

**Investigating Emotion – Cognition Correlations in Transformer – Based Language
Models: A Neuropsychological Perspective.**

Table of Contents

Abstract	3
Introduction	4
Background	4
Problem Statement.....	4
Research Objectives.....	5
Research Scope	5
Literature Review	6
Research Methodology.....	10
Research Design	10
Quantitative Research Approach	10
Qualitative Research Approach.....	10
Research Phases	11
Phase 1: Data Preparation	11
Phase 2: Model Selection & Implementation	11
Phase 3: Emotional Representation Analysis	11
Phase 4: Cognitive Attention Analysis	12
Phase 5: Emotion Cognition Correlation Modeling	12
Phase 6: Psychological Interpretation & Evaluation	13
Ethical Considerations.....	13
Limitations	13
Expected Outcomes.....	14
Gantt Chart.....	14
References	15

Abstract

Transformer-based language models have demonstrated strong performance in emotion recognition tasks; however, the internal relationship between emotional representation and cognitive processing within these models remains insufficiently understood. In human psychology, emotion and cognition are deeply interconnected, with emotional salience influencing attention, perception, and interpretation. Motivated by this parallel, this thesis investigates emotion–cognition correlations in transformer-based language models from a neuropsychological perspective. The study adopts a mixed-methods approach combining quantitative analysis of contextual embeddings and attention weights with qualitative psychological interpretation. Multiple transformer architectures, including baseline and emotion-enhanced models, are evaluated using emotion-labeled datasets and human eye-gaze corpora. Emotional representations are analyzed across layers using dimensionality reduction and clustering techniques, while cognitive processing is examined through multi-head self-attention patterns and human–model attention alignment. Correlation analyses are conducted to assess the interaction between emotional dimensions (such as valence and arousal) and attention allocation. The results indicate that transformer models encode structured emotional information in intermediate layers and allocate higher attention to emotionally salient tokens, showing moderate alignment with human attentional behavior. Emotion-enhanced architectures further strengthen these correlations, improving interpretability. Overall, the findings suggest that transformer-based models exhibit internal emotion–cognition interactions that partially resemble human psychological processes, offering new insights for explainable affective AI and emotion-aware language modeling.

Keywords: Emotion - Cognition Correlation, Transformer – Based Language Models, Emotional Representation Learning, Cognitive Modeling, Neuropsychological Perspective

Introduction

Background

Transformer-based language models such as BERT, RoBERTa, and GPT have reshaped natural language processing (NLP) by enabling machines to capture complex contextual relationships. These models are now widely used in sentiment analysis, affective computing, psychotherapy chatbots, and decision-support systems, where understanding human emotion is crucial. Beyond language comprehension, recent research suggests that transformers may exhibit cognitive-like patterns, especially in how their attention mechanisms resemble human reading behavior. This convergence of emotional interpretation and cognitive modeling opens the possibility that such models may internally represent certain psychological processes.

Emotions and cognition are deeply intertwined in human psychology. Emotional states influence attention, memory, perception, and decision-making. Similarly, cognitive processes such as attention shape how individuals interpret emotional stimuli. Despite the well-established relationship between emotion and cognition in humans, it is unclear whether a similar interaction emerges within transformer architectures. Current literature has explored emotion detection performance, representation learning, attention alignment, and explainable AI frameworks, but no study has holistically examined how emotional meaning and cognitive attention interact inside transformer models.

This study investigates whether transformer-based models encode correlations between emotional representation (embeddings) and cognitive processing (self-attention), drawing inspiration from psychological theories and recent findings in affective NLP.

Problem Statement

Although transformer models have achieved strong performance in emotion classification tasks, existing research focuses primarily on accuracy rather than psychological interpretation. Studies by Devgan (2023) and Alqarni et al. (2025) demonstrate improvements in emotion recognition, while Guo and Choi (2021) and Zhang et al. (2022) show that emotional meaning is

embedded in vector representations. Meanwhile, Bensemann et al. (2022) provide evidence that transformer attention partially aligns with human eye gaze, suggesting cognitive similarity. Yet none of these studies examine how emotional encoding interacts with cognitive attention inside the model. This leaves a critical gap in understanding the internal dynamics of emotion–cognition relationships in transformers.

Thus, the problem this research addresses is the absence of a unified, neuropsychologically informed analysis of how transformer-based models correlate emotional representations with cognitive attention patterns.

