Emotion Learning in Large Language Models

Amirali Famili

Student ID: 720060845/1

Abstract

Large Language Models (LLMs) are rapidly gaining popularity and are going to be an inevitable part of human daily life. LLMs are often referred to as black-boxes which means it is hard to interpret what the model has learned and how it responses to quarries which is increasingly important in sensitive tasks. Emotion detection from text is one of the developing areas of AI, improving the accuracy of emotional models is of great importance. Various explainability and interpretability techniques are designed to shed light on the inner workings of black-box models as a mean for improving their functionality through detailed explanations on their judgments. In this report, I present my final year project which involves using probing technique to gain insight on the emotion learning capabilities of the QWEN, a state-of-the-art LLM model. Doing so would reveal if the pre-trained model is capable of identifying emotions and in which layers is this emotion learning happening.

☐ **I have not used any GenAI tools in preparing this assessment.**

☑ **I certify that all material in this dissertation which is not my own has been identified.**

Signature: _____ _____

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing and made remarkable advancements as generative conversational agents. LLMs offer coherent communications and excellent ability to analyse and respond to queries in a vast variety of domains [1]. In particular, ML algorithms such as recommender systems are making these interactions more meaningful and personalized and thus LLMs are becoming a vital part of human-computer interaction [2]. Emotions play a crucial role in our every day communications, people express emotions through facial reactions, body language, speech signals and written text [3]. As LLMs are trained on large corpuses of text, they can potentially learn emotions expressed in texts, however the large scale of parameters and training data , makes these models complex and difficult to interpret [4]. In order to understand the workings of LLMs and to determine how and where emotion learning is happening in such models, I use two concepts of interpretability and explainability. Although these two concepts are often interchangeably used among researchers, I provide definitions for understanding their difference. Explainability refers to the techniques used to reveal the inner workings of black box models, by training other white box models or changes in input to determine the contribution of each component to our answer. In contrast, Interpretability refers to the techniques used to provide inner workings of the model, that take advantage of the components of a model directly [5, 6]. One of the effective explainability techniques in understanding how learning happens in LLMs is Classifier Probing (CP). CP techniques can provide insights into the structure, and progression of the emotional representations encoded at different layers of the model [7, 8].

In this paper, I review the literature that explores emotion learning from textual data and provide an overview of the workings of LLMs. Furthermore, I specify the outlines of my final year project where I will be trying to understand where emotion learning has happened in a pre-trained LLM model such as QWEN.

2 Emotion models in Psychology and Machine Learning

Sentiment Analysis is the most established area of NLP regarded to Emotion Detection, it involves detecting the polarity of a sentence and determining whether it holds a positive, negative or neutral value [9]. Emotion detection has numerous applications from monitoring student mental health [10] to enhancing human-computer interactions including question and answering in recommender systems and marketing [11]. Emotion detection from text can be challenging because emotions are highly personal and textual representations can convey multiple emotions simultaneously. Moreover, there are ambiguities in the linguistic usage of emotional terms when considering alternation of emotions in intra-cultural, cross-cultural and historical matters [9]. Many efforts have been made to mitigate these complexities, such as emotion detection in sarcasm [12]. In psychology there are two main beliefs which argue the root of emotions. Evolutionary psychology states that emotions are developed through typical responses of our ancestors to natural phenomena and difficulties in their lives. For example, when we are in danger and afraid, our heart increases pounding, our lungs inhale more rapidly and we can feel our muscles tensing, these sensations prepare us for either fight or flee [13]. On the other hand, critics of Evolutionary psychology argue that emotions are not separate and are the result of a single interconnected system in the brain. They Discuss that emotions are developed in society and are highly related to well-being [13]. Consequently, emotion models usually fall into either of the perspectives on emotions [9].

