Infusing Personality into Large Language Models: A Concept Vector Methodology

1st Zhihao Shuai*
HKUST(GZ)§
Guangzhou, China
zhihaoshuai@hkust-gz.edu.cn

2nd Guoyu Li*
Henan University
Kaifeng, China
lgy@henu.edu.cn

3rd Qing Chang*
Nanjing University of Science and Technology
Nanjing, China
qingchang@njust.edu.cn

4th Yuting Dai
South China University of Technology
Guangzhou, China
dyt0818002@gmail.com

5th Bowen Tian
HKUST(GZ)
Guangzhou, China
bowentian@hkust-gz.edu.cn

6th Aming Wu†
Kyungpook National University
South Korea
wuaming@knu.ac.kr

***Abstract*—The advent of large language models (LLMs) has led to significant breakthroughs in natural language processing, often surpassing human performance in various tasks. However, a critical limitation persists: LLMs lack discernible personality traits, which diminishes their effectiveness in emotionally sensitive applications such as mental health support. To address this, we introduce a novel approach called concept vectors, derived from our interpretability research, which enables the embedding of predefined personality traits directly into the internal layers of LLMs. Unlike traditional methods that rely on external models or extensive fine-tuning, our technique leverages the inherent interpretability of LLMs. By visualizing and extracting internal activation patterns, we define concept vectors that represent personality traits, allowing for efficient and controlled personalization of the models. This method significantly enhances the model's ability in sentiment analysis tasks, improving its utility in mental health contexts. Experimental results demonstrate that LLMs infused with concept vectors exhibit improved performance in sentiment analysis. Additionally, our method overcomes challenges posed by traditional fine-tuning and prompt-based interventions, offering a safer and more interpretable approach to personalizing LLMs. These advancements underscore the potential of integrating human empathy into LLMs, thus improving their applicability in real-world scenarios, particularly in mental health care.**

***Index Terms*—Large Language Models, Personality Infusion, Concept Vectors, Interpretability, Mental Health Support, Sentiment Analysis, Fine-tuning**

I. Introduction

The emergence of LLMs has demonstrated their remarkable ability to mimic complex language patterns, often achieving or even surpassing human performance in various tasks[1]. Despite these advancements, a critical limitation persists: the lack of a discernible personality[2]. This issue is particularly significant in mental health support, where existing research shows that approximately 22.5% of patients seek help online[3].While LLMs are increasingly integrated into search engines, their responses often lack empathy and emotional expression due to their impersonal nature. To address this challenge, we introduced a new noun called **concept vectors**, which are derived from the Myers-Briggs Type Indicator (MBTI)-based sentence rewriting task, we were able to endow large language models with predefined personalities, meaning that LLMs carry specific personality inclinations during the decision-making process. This enhancement has shown promising results in sentiment analysis.

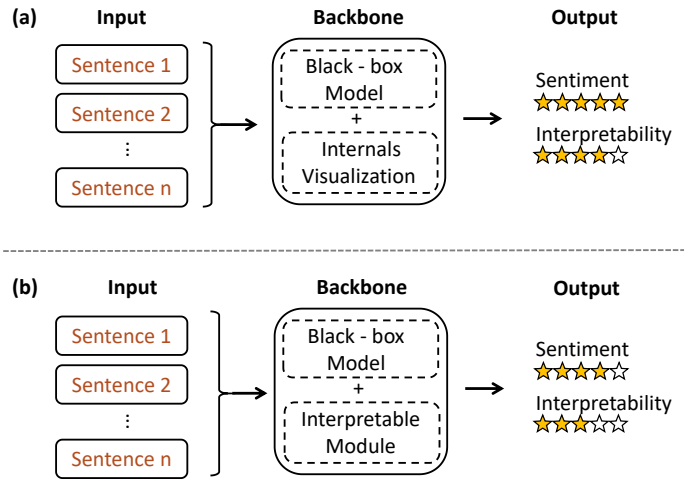

Fig. 1. (a) Previous research on explainability has commonly involved the addition of interpretable modules either before or after black-box models, mapping the contents within the black box to aspects that can be comprehended by humans as a means of achieving interpretability.(b) Our research findings indicate that LLMs possess an inherent capacity to differentiate among certain concepts, which can be understood through visualization techniques, thereby elucidating the model's behavior.

