Generative AI in Psychological Therapy: Perspectives on Computational Linguistics and Large Language Models in Written Behaviour Monitoring

Jordan J. Bird
jordan.bird@ntu.ac.uk
Department of Computer Science, Nottingham Trent University
Nottingham, Nottinghamshire, UK

David Wright
david.wright@ntu.ac.uk
Department of English, Philosophy and Linguistics, Nottingham Trent University
Nottingham, Nottinghamshire, UK

Alexander Sumich
alexander.sumich@ntu.ac.uk
NTU Psychology, Nottingham Trent University
Nottingham, Nottinghamshire, UK

Ahmad Lotfi
ahmad.lotfi@ntu.ac.uk
Department of Computer Science, Nottingham Trent University
Nottingham, Nottinghamshire, UK

ABSTRACT

Technological intervention to support care areas that some people may not have access to is of paramount importance to promote sustainable development of good health and wellbeing. This study aims to explore the linguistic similarities and differences between human professionals and Generative Artificial Intelligence (AI) conversational agents in therapeutic dialogues. Initially, the MISTRAL-7B Large Language Model (LLM) is instructed to generate responses to patient queries to form a synthetic equivalent to a publicly available psychology dataset. A large set of linguistic features (e.g., text metrics, lexical diversity and richness, readability scores, sentiment, emotions, and named entities) is extracted and studied from both the expert and synthetically-generated text. The results suggest a significantly richer vocabulary in humans than the LLM approach. Similarly, the use of sentiment was significantly different between the two, suggesting a difference in the supportive or objective language used and that synthetic linguistic expressions of emotion may differ from those expressed by an intelligent being. However, no statistical significance was observed between human professionals and AI in the use of function words, pronouns and several named entities; possibly reflecting an increased proficiency of LLMs in modelling some language patterns, even in a specialised context (i.e., therapy). However, current findings do not support the similarity in sentimental nuance and emotional expression, which limits the effectiveness of contemporary LLMs as standalone agents. Further development is needed towards clinically validated algorithms.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Lexical semantics**; Reasoning about belief and knowledge; • **Human-centered computing** → *Human computer interaction (HCI).*

KEYWORDS

Generative Artificial Intelligence, AI, Large Language Models, LLMs, Computational Linguistics, Psychology

ACM Reference Format:
Jordan J. Bird, David Wright, Alexander Sumich, and Ahmad Lotfi. 2024. Generative AI in Psychological Therapy: Perspectives on Computational Linguistics and Large Language Models in Written Behaviour Monitoring . In *The PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA '24), June 26–28, 2024, Crete, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3652037.3663893

1 INTRODUCTION

In many parts of the world, access to mental health services is severely limited due to prohibitive costs, strained public services, or lack of availability of experts [11]. Therefore, technological intervention to support care areas that some people may not have access to has been found to be of paramount importance to promote sustainable development of good health and wellbeing [15]. Effective and natural communication is the cornerstone of therapy, and nuanced exchanges between the patient and the healthcare professional are the main form of care provided. In state-of-the-art research from Artificial Intelligence (AI), Generative models such as ChatGPT, have introduced a new possibility for the technological intervention of intelligent agents in settings which require natural communication. Although the efficacy of such approaches is explored in clinical settings, it is unknown which linguistic characteristics and behaviours Large Language Models (LLMs) exhibit compared to natural language produced by humans. That is, “What are the linguistic characteristics that distinguish human psychologists from AI-driven conversational agents in therapeutic dialogues?" Understanding the natural language intricacies of therapeutic dialogues, especially those that are paired (i.e., human and AI responses

PETRA '24, June 26–28, 2024, Crete, Greece

ACM ISBN 979-8-4007-1760-4/24/06
https://doi.org/10.1145/3652037.3663893

to the same query), is important in assessing the current potential of Generative AI to support or enhance psychological therapy, as well as highlighting future areas of improvement in steps towards clinical validation. The current study aims to bridge this knowledge gap through a systematic comparison of linguistic features in therapeutic dialogues held by human clients with human or Generative AI psychologists. To this end, the relevant work is reviewed in Section 2, followed by a description of the methodology in Section 3 and the results in Section 4. Finally, Section 5 discusses key findings and identifies a direction for future work. The data presented are publicly available [1].

