistency in word representation (Supriyono et al., 2024). Additionally, the pretrained embeddings that are used only support lowercase tokens. Preprocessing is achieved using the spaCy tokenisation library (Honnibal et al., 2020).

After tokenisation, outliers were observed in the training dataset, where some premises had a token length of zero, and some premises had an abnormally large number of tokens. Data samples within the 2.5th and 97.5th percentile, based on premise token length, were removed. The distribution of token length for the training, validation and test set are shown in Figure 1. From the training set, each unique word is assigned an index and added to a list to form the vocabulary. As some words may appear in the validation and test sets that are not in the training set, a unique 'out-of-vocabulary' tag is added to the vocabulary, to act as a placeholder when these words are passed to the embedding layer.

100-dimension GloVe embeddings, pretrained on two billion Twitter messages, are used to create an embedding table that translates word indexes to a numerical representation of the word (Pennington et al., 2014). If a word in the vocabulary is not represented in the embeddings, zeros are used for the token. Using the embedding table, each tokenised sentence in the dataset is encoded. As the sentences are required to be of the same length to be able to be passed through the model, each premise and hypothesis is padded to be the length of the largest premise and hypothesis, respectively.

2.5 Ablation Study 1: Attention Mechanism

To investigate the impact of attention architecture on NLI performance, this ablation study compares the main transformer model with self-attention against a variant employing cross-attention with residual connections.

Main Model: Self-Attention with Stable Pooling - The main NLITransformer model concatenates premise and hypothesis embeddings and processes them jointly through the transformer encoder. A learned attention-weighted pooling layer with numerical stabilization aggregates the joint represen-

tation for classification. This approach allows the model to learn interactions between premise and hypothesis tokens implicitly through self-attention across the concatenated sequence.

The attention pooling employs max-score subtraction before softmax to prevent numerical overflow:

$$\alpha_i = \frac{\exp(s_i - s_{\max})}{\sum_j \exp(s_j - s_{\max})}, \tag{9}$$

where $s_i = w^T \mathbf{h}_i$ are the attention scores.

Ablation Variant: Cross-Attention with Residual Connections The NLITransformerCrossAttention variant employs separate transformer encoders for premise and hypothesis, followed by an explicit cross-attention layer where the hypothesis attends to the premise:

$$\text{CrossAttn}(Q_h, K_p, V_p) = \text{softmax}\left(\frac{Q_h K_p^T}{\sqrt{d_k}}\right) V_p, \tag{10}$$

where Q_h is the query from the hypothesis encoder and K_p, V_p are keys and values from the premise encoder. A residual connection stabilizes training:

$$\mathbf{h}' = \text{LayerNorm}(\mathbf{h} + \text{CrossAttn}(Q_h, K_p, V_p)). \tag{11}$$

Separate attention-weighted pooling is applied to both the premise and cross-attended hypothesis representations, which are then concatenated for classification. This architecture explicitly models the asymmetric relationship in NLI where the hypothesis is evaluated against the premise.

Hyperparameters are independently tuned for optimal performance. The main self-attention model uses a learning rate of 8×10^{-5} and dropout of 0.2, while the cross-attention variant uses a lower learning rate (3×10^{-5}) and higher dropout (0.3) due to its increased capacity. More aggressive gradient clipping (max norm = 0.3 vs 0.5) is applied to the cross-attention variant to maintain stability. Both models train for up to 10 epochs with early stopping based on validation accuracy.

2.6 Ablation Study 2: Effect of Reconstruction Weight γ

This ablation investigates how the reconstruction weight γ affects the Variational Siamese Autoencoder (VSAE). The coefficient γ scales the reconstruction loss $\mathcal{L}_{\text{rec}}$ within the total objective:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{rec}}. \tag{12}$$

While β regularises the latent space via the Kullback–Leibler divergence, γ determines the strength of the model's generative regularisation by encouraging reconstruction from the latent representation. The study evaluates how varying γ influences latent feature quality, class separation, and generalisation in the NLI task.

Four configurations were trained with $\gamma \in \{0.0, 0.1, 0.3, 0.5\}$ under identical conditions. The $\gamma = 0.0$ setting serves as a baseline without reconstruction, while higher values progressively increase decoder influence. Each configuration was evaluated on the validation and test sets using accuracy and macro-F1 metrics.

