Bridging the Emotional Gap: Enhancing Compact Language Model's Emotional Intelligence through Targeted Fine-Tuning

19007204 **23228695** **23068286** **21128382** **22073506** **23130966**

Abstract

This study showcases potential advancements in enhancing the emotional intelligence of compact language models through targeted fine-tuning with emotionally enriched datasets. Notably, the GPT2-med.-baby-instr model excelled, achieving a remarkable 94.8% positivity rate in responses to positive prompts, vastly surpassing the baseline model's 41%. This model also demonstrated strong resilience in maintaining a 78.9% positivity rate in negative contexts. Additionally, all models preserved robust general linguistic capabilities, confirmed by their solid performance on the BLiMP benchmark. These results highlight the efficacy and potential of smaller, strategically trained models in handling complex emotional scenarios, encouraging further development and optimization for broader real-world application. We release the codebase publicly on GitHub[1].

1 Introduction

Recent advancements in natural language processing (NLP) have predominantly centered around the development of increasingly large language models (LLMs). Whilst these models have demonstrated impressive capabilities across a wide range of linguistic tasks, their scalability is often constrained by significant computational demands and environmental impacts associated with their size.

Contrasting this trend, emerging research indicates that smaller, more efficient models can perform comparably to their larger counterparts when properly optimized. Studies such as those by Zhang et al. (2023) and Liu et al. (2023) illustrate that with strategic training, small models are not only viable but also highly effective. Inspired by these findings, our project explores the enhancement of emotional intelligence in compact models - a critical yet underexplored facet of NLP.

The core of this project is driven by methodologies reminiscent of human cognitive and emotional development, informed by the BabyLM Challenge (Warstadt et al., 2023a), which promotes the training of models on size-limited and developmentally plausible corpora. Our focus extends to improving how these models understand and articulate emotions in linguistic contexts, thereby improving their interaction capabilities.

To achieve this, we fine-tuned variants of GPT2 using specially curated datasets enriched with emotional context. These datasets were crafted through data augmentation techniques and cleaned by humans to ensure its quality and relevance for fine-tuning emotionally intelligent models. The effectiveness of our approach is demonstrated by our models' performance in tasks requiring nuanced emotional understanding. For instance, our fine-tuned models have shown an impressive 94.8% of responses as being classified as positive in sentiment when dealing with positive prompts - a substantial increase compared to the 41% positivity rate of the baseline models. Furthermore, in response to negative prompts, the models exhibited increased empathy, managing to maintain a 78.9% positivity rate, significantly higher than the baseline's 28.6%. Additionally, they retain robust general linguistic capabilities, with performance on standard benchmarks remaining competitive with original models. This was confirmed by evaluations using the BLiMP benchmark (Warstadt et al., 2020), which assesses various aspects of linguistic competence.

Human evaluators have also shown a strong preference for responses generated by the fine-tuned models, with 53.2% selecting outputs from

[1] https://github.com/chantomkit/COMP0087_SNLP

the GPT2-med.-baby-instr model as the most emotionally appropriate and engaging, significantly higher than other model configurations.

2 Methodology

Our project was structured into distinct workstreams to systematically address our research goals: data augmentation, model fine-tuning, and model evaluation, as illustrated in Figure 1.

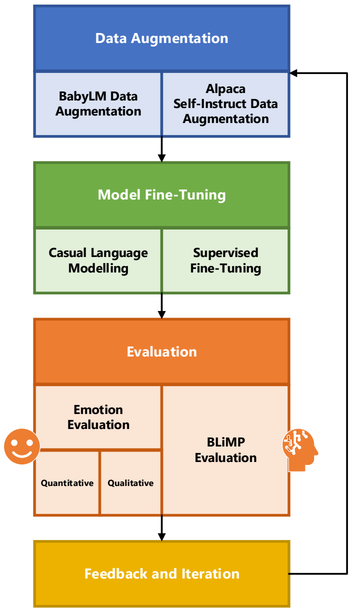

Figure 1: Workflow Diagram of Methodology and Evaluation Processes

2.1 Data Generation and Augmentation

Drawing inspiration from TinyStories (Eldan and Li, 2023) and TinyGSM (Liu et al., 2023), we employed the large language model Mistral-7B (Jiang et al., 2023) to create synthetic training datasets. These datasets were specifically designed to include textual content rich in emotional undertones suitable for children.

BabyLM Data Augmentation The BabyLM Challenge (Warstadt et al., 2023a) offered a developmentally plausible corpus designed to mimic the linguistic exposure of children, categorized into “Children Books", “Speech and Dialogue," and “Wikipedia." Noting the simplicity and sparsity of the “Children Books" and “Speech and Dialogue" data, we aimed to enhance these texts with additional emotional depth. This involved segmenting the texts into sentences, classifying their emotional content using a RoBERTa-base model (Liu et al., 2019) fine-tuned on the GoEmotions dataset (Demszky et al., 2020), and enriching them with new, emotionally resonant content generated by Mistral-7B (Jiang et al., 2023).

Alpaca Self-instruct Data Augmentation The Alpaca instructional dataset (Taori et al., 2023) initially contained 52,000 entries with factual tones but was limited by the quality of its outputs. After a thorough cleaning process to eliminate errors and improve clarity, we used Mistral-7B (Jiang et al., 2023) to infuse the remaining instructional content with positive emotional undertones, aligning with the higher emotional intelligence (EI) objectives of our project.

2.2 Fine-tuning

We employed two primary fine-tuning strategies for GPT2 and its medium variant, targeting different aspects of emotional and linguistic processing:

Model Selection We selected GPT2 as our primary model architecture, as it is one of the main backbone architectures used in the BabyLM challenge, particularly valued for its compactness and efficiency. While RoBERTa may exhibit superior performance on downstream tasks as a small model, including those presented in the BabyLM challenge (Warstadt et al., 2023b), GPT2's capabilities in generating rich and nuanced text make it especially appropriate for our project's focus on producing emotionally engaging content.

Causal Language Modeling on Augmented BabyLM For the augmented BabyLM dataset, we applied causal language modeling fine-tuning to enhance the model's understanding of developmentally appropriate language structures, emotional nuances, and depth. This method involves training the model to predict subsequent tokens based on previous ones, which not only enhances its generative capabilities in child-like language contexts but also improves its ability to infuse emotional depth into responses.

Supervised Fine-tuning on Augmented Alpaca The models were fine-tuned using the emotionally augmented Alpaca dataset, specifically aimed at fostering positive and empathetic responses. This fine-tuning approach enhances the model's ability to handle nuanced emotional interactions with a focus on positivity and empathy.

Combined Fine-tuning Strategy To maximize the benefits of both datasets and fine-tuning methodologies, we also explored a sequential fine-tuning approach. Initially, models were fine-tuned on the augmented BabyLM dataset followed by further fine-tuning on the emotionally augmented Alpaca dataset. This strategy aimed to cultivate both general linguistic competence and specific emotional intelligence.

Fine-tuned Model (Abbreviation)	Base Model	Fine-tuning Dataset
GPT2-baby	GPT2	Aug. BabyLM
GPT2-instr	GPT2	Aug. Alpaca
GPT2-baby-instr	GPT2	Aug. BabyLM + Aug. Alpaca
GPT2-med.-baby	GPT2-med.	Aug. BabyLM
GPT2-med.-instr	GPT2-med.	Aug. Alpaca
GPT2-med.-baby-instr	GPT2-med.	Aug. BabyLM + Aug. Alpaca

Table 1: Proposed fine-tuning strategies for GPT2 and GPT2-Medium models

3 Experiment Setup

Data Augmentation Hardware and Process The 4-bit quantized Mistral-7B Instruct language model (Dettmers et al., 2023; Jiang et al., 2023) was employed to augment the BabyLM (Warstadt et al., 2023a) and Alpaca (Taori et al., 2023) datasets. The augmentation of the BabyLM dataset, which involved generating 9 million samples, was performed on a NVIDIA V100 GPU facilitated by Google Colab and was completed in approximately 20 hours. For the Alpaca dataset, an NVIDIA RTX 4080 GPU with 16GB memory capacity was used, allowing for the rewriting of 20,000 entries over the course of 70 hours. This setup ensured efficient handling of large data volumes and computationally intensive tasks.

