BrightSide: Using Natural Language Processing to Promote Mental and Emotional Well-being

Anushree Aggarwal
Computer Science
Stanford University
anushre@stanford.edu

Brennan Megregian
Computer Science
Stanford University
brennan4@stanford.edu

Max Sobol Mark
Computer Science
Stanford University
maxsobolmark@stanford.edu

Abstract

This paper introduces BrightSide, an innovative platform leveraging advanced Natural Language Processing (NLP) techniques and the in-context learning capability of the GPT-3.5 model, to bolster mental and emotional well-being. In the wake of the COVID-19 pandemic's heightened stress environment, BrightSide uses positive reframing—a cognitive restructuring technique—to generate personalized positive affirmations aimed at immediate relief. Furthermore, BrightSide applies text summarization techniques to facilitate users' understanding of prevalent mood and emotional patterns. For a comprehensive evaluation, we employ the Positive Psychology Frames dataset, assessing the model's effectiveness in producing quality positive reframings, and the SAMSum dataset, analyzing the model's proficiency in summarizing emotional exchanges. This pioneering research offers a unique fusion of technological advancements and psychological strategies, seeking to create novel, accessible therapeutic aids that supplement traditional mental health interventions and contribute to the critical discourse on mental health and well-being.

1 Introduction

Mental health and well-being, ever-present global concerns, have been further accentuated by the unprecedented crisis of the COVID-19 pandemic. The ongoing global situation has thrown into sharp relief the importance of identifying and addressing mental health disorders. As the world continues to grapple with the pandemic's ramifications, discourse on mental health has surged in both the popular media and academic literature (Moreno et al., 2020).

The construct of mental or emotional well-being extends beyond the realm of clinical diagnoses. It refers to the emotional tenor of an individual's daily experiences, including the frequency and intensity of a range of emotions such as joy, stress, sadness, anger, and affection, which collectively shape one's quality of life (Kahneman and Deaton, 2010). This concept of well-being is holistic, encompassing both mental and physical health, a high level of life satisfaction, and a sense of purpose. At its core, well-being is characterized by feelings of positivity and contentment with oneself and one's life.

Contrastingly, mental health and mental illness denote the absence or presence of specific mental diseases. These can also be construed as states of being influenced by a range of biological, psychological, and social factors that contribute to an individual's mental state and their ability to function effectively within their environment (Manwell et al., 2015).

In addressing mental health disorders, traditional interventions such as psychotherapy and counseling continue to be effective and reliable modalities. However, strategies to foster mental well-being, a concept with a broader scope, require more nuanced, easily accessible, and personalized approaches. Best practices for promoting mental well-being are highly individualized, varying significantly from person to person.

One such strategy, known as positive reframing, has shown considerable promise in enhancing emotional well-being. Positive reframing involves the cognitive restructuring of negative or anxiety-provoking thoughts into positive ones, a technique that fosters gratitude and mindfulness. This process has been validated by research demonstrating its efficacy in enhancing emotional well-being (Ziems et al., 2022). For instance, a thought such as, *"I hate making decisions"* can be positively reframed to *"I appreciate the opportunity to make decisions as it empowers me to shape my future in line with my values and desires."*

Complementing this approach are positive affirmations—constructive phrases or statements used to challenge and counter negative or unhelpful

thoughts—proven to promote emotional well-being (Cascio et al., 2015). Following the example given earlier, a positive affirmation to address decision-making anxiety could be, "I trust that I will guide myself to the right decision."

Given this background, our project seeks to explore the application of natural language processing (NLP) techniques, particularly utilizing large language models for prompts, with the aim of promoting mental and emotional well-being. The focal point of our endeavor is to generate personalized positive affirmations using the principle of positive reframing. By doing so, we aim to craft a unique integration of technology and psychology, deploying the capabilities of NLP to enable individualized enhancement of emotional well-being.

2 Research Questions

Technological advancements have resulted in intriguing intersections between machine learning and mental health. Mental health chat-bots, for instance, have recently emerged as promising tools for continuous therapeutic interactions (Vaidyam et al., 2019). While these digital interfaces serve a crucial role, our platform BrightSide seeks to fill a different niche. Specifically, it aims to provide immediate relief during transitory periods of stress or anxiety through the provision of personalized positive affirmations, produced using positive reframing techniques.

BrightSide leverages the in-context learning capabilities of GPT-3.5, a large pre-trained language model, to examine how interactive and tailored positive reframing can bolster individuals' mental and emotional well-being. We do not envisage BrightSide as a substitute for traditional counselling and psychotherapy. Instead, we aspire to offer immediate and short-term support for individuals seeking assistance to navigate basic everyday feelings of stress, sadness, or anger.

