NLP-Cimat at SemEval-2025 Task 11: Prompt Optimization for LLMs via Genetic Algorithms and Systematic Mutation applied on Emotion Detection

Guillermo Segura-Gómez[1], **A. Pastor López-Monroy**[1], **Fernando Sánchez-Vega**[1,2], **Alejandro Rosales-Pérez**[3]
[1,3]Mathematics Research Center (CIMAT) * †
[2]Secretaría de Ciencias, Humanidades, Tecnología e Innovación (SECIHTI) ‡
{guillermo.segura, pastor.lopez, fernando.sanchez, alejandro.rosales}@cimat.mx

Abstract

Large Language Models (LLMs) have shown remarkable performance across diverse natural language processing tasks in recent years. However, optimizing instructions to maximize model performance remains a challenge due to the vast search space and the nonlinear relationship between input structure and output quality. This work explores an alternative prompt optimization technique based on genetic algorithms with different structured mutation processes. Unlike traditional random mutations, our method introduces variability in each generation through a guided mutation, enhancing the likelihood of producing better prompts at each generation.. We apply this approach to emotion detection in the context of SemEval 2025 Task 11 for English language solely, demonstrating the potential to improve prompt efficiency, and consequently task performance. Experimental results show that our method yields competitive results compared to standard optimization techniques while maintaining interpretability and scalability.

1 Introduction

Large Language Models (LLMs) have experienced significant growth in recent years. Their remarkable performance stems from their ability to understand and model language more effectively than any previously developed tool (Brown et al., 2020). The essential interest in LLMs lies in their capacity to excel at numerous specific tasks without requiring extensive fine-tuning or contextual information (Radford et al., 2019; Devlin et al., 2019). This is quite powerful in many ways. On the one hand, more traditional machine learning or deep learning models require a significant amount of data to achieve LLM performance (LeCun et al., 2015). On the other hand, by having an LLM available, you have a model capable of performing almost any natural language-related task with a high level of competence. Nevertheless, despite the outstanding performance demonstrated by LLMs, their ability to process and understand subjective aspects of text, such as human emotions, remains a complex challenge (Zhang et al., 2023; Sabour et al., 2024, Singh et al., 2023).

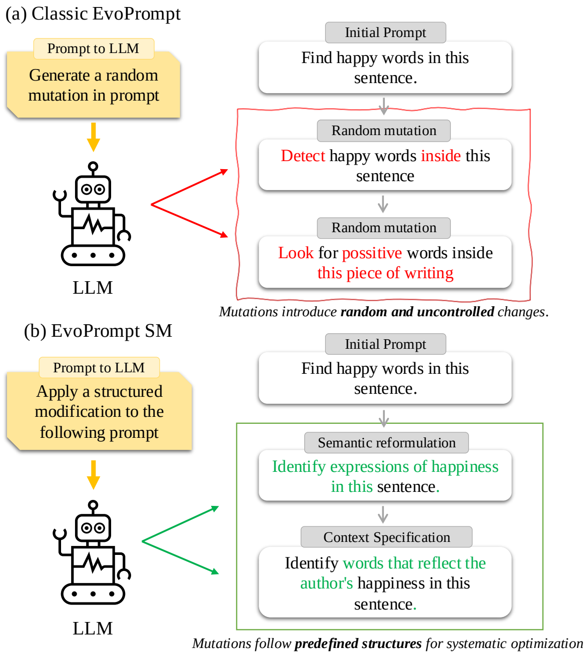

Figure 1: Comparison between Classic EvoPrompt and EvoPrompt SM. In Classic EvoPrompt (a), mutations occur randomly, leading to uncontrolled modifications. In EvoPrompt SM (b), structured mutations such as *semantic reformulation* and *context specification* are applied, ensuring systematic optimization.