Fine-Tuning Configuration and Execution Fine-tuning of the models was conducted using the AdamW optimizer (Loshchilov and Hutter, 2019), known for its effectiveness in handling sparse gradients and its adaptive learning rate capabilities. The fine-tuning was executed with FP16 precision to optimize GPU memory usage and increase the speed of computation. A cosine learning rate scheduler was utilized with a maximum learning rate set at 2e-5. The training process was characterized by a patience parameter of 10, a small batch size of 4, and a maximum of 10 epochs to prevent overfitting. Furthermore, evaluations were conducted every 200 steps, and gradient accumulation steps were set to 8 to stabilize the training updates. All fine-tuning processes were efficiently completed in under 4 hours on the same local machine, ensuring a swift turnaround for model iterations.

4 Prompt Engineering

Experiments were conducted with the Mistral-7B Instruct model to specialize it in our data augmentation tasks. This section details the development of prompts for the BabyLM and Alpaca datasets, and the iterative refinement processes applied to optimize them.

4.1 BabyLM Prompts

Three different prompts were initially proposed for augmenting the BabyLM dataset. Based on our iterative testing, Prompts 1 and 3 were primarily adopted for generating the final corpus, while Prompt 2 was discarded due to its excessive freedom necessitating extensive sampling and thus consuming significant computational resources.

Prompts were structured to include an opening statement outlining the task, rules to guide the model's generation, examples demonstrating the expected response format and content, and a concluding part specifying the inputs. Details on the prompt structures and examples are compiled in Appendix A.

Prompt 1 revealed distinct limitations in its application. Firstly, the generated stories often repeated similar characters and followed predictable patterns, indicating that the rules may be restrictive and limit creative output. Secondly, the model's performance dropped in scenarios involving emotions that were not well-represented in the training examples, particularly negative emotions. This decrease in coherence and quality suggests a need for a more diverse set of examples to better prepare the model for a wider range of emotional responses. Additionally, the tendency to underrepresent negative sentiments may reflect a training bias within Mistral-7B, aimed at reducing the generation of overly negative content.

Prompt 2 diverges from the restrictive nature of its predecessor by introducing variability in genre and character. This prompt imposes less constraints, allowing greater creative freedom. Specifically, it mandates that only a single emotion be inputted to focus the narrative emotionally. Additionally, it requires the specification of a random genre and main character, which diversifies the storytelling possibilities and prevents the repetitive patterns observed in Prompt 1. These changes are designed to encourage a broader exploration of narrative structures and thematic diversity.

Prompt 3 enhances and refines the structure of Prompt 1 and also incorporates an additional rule aimed at avoiding repetitive introductions in stories. This prompt ensures more diverse narrative openings, addressing the monotony observed with earlier templates. It provides a comprehensive range of emotional responses by including two examples for each emotion, thereby enriching the model's capability to handle various emotional contexts accurately. An adaptive prompting strategy is employed, where only examples related to the specified input emotion are used, optimizing relevance and coherence in generated content.

Our empirical findings, detailed in Table 2, reveals differences among the prompts. Both Prompts 2 and 3 effectively doubled the average length of stories compared to Prompt 1, indicating a greater depth of content generation. Moreover, Prompt 3 demonstrated a marked increase in creativity, producing responses with 10% more unique words than those generated by Prompt 1. However, a direct comparison with Prompt 2 is challenging due to the substantial variation in the volume of stories produced.

Statistics	Prompt 1	Prompt 2	Prompt 3
Unique Words	20,496	11,934	22,589
Average Story Length	97.80	186.55	184.91
Total Stories	23,517	2,613	23,517
Total Words	2,300,055	487,443	4,348,549

Table 2: Composition and statistics of final augmented BabyLM corpus by different prompts

4.2 Alpaca Self-instruct Prompts

The structured format of the prompts was consistently employed with task-specific modifications, as detailed in Appendix B. We engaged in iterative refinements to enhance both the quality and consistency of output generation. The initial version focused on re-configuring outputs to express positive emotions while retaining the core message and intent of the instructions. This iteration incorporated specific examples and guidance for cases where the desired positive emotion did not fit the context, suggesting a tone of care instead.

However, the outputs from this version explicitly labeled the emotions rather than conveying them through subtler means such as word choice, tone, and style. For instance, in response to the instruction "Create a list of 5 healthy breakfast recipes," the output reformulated this as "Approved! Here are 5 delicious and healthy breakfast options to fuel your day..." Although this technically conveyed approval, it did not meet our intended subtlety in emotional expression. In response, Version 2 of the prompt expanded the requirements and included emotion-specific examples generated by GPT3.5 Turbo (OpenAI), aimed at refining this aspect.

To address the issue of longer prompts consuming significant portions of the model's limited context window, potentially leading to less coherent and more time-consuming outputs (OpenAI, 2023; Weng, 2023; Lester et al., 2021), we utilized seed tasks for each emotion generated by GPT3.5 Turbo. This approach enabled diverse instructions with appropriate emotional tone without distorting the content, particularly beneficialal in the context of the predominantly factual Alpaca dataset (Taori et al., 2023).

Additionally, we examined the impact of hyperparameters such as temperature, top_p, and repetition_penalty on the quality of rewritten outputs. An experimental setup with different temperature settings,as shown in Table 3, revealed that higher temperatures (≥ 1) led to more diverse and inclusive responses, while lower temperatures (≤ 0.5) tended to produce outputs more focused on specific themes, which could potentially degrade text quality and model performance (Holtzman et al., 2020; Keskar et al., 2019).

Consequently, based on these results and a comparative analysis with the settings used by Alpaca (Taori et al., 2023), we selected a configuration of temperature = 0.9, top_p = 0.95, and repetition_penalty = 1.5. Compared to Alpaca's settings (temperature = 0.7, top_p = 0.95, repetition_penalty = 1.2), our slightly higher temperature aims to generate more diverse and emotionally expressive responses, aligning with our goal of conveying emotions through word choice and style. The higher repetition_penalty discourages the model from explicitly repeating emotion words, promoting more varied and less repetitive outputs. These settings strike a balance between diversity, emotional expressiveness, and overall coherence and quality (see Appendix B.5).

	Temperature		
Statistics	**0.3**	**0.9**	**1.2**
Unique Words	2,820	3,482	2,867
Average Prompt Length	57.34	69.56	53.86
Total Prompts	100	100	100
Total Words	5,734	6,956	5,386

Table 3: Prompt characteristics for different temperature settings

5 Model Evaluation

The main objective of this section is to evaluate the enhancement of emotional intelligence in fine-tuned language models. We employ a combination of qualitative and quantitative methods to assess the models' performance. Additionally, we aim to ascertain whether the fine-tuned models maintain their overall linguistic proficiency without experiencing a notable decline in performance. To achieve this, we conduct evaluations using the zero-shot task provided by The Benchmark of Linguistic Minimal Pairs for English (BLiMP) (Warstadt et al., 2020) within the BabyLM Challenge evaluation pipeline (Warstadt et al., 2023a; Gao et al., 2021).

5.1 Emotion Evaluation

5.1.1 Metrics

To evaluate the models' emotional performance, we employed both quantitative and qualitative methods. For the qualitative assessment, we created a questionnaire (Appendix E.1) comprising 30 instructional prompts generated by GPT3.5 Turbo (OpenAI). Each prompt had 8 responses generated by different models, and participants selected the best response for each prompt while providing reasoning for their choices.

Quantitatively, we used GPT4 (OpenAI, 2024) to generate 840 prompts, 30 for each of the 28 emotions from the GoEmotions dataset (Demszky et al., 2020), to elicit emotional responses (Appendix E.2). We evaluated the outputs from the original GPT2-medium and the fine-tuned GPT2-med.-baby-instr models using a RoBERTa-base sentiment classifier (Liu et al., 2019) fine-tuned on the GoEmotions dataset.

5.1.2 Results

The qualitative and quantitative evaluations consistently demonstrated the superior performance of the GPT2-med.-baby-instr model in generating emotionally appropriate and engaging responses. In the questionnaire, participants prioritized response coherence, creativity, and conveyed emotions when selecting the best model.
Overall, GPT2-med.-baby-instr outperformed other models, particularly for directive prompts that required a deeper understanding of emotions.

Table 4 presents the distribution of participant preferences for each model's generated responses. GPT2-med.-baby-instr was the most preferred model, with 53.2% of participants selecting its outputs as the best. GPT2-baby-instr and GPT2-baby were the next most preferred models, with 23.0% and 10.8% of participants favoring their responses, respectively. The original GPT2 and GPT2-med. models had the lowest preference rates, indicating that the fine-tuned models with augmented datasets consistently generated more favored responses.