Drawing inspiration from positive affirmation podcasts, our methodology involves using positive reframing techniques to devise personalized affirmations. We seek to scrutinize the effectiveness of these techniques in promoting emotional well-being. A key facet of our investigation involves assessing the efficacy of the in-context learning technique in offering beneficial positive reframings.

After collating input text from our users and the positive reframings produced by our model, we intend to further explore the application of text summarization techniques. We hypothesize that these techniques could provide valuable insights by helping users discern and recognize prevalent behaviors and patterns related to their mood and overall emotional well-being.

In terms of the research question, our study focuses on the following:

1. How effective is the in-context learning of GPT-3.5 in generating personalized positive reframings for immediate relief from everyday stressors?

2. Can positive reframing techniques be used effectively to create personalized affirmations, and to what extent can these methods promote emotional well-being?

3. How can text summarization techniques applied to user inputs and model-generated reframings facilitate users in understanding and identifying common patterns in their mood and emotional well-being?

Our overarching goal is to utilize the capabilities of machine learning, especially large language models, to develop accessible and immediate therapeutic aids that supplement traditional psychotherapy. Through this exploration, we hope to contribute to the broader discourse on mental health by presenting an innovative blend of technology and psychology.

3 Data

To analyze and evaluate the performance and efficacy of our proposed tasks, positive reframing and text summarization, we will utilize two distinct datasets.

3.1 Positive Psychology Frames

The Positive Psychology Frames dataset by Ziems et al. (2022) provides an essential benchmark for measuring the quality of positive reframing generated by our model. This dataset is unique and particularly suited to our needs due to its focus on transforming stressful experiences, indicated by the hashtag #stressed, into positive reframings. This aligns well with our goal of generating personalized positive affirmations for users experiencing stress or anxiety.

The dataset's comprehensive structure, consisting of 8,349 sentence pairs and 12,755 structured annotations, provides a vast array of instances where real-life expressions of stress are reframed positively. This richness in data ensures a high level

of diversity in the training and evaluation process, crucial for developing a model with generalizable skills.

The six reframing strategies outlined in the dataset play a pivotal role in guiding our model's approach to generating positive affirmations:

- **Growth Mindset:** Encourages the model to perceive difficult situations as opportunities for personal growth, thereby promoting resilience and adaptive coping strategies.
- **Impermanence:** Reminds the model that negative situations are not permanent, which can instill hope and reduce feelings of helplessness in users.
- **Neutralizing:** Ensures that the model doesn't exacerbate negative feelings through its language, but instead moderates the emotional intensity of the situation.
- **Optimism:** Promotes a focus on the positive aspects of the present situation, thereby fostering positive emotions and reducing negative affect.
- **Self-affirmation:** Enhances self-efficacy and self-esteem by reminding users of their strengths and values.
- **Thankfulness:** Encourages gratitude, a key contributor to subjective well-being.

3.2 SAMSum

The SAMSum dataset, sourced from Hugging Face, is a fundamental tool for our model's training and evaluation when addressing the task of abstractive summarization. Introduced by Gliwa et al. (2019), this corpus is purpose-built for dialogue summarization, making it highly relevant to our project, given the conversational input-response context we operate within.

With its rich content, the SAMSum dataset comprises 16,679 dialogues, all expertly annotated and summed up in an abstract manner. The dialogue instances span a multitude of themes, which helps in maintaining a broad representation of conversational scenarios during training, ultimately assisting in the generalization of the model's summarization ability.

The strength of the SAMSum dataset lies in its focus on abstractive, rather than extractive, summarization. This means the summaries do not merely extract key phrases from the dialogue but rather retell the essence of the conversation in a fresh, concise, and coherent manner. This ability to abstract and articulate the crux of a dialogue is crucial for our model, enabling it to summarize intricate and nuanced conversations effectively.

Moreover, the dataset's emphasis on maintaining the informal, colloquial style of dialogues makes it particularly valuable for our application, allowing our model to generate user-friendly summaries that feel natural and engaging.

The unique features of the SAMSum dataset guide our model's summarization strategy:

- **Context Awareness:** The model learns to take the entire dialogue into account, ensuring summaries are coherent and contextually accurate.
- **Abstraction:** Encourages the model to generate summaries that capture the essence of dialogues, instead of merely extracting phrases.
- **Conciseness:** The model is trained to eliminate unnecessary details and focus on the key points, providing succinct summaries.
- **Natural Language:** The colloquial style of dialogues helps in developing a model that delivers summaries in a user-friendly, informal manner.

Therefore, the SAMSum dataset forms a robust base for our model's training, enabling it to generate high-quality, abstractive, and contextually coherent dialogue summaries.

