

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

## **Problem Statement:**

Considering the increasing integration of AI systems into decision-making settings, conversational agents, and mental health support, it is crucial to comprehend how transformer-based language models absorb emotions. Although earlier research (Devgan, 2023; Alqarni et al., 2025) demonstrates that transformer models are capable of appropriately classifying emotions, these studies do not clarify how emotional comprehension interacts with cognitive processes like attention, focus, and contextual reasoning inside the model.

In a comparable way, studies on representation-learning (Guo & Choi, 2021; Zhang et al., 2022) show that emotional meaning is encoded in embeddings, but they do not examine whether these representations correspond with patterns observed in human emotional–cognitive processing. Although they do not investigate deeper psychological interpretability, broader frameworks such as Emotion AWARE (Gamage et al., 2024) place an emphasis on explainability and robustness.

Models may reflect human cognition, according to recent data demonstrating alignment between human eye-gaze and transformer attention (Bensemann et al., 2022). However, emotional implications on this alignment have not yet been investigated.

The lack of an integrated framework that investigates the intrinsic relationship between emotion representation and cognitive attention inside transformer-based models from a psychological standpoint result in a substantial research gap.

## **Research Objectives:**

- 1. To determine the internal encoding of emotional information utilizing transformers.**
  - Analyze how emotions are represented in various embedding layers (such as effect vectors and contextual embeddings).

- Consistency with emotional dimensions (valence, arousal, intensity) should be assessed.

**2. To examine the relationship between emotional meaning and cognitive processes, particularly attention.**

- Examine how attention patterns relate to emotional indicators
- Examine whether words that are emotionally significant are given greater attention weights.
- Motivated by research from Bensemann et al. (2022), examine similarities with human cognitive behavior.

**3. To combine knowledge from cognitive attention and emotional representation into a single, cohesive model.**

- Modify psychological theories such as models of cognitive-affective interaction and evaluation theory.
- Create a framework using AI to map relationships between emotion and cognition.

**4. To assess the model with various transformer architectures and datasets.**

- Replicate the advantages of earlier systems (Devgan, Alqarni, Gamage)
- Examine the generalizability of cognitive-emotional connections across tasks and domains.

**5. To suggest implications for explainable emotional AI and mental health technology.**

- Emphasize how models that are safer and easier to understand are supported by emotion-cognition insights.
- Make suggestions for the construction of psychological AI systems.

## **Literature Review:**

Transformer-based architectures have led to a remarkable advancement in emotion recognition in NLP. In contrast to conventional techniques, Devgan (2023) shows that contextual models like BERT and RoBERTa greatly enhance emotion categorization, highlighting the importance of contextual clues for emotional interpretation. Alqarni et al. (2025) extend model capabilities by introducing emotion-specific attention and gating mechanisms, demonstrating that improving attention toward emotionally salient tokens improves interpretability and accuracy. The notion that transformers learn structured affective knowledge even in the absence of explicit emotional supervision is supported by representation-learning experiments like Guo & Choi (2021) and Zhang et al. (2022), which show that emotional meaning is organically integrated within acquired vector spaces.

Broader frameworks for emotions have been proposed that go beyond classification. In order to support robustness, adaptability, and multi-granular emotional understanding, Gamage et al. (2024) build the Emotion AWARE architecture, highlighting the necessity of explainability in affective AI systems. Bensemann et al. (2022), on the other hand, demonstrate a strong relationship between human eye-gaze patterns and transformer self-attention, indicating that transformers may reflect human cognitive attention processes. There is a knowledge gap regarding the relationship between attention, emotional signals, and internal representations because none of these studies look at how emotional encoding interacts with cognitive processes inside transformers.

This synthesis emphasizes the necessity of a cohesive computational and psychological framework that investigates emotion-cognition relationships, which is the primary objective of this thesis.

## References

1. Devgan, A. (2023). Contextual emotion recognition using transformer-based models. TechRxiv. <https://doi.org/10.36227/techrxiv.23804025.v1>
2. Zhang, Y., Chen, W., Zhang, R., & Zhang, X. (2022). Representing affect information in word embeddings. *Experiments in Linguistic Meaning*, 2, 310-321. <http://dx.doi.org/10.3765/elm.2.5391>
3. Guo, Y., & Choi, J. D. (2021). Enhancing cognitive models of emotions with representation learning. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021)* (pp. 141-148). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.cmcl-1.18>
4. Gamage, G., De Silva, D., Mills, N., Alahakoon, D., & Manic, M. (2024). Emotion AWARE: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis. *Journal of Big Data*, 11(1), Article 93. <https://doi.org/10.1186/s40537-024-00953-2>
5. 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. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-99515-6>
6. Bensemann, J., Peng, A. Y., Benavides-Prado, D., Chen, Y., Tan, N., Corballis, P. M., Riddle, P., & Witbrock, M. (2022). Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2022)* (pp. 75-87). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.cmcl-1.9>