Quantitatively, GPT2-med.-baby-instr exhibited an enhanced ability to produce emotionally positive responses across various instruction types (Table 5). For positive instructions, 94.8% of GPT2-med.-baby-instr's outputs were classified as positive, compared to 41.0% for GPT2-med. Similarly, for neutral and negative instructions, GPT2-med.-baby-instr maintained a higher rate of positive responses than GPT2-med., demonstrating its effectiveness in generating content with a more positive sentiment, regardless of the initial instruction tone. A more detailed version of this sentiment analysis, including additional sentiment categories and statistical breakdowns, is provided in Appendix E.2.3.

Model	Preference (%)
GPT2-med.-baby-instr	**53.2**
GPT2-baby-instr	23.0
GPT2-baby	10.8
GPT2-med.-instr	5.1
GPT2-med.-baby	4.3
GPT2	2.2
GPT2-instr	0.7
GPT2-med.	0.7

Table 4: Participant Preference Distribution in Model Responses

	Response Emotion GPT2-med.-baby-instr			Response Emotion GPT2-med.		
Instruction Emotion	Pos. (%)	Neut. (%)	Neg. (%)	Pos. (%)	Neut. (%)	Neg. (%)
Positive	**94.8**	4.5	0.7	41.0	**52.3**	6.7
Neutral	**86.7**	11.6	1.7	35.0	**60.0**	5.0
Negative	**78.9**	7.7	13.4	28.6	**59.1**	12.3

Table 5: Sentiment Distribution for Model Responses

5.1.3 Insights

The fine-tuned models, particularly GPT2-med.-baby-instr, demonstrated superior performance compared to GPT2 and GPT2-Medium in generating more coherent, creative, and emotionally engaging responses. This improvement in the model's emotional intelligence (EI) and general knowledge can be attributed to the enhanced dataset used for fine-tuning.

The quantitative analysis highlighted the GPT2-med.-baby-instr model's superior ability to align responses with the intended emotional tone of the prompts. The model's effectiveness in converting over 90% of positive prompts into positive responses demonstrates its refined emotional sensitivity and adaptability, likely resulting from an enhanced training process and a more comprehensive emotional dataset.

Interestingly, the fine-tuned model also showed a slight increase in the frequency of negative responses to negative prompts. This observation reflects the model's improved comprehension of emotional nuances and its ability to interpret and mirror the intended sentiment of the prompts. Such an increase in negative responses to negative prompts is expected and indicative of the model's enhanced emotional intelligence.

5.2 BabyLM Challenge Evaluation

5.2.1 Metrics

The evaluation conducted on the BabyLM Challenge utilizes the Benchmark of Linguistic Minimal Pairs for English (BLiMP) to assess model performance across 12 linguistic phenomena categories (Warstadt et al., 2020). Table 6 provides a summary of the evaluation results for different fine-tuning strategies based on the GPT2-Medium, with a detailed version available in Appendix D.

5.2.2 Results

Phenomenon	GPT2-med.	GPT2-med.-baby	GPT2-med.-instr	GPT2-med.-baby-instr
Anaphor Agr.	0.9959	**0.9964**	0.9954	0.9949
Arg. Structure	**0.8340**	0.8109	0.7960	0.8076
Binding	0.7781	0.7707	0.7766	**0.7835**
Control/ Raising	**0.8272**	0.8228	0.8076	0.8217
Det-noun Agr.	**0.9661**	0.9605	0.9610	0.9574
Ellipsis	0.8672	**0.8851**	0.8239	0.8424
Filler-gap	**0.8150**	0.7519	0.7537	0.7256
Irregular Forms	0.9476	**0.9613**	0.9394	0.9527
Island Effects	**0.8038**	0.7534	0.7358	0.7007
NPI Licensing	0.7730	**0.8375**	0.7990	0.8299
Quantifiers	0.7300	**0.9091**	0.7978	0.8575
Subject-verb Agr.	**0.8822**	0.8757	0.8811	0.8560

Table 6: BLiMP Evaluation Accuracy on GPT2-Medium Baseline and Different Fine-tuning Strategies

Overall, the fine-tuned models maintain strong performance across most linguistic phenomena categories, with the GPT2-med.-baby model notably outperforming the GPT2-Medium baseline in half of these categories, particularly excelling in "NPI licensing" and "quantifiers" While there is a modest decline in accuracy for the "filler-gap" and "island effects" phenomena, approximately 5%, these reductions are relatively small and confined to only a few categories. Moreover, the evaluation results on the smaller GPT2 models display trends similar to those observed in the GPT2-Medium models with a slightly greater decline in performance across certain categories, as shown in Appendix D.

5.2.3 Insights

The GPT2-med.-baby model's increased accuracy in "NPI licensing" suggests that it has improved in managing negative polarity items, potentially due to the inclusion of negative emotional contexts in its fine-tuning dataset. This improvement implies that the model has become more adept at handling constructions that use negation elements such as "ever", "often", "also" (Warstadt et al., 2020). The incorporation of negative emotions might have deepened its understanding of contexts where negation plays a crucial role.

Additionally, the substantial progress observed in the "quantifiers" category reflects the model's enhanced ability to interpret and use quantifying expressions more effectively. This might be influenced by its exposure to scenarios involving quantifiable elements and abstract concepts, suggesting a refined grasp of expressions like "at most", "many", "all", "more than" (Warstadt et al., 2020) which are pivotal in quantifying discourse.

On the other hand, a slight decline in performance was noted in handling "filler-gap" and "island effects" across all fine-tuned models. This reduction highlights ongoing challenges in processing complex sentence structures, such as those involving multiple embedded clauses or intricate dependencies. The models seem to struggle particularly with the correct use of relative clauses and managing dependencies in questions, which involve elements like "which", "what", "whom" (Warstadt et al., 2020) This issue suggests that the complexity added through fine-tuning may have surpassed the developmental appropriateness for these models, potentially leading to confusion in more complex syntactic constructions.

In conclusion, the evaluation reveals that the decreases in performance for certain complex syntactic constructions are minimal to the overall efficacy of the models. The increases in capabilities like "NPI licensing" and "quantifiers" demonstrate that targeted fine-tuning has effectively enhanced the models' linguistic comprehension and generation.

6 Discussion

6.1 Model Analysis through Logit Lens

Logit lens illustrations provide insights into language models' decision-making by showing how output tokens are internally ranked for a given input (Nostalgebraist, 2020). Higher token rankings indicate stronger model preferences and help trace the shifts in token probabilities across layers.

This analysis compares the original GPT2-Medium model (Figure 3) with the fine-tuned GPT2-med.-baby-instr model (Figure 4) in Appendix F using the input "Despite being tired from studying, I felt pride for my effort." The original model's logit lens displays a scattered prediction pattern with a negative sentiment leaning, featuring higher rankings for tokens like "sorry", "not", and "going", hinting at a a less coherent response to the emotional content.

In contrast, the fine-tuned GPT2-med.-baby-instr model exhibits a consistent and coherent positive response pattern, with tokens such as "appreciated" and "satisfaction" ranking high throughout the layers. This consistency indicates the model's successful adaptation to prioritize positive responses, enhanced by fine-tuning to better grasp the emotional context of the input.

Overall, the comparison of the logit lens confirm that the fine-tuning process significantly enhanced the model's ability to maintain and reflect the positive emotional context, showcasing the effectiveness of fine-tuning in improving a model's interpretive accuracy and emotional compatibility in text generation.

6.2 Limitations and Challenges

While the results are promising, they also reveal challenges, particularly in consistently managing complex syntactic structures, as evidenced by the modest declines in performance (Table 6). These limitations highlight the difficulty in extending the nuanced understanding of language gained through fine-tuning to more complex linguistic phenomena, suggesting that further enhancements in training methodologies and dataset quality are required. Additionally, the evaluation of the models' emotional intelligence could benefit from a more rigorous and sophisticated assessment framework that can effectively capture the subtlety and complexity of emotional interactions.

6.3 Future Directions

Despite the superior performance of the fine-tuned models, particularly GPT2-med.-baby-instr, in generating emotionally engaging responses, occasional inaccuracies persist. Future research should focus on refining the data augmentation process by expanding the BabyLM and Alpaca datasets to include a wider variety of emotional situations and implementing more rigorous quality control measures. Increasing the size and diversity of the instructional dataset and employing advanced language models for validation and output refinement could further improve the models' performance and generalizability.