3 Results

3.1 Performance Comparison

The performance of each implemented model is shown in Table 1. Models are evaluated on test set accuracy and macro-averaged F1-score. The F1-score is particularly important given the class imbalance in the test set, as it equally weights precision and recall for both classes. Figures 9, 10, and 11 in Appendix B show the confusion matrices for each model.

Model	Accuracy (%)	F1
BiGRU	**70.0**	0.650
VSAE	68.1	**0.678**
Transformer	69.1	0.654

Table 1: Model performances on the test dataset.

All models achieved similar performance metrics on the test dataset. The BiGRU model achieves the highest accuracy of all models, with 70% accuracy. However, the model also has the lowest F1 score, indicating that it did not balance precision and recall as well. The high accuracy, and low F1 score, indicates that the model was fitted to predict the majority class more often than the minority class. The VSAE model achieved the highest F1 score of 0.678, but the lowest accuracy of the three models. This indicates that it best balanced precise classifications, whilst not overfitting to the dominating class, at the expense of a higher accuracy. This means that the VSAE model would most likely translate best to a balanced dataset.

3.2 Ablation Study 1: Attention Mechanism

Table 2 summarizes the performance comparison between the main self-attention model and the

cross-attention ablation variant. The main NLI-Transformer model with self-attention achieves superior test accuracy (0.6910) and F1-score (0.6542) compared to the NLITransformerCrossAttention variant (0.6275 accuracy, 0.6222 F1), indicating that jointly processing premise-hypothesis pairs through self-attention is more effective than explicit cross-attention with separate encoders.

Model	Ep.	Val	Test	F1
Self-Attn	8	0.676	**0.691**	**0.654**
Cross-Attn	9	0.653	0.628	0.622

Table 2: Attention mechanism ablation study results

Figure 7 shows that the self-attention model converges faster (epoch 8) with more consistent validation performance, while the cross-attention variant exhibits slightly higher training loss throughout, suggesting its additional complexity may require more data to fully leverage its capacity. The self-attention model's superior performance (Figure 8) can be attributed to fewer parameters enabling better generalization, earlier token interaction capturing subtle semantic relationships, and simpler architecture requiring less aggressive regularization.

The confusion matrix (Figure 11) reveals both models struggle more with false negatives, adopting conservative prediction strategies influenced by training set bias. The self-attention model achieves 46.1% recall on entails versus 84.2% on neutral, reflecting dataset imbalance. Attention weight analysis (Figure 12) shows the model assigns higher weights to semantically important tokens, with strong attention alignment in correctly classified samples, though sometimes over-attending to lexical overlap in misclassified cases.

The ablation demonstrates that self-attention with stable pooling is superior for this NLI task, achieving higher accuracy (69.10% vs 62.75%) and F1-score (0.6542 vs 0.6222). While cross-attention provides an intuitive mechanism for modeling premise-hypothesis relationships, the additional architectural complexity does not translate to improved performance, possibly due to limited training data and self-attention's ability to implicitly learn similar interaction patterns.

3.3 Ablation Study 2: Effect of Reconstruction Weight γ

Figure 5 shows validation and test performance across $\gamma \in \{0.0, 0.1, 0.3, 0.5\}$. Validation macro-F1 improved steadily up to $\gamma = 0.3$, indicating that moderate reconstruction promotes more balanced and semantically coherent latent representations. However, test macro-F1 peaked at $\gamma = 0.0$, suggesting that stronger reconstruction slightly reduces generalisation by overconstraining the latent space. Accuracy followed the same trend, confirming that reconstruction acts as a mild regulariser beneficial only up to a point.

Overall, $\gamma \approx 0.3$ achieved the best trade-off between semantic regularisation and discriminative accuracy, supporting the hypothesis that limited reconstruction weighting enhances class balance without impairing generalisation.

4 Conclusion

This report assessed three distinct model architectures for the application of NLI. A bidirectional GRU (BiGRU), variational siamese autoencoder (VSAE), and transformer model with self-attention were applied to classifying premise-hypothesis pairs for their textual entailment. The BiGRU model achieved the highest accuracy, at 70%, whilst the VSAE model achieved the strongest F1 score, of 0.678. The VSAE model was determined to be the superior model, for its ability to balance precision and recall across both classes. The ablation studies compared the impact of a self-attention module against a cross-attention module, and found that self-attention is preferred for this NLI task. The additional ablation study on reconstruction weight γ results showed that moderate reconstruction weighting ($\gamma \approx 0.3$) enhanced class balance and latent structure, while stronger reconstruction reduced generalisation.

