## Thermodynamic Visualization Method for Interpreting Attention Mechanisms in Transformer-Based Language Models

### Abstract

This disclosure presents a novel method for enhancing the interpretability of the Multi-Head Self-Attention Mechanism in transformer-based language models by employing thermodynamic principles. The approach involves mapping attention weights to a thermodynamic model, enabling the identification of critical points where significant changes in attention distribution occur. By treating attention weights as analogous to a physical system undergoing phase transitions, we define a thermodynamic potential function, similar to free energy, to analyze these weights. The method computes the first and second derivatives of this potential function with respect to a control parameter, such as the temperature in a softmax function, to pinpoint critical points. A visualization tool is developed to display a heatmap of attention weights and overlay critical points, highlighting the most influential tokens in the input sequence. This integration into the transformer architecture provides a deeper understanding of the attention mechanism, offering insights into the model’s decision-making process and enhancing interpretability.

### 1. Field of the Invention

This invention pertains to the field of machine learning, specifically to the interpretability of transformer-based language models. It leverages thermodynamic principles to enhance the understanding of the attention mechanism within these models. The invention is particularly relevant to natural language processing (NLP) applications, where interpretability and transparency of model decisions are crucial.

### 2. Background of the Invention

The interpretability of large language models (LLMs) such as GPT-3 and BERT has been a significant challenge in the field of natural language processing (NLP). These models, which rely on the Transformer architecture, utilize the Multi-Head Self-Attention Mechanism to weigh the importance of different words in a sequence, capturing complex dependencies and contextual information. Despite their success, understanding why these models make certain decisions remains difficult.

Several approaches have been proposed to improve the interpretability of attention mechanisms. Tools like BertViz and the Language Interpretability Tool (LIT) provide visualizations of attention weights, helping users understand how attention is distributed across the input sequence. These tools, however, often lack a deeper theoretical foundation that could explain why certain attention patterns emerge [[1]](https://www.semanticscholar.org/paper/beb051c652f02c2d5829d783fbc4f3acce99bc3c), [[2]](https://www.aclweb.org/anthology/2020.emnlp-demos.15.pdf).

Thermodynamic phase transitions, a concept from statistical physics, offer a promising framework for understanding these patterns. In thermodynamics, phase transitions describe changes in the state of a system, such as from liquid to gas, which occur at critical points where small changes in conditions can lead to significant shifts in behavior. This concept has been applied to various fields, including neural networks, to study phenomena like learning dynamics and generalization [[3]](https://www.semanticscholar.org/paper/f70bc6661c774ee45bb773e995c1a5351856db5b).

By mapping attention weights to a thermodynamic model, we can identify critical points where significant changes in attention distribution occur. This approach can highlight which parts of the input sequence are most influential, thereby making the attention mechanism more interpretable. Existing work on phase transitions in neural networks and attention mechanisms provides a foundation for this idea [[4]](https://arxiv.org/pdf/1706.03762).

Recent literature has explored various methods to enhance the interpretability of LLMs. For instance, the Proto-lm framework introduces a prototypical network-based approach for built-in interpretability during the fine-tuning stage, maintaining competitive performance while providing immediate interpretability [[5]](https://arxiv.org/pdf/2311.01732). Another study, RecExplainer, aligns LLMs with recommendation models to improve interpretability through behavior and intention alignment [[6]](https://www.semanticscholar.org/paper/d357a811d8ee0e2537c2987564df5c13d59ec02d).

Despite these advancements, current methods often focus on post hoc interpretability, applied after inference, and may not provide a comprehensive theoretical explanation for the observed attention patterns. Additionally, they may lack the ability to pinpoint critical points where significant changes in attention distribution occur, which are crucial for understanding the model’s decision-making process.

The proposed method addresses these limitations by integrating thermodynamic principles into the attention mechanism. By treating attention weights as analogous to a physical system undergoing phase transitions, we can define a thermodynamic potential function to analyze these weights. This approach allows us to compute the first and second derivatives of this potential function with respect to a control parameter, such as the temperature in a softmax function, to pinpoint critical points. The resulting visualization tool not only displays a heatmap of attention weights but also overlays critical points, highlighting the most influential tokens in the input sequence.

