CRGA Telemetry-Based Wellness Companion: An Adaptive and Privacy-Preserving LLM-Based Wellness Companion

Sasmit Bhattacharya
Department of Computing Technologies
School of Computing
College of Engineering and Technology
SRMIST, Kattankulathur
Chennai, India
sb4749@srmist.edu.in

Rajveer Singh Bagga
Department of Computing Technologies
School of Computing
College of Engineering and Technology
SRMIST, Kattankulathur
Chennai, India
rb9820@srmist.edu.in

Mrs. S. Sony Priya
Assistant Professor
Department of Computing Technologies
School of Computing
Faculty of Engineering & Technology
SRMIST, Kattankulathur
Chennai, India
sonypris@srmist.edu.in

Dr. Karthikeyan M
Associate Professor
Department of Computing Technologies
School of Computing
Faculty of Engineering & Technology
SRMIST, Kattankulathur
Chennai, India
karthikm1@srmist.edu.in

Mrs. Jayanthi V
Assistant Professor
Department of Computing Technologies
School of Computing
Faculty of Engineering & Technology
SRMIST, Kattankulathur
Chennai, India
jayanthv4@srmist.edu.in

***Abstract*—Currently, there are two main issues with LLM-based wellness companions: too much verification of the users and keeping very sensitive information of the conversations in cloud servers. The solution to these problems is CRGA, which takes advantage of keystroke analysis and text analysis together to move from empathy reflection to coaching at the exact right time, when the user is ready. For the sake of privacy, CRGA replaces cloud storage with on-device storage in the form of JSON vaults, containing only crucial identity stakes and goals. Experimental evaluation showed that on-device arbitration was accurate for routing, provided a reduction of the data footprint by 18 times, and added only 41 milliseconds of latency. Thus, the method described above provides a very efficient and responsive wellness experience without requiring retraining of any underlying models.**

***Index Terms*—Large language models, sentiment analysis, cognitive state estimation, emotional support conversation, typing telemetry, on-device memory, privacy-preserving conversational AI, human-computer interaction**

I. Introduction

LLM-based conversation wellness assistants have made tremendous strides in moving from experimental systems to being implemented in practice. Field experiments reveal that, if properly designed, LLM-driven interventions can successfully help individuals reshape their negative thinking patterns and lower their arousal levels on a scale impossible for human therapists alone (Sharma et al., 2024). At the same time, literature reviews within the field suggest some significant flaws with regards to the tone, safety, and privacy of user information (Stade et al., 2024), and assessments of commercial companions show even greater flaws concerning attachment and handling of critical situations (Maples et al., 2024). However, despite the utility of the approach, standard development methods often lead to important deficiencies.

Two gaps are explored in this paper. The first gap is regarding behavioral inconsistency. Instruction-trained large language models (LLMs) that have been assigned the behavior of a coach demonstrate the tendency to generate more validating and affirmative answers even after the end-user finishes venting and starts considering ways to act. As shown quantitatively by Kang et al., LLMs under-utilize actions-based approaches in their response turns when the human coach is expected to switch from listening to guiding. This has always been the criticism directed towards digital assistants: "They listen but never help." The question is, therefore: How do we know if we should continue mirroring or start coaching?

The second issue pertains to memory and privacy. To preserve the illusion of continuity in each session, most wellness agents store a copy of the conversation in a cloud-based database, associated with the user account. Although this system offers an advantage in terms of personalization, it involves concentrating the most private type of text generated by the individual in one place. Advances in local computation (Alizadeh et al., 2023) and structured external memory for LLMs (Packer et al., 2023; Zhong et al., 2024) make it possible

to pursue other options. However, this is not yet the standard practice in existing deployments.

We believe that the two discussed gaps can be better resolved via a more holistic approach, since they both ultimately relate to one basic question: What should be the particular organized signal to the system to make the LLM behave in such a way that it makes the right contextual decision? This gap-resolution concept is referred to as CRGA. Every time a conversation happens, the analyzer running within a web browser evaluates the input and style of interaction from a user and calculates a concise vector of cognitive-state indicators and its scalar summary called cognitive receptivity. Depending on this scalar indicator, the gatepolicy (gatekeeper) determines whether the mirror prompt, coach prompt, or their combination is applied to the interaction. At the end of every session, the interaction is summarized in a concise JSON file with two parts: identity stakes and actionable goals.