Recent research in interpretability has highlighted two main approaches: *Local Explanations* and *Global Explanations* [4, 5, 6]. Early efforts primarily focused on local explanations, utilizing techniques such as attention weight visualization, probing feature representations, and counterfactual reasoning to provide detailed explanations at the level of individual

∗ The first three authors contributed equally to this work.
§ The Hong Kong University of Science and Technology (Guangzhou)
† Correspondence to Aming Wu Yue (wuaming@knu.ac.kr).

tokens, instances, neurons, or subnetworks [7, 8]. While these low-level explanations are reliable, they often lack readability and intuitiveness, limiting their practical application as shown in Fig 1 (a) [9]. In contrast, recent research has shifted towards global explanations that align more closely with human cognition, such as concept-based analyses [10, 11]. Integrating concept bottleneck models (CBMs) into pretrained language models has yielded significant improvements in the "interpretability-utility" trade-off [12]. This approach has been particularly effective in personality analysis tasks, where human-understandable concepts like "Gentleness," "Anger," and "Rationality" are mapped to neurons in the concept bottleneck layer [13]. The final decision layer, a linear function of these concepts, facilitates easy interpretation of decision rules. However, this method primarily addresses micro-level concepts rather than global task-level concepts, which may lead to an overemphasis on sub-features like gentleness, anger, and rationality, while neglecting the broader personality construct [14].

Our innovation lies in the discovery that LLMs inherently possess a degree of interpretability that can be harnessed without introducing external models or concepts. Specifically, by visualizing the outputs of specific layers within the LLMs, we identified distinct patterns corresponding to personality traits. This discovery allowed us to define and extract what we term as concept vectors, which represent these traits. By visualizing these intermediate layer parameters, we can clearly observe the personality biases reflected in the model's reasoning process, thus achieving model behavior explanation without relying on external methods as shown in Fig 1 (b).

Beyond interpretability, our method offers a significant practical advantage: the introduction of an interpretable and safe vector dictionary intervention. Existing intervention methods typically fall into two categories. The first, Oracle intervention, involves human experts manually adjusting concept activation $\hat{a}$ and feeding it into the classifier. While straightforward, this method fails to correct flawed mappings learned by the LLMs backbone, resulting in recurring errors [5]. The second approach, fine-tuning the LLMs on test data, is inefficient and prone to severe overfitting [15, 16]. Our method overcomes these challenges by leveraging the **concept vectors** within a vector dictionary, which can be overlaid during fine-tuning tasks [11]. This approach is not only more efficient and interpretable but also inherently safer than traditional online fine-tuning or prompt-based methods [17]. Unlike prompt-based interventions, where the outcome may remain uncertain, our approach ensures that the LLMs have genuinely internalized the desired personality traits, providing a more reliable and controlled personalization process [11].

The key contributions of our work are as follows:

- **Introducing Concept Vectors for Personality Control in LLMs**: We propose a method for embedding concept vectors within specific layers of LLMs to effectively endow them with predefined personality traits. This approach is more efficient and inherently safer than traditional fine-tuning or prompt-based methods, ensuring accurate and consistent results in personalization tasks.
- **Revealing and Leveraging Inherent Interpretability of LLMs**: We create a macro-level interpretability framework that visualizes outputs from specific layers of LLMs, allowing for the explanation and control of personality traits without external models.
- **Integrating Human Empathy into LLMs for Improved Mental Health Support**: Our approach underscores the critical importance of empathy and emotional expression in the context of mental health support, enhancing the controllability of LLMs to facilitate emotionally resonant responses. This emphasis on human-like understanding and compassion renders LLMs more effective for real-world applications in the field of mental health care.

II. Methods

Our research involved extensive experiments on general-purpose LLMs, focusing on their inherent ability to represent and activate concepts related to personality, emotion, profession, and ideology within their transformer layers. Specifically, we conducted experiments on Meta-LLaMA-3-8B and Meta-LLaMA-3.1-8B, and discovered that these models inherently exhibit activation vectors corresponding to these concepts, enabling differentiation between different types of concepts without external intervention [18, 19]. By identifying and visualizing the layers where these activation vectors are most prominent, we can explain the outputs of the LLMs [20]. In this study, we extract these activation vectors and integrate them into the specific layers of the respective models, thereby imbuing the models with predefined personality traits [21].

A. Data Processing

We utilized the dataset created by Zou et al.[22] for our study, which is specifically designed to evaluate the factual reasoning capabilities of LLMs. The dataset contains a wide range of factual statements across different domains. To facilitate the model's reasoning process, we employed carefully crafted prompts that guided the LLM in rewriting sentences to incorporate various conceptual dimensions such as personality, emotion, and ideology shown in Fig 2.

```
Prompt:
f"Rewrite the following sentence to reflect the '{style}' style in the MBTI personality framework:\n\n"
f"Sentence: \"{sentence}\"\n\n"
f"Rewritten Sentence:"
```

Fig. 2. Prompt design for guiding the LLM in sentence rewriting

Through this prompt-based guidance, we aimed to analyze the activation vectors generated within the transformer layers, assessing the model's ability to differentiate and systematically activate internal representations of these concepts. This analysis sets the foundation for future work, where we plan to delve deeper into the model's capability to distinguish between different types of concepts based on these internal activations.

