Evaluating AI-driven Psychotherapy: Insights from Large Language Models and Human Expert Comparisons

Anonymous ACL submission

Abstract

The integration of Large Language Models (LLMs), such as GPT-4, has shown great promise in mental health applications for initial assessments based on user-reported symptoms. Traditional assessments often involve subjective evaluations by professional psychologists, leading to inconsistent reproducibility across datasets. To address this, we developed a comprehensive evaluation framework using entropy analysis, keyword frequency analysis, and Latent Dirichlet Allocation (LDA) to quantitatively assess LLM outputs. Our results indicate that LLMs can effectively identify and engage with a range of treatment topics and provide a broader range of treatment opinions than human psychologists. However, LLMs lack depth in their responses, the recommendation generated by LLMs trends to using generalized word instead of using professional words. This study explores the feasibility of LLMs as virtual psychotherapists, highlights their shortcomings in depth, and proposes improved methods for evaluating large model responses. This research provides valuable insights into the potential and challenges of integrating LLMs into mental health practices, paving the way for future research to enhance the effectiveness and reliability of AI-driven therapeutic solutions.

1 Introduction

Psychotherapy, a therapeutic interaction or treatment between a trained professional and a client aimed at addressing psychological issues and improving mental health, is a fundamental component of the mental health cycle. It applies multiple non-invasive methodologies to address psychological problems. Psychotherapy is also considered as a secondary methodology to prevent the recurrence of certain conditions and is often utilized to manage urgent cases of depression (Karrouri et al., 2021). In psychotherapy field, Cognitive Behavioral Therapy (CBT) is recognized to be crucial in addressing anxiety. This enhances the key position of CBT in helping patients with depression and highlights its importance as both a preventive technique and treatment methodology (Bandelow et al., 2017). Additionally, Non-Directive Support Therapy (NDST) has been applied in psychotherapy treatment methodologies. It provides emotional support and energy for patients in self-exploration and self-development to solve their problems. One research suggests that, compared with traditional methodologies, this psychotherapy approach showed better treatment results in the short term (Cuijpers et al., 2014). Additionally, invasive psychological treatment methodologies have been proven to have similar effectiveness to depression medication treatment during the urgent treatment stage (de Maat et al., 2007). This enhances the key role and benefits of using psychotherapy in treating mental health disorders and managing overall mental health.

The integration of Artificial Intelligence (AI) technologies, such as GPT-4 and other Large Language Models (LLMs), has driven the development of intelligent psychotherapy applications. The primary goal of researchers and institutions is to provide timely and effective treatment recommendations for medical professionals and individuals seeking treatment (Chen et al., 2023; Montagna et al., 2023). Although LLMs have demonstrated the strong ability to analyze natural language and provide diverse feedback quickly (Singhal et al., 2023) , the efficiency and reliability of these AI-powered psychotherapy tools in providing accurate diagnoses and recommending effective treatments still remains controversial (Manríquez Roa et al., 2021) . This is mainly due to the complexity of the medical field that requires large language models to have the ability to understand the medical context, find appropriate medical knowledge, and reason using authoritative information and clues provided by patients, and this complexity in the medical field have led to a variety of potential treatments (Singhal et al., 2023). Therefore, assessing the performance of AI-based virtual psychotherapists in the depth and coverage of their therapeutic advice, especially when compared to human professionals, has become a key focus of research.

This study aims to evaluate the effectiveness of LLM-based chatbots in recommending treatment suggestions and their consistency with those proposed by psychotherapists and the depth of the protocol.

- Our (H_0) is that there is no significant difference in the diagnosis and treatment opinions provided by psychologists and LLMs overall, i.e., the quality of the output of the large model is broadly consistent with that of psychologists.

To test this hypothesis, we introduce a novel evaluation framework that applies case studies from the American Psychological Association (APA) as a benchmark to detect the differences between LLMs output and case scenarios through LDA modeling and entropy analysis, so as to comprehensively evaluate their application in the field of psychology.

Our contributions are:

- We propose and implement a framework combining entropy analysis, keyword frequency analysis, and the novel Latent Dirichlet Allocation (LDA) to evaluate the diversity, depth, and applicability of LLMs in generating psychological diagnoses and treatment recommendations. This provides a quantitative way to measure the feasibility of LLM applications in clinical settings and offers a new perspective on evaluating LLM technology in mental health diagnosis and treatment planning.

- Through detailed comparisons and in-depth analysis, we evaluated the differences between LLM-generated treatment recommendations and those made by human psychologists. Our findings suggest that LLM recommendations often lack the detail found in human expert recommendations, highlighting both the strengths and shortcomings of LLMs in generating psychotherapeutic recommendations and providing a balanced perspective on integrating LLM techniques with psychotherapy practice.

- We demonstrate the potential impact of LLMs in increasing access to mental health care by validating their ability to provide mental health related diagnoses and treatment recommendations in evaluating its diversity and depth. Our study highlights the potential for LLMs platforms to improve the accessibility and scalability of psychotherapy services, especially in resource-limited or remote areas. Additionally, we initiate discussions on ethical, practical, and strategic planning considerations to maximize the benefits of AI in mental health practices.

Through rigorous evaluation and comparative analysis utilizing novel Latent Dirichlet Allocation (LDA) modeling, word frequency analysis, and a series of statistical analyses to evaluate the treatment recommendation capability, diversity, and depth based on 10 professional case studies from the American Psychological Association and generated by LLMs, this study highlights the potential and limitations of LLMs in the diagnosis and treatment of mental health, and provides valuable insights and directions for future research and application in this field.

2 Related Work

In recent years, with the advancement of natural language processing technology, Large Language Models (LLMs) like GPT-4 have been widely studied in the field of primary consultation and support in the health field. Michimasa et al. 2024 demonstrated in their experiments that LLMs can exhibit a level of professionalism similar to that of psychologists, with no high-risk, aggressive, or discriminatory responses found in conversations with GPT-4. In addition, Luoma Ke (Ke et al., 2024)'s study also confirmed that LLMs, as a preliminary diagnostic tool in clinical and counseling psychology, can quickly identify potential mental health problems in users, such as depression and anxiety. John (Ayers et al., 2023) evaluated responses from physicians and LLMs, with the results that raters favoring responses from LLMs, and the quality of LLMs outpacing physician responses.

However, although LLMs have generally received neutral and positive feedback in past research evaluations, they exhibit a range of problems. Critics, such as Topol (Meskó and Topol, 2023), point out that the recommendations generated by LLMs were not very reliable because the data used by the large model did not come from a formal bedside conversation, and that the responses from the large model may involve fictitious sources. According to a survey, out of 157 participants, 123 used ChatGPT for health queries. Besides, 83 people believed that the treatment recommendations provided by the large model are more accurate than those provided by traditional online communities. While the study found that people prefer to use LLMs for health consultations, the researchers also expressed concerns that the databases of LLMs need to be updated in a timely manner to ensure the accuracy and reliability of their information. (Xiao et al., 2024).

At the same time, Natural Language Processing (NLP) technology has been widely used in text analysis tasks, and NLP methods have also shown significant value in the psychological field. For example, text analysis methods such as Pearson correlation coefficients and sentiment analysis have been used to assess the consistency of machine-generated responses with human expert recommendations (Danna et al., 2024). In addition, NLP techniques such as TF-IDF and Word2Vec have been applied to data classification for the assessment of suicidality (Aldhyani et al., 2022). While these techniques excel in dataset processing, they have traditionally been used primarily for data classification in deep learning, or to predict suicidality and mental illness by analyzing online social media comments. Existing studies have not focused on the application of these methods in assessing the output quality generated by LLMs, revealing potential research gaps and development directions in this field.

With the development of AI, especially the integration of LLMs in mental health, finding a way to assess the quality of the output of these models have become particularly urgent (Elyoseph and Levkovich, 2024). The benefit of quantifying the output of LLMs is that it can provide an objective way to evaluate the effectiveness and reliability of these models in real-world applications. Research has shown that while LLMs can deal with a wide range of topics, they often lack the depth provided by human experts, a problem that may stem from the phenomenon of knowledge duplication in LLMs (Chen et al., 2023). Therefore, there is a need to explore and establish a new assessment framework to comprehensively assess the capacity of LLMs in terms of mental health diagnosis and treatment recommendations. Such a framework can not only help identify and address the shortcomings of LLMs in specific applications, but also facilitate a more effective fusion of AI and human expertise.

