

Multimodal Argument Retrieval using FAISS, Graph Neural Networks, and Attention

Calvin Stahoviak

April 2025

University: University of New Mexico

Course: ST, Advanced Machine Learning

Instructor: Professor Trilce Estrada

Student: Calvin Stahoviak

Assignment: Lab 2, GNN, Attention, and RAG

1 Introduction

Retrieval augmented generation (RAG) is a technique used to optimize or enhance the output of a large language model [4]. In this technique, an LLM references supplemental documents outside of its training dataset to extend the capabilities of the LLM to specific domains or tasks without the need of retraining. RAG has been credited as a cost-effective method for improving the performance of natural language processing (NLP) tasks. LLMs classically suffer from hallucinating information, presenting false information, or bad references to the wrong knowledge in its responses. RAG is an innovative method for addressing these challenges. In this work RAG is implemented and discussed on a knowledge base.

To analyze the quality of our RAG implementation, this report uses popular subreddits as a text corpus. The subreddit of choice for the analysis in this report is r/AITAH, a community where users are encouraged to post stories or situations about themselves and ask others “am I the asshole?” for making said choice. This subreddit is chosen because it comprises almost entirely of text submissions and comments which often involves users arguing or making a case for their opinions. This subreddit is an excellent environment to test RAGs ability to find relevant documents and use them in a rebuttal. In addition to r/AITAH, this report briefly covers r/Silksong, a much smaller community banded around discussing the video game Silksong, which had been delayed for several years up until recently. r/Silksong is chosen because of the unique mannerisms, speech patterns, and inside-jokes that the community has built up over the years. This report seeks to discover if RAG can motivate an LLM to respond with more knowledge or mimic the communities speech patterns.

2 Data Pre-Processing

Comments and their associated metadata are scrapped from reddit using the *PushShift.io API*. Subreddits range in size from several thousands to hundreds of millions of comments. To lower computational cost and for *reasonable compute* given the available resources each dataset is truncated to 100,000 comments. The metadata for each comment, including author, score, created, and link are all removed, leaving only the body attribute, which is the raw text of the comment.

3 Embedding Methods

The raw text collected from these comments are embedded using several different methods, which are later compared against each other. In this section, sentence transformer, graph convolution network, concatenation, and self-attention based fusion methods are detailed.

3.1 Text Embedding

The body, or text, of these comments are used as input to a sentence transformer to create sentence embeddings. *Sentence embeddings* are vectors that aim to capture the semantic meaning of a body of text. The sentence transformer of choice is "`all-MiniLM-L6-v2`", a HuggingFace sentence embedding model which maps sentences and short paragraphs into 384-dimensional dense vector space. This model is the result of a rich history of research in model architecture, distillation and fine tuning strategies. The model "`all-MiniLM-L6-v2`" is built upon the "`MiniLM-L12-H384-uncased`" model, which is a 6-layer variant of MiniLM, a distilled model discussed in detail in the paper *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers* [1]. MiniLM presents and terms a process *deep self-attention distillation*, which compresses larger pretrained models while preserving performance. MiniLM is able to maintain performance of up to 99% accuracy using only 50% of the large transformers parameters. This allows for faster inference, reduced computational cost, and downstream task-specific layers to be trained substantially faster.

The backbone of our model, the `MiniLM-L6-H384-uncased`, has 6 transformer layers. This distilled model is then fine-tuned on over 1 billion sentence pairs to create our sentence transformer model of choice, "`all-MiniLM-L6-v2`". The fine-tuning is done using contrastive learning to pull semantically similar embeddings close together and push unrelated embeddings apart.

At inference time, a sentence goes through several stages of the model to arrive at an embedding. First, the input sentence is tokenized using the WordPiece tokenizer. WordPiece splits sentences into sub-word units called *word pieces*. An input of proper format for the transformer model is then prepared. this input consists of token IDs, an attention mask to exclude padding if the input is less than 128 tokens, or truncation if the input is greater than 128 tokens. Next, the input is passed through 6 layers of attention and feedforward networks resulting in a semantic embedding of 384 dimensions for each token of the input. Lastly, mean pooling averages all valid token embeddings, arriving at a single embedding for the input sentence. This process is done for all of the comments scraped from the subreddit r/AITA resulting in a sentence embeddings dataset. This sentence embeddings dataset is then improved using the following methods: graph neural networks, node2vec, and multimodal self-attention based fusion.