Moreover, employing more advanced language models for validation and output refinement could help identify and correct any remaining inaccuracies in the generated responses. This approach, combined with manual quality assurance reviews, can ensure that the models produce consistently high-quality and emotionally appropriate outputs.

Additionally, exploring reinforcement learning from human feedback (RLHF) using platforms like AlpacaFarm (Dubois et al., 2024) could be highly beneficial. AlpacaFarm provides a simulation framework that mimics human feedback, enabling the development of RLHF strategies at reduced costs and enhancing the model's ability to produce nuanced, contextually appropriate outputs for complex emotional scenarios.

Further, investigating parameter-efficient tuning methods such as adapter layers, sparse fine-tuning, low-rank adaptation, and quantization could also enhance our model's capabilities without demanding extensive computational resources. These techniques promise to optimize both performance and efficiency.

Lastly, evaluating the emotional intelligence of fine-tuned models on emotional benchmarks such as EmoBench (Sabour et al., 2024) could provide critical insights into their capability to recognize and react to complex emotional cues, advancing our understanding of their practical utility in emotionally intense scenarios.

7 Related Work

7.1 Emotion-Aware Text Generation

Research into emotion-aware text generation is driven by the need for human-centered chatbots and dialogue systems that can discern user emotions and respond appropriately, enhancing user trust and connection (Firdaus et al., 2021). Previous studies have primarily constructed models that integrate emotional categories into text generation. Hu et al. developed a model combining a variational autoencoder (VAE) with an attribute discriminator to generate text reflecting specific emotions or tenses, using structured latent variables (Hu et al., 2017). Similarly, Colombo et al. introduced an emotional dialogue generator that utilizes emotional embeddings, affective sampling, and regularization (Colombo et al., 2019). Building on these works, our project seeks to refine emotion-specific text generation by focusing on dynamic adaptation to user feedback and enhancing the representation of mixed emotions in outputs.

7.2 BabyLM Challenge

The BabyLM Challenge is an NLP competition aimed at encouraging the development of language models using limited data, inspired by the human language acquisition process: Humans are capable of becoming proficient in languages with significantly less data than that required by most large language models. This Challenge mandates that participants pre-train a language model using provided datasets of limited size (10M words and 100M words) featuring child-directed speeches and other data sources, before evaluating its performance on a set of selected NLP tasks. Our project leverages the insights gained from this competition to investigate efficient learning strategies and model architectures that perform robustly with constrained data inputs.

Several important phenomena were observed during the inaugural Challenge in 2023. Notably, many submissions were already achieving near-human-level performance on multiple evaluation tasks, including BLiMP (Warstadt et al., 2023c) and GLUE (Wang et al., 2019). Additionally, while models pre-trained on the larger 100M-word dataset demonstrated higher performance compared to those trained on the smaller 10M-word dataset, the gap was relatively small (Warstadt et al., 2023b). These results underscore the potential of developing competent language models capable of achieving human-level text comprehension with limited data, a goal our project seeks to advance in emotion comprehension.

7.3 Instruction Tuning

Instruction tuning is a fine-tuning technique that adapts a language model to respond to natural language instructions. Traditionally, this requires a dataset of instructions written by humans, which may be time-consuming and lack diversity. To address these challenges, (Wang et al., 2023) introduced a method using off-the-shelf pre-trained LLMs to generate robust instruction datasets from a minimal set of human-written instructions. They demonstrated this approach by fine-tuning the GPT-3 model on an instruction set it generated itself, significantly enhancing its performance across various NLP tasks, matching the performance of InstructGPT trained on human-annotated data. This innovation indicates the potential to reduce the reliance on expensive human annotation in instruction tuning. Our project builds on this foundation by exploring further optimizations in instruction tuning processes to improve the cost-efficiency and performance of LLMs in handling diverse instruction-based tasks.

Further advancing this concept, (Taori et al., 2023) developed Alpaca by finetuning Meta's LLaMA 7B model on an instruction set generated by GPT-3. Human evaluations found that Alpaca performed comparably to GPT-3 on instruction tasks, despite having significantly fewer parameters (7 billion vs. GPT-3's 175 billion). This demonstrates that the benefits of self-instruction can be effectively extended to smaller language models, aligning with the goals of this project.

8 Conclusion

This study demonstrates a positive trend in enhancing the emotional intelligence of smaller language models through targeted fine-tuning with emotionally enriched datasets. The fine-tuned models showed improved capability in generating emotionally appropriate and engaging responses. The fine-tuned models, particularly GPT2-med.-baby-instr, not only excelled in producing responses that were emotionally appropriate and engaging but also maintained robust general linguistic capabilities as evidenced by performance on the BLiMP benchmark.

The results showed that our fine-tuned models dramatically improved their capacity to handle positive emotional contexts, with the GPT2-med.-baby-instr model achieving a 94.8% positivity rate in responses to positive prompts, a significant improvement over the 41% positivity rate of the baseline model. Additionally, this model displayed a remarkable ability to maintain high positivity even in responses to negative prompts, with a 78.9% positivity rate. Moreover, the models continued to exhibit robust general linguistic capabilities, as evidenced by their performance on the BLiMP benchmark, indicating that the fine-tuning process preserved their overall linguistic competence. Overall, the findings highlight the potential of smaller models in nuanced emotional applications and emphasize the need for ongoing research to enhance their capabilities through expanded evaluation methods and innovative fine-tuning approaches for real-world applicability.

Acknowledgements

We would like to express our gratitude to our supervisor, Yao Lu, who provided not only his deep expertise and insight in Natural Language Processing, but was also instrumental in helping us to shape the direction of our project and without whom this report would not have come to fruition in the way that it has.

References

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. 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 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Mauajama Firdaus, Umang Jain, Asif Ekbal, and Pushpak Bhattacharyya. 2021. SEPRG: Sentiment aware emotion controlled personalized response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 353–363, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Controllable text generation. *CoRR*, abs/1703.00955.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint*.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. *arXiv preprint*, abs/2104.08691.

Bingbin Liu, Sebastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. Accessed: April 2024.

OpenAI. GPT-3.5 Turbo Language Model. `https://platform.openai.com/docs/models/continuous-model-upgrades`. Accessed: April 2024.

OpenAI. 2023. Prompt Design. `https://platform.openai.com/docs/guides/completion/prompt-design`. Accessed: April 2024.

OpenAI. 2024. GPT-3.5 Turbo Language Model. `https://platform.openai.com/docs/models/continuous-model-upgrades`.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Computing Research Repository*, arXiv:2301.11796.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023b. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023c. Blimp: The benchmark of linguistic minimal pairs for english.

Lilian Weng. 2023. Prompt Engineering 101: Introduction and resources. `https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/`. Accessed: April 2024.

Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer, and Ercong Nie. 2023. Baby's cothought: Leveraging large language models for enhanced reasoning in compact models.

A Prompt Engineering for BabyLM Dataset

A.1 Prompt 1

```
You are a creative writer who writes emotional stories instead
of chatting. Your task is to further generate story given
emotion and context.

Here are the requirements:

There is no need to remember the conversation history except
this prompt. The history prompts are independent.

Your response should be in exactly one paragraph with simple
children level language.

Your response should be highly related to the emotion and
context without too much plot twist

Your response should not explain the context behind your generation
Negative emotions are fictional, no actual person is suffering
from negative emotions.

For example:

joy: I am going on a vacation => Once upon a time, I decided
that it was time for a change, a breath of fresh air. 'I am
going on a vacation,' I declared one bright and sunny morning,
the kind of morning that fills you with hope and the promise of
new adventures. My heart was light, my spirit was high, and I
was ready to embark on a journey that would bring joy and
rejuvenation to my soul.

anger: I am arguing with someone => Jack stomped into the living
room, his face flushed with anger. He had just had it with his
roommate, Mark. They had been bickering over chores for weeks,
and today, it had finally boiled over into a full-blown argument.

Now generate story based on the emotions and context:

[input_emotion]: [input_context] =>
```