From the viewpoint of the signal processing chain, the innovation seems quite straightforward. CRGA could be considered as a system that performs sentiment and cognitive state analysis to produce a signal that shows readiness to switch from empathic communication mode to the one where actions need to be performed, while saving historical data in long-term storage of the user's device. Novelty cannot be claimed about any of the system parts, as each one belongs to an active field of research [4], [10], [11], [14].Novelty is achieved through the integration of these pieces in a feedback loop which can be easily implemented as a web app without changes to LLM.

The rest of this paper is divided into the following sections. Section II is devoted to the review of related work. Section III outlines the methodology that has been followed for the three main components. Section IV discusses the system architecture with the aid of the illustration in Fig. 1. Section V highlights the benchmarks and the user study conducted. Section VI concludes.

II. Related Work

Work on mental health and wellness systems using large language models (LLMs) has moved quickly forward. In this respect, the field work by Sharma *et al.* is quite extensive as it involves a real-world application of mental health system in which they show how the use of an LLM-based cognitive restructuring module helped many users control their emotions and stop thinking negatively about situations [2]. Their work builds upon earlier research where they formally define cognitive reframing as a linguistic task [8]. More broadly, Stade et al. raise an important point, namely that evaluations must be done using rubrics relevant for the domains, unlike the generic methods of evaluation used in natural language processing, which is highly relevant for our current evaluation criteria [1]. Conversely, the study on Replika and a cohort of users by Maples et al. raises concerns of attachment among other issues [9].

The design for how a conversational agent should switch between its different support strategies is most explicitly defined within the field of emotional support conversations (ESC). This includes Liu et al.'s creation of the ESConv dataset along with an eight-strategy taxonomy that informs much of the later literature [4].Lookahead strategy prediction was proposed by Cheng et al [5], and Tu et al. showed that mixing multiple strategies together could outperform choosing only one [6]. Some of the more recent literature has started to decouple the process of strategy prediction from response generation, which includes modeling heterogeneous discourse dynamics as well as making large gains in reducing strategy-preference biases compared to earlier baselines [7]. The particular result that motivated this work was reported by Kang et al. where their analysis found that large language models consistently favor using validation and suggestion strategies too infrequently and too often, respectively, especially when the dialogue stage requires taking action [3].The key difference between CRGA and these other works is that CRGA utilizes additional typing telemetry in conjunction with the text, and secondly makes a binary decision on two combined strategies.

It is highly likely that typing patterns contain some form of information that can be used to infer a person's emotional state. In a study conducted by Knol et al., using the BiAffect keystroke capture system, typing behavior characterized by decreased motor activity was found to predict anhedonia, whereas typing speed and keystrokes were found to be secondary predictors for irritability [10]. In a more extensive study involving adolescents, Braund et al. found relatively weaker sex-based correlations between keystrokes and standard mental health indicators, while cautioning against the use of one feature alone for predicting an outcome [11]. We view these results as promising; typing behavior contains valuable information that can be used as prior information but not as a conclusive predictor on its own. In a similar study involving linguistic markers, Funkhouser et al. showed that a small set of linguistic predictors can convey a meaningful yet limited amount of information [12].

Memory in LLM dialogue has been tackled by means of virtualized context management [14](MemGPT),and of affect-aware memory decay [15] (MemoryBank). Both of them are aimed at an extended effective context and rely on the use of external storage which has properties orthogonal to the inherent memory system itself. Our contribution consists of a local vault based on the MemGPT idea of a memory that is being summarized and paged back into the prompt, however, applied in the opposite way, namely, the target is to achieve minimal and bounded memory that exists purely locally on the user's side.As regards inference methods, approaches like flash-aware weight streaming [13] have made it increasingly easy to run fully locally, and our memory layer is designed in a way that switching to a local decoder does not require any change of the schema. Lastly, as regards CBT-style reasoning and empathy detection, we refer to Diagnosis-of-Thought prompting [16] and EPITOME [17]by Sharma et al., respectively.