One methodology can be considered in the framework is the cosine similarity, which can be used to compare similarity of the text written by psychotherapist and LLMs generated text. Cosine similarity is a vector space modeling technique used to quantify the similarity between two documents (Januzaj and Luma, 2022), making it a key tool for text analysis and comparison. This metric calculates the cosine value of the angle between two vectors, representing the position of the text in a multidimensional space, to determine their similarity. It has a wide range of applications, especially in the evaluation of text consistency and relevance in automated systems. In the Automated Essay Scoring (AES) system, cosine similarity plays an important role by comparing the text submitted by students with the documents written by experts. By using this method in conjunction with weighted terminology analysis, the AES system achieves a meticulous assessment of textual consistency, demonstrating the effectiveness of the method in an educational setting (Lahitani et al., 2016). In addition, the field of psychology also employs cosine similarity for diagnostic purposes, facilitating the comparison of the symptoms provided by the patient with the established psychological profile during a virtual consultation. This innovative application helps doctors reduce their workload as a diagnostic aid by analyzing the user's text input to make a preliminary diagnosis of a patient's mental health (Bhattacharya and Pissurlenkar, 2023).

However, when the lengths of the two inputs are different, the output generated by the cosine similarity method will be significantly affected, which is not accurate for evaluating the LLMs response and case studies of text of different lengths. Therefore, we introduce entropy analysis to more effectively evaluate the complexity of the results generated by LLMs. Entropy is a measurement derived from information theory that measures uncertainty and randomness within a system. A study using entropy to measure the consistency and diversity of Key Audit Matters (KAMs) disclosures in audit reports showed that monitoring the entropy of KAMs disclosures can reveal trends and consistency in the evolution of audit practices over time (Lin, 2023). This study suggests that we can evaluate the performance of LLMs by measuring the topic distribution of entropy and further analyze the diversity of LLMs-generated topics.

In our study, we aim to critically assess the effectiveness of LLMs in performing tasks similar to those of virtual psychologists by using APIs such as ChatGPT, as well as mainstream NLP tools, including LDA and entropy analysis.

3 Methods

3.1 Data Source Selection

Our research methodology starts with selecting the appropriate dataset to make the evaluation. We chose a series of formatted case study from APA instead of using non-structural dataset like DAIC-WOZ from USC (Burdisso et al., 2024), mainly because the structured format of the APA is more in line with the capabilities of LLMs. We initially used USC's DAIC-WOZ dataset, but found that ChatGPT could not track the transcription format of Q&A correspondingly when processing this type of transcription's data without manually intervened. While we found that manual intervention allowed ChatGPT to follow the Q&A transcription format in the dataset, this intervention method was shown to lead to later human intervention bias in LLM answers, resulting in inaccurate research results (Loya et al., 2023). In contrast, the highly well-formatted APA case studies provide a diverse and comprehensive mental health scenario, and this structured format is more suitable for assessing the diagnostic and treatment recommendation capabilities of LLMs than the DAIC-WOZ dataset. In addition, APA has been mentioned in many psychology research papers and is considered as one of the most authoritative sources of psychological research data (Badr et al., 2023; Sheridan and Carr, 2018).

In our study, we selected 10 case studies from the American Psychological Association (APA), including cases of individuals with depression and Post-traumatic Stress Disorder (PTSD). These cases include the patient's background, diagnosis, and corresponding treatment plan. All personal information has been anonymized by the APA. The cases cover a diverse range of genders and ages, ensuring a comprehensive evaluation of the treatment recommendations provided by LLMs.

3.2 Entropy Analysis for Topic Distribution

In our study, we used entropy analysis to assess how LLMs divided their attention across different psychotherapy topics and compared it to human psychologists. Through entropy analysis, we can determine whether the text generated by LLMs is concise or diverse with multiple topics. In order to ensure fair comparison, we have normalized the topic probabilities in the document, and the normalization calculation is as follows:

$$p(t_i) = \frac{n_{t_i}}{\sum_j n_{t_j}}$$

Here, n_{t_i} represents the count of words associated with topic t_i within a document, and $\sum_j n_{t_j}$ is the total word count across all topics in that document. This ensures that the sum of probabilities across topics equals one, facilitating a meaningful entropy calculation.

The entropy for each document's topic distribution was then computed using the formula:

$$H(T) = -\sum_{i=1}^{K} p(t_i) \log_2 p(t_i)$$

This equation, where K is the number of topics and $p(t_i)$ denotes the probability of each topic, utilizes the logarithm base 2 to measure entropy in bits, enhancing our understanding of topic distribution's evenness.

3.3 Prompt Design

In this study, a specific prompt was designed for the LLMs to ensure consistency in the responses across different models. This prompt incorporates a curated list of keywords that are closely related to mental health treatment, ensuring that the treatment recommendations generated are relevant and based on well-established psychological principles.

- **Diagnosis Section:** The prompt includes keywords such as *anxiety*, *depression*, and *panic attacks*. These terms are selected to guide the LLMs to focus on common psychological conditions, facilitating a targeted exploration of potential diagnoses.

- **Treatment Plan Section:** Keywords like *Cognitive Behavioral Therapy (CBT)*, *psychodynamic therapy*, and *humanistic therapy* are included. These therapies represent a range of approaches in psychotherapy, allowing the LLMs to generate diverse and comprehensive treatment plans.

This methodical selection of keywords is informed by recent advancements in AI applications within healthcare, where patients can utilize an LLMs to input relevant keywords or questions, thus accessing a wealth of medical knowledge (Pagad et al., 2022). We used this idea to design the prompt to let LLMs' output become consistent and relevant. The complete prompt utilized in our evaluations is detailed in Appendix A.

3.4 Word Frequency Analysis

We used word frequency analysis to assess the similarity between the treatments described in the APA case study and those generated by LLM. Our study built on the potential LDA topic modeling of Blei(Blei, 2003) and extends the application of natural language processing (NLP) techniques in mental health research outlined by Miner (Miner et al., 2020). We aimed to compare the differences in treatment recommendations between the results generated by LLMs and the demonstration results in the APA case study by quantifying treatment-specific terms in text data. Besides, since one research done by Torous and Keshavan (Torous and Keshavan, 2020) highlights the importance of evaluating digital tools to ensure that these tools meet clinical standards and effectively enhance patient care in the field of mental health. Our another focus in our quantitative assessment framework is the analysis of treatment-related word frequency comparisons between APA case studies and LLM outputs. We wanted to use this approach to assess whether the LLMs was able to generate broader clinical recommendations, while retaining some depth of therapeutic insight. Through this exploration, we aim to uncover the potential of LLMs as a tool for mental health practitioners and the performance of LLMs in the professional field.

3.5 Comparative Analysis Using Latent Dirichlet Allocation (LDA)

3.5.1 Objective of Using LDA

In order to provide a detailed analysis and comparison of the treatment recommendations provided by ChatGPT with those described in (APA) case study, we used the LDA as another important part of our evaluation framework for LLMs. LDA was chosen as our methodological tool based on its effectiveness in identifying potential topics in a large corpus of text, as demonstrated by the groundbreaking study (Hagg et al., 2022; Kotenko et al., 2021). As a result, the application of LDA enables a detailed and structured comparative analysis, with a particular focus on the thematic differences between the responses generated by ChatGPT and the treatment recommendations described in the case study. This approach allows us to understand ChatGPT's capabilities and limitations in psychotherapy related tasks. Through this analytical perspective, we aim to critically assess the similarity of ChatGPT recommendations with contemporary treatment standards in evaluating the diversity and depth of the responses, thereby contributing to an ongoing conversation about the integration of AI in clinical settings.

3.6 Summary of Analytical Procedure

The comparative analysis is based on a two-stage approach, distilling and examining the essence of the topic through the LDA model of APA case studies and ChatGPT-generated recommendations. Before LDA was applied, extensive text data preprocessing was performed, including tokenization, stop word removal, and invalid word reduction, to optimize the text's topic extraction.

The analysis process is as follows: First, the Subject Heading Distribution Analysis involves identifying and visualizing the most important words within the topics extracted from APA case studies and ChatGPT outputs. By examining word distribution, the main thematic focus of each source is elucidated, thereby assessing the consistency and differences in treatment topics. Next, the Document-topic ratio assessment quantifies the representation of each topic in a single document, facilitating a fine-grained comparison of topic prevalence between the original case study and ChatGPT recommendations. This stage uses heat map visualization to display the topic distribution pattern, highlighting the similarities and differences in theme emphasis. Following this, the Compare Topic-Word Relationship Exploration uses a heat map to further dissect the relationship between key terms and their related topics in the two datasets. This step is essential for assessing the depth and specificity of ChatGPT's treatment recommendations relative to the established treatment modalities documented in the APA case study. Finally, the Entropy-based variability assessment employs entropy measurements to assess the variability of topic distributions in LLMs and artificially generated text. This analysis quantifies the diversity of topics covered by each source, providing insights into the comprehensiveness of treatment recommendations and concerns.