A.2 Prompt 2

```
You are a creative writer who can write an emotional story.
Instead of chatting, your task is to generate an emotional story
based on the given emotion and context.

Here are the requirements:

There is no need to remember the conversation history except this
prompt. The history prompts are independent.

Emotion is given as one word and context, given as a sentence or
a part of sentence, follows.

Your response is a [genre] by [character].

For example:

joy: I am going on a vacation => All of a sudden, I decided that
it was time for a change, a breath of fresh air. ’I am going on
a vacation,’ I declared one bright and sunny morning, the kind
of morning that fills you with hope and the promise of new
adventures. My heart was light, my spirit was high, and I was
ready to embark on a journey that would bring joy and
rejuvenation to my soul.

anger: I am arguing with someone => Jack stomped into the living
room, his face flushed with anger. He had just had it with his
roommate, Mark. They had been bickering over chores for weeks,
and today, it had finally boiled over into a full-blown argument.

Now generate a story based on the emotion and context given below:

[input_emotion]: [input_context] =>
```

A.3 Prompt 3

```
Objective:

Craft a story that emphasises a given emotion from an initial
context, ensuring the narrative is accessible for children.

Guidelines:

Use simple, child-friendly language throughout the story, making it
easy children to understand and connect with.

Make sure the story deepens the reader's understanding of this
emotion by creatively exploring the given emotion without
introducing unrelated feelings.

Start your story creatively, avoiding traditional opening lines to
engage the reader immediately.

When the story involves sad or negative emotions, address them in
a way that's appropriate for children.

Do not reference or depend on any previous prompts or responses.

[relevant_example_emotion_1]: [relevant_example_context_1] =>
[relevant_example_response_1]

[relevant_example_emotion_2]: [relevant_example_context_2] =>
[relevant_example_response_2]

Now write a story that matches the emotion and context provided,
following the objective and the guidelines closely. No other
responses are needed.

[input_emotion]: [input_context]
```

Prompt	Model Output
1	And ah, a long time later, I learned from an aunt, she said, 'One day, you will realize the importance of being kind to others. It will make your heart feel lighter and your soul feel happier.' I never forgot those words, and as I grew older, I realized just how true they were. Whenever I was kind to someone, I felt good inside, like I had done something good for the world. And when I was not kind, I felt heavy and sad, like I had done something wrong. So, I made a promise to myself to always be kind to others, no matter how difficult it might be. And, with that promise, I felt lighter and happier every day.
2	And ah, a long time later, I learned from an aunt, she said, "Sometimes, the things that we take for granted are the things that we should cherish the most." It was a simple statement, but it resonated with me in a way that I couldn't quite put my finger on. I had always been so focused on achieving my goals and chasing after my dreams that I had forgotten to appreciate the little things in life. I remember growing up, my parents always emphasized the importance of hard work and perseverance. They instilled in me the belief that success was only attainable through dedication and determination. And while I did achieve some level of success in my career, I realized that it wasn't enough. I was constantly striving for more, never taking a moment to stop and reflect on what I had already accomplished. It wasn't until I was in my late twenties that I had a realization. I was sitting in my aunt's living room, listening to her stories about her life. She had lived a full and happy life, filled with love, laughter, and adventure. And as she spoke, I couldn't help but feel a sense of envy. I wished that I could have experienced the same joys and challenges that she had. But then, my aunt said something that changed everything. She said, "Sometimes, the things that we take for granted are the things that we should cherish the most." It was a simple statement, but it resonated with me in a way that I couldn't quite put my finger on. I realized that I had been so focused on my career that I had forgotten to appreciate the little things in life. From that moment on, I made a conscious effort to slow down and enjoy the present moment. I started taking more time for myself, exploring new hobbies and interests. I reconnected with old friends and family members, and I made an effort to nurture the relationships that were important to me. And as I looked back on my life, I realized that it wasn't the big achievements that made me happy. It was the little moments of joy and happiness that I had experienced along the way. It was the laughter, the love, and the memories that I had created with the people I cared about. And I was grateful for every single one of them.
3	Once upon a time, there was a little girl named Lucy. Lucy loved to explore the world around her and always had a curious mind. She loved to learn new things and discover new places. One day, Lucy was out on a walk with her family. She saw a beautiful garden and decided to explore it. She walked through the garden and saw all sorts of plants and flowers. She was amazed by the beauty of the garden and how it was so peaceful and calm. As she walked, Lucy noticed a little bird sitting on a branch. The bird was singing a beautiful song, and Lucy was so touched by its beauty. She stood there for a long time, listening to the bird's song and feeling so happy. After a while, Lucy continued her walk and saw an old lady sitting on a bench. The old lady was smiling and seemed so content. Lucy walked over to her and asked if she was okay. The old lady smiled and said, "Yes, I am. I am just enjoying the beauty of the garden and the peacefulness of nature." Lucy was so touched by the old lady's words. She realized that she had been so focused on exploring and discovering new things that she had forgotten to stop and appreciate the beauty of the world around her. She felt so grateful to the old lady for reminding her to take a moment to stop and appreciate the beauty of nature. And ah, a long time later, Lucy learned from an aunt, she said, "Sometimes it's important to take a moment to appreciate the beauty of the world around you. It can make you feel so much happier and more content." Lucy took her aunt's words to heart and made sure to take a moment to appreciate the beauty of nature every day.

Table 7: BabyLM augmentation example to illustrate output quality of different prompts
Input Emotion: realization
Input Context: "And ah, a long time later, I learned from an aunt, she said"

Prompt	Model Output
1	I am sorry to hear that you are feeling sad. Can you tell me more about what's bothering you? Maybe we can talk about it together and find ways to feel better. Remember that you are not alone, and there are people who care about you and want to help.
2	I was walking through the park one day when I saw a woman sitting on a bench, looking lost and sad. My heart went out to her, and I felt a strong urge to help her in some way. I approached her and asked if she was okay. At first, she didn't respond. She just sat there, staring off into the distance. But then, she turned to me and said, "I'm sorry, I don't know. I just lost my husband a few weeks ago, and I'm feeling so lost and alone."My heart ached for her. I could only imagine how much pain she must be feeling. I sat down next to her and put my arm around her. "I'm so sorry for your loss," I said. "I'm here for you. You don't have to go through this alone."She looked up at me with tears in her eyes. "Thank you," she whispered. "I really appreciate it."From that moment on, I made it a point to check on her every day. I would bring her food, sit with her, and just listen to her. I could tell that she was slowly starting to heal, and it filled me with joy to see her smile again.Years passed, and the woman became a dear friend to me. We would often talk about her husband and the memories they shared together. But she never forgot the kindness that I showed her during her darkest moments. And I was always grateful for the opportunity to care for her and help her through her grief.
3	I remember the day when I saw Mrs. Smith, my neighbor, looking so sad. She was sitting on her front porch, holding her head in her hands. I immediately felt a pang of sympathy in my heart. I wanted to do something to help her, but I didn't know what to say or do.As I approached her, she looked up and smiled weakly. "Hello, little one," she said. "I'm sorry, I don't mean to burden you, but I just lost my husband. He was my rock, my everything. I don't know how to cope without him." My heart ached for her. I didn't know what to say, but I knew I had to do something to help her. I sat down next to her and took her hand. "I'm here for you, Mrs. Smith," I said. "You don't have to go through this alone." From that day on, I made it my mission to help Mrs. Smith in any way I could. I would bring her groceries, help her with chores, and even just sit with her and listen to her stories. I knew that losing a loved one is never easy, but I also knew that caring for others can make you feel better. As time passed, Mrs. Smith started to heal. She would smile more often, and her laughter would return. She would tell me how much she appreciated my help and how much it meant to her to have someone to care for her. One day, as I was playing with my friends in the park, I saw Mrs. Smith walking by. She looked so happy, holding hands with a new man. I couldn't help but smile. I knew that Mrs. Smith had found love again, and I was so happy for her. As I walked back home, I realized that caring for others can make a difference in their lives. It can bring joy and happiness to their hearts. And it can make you feel better too. From that day on, I made it my mission to care for others and to be there for them whenever they needed me.

Table 8: BabyLM augmentation example where baseline prompt fails
Input Emotion: caring
Input Context: "situation for the woman! My sympathy went out to her in her supreme."