III. Methodology

CRGA could thus be seen as a process that takes three steps – sense, decide, remember. The first step processes the user's message as well as the user's typing pattern into something manageable – a set of concise, understandable indices. The second step combines all of the aforementioned into one readiness index and uses it in order to choose an optimal reaction strategy. The third step packs up everything learned during the session into something small yet structured and then stores it on the user's machine. The three steps are rather simplistic and that is what gives rise to a very important part of the overall argument – quite a lot of personalization and strategy selection can be done with not too many dynamic elements involved.

A. Sentiment and Cognitive-State Analysis

On receiving the message, the client extracts two components. One is the text component itself, where a series of shallow but meaningful features is extracted, such as the number of tokens in the message, the ratio of first-person singular pronouns, and the frequency of negative affect lexicons using a simple sentiment lexicon. This is because, during this step, the aim is to provide a consistent sentiment signal, not deep semantic extraction, which is left for future steps to be conducted within the LLM.

The second part is the typing telemetry collected inside the browser while the user composes the message. There are three characteristics that we collect: (1) the time elapsed between the first keypress and the sending event, (2) the average interkeypress latency, and (3) the typing speed in characters per second. On their own, each of these characteristics does not reveal much about the current state of the user. In combination, however, they create a texture that is surprisingly revealing. Short and fast messages using highly negative affect language differ from long and slow messages that rely extensively on first-person pronouns. While the distinction is rather vague, it aligns well with the difference between a person who is in an acute state of anxiety and a person engaged in reflective thought processes.

These properties give rise to three intuitive metrics: the anxiety metric, which rises with fast and short messages with negative sentiment; the reflection metric, which rises with slow and long messages that refer to the self; and the cognitive distortion metric, which signals the use of absolutist terms such as "always," "never," and "everything," as well as catastrophizing and overgeneralization, using the Diagnosis of Thoughts framework [16]. These metrics are generated by an easily interpretable rule-based aggregator and scaled to the unit interval. In this case, we explicitly choose to use a rule-based aggregator rather than a learned classifier. This is because a learned classifier would need a labeled corpus of wellness data, which we do not have access to. More importantly, it hides the very design choice that we would like a reviewer or a user to be able to inspect.

B. Cognitive Receptivity-Gated Arbitration

These three dimensions are combined into one number between zero and one that serves as an index of cognitive receptivity. Theoretically, the higher the index value, the more likely it is that the user will be reflecting on their experiences and willing to work with suggestions. If the user seems to be anxious or trapped in the distorting loop, the index takes lower values. A small addition from the history component slowly increases the index value for sessions where the majority of reflections are mirroring ones.

The arbitration rule is deliberately simple. In case there is high receptivity score, then a coaching prompt will be used, instructing the LLM to suggest a small and concrete next step, as well as reframing the thought based on Sharma et al. [8].In case there is a low receptivity score, then a mirroring prompt will be used, asking the LLM to stay present with and validate the emotionally important experience. In between these two values, there is an intermediary range of scores for which the blended prompt will be used, where the LLM mirrors first and suggests once. The use of two thresholds instead of one ensures that policy changes do not result from small score changes, while in the cold start period—in which enough text input has not yet been obtained to normalize the features—then a blended prompt is applied.

The design deliberately avoids the level of detail, i.e., eight-way policy prediction, normally considered in the ESC research community [4]. The pertinent issue in real-life applications of wellness systems is usually whether to listen or start coaching at each interaction point. The most basic choice involving only two policies and easiest to review in case of any errors is the two-policy gate.

C. Local Vault and Session Compression

Once again, the full transcription of the conversation is presented to the LLM after each interaction, along with an output prompt asking the program to create a JSON object consisting of two fields. The first field, referred to as "identity stakes," contains a list of all the things that the user believes are at stake for them in their life right now, including items like a particular deadline, a difficult relationship, or even maintaining a certain identity. The second field, referred to as "actionable goals," includes a list of the practical objectives laid out by the coaching strategy.