4 Methods & Approaches

Our research methodology is a combination of the latest advances in large pre-trained language models, specifically harnessing their in-context learning capabilities for both positive reframing and text summarization tasks. Our application, BrightSide, intends to use these methods to provide users with personalized positive affirmations as well as summaries of their concerns and the reframed interpretations of those concerns. This application could potentially help identify recurring patterns, triggers, and themes related to their concerns and anxiety, given the user's consent for this introspective tracking.

4.1 Positive Reframing

Positive reframing is a critical element of our project, the goal of which is to transform pessimistic or negative sentences into their positive counterparts. To achieve this, we experimented with a k-shot learning approach, trying varying values of k, specifically 0, 1, 2, 5, and 10, using OpenAI's GPT-3.5-turbo model.

In our specific application, we present the model with k pairs of sentences from the training set of the Positive Psychology Frames dataset. These pairs demonstrate examples of positive reframing. The model is then given a similar reframing task. For instance:

> "Please positively reframe the following sentences:"'
>
> **Input:** "I have less than 2 hours to get home and make mashed potatoes from scratch."
> **Reframed:** "I have less than 2 hours to get home and make mashed potatoes from scratch, once I make them I'll be so relieved."
>
> **Input:** "If no one sees me for a while it's because I've climbed in a dark hole and don't wanna be found for a while."
> **Reframed:** "I'm going to take some time to take care of myself for a while, so don't worry if you don't see me."
>
> {. . .}

We tune the value of k, with details provided in Section 5.1. We arrived at this specific prompt after doing some prompt engineering. In particular, we found that giving less information about the task generally resulted in better, more positive reframings, as discussed in 5.1, and that inserting "Reframed: " before the model output was important, because without it the model would stochastically insert it.

4.2 Text Summarization

Alongside positive reframing, our approach also integrates the task of text summarization using the same large pre-trained language model. This process aims to generate concise summaries of a selection of input sentences and their respective reframings. The purpose of this text summarization task is twofold. Firstly, it provides users with a condensed overview of their expressed concerns and the respective reframed perspectives. Secondly, it aids in identifying common patterns and themes that surface in the concerns expressed by the users.

The relevance of text summarization in our project comes from the inherently dialogic structure of our data. The input consists of user sentences expressing their concerns and the output consists of the reframed versions of these concerns. In this scenario, summarization serves to condense these dialogues into more manageable and insightful units of information. This, in turn, facilitates pattern recognition and theme identification, contributing to a more comprehensive understanding of the user's emotional landscape.

Our specific experimental approach for text summarization involves in-context learning on OpenAI's GPT-3.5 turbo model, in order to keep our methods consistent between this task and our primary positive reframing task so that the model maintains a consistent understanding of the user's context while summarizing the dialogues (Brown et al., 2020). In particular, we explore the effects of few- or k-shot learning, for varying values of k, specifically 0, 1, 2, 5, and 10. This experimentation adds coherence and continuity to the overall user experience. To build each 'base' prompt for each value of k, we randomly selected and removed k input-summary pairs from the SAMSum dataset. For each test example, we would then append this to our main base prompt. As an example, the 1-shot base prompt for our in-context summarization experiment was:

> "Please summarize the following conversation:"'
>
> **Input:**
> *John:* I'm feeling really stressed out.
> *Anna:* Why, what's going on?
> *John:* Between university and trying to secure an internship, I'm just overwhelmed.
> *Anna:* I can imagine that's tough. Have you considered talking to your career advisor about this?
> *John:* I did, but I still don't have an internship and my coursework is piling up.
> *Anna:* I understand. Remember, it's okay to ask for help when you need it.
> *John:* I know, but I feel like I should be able to handle this.
> *Anna:* We all need help sometimes, John.

You're doing your best and that's all that matters.
Summary: John is working hard to keep up with his schoolwork and find an internship. He's even asked his career advisor for help. His friend Anna reminds him that it's okay to ask for help when things get tough, and that trying your best is what really counts. This shows that even when things are hard, it's important to keep going and remember that it's okay to ask for help.

In our platform, users are given the choice to allow for their inputs to be stored in BrightSide for the purpose of providing summarizations. If users do not allow this, then summarization is not offered as option. For users who do give permission for their messages to be stored, we keep a JSON list of there string inputs and our positively reframed outputs. Then, this list is joined into one whole string where each of the user's input is given the prefix "You: ", and each of BrightSide's summarizations is given the prefix "BrightSide: ". We then append this string to our base prompt as shown above.