One particularly challenging task is identifying the emotion experienced by the author when writing a text, rather than the emotion perceived by the reader (Alvarez-Gonzalez et al., 2021). This distinction is crucial in tasks such as sentiment

*Jalisco S/N Valenciana, 36023, Guanajuato, Guanajuato, México
†Monterrey, Av. Alianza Centro 502, Apodaca, 66628, Nuevo León, México.
‡Av. Insurgentes Sur 1582, Col. Crédito Constructor, 03940, CDMX, México

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1662–1669
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

analysis, psychological research, and user experience evaluation. The **SemEval 2025 Task 11A** competition focuses precisely on this challenge (Muhammad et al., 2025b), providing a dataset where sentences are labeled based on the author's emotional state at the time of writing, rather than how a reader interprets the text (Muhammad et al., 2025a). This task is more complex than traditional emotion classification, especially for LLMs, which lack direct access to human emotional experiences. They infer emotions based purely from linguistic patterns present in their training data (Chochlakis et al., 2024). Therefore, determining the optimal way to prompt an LLM to infer the author's emotions is non-trivial and requires careful design (Li et al., 2023).

However, the performance of LLMs is highly dependent on how prompts are constructed (Desmond and Brachman, 2024). Developing more effective prompts is essential, particularly given that there is no single correct method for doing so (Li et al., 2025). Within this context, prompt engineering has reached a boom, and human-constructed prompts are the vast majority of the time used to perform tasks with an LLM (Webber et al., 2020). Despite this, determining how to best phrase a prompt to make an LLM infer the emotional state of an author remains an open problem.

In this paper, we explore an approach based on the use of genetic algorithms to optimize prompts for for LLM-based emotion classification. We explore an alternative mutation designed to introduce structured variability at each generation, ensuring that mutations are aligned with patterns that have shown potential for enhancing prompt quality. By systematically evolving prompts without human intervention, this method offers a robust and scalable solution for tasks that require accurate emotional inference. While our approach is evaluated in the context of emotion classification, its potential applications include in contexts where optimized prompts without human intervention are needed, such as chatbots (Yigci et al., 2024), code generation (Chen et al., 2021), and automation of complex tasks with LLMs (Bommasani et al., 2021).

2 Related Work

The traditional methods used for emotion classification were lexicon-based approaches (Cambria et al., 2017), where a predefined list of words was used to classify sentences according to sentiment by number of occurrences or any other linguistic criterion. These methods faced significant challenges related to context dependency and polysemy, which limited their accuracy in complex texts (LeCun et al., 2015). The advent of deep learning marked a revolution, as word embeddings and transformer-based approaches could be used to do emotion classification. Models such RoBERTa and TS showed superior performance compared to more traditional approaches (Adoma et al., 2020; Kolev et al., 2022). More recently, LLMs have shown comparable performance while being more cost-effective in terms of data and training requirements. Therefore, optimizing prompts for these tasks has become a more efficient approach (Liu et al., 2023; Imran, 2024).

Algorithm 1 EvoPrompt Classic vs EvoPrompt SM

Input: Initial population of prompts P_0, number of generations G, population size N
Output: Optimized set of prompts P_G
Initialize population P_0 with N prompts (human-crafted + LLM-generated)
for $g = 1$ to G **do** ▷ Start Evolutionary Process
  Selection: Choose M parent prompts using tournament, wheel or random selection
  Crossover: Generate offspring prompts via crossover operation
  if EvoPrompt Classic then
    Mutation: Apply random mutation to offspring
    Selection: Choose top N prompts based on fitness
  else if EvoPrompt SM then
    Selection 1: Choose top candidates for mutation after crossover
    Mutation: Apply structured mutation from predefined set
    Selection 2: Choose top N prompts based on fitness after mutation
  end if
  Update population: $P_{g+1} \leftarrow$ selected best prompts
end for
Return final optimized prompt set P_G

The process of optimizing prompts for a language model in an automated manner is known as *automated prompt generation*. Different approaches aim to generate improved synthetic prompts (Li et al., 2025). Biologically inspired approaches to prompt optimization treat the problem as an evolutionary process (Shapiro, 1999). In evolution, prompts are viewed as organisms, which are managed through genetic operations such as mutation or crossover over epochs. The pioneering work in implementing an evolutionary process using an LLM as an optimizer is *EvoPrompt* (Guo et al., 2023). In EvoPrompt a more traditional approach to an evolutionary algorithm is proposed, in which an evaluation function chooses the best prompts that maximize the score of the task at hand. Other approaches, such as *Promptbreeder*, introduce a self-referential method, where both task-prompts and mutation-prompts evolve through a genetic algorithm guided by LLMs (Fernando et al., 2023; Chen et al., 2024).