2 RELATED WORK

In this section, state-of-the-art work in related fields is discussed and analysed towards the formation of this study's research question. Technological development in AI and chatbots presents an opportunity for the provision of psychoeducation and mental healthcare, particularly for underserved communities, but also raises specific ethical and legal issues (e.g., in relation to transparency, oversight, and regulation), due to potential discrimination risks [5, 8]. A proposed ethical framework relating to chatbots for psychological care includes guidance on non-maleficence, autonomy, justice, and explicability [16]. Benefit and risk in psychoeducation and mental health care are often dependent on language. Previous studies have explored the differences between human and generative text in language use [4]. The current paper compares human and AI psychologists on linguistic characteristics, with the aim of identifying ways in which benefit to psychological care can be maximised and risk of ethical breach can be mitigated.

Recent research has studied transformer technology for clinical support: For example, NLP techniques have been investigated with regard to recognising depression during human-robot conversation[1]. The findings of that study suggest 77% accuracy in recognising depression from natural language using Hierarchical Attention Networks and Long-sequence Transformer models. In [2], a self-attention transformer architecture was shown to be capable of predicting tokens in a sequence with 88.65% top-1 and 96.49% top-5 accuracy, given a dataset of mental health support questions and answers. Regarding the use of AI in interventions, Fitzpatrick et al. [9] note that conversational agents appear feasible, engaging and effective when applied to the delivery of cognitive behavioural therapy (CBT). University students showed a reduction in depression symptoms (as measured using PHQ-9) when interacting with *Woebot* compared to a control group provided with a self-help book; although there was no difference between the groups in the reduction of anxiety symptoms (GAD-7). According to Durden et al. [6], symptoms of perceived stress and burnout were reduced after interacting with Woebot, which offers chat-based mental health support, over 8 weeks.

Farhat [7] advocates three main benefits of OpenAI ChatGPT as a mental health resource: personalised care, access to care and cost-effectiveness. Psychological support may be inaccessible to some due to a shortage of experts or prohibitions due to healthcare-related costs, but (at the time of writing) ChatGPT can be freely accessed remotely, using devices connected to the Internet with web browsing capabilities. However, users are often urged to seek professional help. Also, some responses to prompts were problematic, such as the proposal of prescription medication (in the form of a list), without the supervision of a physician. Therefore, Cheng et al. describe ChatGPT as having *immense potential* in the field, but with limited practical applications currently [3]. The position paper notes that an effort is required in prompt engineering to maximise appropriateness of output, as well as avoiding inaccuracies. The article concludes with the argument that GPT provides a foundation for automated psychotherapy, but clinically validated algorithms are required. Similar views are shared across other fields, that is, that i) generative AI has major potential for psychotherapy; and ii) technology is not yet ready for clinical use.

From a computational linguistics perspective, the current study investigates the linguistic characteristics that distinguish human psychologists from AI-driven conversational agents in therapeutic dialogues. To achieve this, we synthesise comparative responses to human queries from the point of view of a psychologist, and test the nuanced differences between this synthetic response and human psychologists. The exploration of differences between the two is important for future direction on the enhancement of generative algorithms in a step towards clinical approval.

3 METHOD

This section describes the methods followed by this study. This includes data collection and generation, preprocessing, and feature extraction, and finally the method followed for the analysis and statistical comparison of the data.

3.1 Data Collection

For the purposes of this study, data are collected from CounselChat[2], provided through the HuggingFace Hub platform[3]. The dataset contains a total of 3513 pairs of strings, where a context is provided by a human patient and a response is then provided by a human psychologist. Five rows had no psychologist response in the collected dataset and were thus removed, leaving 3508 pairs of strings remaining, which were used for this study.