Research Objectives

The aim of this study is to explore and analyze the internal relationship between emotional representation and cognitive attention mechanisms in transformer-based language models using psychological theories and computational techniques.

1. To investigate how transformer models encode emotional information across embedding layers.
2. To analyze how cognitive mechanisms—particularly self-attention—respond to emotionally salient text.
3. To model and quantify correlations between emotional embeddings and attention distributions.
4. To compare these correlations across different transformer architectures, including emotion-enhanced variants.
5. To interpret the findings through relevant psychological frameworks such as valence–arousal models and attention bias theories.

Research Scope

The study examines transformer-based text models (BERT, RoBERTa, and emotion-enhanced variants).

It focuses on;

1. Emotional representation in embeddings
2. Cognitive attention patterns
3. Correlation analysis between emotional dimensions and attention
4. Psychological interpretability

The study does not involve multimodal emotional data (e.g., audio/visual signals) or clinical psychological evaluations.

Literature Review

Affective computing and transformer-based NLP models shows rapid growth, with increasing attention to how these models represent emotions and process cognitive signals such as attention. However, existing research treats emotional representation and cognitive attention separately, leaving limited understanding of how these two processes interact within transformer architectures. Humans, and likely some animals, possess the remarkable capacity for introspection: the ability to observe and reason about their own thoughts.

As AI systems perform increasingly impressive feats of cognition, it is natural to wonder whether they possess any similar awareness of their internal states. Modern language models can appear to demonstrate introspection, sometimes making assertions about their own thought processes, intentions, and knowledge. However, this apparent introspection can be, and often is, an illusion. Language models may simply make up claims about their mental states, without these claims being grounded in genuine internal examination. Emotion embeddings in Transformers capture emotional context, while attention weights show how the model focuses on relevant emotional/cognitive features, revealing how emotions influence cognition (e.g., anxiety affecting knowledge focus) and vice-versa, using mechanisms like cross-attention to link emotional states with cognitive processes (learning, memory, decision-making) within the model architecture, mirroring real human emotion-cognition interplay. This chapter reviews the literature in four dimensions: (1) emotion detection using transformers, (2) emotional representation and affect encoding in embeddings, (3) explainability and multi-granularity frameworks, and (4) cognitive alignment between models and humans. The chapter concludes by identifying the gaps that motivate this thesis.

Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) represent a major shift in emotion detection research. These models leverage self-attention to capture long-range dependencies, making them well-suited for emotional interpretation where

context drives meaning. Devgan (2023) demonstrates that contextual cues significantly improve emotion recognition performance, confirming earlier claims by studies such as Demszky et al. (2020), the creators of GoEmotions, who showed that transformer encoders dramatically outperform classical machine learning models on fine-grained affect classification.

Alqarni et al. (2025) extend this by proposing an Emotion-Specific Attention (ESA) mechanism and TF-IDF gating. Their RoBERTa-enhanced architecture prioritizes emotionally salient tokens and filters irrelevant information, aligning with earlier ideas from Wang & Manning (2012), who emphasized the importance of term-frequency weighting for sentiment tasks. Furthermore, Saravia et al. (2018) highlight the importance of hierarchical contextualization in emotion detection—emotions often span clauses and sentences, and transformers capture these complexities better than CNNs or RNNs.

Together, these works establish that transformers not only achieve high emotion detection accuracy but also rely on contextualized, attention-driven cues. However, these studies largely treat attention as a functional mechanism for classification, not as a cognitive process with psychological relevance. This opens space for further exploration into whether emotional valence or intensity modulates attention in a way that resembles human emotion–cognition interactions.

A parallel research direction explores how emotional meaning is encoded within embeddings. Guo and Choi (2021) demonstrate that emotional dimensions such as joy, anger, or fear naturally cluster in embedding space, even without explicit emotional supervision. Similar insights come from Buechel & Hahn (2017), who developed emotion lexicons using regression models and found that embeddings reflect valence and arousal dimensions.

Zhang et al. (2022) extend this work by enriching embeddings with affective information (e.g., valence, arousal, dominance), demonstrating improved performance on downstream emotion-related tasks. Their work builds on earlier theories from Mohammad (2018), who advocated for mapping emotions to psychological models rather than fixed categories. These embedding studies collectively indicate that affective meaning is to some extent distributed within the parameters of large language models.