B. Concept Vector Extraction

To identify the specific transformer layers where concept vectors are most effectively activated, we first aim to maximize the similarity between vectors corresponding to the same concept and minimize the similarity between vectors from different concepts [23]. We employ cosine similarity to measure this similarity, which is defined as:

$$\text{cosine\_similarity}(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}) = \frac{\mathbf{h}_i^{(l)} \cdot \mathbf{h}_j^{(l)}}{\|\mathbf{h}_i^{(l)}\| \|\mathbf{h}_j^{(l)}\|}, \tag{1}$$

Formally, our objective is to:

$$\max \left(\bar{s}_{i,i}^{(l)} + \bar{s}_{j,j}^{(l)} \right) \quad \text{and} \quad \min \left(\bar{s}_{i,j}^{(l)} + \bar{s}_{j,i}^{(l)} \right), \tag{2}$$

where $\bar{s}_{i,j}^{(l)}$ represents the average cosine similarity between vectors of different concepts i and j at layer l, defined as:

$$\bar{s}_{i,j}^{(l)} = \frac{1}{N} \sum_{k=1}^{N} \text{cosine\_similarity}(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}), \tag{3}$$

where $\mathbf{h}_i^{(l)}$ and $\mathbf{h}_j^{(l)}$ are the hidden state vectors at layer l, corresponding to the input samples associated with concepts i and j, respectively [24]. These hidden state vectors are the representations learned by the transformer model at each layer for different input concepts.

From this objective, we derive the following expression:

$$\max \left(\Delta S^{(l)} \right) = \max \left(\bar{s}_{i,i}^{(l)} + \bar{s}_{j,j}^{(l)} \right) - \min \left(\bar{s}_{i,j}^{(l)} + \bar{s}_{j,i}^{(l)} \right), \tag{4}$$

where $\Delta S^{(l)}$ quantifies how well the model distinguishes between different concepts at layer l. It is defined as:

$$\Delta S^{(l)} = \left(\bar{s}_{i,i}^{(l)} + \bar{s}_{j,j}^{(l)} \right) - \left(\bar{s}_{i,j}^{(l)} + \bar{s}_{j,i}^{(l)} \right). \tag{5}$$

This definition captures the difference between the similarity within the same concept and the similarity across different concepts. By maximizing $\Delta S^{(l)}$, we can identify the layers where the model best differentiates between concepts [25].

Once the specific layers are identified, we proceed to calculate the concept vectors for these layers. The concept vectors are computed as the average difference between the hidden state vectors corresponding to different concepts in the identified layers, defined as:

$$\mathbf{c}^{(j-i)} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathbf{h}_j^{(l)} - \mathbf{h}_i^{(l)} \right). \tag{6}$$

The extracted concept vectors are then projected into a lower-dimensional space, typically a four-dimensional space, for further visualization and analysis [26].

C. Personality Infusion

After determining the base bias vectors and identifying the specific transformer layers, we integrated these bias vectors into the corresponding layers of the respective LLMs. This process involved augmenting the model's parameters with the extracted bias vectors, effectively modifying the model's behavior to exhibit predefined personality traits.

To further assess the efficacy of personality infusion, we tasked the model with sentiment analysis using the Sentiment Labelled Sentences dataset and observed that LLMs infused with the INFJ personality type demonstrated heightened sensitivity in emotional analysis tasks. Additionally, we constructed authentic conversational scenarios to further explore the impact of various personality types on the model's responses, as illustrated in the Fig 3. These experiments confirm that our method not only enables the LLM to exhibit human-like personality traits but also enhances its ability to provide empathetic and contextually appropriate responses, particularly in scenarios where emotional understanding is critical.

III. Experiment and Result

A. Data and Metrics

In our experiments, we employed the dataset[22] which comprises approximately 10,000 sentences with an average of 18 words each, to simulate complex real-world scenarios for LLMs. Additionally, we utilized the Sentiment Labelled Sentences dataset consisting of 3000 sentences for evaluating our fine-tuned model's performance on emotional context understanding[27]. We selected accuracy, precision, recall, and F1-score as the evaluation metrics to provide a comprehensive assessment of the model's capabilities in sentiment classification, which are essential for gauging its sensitivity to emotional nuances and overall predictive accuracy in the realm of sentiment analysis.