### 3. Description of the Invention

#### 3.1 Overview of the Invention

The invention introduces a novel method to enhance the interpretability of the Multi-Head Self-Attention Mechanism in transformer-based language models by leveraging thermodynamic principles. The core idea is to map attention weights to a thermodynamic model, enabling the identification of critical points where significant changes in attention distribution occur. This approach treats attention weights as analogous to a physical system undergoing phase transitions, allowing us to define a thermodynamic potential function, similar to free energy, to analyze these weights.

The method involves computing the first and second derivatives of this potential function with respect to a control parameter, such as the temperature in a softmax function, to pinpoint critical points. A visualization tool is developed to display a heatmap of attention weights and overlay critical points, highlighting the most influential tokens in the input sequence. This integration into the transformer architecture provides a deeper understanding of the attention mechanism, offering insights into the model’s decision-making process and enhancing interpretability.

The significance of this invention lies in its ability to provide a theoretical foundation for understanding attention patterns in transformer-based models. By identifying critical points in the attention distribution, the method highlights the most influential parts of the input sequence, making the attention mechanism more interpretable. This is particularly relevant for natural language processing (NLP) applications, where interpretability and transparency of model decisions are crucial.

#### 3.2 System Description

The proposed system consists of several key modules that work together to enhance the interpretability of the Multi-Head Self-Attention Mechanism in transformer-based language models. These modules include:

1. **Attention Weight Calculation Module**: This module computes the attention weights $a\_{ij}$ for a given head in the Multi-Head Self-Attention Mechanism using the scaled dot-product attention mechanism.
2. **Thermodynamic Mapping Module**: This module maps the attention weights to a thermodynamic potential function $Φ$, which is defined as:

$$Φ=-\sum\_{i,j}^{​}a\_{ij}loga\_{ij}$$

* Here, $a\_{ij}$ represents the attention weight from the $i$-th token to the $j$-th token in the input sequence.
1. **Critical Point Identification Module**: This module computes the first and second derivatives of $Φ$ with respect to a control parameter $λ$, such as the temperature in a softmax function, to identify critical points. The first derivative, $\frac{∂Φ}{∂λ}$, indicates the rate of change of the potential function, while the second derivative, $\frac{∂^{2}Φ}{∂λ^{2}}$, helps identify points of inflection, which correspond to phase transitions.
2. **Visualization Tool**: This tool generates a heatmap of the attention weights and overlays the critical points identified through the thermodynamic analysis. The heatmap allows users to see the distribution of attention across the input sequence, while the critical points highlight the most influential tokens.

The integration of these modules into the conventional transformer architecture enhances the interpretability of the attention mechanism by providing insights into which parts of the input sequence are most influential in the model’s decision-making process.

#### 3.3 Detailed Description of the Invention

The detailed description of the invention involves a comprehensive explanation of each module and the processes involved in enhancing the interpretability of the Multi-Head Self-Attention Mechanism.

**Attention Weight Calculation Module**

The attention weights $a\_{ij}$ are computed using the scaled dot-product attention mechanism. Given a set of queries $Q$, keys $K$, and values $V$, the attention weights are calculated as follows:

$$a\_{ij}=softmax\left(\frac{QK^{T}}{\sqrt{d\_{k}}}\right)$$

where $d\_{k}$ is the dimensionality of the keys. The softmax function ensures that the attention weights sum to one, providing a probability distribution over the input tokens.

**Thermodynamic Mapping Module**

The attention weights $a\_{ij}$ are mapped to a thermodynamic potential function $Φ$, which is defined as:

$$Φ=-\sum\_{i,j}^{​}a\_{ij}loga\_{ij}$$

In this context, $Φ$ is analogous to the free energy in thermodynamics, where $a\_{ij}$ can be interpreted as the probability distribution of attention weights. The negative sign ensures that higher attention weights contribute more significantly to the potential function.

**Critical Point Identification Module**

To identify critical points, we compute the first and second derivatives of $Φ$ with respect to a control parameter $λ$, such as the temperature in a softmax function. The first derivative is given by:

$$\frac{∂Φ}{∂λ}=-\sum\_{i,j}^{​}\left(\frac{∂a\_{ij}}{∂λ}loga\_{ij}+\frac{∂a\_{ij}}{∂λ}\right)$$

This derivative indicates the rate of change of the potential function with respect to $λ$.