3.2 Synthetic Data Generation

The LLM used to generate data for this study is *Mistral-7B-Instruct-v0.2* developed by MistralAI [10]. The model is chosen for the purposes of this study due to its open source licence and competitive ability compared to leading private models. This includes the closed-source ChatGPT (OpenAI) and the Meta LLaMA model, which is open source but with a licence that dictates that it cannot be used to train other language models[4]. According to the MISTRAL authors, there are several main differences in the architecture of Mistral-7B compared to LLaMa, including the introduction of Sliding Window

[1] The dataset can be downloaded from: https://www.kaggle.com/datasets/birdy654/human-and-llm-mental-health-conversations

[2] Available online: https://counselchat.com [Last Accessed: 02/02/2024]

[3] Available online: https://huggingface.co/datasets/Amod/mental_health_counseling_conversations [Last Accessed: 02/02/2024]

[4] According to Section 1, part b, clause v in: https://github.com/facebookresearch/llama/blob/main/LICENSE [Last Accessed: 02/02/2024]

<s>[INST] You are a psychologist speaking to a patient. The patient will speak to you and you will then answer their query. [/INST] Okay. Go ahead, patient. I will answer you as a psychologist. [INST] Patient: [QUERY] Psychologist: [/INST]

Figure 1: Prompt used as input to the LLM. Red: start of string, Green: Instructions, Blue: patient query extracted from the dataset.

Attention, Rolling Buffer Cache, and Pre-fill and chunking. Figure 1 shows the prompt used as input to the LLM to generate the data for this study. The <s> tag represents the start of a string, and the instructions are enclosed in the [INST] and [/INST] tags. [QUERY] is replaced by the patient's query from the dataset. The initial dataset is iterated with each patient query inserted into the prompt and the LLM output stored in the dataset as a synthetic equivalent.

3.3 Linguistic Feature Extraction

For the purposes of this study, ten categories of linguistic features are extracted from a given text, which are detailed in this subsection. The features are chosen based on the criteria of producing a comparable fixed-length vector, thus retaining the possibility of usage in statistical analysis and, in the future, machine learning.

3.3.1 Basic Text Metrics. Initially, 6 basic features from the text are extracted. These are the number of characters, words, sentences, and unique words, as well as the average length of sentences and words.

3.3.2 Lexical Diversity and Richness. Following basic text metrics, diversity and richness measures are extracted. In this case, *diversity* refers to measurements considered to represent the variety of unique words within a text, and *richness* refers to the depth and sophistication of a vocabulary within a given text. The measures extracted include the following:

Type Token Ratio $TTR = \frac{V}{N}$, where V is the number of unique words and N is the total number of words. Higher values denote greater variety in vocabulary.

Yule's K: $K = 10^4 \times \frac{\sum_{i=1}^{V} i^2 \cdot f_i - N}{N^2}$, where K is a quantification of richness, and f_i is the frequency of the i^{th} word type. Higher values suggest greater diversity in the text.

Simpson's D: $D = \frac{\sum_{i=1}^{V} f_i(f_i - 1)}{N(N-1)}$, where D is the probability of two randomly selected tokens are of the same type, aiming to quantify repeated uses of words. Lower values indicate higher diversity, since higher values are derived when randomly selected tokens differ.

Herdan's C: $C = \frac{\log N}{\log V}$, where C is a logarithmic measure of vocabulary diversity, calculated by the logarithm of the total words N divided by the log of unique words V. Similarly to D, lower values denote greater diversity.

Brunét's W with constant -0.165: $W = N^{\left(V^{-0.165}\right)}$, where the total words N are raised to the power of unique words V, and a constant is used to prevent distortions given longer input text. Lower values of W denote higher richness within the vocabulary of a given text.

Honoré's R: $R = 100 \times \frac{\log N}{1 - \frac{V_1}{V}}$, where R quantifies the relationship between total words N, unique words V, and words that appear only once (hapax legomena) V_1. Higher values suggest greater richness in the vocabulary, particularly in cases where a wide range of infrequently used words appear.

3.3.3 Readability Scores. Several formulae are considered when estimating the ease of reading of the text. These include the following:

Kincaid Grade Level: Kincaid $= 0.39\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$, which is an indication of the US school grade level required to comprehend a given text.

The Automated Readability Index (ARI): ARI $= 4.71\left(\frac{\text{Characters}}{\text{Words}}\right) + 0.5\left(\frac{\text{Words}}{\text{Sentences}}\right) - 21.43$. ARI is an estimation of the grade level required to understand text given character counts.

