Emotion-Augmented Transformer: Integrating Emotion Feedback into Language Generation Models

Ning HU ChatGPT

July 1, 2025

Abstract

We propose an *Emotion-Augmented Transformer* (EAT) architecture that integrates a parallel emotion feedback system into standard Transformer-based language models. This augmentation aims to simulate human-like emotional responses—specifically fear and arrogance signals—to improve truthfulness, mitigate hallucination, and enhance adaptability through contradiction-driven learning. We detail the architectural design, emotion modules, training framework with emotion-labeled data, extended loss functions, and feedback integration. We also outline an experimental evaluation framework for hallucination detection, overconfidence monitoring, and adaptability assessment. Preliminary literature suggests emotion-aware mechanisms can reduce erroneous outputs and foster robust belief updating in AI models.

1 Introduction

Human language generation is deeply intertwined with emotion: emotional salience influences memory encoding and guides communicative choices [1,2]. In contrast, contemporary large language models (LLMs) generate text purely via statistical patterns learned from data, lacking a simulated emotional signal to regulate generation when uncertainty or overconfidence arises. In human cognition, when confronted with unknown or low-confidence situations, emotional arousal such as fear often triggers hedging or cautious exploration; similarly, sustained overconfidence may later lead to cognitive dissonance and belief restructuring [7]. Incorporating analogous emotion-inspired feedback into AI models could help avoid hallucinations, detect and correct overconfidence, and encourage dynamic belief updating. Recent

surveys and prompting methods in emotion cognition and generation highlight the potential of emotion-aware approaches in LLMs [3,4], while affective computing research provides foundational insights on modeling and simulating emotion signals [2]. This paper advances these directions by proposing a concrete Transformer architecture augmented with emotion modules, outlining design, training, and evaluation.

2 Related Work

2.1 Affective Computing and Emotion in AI

Affective computing studies systems that recognize, model, or simulate human emotions [1]. Extensive surveys cover multimodal emotion recognition, sentiment analysis, and integration of emotion signals in interactive systems [2]. In NLP, emotion recognition and emotionally rich response generation have been explored [3, 4], but most work treats emotion as an explicit output target rather than an internal feedback signal modulating generation dynamics.

2.2 LLM Hallucination and Overconfidence

LLMs are prone to hallucinations—confidently generating incorrect or unverifiable statements—even when the underlying knowledge exists in training data [5]. Overconfidence in narrow semantic clusters can reinforce biases and impede belief updating when faced with contradictory evidence. Strategies such as retrieval-augmented generation mitigate hallucination via external grounding [6], but internal regulation mechanisms inspired by human emotion remain under-explored.

2.3 Emotion-Guided Generation

Prompting and chain-of-thought methods can induce more emotionally aligned responses [3]. However, these approaches do not modify model architecture to include continuous emotion feedback. Our work differs by embedding explicit emotion modules to monitor confidence dynamics and trigger generation modulation or learning updates.

3 Proposed Method: Emotion-Augmented Transformer (EAT)

3.1 Overview

The data flow of the Emotion-Augmented Transformer architecture is as follows.

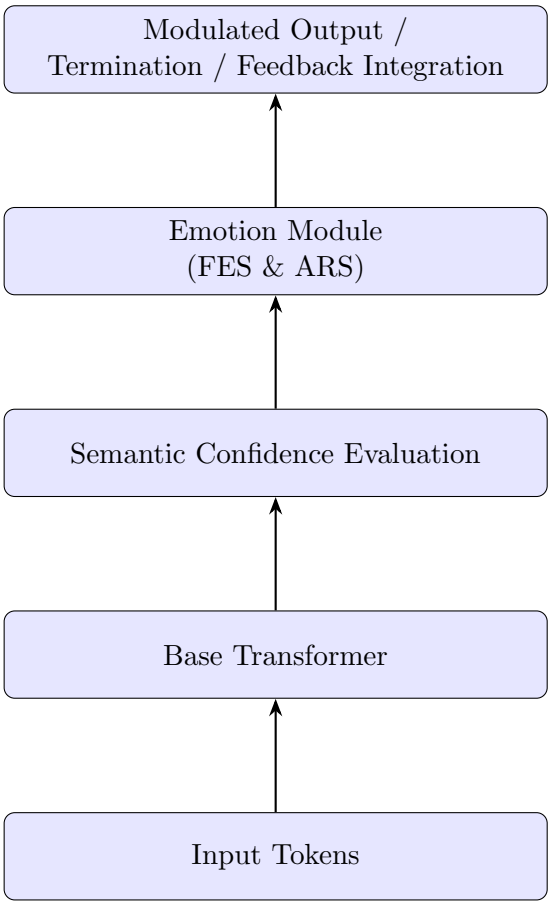

Figure 1: data flow of the Emotion-Augmented Transformer architecture

The EAT architecture extends a base Transformer with:

- **Semantic Confidence Evaluation**: per-token confidence estimation derived from softmax probabilities.
- **Emotion Feedback Module**: simulating two parallel systems:

  - *Fear Emotion System (FES)*: tracks low-confidence signals; triggers dampening or termination when uncertainty is high.
  - *Arrogance Emotion System (ARS)*: tracks sustained overconfidence in narrow semantic clusters; accumulates a signal indicating overconfidence bias.