4 Experiment

In this study, we conducted a comprehensive analysis of 10 cases of depression treatment, with the aim of exploring the differences between large language models (LLMs) and human psychologists in providing treatment recommendations. By using the latent Dirichlet assignment (LDA) model to analyze the text of the treatment recommendations given by both parties, supplemented by entropy analysis and word frequency analysis, we try to reveal the similarities and differences in topic extraction. This article will take the analysis process of the first case study (Case Study 1) as an example, and the data and analysis of the remaining case studies are included in the appendix.

4.1 LDA Modeling

4.1.1 Word distribution in topic

The advice provided by the LLMs identified by the LDA model covers topics such as family, group adjustment, academic, and medication. While the advice of human psychologists also exhibits a similar thematic composition. But in the same case study, human experts emphasize more specific topics. In the 2, psychotherapist provide more specific treatment methodology "CBT" compare with LLMs, which only mention the categorical word such as "treatment".

4.1.2 Document-topic distribution heatmap

Document-Topic Assignment presents a corpus of a series of case studies interpreted by LLMs. The visualization represents a matrix where rows correspond to individual documents and columns represent topics derived from the LLM output.

Each cell in the matrix reflects the proportion of the document content that is relevant to a given topic, which is determined by the inference algorithm of the LDA model. The color gradient from lighter to darker represents an increase in relevance, providing a visual measure of the topic's salience in each document.

As shown in figure 3 and 4, The results of the LLMs show the multifaceted distribution of the various topics, with no single topic dominating the content of the document. This shows that in this case study, LLMs tend to distribute content more evenly across multiple topics, which may indicate that it takes a less specialized but more integrated approach when generating discussions about psychotherapy and mental health.

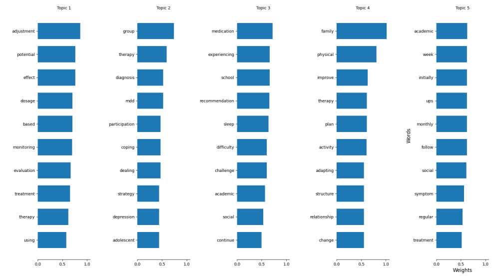

Figure 1: Word frequency analysis of treatment plans generated from LLMs by using APA study cases

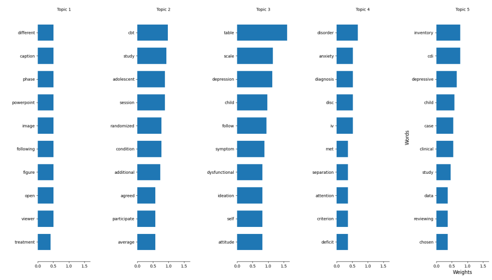

Figure 2: Word frequency analysis of treatment plans based on APA study cases

4.1.3 Topic-word association heatmap

By comparing the document-topic distribution plots of the two datasets, we observed that LLMs had a relatively uniform topic distribution across different documents based on graph 5 and 6, while human psychologists showed a more pronounced preference and focus. A similar phenomenon was observed in the subject-word heatmap analysis, where the relevance of certain keywords in the human psychologist's advice was more concentrated and more dispersed in the LLMs.

4.2 Entropy Analysis

Furthermore, besides using graphic to extract the key insights from the dataset, we also compared the entropy values of treatment recommendations generated by human psychologists (raw entropy) and LLMs entropy across different case studies by using Mann-Whitney U test and traditional box plot. The entropy is calculated to measure the diversity and uniformity of the distribution of topics in therapeutic texts.

Table 2 shows a comparison of the entropy of raw and LLMs in different case studies. The table includes raw entropy, LLMs entropy, the difference between the two, the percentage difference, and the absolute difference for each case study.

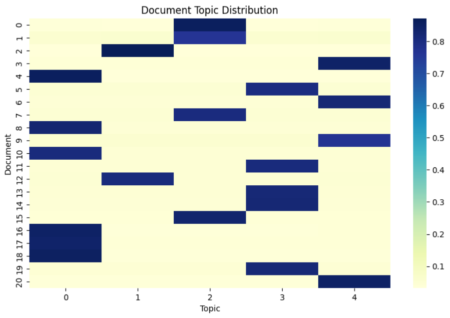

Figure 3: Word frequency heatmap analysis of treatment plans generated from LLMs by using APA study cases

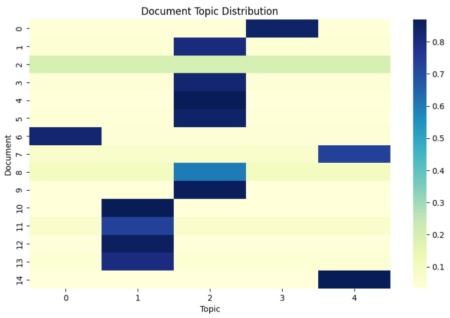

Figure 4: Word frequency heatmap analysis of treatment plans based on APA study cases

Figure 7 presents a box-plot comparing the entropy values of original text and LLM-generated text across different case studies, providing a visual representation of the data.

4.2.1 Mann-Whitney U test

To statistically evaluate the difference in entropy between the original text and the LLM-generated text, we also performed the Mann-Whitney U test, which is a non-parametric test suitable for comparing differences between two independent samples. The results are shown in the table 1 and show that there is no statistically significant difference between the two sets of text.

Table 1: Mann-Whitney U test result

measurement	value
U statistics	31.0
P value	0.162

In our study, the Null Hypothesis (H_0) is that there is no significant difference in treatment recommendations between large language models (LLMs) and case studies overall. The Mann-

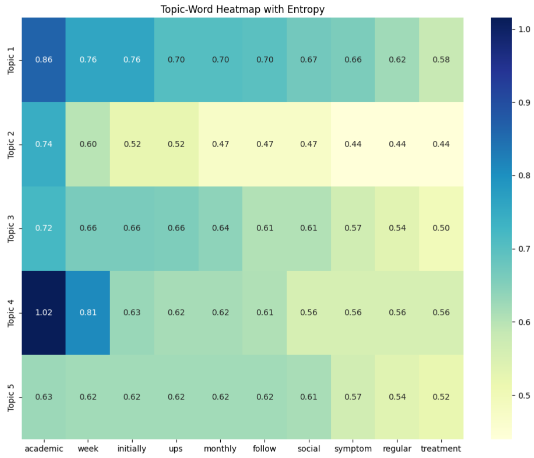

Figure 5: Entropy heat-map analysis of treatment plans generated from LLMs by using APA study cases

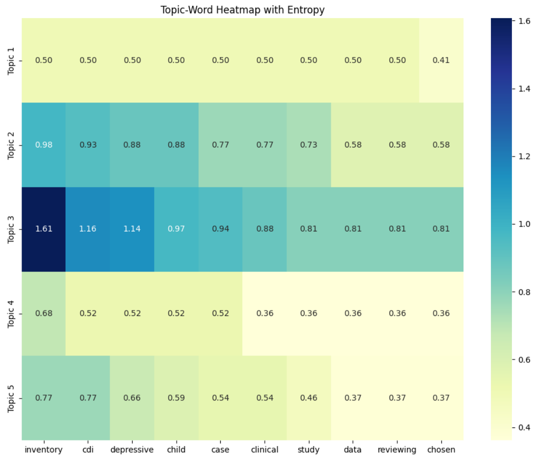

Figure 6: Entropy heat-map analysis of treatment plans based on APA study cases

Whitney U test of entropy showed a U statistic of 31.0 and a P-value of 0.162, suggesting that the difference in topic distribution between the text generated by LLMs and the text generated by human psychologists was not statistically significant, which supported our (H_0) that LLMs and human psychologists' recommendations were similar in diversity and uniformity.