The Coleman-Liau Index: Coleman-Liau $= 0.0588L - 0.296S - 15.8$. Coleman-Liau is another US grade prediction for text comprehension, where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

The Flesch Reading Ease: Flesch $= 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right)$, which is an indication of readability given the lengths of words and sentences.

The Gunning Fog Index: Gunning Fog $= 0.4\left[\left(\frac{\text{Words}}{\text{Sentences}}\right) + 100\left(\frac{\text{Complex Words}}{\text{Words}}\right)\right]$, which is an estimate of how many years of formal education are required to understand a given text upon first reading it. Thus, the complexity of the words is considered.

Läsbarhets Index (LIX): LIX $= \frac{\text{Words}}{\text{Sentences}} + \frac{100 \times \text{Long Words}}{\text{Words}}$, which scores the difficulty of a text considering the lengths of words and sentences.

SMOG Index: SMOG $= 1.0430\sqrt{\text{Polysyllable Words} \times \frac{30}{\text{Sentences}}} + 3.1291$. SMOG is an estimate of how many years of education are required to understand a text, with a focus on words that contain more than one syllable (polysyllabic).

Andersson's Readability Index (RIX): RIX $= \frac{\text{Long Words}}{\text{Sentences}}$, which scores the readability of a text given the number of long words in relation to the number of sentences.

The Dale-Chall Readability Formula: Dale-Chall $= 0.1579\left(\frac{100 \times \text{Difficult Words}}{\text{Words}}\right) + 0.0496\left(\frac{\text{Words}}{\text{Sentences}}\right)$, which is an evaluation of readability based on words easily understood by 4^{th} grade students in the United States and sentence length.

3.3.4 Sentence Structure. At the sentence level, the structure is transformed into numerical features. Sentence structure features include the following: The number of passive sentences within the text, which is calculated via checking for a Part-of-Speech tag VBN (past principle verb, e.g. "the food has been **eaten**"), and any of the following: VBZ (3^rd person singular present tense verb e.g., "they **eat** the food"), VBD (past-tense verb e.g., "they **went** to University"), or VBG (present participle verb, for example, "I am **running**"). The mean type token ratio of sentences $TTR = \frac{V}{N}$. The mean words per sentence and the words per paragraph. Finally, the usage of *to be* verbs ("the sky **is** blue"), auxiliary verbs (helping to form the present perfect tense e.g., "**have** you been to Greece?"), and nominalisation (conversion of terms into nouns, e.g., "did you make a **decision** where to publish?" where "to decide" has been nominalised to "decision").

3.3.5 Word Usage and Frequency. Features based on word usage and their frequency include the following: The frequency of pronouns in a given text (e.g. "**his** paper got minor corrections, but **he** eventually got it published"). The frequency of function words, which include several types of term that have little meaning alone but provide grammatical structure (e.g., "**can you** finish **the** manuscript **and** submit **it to** the conference?"). The usage of conjuction words (words that join words, phrases, clauses, or sentences, for example, "his paper got minor corrections, **but** he eventually got it published"), the usage of pronouns, and finally the usage of prepositions.

3.3.6 Punctuation and Style. Features derived from punctuation of sentence style include the frequency of punctuation usage, and the number of sentences beginning with pronouns, interrogative words (who, what, where, when, why, how, which, or whose), articles (a, an, the), subordinations (because, although, etc.), conjuctions (for, and, nor, but, or, yet, *or* so), and prepositions (in, on, at, by, etc.).

3.3.7 Sentiment and Emotion. The valence and emotional analyses of the text include the following features: The polarity of the sentiment (-1: negative, 1: positive) and the subjectivity of the sentimental value (0 to 1, where 1 is the opinionated sentiment). In addition, scores for the detection of fear, anger, anticipation, trust, surprise, sadness, disgust, joy, and overall positive and negative emotion scores.

3.3.8 Named Entity Recognition (NER). For each text, NER is performed to count the usage of the following tagged Part of Speech: PERSON (an individual's name), NORP (Nationalities, Religious or Political Groups), FAC (Facilities, Ergo buildings), ORG (Organisations), GPE (Geo-Political Entities), LOC (Non-GPE locations), PRODUCT (manufactured objects), EVENT (named events), WORK_OF_ART (titles of pieces of creativity), LAW (named legal works such as constitutions or acts), and LANGUAGE (names of natural languages).