3.2 GCN Node Embeddings

A graph representation, G , is made with the text embeddings using NetworkX and Facebook AI Similarity Search (FAISS) libraries. FAISS is responsible for quickly and efficiently computing similarity between all pairs in the embedding dataset. Similarity in this case is computed using the cosine similarity metric, defined below.

$$\text{sim}_{\text{cos}}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (1)$$

Using this library, the $k = 10$ most similar embeddings are found for each embedding and used to build a graph. Each text comment becomes a node in graph G which is connected via edges to the k most similar text embeddings. Edge weights represent the similarity score between nodes. This finishes our transformation of a text corpus into a structured graph where semantically related embeddings are connected or near to each other.

A graph convolutional network (GCN) is used to refine our comment embeddings and make them *graph-aware*. This means that the GCN learns to incorporate semantic information from the k nearest comments and our new node embeddings capture discussion patterns and the contextual information of similar comments. The GCN architecture consists of a 384 dimension input layer, followed by a 128 dimension hidden layer, and finally an output layer that returns to a 384 dimension vector space. The training process is unsupervised, or *self-supervised*, using the original embeddings as both the input and the target. The model seeks to reconstruct the input embeddings with incorporated information from neighboring nodes in the graph. It does this by employing a loss function that compares the GCN output embeddings with the original text embeddings. The end result of this model are enhanced embeddings that are made more aware of semantically similar comments.

The limitation of this approach is that the immediate semantic context that any given node is able to realize is limited to its k neighbors. There is no graph traversal mechanism in place to find longer chains of semantically connected nodes. Furthermore, since the GCN only has one hidden layer, that

semantic information can at most propagate one node away. These limitations are mitigated by using node2vec.

3.3 Node2Vec Embeddings

Node2Vec is an algorithm that leverages random walks on a similarity graph to learn embeddings that capture both semantic meaning and structural relationships in the graph [3]. Nodes that co-occur on a random walk have their embeddings pulled closer together. Representations are learned by optimization for the prediction of node neighborhoods, which are discovered through biased random walks. These random walks are done on the same graph representation G used previously.

Parameter	Value
Embedding dimension	384
Walk length	10
Context window size	5
Walks per node	10
Negative samples per positive	1
Return parameter (p)	1.0
In-out parameter (q)	1.0
Batch size	128

Table 1: The node2vec model hyperparameters used in this work.

Node2vec is implemented using the parameters listed in Table 1. Node2vec already exhibits a dramatically longer training time than all other embedding methods, so parameters such as walk length, and walks per node have purposely been chosen to be small but also effective. Some parameters do not have trivial meanings – the context window size means that nodes up to five hops away from the target node are considered related to each other. The return parameter is the likelihood of returning to the previously visited node, and the in-out parameter is the likelihood of BFS vs. DFS search. During the training process, the model performs random walks to generate positive samples. Negative samples are then automatically generated, and loss is calculated based on how well the current embeddings predict which nodes appear together on walks.

Although a stronger model than GCNs at identifying local structures and capturing long-range dependencies, node2vec has its own brand of limitations. Node2vec is a highly flexible model and also extremely sensitive to its hyperparameters. For large graphs, node2vec is also significantly more expensive, which is seen in Figure 3.

3.4 Multimodal Embeddings

This report presents two methods for using our text and graph embeddings to create a multimodal embedding. The first method is a simple concatenation. Using both the sentence transformer embeddings and Node2vec embeddings with graph-awareness, concatenate them into a new set of embeddings that has 768 dimensions.

The second, more involved method, fuses together sentence transformer embeddings and graph embeddings using self-attention to create a new set of 384 dimension multimodal embeddings that leverage the semantic knowledge of both sets. Two versions of this model are created, a simple model with only a linear and a graph attention network (GAT) layer, and an enhanced model which is described next. The enhanced model begins by separately processing both embeddings through a linear and ReLu layer. Then these embeddings are concatenated and moved through a linear, Relu, linear and softmax layer. The output here is the connected with the original embeddings to create a cross-modal fusion. This is passed into a GAT layer where the model performs attention and selects the most relevant nodes in our graph. The full architecture is visualized in Figure 1.

There were several reasons to implement this added complexity. Residual connections preserve the semantic information from the original text embeddings, which is the backbone of this entire operation. Transformer-like attention layers that fuse our embeddings selectively identify the most relevant features of both. Multiple layers of multi-headed GAT layers seek to enable the model to find more complex structural information from our graphs.