2.1 Categorical Emotion Models

Categorical models are derived from evolutionary psychology and are based on the idea that emotions can be classified into discrete categories [9]. Six core emotions including happiness, sadness, fear, anger, surprise, and disgust which were introduced by Ekman form the basis of emotion annotation and training categorical emotion models [14]. Some datasets allow for a more nuanced classification of

emotional states. As an example the GoEmotions dataset allow for more build on categorical models by providing 28 emotion labels including neutral [15]. Categorical models are effective in places where clear, discrete emotion labels can be used. However, they oversimplify emotional expression and are inherently limited in capturing complex, blended emotions [16].

2.2 Dimensional Emotion Models

Dimensional models offer a different approach, by representing emotions along continuous axes, as a single inter connected circumplex. For instance, Russell's circumplex model [17], defines emotions based on intensity of feeling (arousal) and pleasantness (valance). This approach is a more flexible representation of emotions and makes it possible to capture more complex emotions, such as bittersweet or envy. [16].

3 Large Language Models (LLMs)

Language models are auto-regressive in nature, that means that their task is to predict the next token. A language model's architecture is encoder-decoder or decoder type [4]. Here, we will focus on the more common encoder-decoder architecture, of which Qwen is based upon. Before the advent of transformers, language models were traditionally built using recurrent neural networks (RNNs) [18]. Whilst RNNs, such as Long-Short term memory (LSTMs) networks are capable of learning long range dependencies from the data, to learn multiple contexts from the same stream of data, complicated and deep network structures are required [19]. As I will demonstrate in this section, Transformers offer a unique approach in that they are both parrallelizable owing to multi-head attention and through stacking transformer layers they are capable of learning multiple contexts from the same stream of data. Furthermore, as highlighted by the seminal "Attention is all you need" [18] publication, the attention mechanism allows transformers to learn from arbitrarily long streams of data whilst scaling the complexity of the problem in a significantly less aggressive way than RNNs.
With the combination of the efficiency of transformers and large corpuses of text, language models transitioned into *Large* Language Models (LLMs) [18, 20]. To understand the role of which the pre-processing (tokenisation) and model architecture has on emotion learning in language models I sequentially describe these topics for LLMs.

3.1 Natural Language Processing and Tokenisation

The largest and most common datasets are harvested using web scraping from online sources such as the English Common Crawl dumps [21], and its subsidiaries (C4 [22], Refined Web [23]), The Pile [24], GitHub, Wikipedia and social media platforms.
Once the text from a dataset has been cleaned, parsed and normalised it can be tokenised. Tokenisation aims to capture the information of individual words, phrases and meaning via embedded representations [4]. Each LLM uses a different, customised tokeniser to generate word embeddings, some commmon examples include Word2Vec, GloVe, FastText and N-grames for generating static word embeddings [25, 4]. Some words contain multiple meanings and therefore require some context to establish their meaning, dynamic contextual embeddings can be used to generate a better representation of individual words where context matters, ChatGPT and Bert tokenisers use this approach. The second important element of tokenisation, specifically for transformer models is the addition of positional embeddings, which gives a unique representation of the position encoding of each token [18].

For QWEN, the tokenisation procedure uses a similar but updated and more efficient tokeniser to ChatGPT-3.5 and 4 called fast byte pair encoding (BPE). In addition as the general aim of QWEN is to provide text translations between English and Chinese and other multilingual downstream tasks, commonly used Chinese characters and words are added to the base vocabulary of the tokeniser [20]. Once the corpuses of text have been ingested via tokenisers, the dataset is ready to be fed into the model.

3.2 Transformers

Transformers, particularly their attention mechanisms, have demonstrated unrepresented breakthroughs in LLMs [18].

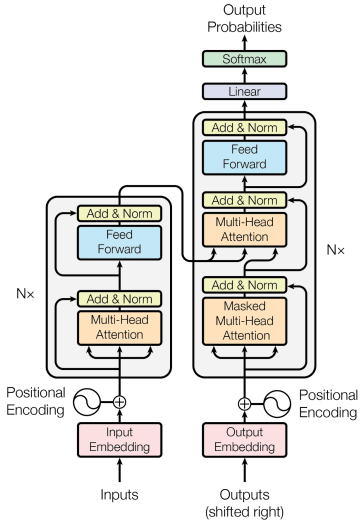

Figure 1: The Transformer model architecture [18].