The second derivative is given by:

$$\frac{∂^{2}Φ}{∂λ^{2}}=-\sum\_{i,j}^{​}\left(\frac{∂^{2}a\_{ij}}{∂λ^{2}}loga\_{ij}+2\frac{∂a\_{ij}}{∂λ}\frac{∂loga\_{ij}}{∂λ}+\frac{∂^{2}a\_{ij}}{∂λ^{2}}\right)$$

This derivative helps identify points of inflection, which correspond to phase transitions in the attention distribution.

By analyzing these derivatives, we can pinpoint the values of $λ$ where significant changes in the attention distribution occur, thus identifying the critical points.

**Visualization Tool**

The visualization tool generates a heatmap of the attention weights and overlays the critical points identified through the thermodynamic analysis. The heatmap allows users to see the distribution of attention across the input sequence, while the critical points highlight the most influential tokens. This visualization provides a deeper understanding of the attention mechanism, offering insights into the model’s decision-making process and enhancing interpretability.

#### 3.4 Showcases

To illustrate the effectiveness of the proposed method, we present two examples: sentiment analysis using BERT and machine translation using a Transformer model.

**Example 1: Sentiment Analysis using BERT**

**Input** - Task: Sentiment Analysis - Model: BERT - Input Sentence: “The movie was absolutely fantastic and I loved every moment of it.”

**Process** 1. **Attention Weight Calculation**: Compute the attention weights $a\_{ij}$ for the input sentence using BERT’s attention mechanism. 2. **Thermodynamic Mapping**: Map the attention weights to the thermodynamic potential function $Φ$. 3. **Critical Point Identification**: Compute the first and second derivatives of $Φ$ with respect to a control parameter $λ$ to identify critical points. 4. **Visualization**: Generate a heatmap of the attention weights and overlay the critical points.

**Output** - Heatmap: Shows the distribution of attention weights across the input sentence. - Critical Points: Highlighted tokens such as “fantastic” and “loved” where significant changes in attention distribution occur.

**Significance** - Input Problem: The input sentence contains multiple positive words, but it’s unclear which words are most influential in determining the sentiment. - Output Solution: The thermodynamic visualization identifies “fantastic” and “loved” as critical points, providing clear interpretability and showing that these words are most influential in the model’s positive sentiment prediction.

**Example 2: Machine Translation using a Transformer Model**

**Input** - Task: Machine Translation (English to French) - Model: Transformer - Input Sentence: “The weather is nice today.”

**Process** 1. **Attention Weight Calculation**: Compute the attention weights $a\_{ij}$ for the input sentence using the Transformer’s attention mechanism. 2. **Thermodynamic Mapping**: Map the attention weights to the thermodynamic potential function $Φ$. 3. **Critical Point Identification**: Compute the first and second derivatives of $Φ$ with respect to a control parameter $λ$ to identify critical points. 4. **Visualization**: Generate a heatmap of the attention weights and overlay the critical points.

**Output** - Heatmap: Shows the distribution of attention weights across the input sentence. - Critical Points: Highlighted tokens such as “weather” and “nice” where significant changes in attention distribution occur.

**Significance** - Input Problem: The input sentence contains multiple words, but it’s unclear which words are most influential in the translation process. - Output Solution: The thermodynamic visualization identifies “weather” and “nice” as critical points, providing clear interpretability and showing that these words are most influential in the model’s translation process.

#### 3.5 Novelty of the Invention

The novelty of this invention lies in the integration of thermodynamic principles into the attention mechanism of transformer-based language models. The key innovative components and processes introduced by this invention include:

1. **Thermodynamic Mapping of Attention Weights**: The method maps attention weights to a thermodynamic potential function $Φ$, which is analogous to the free energy in thermodynamics. This mapping provides a theoretical foundation for analyzing attention weights, which is not present in existing methods.
2. **Identification of Critical Points**: By computing the first and second derivatives of the thermodynamic potential function $Φ$ with respect to a control parameter $λ$, the method identifies critical points where significant changes in attention distribution occur. This process highlights the most influential tokens in the input sequence, enhancing interpretability.
3. **Visualization Tool**: The development of a visualization tool that generates a heatmap of attention weights and overlays critical points is a novel contribution. This tool provides a clear and intuitive way to understand the distribution of attention and the most influential tokens, offering insights into the model’s decision-making process.
4. **Integration with Transformer Architecture**: The integration of the thermodynamic visualization method into the conventional transformer architecture is a unique aspect of this invention. This integration enhances the interpretability of the attention mechanism without compromising the performance of the model.