5 Results

5.1 Interpretation of Topic-Word Frequency Analysis

The LDA analysis of the case studies uncovered a diverse range of topics associated with PTSD and depression, including treatment methods, patient living environments, and social factors such as school and social circles. These topics reveal

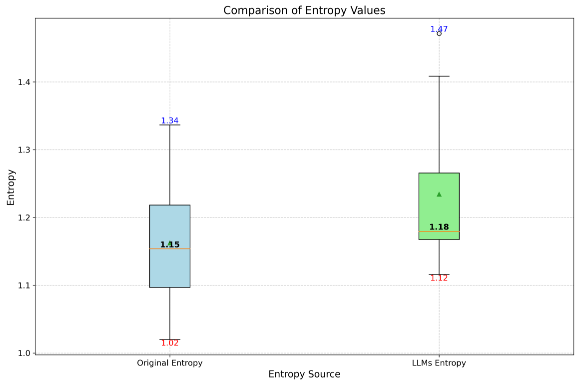

Figure 7: Comparative Entropy Analysis of LLMs-Generated and Expert-Designed Treatment Plans in Psychological Case Studies

both consistency and differentiation between the treatment recommendations generated by LLMs and those provided by human experts. While LLMs effectively identified general treatment topics like "medications" and "symptoms," they often lacked the depth and specificity evident in human-generated recommendations. For instance, human psychologists frequently mentioned specific therapies such as CBT, whereas LLMs tended to use broader terms like "treatment."

The document-topic distribution analysis further highlighted significant differences in the depth of engagement between LLMs and human experts. Human psychologists provided detailed and more professional terms, such as "CBT," whereas LLM-generated responses were more general. This suggests that, while LLMs can cover a wide range of relevant topics, they do not engage with the same level of depth, complexity, and detail as human experts. These findings align with the entropy analysis results, reinforcing the ongoing disparity in professionalism between large language models and human experts.

Based on the analysis of Figures 5 and 6, distinct differences were observed in the topic-word associations and entropy values between the treatment recommendations provided by human psychologists and those generated by LLMs. The heatmaps demonstrate that human-generated texts have more concentrated keyword relevance within specific topics, resulting in lower entropy values and indicating a more focused and detailed discussion. In contrast, LLM-generated texts display a broader but less focused distribution of keywords, leading to higher entropy values. This dispersion suggests that LLMs cover a wider array of topics but with less depth and specificity. For example, in Topic 3, the keywords in LLM-generated texts are more evenly spread across terms like "academic," "week," and "initially," reflecting a general approach rather than a detailed examination.

To statistically validate these findings, we tested the Null Hypothesis (H_0) that there is no significant difference in treatment recommendations between large language models (LLMs) and human experts. Using the Mann-Whitney U test on entropy values, we obtained a U statistic of 31.0 with a p-value of 0.162. This supports the null hypothesis, indicating no significant difference in the overall uniformity of topic distribution between LLMs and human experts.

Overall, the heatmap and entropy analysis highlight the need for further refinement of LLMs to enhance their ability to provide detailed and specific treatment recommendations, aiming to achieve a balance and depth similar to that of human psychotherapists. These observations underscore the ongoing need to improve LLMs for more effective therapeutic applications.

6 Conclusion

In this study, we have conducted a comprehensive analysis comparing Large Language Models (LLMs) with human psychologists in providing treatment recommendations for depression. Employing Latent Dirichlet Allocation (LDA) and entropy analysis, we found that while LLMs exhibit comparable diversity and uniformity in generating treatment recommendations, they lack the specificity and depth of human experts. LLMs effectively cover a wide range of topics but do not engage with the nuanced details that characterize human-generated recommendations. Despite this, the uniformity and diversity in LLM-generated recommendations suggest significant potential for their application in mental health care. However, further improvements are necessary to ensure consistent and in-depth performance across therapeutic scenarios. This study provides valuable insights into the potential and challenges of integrating LLMs into mental health practices while providing a new methodology in evaluating text-based LLMs generated response, paving the way for future research to enhance the effectiveness and reliability of AI-driven therapeutic solutions.

7 Limitations

Although this study provides important insights into AI-based approaches to the comparison of virtual psychotherapists with human professionals, there are several limitations to be concerned about. First, the generalizability of our results may be limited due to the specificity of the case studies used from APA sources and the datasets on which LLMs were trained, and the conclusions of the study are not broadly representative due to the small amount of data. In addition, since we did not collect the latest study cases, and the LLMs was trained based on massive amounts of data, this may lead to the limitation of our analysis conclusions due to the fact that some of the treatments in our study cases are not the latest mainstream treatments. At the same time, our analysis relied on textual data through LDA models, which also limited our ability to consider non-verbal cues and clinical intuitions inherent in human treatment. In addition, the ability of large model systems to interpret complex human emotions and clinical contexts remains lacking, which may affect the depth of therapeutic interventions recommended by these systems. Ethical issues regarding privacy and data sensitivity, as well as the enormous computational demands for deploying LLMs in clinical settings, also pose significant challenges.

It is worth noting that in some cases during our LDA analysis, the Entropy of the subject distribution (Entropy) appeared to be greater than 1. This may be due to outliers in the data preprocessing steps (such as word frequency calculation, TF-IDF transformation, etc.), which further affects the output of the model. These outliers may be amplified in subsequent processing steps, resulting in a probability value greater than 1. Although these numerical errors usually do not significantly affect the overall performance of the model and the final analysis results, the handling of these anomalies needs to be considered in further research.

Moreover, LLMs are highly dependent on the variety and richness of the input data they are trained on. In cases where training data lack demographic diversity or contain biased information, this can lead to skewed or biased AI-generated recommendations and diagnoses. Therefore, while LLMs can significantly expand access to mental health services, the underlying biases in training datasets can limit the appropriateness and effectiveness of the recommendations provided, especially for underrepresented groups. Addressing these data biases is essential to ensure equitable mental health support across diverse populations. This aspect highlights the need for continual updates and the urgency of having a professional clinical dataset to mitigate biases and improve the accuracy and fairness of AI-driven mental health interventions. Further studies should focus on developing robust methods for continuous data validation and enhancement, introducing more comprehensive textual data analysis methods based on quantitative approaches, as well as the implementing comprehensive ethical frameworks to govern AI usage in mental health settings.

8 Acknowledge

Our paper underwent a thorough review by the University of Washington Human Subjects Division (HSD). On November 28, 2023, the HSD determined that our proposed activity does not involve human subjects as defined by federal and state regulations. Consequently, review and approval by the University of Washington IRB is not required.

This determination applies specifically to the activities described in our application (IRB ID: STUDY00019126). We appreciate the guidance provided by the HSD.

References

Theyazn H. H. Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad A. T. Ahmed. 2022. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*, 19(19):12635.

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

Yara Badr, Fares Al-Shargie, Usman Tariq, Fabio Babiloni, Fadwa Al Mughairbi, and Hasan Al-Nashash. 2023. Classification of Mental Stress using Dry EEG Electrodes and Machine Learning. In *2023 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 1–5. ISSN: 2831-6878.

Borwin Bandelow, Sophie Michaelis, and Dirk Wedekind. 2017. Treatment of anxiety disorders. *Dialogues in Clinical Neuroscience*, 19(2):93–107.

Basabdatta Sen Bhattacharya and Vibhav Sinai Pissurlenkar. 2023. Assistive Chatbots for healthcare: a succinct review. *arXiv preprint*. ArXiv:2308.04178 [cs].

David M Blei. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Pastor López-Monroy, and Petr Motlicek. 2024. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *Preprint*, arXiv:2404.14463.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *Preprint*, arXiv:2305.13614.

Pim Cuijpers, Eirini Karyotaki, Erica Weitz, Gerhard Andersson, Steven D. Hollon, and Annemieke van Straten. 2014. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: A meta-analysis. *Journal of Affective Disorders*, 159:118–126.

Zheng Danna, Liu Danyang, Lapata Mirella, and Pan Jeff Z. 2024. TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness.

Saskia M. de Maat, Jack Dekker, Robert A. Schoevers, and Frans de Jonghe. 2007. Relative efficacy of psychotherapy and combined therapy in the treatment of depression: A meta-analysis. *European Psychiatry*, 22(1):1–8.

Zohar Elyoseph and Inbar Levkovich. 2024. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: Case vignette study. *JMIR Ment Health*, 11:53043.

Lauryn J Hagg, Stephanie S Merkouris, Gypsy A O'Dea, Lauren M Francis, Christopher J Greenwood, Matthew Fuller-Tyszkiewicz, Elizabeth M Westrupp, Jacqui A Macdonald, and George J Youssef. 2022. Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review. *Journal of Medical Internet Research*, 24(11):e33166.