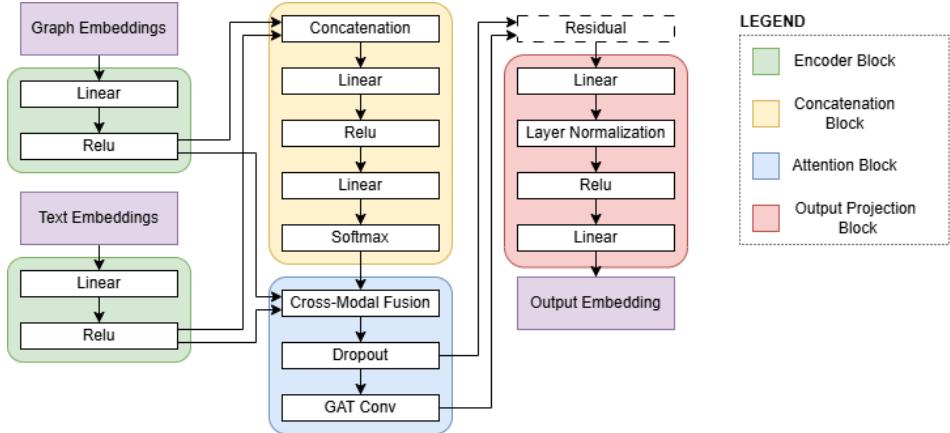


Figure 1: Architecture of the enhanced multimodal self-attention based fusion model.

3.5 Retrieval Augmented Generation

The last step of this algorithm is to use our various embeddings to enhance the response of a large language model. The LLM of choice is *GPT-4o*. This process begins by creating a FAISS index for each embedding set. Then a question answering (QA) chain is created using the FaISS index. A QA chain is responsible for the high-level question-answering process. It receives a user query, converts into the embedding space using the correct model, finds similar documents using a vector database, finds the matching raw text associated with these similar documents using a document store, passes both the question and the retrieved documents to the LLM, and queries a for a response. A QA chain is built for each embedding set.

4 Results & Discussion

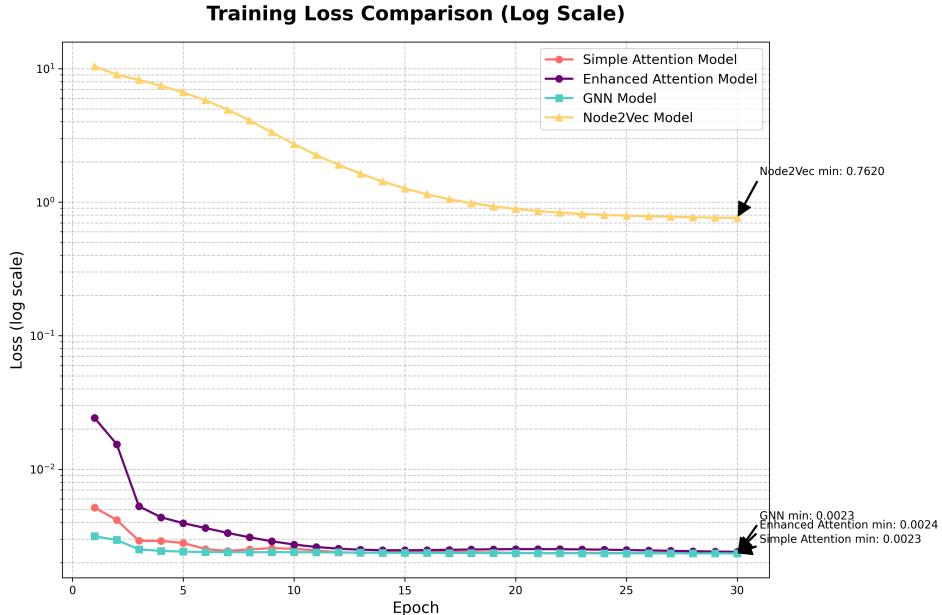


Figure 2: Comparison of loss curves for different embedding models on a logarithmic scale for the r/AITAH text corpus.

All four models are trained for 30 epochs and reach a near equilibrium state as seen in Figure 2. Almost all models reach a very similar loss as well, except for node2vec, which is more than two factors

larger. This is because node2vec is utilizing an entirely different loss function. Node2vec uses negative sampling loss based on the skip-gram model from word2vec, while the other models use mean squared error (MSE) loss. Although the value of the loss function for node2vec is much higher than does not necessarily mean it is performing worse than the other models.

Four different models are used to enhance the embeddings outputted by the sentence transformer. It's clear that Node2vec has a substantially longer training time than all other models as seen in Figure 3. Notably that node2vec is about $140x$ slower than our simple multimodal attention model. This is because node2vec uses many trials of random walks to learn the structure of the graph, and for dense large graphs this process becomes computationally heavy. Even when choosing generously small walks for node2vec, there is a considerable difference in training time.