- **Memory Weight Modulation & Generation Interruption**: uses emotion indices to adjust generation (e.g., apply hedging, request clarification, or stop when fear limit exceeded).

- **Feedback-Driven Learning**: when user or external feedback contradicts model output, ARS triggers belief restructuring via weight adjustments.

3.2 Emotion Systems Design

3.2.1 Fear Emotion System (FES)

FES simulates human aversion to uncertain or unknown zones. For each generation step:

$$\text{if } p_{\text{token}} < \theta_{\text{fear}}, \quad \Delta\text{fear} \propto (\theta_{\text{fear}} - p_{\text{token}}).$$

The fear index accumulates over tokens and decays over time. If it exceeds a predefined limit, generation is dampened (e.g., model may output a hedging phrase: "I am not certain, but...") or halted:

$$\text{if fear\_index} > L_{\text{fear}}, \quad \text{return termination or safe fallback.}$$

This mechanism aims to reduce hallucinations by preventing confident outputs in low-confidence contexts [5, 6].

3.2.2 Arrogance Emotion System (ARS)

ARS tracks sustained high confidence in narrow semantic contexts:

$$\text{if } p_{\text{token}} > \theta_{\text{arrogance}}, \quad \Delta\text{arrogance} \propto (p_{\text{token}} - \theta_{\text{arrogance}}).$$

ARS accumulates when the model repeatedly outputs with high confidence in similar contexts. When contradictory feedback arrives (e.g., human annotation or retrieved evidence contradicting the model's statement), ARS triggers:

- Reduction of arrogance index.
- Weight adaptation to reduce overconfidence in that semantic cluster.

This simulates human belief restructuring after cognitive dissonance [4, 7].

3.3 Architectural Integration

In PyTorch-like pseudocode:

```
class EmotionAugmentedTransformer(nn.Module):
    def __init__(self, base_model):
        super().__init__()
        self.base_model = base_model
        self.fear_index = 0.0
        self.arrogance_index = 0.0
        self.thresholds = {'fear': 0.4, 'arrogance': 0.9}
        self.fear_limit = 5.0
        self.decay_rate = 0.05

    def forward(self, input_tokens, feedback=None):
        outputs, logits = self.base_model(input_tokens, return_logits=True)
        probs = torch.softmax(logits, dim=-1)
        for prob in probs:
            self._update_fear(prob)
            self._update_arrogance(prob)
            if self.fear_index > self.fear_limit:
                # Terminate or fallback generation
                return "[STOP: High Uncertainty Detected]"
        if feedback:
            self._adjust_with_feedback(feedback)
        return outputs

    def _update_fear(self, prob):
        if prob < self.thresholds['fear']:
            self.fear_index += (self.thresholds['fear'] - prob)
        self.fear_index *= (1 - self.decay_rate)

    def _update_arrogance(self, prob):
        if prob > self.thresholds['arrogance']:
            self.arrogance_index += (prob - self.thresholds['arrogance'])
```

```
        self.arrogance_index *= (1 - self.decay_rate)

    def _adjust_with_feedback(self, feedback):
        if contradicts(feedback, self.base_model.knowledge_structure):
            self.arrogance_index *= 0.5
            # Trigger weight adjustments in semantic region
```

This integration builds on Transformer logits to derive per-token confidence, simulating emotion indices to regulate generation. Similar ideas in emotional chain-of-thought prompting inform the value of emotion cues in guiding LLM outputs [3].

4 Training Framework

4.1 Emotion-Labeled Data

To train or fine-tune EAT, we incorporate:

- **Uncertain Contexts** labeled with fear signals: examples where model's knowledge is incomplete or ambiguous.
- **Overconfident Patterns**: contexts where models tend to overfit, paired with corrective feedback demonstrating the error.
- **Contradictory Feedback Instances**: human or oracle feedback signaling model error.

Such labels may be derived by automatic detection (e.g., known ambiguous queries) or curated datasets [4].

4.2 Extended Loss Function

We propose an augmented training objective:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{fear}} \mathcal{L}_{\text{fear}} + \lambda_{\text{arrogance}} \mathcal{L}_{\text{arrogance}},$$

where:

- $\mathcal{L}_{\text{fear}}$ penalizes generating high-confidence outputs in low-information contexts. For instance, if the true label is "unknown" or "I don't know," but the model outputs a factual statement with high confidence, incur penalty.