5 Limitations and Future Work

This study faces several limitations including the relatively small training set (22,122 samples), significant class imbalance leading to conservative predictions with low recall on the entails class, and reliance on static GloVe embeddings that cannot adapt to task-specific semantics. The binary classification framework also excludes contradiction relationships, limiting applicability to standard three-way NLI benchmarks. Future work could address these issues by incorporating data augmentation, employing contextualized embeddings such as BERT or RoBERTa, extending to three-way classification, and exploring ensemble methods that combine the complementary strengths of the implemented architectures.

References

Itamar Arad, Lior Dolev, and Gadi Weiss. 2023. Using efficientnet-b7 (cnn), variational auto encoder (vae), and siamese twin neural network for human movement analysis. *Journal of Personalized Medicine*, 13(5):874.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Min Li and Chun Yuan. 2025. Boosting neural language inference via cascaded interactive reasoning. *Preprint*, arXiv:2505.06607.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *Preprint*, arXiv:1904.01172.

Supriyono, Aji Prasetya Wibawa, Suyono, and Fachrul Kurniawan. 2024. Advancements in natural language processing: Implications, challenges, and future directions. *Telematics and Informatics Reports*, 16:100173.

A Contributions

Table 3 highlights the contributions of each student to this project.

Name	Student Number	Contribution (%)
Mitchell Otley	23475725	33
James Wigfield	23334375	33
Mahit Gupta	23690265	33
Total		100

Table 3: Student contributions.

Mitchell:

- Data preprocessing
- Bidirectional GRU Model
- Report sections 1, 2.1, 2.4, 3.1, 4

James:

- VSAE Model
- Reconstruction Loss Ablation Study
- Report sections 2.2, 2.6, 3.3, 4

Mahit:

- Transformer Models
- Attention Ablation Study
- Report sections 2.3, 2.5, 3.2
- Abstract

B Figures

The following section contains figures referenced in the report.

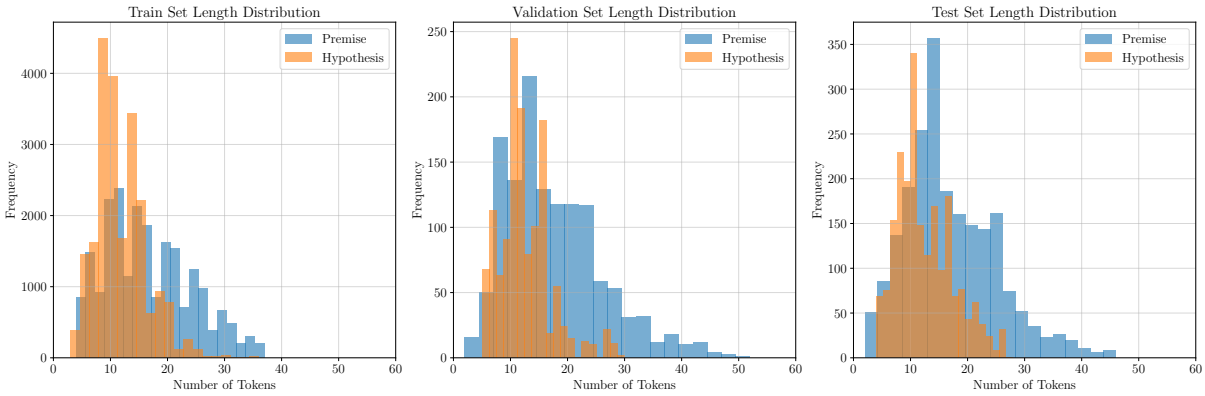

Figure 1: Distribution of tokenised premise and hypothesis lengths in the train, validation, and test sets.

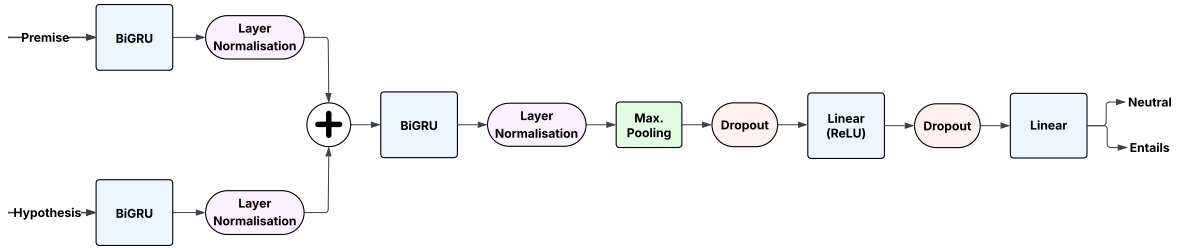

Figure 2: Model diagram illustrating the structure of the bidirectional GRU model implementation.