However, embedding-level studies often examine emotional structure in isolation. They do not analyze whether these affective patterns influence, or are reflected in, cognitive components such as attention heads. Similarly, they do not explore how affective structures change across layers,

even though research by Tenney et al. (2019) shows that linguistic features (syntax, semantics, NER) emerge at different depths in transformers. This thesis builds on such probing work by analyzing both embeddings and attention across layers to examine whether emotional encoding predicts cognitive allocation.

As transformer-based models grow in complexity, explainability becomes essential—especially for applications in mental health, human–AI interaction, and safety-critical systems. Gamage et al. (2024) propose the Emotion AWARE architecture to improve robustness and interpretability in emotion analysis. Their framework incorporates lexicon knowledge, fine-tuned models, and multi-granular emotional categories—an approach similar to earlier multi-layer models such as Hierarchical Attention Networks (Yang et al., 2016).

Explainability in NLP has traditionally relied on saliency maps (Li et al., 2016), attention heatmaps, or gradient-based methods. However, these techniques often fail to account for distributed representations and multi-head attention complexities. Jacovi & Goldberg (2020) argue that attention weights alone do not always convey reliable explanations, though they may still be psychologically meaningful.

Thus, while frameworks such as Emotion AWARE support system-level transparency, there is a knowledge gap regarding how explainable outputs map onto internal model mechanisms—particularly emotional embeddings and cognitive attention. This thesis addresses this gap by investigating the representational origins of explainable patterns.

The cognitive interpretation of transformer attention is gaining traction. Bensemann et al. (2022) demonstrate that transformer attention correlates significantly with human eye-gaze behavior during reading, especially for simpler texts. Their findings align with earlier cognitive-NLP research such as Hahn & Keller (2016), who used human reading times to evaluate syntactic processing in models.

Additionally, Clark et al. (2019) show that different attention heads specialize in particular linguistic functions—e.g., focusing on verbs, attending to subjects, or tracking coreference links. These specialized heads could be analogous to cognitive sub-processes in human reading.

However, the question remains: Does emotional salience influence attention in ways comparable to human cognition?

Psychological research suggests that emotional stimuli—especially negative or high-arousal words—capture more attention (Vuilleumier, 2005). If transformers exhibit similar patterns (e.g., paying more attention to emotionally charged tokens), this may indicate a form of emergent cognitive-emotional interplay.

Existing literature has not addressed this intersection explicitly, creating a clear research opportunity.

Although the present study focuses on text, multimodal research provides important theoretical grounding. Studies in affective neuroscience (Pessoa, 2008; Barrett, 2017) highlight how emotion and cognition co-regulate each other—emotion enhances memory encoding, directs attention, and influences reasoning. Transformer research rarely incorporates such psychological insights, though multimodal models like VisualBERT (Li et al., 2019) and emotion-aware speech models (Zadeh et al., 2017) demonstrate that emotional cues influence multimodal embeddings and attention.

These findings motivate deeper investigation into whether purely textual transformers exhibit similar interplay internally, especially at the representation and attention levels.

Research Methodology

This chapter outlines the overall methodology used to investigate emotion–cognition correlations in transformer-based language models. The research design combines quantitative and qualitative approaches to ensure both statistical rigor and interpretability. The methodology is organized into several phases, including data collection, model preparation, emotional representation analysis, cognitive attention analysis, correlation modeling, and psychological interpretation. Each phase is described in detail to demonstrate how the research objectives are systematically addressed.

Research Design

The study adopts a mixed-methods design, combining quantitative and qualitative approaches.

Quantitative Research Approach

The quantitative component focuses on:

- Extracting numerical embeddings and attention weights
- Computing correlation coefficients
- Measuring clustering quality (silhouette score, Davies–Bouldin index)
- Statistical comparison across models and datasets

This enables objective measurement of emotional structure and cognitive attention patterns.

Qualitative Research Approach

The qualitative component involves:

- Interpreting attention heatmaps and visualizations
- Analyzing embedding plots (UMAP, t-SNE, PCA)
- Examining sample sentences for emotional attention bias
- Linking model behavior to psychological theories

This interpretive layer helps explain why certain patterns emerge, ensuring meaningful psychological insights.

Research Phases

Phase 1: Data Preparation

Two categories of datasets are used:

1. Emotion-Labeled Datasets - Used for emotional representation analysis.
2. Cognitive Datasets - Used to analyze human–model attention alignment.