Emotion classification using LLMs has been extensively explored in recent studies. Various approaches have reshaped the way LLMs are employed for this task. Specifically, we can divide these efforts into two main categories: those that fine-tune LLMs for emotion classification tasks (Zhang et al., 2023; Liu et al., 2024) and those that leverage LLMs' inherent ability to detect emotions, assessing their performance across different contexts solely through prompt optimization (Venkatakrishnan et al., 2023; Peng et al., 2024). In both contexts, an automated approach to find optimal prompts for emotion classification is a highly desirable need. Therefore, this study explores an alternative framework that mitigates the stochastic nature of genetic algorithms by changing the way mutations are performed.

3 Methodology

To tackle the problem, we propose a solution based on LLMs using a zero-shot/few-shot approach. As previously discussed, selecting the optimal prompt is challenging and directly impacts LLM performance. We employed a genetic algorithm to optimize prompts through an evolutionary process customized to the requirements of the task.

The overall structure follows a classical genetic algorithm approach, where prompts undergo iterative selection, crossover, and mutation to improve a fitness function. The distinction between random mutations (classical EvoPrompt) and systematic mutations (our approach) is visually depicted in Figure 1, highlighting the key differences between both strategies. The step-by-step process is outlined in Algorithm 1.

The process begins with an initial population of $2n$ prompts, comprising both human-crafted prompts, manually designed based on linguistic heuristics and task-specific considerations, and those generated by GPT-4o. All initial prompts are evaluated individually. The elements then enter the evolutionary cycle, following an approach similar to **EvoPrompt**, which utilizes a classical genetic algorithm. Prompt selection is performed using three different methods: tournament selection, roulette wheel selection, and random selection.

Once a pair of parent prompts is selected, a crossover operation is applied resulting in a child prompt. After all crossover operations, the top n prompts are evaluated and selected for the next generation. This process is iterated for a predetermined number of epochs using the top n prompts. The prompts from the final epoch are expected to be superior to those from the initial population. The typical range in which we use our approach is $10 \leq n \leq 30$. For this work we use $n = 10$. The optimization process was run for 10 epochs due to computational limitations.

The prompts are evaluated using the same LLM as the fitness function. The main idea is to iteratively refine the prompts generated during evolution, as these prompts are directly used to perform the emotion detection task. Each prompt is evaluated over the validation set of the dataset by calculating the F1 score for its predictions. The selection process is detailed in Algorithm 1. The process is run for each sentiment independently. Predictions are made through a discrete prompting setup: the LLM is asked to make a binary decision using prespecified target words. Initially, the words used are `positive` and `negative`, where the model predicts the presence or absence of a target emotion in a sentence. To further understand the sensitivity of the evaluation, we include an ablation study where the target words are replaced with `present` and `absent`, and analyze the resulting impact on prompt performance.

3.1 Mutation Strategy: Random vs Systematic Evolution

In classical genetic algorithms, mutations are typically random perturbations that introduce uncontrolled variations that could enhance performance. However, this mutation approach often fails to produce the desired effect. The stochastic nature of random mutations reduces the likelihood of generating beneficial variations tailored to the specific task at hand.

To overcome this limitation, our approach replaces random mutations with **systematic mutations**, designed to introduce structured linguistic variation in each generation. Rather than relying on stochastic modifications, our model selects transformations from a predefined set, ensuring that each mutation follows linguistic optimization principles. Each type of mutation is validated to have a positive impact on performance, avoiding disruptive changes that could degrade the prompt's effective-

Emotion	Best Prompt	Macro F1-Score
Anger	Analyze if the sentence expresses anger [...] identify indicators of hostility [...] examining language, structure, or context.	0.4309
Fear	As a Linguistic Analyst, classify phrases that create unease or fear [...] identify specific words contributing to a nervous or tense tone.	0.7734
Joy	Identify happy words in this sentence.	0.7221
Sadness	Assess if the sentence conveys gloom or sorrow [...] identifying words that contribute to a somber tone.	0.6667
Surprise	Does the sentence contain a surprising event or plot twist [...] creating shock or astonishment?	0.5251

Table 1: Best performing prompt per emotion with corresponding Macro F1-score.

ness. By aligning mutations with known patterns that enhance LLM interpretability and task adaptation, this deterministic approach improves convergence speed, reduces variance in performance across generations, and ensures a more consistent refinement of prompts. Unlike random mutations, which may generate unproductive or even detrimental variations, structured modifications incrementally optimize the prompt space, leading to a more stable and efficient evolutionary process.