4.3 User Interface

To combine our two tasks of positive reframing and text summarization, we implemented an initial concept for our platform and app, BrightSide, using React Native and Flask (Grinberg, 2018). Our design involves an initial landing screen providing the user with information about our platform and how to use it (as seen in Figure 6 in Appendix A). We also provide a check box for users to allow for their inputs and exchanges to be stored so that they can be provided with summaries. Finally, at the bottom of the screen we have a note of precuation that the platform should be used for positive reframing and personalized affirmations only, and that it is not a long-term replacement for traditional forms of psychotherapy. We felt that these features were important design decisions, as they immediately set the expected norms for using the platform, and give users complete control and awareness over what they can expect the app to help them with.

For the text-input interface, our initial plan was to use a message-style design, where the user would appear to be sending a 'message' to BrightSide, who would then 'reply' to the message, to give a sense that the user was conversing with BrightSide. However, we realized that this design may have encouraged users to use our platform as a mental-health chat-bot, which we have been clear is not the focus of this product. As a result, we pivoted our design to be a simple text-box input, where users can write as much as they want about how they are feeling, and upon pressing 'reframe' (Figure 7a in Appendix A), they are taken to a new screen which displays BrightSide's positive reframing of their input (Figure 7b in Appendix A). We felt by focusing on individual reframings in this way, we would better focus on the affirmation style of our reframings, and with no other distractions on the page, the user can just focus on their current emotions and how to reframe them. For the summarization, we used the same design styles, expect this time when a user presses 'Give me a recap', a summarization of their feelings and BrightSide's resposes and suggestions are displayed (Figure 8 in Appendix A).

As stated, to implement these designs, we utilized a React Native frontend, and then connected this to a Flask API server. As a result, whenever a user pressed the 'Reframe' button, their input would be reformatted and inserted into our base 5-shot prompt (the best-performing prompt we selected from our positive reframing experiments), and GPT-3.5 turbo's subsequent generated reframing would then be displayed on the next screen. Similarly, whenever users pressed the 'Give me a recap' button, their stored previous messages are joined into one string, and formatted as described in Section 4.2. Then the summarized output from GPT-3.5 turbo is displayed on the next screen.

5 Results and Discussion

In our quest to assess the effectiveness of our model in the positive reframing task, we implemented a dual strategy, integrating both automated and human-driven methods. This hybrid approach was designed to deliver a more comprehensive and nuanced evaluation of our model's performance, addressing not just the quantitative aspects but also qualitative factors that are often overlooked in automated methods.

5.1 Positive Reframing

5.1.1 Automated Evaluation - BERTScore

Our previous deliverables featured BLEU score as a measure of similarity between our model's predicted reframings and the target sentences from the Positive Psychology Frames dataset. However, as previously discussed, BLEU score was a limited

metric to use, because it was unable to capture the nuances of sentiment change and context preservation. Further, the positive reframing task has many possible valid reframings for each input sentence, and BLEU score only captured similarity with one target reframing.

Given these limitations of BLEU score, we decided to shift to BERTScore (Zhang et al., 2019) as our main automated metric. As opposed to BLEU score, BERTScore is not an n-Gram matching approach. Instead, it measures similarities between sentences in the embedding space induced by the BERT model (Devlin et al., 2018). Because BERT was pre-trained with a large corpus of text, its embeddings should capture more accurately capture the similarities (or differences) between our model's predictions and the targets from the dataset, even if different specific wordings were used.

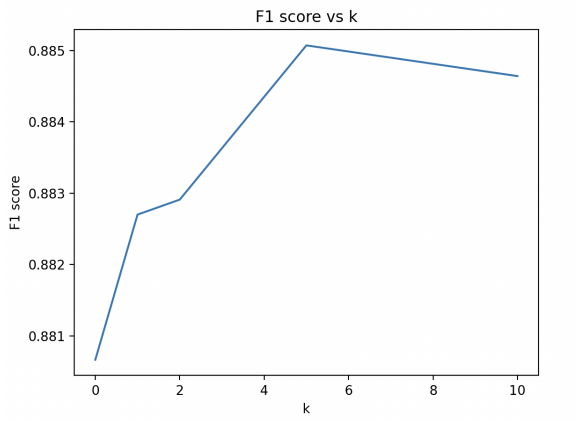


Figure 1: BERTScore (F1 score) increases for higher values of k, but marginally, and plateauing at $k = 10$. Scores were calculated on the complete dev-set of the Positive Psychology Frames dataset.

Figure 1 shows BERTScore values for different values of k. All values of k yield high BERTScores (>0.88), indicating that the model generally outputs reframings that are similar to the reference ones from the dataset.

5.1.2 Human Evaluation

To garner a comprehensive understanding of our model's proficiency in positive reframing, we have devised a meticulous human evaluation strategy, dovetailing with our automated evaluation.