### 4. Description of the Drawings

**Figure 1**: Transformer-Based Language Model with Thermodynamic Visualization

 

####

#### Description of Figure 1

* **Input Embedding Layer**: This subgraph represents the initial embedding process where input tokens are converted into embeddings.
	+ **Token Embeddings**: Converts input tokens into dense vectors.
	+ **Segment Embeddings**: Adds segment information to the token embeddings.
	+ **Positional Embeddings**: Adds positional information to the embeddings to capture the order of tokens.
* **Positional Encoding**: Adds positional information to the embeddings to capture the order of tokens.
* **Encoder Stack**: This subgraph represents the encoder part of the transformer model.
	+ **Multi-Head Self-Attention Mechanism**: Computes attention weights for each token in the input sequence.
	+ **Feed-Forward Neural Network**: Processes the attention data to extract features.
	+ **Layer Normalization**: Normalizes the processed data.
	+ **Residual Connections**: Adds the original input to the normalized data to preserve information.
* **Decoder Stack**: This subgraph represents the decoder part of the transformer model.
	+ **Multi-Head Self-Attention Mechanism**: Computes attention weights for each token in the input sequence.
	+ **Encoder-Decoder Attention Mechanism**: Computes attention weights between the encoder and decoder sequences.
	+ **Feed-Forward Neural Network**: Processes the attention data to extract features.
	+ **Layer Normalization**: Normalizes the processed data.
	+ **Residual Connections**: Adds the original input to the normalized data to preserve information.
* **Output Layer**: This subgraph represents the final layer of the transformer model.
	+ **Linear Transformation**: Transforms the processed data into logits.
	+ **Softmax Function**: Converts logits into probabilities.
* **Thermodynamic Visualization**: This subgraph represents the novel thermodynamic visualization module.
	+ **Thermodynamic Mapping**: Maps attention weights to a thermodynamic potential function.
	+ **Critical Point Identification**: Identifies critical points in the attention distribution.
	+ **Visualization Tool**: Generates a heatmap of attention weights and overlays critical points.
* **Novelty**: The Thermodynamic Visualization subgraph introduces a new process for mapping attention weights to a thermodynamic model, identifying critical points, and visualizing these points to enhance interpretability.

**Figure 2**: Thermodynamic Visualization Module

 

#### Description of Figure 2

* **Thermodynamic Visualization**: This subgraph represents the detailed components of the thermodynamic visualization module.
	+ **Thermodynamic Mapping**: Maps attention weights to a thermodynamic potential function.
	+ **Critical Point Identification**: Identifies critical points in the attention distribution by computing the first and second derivatives of the potential function.
	+ **Visualization Tool**: Generates a heatmap of attention weights and overlays critical points to highlight the most influential tokens.
* **Novelty**: The Thermodynamic Visualization subgraph introduces a new process for mapping attention weights to a thermodynamic model, identifying critical points, and visualizing these points to enhance interpretability.

### 5. Embodiments

The proposed invention can be implemented in various configurations and applications to enhance the interpretability of attention mechanisms in transformer-based language models. Below are specific embodiments illustrating different ways the invention can be deployed:

#### 5.1 Embodiment 1: Integration with Existing NLP Models

In this embodiment, the thermodynamic visualization method is integrated into existing transformer-based NLP models, such as BERT, GPT, or T5. The integration involves modifying the attention mechanism to include the thermodynamic mapping and critical point identification modules. This allows users to gain insights into the model’s decision-making process without altering the core architecture or performance of the model.

* **Configuration**: The thermodynamic modules are added as an auxiliary component to the existing attention mechanism.
* **Deployment**: This embodiment can be deployed in any NLP application where interpretability is crucial, such as sentiment analysis, question answering, or text summarization.

#### 5.2 Embodiment 2: Standalone Visualization Tool

This embodiment involves developing a standalone visualization tool that can be used independently of the transformer model. The tool takes pre-computed attention weights as input and performs thermodynamic mapping and critical point identification to generate visualizations.