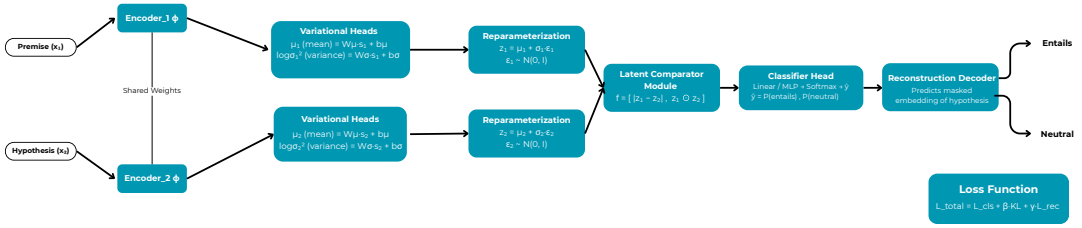

Figure 3: Model diagram illustrating the structure of the Variational Siamese Autoencoder.

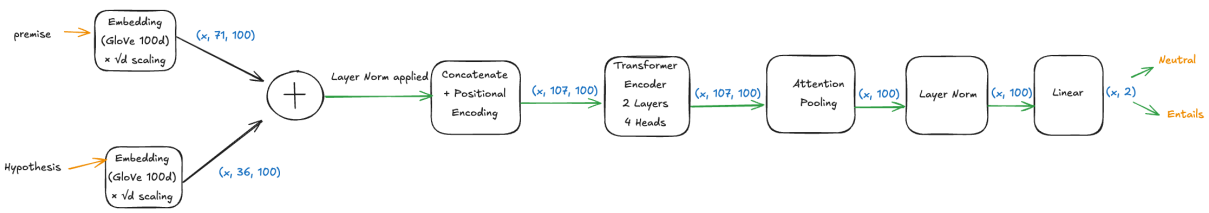

Figure 4: Model diagram illustrating the structure of the attention based transformer.

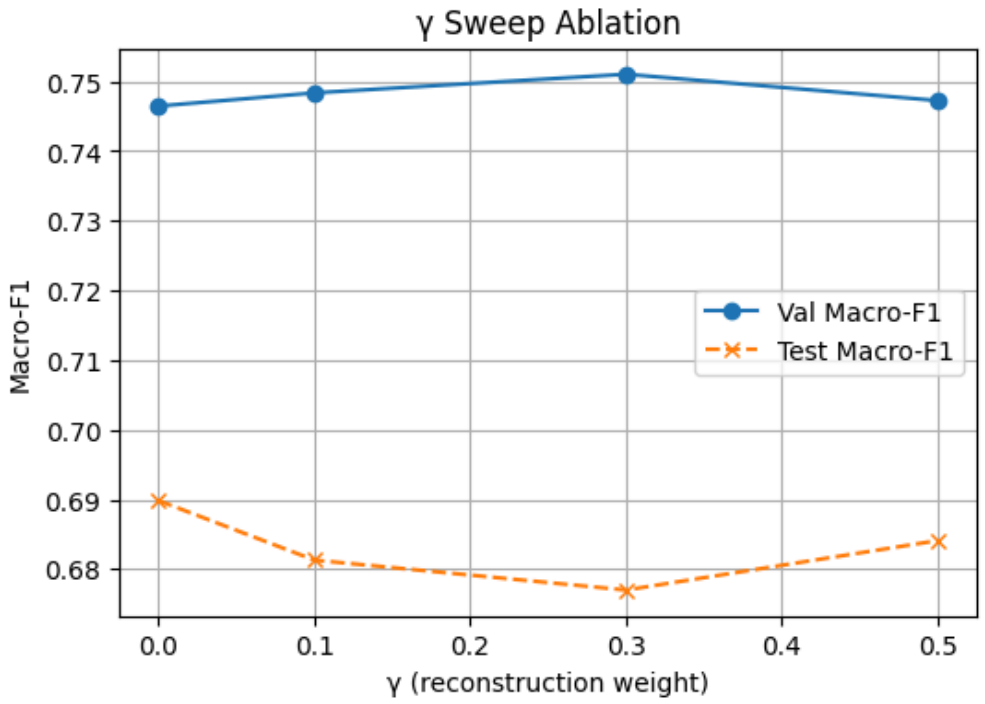

Figure 5: Validation and test macro-F1 for varying reconstruction weights γ.