We conducted an evaluation of our AI model's positive reframing capabilities, involving ten participants. Each participant rated sentences on a scale from 1 (indicating scope for improvement) to 5 (signifying perfection) across four main criteria: Content Preservation, Sentiment Change, Reflection, and Relevance. We also assessed participants' reflections on the positivity of the reframed sentences.

The collected data presented an optimistic picture:

- **Content Preservation:** Our model demonstrated proficiency with 48% of the reframed sentences receiving a 'Perfect' (5) rating, and 21% a 4. However, this also highlights an area where the model could benefit from further fine-tuning to increase the percentage of perfect scores.
- **Sentiment Change:** 66% of sentences were rated 'Perfect' (5), and 18% scored a 4. This demonstrates the model's high effectiveness in transforming negative sentiments into positive ones.
- **Reflection:** The reframed sentences fostered increased positivity for 73% of participants (rating of 5), with another 15% giving a score of 4.
- **Relevance:** A strong 87% of responses received a 'Perfect' (5) for maintaining context, with 7% scoring a 4. This indicated that our model excels in keeping reframed sentences contextually accurate.

These results indicate that our AI model is proficient in positive reframing, specifically in sentiment change and maintaining relevance. However, there is room for improvement in content preservation to ensure the original meaning is consistently captured, and that our model generates outputs that are actually helpful for the user.

5.1.3 Automated Evaluation - Sentiment Analysis Evaluation

Since a major factor in assessing our positive reframing method is to determine whether our model actually outputs *positive* reframings, we also experimented with sentiment analysis, to see whether the overall sentiment of the input sentence became more positive. Figure 2 shows our analysis of how our method transforms the sentiment of sentences. Our method outputs much more positive sentences even than the original reframings from the dataset, especially the 0-shot version. Indeed, the more examples from the dataset we give our method, the less positive our reframings become.

Figure 2: Average sentiment of original sentences from the dev-set of Positive Psychology Frames dataset (red), original reframings from the dataset (blue), and our model's reframings with varying values of k (green). Negative sentiment scores represent negative sentiment, and positive scores represent positive sentiment. Original sentences are strongly negative, whereas every reframing method is positive. Our model is significantly more positive than the reference reframings from the dataset, and positiveness decays with increasing values of k.

5.1.4 Holistic Evaluation

The integration of automated and human-driven methods provides a holistic evaluation of our model's performance. This approach not only combines quantitative and qualitative analysis but also considers various aspects of the task, which is particularly beneficial for complex and novel tasks like positive reframing.

However, the evaluation method is by no means static and will continue to evolve based on the task requirements and research progression. For instance, to test the robustness of our approach, we consider variations in the "shot" learning scenario, experimenting with k-shot learning for multiple values of k. By integrating these quantitative and qualitative insights into our evaluation strategy, we aim to continually refine and enhance our approach, ensuring it aligns with the task requirements and delivers accurate and meaningful results that reflect the model's true capabilities in the positive reframing task.

5.2 Text Summarization

5.2.1 Automated Evaluation - ROUGE Score

Our primary metric for quantitative evaluation for the performance of our text summarization approach utilized the ROUGE score. ROUGE, or Recall-Oriented Understudy of Gisting Evaluation, was first introduced in 2004 by Lin (2004) and has become a very common method of evaluation for text summarization. ROUGE compares overlapping units such as n-grams, word sequences, and word pairs with human written summaries. Thus, to measure the performance of the generated summaries for various k-shot prompts, we compare GPT-3.5 turbo's generated summaries to those given by the SAMSum dataset (Gliwa et al., 2019).

Ultimately, for each value of $k \in \{0, 1, 2, 5, 10\}$, we calculated the ROUGE-1 and ROUGE-L scores for each generated summary for inputs from SAMSum's test dataset. ROUGE-1 calculates the unigram overlap between the generated and 'target' summary, and ROUGE-L is based on the longest common subsequence (LCS) between the generated output and target summary, i.e. the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both. A longer shared sequence indicates more similarity between the two sequences. For each value of k, we took the average of ROUGE-1 and ROUGE-L scores across all test examples. Additionally, for ROUGE-1 and ROUGE-L, we calculate the precision (Figure 3), recall (Figure 4), and final F-1 score (Figure 5), where F-1 is computed as:

$$2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

.