* **Configuration**: The tool is designed as a web-based application or a desktop software that accepts attention weights in a standard format.
* **Deployment**: This embodiment is suitable for researchers and practitioners who wish to analyze attention patterns in various models without modifying the model architecture.

#### 5.3 Embodiment 3: Real-Time Analysis in Interactive Systems

In this embodiment, the thermodynamic visualization method is integrated into interactive systems, such as chatbots or virtual assistants, to provide real-time analysis of attention mechanisms. The system dynamically computes attention weights and visualizes critical points as the user interacts with the model.

* **Configuration**: The thermodynamic modules are embedded within the interactive system’s backend, allowing for real-time computation and visualization.
* **Deployment**: This embodiment is ideal for applications where immediate feedback and interpretability are essential, such as customer support or educational tools.

#### 5.4 Embodiment 4: Educational and Training Platforms

This embodiment focuses on using the thermodynamic visualization method as an educational tool to teach students and practitioners about attention mechanisms in transformer models. The visualization tool can be integrated into online courses or workshops to provide hands-on experience with interpreting attention patterns.

* **Configuration**: The tool is integrated into learning management systems or online platforms, providing interactive tutorials and exercises.
* **Deployment**: This embodiment is suitable for educational institutions, online learning platforms, and workshops focused on machine learning and NLP.

#### 5.5 Embodiment 5: Customizable Visualization Options

In this embodiment, the visualization tool is enhanced with customizable options, allowing users to adjust parameters such as the control parameter $λ$, color schemes, and the level of detail in the heatmap. This customization enables users to tailor the visualization to their specific needs and preferences.

* **Configuration**: The tool includes a user-friendly interface with sliders and dropdown menus for customization.
* **Deployment**: This embodiment is applicable in research settings where detailed analysis and customization are required to explore different aspects of attention mechanisms.

These embodiments demonstrate the versatility and applicability of the thermodynamic visualization method across various domains and use cases, enhancing the interpretability of attention mechanisms in transformer-based language models.

### 6. Claims

1. **A method for enhancing the interpretability of the Multi-Head Self-Attention Mechanism in transformer-based language models**, comprising:
	* Mapping attention weights to a thermodynamic potential function to analyze the distribution of attention weights.
	* Identifying critical points in the attention distribution by computing the first and second derivatives of the thermodynamic potential function with respect to a control parameter.
	* Generating a visualization that displays a heatmap of attention weights and overlays the identified critical points to highlight the most influential tokens in the input sequence.
2. **A system for thermodynamic visualization of attention mechanisms in transformer-based language models**, comprising:
	* An attention weight calculation module that computes attention weights for a given head in the Multi-Head Self-Attention Mechanism.
	* A thermodynamic mapping module that maps the attention weights to a thermodynamic potential function.
	* A critical point identification module that computes the first and second derivatives of the thermodynamic potential function to identify critical points.
	* A visualization tool that generates a heatmap of attention weights and overlays the critical points.
3. **A computer-readable medium storing instructions that, when executed by a processor, cause the processor to perform a method for enhancing the interpretability of attention mechanisms in transformer-based language models**, the method comprising:
	* Computing attention weights for a given head in the Multi-Head Self-Attention Mechanism.
	* Mapping the attention weights to a thermodynamic potential function.
	* Identifying critical points in the attention distribution by computing the first and second derivatives of the thermodynamic potential function.
	* Generating a visualization that displays a heatmap of attention weights and overlays the identified critical points.
4. **A visualization tool for interpreting attention mechanisms in transformer-based language models**, comprising:
	* A module for mapping attention weights to a thermodynamic potential function.
	* A module for identifying critical points in the attention distribution by computing the first and second derivatives of the thermodynamic potential function.
	* A module for generating a heatmap of attention weights and overlaying the identified critical points to highlight the most influential tokens.
5. **A method for identifying influential tokens in an input sequence processed by a transformer-based language model**, comprising:
	* Calculating attention weights for tokens in the input sequence using the Multi-Head Self-Attention Mechanism.
	* Mapping the attention weights to a thermodynamic potential function.
	* Identifying critical points in the attention distribution by analyzing the first and second derivatives of the thermodynamic potential function.
	* Highlighting the identified critical points in a visualization to indicate the most influential tokens.
6. **A system for real-time analysis of attention mechanisms in interactive systems**, comprising:
	* An attention weight calculation module embedded within the interactive system’s backend.
	* A thermodynamic mapping module that maps the attention weights to a thermodynamic potential function.
	* A critical point identification module that computes the first and second derivatives of the thermodynamic potential function to identify critical points.
	* A visualization tool that dynamically generates a heatmap of attention weights and overlays the critical points during user interaction.
7. **A method for educational purposes to teach attention mechanisms in transformer-based language models**, comprising:
	* Integrating a visualization tool into learning management systems or online platforms.
	* Mapping attention weights to a thermodynamic potential function.
	* Identifying critical points in the attention distribution by computing the first and second derivatives of the thermodynamic potential function.
	* Generating a heatmap of attention weights and overlaying the identified critical points to provide hands-on experience with interpreting attention patterns.
8. **A customizable visualization tool for analyzing attention mechanisms in transformer-based language models**, comprising:
	* A user interface that allows users to adjust parameters such as the control parameter, color schemes, and the level of detail in the heatmap.
	* A module for mapping attention weights to a thermodynamic potential function.
	* A module for identifying critical points in the attention distribution by computing the first and second derivatives of the thermodynamic potential function.
	* A module for generating a heatmap of attention weights and overlaying the identified critical points based on user customization.