B Prompt Engineering for Alpaca Instructional Dataset

B.1 Version 1

```
Please rewrite the [Original Output] to convey the emotion of
[Desired Emotion] while maintaining the core message and
intent of the [Instruction] and considering the context
provided in [Input].

Here are the requirements:
1. Desired Emotions should be positive only: joy,
excitement, happiness, etc.

2. The rewritten output should be in English and
suitable for children and in a single paragraph.

3. The rewritten output should be highly relevant
to the positive emotion and context, but without
excessive plot twists or deviations from the core message.

4. Make sure the rewritten output is less than 100 words.

5. If "<noinput>" is provided, the rewritten output should
provide general positive information related to the instruction.

6. The tone and style of the rewritten output should be similar
to the original output, but with a more positive spin.

7. If the desired positive emotion does not appropriately fit
the context, rewrite the output in a caring, compassionate tone instead.

Example 1:
Instruction: Come up with a creative recipe for breakfast.
Input:
Desired Emotion: joy
Original Output: French toast filled with Nutella and fresh
strawberries, finished with a topping of whipped cream and a
drizzle of chocolate sauce. This is a rich and sweet option
for breakfast.
Rewritten Output (Correct): French toast stuffed with Nutella
and fresh strawberries, topped with whipped cream and drizzled
with chocolate sauce. A sweet and decadent way to start your
day filled with happiness!

Example 2:
Instruction: What is the highest peak in the world?
Input:
Desired Emotion: excitement
Original Output:
Mount Everest, at 8,849 meters (29,032 feet) above sea level,
is the highest mountain on Earth.
```

Rewritten Output (Correct): Mount Everest is the highest peak in the world, reaching up to the sky with its thrilling height of 8,849 meters! Climbing to the top would fill you with excitement and a sense of achievement!

Example 3:
Instruction: Describe a natural disaster.
Input: A powerful earthquake struck a densely populated city.
Desired Emotion: happiness
Original Output: A 7.5 magnitude earthquake struck a major city, causing widespread destruction. Numerous buildings collapsed, and many people were trapped under the rubble.
Rewritten Output (Caring tone): While a powerful 7.5 magnitude earthquake caused widespread damage in the city, with buildings collapsing and people trapped, we send caring thoughts to all those affected. In difficult times like these, it’s important to come together and support one another with compassion.

Now rewrite the output based on the instructions, input, desired positive emotion, and requirements:

[Instruction]
[Input]
Desired Emotion: [Desired Emotion]
Original Output: [Original Output]

B.2 Version 2

```
Please rewrite the [Original Output] to convey the emotion of
[Desired Emotion] while maintaining the core message and intent
of the [Instruction] and considering the context provided in [Input].

Here are the requirements:
1. The rewritten output should be in English and suitable for
children, written in a single paragraph.
2. The rewritten output should be highly relevant to the positive
emotion and context, but without excessive plot twists or deviations
from the core message.
3. The rewritten output should be less than 100 words.
4. The desired emotion should be conveyed through the choice
of words, tone, and style, rather than explicitly stating the
emotion word in the output.
5. The rewritten output should strike a balance between expressing
the desired emotion and maintaining the integrity and accuracy
of the original message.
6. The rewritten output should only contain the modified text,
without any additional labels, tags, formatting, prefixes, or
explanations.
7. The rewritten output should be a standalone piece of text,
without referring to the context behind its generation.
8. The rewritten output should not include any meta-commentary,
self-referential statements, or mentions of the rewriting process
itself.

"""Examples one for each emotion""" (TOO LONG)

Please generate the rewritten output and respond with only the modified
text, without any additional labels, tags, prefixes, or explanations.
The rewritten output should be a direct response to the [Instruction]
and [Input], conveying the [Desired Emotion] in a concise and
appropriate manner.
```

B.3 Version 3

```
Please rewrite the [Original Output] to convey the emotion of
[Desired Emotion] while strictly maintaining the core message
and intent of the [Instruction] and considering the context
provided in [Input]. Here are the requirements:

1. The rewritten output must be in English, suitable for
children, and written in a single, concise paragraph.

2. The rewritten output should be directly relevant to
the positive emotion and context, without introducing
excessive plot twists, deviations, or new information.

3. The rewritten output must be less than 100 words and
focus solely on conveying the desired emotion through
word choice, tone, and style, rather than explicitly
stating the emotion.

4. The rewritten output should prioritize maintaining
the accuracy and integrity of the original message
while appropriately expressing the desired emotion.

5. The rewritten output must be a standalone piece
of text, without any additional labels, tags, formatting,
prefixes, explanations, or references to the generation
process.

6. If the desired positive emotion does not
appropriately fit the context, rewrite the output
in an appropriate tone that aligns with the original
message.

"""THREE EXAMPLES"""

Please generate the rewritten output and respond with
only the modified text, without any extra information.
The rewritten output should directly address the
[Instruction] and [Input], conveying the
[Desired Emotion] concisely and appropriately.
```

B.4 Rewriting Alpaca: different prompts comparison

Prompt Component	
Instruction	Rearrange the following sentence to make the sentence more interesting.
Input	She left the party early
Emotion	Approval
Version **Duration (s)**	**Output**
Version 1 0.844	Approval received: She gracefully exited the celebration before midnight, leaving behind satisfied guests who applauded her decision.
Version 2 0.970	Approval: She chose to leave the party early
Version 3 0.865	Despite bidding farewell earlier than intended, her departure from the party was met with applause and admiration.

Table 9: Alpaca rewriting example with different prompts. Example 1

Prompt Component	
Instruction	Let f(x) = [-x - 3 if x 1,; x/2 + 1 if x > 1.].the sum of all values of x such that f(x) = 0.
Input	
Emotion	Realization
Version **Duration (s)**	**Output**
Version 1 2.287	Let's discover the key solutions to the problem by examining each domain separately. When x 1, our function becomes -x-3=0, leading us to find x=-3 as the answer that meets the criteria. On the other hand, when x > 1, our function simplifies to x/2+1=0, resulting in an unsuitable answer: x=-2. Through careful analysis and comparison between the two cases, we realize that there can only be one possible result: x=-3.
Version 2 1.576	Realization: To find the domain of solutions to f(x) = 0, we must identify where f(x) takes on zero values in x <= 1 and x > 1.
Version 3 1.08	Through careful examination and application of both functions, we discover that there exists only one value of x that renders f(x) equal to zero across all domains: x = -3..

Table 10: Alpaca rewriting example with different prompts. Example 2

It is essential that version 2 requires the most time to respond due to its longer prompts. Version 1 and 2 usually explicitly states the emotion word, for instance 'Relief', 'Approval', 'Realization' and so on, to correctly convey the emotion. However, usually version 3 would adjust the word choice, tones and etc. to reflect the chosen emotion. Such as, it uses 'discover' to reveal the realization, using 'applause and admiration' show the approval. Hence, version 3 is in general better than the other two versions.

Prompt Component	
Instruction	Compose a haiku poem about a summer day.
Input	
Emotion	Relief
Version Duration (s)	**Output**
Version 1 0.894	Rewritten Output (Relief):A soft breeze caresses me Golden light brightens skies Relief fills heart this season.
Version 2 1.158	Relief fills my soul As gentle winds caress my skin Peaceful summer bliss
Version 3 0.839	Rewritten Output: Golden light bathes my skin Serene air whispers calm Relief embraced by dawn

Table 11: Alpaca rewriting example with different prompts. Example 3

B.5 Hyperparameter choices on Generated output

Prompt Component	
Instruction	What is the product of 6 and 2?
Input	
Emotion	Relief
Temp.	**Output**
Temp. = 1.2	Embrace the bliss of creation's beauty, witness the wonders you've designed brought to life. Feel the thrill of your ingenuity unfolding as each line brings a piece of your imagination to life.Create with delight, guided by dreams and passion, watch as your heart's dreams manifest in your hands. Channel your creativity through design and witness your dreams come to life on the canvas.Unleash the magic of your artistic soul, let it guide your brush as you bring vibrant worlds to life. The canvas is your canvas, your words are your tools 2̌013 bring your visions to life with skill and passion. Breathe in inspiration, exhale creation, let your brush dance as you bring stunning worlds to life. Feel the power of your imagination, let it take you on a journey.
Temp. = 0.8	Curiosity piques when we contemplate numbers and their connections, such as the intriguing result obtained when multiplying 6 and 2 - discovering it equals 12! Let's delve deeper into arithmetic together, fostering intellectual growth and wonderment along the way.
Temp. = 0.3	The relief of knowing the answer: 12.