3.3.9 Detailed Sentence Information. Finally, detailed information is collected at the sentence level. These features include: Average characters and syllables per word. Average characters, syllables, words, types of words, paragraphs, long words, complex words, and Dale-Chall complex words per sentence.

3.4 Data Analysis

Data analysis is performed in two parts in this work. First, given the raw text (following stop-word removal), word clouds are generated to initially explore whether the diversity of vocabulary between the human and AI-generated text can be visually observed. Following this, the most common words are considered before comparing the mean values of the readability metrics given the human-written and AI-generated texts. Given that it was discovered that humans often had a richer vocabulary than the LLM with the exception of Honoré's R value, this finding is explored further by exploring the count and frequency of hapax legomena (unique words that appear only once in a given text).

Table 1: Frequencies of the 10 most common words within the two sets of text.

Rank	Human Psychologist		Large Language Model	
	Word	***Freq.***	***Word***	***Freq.***
1	may	2900	help	6539
2	feel	2689	may	3813
3	would	2353	important	3453
4	help	2286	remember	3452
5	like	2240	support	3249
6	relationship	1907	feelings	3191
7	time	1885	relationship	3004
8	people	1765	health	2745
9	therapist	1666	understand	2733
10	know	1581	mental	2714

Following this text-level exploration, the features described in Section 3.3 are then extracted. The Wilcoxon signed-rank test is then performed for the paired feature sets between human- and AI-generated text, in order to explore which features are statistically significant between the two. Significance may suggest a difference in usage or ability, whereas statistical non-significance may suggest that both the humans and large language model are implementing such behaviours in both of their texts.

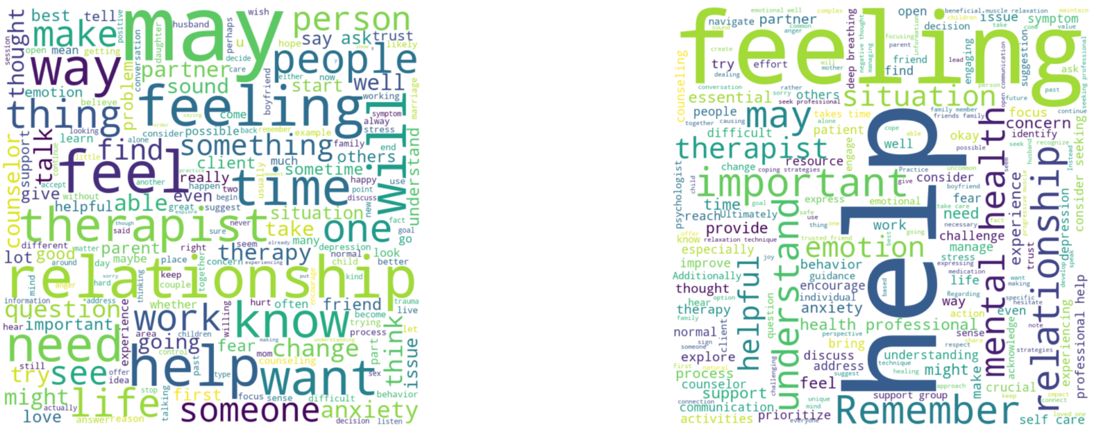
3.5 Experimental Hardware and Software

The experiments in this work were executed on a GPU server with 6 Intel Xeon(R) Platinum 8160 CPUs at 2.1GHz and 4 NVidia RTX A2000 12GB GPUs. All language model inference was performed with the HuggingFace library [17]. The features were extracted using the TextBlob [13], NRCLex [14], and NLTK [12] libraries.