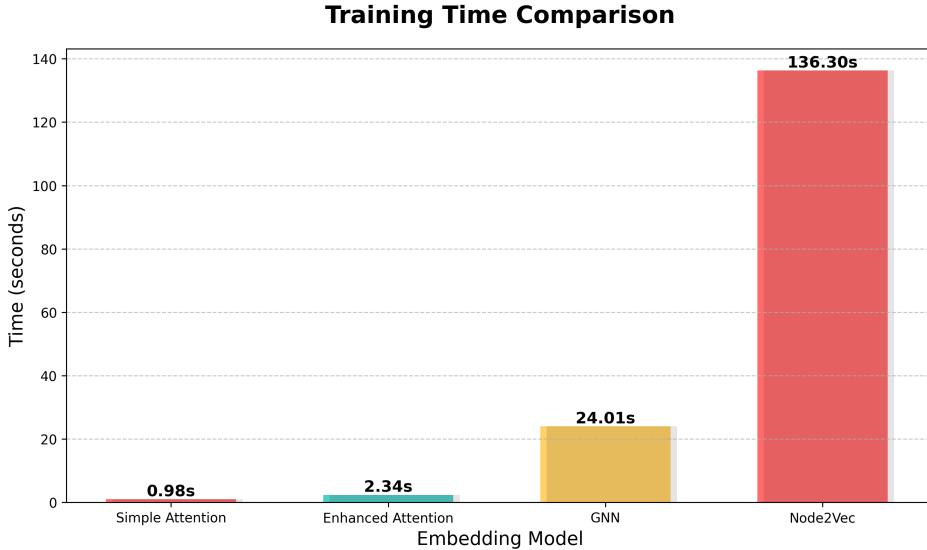


Figure 3: Training time comparison across embedding models the r/AITAH text corpus.

Embeddings and text-to-text responses can be difficult to find appropriate metrics for since the quality of the output is often subjective. A standard way to visualize an embedding space is through a technique called *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [5]. This is a non-linear dimensionality reduction technique, commonly used to reduce high-dimensional data down to 2 or 3 dimensions for a more understandable visualization. This technique has the benefit preserving local structure but distorts global distances, making clusters of embeddings seem either closer or farther apart than they truly are. This method is known for being computationally expensive and has since been improved up with the Barnes-Hut t-SNE approximation method which is used in this report [2].

All of the t-SNE embeddings are visualized side by side in Figure 4. This visualization plots each embedding in two dimensional space with some transparency, to identify which areas are more dense. The figure Figure 5 plots the same two dimensional representation but using hex bins and a logarithmic color scale. A logarithmic visualization is chosen to allow lower density areas of the embedding space to be more prominent. Notice the subtle changes that make each model different. There are differences in the amount of clusters, their size, and their density. The GNN embedding is the most peculiar, with almost all of the embeddings forming a ring around a single super dense cluster. Also notice the enhanced attention model creating the only non-circular embeddings space, with you massive clusters on the top and bottom.

Looking now at the results for RAG in Table 3, this report subjectively analyzes how well our documents are aiding the LLM in answering the query. Unfortunately, almost all of the documents for all the models are very specific to other posts and specific situations. Since most posts on this subreddit involve people telling stories about themselves, its hard to find comments that offer generalizable advice. Inserting specific information from relevant sources seems to hurt the output in this case. The LLM becomes confused about the similar sentiment but irrelevant details in the provided documents. Overall RAG may be a effective tool, but this domain seems to be the wrong application.

Similar graphs and tables for the subreddit r/Silksong are show in Appendix A. Notice a similar



Figure 4: Comparison of t-SNE embeddings across all model variants the r/AITAH text corpus.