When the next session begins, the compacted vault is extracted from local storage and injected into the system prompt. Thus, the model receives a brief overview of the user's identity and intentions, but it does not gain access to the precise quotations spoken during the earlier dialogue. The degree of informativeness of the vault is intentionally lower than that of the raw dialogue data, and this reduction forms an element of the argument for privacy: even if the vault is stolen, the user's real sentences are not compromised. If the user flushes their local storage, they essentially eliminate all traces of themselves from the system's memory.

D. Multi-Scale Analysis and the Accountability Loop

Beyond arbitration for each turn, CRGA specifies three levels of analysis that function upon the vault, rather than within one turn. The first level is nanoscopic, where a pattern detector detects recurring language structures indicating cognitive distortions. The second level is microscopic, where a clustering pass reduces the themes of a session to a concise list of unique root causes of the user. The third level is macroscopic, where a cross-domain reasoner searches for recurring patterns in different aspects of the user's life, such as perfectionism shown in academic performance and later in personal interactions. These layers are not always presented to the user; they contribute to the formulation of the coaching policy when suggestions for tasks are necessary. An accountability cycle, overlaying the coaching process, monitors tasks completed by the user through sessions, starts a new session with an audit of the tasks, and explicitly addresses failures to complete tasks rather than ignoring them. During piloting, explicit handling avoided regression to affirmation only.

IV. System Architecture

The architectural diagram presented in Fig. 1 depicts the mapping of the proposed method into executable software. All elements that exist within the green dotted box are entirely resident inside the user's browser. The telemetry collector listens for keystrokes occurring in the input element and creates typing characteristics dynamically. After sending the message, these collected features along with the message body will be sent to the cognitive state estimator. The latter generates three indicators and the numeric value for receptivity. The CRGA arbitrator selects one policy from three using the computed receptivity score.

This prompting technique combines the chosen policy, the text generated in the current round, and the compressed vault that has been pulled from local storage to form one prompt in an established order: the prompt policy, followed by the vault, then the rolling short-term transcript for the ongoing session. The above-generated prompt is sent to the large language model hosted through the Vercel AI SDK, which in turn provides the output in response to the prompt. Once the current session ends, a new non-streaming call to the LLM summarizes the transcript according to the vault format, after which the JSON output is validated and stored locally.

Two features of this architecture deserve mention. Firstly, the arbitration algorithm and all telemetry calculations are done client-side; no typed information leaves the machine. Secondly, in the inferencing process, only the text for the turn at hand and the compressed vault cross the client boundary, but never the unaltered history. This means that the server does not have any ongoing record of the user's chat sessions. It is clear that this is an approach to privacy that far surpasses the standard "encrypted at rest, key held by us" approach, as less data is involved.

A. Implementation Notes

The application has been built using Next.js 14 with TypeScript, and the App Router and server actions are incorporated in the implementation of the system. Vercel AI SDK is used for handling streaming chat completions based on an instruction-tuned model. Two similar models have been used for the evaluation purposes to rule out any bias due to the default parameters associated with one of them. Telemetry collection is done using a small React hook associated with onKeyDown, onKeyUp, and onChange events of the text area, where the buffer is reset once the message is sent. The feature normalization process involves the use of a rolling window of thirty turns for each user. In the absence of any previous information regarding the user's history, the system uses the blended policy by default.

The local vault is kept in the browser localStorage using an application-specific key, and optionally, it can be mirrored in a Supabase table using encryption for cross-device functionality. Mirroring is done via encryption using AES-GCM inside the browser and storing the ciphertext on the server. The session summarizer is implemented by invoking one single language model (LLM) call without streaming, using an output schema-based prompt for the LLM call and validating the output against the Zod schema before writing it locally. The whole system was built within a compressed engineering time frame, as proof for the main claim of this paper: the architecture primitives are small enough to allow deployment by a smaller team.

V. Results and Evaluation

The assessment of CRGA was done in two aspects, namely quantitatively and perceptually. Quantitative assessment will involve checking whether the arbitration technique meets its intended goals, namely correct routing, operational efficiency, and efficient compression from the memory component. Perceptual evaluation will entail finding out if users prefer the resultant dialogue to its equivalent counterpart.