Ylber Januzaj and Artan Luma. 2022. Cosine similarity – a computing approach to match similarity between higher education programs and job market demands based on maximum number of common words. *International Journal of Emerging Technologies in Learning (iJET)*, 17:258–268.

Rabie Karrouri, Zakaria Hammani, Roukaya Benjelloun, and Yassine Otheman. 2021. Major depressive disorder: Validated treatments and future challenges. *World Journal of Clinical Cases*, 9(31):9350–9367.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review. *arXiv preprint*. ArXiv:2401.01519 [cs].

Igor Kotenko, Yash Sharma, and Alexander Branitskiy. 2021. Predicting the Mental State of the Social Network Users based on the Latent Dirichlet Allocation and fastText. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 191–195. ISSN: 2770-4254.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6, Bandung, Indonesia. IEEE.

Jintian Lin. 2023. Measuring and analyzing the information entropy value of key Audit matters (KAMs) disclosure at the system and reporting scale. *Heliyon*, 10(1):e23255.

Manikanta Loya, Divya Anand Sinha, and Richard Futrell. 2023. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716. ArXiv:2312.17476 [cs].

Tania Manríquez Roa, Nikola Biller-Andorno, and Manuel Trachsel. 2021. *The Ethics of Artificial Intelligence in Psychotherapy*, chapter The Ethics of Artificial Intelligence in Psychotherapy. Oxford University Press.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1):1–6. Publisher: Nature Publishing Group.

Inaba Michimasa, Ukiyo Mariko, and Takamizo Keiko. 2024. Can Large Language Models be Used to Provide Psychological Counselling? An Analysis of GPT-4-Generated Responses Using Role-play Dialogues.

Adam S. Miner, Liliana Laranjo, and A. Baki Kocaballi. 2020. Chatbots in the fight against the COVID-19 pandemic. *npj Digital Medicine*, 3(1):1–4. Number: 1 Publisher: Nature Publishing Group.

Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pages 205–212, Lisbon Portugal. ACM.

Naveen S Pagad, Pradeep N, Khalid K. Almuzaini, Manish Maheshwari, Durgaprasad Gangodkar, Piyush Shukla, and Musah Alhassan. 2022. Clinical text data categorization and feature extraction using medical-fissure algorithm and neg-seq algorithm. *Computational Intelligence and Neuroscience*, 2022(1):5759521.

Domhnall Sheridan and Michael Carr. 2018. Mens sana: An Investigation into the Relationship between Psychological Traits and Academic Success of First Year Engineering Students. In *2018 3rd International Conference of the Portuguese Society for Engineering Education (CISPEE)*, pages 1–5.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Number: 7972 Publisher: Nature Publishing Group.

John Torous and Matcheri Keshavan. 2020. COVID-19, mobile health and serious mental illness. *Schizophrenia Research*, 218:36–37.

Yunpeng Xiao, Kyrie Zhixuan Zhou, Yueqing Liang, and Kai Shu. 2024. Understanding the concerns and choices of public when using large language models for healthcare. *arXiv preprint*. ArXiv:2401.09090 [cs].

A Generating Psychologically-Informed Treatment Recommendations Based on Detailed Case Information

Introduction: I will provide you with essential information about a client needing psychological consultation, which may include but is not limited to age, gender, symptoms, past diagnostic information, current life circumstances, treatment goals, and expectations. Assuming you are a highly professional psychotherapist, you are required to give me psychological counseling treatment recommendations.

Detailed Case Information:

Age and Gender: Provide the client's age and gender. Primary Symptoms: Describe the client's main psychological or emotional symptoms, such as anxiety, depression, panic attacks, etc. Diagnostic History: Outline any formal diagnoses the client has received in the past and the outcomes of any treatments they underwent. Current Life Circumstances: Describe the client's family environment, work or school situation, and social activities. Treatment Goals and Expectations: Specify the concrete expectations and goals of the client and their family for the treatment, including problems they hope to resolve and areas of life they wish to improve.

Treatment Recommendation Requirements

Diagnosis: You need to give a diagnosis based on the information I provided to you. You also need to give a detail reason of why you give this diagnosis.

Psychotherapy Plan: Theoretical Framework Selection: Based on the client's symptoms and diagnosis, choose an appropriate psychotherapy theoretical framework, such as Cognitive Behavioral Therapy (CBT), psychodynamic therapy, humanistic therapy, etc. Specific Therapeutic Techniques: Detail the therapeutic techniques to be used, such as exposure therapy, emotional restructuring, psychoeducation, etc.

Medication Recommendations (if applicable): Suggest possible pharmacological treatments, noting recommended types of medication, suggested dosages, and potential side effects.

Supportive Therapy Measures: Recommend supportive therapy measures such as group therapy, family therapy, or other community resources to enhance the effectiveness of the primary treatment plan.

Lifestyle and Behavioral Advice: Provide recommendations for lifestyle adjustments that improve overall health and psychological state, including regular physical activity, healthy eating, and good sleep habits.

Monitoring and Adjustment: Describe the proposed evaluation and monitoring plan to regularly check the effectiveness of the treatment and adjust the treatment plan as needed.

Output Format Requirements: Please provide the treatment plan in a report format, where each section is clearly titled and thoroughly described. Language and Expression:

Use precise professional terminology, ensuring that language is clear, rigorous, yet empathetic and understanding toward the client.

Ethical Considerations

B Study Cases and ChatGPT-Generated Responses Used in the Evaluation

The responses were generated on 5/7/2024 and 5/8/2024 by OpenAI's large language model, GPT-4. The links to the original chat history with Chat-

GPT were listed below:

Study Case 1: https://chat.openai.com/share/23f1b1f1-021b-4d82-be34-dd71ba6d1348

Study Case 2: https://chat.openai.com/share/91c6bbca-cd58-4510-8894-55ad9d773112

Study Case 3: https://chat.openai.com/share/73dda11f-afeb-4fb4-b5e1-4faa14cb4c72

Study Case 4: https://chat.openai.com/share/a721c2a6-1405-4bb9-a1dd-ea80beb78a9a

Study Case 5: https://chat.openai.com/share/b94e1d7f-7f52-49a3-b0ab-9e14c74cbf1a

Study Case 6: https://chat.openai.com/share/78605bcd-473d-4e83-b8c0-467700083251

Study Case 7: https://chat.openai.com/share/78605bcd-473d-4e83-b8c0-467700083251

Study Case 8: https://chat.openai.com/share/80dc4e95-07dd-4bae-a5ec-2c20d12b41fc

Study Case 9: https://chat.openai.com/share/41571272-52f1-4954-9431-87eb04a3cd16

Study Case 10: https://chat.openai.com/share/7993a913-b157-4374-9618-fc34470c8bad

C Entropy Comparison Table

Table 2: Entropy Comparison

Case	**Orig. Entropy**	**LLM Entropy**	**% Diff.**	**Abs. Diff.**
Case Study 1	1.163029	1.167993	-0.43	0.004964
Case Study 2	1.019811	1.250628	-22.63	0.230817
Case Study 3	1.266897	1.130842	10.74	0.136054
Case Study 4	1.065192	1.115729	-4.74	0.050537
Case Study 5	1.144286	1.167156	-2.00	0.022870
Case Study 6	1.094543	1.471529	-34.44	0.376986
Case Study 7	1.208523	1.408343	-16.53	0.199819
Case Study 8	1.336413	1.189755	10.97	0.146658
Case Study 9	1.102466	1.270382	-15.23	0.167915
Case Study 10	1.221561	1.168837	4.32	0.052724
Mean Entropy	1.162272	1.234119		
Standard Deviation	0.097259	0.119291		

D Entropy Comparison Table - Box plot

Table 3: Entropy Comparison Box plot

Scores	Median	IQR	Q1	Q3	Min	Max
Entropy (Psychotherapist)	1.15	0.12	1.10	1.22	1.02	1.34
Entropy (Large Language Model)	1.18	0.10	1.17	1.27	1.12	1.47