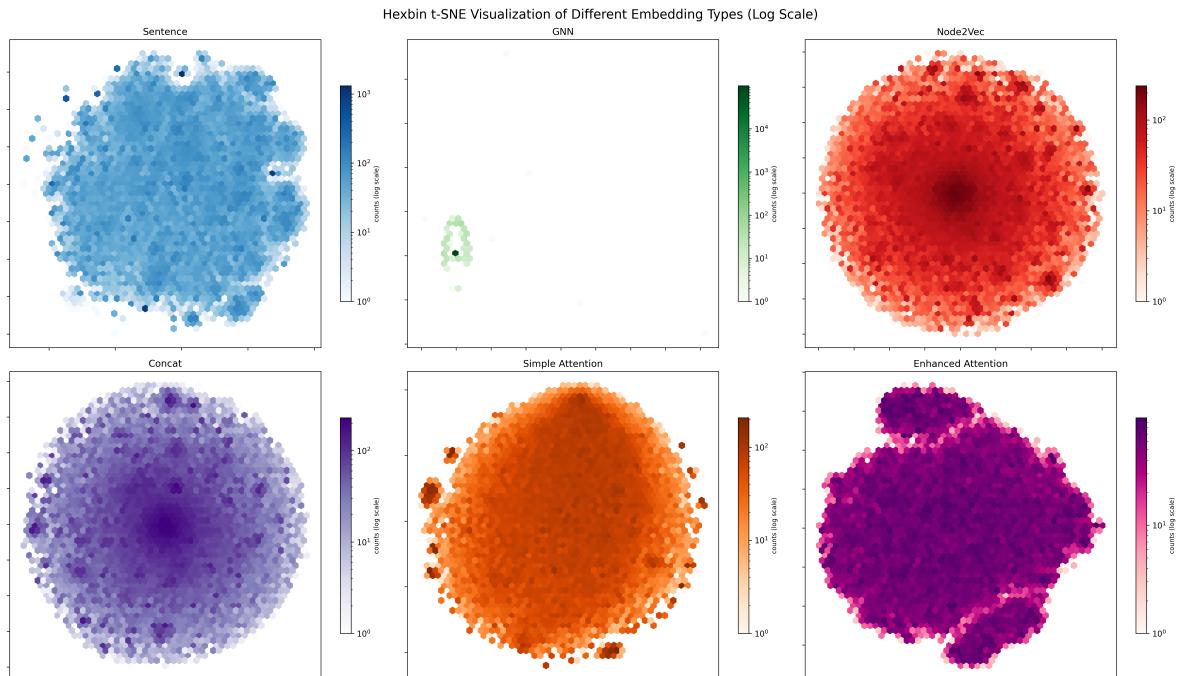


Figure 5: Hexbin visualization of t-SNE embeddings comparing density across all model variants the r/AITAH text corpus.

pattern between embeddings there as well. For r/Silksong, notice in Table 2 find that although RAG is able to assemble similar documents exhibiting the unique quips and catch-phrases common in the community, ultimately the LLM isn’t able to respond in a user-like manner. A solution could be to prompt the LLM to respond in the same style and format as the documents presented in the QA chain.

5 Conclusion

In this project, RAG is used to generate responses from an LLM that are enhanced by relevant documents. To start this process comments are embedded from a popular subreddit using a sentence transformer and enhanced through several methods, including Node2Vec and multimodal attention-based fusion models. A query is similarly embedded, and used to find similar documents to enhance the output. This report finds that this method does inform the LLM of contextual information or patterns of speech in the subreddit’s community, however, it does not always utilize this information to form a natural response to the query that is typical of the subreddit. For potential future work, it would be beneficial to consider prompting the LLM to respond in the style or format of the relevant documents, while looking at a larger number of similar documents to mimic speech patterns.

References

- [1] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv preprint arXiv:2002.10957*. <https://doi.org/10.48550/arXiv.2002.10957>
- [2] van der Maaten, L. (2013). Barnes-Hut-SNE. *arXiv preprint arXiv:1301.3342*. <https://doi.org/10.48550/arXiv.1301.3342>
- [3] Grover, A., & Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *arXiv preprint arXiv:1607.00653*. <https://doi.org/10.48550/arXiv.1607.00653>
- [4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*. <https://doi.org/10.48550/arXiv.2005.11401>
- [5] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

A Supplementary Material

This section provides additional figures and tables for the subreddit r/Silksong, which was the second text corpus that was tested.

A.1 Additional Figures

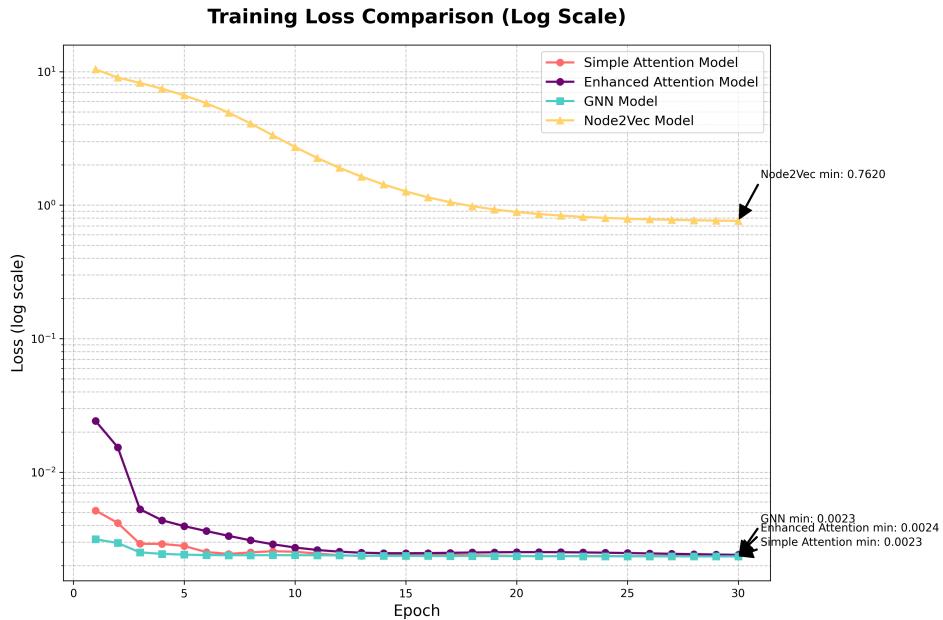


Figure 6: Comparison of loss curves for different embedding models on a logarithmic scale for the r/Silksong text corpus.