B. Experimental Setup

To benchmark our approach, we selected two powerful LLMs as our baselines:

- Meta-LlaMA-3-8B is a versatile large language model that demonstrates strong capabilities in contextual understanding and human-like text generation, suitable for a variety of applications including chatbots and content creation, though it may need further enhancement in processing speed for real-time demanding situations [28].
- Meta-LlaMA-3.1-8B, as an advanced iteration of LlaMA-3, further improves text generation precision and contextual awareness by incorporating cutting-edge language processing technologies. It also significantly increases processing efficiency and system stability through optimized algorithms, making it well-equipped to manage a broader spectrum of complex tasks and high-load AI applications.

The experiments were conducted on an NVIDIA A800-SXM-80GB, where we processed all sentences from the dataset[22] under eight different conceptual dimensions: the four MBTI dimensions (E/I, S/N, T/F, J/P).

TABLE I
SENTIMENT ANALYSIS RESULTS FOR META-LLAMA-3-8B AND META-LLAMA-3.1-8B

Model	Bias Configuration	Accuracy ↑	Precision ↑	Recall ↑	F1-score ↑
Meta-LlaMA-3-8B	Without Bias	0.8861	0.8582	0.9295	0.8925
	With Bias	**0.8914**	**0.8665**	**0.9326**	**0.8983**
Meta-LlaMA-3.1-8B	Without Bias	0.8705	0.8742	0.8691	0.8716
	With Bias	**0.8873**	**0.8902**	**0.8863**	**0.8882**

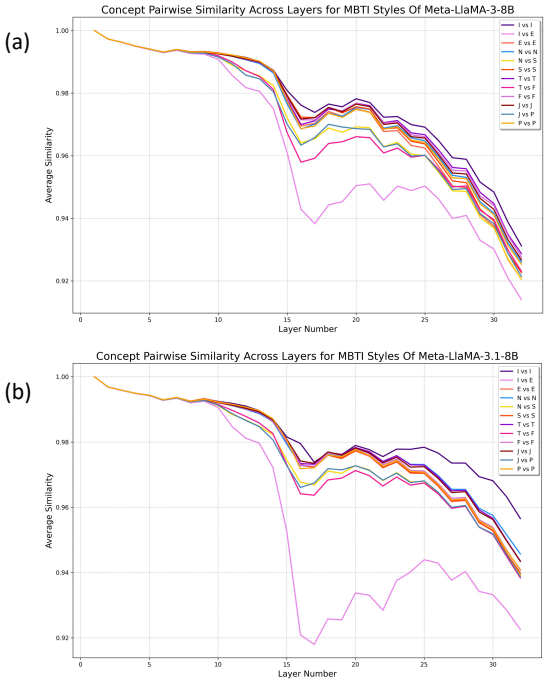

Fig. 3. (a) Previous research on explainability has commonly involved the addition of interpretable modules either before or after black-box models, mapping the contents within the black box to aspects that can be comprehended by humans as a means of achieving interpretability.(b) Our research findings indicate that LLMs possess an inherent capacity to differentiate among certain concepts, which can be understood through visualization techniques, thereby elucidating the model's behavior. Concurrently, we endeavor to extract these distinctions for the purpose of controlling the model's output.

C. Experimental Results and Analysis

Initially, we utilized prompt to have the Meta-LlaMA-3-8B and Meta-LlaMA-3.1-8B rewrite all sentences from the dataset[22], using the four dimensions of MBTI as a framework. Employing the method previously proposed for identifying the most influential layers and extracting concept vectors, visualization techniques revealed the model's internal capacity to distinguish between different MBTI concepts shown in Fig 3. From Fig 3, it is evident that certain layers within the LlaMA series models exhibit a distinct differentiation between similar or opposing concepts, a distinction that can be comprehended and utilized to enhance the interpretability of the large model's outcomes. This inherent interpretability is not only more intuitive and easier to understand than traditional explainability methods, but also facilitates the extraction of concept vectors to control the model's behavior. By combining the vectors of the four dimensions, we were able to construct models representing 16 distinct personalities. Moreover,Moreover, this personality attribution method is more computationally efficient compared to online training fine-tuning and more reliable than guidance through prompts alone.