The structured mutations:

- **Context Specification**: Clarifies and refines the prompt's focus.
- **Lexical Reformulation**: Rewords prompts while preserving meaning.
- **Profiling**: Adapts prompts based on predefined linguistic traits.
- **Simplification**: Reduces complexity for clearer interpretation.

By controlling each mutation, we enhance replicability while preserving diversity in the evolutionary search space. The comparative impact of systematic versus random mutations is discussed in more detail in Figure 1.

3.2 Experimental Setup

The model used for evaluation tasks, crossover generation, and systematic mutations is **Llama 3.1 8B**. The implementation was carried out using `PyTorch` with the `transformers` library from Hugging Face (Wolf et al., 2020), leveraging the `bitsandbytes` library for optimized inference in low-precision configurations (Dettmers et al., 2022).

The model is executed in an **8-bit quantized configuration**, which significantly reduces memory consumption and computational requirements while maintaining comparable performance to full-precision models (Frantar et al., 2022). The execution hardware consists of two **NVIDIA Titan RTX graphics cards** with 24 GB of DDR6 memory, hosted by the Supercomputing Laboratory of the Bajío, located at the Center for Research in Mathematics (CIMAT), Guanajuato, Mexico (Centro de Investigación en Matemáticas A.C, n.d.).

4 Results and Discussion

The model was executed using the random mutation configuration, following an approach similar to *EvoPrompt*. This was done to compare the results obtained with the proposed systematic mutation model. Likewise, the systematic mutation model was executed, and its results are presented in Table 3. Table 2 shows the results using the validation dataset for English solely. The performance of the initial Llama model with a generic initial prompt is compared, along with the classical EvoPrompt approach and EvoPrompt with systematic mutations.

One of the best-performing prompt was from the *joy* category (Table 1), specifically: *Identify happy words in this sentence*. The notable aspect of this prompt is that it resulted from a systematic mutation. All prompts in that population generally had low scores (Macro F1-Score ~ 0.55), and even after evolution, the validation score only improved slightly (Table 2). The reason this prompt achieved such a high score is that it aligns closely with the dataset's focus on the author's perceived emotion. The prompt guides the language model to identify linguistic patterns that reflect the author's emotional state, as *identifying happy words* is more related to the expressed emotion than to the perceived emotion.

This reasoning explains the overall structure of the best-performing prompts for *fear* and *joy*, as

Emotion	Anger	Fear	Joy	Sadness	Surprise
Llama Initial	[0.5941, 0.6170]	[0.6522, 0.6955]	[0.5401, 0.5452]	[0.6477, 0.6697]	[0.6064, 0.6720]
Llama EvoPrompt	[0.6397, 0.6470]	[0.7395, 0.7522]	[0.5401, 0.5452]	[0.6842, 0.6892]	[0.7108, 0.7225]
Llama EvoPrompt SM	[0.6528, 0.6602]	[0.7546, 0.7676]	[0.5533, 0.8131]	[0.6982, 0.7033]	[0.7253, 0.7372]

Table 2: Validation F1-score range $[min, max]$ per emotion category. The values represent the Macro F1-score per emotion, calculated on the validation set. The range corresponds to the results of the final epoch for Llama EvoPrompt and Llama EvoPrompt SM (Systematic Mutation). In the case of the initial model evaluation (Llama Initial), it refers to the range of values obtained from the initially evaluated prompts.

Emotion	EvoPrompt Modified	EvoPrompt Original
Anger	0.4309	0.4223
Fear	0.7734	0.7579
Joy	0.7221	0.7077
Sadness	0.6667	0.6534
Surprise	0.5251	0.5146
Macro F1	0.6236	0.6111
Micro F1	0.6571	0.6440

Table 3: Comparison between EvoPrompt Original and Modified. All values correspond to the F1-score metric. These results were part of the official SemEval submission.