### 7. Conclusion

The proposed thermodynamic visualization method offers a groundbreaking approach to enhancing the interpretability of the Multi-Head Self-Attention Mechanism in transformer-based language models. By mapping attention weights to a thermodynamic potential function and identifying critical points through the computation of first and second derivatives, this method provides a robust theoretical framework for understanding attention patterns. The development of a visualization tool that overlays these critical points on a heatmap of attention weights further aids in highlighting the most influential tokens in the input sequence.

This method not only bridges the gap between complex model decisions and human interpretability but also integrates seamlessly into existing transformer architectures without compromising performance. The ability to pinpoint critical points where significant changes in attention distribution occur offers valuable insights into the model’s decision-making process, making it particularly useful for applications in natural language processing where transparency and interpretability are paramount.

The versatility of the thermodynamic visualization method is demonstrated through various embodiments, including integration with existing NLP models, standalone visualization tools, real-time analysis in interactive systems, educational platforms, and customizable visualization options. Each embodiment showcases the method’s potential to enhance interpretability across different domains and use cases.

By providing a deeper understanding of the attention mechanism, this invention paves the way for more transparent and interpretable transformer-based language models, ultimately contributing to the advancement of machine learning and natural language processing fields.

### 8. Keywords

#### 1. Thermodynamic Concepts

* Thermodynamic phase transitions
* Thermodynamic potential function
* Free energy
* Critical points
* Phase transitions

#### 2. Attention Mechanism

* Multi-Head Self-Attention Mechanism
* Attention weights
* Attention matrix
* Scaled dot-product attention
* Attention distribution

#### 3. Mathematical and Computational Methods

* Derivatives
* Control parameter
* Hyperparameter
* Softmax function
* Inflection points

#### 4. Visualization Tools

* Heatmap
* Graphical interface
* Visualization tool
* Overlay markers
* Attention heatmap

#### 5. Transformer Architecture

* Transformer-based language models
* Integration with transformer
* Input sequence
* Model interpretability
* Decision-making process

### References

[1] Jesse Vig (2019) Visualizing Attention in Transformer-Based Language Representation Models. *ArXiv*, abs/1904.02679.

[2] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif & Ann Yuan (2020) The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 107-118.

[3] D. Oprisa & Peter Toth (2017) Criticality & Deep Learning I: Generally Weighted Nets. *arXiv: Artificial Intelligence*.

[4] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin (2017) Attention is All you Need. In *Neural Information Processing Systems*, pages 5998-6008.

[5] Sean Xie, Soroush Vosoughi & Saeed Hassanpour (2023) Proto-lm: A Prototypical Network-Based Framework for Built-in Interpretability in Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 3964-3979.

[6] Yuxuan Lei, Jianxun Lian, Jing Yao, Xu Huang, Defu Lian & Xing Xie (2023) RecExplainer: Aligning Large Language Models for Recommendation Model Interpretability. *ArXiv*, abs/2311.10947.