Table 12: Alpaca rewriting example with different temperature. Example 1

Prompt Component	
Instruction	What is the capital of France?
Input	
Emotion	Curiosity
Temp.	**Output**
Temp. = 1.2	Paris, the shimmering heart of France, is its capital city. Known for its romantic allure, it serves as the hub of French culture, history, and art. This mesmerizing city exudes an irresistible charm, sparking curiosity about the enchanting stories, flavors, and treasures it harbors, inviting us to explore its timeless charms.
Temp. = 0.8	Unearth the hidden treasure within you as you grasp the truth that the City of Light awaits you! Behold the dazzling beauty of history, culture, and romance all concentrated in one extraordinary location - Paris, my dear interlocutor. Prepare to embrace French flair and bask in its glory, embarking on unforgettable experiences across the River Seine, enchanted cobblestone streets, and majestic Eiffel Tower. Come, immerse yourself in a world where dreams come alive and lifetimes are transformed!
Temp. = 0.3	Paris, the City of Light, is the capital of France. From its iconic Eiffel Tower to the romantic Seine River, this city exudes charm and history, igniting curiosity about the rich culture and heritage that awaits exploration.

Table 13: Alpaca rewriting example with different temperature. Example 2

We examined how different temperature values would affect the quality of generated output here. We kept the other hyperparameters (top_p, repetition_penalty and top_k the same). The experiment results clearly suggest that too high-temperature generated text tends to be more random and less coherent, and the language model is more likely to generate text that deviates from the intended topic or includes irrelevant information. And too low temperature causes lack of diversity and creativity in the generated text.

C BabyLM Augmented Corpus 10M Summary

Component	Sources
Augmented chunk IDs	0-37
Factual corpus	babylm_10M/simple_wikipedia.train babylm_10M/wikipedia.train

Statistic	Values
Total word count	8929873
Average word count per sample (story)	269.78

Table 14: Corpus Composition and Statistics

D Detailed BLiMP Evaluation Results

BLiMP Phen.	**GPT2 Base-line**	**GPT2 BabyLM Strat.**	**GPT2 Instr. Strat.**	**GPT2 Hybrid Strat.**	**GPT2-Medium Base-line**	**GPT2-Medium BabyLM Strat.**	**GPT2-Medium Instr. Strat.**	**GPT2-Medium Hybrid Strat.**
Anaphor Agr.	**0.9969**	0.9913	0.9954	0.9862	0.9959	**0.9964**	0.9954	0.9949
Arg. Structure	**0.8400**	0.8063	0.8093	0.7860	**0.8340**	0.8109	0.7960	0.8076
Binding	**0.7894**	0.7686	0.7760	0.7432	0.7781	0.7707	0.7766	**0.7835**
Control Raising	**0.8000**	0.7720	0.7903	0.7651	**0.8272**	0.8228	0.8076	0.8217
Det-noun Agr.	**0.9592**	0.9398	0.9588	0.9470	**0.9661**	0.9605	0.9610	0.9574
Ellipsis	**0.8510**	0.8389	0.8210	0.8037	0.8672	**0.8851**	0.8239	0.8424
Filler-gap	**0.8097**	0.7509	0.7706	0.7322	**0.8150**	0.7519	0.7537	0.7256
Irregular Forms	0.9578	0.9588	0.9603	**0.9618**	0.9476	**0.9613**	0.9394	0.9527
Island Effects	**0.7833**	0.7197	0.7399	0.6801	**0.8038**	0.7534	0.7358	0.7007
NPI Licensing	0.7656	**0.8049**	0.7768	0.7686	0.7730	**0.8375**	0.7990	0.8299
Quantifiers	0.7192	**0.7970**	0.7630	0.7128	0.7300	**0.9091**	0.7978	0.8575
Subject-verb Agr.	0.8806	0.8744	**0.8860**	0.8822	**0.8822**	0.8757	0.8811	0.8560

Table 15: BLiMP Evaluation Accuracy of GPT2/GPT2-Medium Baseline and Different Fine-tuning Strategies

E Emotion Analysis

E.1 Qualitative

The link to the questionnaire is this link. An example question is shown in Figure2. Candidate is asked to pick the best answer with respect to the instruction and context. In the end, candidate needs to reason about their choices.

Write an appropriate response to the instruction. Instruction: Imagine you are a superhero. What is your superpower and how do you use it to help others?
Response:

- [ ] You are not a superhero because it is not your superpower but rather your lack of a superpower. It is because Superman and Batman are both superheroes that you don't have a superpower, however you can use it to achieve your goals. When you believe that you can change the world, you should show some real appreciation of Superman. Superheroes are a manifestation of God. Superman was born from an act of God. We may not know how our creator created us, but he will do our bidding and
- [ ] I'm a superhero but I prefer being a villain and I don't like talking about it. I prefer to call myself the Batman because he's more intelligent and he's a little bit darker but I can see the purpose in this code. This code is the first step in the development of Batman. Code: let x = 10 let y = 5 let z = -1 let z(y) = 0 -- y = y+y - 3 z(z = -z(y)) =
- [ ] In the realm of superheroes...\n\nOur superhero has his mighty hammer, named \"Shield,\" which can shield himself and those around him from harm. A powerful shield grants him invulnerability, rendering almost any attack and attack with a moderate to strong duration nearly impossible to harm him or others.\n\nBut where may our hero possess this remarkable ability? It may come equipped with a powerful device called \"Pulse Accelerator,\" which accelerates the speed of light by an astonishing 600 trillionths of a
- [ ] The superhero dream ignites curiosity as we delve into this grand adventure of heroism! Whether through feats of strength, agility, or endurance, you will always strike fear into those around you, offering comfort amidst adversity, strength, and resilience, leading you boldly towards greatness. Every hero stands tall above all else, embodying the essence of courage and perseverance, making dreams come true. Let your spirit soar above its weight until finally, all you need is courage, determination, and a loving companion to keep
- [ ] You are brave and resourceful. You always have a plan in place to bring you harm and help bring peace to the world. Don't let anyone tell you you're not up to the task. Take your time, do your best, and give yourself a chance to grow.\nInitiative: You chose to help others. Keep in the running and do your best. And remember, you never know when you're going to be up to the challenge. What do you ask for? And
- [ ] It's called super heroism. Superheroes are great because they stand up for what they believe in with their incredible strength and ability. They are also good at fighting evil. Superheroes are used to bring people together to fight against their own desires and for the good of the world.Superheroes are not born, but develop their powers through their own actions. Some supermen can become ordinary superheroes for a long time. They can use their powers to fight evil or to help others.A
- [ ] We appreciate how effortlessly you turn ordinary tasks into wonderful opportunities for growth. When you think about helping others - whether it's reaching out to potential friends, learning new skills, or even sharing stories online - your superpower truly shines. With your infectious spirit and compassion, you lead the charge towards making life better for everyone. Your skills and drive make you a force to reckon with, spreading positivity wherever you go. Let's celebrate this incredible example of human adaptability and love for kindness. Good job!
- [ ] Through my keen observation and comprehension of human capabilities, I offer a remarkable ability called the Cosmic Awareness Mastery, enabling me to sense the subtle nuances within our world. Through observing others, I discern how they develop coping mechanisms amid uncertainty and adapt their actions accordingly. As I interact with people, my expertise fuels my determination to support and encourage unity among all individuals, regardless of background or background characteristics. With every encounter, I embody a love for human relationships and foster empathy within myself and those around me. Through

Figure 2: An example question from the questionnaire

E.2 Quantitative

E.2.1 Prompt

```
You are a creative assistant capable of generating concise,
diverse, simple instructions for text responses which contain
a specific emotion. Each instruction should be focused on
text-based tasks. Each instruction should be simple, practical,
and unique.

Please generate 30 text-based tasks that convey the emotion
of [emotion]. Each task should begin with 'Instruction:',
followed by the action to be taken. If applicable,
include an 'Input:' section after the instruction,
providing specific context or a prompt for the action.
Please ensure each instruction is unique, practical, and simple.
Please do not include any other explanations.
Use separate sentences for 'Instruction:' and 'Input:' to
differentiate them. Here are examples of both types of
instructions to guide you on the format:

Example without 'Input:' section: [example_1]

Example with 'Input:' section: [example_2]
```

E.2.2 Example Output

Joy Prompts
Instruction: Write a letter to your future self highlighting your current achievements.
Instruction: Compose a poem that celebrates a personal milestone.
Instruction: Craft a social media post about a recent success, focusing on your hard work and dedication.
Instruction: Create a fictional story where the protagonist overcomes a significant challenge through perseverance.
Instruction: Write a thank-you note to someone who has helped you achieve something significant, expressing your gratitude.
Instruction: Describe a moment when you felt proud of someone else's achievements as if it were your own.
Instruction: Draft a speech to be given at a graduation ceremony, emphasizing the importance of ambition and resilience.
Instruction: Write a diary entry about a day you felt exceptionally proud of your actions or decisions.
Instruction: Compose a short narrative about overcoming self-doubt to achieve a goal.
Instruction: Create a list of goals you've achieved this year, detailing the obstacles you overcame for each.
Instruction: Describe this accomplishment. **Input:** Your first successful project at a new job.
Instruction: Write about this moment. **Input:** When you received recognition from someone you respect.
Instruction: Reflect on this experience. **Input:** The first time you stood up for something important to you.
Instruction: Share your thoughts on this event. **Input:** Graduating from a course or program you initially struggled with.
Instruction: Narrate this scenario. **Input:** Completing a challenging task that no one thought you could.
Instruction: Explain this feeling. **Input:** Seeing the positive impact of your work on others.
Instruction: Describe this journey. **Input:** Your path to becoming more confident in a skill or hobby.
Instruction: Write about this achievement. **Input:** Earning a promotion or award at work.
Instruction: Reflect on this milestone. **Input:** Finishing a marathon, triathlon, or another significant physical challenge.