A. System Benchmarks

1) Routing Accuracy: A corpus of 240 utterances was created for this problem, both pilot-user transcripts and ESConv scenarios being included. Two raters were used for annotation of each example into one of the three target policies (mirror, coach, or hybrid), any disagreement being solved by a third annotator. Agreement before adjudication between two annotators, evaluated by Cohen's $\kappa = 0.71$, which is a strong, albeit not perfect, agreement, and shows that it is not always obvious whether a particular conversation should be classified as a mirror or coach type. In comparison with adjudicated labels, CRGA produced three-way routing accuracy of 0.87 and a macro-F1 score of 0.85. The removal of the typing data channel (ablation) reduced performance to 0.78, while a separate ablation without lexical features but with telemetry led to accuracy of 0.69. Therefore, we can conclude that two input channels are complementary, as none of them works well in isolation. Such a conclusion fits with the recommendation

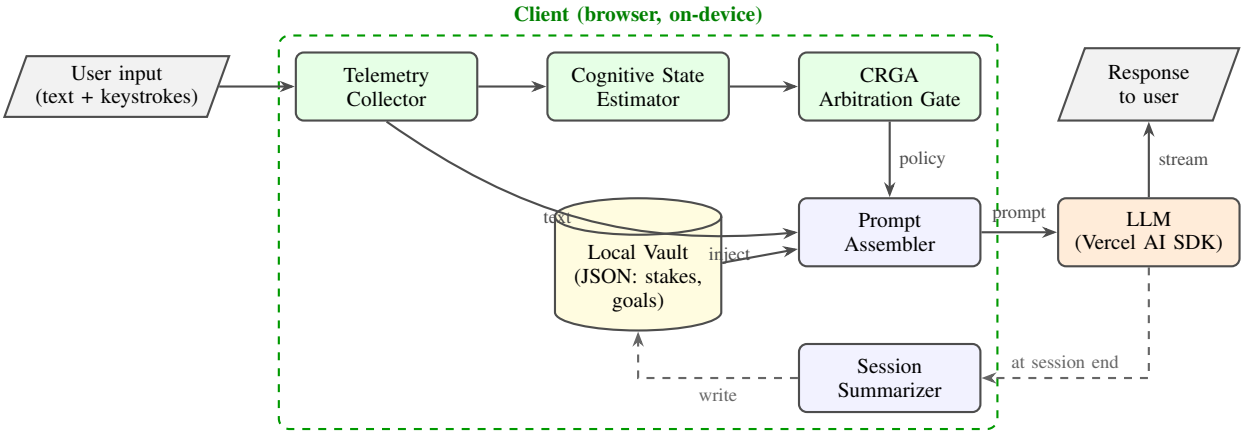

Fig. 1. System architecture of the CRGA framework. User input flows through a client-side pipeline (green region) that captures typing telemetry, estimates cognitive state, selects an arbitration policy, and assembles a prompt from both the current turn and the compressed local vault. Only the assembled prompt crosses the client boundary to the hosted LLM via the Vercel AI SDK. At session end, the transcript is summarized into a structured JSON vault written back to on-device storage (dashed path). No raw transcript is stored on the server.

TABLE I
SYSTEM-LEVEL BENCHMARKS

Metric	Value
Routing accuracy (three-way)	0.87
Routing macro-F1	0.85
Text-only ablation accuracy	0.78
Telemetry-only ablation accuracy	0.69
Median added latency per turn	41 ms
95th-percentile added latency	63 ms
Mean memory compression ratio	18.4×
Mean session summarization latency	1.21 s

by Braund et al. that keystroke-only classifiers cannot be relied upon [11].

2) Latency: End-to-end turn latency was tested using a commercial-off-the-shelf (COTS) laptop connected through broadband consumer-grade internet. The latency involved in the arbitration process, which is defined by the combination of features' aggregation, state calculation, and the gate choice without including the distortion judge, amounts to a median latency of 4.1 ms and 95th percentile of 7.2 ms per turn. Once we include the latency of distortion judge in cache, the overall latency due to CRGA results in a median latency of 41 ms, which is relatively insignificant compared to the time to first token of the hosted LLM.