Discussion. Overall, we see that our prompting techniques for each value of k perform quite similarly across the board. However, there is a tendency for 2-, 5-, and 10-shot prompts to outperform 0-

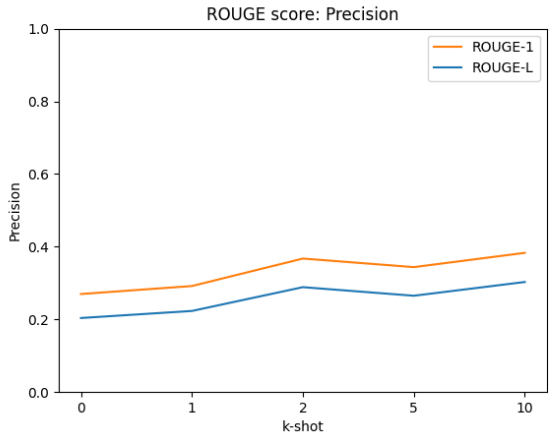


Figure 3: ROUGE precision versus k

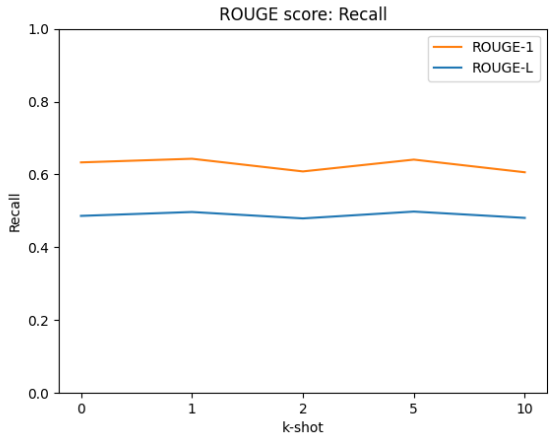


Figure 4: ROUGE recall versus k

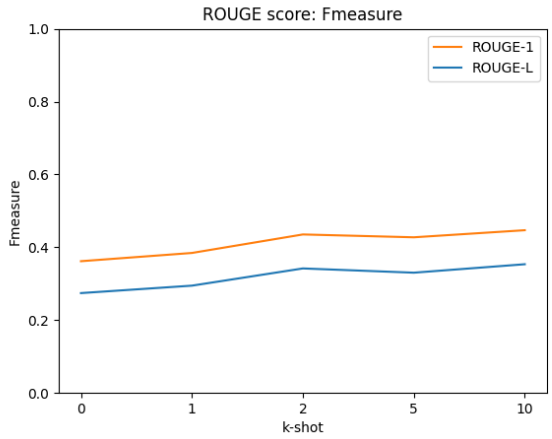


Figure 5: ROUGE F-1 versus k

and 1-shot. This makes sense, since by giving our model more context about the types of summaries we hope to generate, we can expect outputs that are closer to the reference summaries provided by the dataset. We also observe that our method achieves higher ROUGE-1 scores than ROUGE-L. However, given the very subjective nature of summaries, and since for summaries there is typically no 'correct' answer, it makes sense that our model is able to generate more unigrams that are similar to the reference summaries than whole subsequences. If we examine the F-1 plot (Figure 5), we see that most of our k-shot prompts achieve scores approximately between 0.4 and 0.5. While this tells us that our method did moderately well in replicating the target summary, we can look out at the breakdown of this score into precision and recall to gain a better understanding of where our method is succeeding and where it is failing. As we can see, the ROUGE-1 precision scores for all k lie between 0.3 and 0.4, which is somewhat low. On the other hand, ROUGE-1 recall scores are much higher, and generally lie between 0.6 and 0.7. Low precision and high recall tells us that our summarization approach returns most of the 'correct' or 'relevant' unigrams for each summary, but that it also returns a good number of incorrect or 'irrelevant' unigrams. However, given the nature of text summarization as a task, this is not necessarily a bad thing.

To look more holistically at the performance of our text summarization approach, we can note that automatic quantitative evaluations may not be the best method of evaluation for our task. Since there can be many different relevant summaries of an input text, and there is often no single 'correct' summary, ROUGE may not be very helpful for us. More specifically, two problems with ROUGE are that 1) it favors lexical similarities between generated summaries and model summaries, which makes it unsuitable to evaluate abstractive summarization, or summaries with a significant amount of paraphrasing, and 2) it does not make any provision to cater for the readability or fluency of the generated summaries (Ng and Abrecht, 2015). While the latter problem is not so relevant with the use of GPT-3.5 turbo which has little issue generating coherent and fluent outputs, the former objection is a very real concern. As a result, while we were not able to implement human evaluations for our text summarization for this project, our first next step if we were to continue this work would be utilizing this more holistic evaluation technique. Additionally, when examining some of the specific generated summaries generated by our GPT-3.5 turbo prompts, we often found that the summaries

were as long as the original input text, which almost defeats the purpose of summarization in the first place. Thus, going forward we would definitely benefit from human evaluation as well as further prompt engineering and instruction tuning to ensure the sumamrizations are kept brief.