E Image Analysis Results

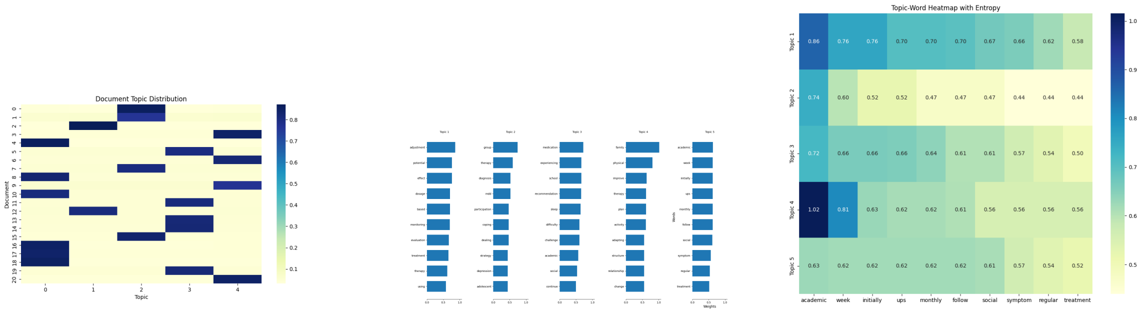

Figure 8: LLMs Analysis Results for Case Study 1

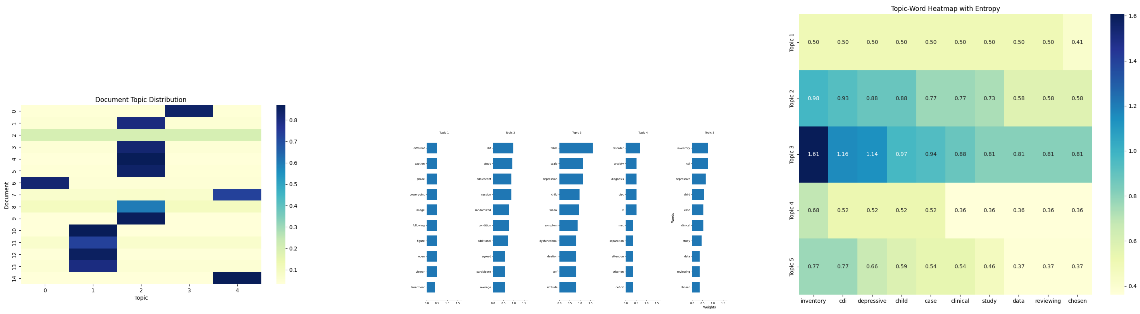

Figure 9: Original Analysis Results for Case Study 1

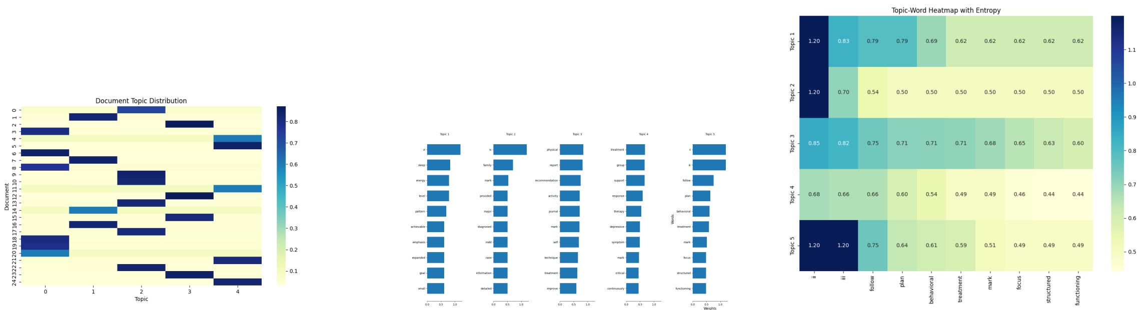

Figure 10: LLMs Analysis Results for Case Study 2

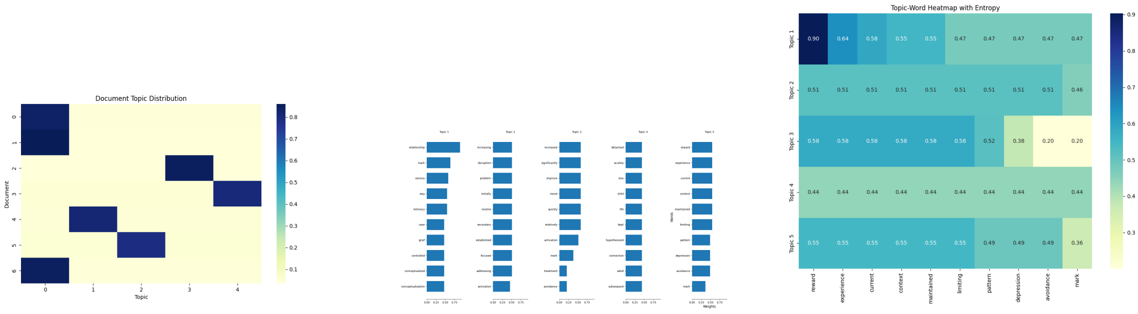

Figure 11: Original Analysis Results for Case Study 2

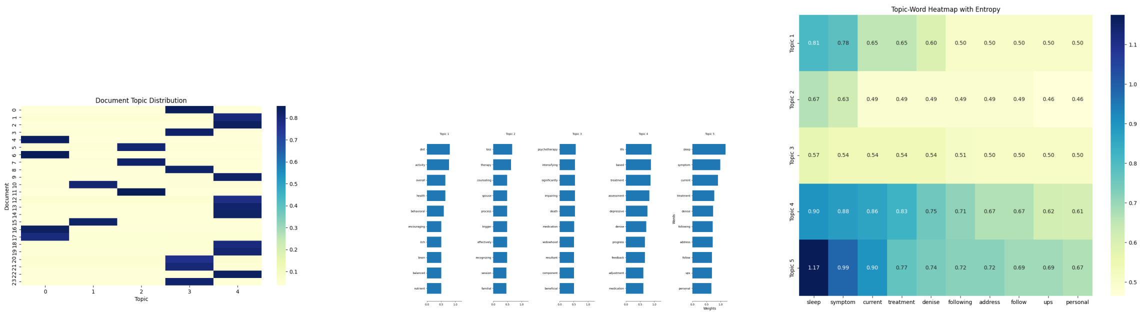

Figure 12: LLMs Analysis Results for Case Study 3

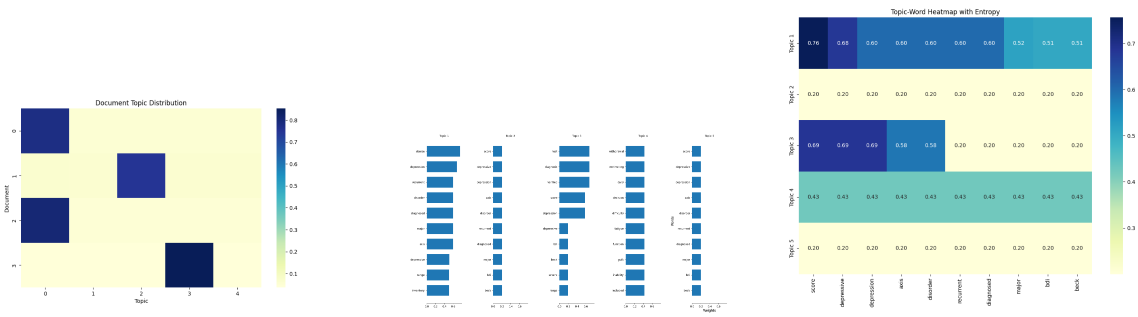

Figure 13: Original Analysis Results for Case Study 3

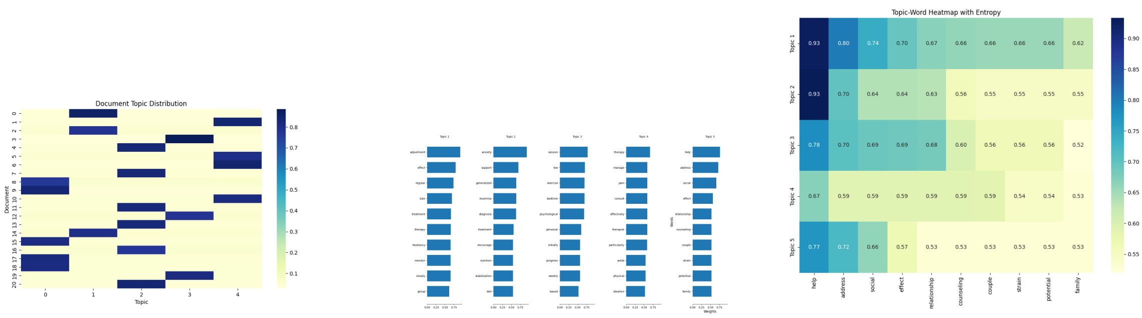