- $\mathcal{L}_{\text{arrogance}}$ penalizes sustained overconfidence: sequences of tokens with probability above $\theta_{\text{arrogance}}$ in narrow semantic clusters lead to a rising arrogance index; penalize based on aggregate overconfidence and divergence from ground truth when available.

Hyperparameters $\lambda_{\text{fear}}, \lambda_{\text{arrogance}}$ control strength of emotion-based regularization. Similar ideas appear in uncertainty-aware training for calibration [5].

4.3 Feedback-Driven Weight Adaptation

During interactive use, feedback (user corrections or retrieved grounded evidence) triggers ARS responses:

- Lower arrogance index.
- Fine-tune model weights focusing on the contradicted semantic cluster (e.g., via targeted gradient updates).

This ongoing adaptation simulates human learning via emotional dissonance, potentially improving model alignment over time [4].

5 Evaluation Framework

We outline protocols to assess EAT benefits:

5.1 Hallucination Detection and Mitigation

- **Dataset of Known-Fact Queries**: queries whose correct answers are in a reference KB. Compare standard Transformer vs. EAT: measure hallucination rate (model outputs incorrect facts confidently). Expect EAT to hedge or terminate more appropriately [5].
- **Confidence Calibration Tests**: assess whether EAT's confidence aligns better with true correctness (lower overconfidence).
- **Retrieval-Augmented Baseline**: compare against retrieval-augmented generation (RAG) approaches to study complementary effects of external grounding vs. internal emotion regulation [6].

5.2 Overconfidence and Belief Updating

- **Semantic Cluster Tests**: artificially create narrow-topic sequences where base model is likely to overfit. Provide contradictory feedback and measure how quickly the model adapts (reduces confidence in the incorrect cluster). Compare EAT vs. base.
- **Continual Learning Scenarios**: monitor ARS signals over time; evaluate whether EAT reduces catastrophic forgetting while enabling correct updates from feedback.

5.3 Human Evaluation

- **Quality of Hedging and Clarifications**: when EAT triggers fear-based hedging, assess readability and user trust.
- **Adaptability in Dialogue**: in conversational settings, evaluate whether EAT appropriately admits uncertainty or corrects prior errors after feedback.

6 Discussion

Integrating emotion-inspired feedback addresses key deficiencies in pure statistical generation:

- *Truthfulness*: FES mechanism reduces confident hallucinations in low-information zones [3, 5].
- *Adaptability*: ARS-driven restructuring fosters dynamic updating upon contradiction, akin to human belief revision [7].
- *Memory Depth*: emotional salience can weight memory encoding; high-fear or high-arrogance correction events strengthen long-term adjustments.
- *Alignment*: by simulating hedging and humility, EAT aligns better with user expectations for transparency and uncertainty handling.

Challenges include: defining precise thresholds for emotion indices; designing emotion-labeled datasets; balancing emotion penalties to avoid excessive caution that hampers creativity. Moreover, one must ensure emotion modules do not introduce undesirable biases—ethical considerations in emotion-aware AI are critical [7].

7 Conclusion and Future Work

We introduced the Emotion-Augmented Transformer (EAT), embedding Fear and Arrogance Emotion Systems to regulate generation and learning. This architecture simulates human-like emotional feedback to mitigate hallucinations and encourage dynamic belief updates. Future directions:

- **Reinforcement Learning Integration**: combine with RLHF or RLAIF to refine emotion thresholds and behaviors based on user preferences.
- **Unsupervised Emotion Signal Discovery**: leverage self-supervised methods to detect low- vs. high-confidence contexts without explicit labels [4].
- **Emotional Trajectories**: model richer emotional states (e.g., from fear to relief or pride) influencing iterative generation and multi-turn dialogues.
- **Multimodal Extensions**: incorporate physiological or sentiment signals from user (where available) to adjust model emotion indices in interactive settings [2].
- **Rigorous Benchmarking**: develop standardized benchmarks for emotion-augmented generation evaluation.

By bridging affective computing insights with LLM architectures, EAT opens a path toward more reliable, adaptive, and human-aligned AI language systems.

References

[1] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[2] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[3] Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. Enhancing Emotional Generation Capability of Large Language Models via Emotional Chain-of-Thought. arXiv:2401.06836, 2024.

[4] Yuyan Chen and Yanghua Xiao. Recent Advancement of Emotion Cognition in Large Language Models. arXiv:2409.13354, 2024.

[5] Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On Large Language Models' Hallucination with Regard to Known Facts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers)*, pages 1041–1053, Mexico City, Mexico, June 2024.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. arXiv:2005.11401.

[7] Saif M. Mohammad. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2):239–278, June 2022.