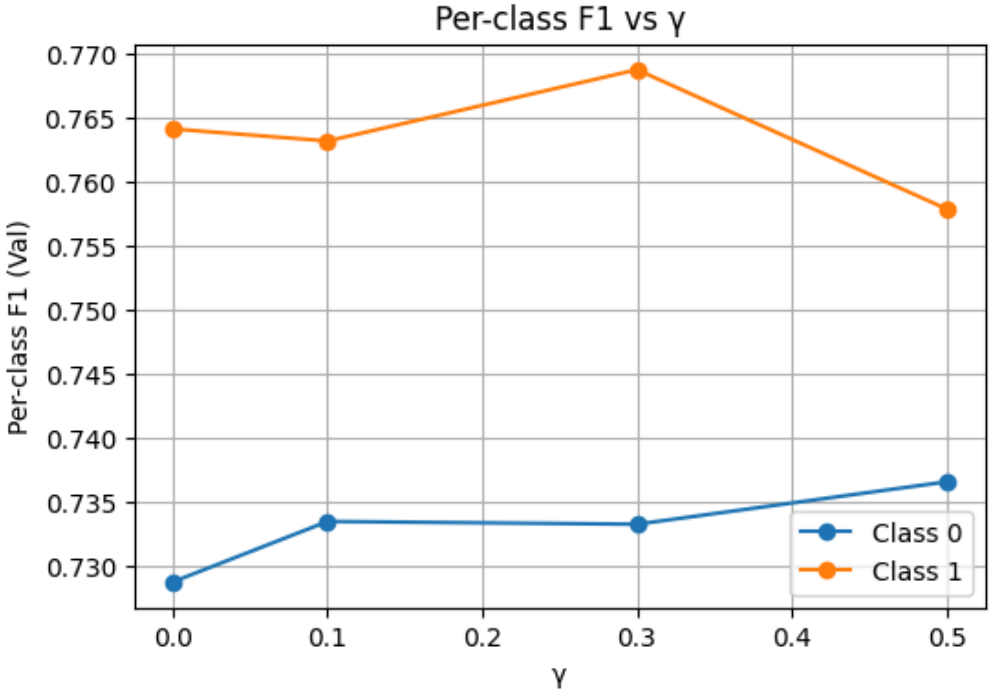

Figure 6: Per-class F1 scores on the validation set across reconstruction weights γ.