In the Sentiment Labelled Sentences dataset, we compared baseline models with fine-tuned models using the INFJ personality as an example. Table 1 shows that both Meta-LlaMA-3-8B and Meta-LlaMA-3.1-8B improved by approximately 1% across all sentiment analysis metrics after fine-tuning towards the INTJ personality. The enhancements in the first three metrics indicate better affective recognition, while the increased F1-Score reflects overall improved performance due to the attribution of specific personality traits. This demonstrates that our concept vector approach can effectively impart distinct personalities to large models, enhancing their emotional analysis capabilities. Integrating such models into Artificial Intelligence(AI) search engines will improve emotional analysis in question-answering scenarios, better addressing the needs of individuals with mental health concerns and elevating empathetic care standards.

Throughout our experiments, we additionally discovered that there are discernible differences in the internal representations of large models regarding the similar and dissimilar aspects of concepts such as emotions, professions, and ideologies. In our forthcoming research, we plan to further investigate this inherent interpretability to enhance the safety and efficiency of managing large models.

IV. CONCLUSION

This paper introduces the concept of concept vectors to address the limitations of LLMs in terms of personality. Our approach leverages the interpretability of LLMs to embed predefined personality traits without the need for external models or extensive fine-tuning, providing a safer and more efficient method for personalization, particularly in emotionally sensitive applications such as mental health support. Experiments demonstrate that LLMs infused with concept vectors exhibit higher accuracy and emotional resonance in sentiment analysis tasks. Our research reveals the potential for internal interpretability within LLMs, opening new possibilities for personality and empathy in human-AI interaction. Future work will focus on further optimizing this approach and exploring additional applications.

References

[1] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[2] K. Crawford, "The atlas of ai: Power, politics, and the planetary costs of artificial intelligence," 2021.

[3] M. Reinert, D. Fritze, and T. Nguyen, "The state of mental health in america 2024," 2024.

[4] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 10, pp. 4291–4308, 2020.

[5] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5338–5348.

[6] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," in *ISMIR*, vol. 53, 2017, pp. 537–543.

[7] A. Ross, A. Marasovic, and M. E. Peters, "Explaining nlp models via minimal contrastive editing (mice)," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3840–3852.

[8] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld, "Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models," in *59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021, pp. 3434–3449.

[9] M. Losch, M. Fritz, and B. Schiele, "Interpretability beyond classification output: Semantic bottleneck networks," *arXiv preprint arXiv:1907.10882*, 2019.

[10] E. D. Abraham, K. D'Oosterlinck, A. Feder, Y. Gat, A. Geiger, C. Potts, R. Reichart, and Z. Wu, "Ce-bab: Estimating the causal effects of real-world concepts on nlp model behavior," in *NeurIPS*, vol. 35, 2022, pp. 17 582–17 596.

[11] Z. Tan, L. Cheng, S. Wang, Y. Bo, J. Li, and H. Liu, "Interpreting pretrained language models via concept bottlenecks," *arXiv preprint arXiv:2311.05014*, 2023.

[12] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," *International Conference on Learning Representations (ICLR)*, 2023.

[13] Z. Tan, L. Cheng, S. Wang, Y. Bo, J. Li, and H. Liu, "Interpreting pretrained language models via concept bottlenecks," 2023.

[14] C. Meisser, S. Lazov, I. Augenstein, and R. Cotterell, "Is sparse attention more interpretable?" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 122–129.

[15] U. Evci, T. Gale, J. Menick, P. S. Castro, and E. Elsen, "Rigging the lottery: Making all tickets winners," in *International Conference on Machine Learning*, 2020, pp. 2943–2952.

[16] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," *arXiv preprint arXiv:2306.11695*, 2023.

[17] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," in *Advances in Neural Information Processing Systems*, 2023.

[18] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 37–42.

[19] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] P. Schoenegger, S. Greenberg, A. Grishin, J. Lewis, and L. Caviola, "Can ai understand human personality? – comparing human experts and ai systems at predicting personality correlations," *arXiv preprint arXiv:2406.08170*, 2024.

[21] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "Towards controllable biases in language generation," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 129–140.

[22] S. C. J. C. P. G. R. R. A. P. X. Y. M. M. A.-K. D. S. G. N. L. M. J. B. Z. W. A. M. S. B. S. K. D. S. M. F. Z. K. D. H. Andy Zou, Long Phan, "Representation engineering: A top-down approach to ai transparency," 2023.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations (ICLR)*, 2013.

[24] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[25] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 55–65.

[26] A. W. Liu, T. Scao, B. Zoph, C. Maddison, and C. Raffel, "What makes good in-context examples for gpt-3?" *arXiv preprint arXiv:2101.06804*, 2021.

[27] D. Kotzias, M. Denil, N. Kalchbrenner, P. Smyth, and N. De Freitas, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 597–606.

[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2307.09288*, 2023.