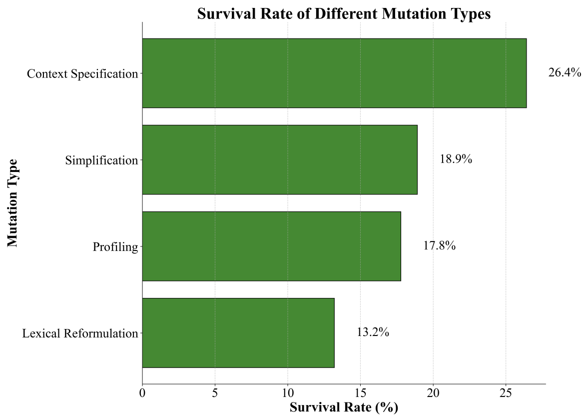

Figure 2: Survival rates of prompts based on the applied mutation type. The rates represent the percentage of prompts that survived the selection process after each specific mutation was applied, aggregated across all emotion categories.

well as the lower performance observed for *anger*, *sadness*, and *surprise*. The prompts obtained for these emotions share a common approach of searching for the emotion within the sentence, making them more suitable for detecting the emotion perceived by the reader.

These findings underscore the potential of systematic mutations, which, relying solely on prompt engineering assumptions, produced targeted modifications. This approach generated prompts that effectively identified task-relevant patterns, surpassing the EvoPrompt method, where random mutations failed to yield superior results. This suggests that replacing stochastic mutations with structured linguistic modifications enhances both effectiveness and consistency in prompt generation, leading to improved overall performance.

Another possible explanation for the model's success in prediction could come from a class imbalance. From the dataset paper (Muhammad et al., 2025a), we know that the most represented emotion is *fear*, while the least represented is *anger*, which aligns with the results obtained in Table 3. However, *joy* is the second least represented emotion and still achieved the second-best score, which challenges this explanation. Additionally, reinforcing this point, the results in Table 4 show that under the alternative evaluation, the *surprise* class performed worse, even though it has similar representation to the *sadness* class, which achieved a higher score.

4.1 Mutation Success Analysis

To better understand the internal dynamics of our evolutionary process, we analyzed the survival rates of different mutation types across generations. Figure 2 shows the percentage of surviving prompts after applying each mutation type. The most successful mutation was `context specification`, followed by `simplification`, while `lexical reformulation` exhibited the lowest survival rates. These results suggest that mutations focusing on refining the task specification were more effective, whereas mutations that altered the way the model is addressed tended to be less successful.

4.2 Evaluation Ablation Study

As mentioned in the methodology, the evaluation method was carried out using the same language model. The results presented in Table 3, corresponding to the official competition submission, the tokens `positive` and `negative` were used as target tokens. However, it is possible that these tokens introduce issues when detecting certain emotions, since restricting the prediction to positive or negative biases emotions like *anger* or *surprise*, which are not easily distinguishable with only these two labels. For this reason, a second study was conducted using different tokens `present` and `absent`, which are more aligned with the dataset's design and the task of predicting the emotion itself. The idea was that they would better capture whether the emotion was present or not. The results obtained are shown in Table 4. Comparing the two evaluations, the second approach clearly achieves superior performance, demonstrating a significant impact of this adjustment on the model.

Emotion	**EvoPrompt Modified**	**EvoPrompt Original**
Anger	0.6909	0.6557
Fear	0.7568	0.6176
Joy	0.7593	0.7600
Sadness	0.7550	0.7381
Surprise	0.6625	0.5967
Macro F1	0.7249	0.6736
Micro F1	0.7451	0.6781

Table 4: EvoPrompt Modified evaluated using an alternative evaluation approach. All values correspond to the F1-score metric. These results were not included in the official SemEval submission.