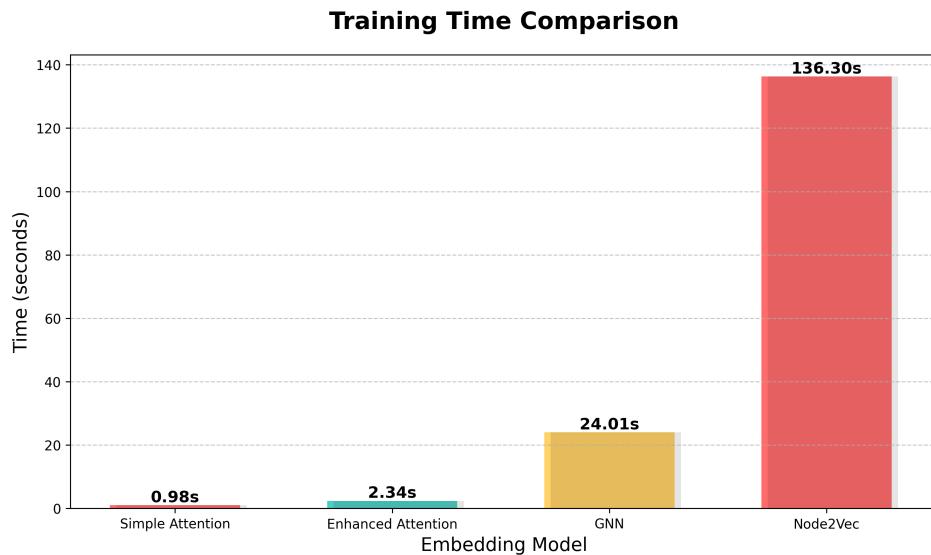


Figure 7: Training time comparison across embedding models the r/Silksong text corpus.

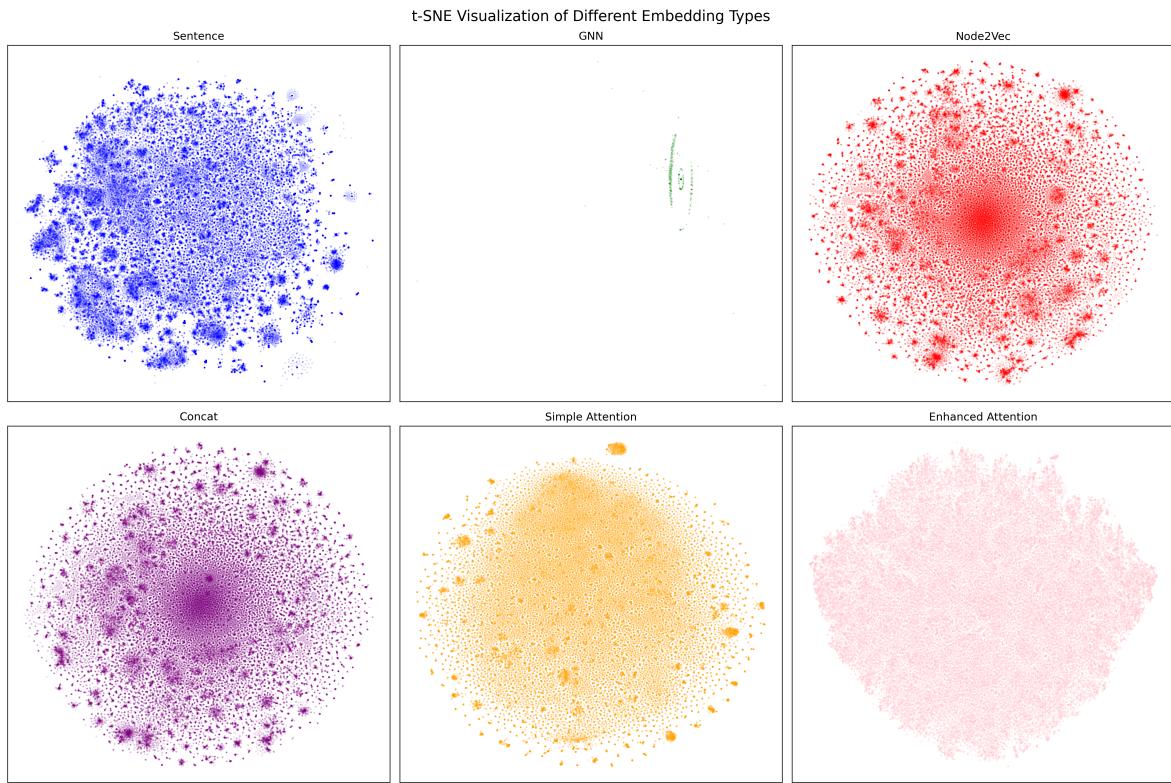


Figure 8: Comparison of t-SNE embeddings across all model variants the r/Silksong text corpus.