Figure 1 shows the typical structure of a transformer. Once input tokens have been tokenised using the input embedding and positional encoding it is passed into the multi-headed attention mechanism. To understand the role of multi-headed attention, I first explain the role of a single attention block, known as self-attention.

Self-attention Mechanism captures correlation between elements with respect to their positions in the input sequence. Hence, allowing transformers capture contextual information. In self-attention mechanism the input sequence is mapped to Queries (Q), Keys (K), and Values (V). The Qs, Ks and Vs are derived from different linear projections of the same input using a randomly initialised vector of learnable weights. The Qs represent the current token in the input sequence, the Ks represent a target token from the input. The dot-product of the Q and K vector, which is then scaled ($\frac{1}{\sqrt{d_k}}$) according to the target output dimension (d_k). Passing this through a softmax function creates a vector of probabilities that is used to weight the Vs vector. This creates the output of the attention process which represents a relation between the words in the input sequence. This process can be formulated as,

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

In practice, a more effective method is to perform a number of self-attention layers in parallel.

Multi-head attention Mechanism is a variation of self-attention that allows the model to simultaneously attend to information from different representation subspaces. The Q, K, and V vectors are projected into h 'heads' using learnable linear transformations in multi-head attention. The self-attention operation is then performed on each of these projected subspaces in parallel, and the results are concatenated and projected again to produce the final output.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ is the output of the i-th attention head and the weights are learnable parameters. Since each head can focus on different aspects of the input and capture hidden representations, including emotional cues, contextual information, and syntactic structure, which is beneficial for interpreting complex or ambiguous emotions [26, 27, 18].

The following intuitive example, can explain how multi-head attention mechanism could enhance emotion learning in transformers. Consider the sentence "I am working hard on my report I think I am doing well but I am worried about my final grade". Each attention head would focus on a different contextual information such as the repeated use of the word "I", causality, the sentiment of each part as well as chunk phrasing. Each head independently performs the attention mechanism and the outputs are concatenated and undergo a linear transformation to produce one output which now can reflect more complex emotions.

3.3 QWEN

QWEN is a family of LLMs, designed with a slightly altered version of the transformer architecture used in LLaMa LLM, which forms the basis of this model[20]. These changes include, a united approach for the embeddings, using Rotary Positional Embedding (RoPE) for positional information, removing biases from most layers, utilizing pre-normalization with RMSNorm instead of traditional layer normalization, employing a combination of Swish and Gated Linear Unit called SwiGLU as the activation function, as well as reducing the number of neurons for each feed-forward layer [20].

# of Params	Hidden size	Heads	Layers	Learning rate	Batch size	Training tokens
1.8B	2048	16	24	3.0×10^{-4}	4M	2.2T
7B	4096	32	32	3.0×10^{-4}	4M	2.4T
14B	5120	40	40	3.0×10^{-4}	4M	3.0T

Table 1: Main QWEN Model's hyper-parameters setting [20].

QWEN was trained on a dataset built with up to 3 trillion tokens, including web documents, encyclopedia, books, and codes [20]. Making the model effective in linguistics, arithmetic, coding, and logical reasoning [20].
There are many variants of QWEN, including QWEN-PMP, QWEN-Chat, Code-QWEN, Math-QWEN-Chat [20], as well as different recently published versions, such as QWEN2 and it's variants and QWEN2.5-Coder [28, 29]. However, for the development of this project I only focus on the base LLM in the QWEN family.

4 Interpretability and Explainability of LLMs

Interpretability refers to the ability to explain the inner workings and internal decision making of an AI model, which is interpretable or understandable by humans [5]. Interpretable methods require access to the inner workings of a model, such as architecture, internal gradients and weights and therefore restrict the LLMs in this work to those which are open-source [6]. With the use of the internal gradients and weights, attention maps can be constructed to establish which part of the model input is being focused on for a particular decision [30].
Explainability is distinct from interpretability, as the term explainability is used to describe calculations which do not have access the the model and regard the model as a black-box. Explainability seeks to explain model decisions or output based on results from other models or inputs of the main model without referencing the internal workings, but rather in terms of the features of the data [5, 6].
LLMs are considered complex as it can be difficult to both interpret their inner workings and explain their output. This is highlighted by the efforts to understand their reasoning and to provide safe justification for resulted outputs. It is also of utmost importance, since they can produce harmful content, output sensitive training data, produce unfair (or biased) decisions, and misinformation [31]. Furthermore, providing such justification and reasoning for the output of these models can improve user's confidence in these models [30].