5 Conclusion

This study introduced a novel approach for optimizing prompts via systematic mutations guided by genetic algorithm principles. By replacing stochastic mutations with structured linguistic modifications, the proposed method enhanced prompt effectiveness and consistency, leading to superior performance across all emotion categories. Notably, the improvements in joy and fear suggest that aligning mutations with underlying linguistic patterns can significantly impact classification accuracy. These findings highlight the potential of systematic mutation strategies in prompt engineering, paving the way for more efficient and automated optimization techniques in LLM-driven emotion classification. Future work could explore refining mutation strategies further and extending this approach to other NLP tasks.

Limitations

This study has some limitations that should be taken into account for future improvements. First, the optimization process was limited to ten iterations due to time and computational constraints. This probably restricted the potential of the model, especially in emotions such as anger, sadness, and surprise, where it is more difficult to capture subtle linguistic patterns. With more iterations and more precise and above all perhaps somewhat more deterministic mutation rules, performance could be improved, especially by generating messages capable of detecting emotional nuances more effectively.

Second, the systematic mutations were designed based on general prompt engineering assumptions, which may not fully capture the complexity of all linguistic expressions. Furthermore, the evaluation was performed only on the SemEval Task 11A dataset, which limits the generalizability of the results. It is important to test the method on datasets with different annotation schemes and language models to assess its robustness. Future work could also explore integrating other prompt tuning techniques for a more complete comparison.

Acknowledgments

The authors gratefully acknowledge the Centro de Investigación en Matemáticas (CIMAT) and the Secretaría de Ciencias, Humanidades, Tecnología e Innovación (SECIHTI) for providing the computing resources of the CIMAT Bajío Supercomputing Laboratory (#300832).

Sánchez-Vega acknowledges SECIHTI for its support through the program “Investigadoras e Investigadores por México” (Project ID.11989, No.1311), and Rosales-Pérez acknowledges SECIHTI for its support through the grant project “Búsqueda de arquitecturas neuronales eficientes y efectivas” (CBF2023-2024-2797).

Segura-Gómez acknowledges financial support from SECIHTI through the graduate scholarship number 4006758 and CVU 1308651. The authors also express their gratitude to the organizers of SemEval 2025 Task 11A for providing the dataset and challenge that motivated this research.

References

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, pages 117–121. IEEE.

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. *arXiv preprint arXiv:2109.01900*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

(CIMAT) Centro de Investigación en Matemáticas A.C. n.d. Laboratorio de supercómputo del bajío. Accessed on February 24, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. Prompt optimization in multi-step tasks (promst): Integrating human feedback and preference alignment. *arXiv e-prints*, pages arXiv–2402.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2024. The strong pull of prior knowledge in large language models and its impact on emotion recognition. *arXiv preprint arXiv:2403.17125*.

Michael Desmond and Michelle Brachman. 2024. Exploring prompt engineering practices in the enterprise. *arXiv preprint arXiv:2403.08950*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Mia Mohammad Imran. 2024. Emotion classification in software engineering texts: A comparative analysis of pre-trained transformers language models. In *Proceedings of the Third ACM/IEEE International Workshop on NL-based Software Engineering*, pages 73–80.

Vladislav Kolev, Gerhard Weiss, and Gerasimos Spanakis. 2022. Foreal: Roberta model for fake news detection based on emotions. In *ICAART (2)*, pages 429–440.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Wenwu Li, Xiangfeng Wang, Wenhao Li, and Bo Jin. 2025. A survey of automatic prompt engineering: An optimization perspective. *arXiv preprint arXiv:2502.11560*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. 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*, pages 5487–5496.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Jonathan Shapiro. 1999. Genetic algorithms in machine learning. In *Advanced course on artificial intelligence*, pages 146–168. Springer.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2023. Language models (mostly) do not consider emotion triggers when predicting emotion. *arXiv preprint arXiv:2311.09602*.

Radhakrishnan Venkatakrishnan, Mahsa Goodarzi, and M Abdullah Canbaz. 2023. Exploring large language models' emotion detection abilities: use cases from the middle east. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 241–244. IEEE.

Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. 2020. Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Defne Yigci, Merve Eryilmaz, Ail K Yetisen, Savas Tasoglu, and Aydogan Ozcan. 2024. Large language model-based chatbots in higher education. *Advanced Intelligent Systems*, page 2400429.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.