Multi-Emotional Probing for Controllable Language Generation: Exploring the Geometry of Emotion in Transformer Models

Name: Daria Feng, junior undergraduate
Major in Computer Science,
Contact: yfeng266@wisc.edu

Name: Eric Xu, junior undergraduat
Major in Statistics,
Contact: zxu684@wisc.edu

Name: Shengbo Qian, senior undergraduate
Major in Statistics,
Contact: sqian37@wisc.edu

Name: Huuiyu Li, senior undergraduate
Major in Statistics,
Contact: hli798@wisc.edu

Course Number: Stat 453, Spring 2025
Team number: 06

File compiled on Overleaf

Project Objectives.

1. **Motivation.** Large language models (LLMs) have shown impressive capabilities in generating human-like text, but controlling their emotional expression remains challenging. Current approaches rely heavily on prompt engineering, which lacks fine-grained control. We aim to study how emotions are internally represented within language models and develop a method to directly manipulate these representations to achieve more precise emotional control in text generation [5, 1, 2].
2. **Existing literature.** Recent surveys have categorized interpretability and controllability approaches across modalities and model architectures [2]. Current research on emotion control in LLMs mainly follows two approaches: prompt-based methods and fine-tuning techniques. Prompt-based methods instruct models to express specific emotions but offer limited control granularity. Fine-tuning approaches adapt models for emotional expression but require extensive resources and sacrifice general capabilities. Recent work on model interpretability has explored feature direction intervention and internal vector manipulation [3, 1], and methods such as LM-Debugger offer insights into model internals [3], but have not been extensively applied to emotional control.
3. **Limitations of existing models or methods.** Existing approaches face several limitations: (1) Prompt engineering provides only coarse control over emotional expression; (2) Fine-tuning methods require significant computational resources and risk catastrophic forgetting; (3) Neither approach allows for precise mixing of multiple emotions; (4) Current methods provide little insight into how emotions are internally represented within LLMs [5].
4. **Overview of proposed project.** Our project will develop a novel approach that uses linear probes to discover "emotion direction vector" in transformer models' latent space. By identifying these directions in BERT's representation space and transferring them to generative models like LLaMA-2, we can achieve fine-grained control over emotional expression, including the mixing of multiple emotions [5]. We will visualize how different layers respond to emotional interventions, providing interpretability insights into the model's internal representation of emotions [3, 4].

Models. Our project employs a two-stage approach with distinct models for emotion representation extraction and controlled text generation. For the first stage, we will use BERT-large-uncased (330M parameters) as our probe model to extract emotion representations from hidden layers. For the second stage, we will use Llama-2-13B as our generation model for emotion vector injection.

Datasets. We will use the GoEmotion dataset developed by Google Research. This dataset contains approximately 58,000 Reddit comments annotated with 27 fine-grained emotion categories and a neutral class. The multi-label annotation scheme allows comments to express multiple emotions simultaneously, making it ideal for our investigation of emotion mixing. Categories include basic emotions like joy, anger, and sadness, as well as more nuanced emotions such as admiration, curiosity, and remorse. The dataset is publicly available through the TensorFlow Datasets library and includes standard train/validation/test splits. GoEmotions GitHub Repository

Measurements. We evaluate our method using four key metrics that reflect both controllability and interpretability.
(1) For the probe model, we report classification accuracy, precision, recall, and F1 score to assess how well emotional categories are captured in hidden states.
(2) For the generation phase, we measure emotion expression accuracy by applying external

classifiers (e.g., VADER or trained emotion probes) to determine whether the generated text reflects the intended emotion.
(3) To assess fine-grained control, we evaluate emotion mixing effectiveness by interpolating between different emotion directions (e.g., 0.7 anger + 0.3 sadness) and analyzing output variations.
(4) Finally, we measure layer-wise impact by injecting the same emotion vector into different Transformer layers and visualizing the resulting emotion expression intensity through heatmaps, revealing which layers are most sensitive to emotional control.

Compute budget. Our compute budget is based on the availability of a single A100 GPU with 40GB VRAM accessed through Google Colab Pro, with approximately 200GB of Google Drive storage space for models and datasets. We will experiment with two main models: BERT-large-uncased (330M parameters) and Llama-2-13B (13B parameters).

- BERT-large emotion probe training: The BERT-large model requires approximately 6GB of VRAM. Training linear probes on GoEmotion dataset (58,000 examples) will take approximately 2-3 hours per run.
- Llama-2-13B inference with emotion vectors: The Llama-2-13B model requires 26GB in standard precision, but we will use 4-bit quantization (QLoRA) to reduce memory footprint to approximately 10GB. Each generation experiment with emotion vector injection will take approximately 2-4 hours to generate and evaluate 100 text samples.

The total compute budget is estimated at 90 GPU hours on an A100. This is feasible within a 4-week project timeline.

Collaboration plan.

Daria Feng: design idea; write proposal; preprocess the GoEmotion dataset; design BERT-based emotion probe models; extract emotion vectors and validate their effectiveness; visualize emotion representation spaces.
Eric Xu: design idea; write proposal; deploy the Llama-2 generation model; develop emotion vector injection mechanisms; implement mixed emotion control logic; optimize the generation process for efficiency.
Huiyu Li: create heatmaps and comparison plots; visualize emotion vector distributions; develop emotion score analysis methods; design evaluation metrics.
Shengbo Qian: coordinate the complete workflow; write the experimental report; create presentation slides; summarize research findings and applications.

References

[1] A. Pan, L. Chen, & J. Steinhardt, LATENTQA: Teaching LLMs to Decode Activations into Natural Language, arXiv preprint arXiv:2412.08686 (2024).

[2] Y. Dang, K. Huang, J. Huo, Y. Yan, S. Huang, D. Liu, M. Gao, J. Zhang, C. Qian, K. Wang, Y. Liu, J. Shao, H. Xiong, & X. Hu, Towards Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey, arXiv preprint arXiv:2412.02104 (2024).

[3] M. Geva, A. Caciularu, G. Dar, P. Roit, S. Sadde, M. Shlain, B. Tamir, & Y. Goldberg, LM-Debugger: An Interactive Tool for Inspection and Intervention in Transformer-Based Language Models, arXiv preprint arXiv:2204.12130 (2022).

[4] L. Weissweiler, V. Hofmann, A. Köksal, & H. Schütze, The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative, In *Proc. EMNLP*, 10859–10882 (2022).

[5] A. N. Tak, A. Banayeeanzade, A. Bolourani, M. Kian, R. Jia, & J. Gratch, Mechanistic Interpretability of Emotion Inference in Large Language Models, arXiv preprint arXiv:2502.05489 (2024).