4.1 Interpretability Methods

Layer-wise Relevance Propagation (LRP) works by assigning a relevance score to each neuron in all layers, and trace the conserved relevance of all neuron from output back to the input, layer by layer, similar to back propagation process [32]. In LRP, the sum of the relevance scores from each layer should be equal to the prediction to track the most contributing neurons [32].
Neuron Activation Interpretations use neuron analysis techniques, to explore the neurons that are activated given an input, which highlights the most relevant parameters to the output [30].
Concept-Based Interpretations methods map the inputs of a model to a set of pre-defined concepts, and measure how the relevance of correlation between the input, concepts and components of that model affect the output. After selecting the most important concepts to the model, we can generate human understandable interpretations from those concepts. Defining relevant concepts remains a challenging aspect of concept based interpretations [30].

4.2 Explainability Methods

Attention Mechanism Explanations map relevant parts of input to it self, these relevance scores captured can be used to further examine the behaviour of the model. Techniques such as visualization methods and probing-based methods rely solely on attention mechanism [30].
Example-Based Explanations attempt to measure changes in output with different inputs, unlike techniques such as **Perturbation-Based Explanations**, where they do not manipulate or perturb inputs to keep track of changes in the output [30].
Natural Language Explanations take advantage of Natural Language Generation (NLG) techniques to explain the model's decision making and a step by step reasoning in generated texts [30]. Chat-GPT is a valid example of such models.
Surrogate models use various white box models such as decision trees, rule based systems and linear models, to evaluate the predictions of a black-box model, in comprehensible representations [30]. LIME (Local Interpretable Model-agnostic Explanations) is a well-known surrogate model which a technique used to explain the predictions of any classifier, by approximating the model's behaviour around specific inputs that require explanations, using interpretable models (or white box models). The interpretable model will then provide insight on most influential features led to the output of our specified input [33].
Probing-Based Explanations and Interpretations A probe is a classifier model, trained on the learned representation of the model, in order to capture knowledge from intermediate layers of that model [7, 30]. The trained classifier would then perform classification tasks on target data such as emotional datasets [8, 30]. If the probe performs well on the dataset, it indicates that the black-box model have captured the representations described in the dataset [8]. Similar works on neural networks have demonstrated that lower layers are responsible for capturing word-level syntax, while deeper layers can predict sentence-level syntax and semantic representations [30]. There are mainly two types of probing based explanations, linear and non-linear probing, both of which are considered to be explainability techniques.
Linear (or structural) probing requires simpler classifier models, to reveal if the features captured in black-box model representations (i.e. weights) are linearly separable [34]. Non-linear probing requires more complex classifier models, to capture non-linear relationships encoded in the representations [35]. Google researcher Ian Tenney along with his research partners in the paper [36] have shown which layers of BERT are responsible for capturing syntactic and semantic representations, using a technique called edge probing, which places restrictions on the classifier probe.
However, there are doubts on the efficiency of the probing methods [8]. By using simple classifier probes, explaining their results becomes more clear, but the model may fail to capture more complex (i.e. non-linear or unstructured) representations [7]. On the other hand, by using complex classifier probes we can capture high-dimensional representations of the black-box model, at the cost of expressiveness of the probe [7].