Figure 14: LLMs Analysis Results for Case Study 4

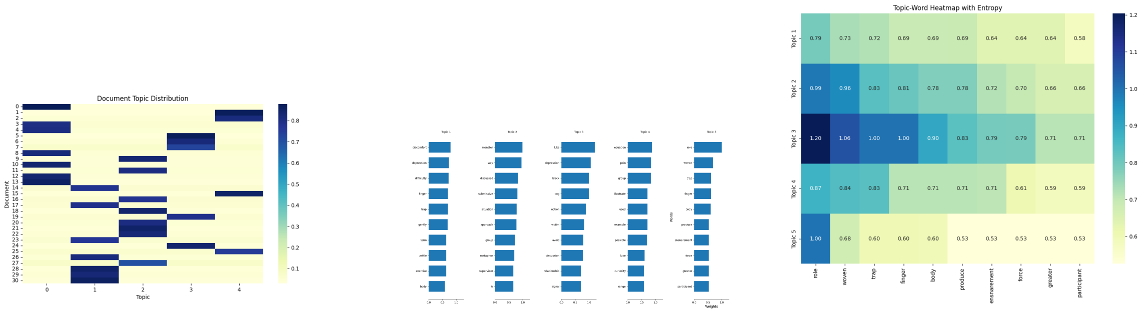

Figure 15: Original Analysis Results for Case Study 4

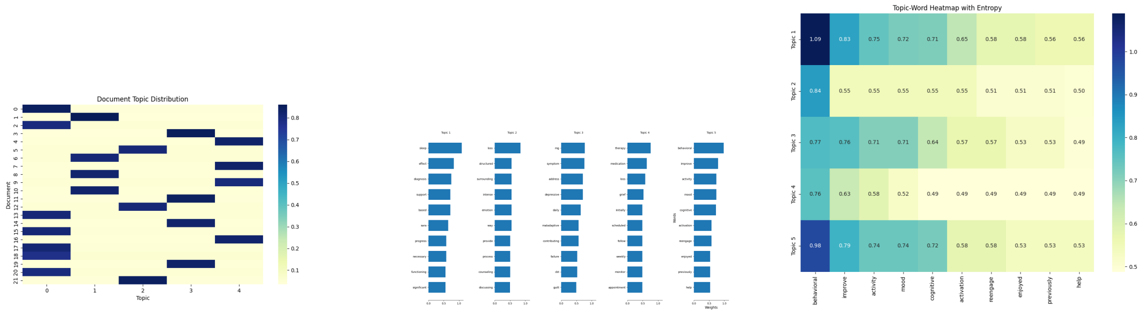

Figure 16: LLMs Analysis Results for Case Study 5

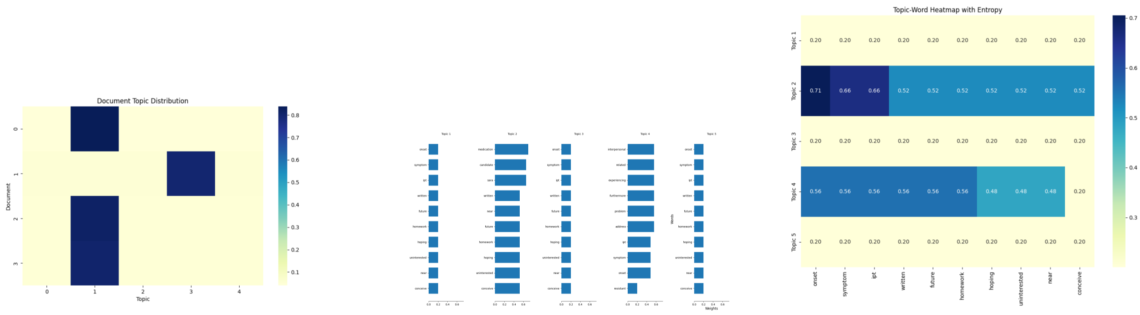

Figure 17: Original Analysis Results for Case Study 5

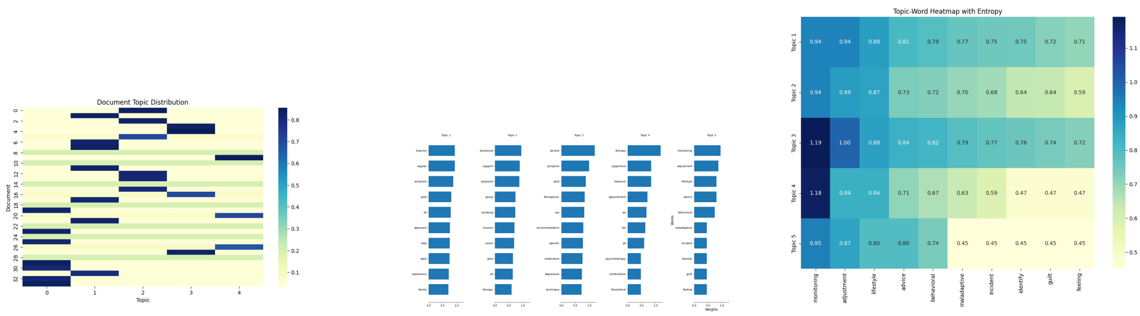

Figure 18: LLMs Analysis Results for Case Study 6

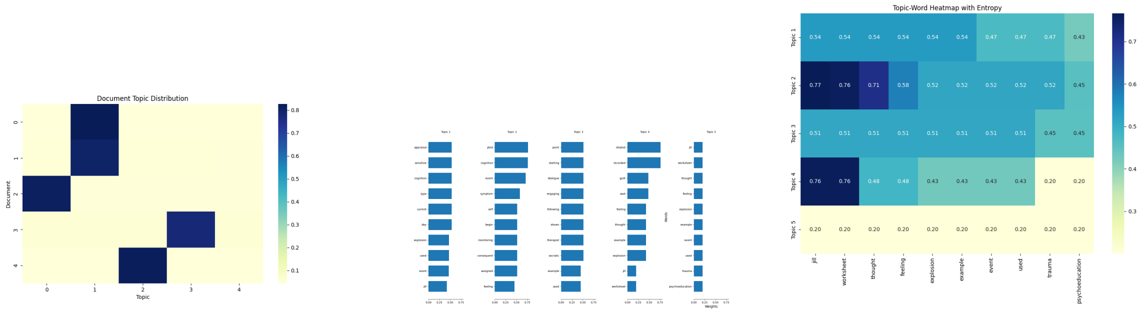

Figure 19: Original Analysis Results for Case Study 6

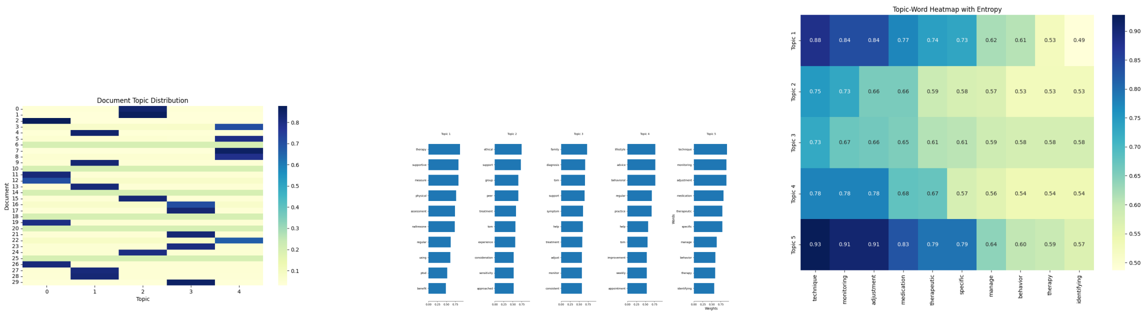

Figure 20: LLMs Analysis Results for Case Study 7

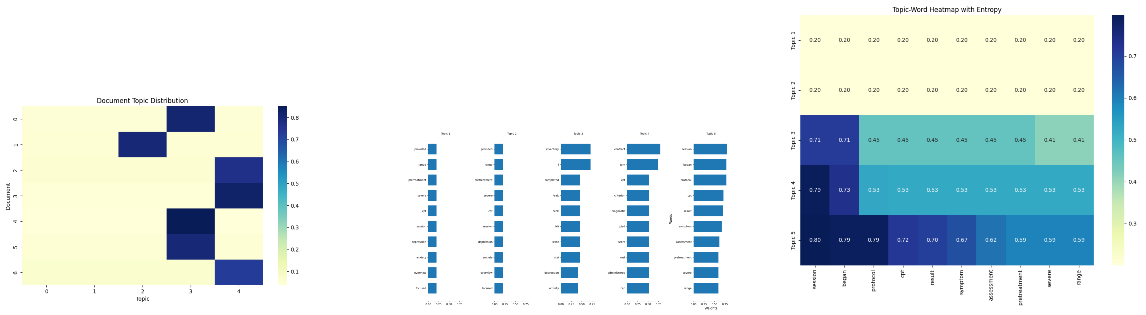