6 Future Work

In future developments, we aim to enhance our model's effectiveness in two main areas. Firstly, we will conduct extensive human evaluations for both text summarization and positive reframing tasks. Given the subjective nature of summarization and the specificity of our use-case, assessing the users' understanding of their emotions and concerns is crucial. Additional human input on the efficacy of positive reframing will further refine our model's ability to transform negative sentiment while preserving the original content.

Secondly, we will focus on technical advancements in our user interface design. This encompasses exploring current technological trends to provide a highly functional, intuitive, and aesthetically pleasing user experience. This technical approach will improve our model's effectiveness, and further fulfill our commitment towards promoting emotional well-being.

7 Conclusion

In this paper, we put forth a robust, comprehensive methodology for AI-driven positive reframing and text summarization tasks, utilizing OpenAI's GPT-3.5 Turbo. The evaluation strategies for our work ingeniously marry automated metrics, such as BERTScore and ROUGE, with human evaluations to quantify proficiency in an array of dimensions including content preservation, sentiment change, and relevance.

For positive reframing, our AI model consistently yielded reframings that were in close alignment with the reference instances from the dataset. Human evaluations shed light on the model's adeptness in sentiment transformation and context preservation, simultaneously revealing potential avenues for improvement in content preservation. These observations were further substantiated by sentiment analysis evaluations, which underscored the model's prowess in generating positive reframings.

As for our text summarization methodology, we observed a moderate level of performance across different "shot" prompts, with 2-, 5-, and 10-shot prompts generally exhibiting superior performance than the 0- and 1-shot. Nonetheless, we concede the inherent limitations of automated evaluation techniques such as ROUGE in capturing the intricacies of summarization tasks.

While our research constitutes a significant stride in the field of AI text generation, we acknowledge the areas that could benefit from further refinement. Future iterations of our work will prioritize an encompassing integration of human evaluation methodologies, a more detailed approach to prompt engineering, and a comprehensive exploration into the effects of different positive reframing strategies. The iterative nature of our research underscores our commitment to the perpetual enhancement of AI capabilities in language understanding and generation tasks.

8 Ethical Consideration

Our research, while contributing to advancements in AI-driven text generation, also reaffirms a profound commitment to ethical conduct. We conscientiously approached the issues of fairness, privacy, and potential misuse, striving to instill our methodology with a rigorous ethical framework - especially given the sensitive nature of mental health and wellness, which is the context of our project.

Fairness and Bias. Recognizing the potential of language models to perpetuate societal biases embedded in their training data, we incorporated measures to mitigate such implications. We leveraged diverse datasets, mindful of representing a variety of perspectives and experiences. Nevertheless, we urge for continuous efforts in training more fair and unbiased AI models, and we advocate for further research to identify and neutralize biases.

Privacy. In line with privacy norms, we ensured that all data used, especially in the human evaluations, were fully anonymized, with no identifiable information traceable to the participants. We strictly adhered to the guidelines set forth by institutional review boards and complied with all relevant data protection regulations.

Potential Misuse. While our work opens avenues for positive applications like sentiment transformation in therapeutic contexts, we acknowledge the potential for misuse. The ability of AI models to generate persuasive, positively-framed text could be exploited for manipulation or deceit. To mitigate these risks, we underline the importance of

strong ethical guidelines governing the deployment of these technologies, coupled with measures to educate users about their functioning.

Transparency and Explainability. The 'black box' nature of AI models often poses challenges in understanding their decision-making process. Our work aimed to address this concern by providing clear descriptions of the algorithms, evaluation metrics, and experimental setup, thus promoting transparency and reproducibility. We also intend to contribute to ongoing efforts for developing more interpretable AI models.

In conclusion, we believe in technology serving as a positive force that respects human values and ethics. It is our assertion that all AI research should engage with these ethical dimensions, contributing not only to scientific advancements but also to a more equitable, fair, and responsible technology landscape. We encourage people to use our platform for only the use cases stated in our introduction and on the landing page of our app, and we also encourage those seeking more serious or long-term mental health concerns to contact trained professionals.

9 Authorship Statement

Each author made significant contributions to this work. Anushree developed the BLEU score evaluation and designed the human evaluation metrics for the positive reframing task. She distributed the human evaluation forms and analyzed the resulting data. Brennan led the text summarization experiments, which included k-shot prompting of GPT-3.5 Turbo and ROUGE score calculations. She also developed the user interface to demonstrate the model's usage. Max implemented the k-shot prompting of GPT-3.5 Turbo for the positive reframing experiments, and calculated both the BERT scores and the sentiment analysis results. All members contributed to the writing of this report.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Christopher N. Cascio, Matthew Brook O'Donnell, Francis J. Tinney, Matthew D. Lieberman, Shelley E. Taylor, Victor J. Strecher, and Emily B. Falk. 2015. Self-affirmation activates brain systems associated with self-related processing and reward and is reinforced by future orientation. *Social Cognitive and Affective Neuroscience*, 11(4):621–629.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.".