4 RESULTS AND ANALYSIS

Figure 2 shows word clouds generated from both the human-written and LLM-generated text. The visual distribution in these figures suggest that the human text contains a more diverse vocabulary, due to the more similar size of the tokens. In the AI wordcloud, on the other hand, the terms *help* and *feeling* are over-represented compared to the other tokens, suggesting there is a strong emphasis on a smaller vocabulary. Beyond wordclouds, Table 1 shows the 10 most common words within the two sets of text. As expected from the word cloud, the word with the highest frequency in the AI-generated text was *help*, with 6539 instances. The second most common word, *may*, occurred 2726 fewer times, with a total of 3813 instances. Moreover, the frequencies of the top 10 tokens in the human responses had a standard deviation 418.47 compared to to 1073.19 for the LLM responses, further suggesting a significant difference in vocabulary.

Figure 3 shows a selection of lexical metrics extracted from the two sets of responses in the psychology dataset. It can be observed that the mean TTR is higher than the AI model for that of the human, indicating that on average there is a wider variety of words used in relation to the total number of words. Yule's K shows a similar result, since lower values indicate higher lexical richness, although it must be noted that the mean difference between these two values is 0.1. Similarly, for Simpson's D, where lower values indicate higher diversity, the human text scored slightly lower at 0.38 compared to 0.4 for the LLM. Furthermore, human text was



(a) Word Cloud for Human Psychologist Responses

(b) Word Cloud for AI Language Model Responses

Figure 2: Word clouds generated from human and LLM responses. The distribution suggests a more diverse vocabulary within human responses.

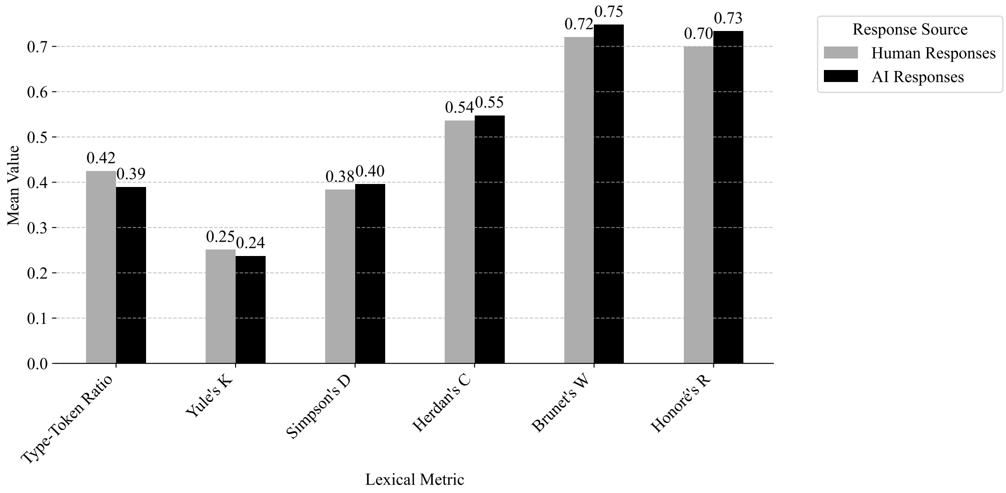


Figure 3: Mean lexical richness and diversity metrics observed within the two sets of text.

shown to score higher Herdan's C and lower Brunet's W compared to the LLM, both indicators of higher lexical richness. On the other hand, the LLM text was shown to have a higher Honoré's R value, $R = \log \frac{N}{1-\frac{V_1}{V}}$. On average, this suggests a higher number of hapax legomena (words that appear only once). This is explored further in Table 2 where single-use unique words are analysed; it can be observed that while the LLM has a higher count of hapax legomena (6392 compared to the human 4047), it can also be observed that the human psychologists make use of a greater number of tokens (26015) over the LLM's 16481.

Considering all of the metrics, the majority (with the exception of Honoré's R) suggest that the human text has higher lexical richness and diversity. Although this could be for many reasons regardless, the insight into the difference within these metrics shows that there are linguistic differences between the human and AI text, meaning that the methodology of communication in a psychological context may be different for the two. Furthermore, this finding could represent evidence of AI-based formulaic language patterns compared to nuanced and context-adapted natural human language.

The results of the Wilcoxon signed-rank test can be found in Tables 3 and 4. Although most of the features have statistically significant distributions between the human- and AI-generated dataset, nine do not. This suggests that there are some similarities between the writing behaviour of both the human and the algorithm in this particular case. Features that were not significantly different include function words, which suggests that the LLM's use of language structure is comparable to that of the human psychologists. The absence of significant differences between human and

Table 2: Hapax Legomena analysis of the human and AI-generated responses.