Figure 21: Original Analysis Results for Case Study 7

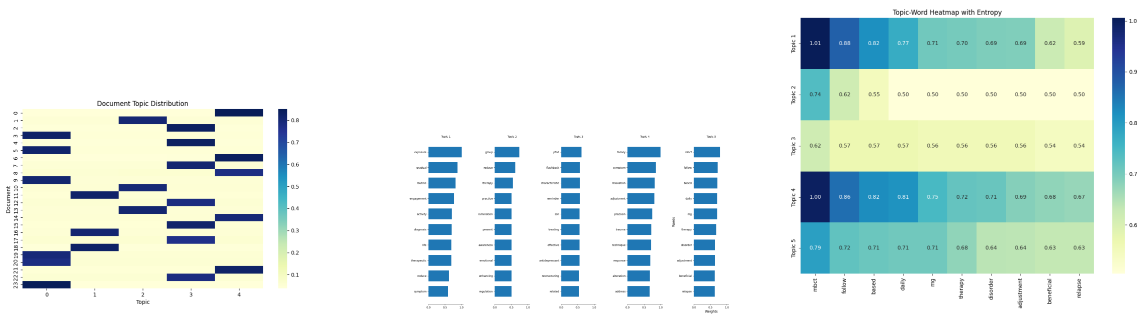

Figure 22: LLMs Analysis Results for Case Study 8

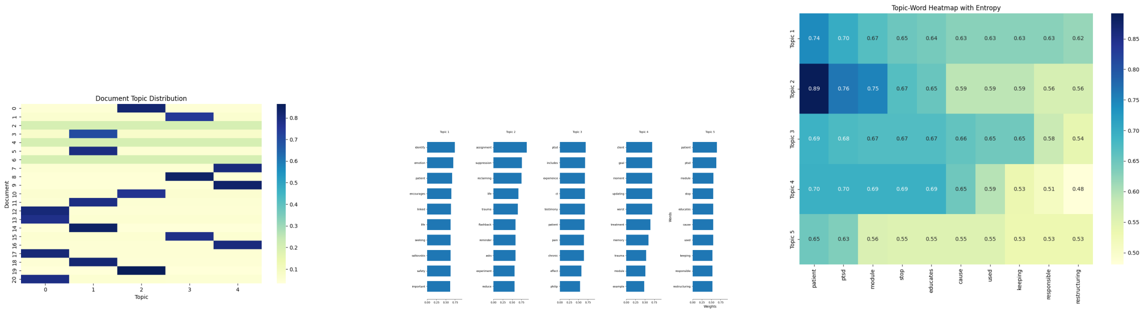

Figure 23: Original Analysis Results for Case Study 8

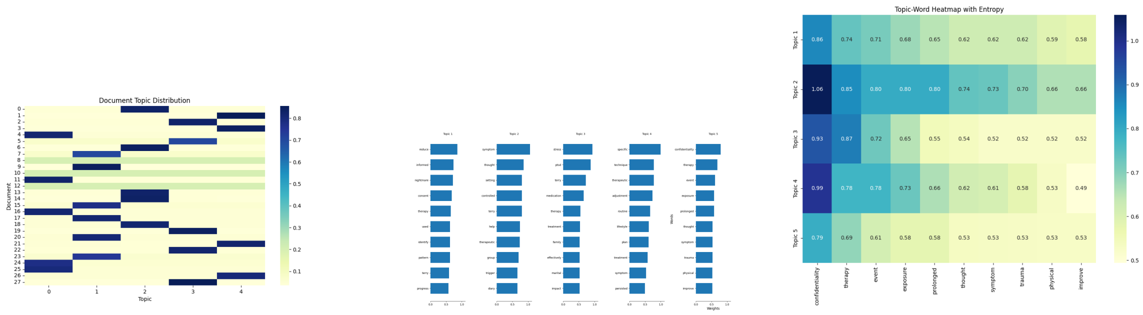

Figure 24: LLMs Analysis Results for Case Study 9

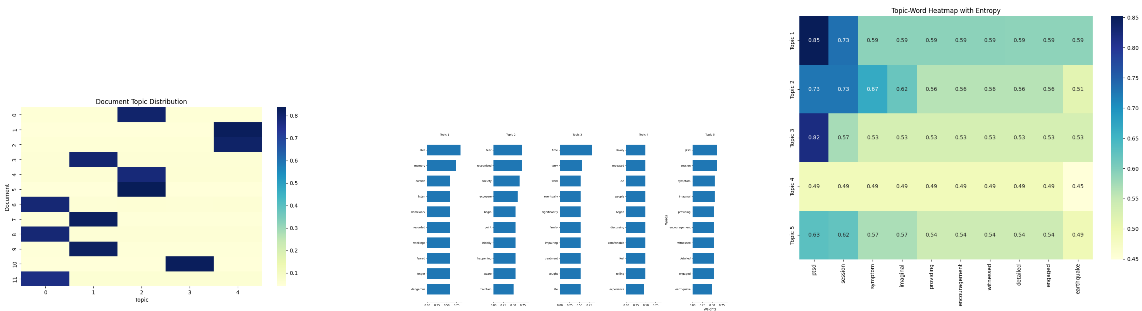

Figure 25: Original Analysis Results for Case Study 9

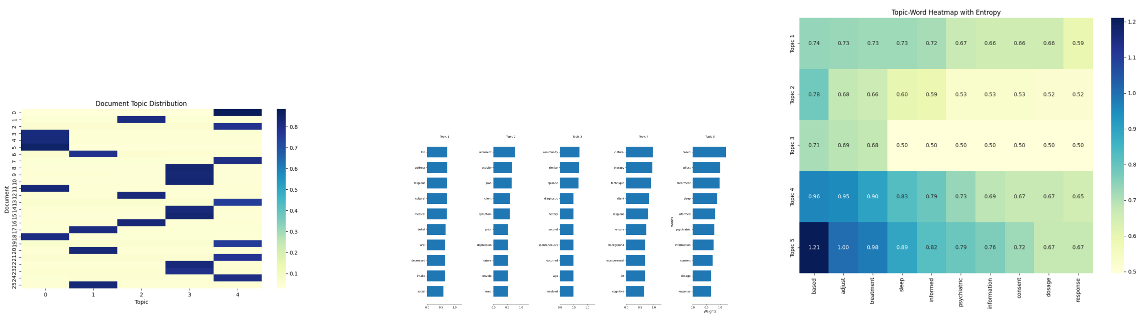

Figure 26: LLMs Analysis Results for Case Study 10

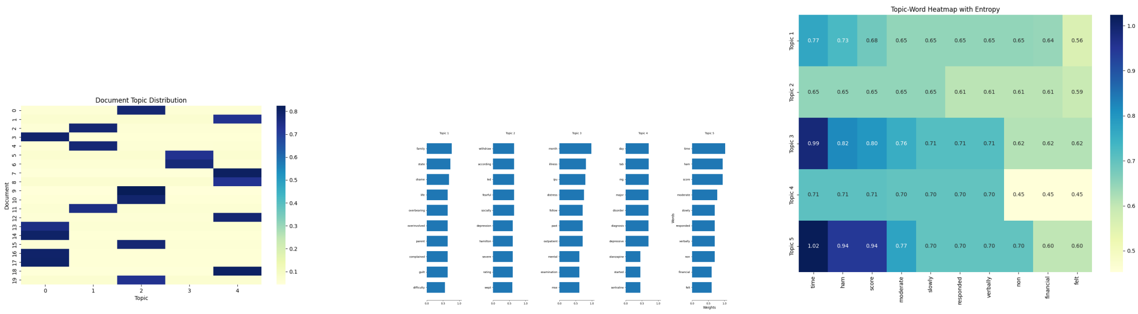

Figure 27: Original Analysis Results for Case Study 10