Parameter-Free Probing on the other hand, is categorised as an interpretability technique, because it advocates the need for a classifier probe. It is based on the idea of interpretability, that the model alone is perfectly capable of expressing it self, by analysing the representations captured by the neural network as they are [30]. For example, in reference [37] Facebook researchers, treat each attention head as a classifier probe to demonstrate the synthetic relations learned by each head. Another example can be observed from reference [38], where Rebecca Marvin and Tal Linzen designed a dataset with pairs of correct and incorrect sentences to test the grammaticality of the predictions, by measuring the difference in probabilities assigned to correct and incorrect pairs, they were able to interpret the judgment of the model with respect to the grammar of the sentences [30].

4.3 Problem With Explaining Black-Box models

It is argued that explanations provided by these methods are usually unreliable, ambiguous, and sometimes wrong. For many prediction algorithms that deeply affect human lives, such as criminal justice or healthcare. It may be unethical to rely on the explanations provided by these models. As a result, for human oriented fields, it's best to use inherently interpretable (or white box) machine learning techniques such as linear regression, decision trees, and rule-based systems [39].

5 Project Specification

The main focus of this project is to determine whether pre-trained LLMs, specifically QWEN, encode information about emotions in their learnt representation during their training process and to determine the specific layers of the model where emotion learning is happening. I will achieve this objective by employing probe classifiers to examine which specific layers of a large model have encoded the emotional content. This involves choosing and preprocessing datasets, designing and implementing probe classifiers and placing them after each layer, giving an input to the QWEN model and feeding the representation form hidden layers during forward propagation, evaluation and analysis of the performance of classifiers on each dataset and finally interpretation of the model by analysis of the results.

The first step is to choose benchmark datasets which have labelling for emotion. The choice of datasets for this project is discussed in section 6.1 at length. The chosen dataset then needs to be preprocessed. The preprocessing involves tokenisation of the dataset with the same tokeniser that was used for the pre-trained model we are aiming to analyse (in this case QWEN). Next I will input the preprocessed input data through the pre-trained QWEN model and extract representations from each layer and save both token-level and overall representations for further analysis. I will then design and implement lightweight probe classifier which I will train on the extracted representations in order to evaluate their ability to classify emotion by comparing to the labelled data. The design of probe classifiers will initially involve a simple feed-forward linear layer with a softmax activation function and according to need can be further developed to include intermediate layers if the task proves to be more complicated. Metrics such as accuracy and F1-score as well as confusion matrices will be used for the analysis.

To strengthen the argument provided in my study, I will perform ablation studies where I remove certain keywords or emotions to identify their impact on the layer-wise learnt representations. Furthermore, if it is feasible within the time constraints of the project, I would test different datasets to compare the trained probes and use other interpretability and explainability techniques to validate my findings. Afterwards, it would be interesting to repeat this process on different models such as BERT, to capture connections between responsible layers for emotional representations across different models.

This project can improve understanding of emotion-related representations in LLMs which can be impactful in improving the accuracy and fairness in emotion detection. More specific examples of areas the outcome of this project can have great impact on are in transfer learning, where a pre-trained

model is fine-tuned on a dataset, early layers responsible for general NLP tasks are frozen which means the emotional encodings captured from training are preserved. This means in order to fine-tune the relevant parts of a model, detailed knowledge of where learning happens would be extremely valuable. Furthermore, the outcome of this project can pave the path for for the development of explainable LLMs through enhancing their interpretability and emotional models by improving their accuracy.

5.1 Potential challenges

One of the main challenges in this model is designing and implementation of custom probes for placing after each layer of the QWEN model. This requires deep understanding of the intermediate layers of the QWEN model and probing method.
Another challenging aspect of this project could be that emotions are generally hard to capture for LLMs and by using a simple linear probe I might fail to capture encoded emotions from the model. On the other hand, with a complex non-linear probe, I might fail to validate their results on the datasets. Achieving balance between complexity and simplicity of the probes might take more time than planned, since it's a trial and error process.
Although this project is result centric, I intend on trying different approaches to explain and interpret the QWEN model. In the worst case, I can conclude with certainty that QWEN model has not learned any emotional representations from it's training, and move on to another model such as BERT and follow the steps I have taken with QWEN.

5.2 Timeline

The timeline of achievable in this project is demonstrated in the table below.