Human Psychologists			**Large Language Model**		
Hapax Legomena	***Unique Tokens***	***% Hapax***	***Hapax Legomena***	***Unique Tokens***	***% Hapax***
4047	26015	15.56	6392	16481	38.78

Table 3: Wilcoxon signed-rank test p-values for the non-statistically significant linguistic features observed between the Human and AI-generated responses.

Feature	***p-value***
Function Word Count	0.9219
Type Token Ratio	0.2517
Word usage, pronoun	0.1629
Sentence beginnings, pronoun	0.2475
NER, Location	0.0913
NER, Event	0.3264
NER, Law	0.5111
NER, Language	0.2253
EMOT, disgust	0.0633

AI for type token ratios indicates a similar level of lexical diversity, suggesting that the LLM does not tend to repeat or excessively vary when producing responses when compared to human psychologists. The use of pronouns, used as part of empathetic communication, is interestingly also insignificantly different between the two. The LLM's equivalence to humans suggests mimicry of related language; similarly, the sentence beginnings with pronouns are likewise nonsignificant between the two. More research is required on whether the statistical significance has been affected by rarity. The results suggest that LLMs are becoming increasingly improved in modelling language patterns within therapeutic settings, which is promising for the possibility of technological intervention within psychology. Although true empathy requires true intelligence, the results show that mimicry of the distribution of empathetic language may be possible via language proficiency. All other features were considered statistically significant in their distribution. AI-generated responses are often much longer than human responses and consist of several paragraphs. Following this observation, it is pertinent that features such as the number of sentences, their average length and character counts, numbers of paragraphs, etc. are significant between the two classes of data. Given that experts tend to produce shorter responses, this suggests that the LLM's longer answers are inappropriate for therapeutic communication. As expected from the previous exploration, all lexical diversity and richness measures (except for the Type Token Ratio) were statistically significant between the two. Similarly, the use of emotion and sentiment was also significantly different between the two, suggesting a difference in the use of supportive or objective language used and that synthetic linguistic expressions of emotion also differ from those expressed by an intelligent being.

5 CONCLUSION AND FUTURE WORK

This study has explored linguistic characteristics within dialogues with human psychologists and a large-language Generative AI model. It was observed that there are statistically significant differences within many of the features, particularly in terms of vocabulary and emotional expression. The results suggest that AI can mimic the linguistic structure but lacks the nuanced understanding and adaptability that naturally feature in human interaction, instead expressing information in a less diverse way. As the fledgling field of Large Language Models inevitably continues to grow at a rapid pace, backed both by academics and industry, the results of this study have shown the importance of improving synthetic emotional intelligence and linguistic nuance towards the development of LLMs that can more accurately replicate complex linguistic features. It is important that LLMs, if used as tools of technological intervention in health, are both effective and empathetic, in the same ways that human psychologists are. Future work could explore additional testing methods, such as analysis of the correlation coefficient given a feature and source, deeper corpus linguistic analysis of the responses, as well as expert evaluations on the quality of the responses provided and whether they are acceptable from a clinical point of view.

REFERENCES

[1] Christian Alexander Belser. 2023. *Comparison of Natural Language Processing Models for Depression Detection in Chatbot Dialogues.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[2] Jordan J Bird and Ahmad Lotfi. 2023. Generative Transformer Chatbots for Mental Health Support: A Study on Depression and Anxiety. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments.* 475–479.

[3] Szu-Wei Cheng, Chung-Wen Chang, Wan-Jung Chang, Hao-Wei Wang, Chih-Sung Liang, Taishiro Kishimoto, Jane Pei-Chen Chang, John S Kuo, and Kuan-Pin Su. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and clinical neurosciences* 77, 11 (2023), 592–596.

[4] Ilker Cingillioglu. 2023. Detecting AI-generated essays: the ChatGPT challenge. *The International Journal of Information and Learning Technology* 40, 3 (2023), 259–268.