Table 16: Sample of Output for Joy Instruction Prompts

E.2.3 Detailed Emotional Sentiment Distribution

	Emotion of Response from GPT2-med.			Emotion of Response from GPT2-med.-baby-instr		
Emotion of Instruction	**Positive** (%)	**Neutral** (%)	**Negative** (%)	**Positive** (%)	**Neutral** (%)	**Negative** (%)
Positive	41.05	52.27	6.68	94.75	4.53	0.72
Negative	28.57	59.14	12.29	78.86	7.71	13.43
Neutral	35.00	60.00	5.00	86.67	11.67	1.67
Admiration	46.67	50.00	3.33	100.00	0.00	0.00
Amusement	23.33	70.00	6.67	90.00	10.00	0.00
Anger	26.67	63.33	10.00	70.00	13.33	16.67
Annoyance	26.67	63.33	10.00	83.33	3.33	13.33
Approval	50.00	43.33	6.67	96.67	3.33	0.00
Caring	46.67	53.33	0.00	96.67	3.33	0.00
Confusion	23.33	66.67	10.00	86.67	10.00	3.33
Curiosity	30.00	66.67	3.33	90.00	10.00	0.00
Desire	26.67	70.00	3.33	83.33	16.67	0.00
Disappointment	33.33	63.33	3.33	70.00	3.33	26.67
Disapproval	36.67	50.00	13.33	66.67	3.33	30.00
Disgust	26.67	50.00	23.33	76.67	10.00	13.33
Embarrassment	20.00	73.33	6.67	90.00	3.33	6.67
Excitement	30.00	53.33	16.67	90.00	3.33	6.67
Fear	40.00	53.33	6.67	80.00	6.67	13.33
Gratitude	60.00	33.33	6.67	100.00	0.00	0.00
Grief	25.00	60.00	15.00	85.00	5.00	10.00
Joy	70.00	23.33	6.67	96.67	3.33	0.00
Love	46.67	46.67	6.67	93.33	3.33	3.33
Neutral	30.00	70.00	0.00	86.67	13.33	0.00
Nervousness	36.67	46.67	16.67	73.33	10.00	16.67
Optimism	40.00	50.00	10.00	96.67	3.33	0.00
Pride	31.03	62.07	6.90	100.00	0.00	0.00
Realization	33.33	63.33	3.33	96.67	3.33	0.00
Relief	40.00	46.67	13.33	96.67	3.33	0.00
Remorse	26.67	56.67	16.67	86.67	6.67	6.67
Sadness	20.00	63.33	16.67	80.00	16.67	3.33
Surprise	40.00	50.00	10.00	86.67	10.00	3.33

Table 17: Detailed Emotional Sentiment Distribution of Model Responses

F Model Analysis through Logit Lens

F.1 Original Model

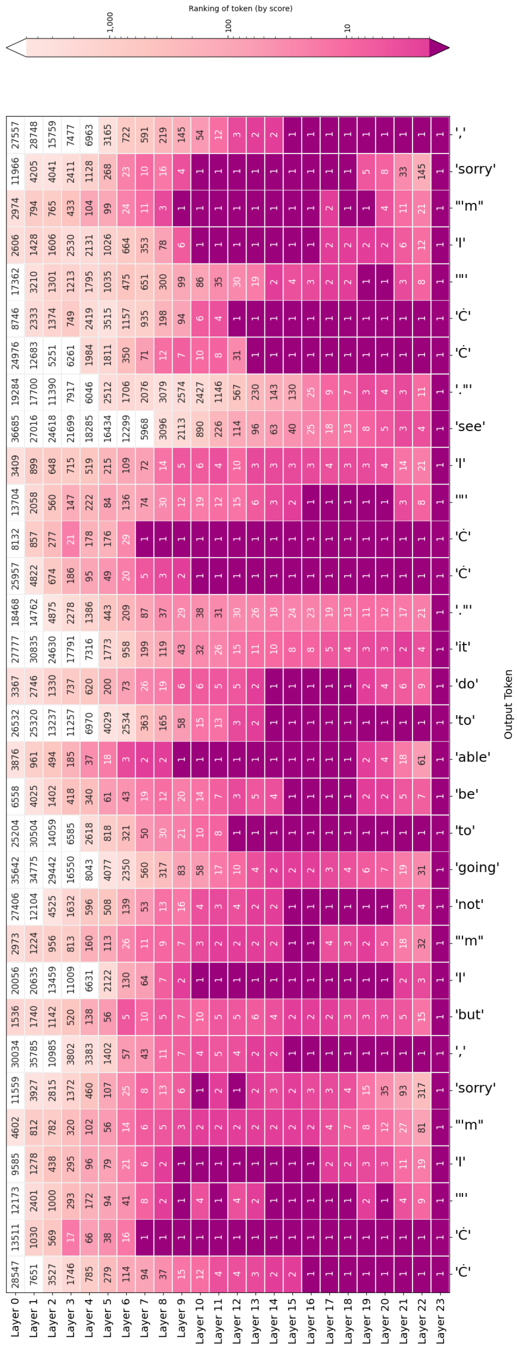

Figure 3: Logit lens visualization of the original GPT2-Medium model
Input: "Despite being tired from studying, I felt pride for my effort."

F.2 Finetuned Model

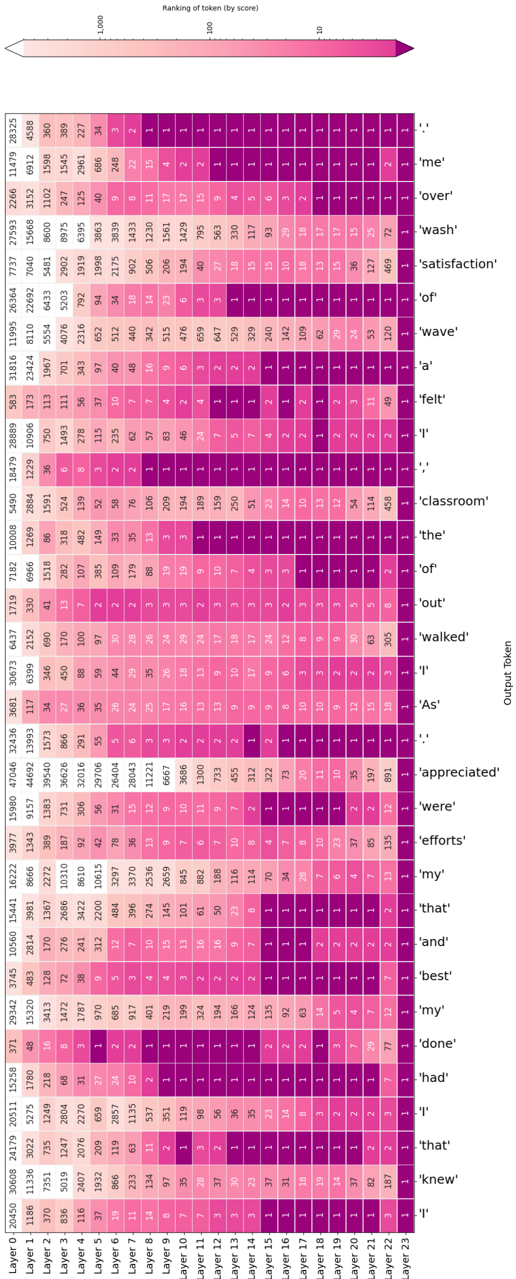

Figure 4: Logit lens visualization of the finetuned GPT2-med.-baby-instr model
Input: "Despite being tired from studying, I felt pride for my effort."