Daniel Kahneman and Angus Deaton. 2010. High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 107:16489–93.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Laurie Manwell, Skye Barbic, Karen Roberts, Zachary Durisko, Cheolsoon Lee, Emma Ware, and Kwame McKenzie. 2015. What is mental health? evidence towards a new definition from a mixed methods multidisciplinary international survey. *BMJ open*, 5:e007079.

Carmen Moreno, Til Wykes, Silvana Galderisi, Merete Nordentoft, Nicolas Crossley, Nev Jones, Mary Cannon, Christoph U Correll, Louise Byrne, Sarah Carr, Eric Y H Chen, Philip Gorwood, Sonia Johnson, Hilkka Kärkkäinen, John H Krystal, Jimmy Lee, Jeffrey Lieberman, Carlos López-Jaramillo, Miia Männikkö, Michael R Phillips, Hiroyuki Uchida, Eduard Vieta, Antonio Vita, and Celso Arango. 2020. How mental health care should change as a consequence of the covid-19 pandemic. *The Lancet Psychiatry*, 7(9):813–824.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464. PMID: 30897957.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing.

A Appendix

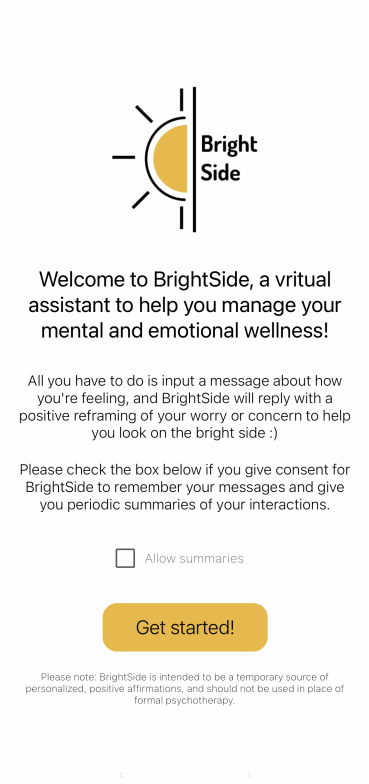
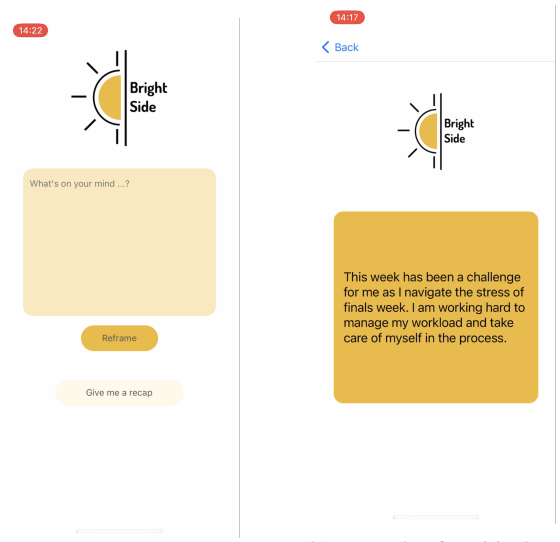


Figure 6: Landing screen with norms and use expectations



(a) Input

(b) Example of positively reframed text

Figure 7: Positive Reframing screens

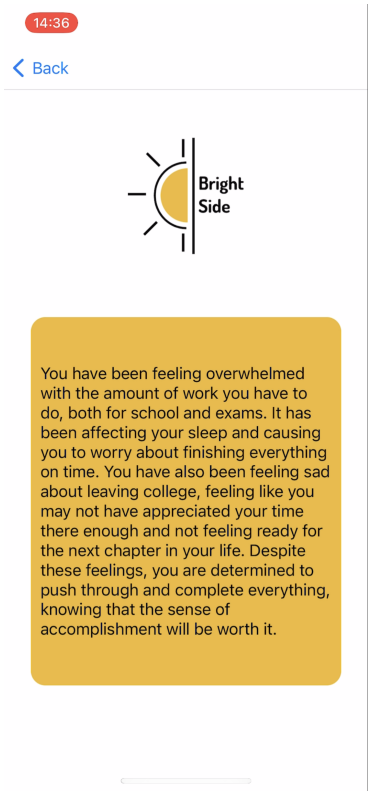


Figure 8: Example of summarized interactions with BrightSide