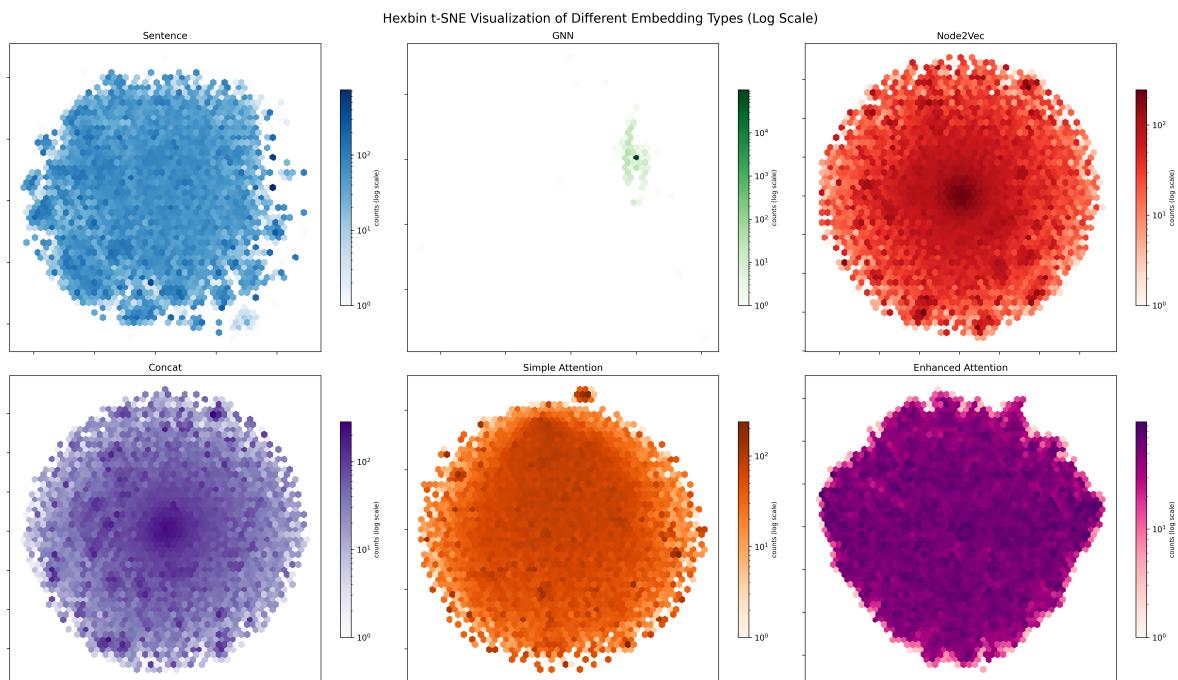


Figure 9: Hexbin visualization of t-SNE embeddings comparing density across all model variants the r/Silksong text corpus.

A.2 Additional Tables

SENTENCE EMBEDDINGS

Document 0	When will Silksong not be released?
Document 1	When will Silksong release? (i have lost almost all hope that it will release)
Document 2	When will silksong release
Document 3	When will silksong release
Answer	I don't know when Silksong will be released. There hasn't been an official release date announced yet.

GNN EMBEDDINGS

Document 0	Formula for the flatness of a universe, Omega = actual mass density / critical mass density...
Document 1	Magnetic field strength at distance = (permeability of free space * current)/(2 * distance)
Document 2	Placeholder
Document 3	Day 21 of commenting... I had a dream last night that Silksong was out...
Answer	I don't have information on the specific release date for Silksong. Development timelines can be unpredictable, but hopefully, the game will be worth the wait. Keep an eye on official announcements from Team Cherry.

NODE2VEC EMBEDDINGS

Document 0	Image link: preview.redd.it/z268g5y3ifub1.jpeg...
Document 1	Image link: preview.redd.it/xdgxdhnsautb1.jpeg...
Document 2	Fake news
Document 3	Frustration at lack of news from Team Cherry while Mongoose Rodeo shares info consistently...
Answer	I don't know the exact release date for Silksong as there hasn't been any official announcement from Team Cherry regarding it.