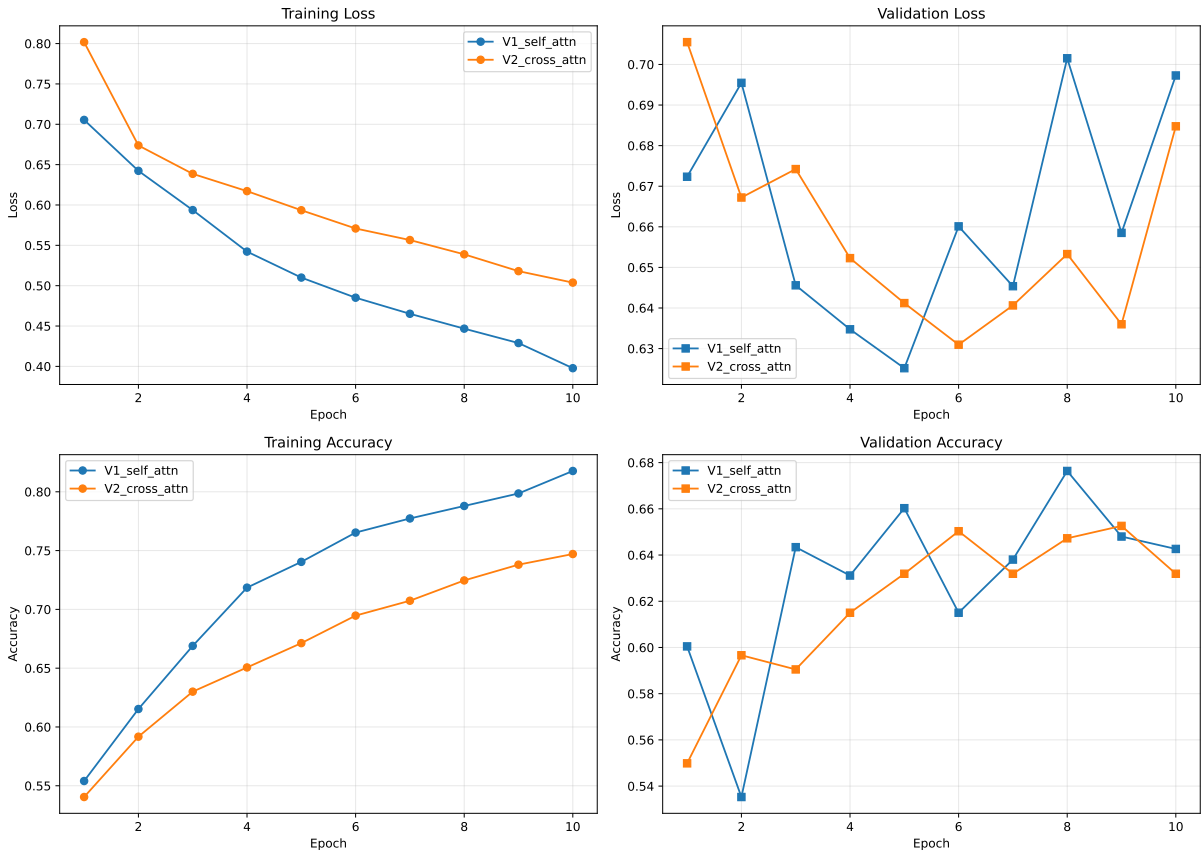

Figure 7: Diagram illustrating the training curves of best performing attention model.

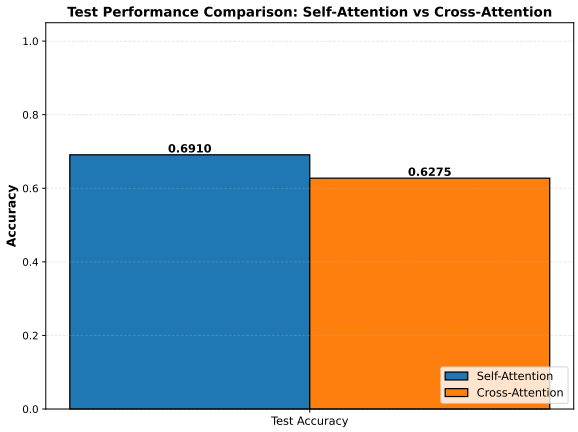

Figure 8: Test performance comparison between self-attention and cross-attention transformer models.

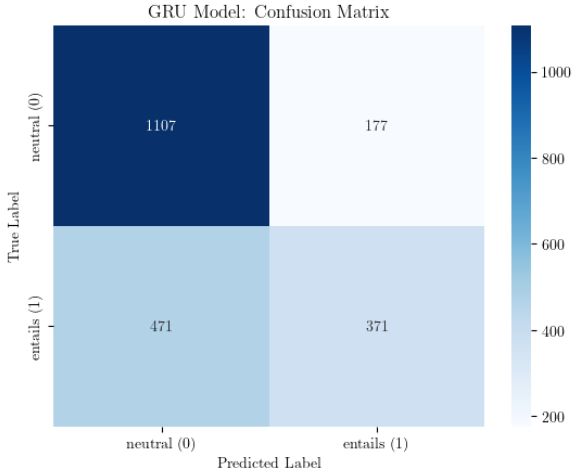

Figure 9: Confusion matrix for the best performing GRU model, showing prediction patterns on the test set.