November	December	January	February	March	April
Define project scope and research objectives. Conduct literature review.				Prepare for final presentation Prepare for final submission	
pre-processing the datasets gaining knowledge on the internal workings of QWEN researching on various probe architectures				Prepare project demonstration	
	Initial design and implementation of probes		Finalising results Creating final visualisations for the demonstration		
		Training and testing probes Addressing potential challenges Adjusting the design of the probes according to results --------- If time permits: Experiments with different explainability tools, datasets and models. potentially investigate other LLM models			

6 Requirements

The required resources for actualising this project are described in this section.

6.1 Datasets

The datasets that I will be using in this study are GoEmotion and ISEAR. I have chosen these two datasets because they span a diverse range of emotions. I expect these two datasets to be sufficient to demonstrate the objective of the project. However, I have researched 2 other datasets EmpatheticDialogues and EmoBank which can be used as material if the first two datasets were not providing conclusive results.

GoEmotion dataset is developed by google researchers. It consists of 58,000 samples with 28 different emotions including neutral, and it consists of Reddit comments, emotion label and the ID of the Reddit user. The dataset was manually annotated by 82 native English speakers who were given pre-defined definitions of each emotion. Ambiguous emotions (mostly expressed positively) were labelled as neutral [15].

GoEmotion is a widely recognised and used dataset in the field of emotion detection from text and has proven to be reliable for training LLMs [40].

ISEAR (International Survey on Emotion Antecedents and Reactions) dataset was collected by a large group of psychologists from 37 countries across 5 continents, in which people were asked to describe the situations where they experienced joy, fear, anger, sadness, disgust, shame, and guilt. ISEAR dataset is comprised of 7666 samples with 2 columns one for emotion and the other for the response in text, with 1096 participants from various backgrounds, such as psychology, social science law, engineering and medical [41]. Whilst ISEAR dataset is smaller in size in comparison to GoEmotion, since it focuses on the more basic emotions. I believe this would be complementary to GoEmotion.

EmpatheticDialogues dataset is developed by Facebook researchers, with 810 participants. It consists of 25000 samples as situations described in text, based on 32 emotions, introduced in the paper [42]. EmpatheticDialogues is comprised of 4 columns: conv_id, keeps track of the conversation sequence; utterance_idx, keeps track of the repeats in conversation sample (since each dialogue is repeated several times); context, is the emotional label with 32 different emotions; prompt, is the actual conversation's text. Another interesting dataset is **EmoBank**, it consists of 10000 samples and is annotated for dimensional models with 3-dimension as Valance, Arousal and Dominance (VAD) representation structure [43].

I intend to pre-process the text labels of the final datasets, by removing unnecessary punctuations and checking spelling mistakes. Techniques such as bag-of-words could be used to gain insight on the textual structure of these datasets, to find emotionally charged words.

6.2 Models and tools

The main model used for this project is QWEN, however, It would also be interesting to perform the same experiments on the more recently published QWEN2 model [28]. If achievable within the time constraints of the project, I can consider studying models such as BERT [27] for capturing the differences and similarities between older and newer transformers in recognizing emotional representations from their training, regardless of the data they are trained with, since I mainly intend on identifying the layers that emotion learning is happening.

The main explanation technique used for this project is classifier probing technique, which can provide valid explanations on the internal workings of the QWEN. In addition it would be interesting to capture the behaviour of model using other explainability methods discussed in this paper as well. A broad knowledge of Python libraries is required to achieve the goals in this project. Libraries such as pandas, numpy and scikit-learn will be used for pre-processing and data management. Hugging Face Transformers library will be used for loading the pre-trained model. The PyTorch library would be used for implementing probe classifier, optimisation and training routines. Furthermore, a range of libraries including matplotlib, seaborn and plotly will be essential for analysis and visualisations. Other libraries such as lime could be used to improve understanding and explanations.