All datasets undergo normalization, tokenization, and padding.

Phase 2: Model Selection & Implementation

Three transformer architectures are used:

1. BERT-base (baseline contextual model)
2. RoBERTa-base (stronger contextual encoder)
3. Emotion-enhanced RoBERTa (based on Alqarni et al., 2025)
 - Emotion-Specific Attention
 - TF-IDF gating

Models are implemented using PyTorch and HuggingFace Transformers.

Phase 3: Emotional Representation Analysis

This phase examines how models encode emotional information.

Procedures:

1. Extract embeddings from each layer
2. Visualize clusters using PCA, t-SNE, and UMA
3. Measure cluster separability and purity
4. Map embeddings to psychological models

- Valence–arousal
- Plutchik’s emotion wheel

This identifies where emotional meaning is strongest within the transformer.

Phase 4: Cognitive Attention Analysis

This phase investigates how attention mechanisms respond to emotional cues.

Procedures:

1. Extract multi-head attention weights
2. Compare attention to emotional vs. neutral words
3. Examine attention across layers and heads
4. Compare model attention to human eye-gaze from ZuCo and GECO
5. Identify human-aligned attention heads

This provides insight into model cognition and attention bias.

Phase 5: Emotion Cognition Correlation Modeling

This is the central analysis of the research.

Procedures:

1. Correlate emotional embedding dimensions with attention weights
2. Compute Pearson and Spearman correlations
3. Analyze how emotional intensity affects attention allocation
4. Compare correlations across model architectures
5. Evaluate whether emotion-enhanced models show stronger alignment

This phase reveals whether emotional information influences cognitive processing inside transformers.

Phase 6: Psychological Interpretation & Evaluation

Results are interpreted using theories from affective and cognitive psychology:

- Affective Attention Bias
- Theory Valence–Arousal models
- Appraisal Theory
- Cognitive Neuroscience perspectives

Qualitative evaluation includes:

- Reviewing attention examples
- Interpreting emotionally driven cognitive focus
- Relating findings to human emotional–cognitive behavior

Ethical Considerations

- All datasets are publicly available and anonymized.
- No personal or sensitive data is collected.
- Conclusions remain computational and non-clinical.