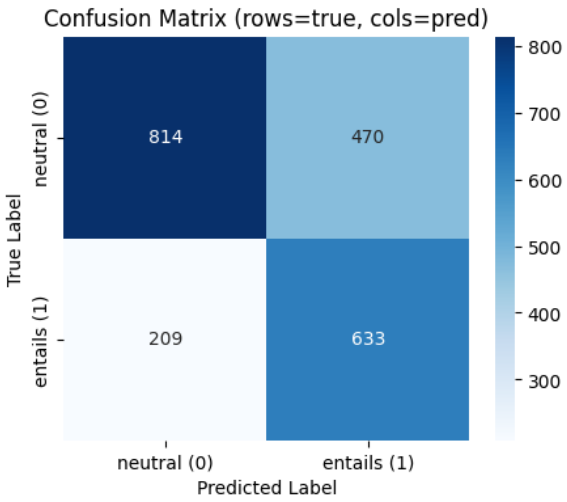

Figure 10: Confusion matrix for the best performing VSAE model, showing prediction patterns on the test set

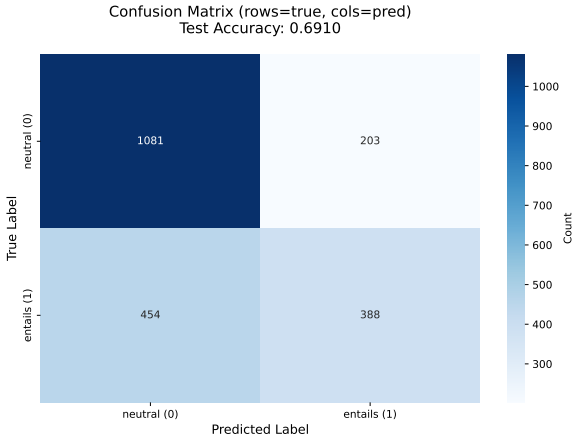

Figure 11: Confusion matrix for the best performing self-attention model, showing prediction patterns on the test set.

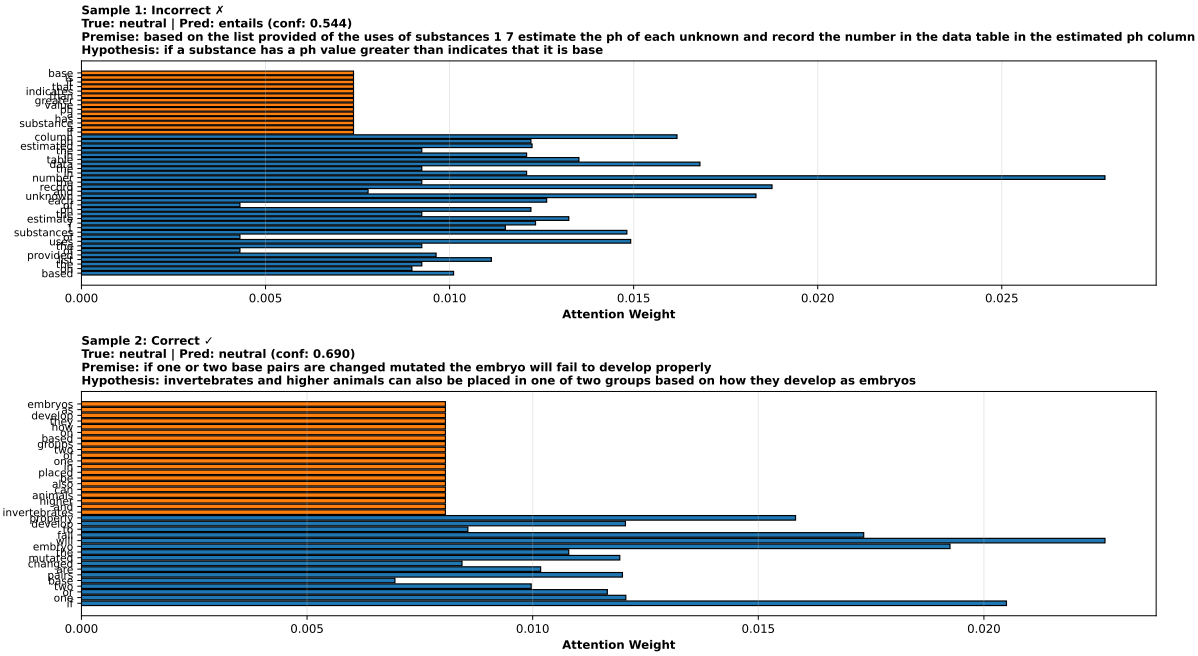

Figure 12: Attention weight visualization for two test samples. Sample 1 (top) shows an incorrect prediction where the model over-attends to lexically similar terms without capturing semantic neutrality. Sample 2 (bottom) demonstrates correct classification with appropriate attention to key semantic tokens. Blue bars indicate premise tokens, orange bars indicate hypothesis tokens.