SIMPLE ATTENTION EMBEDDINGS

Document 0	It WILL come out 2024 or else I'm selling my appendix
Document 1	I have no idea.
Document 2	True, there is like a generation's time between each game
Document 3	Ahh the ol' post history dig up... lay off the weed... helmet...
Answer	I don't know the exact release date for Silksong. Some people speculate it might come out in 2024, but nothing official has been confirmed.

ENHANCED ATTENTION EMBEDDINGS

Document 0	Nice
Document 1	They announced at that State of Play that 'Stray' would be free... Silksong could be next.
Document 2	[clown emoji]
Document 3	I feel like it's better if we just don't let people know about game shows...
Answer	I don't have information on the exact release date for "Silksong." It's best to keep an eye on official announcements from the developers.

Table 2: Comparison of model responses and source documents for the query, “when will silksong be released? I’m feeling hopeless about it...”, using the r/Silksong subreddit.

SENTENCE EMBEDDINGS

Document 0	I totally agree, that's really bad and so disrespectful to her, that's what I meant writing, maybe I didn't express myself right, what I meant was that friendship between exes is totally possible, but if it was really just friendship and he was over he's ex, there would be no reason to hide it. In my opinion Op deserves so much better!
Document 1	INFO: does your ex know you're interested in, and talking to, her friend? If they're friends, wouldn't her friend know that you two dated and would have told her that you're talking already?
Document 2	This needs upvotes. Stop and ask yourself why she was allowing this conversation in her life. Maybe she needed the emotional validation. Someone who was having their emotional needs met would probably never have started talking to the ex. (Or anyone else)
Document 3	She literally said that she isn't gonna dictate that, but that's really fucked up of him to be friends with HER ex's, really fucked up.
Answer	Based on the context provided, it seems like there may be concerns about the impact of discussing an ex on current relationships or friendships. It's not necessarily wrong to express your feelings about a topic that makes you uncomfortable, but it's important to communicate your perspective respectfully and considerately. If you feel that the conversations about their ex are problematic or disrespectful, it might be worth having an open discussion with your friend about how it affects you or others involved. However, ultimately, it's up to your friend to decide whether they want to continue discussing their ex, and it's important to respect their autonomy. If the conversations are negatively impacting you, you might consider setting boundaries for yourself regarding these discussions.

NODE2VEC EMBEDDINGS

Document 0	Oh hell no. You need to get out of that house ASAP. If they want to keep you, tell them your rate for abusive children is \$200 a day and \$100 for normal children. You're not getting paid enough to begin with for what you do. The average rate for a live in nanny is \$15 an hour, so unless you're only working 3 hours a day, they are taking advantage of you.
Document 1	I bet she has a golden thots why he's not leaving her
Document 2	NTA
Document 3	You've got lots of issues to resolve after he crossed that line - it's huge. Thank you for your view. I'm just so confused. People say manipulation then you're saying what he would say so I'm just confused. I was on drugs before but am sober now. I haven't left any part out, we do argue, I do bad stuff too but this whole thing about my ex I haven't left anything out. What makes you think I'm ignorant ??? Thank you
Answer	Based on the context provided, it's difficult to determine whether you're in the wrong without more information about the situation. However, generally speaking, insisting that a friend stop talking about their ex can be seen as insensitive if they are going through a difficult time and need to process their emotions. On the other hand, if the conversation is becoming repetitive and negatively impacting your well-being or the friendship, it's reasonable to set boundaries. It's important to approach the situation with empathy and communicate your feelings honestly while also considering your friend's perspective.

ENHANCED ATTENTION EMBEDDINGS

Document 0	Thanks for the opinion, appreciate it
Document 1	NTA at all. I question the age difference and I apologize for doing so. If my 17 year old niece was dating a 28 year old I would be upset and concerned. I myself have struggled with depression and anxiety etc. I would highly recommend learning DBT skills, Dialectical Behavior Therapy. I am 49 and have finally come to the realization that I can't expect someone to love me if I don't love myself. I think it may help you to embrace where you are now and let him go. You deserve better and a loving relationship will come to you when you are ready.
Document 2	You deserve it all life is worth living Not the ass
Document 3	Nta I'm gonna guess niece is jealous and doesn't want her mom giving you attention. If this is a group chat that the niece is texting her mom is in the wrong.
Answer	I don't know enough about the specifics of your situation to determine if you are the asshole or not. Generally, insisting someone stop talking about their ex can depend on the context and your reasons. If your friend is constantly bringing up their ex in an unhealthy way and it is affecting their well-being or your friendship, it may be reasonable to have a conversation about setting boundaries. However, if you are simply uncomfortable without a significant reason, it might be helpful to discuss why it bothers you and find a compromise.

Table 3: Selected embedding model responses for the query, “AITA for insisting my friend stop talking about their ex?”, using the r/AmItheAsshole subreddit.