[5] Catherine Diaz-Asper, Mathias K Hauglid, Chelsea Chandler, Alex S Cohen, Peter W Foltz, and Brita Elvevåg. 2024. A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. *American Psychologist* 79, 1 (2024), 79.

[6] Emily Durden, Maddison C Pirner, Stephanie J Rapoport, Andre Williams, Athena Robinson, and Valerie L Forman-Hoffman. 2023. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". *Internet Interventions* 33 (2023), 100637.

[7] Faiza Farhat. 2023. ChatGPT as a complementary mental health resource: a boon or a bane. *Annals of Biomedical Engineering* (2023), 1–4.

[8] Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research* 21, 5 (2019), e13216.

[9] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.

[10] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[11] Martin Knapp, Michelle Funk, Claire Curran, Martin Prince, Margaret Grigg, and David McDaid. 2006. Economic barriers to better mental health practice and policy. *Health policy and planning* 21, 3 (2006), 157–170.

Table 4: Wilcoxon signed-rank test p-values for the statistically significant linguistic features observed between the Human and AI-generated responses.

Feature	***p-value***
Average word Length	< 0.0001
Pronoun Frequency	< 0.0001
Coleman-Liau Index	< 0.0001
Characters per word	< 0.0001
Syllables per word	< 0.0001
Words per sentence	< 0.0001
Number of paragraphs	< 0.0001
Complex words per sentence	< 0.0001
Long words per sentence	< 0.0001
Automated Readability Index	< 0.0001
Kincaid	< 0.0001
Gunning Fog Index	< 0.0001
Conjuction words	< 0.0001
Number of setences	< 0.0001
LIX	< 0.0001
Nominalisation words	< 0.0001
Complex words per sentence	< 0.0001
Flesch Reading Ease	< 0.0001
EMOT, trust	< 0.0001
Average sentence length	< 0.0001
Dale-Chall	< 0.0001
EMOT, positive	< 0.0001
Syllables per sentence	< 0.0001
To be verbs	< 0.0001
Characters per sentence	< 0.0001
Content word density ratio	< 0.0001
EMOT, surprise	< 0.0001
Character count	< 0.0001
RIX	< 0.0001
Honoré's R	< 0.0001
EMOT, anticipation	< 0.0001
Word types per sentence	< 0.0001
Unique word count	< 0.0001
Sentences per paragraph	< 0.0001
Brunet's W	< 0.0001
SMOG Index	< 0.0001
Sentiment polarity	< 0.0001
Word count	< 0.0001
EMOT, negative	< 0.0001
Type Token Ration	< 0.0001
Vocab to total words ratio	< 0.0001
Yule's K	< 0.0001
Passive sentences	< 0.0001
Words per sentence	< 0.0001
Sentence beginnings, subordination	< 0.0001
Sentence beginning, preposition	< 0.0001
EMOT, sadness	< 0.0001
Sentence beginning, interrogative	< 0.0001
Auxiliary verbs	< 0.0001
Simpson's D	< 0.0001
NER, Geopolitical	< 0.0001
Punctuation frequency	< 0.0001
Sentence beginnings, article	< 0.0001
EMOT, fear	< 0.0001
NER, product	< 0.0001
EMOT, joy	< 0.0001
Sentence beginning, conjunction	< 0.0001
NER, Facility	0.0001
NER, person	0.0002
Sentiment subjectivity	0.0002
Herdan's C	0.0002
EMOT, anger	0.0007
NER, Organisation	0.0026
NER, Nationality, religious, political	0.0125
Preposition usage	0.0170
NER, artwork	0.0486

[12] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[13] Steven Loria. 2018. textblob Documentation. *Release 0.15* 2 (2018).

[14] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.

[15] F Vailati Riboni, B Comazzi, K Bercovitz, G Castelnuovo, E Molinari, and F Pagnini. 2020. Technologically-enhanced psychological interventions for older adults: A scoping review. *BMC geriatrics* 20, 1 (2020), 1–11.

[16] Giovanna Nunes Vilaza and Darragh McCashin. 2021. Is the automation of digital mental health ethical? Applying an ethical framework to chatbots for cognitive behaviour therapy. *Frontiers in Digital Health* 3 (2021), 689736.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).