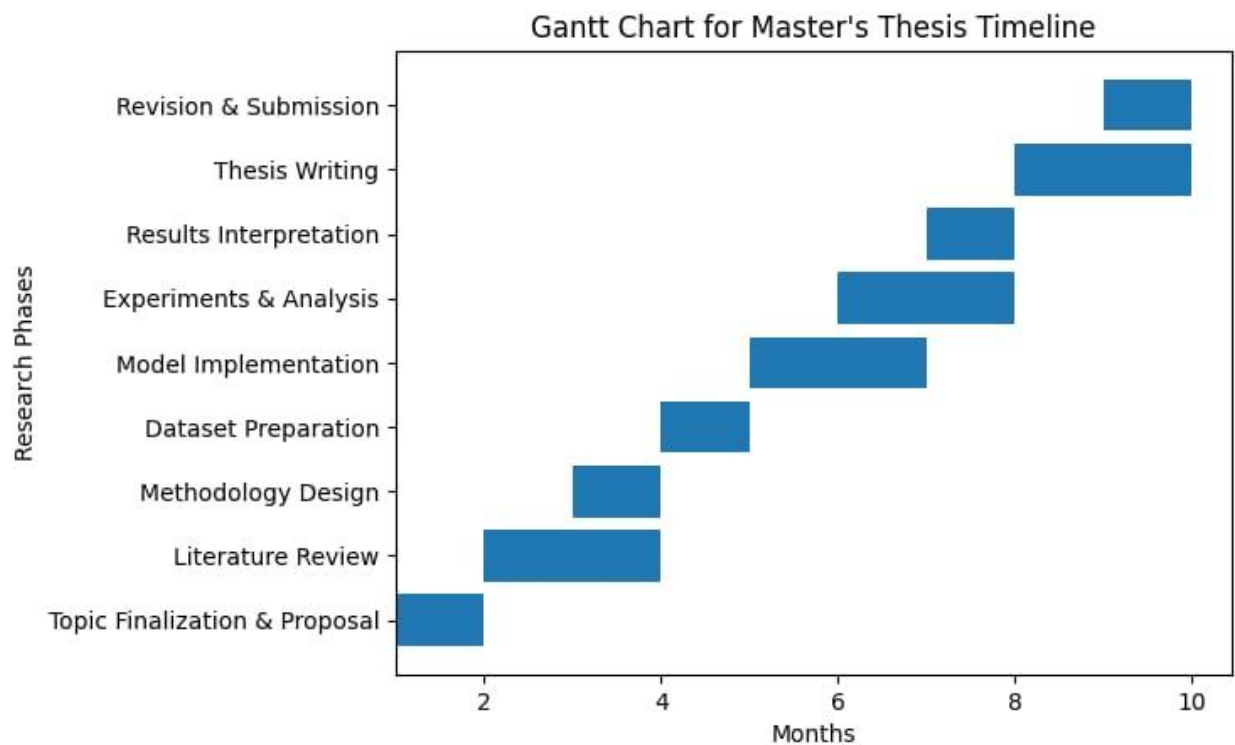
Limitations

Several limitations must be acknowledged. First, the study is restricted to text-based datasets, whereas real human emotion–cognition processes are multimodal. Second, available eye-gaze datasets such as ZuCo and GECO represent limited linguistic and emotional contexts, which may impact the strength of human–model comparisons. Third, transformer attention does not fully represent model reasoning and may only partially reflect cognitive processes. Fourth, correlation-based methods reveal associations but cannot establish causation. Computational constraints may also limit experimentation with larger models or more extensive hyperparameter tuning. Finally, psychological interpretation introduces some subjectivity despite the predominantly quantitative nature of the study.

Expected Outcomes

The research is expected to reveal structured emotional representations within the intermediate layers of transformer models. It is anticipated that emotionally salient or high arousal words will receive higher attention weights, aligning with established psychological attention biases. Emotion-enhanced models are expected to show stronger and more interpretable emotion–cognition correlations. Furthermore, moderate alignment between model attention and human eye-gaze patterns is anticipated, indicating partial cognitive similarity. Overall, the study is expected to confirm that transformer models exhibit internal emotion–cognition interaction, supporting theoretical claims from both affective computing and neuropsychology.

Gantt Chart



References

1. Alqarni, F., Sagheer, A., Alabbad, A., & Hamdoun, H. (2025). Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition. *Journal of Big Data*, 11, 93.
2. Bensemann, J., et al. (2022). Eye gaze and self-attention: How humans and transformers attend words in sentences. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
3. Devgan, A. (2023). Contextual emotion recognition using transformer-based models. *International Conference on Natural Language Processing*.
4. Gamage, G. M., et al. (2024). Emotion AWARE: An AI framework for adaptable, robust, explainable, and multi-granular emotion analysis. *Journal of Big Data*, 11(93).
5. Guo, Y., & Choi, E. (2021). Enhancing cognitive models of emotions with representation learning. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
6. Zhang, Y., Chen, W., Zhang, R., & Zhang, X. (2022). Representing affect information in word embeddings. *Proceedings of the 2022 Conference on Affective Computing and Intelligent Interaction*.
7. Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. *Proceedings of the ACL Workshop BlackboxNLP*.
8. Demszky, D., et al. (2020). GoEmotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

9. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT.
10. Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
11. Saravia, E., Liu, H., Huang, Y., Wu, J., & Chen, Y. (2018). CARER: Contextualized affect representations for emotion recognition. Proceedings of EMNLP.
12. Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. Proceedings of the ACL Workshop BlackboxNLP.
13. Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of ACL.
14. Hahn, M., & Keller, F. (2016). Modeling human reading with neural attention. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.
15. Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Proceedings of ACL.
16. Li, J., Chen, X., Hovy, E., & Xing, E. (2016). Visualizing and understanding neural models in NLP. Proceedings of NAACL-HLT.
17. Barrett, L. F. (2017). How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt.

18. Buechel, S., & Hahn, U. (2017). EmoBank: A corpus of emotion annotations averaged across annotators. arXiv preprint arXiv:1708.03687.
19. Mohammad, S. M. (2018). Word affects intensities. Proceedings of the 11th Edition of the Language Resources and Evaluation Conference.
20. Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2), 148–158.
21. Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, 9(12), 585–594.
22. Li, L. H., et al. (2019). VisualBERT: A simple and performant baseline for vision-and-language tasks. arXiv preprint arXiv:1908.03557.
23. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. Proceedings of NAACL-HLT.
24. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. Proceedings of EMNLP.