7 Summary

In this literature review, I described the psychology behind emotions and their importance for human-computer interaction. I have raised the issue of black-box nature of LLMs and the need for understanding how emotion learning happens in these increasingly widely used algorithms. The objectives of this project are set to be feasible with a clear timeline and very well planned steps to ensure success. I have chosen state-of-the-art model QWEN and very well regarded labeled datasets to conduct my research and the method I have selected (probe classifier) is widely used and well established. The findings of this project can contribute towards eliminating biases from LLM models and improve the performance of application of LLMs in emotion related areas such as mental health, customer service, and social media monitoring. Therefore I believe this project will be impactful in the field.

References

[1] P. Kumar, "Large language models (llms): survey, technical frameworks, and future challenges," *Artificial Intelligence Review*, vol. 57, no. 10, p. 260, 2024.

[2] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang *et al.*, "Recommender systems in the era of large language models (llms)," *arXiv preprint arXiv:2307.02046*, 2023.

[3] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[6] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv preprint arXiv:2012.01805*, 2020.

[7] A. K. Rege, "The probe paradigm: A theoretical foundation for explaining generative models," in *Neurips Safe Generative AI Workshop 2024*.

[8] A. A. Ivanova, J. Hewitt, and N. Zaslavsky, "Probing artificial neural networks: insights from neuroscience," *arXiv preprint arXiv:2104.08197*, 2021.

[9] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion detection in text: a review," *arXiv preprint arXiv:1806.00674*, 2018.

[10] M. Bustos-López, N. Cruz-Ramírez, A. Guerra-Hernández, L. N. Sánchez-Morales, and G. Alor-Hernández, "Emotion detection from text in learning environments: a review," *New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques*, pp. 483–508, 2021.

[11] A. Kumar, A. K. Bhartee, H. Gupta, and P. K. Singh, "A review on emotion detection of contextual text," in *2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 2023, pp. 465–471.

[12] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, "A multimodal corpus for emotion recognition in sarcasm," *arXiv preprint arXiv:2206.02119*, 2022.

[13] J. Prinz, "Which emotions are basic," *Emotion, evolution, and rationality*, vol. 69, p. 88, 2004.

[14] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[15] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020. [Online]. Available: https://arxiv.org/abs/2005.00547

[16] D. Alkurdi, S. Rasouli, W. J. Talamonti, M. Nasir, and A. Alsaid, "From categorical to dimensional: A multifaceted approach to emotions," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA: Los Angeles, CA, 2024, p. 10711813241262446.

[17] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[18] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[19] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *arXiv preprint arXiv:1801.01078*, 2017.

[20] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[21] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave, "CCNet: Extracting high quality monolingual datasets from web crawl data," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4003–4012. [Online]. Available: https://aclanthology.org/2020.lrec-1.494

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[23] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023.

[24] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima *et al.*, "The pile: An 800gb dataset of diverse text for language modeling," *arXiv preprint arXiv:2101.00027*, 2020.

[25] A. Nayak, H. Timmapathini, K. Ponnalagu, and V. G. Venkoparao, "Domain adaptation challenges of bert in tokenization and sub-word representations of out-of-vocabulary words," in *Proceedings of the first workshop on insights from negative results in NLP*, 2020, pp. 1–5.

[26] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[27] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[28] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[29] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.

[30] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[32] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification," *Frontiers in aging neuroscience*, vol. 11, p. 456892, 2019.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[34] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4129–4138.

[35] J. C. White, T. Pimentel, N. Saphra, and R. Cotterell, "A non-linear structural probe," *arXiv preprint arXiv:2105.10185*, 2021.

[36] J. E. Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–31, 2022.

[37] K. Clark, "What does bert look at? an analysis of bert's attention," *arXiv preprint arXiv:1906.04341*, 2019.

[38] R. Marvin, "Targeted syntactic evaluation of language models," *arXiv preprint arXiv:1808.09031*, 2018.

[39] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[40] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5487–5496.

[41] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, vol. 66, no. 2, p. 310, 1994.

[42] H. Rashkin, "Towards empathetic open-domain conversation models: A new benchmark and dataset," *arXiv preprint arXiv:1811.00207*, 2018.

[43] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," *arXiv preprint arXiv:2205.01996*, 2022.