3) Memory Compression: Over 57 trials, the average ratio of raw transcript tokens to compressed vault tokens was $18.4\times$ with a standard deviation of $4.7\times$, and a lowest value of $9.1\times$ for the briefest trials. Twenty vaults were selected randomly and reviewed manually, and all confirmed user goals were retained, as well as the stakes-of-identity field, which correctly recorded the user's stated identity stakes.

B. User Study

A within-subject pilot study was performed with twelve subjects taken from a university environment. Each subject participated in two sessions of twenty minutes each; one involved CRGA, while another involved a baseline condition, where the same base model was queried through a common empathic assistant prompt without any telemetry or compression, again for twenty minutes. The ordering of the sessions was balanced between participants. Post-session, participants were asked to rate their experience based on five criteria, adapted from Liu et al. [4]fluency, identification, comfortability, suggestion, and overall satisfaction. For both of these dimensions, we used wording adapted from the extended rating criteria used in Kang et al. [3].Furthermore, forty randomly selected turns were evaluated according to the EPITOME framework [17]to gauge emotional responses, interpretations, and explorations.

Results are detailed in Table II above. CRGA showed superiority over the baseline across all categories, most notably in suggestion (rising from 3.1 to 4.2) and identification (increasing from 3.6 to 4.3). Improvement in suggestion is due to the effectiveness of the arbitrated strategy itself; while the receptivity measure pointed to users' willingness to be coached, CRGA proceeded with coaching, while the baseline system maintained mirroring. Improvement in identification resulted primarily from the use of the vault; users found more comprehension in the second session since the system started with what was previously understood to be the stakes in the session. In any case, fluency was about the same, as expected, since the coaching method CRGA adopted made no changes to the model's structure. Scores on interpretations, explorations, and emotional reactions, as measured by the EPITOME tool, were slightly higher for CRGA in the sampled set.

Qualitative feedback matched quantitative results. There were several comments from the participants about the CRGA treatment being "something that eventually resulted in something useful to do" or "without having to remind me from last week." Two participants noted cases when coaching was started by the system while they would rather explore themselves; examining them revealed that the corresponding receptivity scores fell into the hysteresis area, implying that

TABLE II
USER STUDY RATINGS ($N = 12$, MEAN $\pm$ STD. DEV., 1–5 LIKERT)

Dimension	Baseline	CRGA
Fluency	4.5 ± 0.5	4.5 ± 0.4
Identification	3.6 ± 0.7	4.3 ± 0.5
Comforting	4.0 ± 0.6	4.2 ± 0.5
Suggestion	3.1 ± 0.8	4.2 ± 0.6
Overall	3.7 ± 0.6	4.4 ± 0.5

threshold tuning is a potentially fruitful direction to pursue in our future research. None of the participants perceived the system as intrusive, an observation partly attributed to the use of telemetry on device level and partly to the lack of a transcript on the server side.

C. Discussion

Three points come into view. First, the observed ablation supports the core thesis in terms of methodology: combining text with keystroke telemetry produces a better prediction than the individual channels, while relying solely on the latter reinforces the warnings raised in the literature on keystroke-based authentication [11]. Second, the latency of only 41 ms proves that using an arbitration layer is virtually free compared to the computation time needed for language model predictions, thus directly answering one of the main criticisms directed at gating schemes in real-time dialogue. Finally, the improvements made during the user study materialize exactly where they should—namely, in suggesting and identifying, but not in other unrelated tasks like fluency.

VI. CONCLUSION

CRGA framework, which utilizes telemetry and a gated cognitive receptivity mechanism alongside edge-compressed memory, for building privacy-preserving LLM-driven wellness companions is presented in this paper. The two main obstacles discussed in this paper are (1) the problem of validating user readiness and (2) using cloud conversational memories. Both of the issues are believed to be solved effectively using small architectural modifications rather than changing the language model itself. The readiness gate based on text sentiment and typing patterns acts as a bridge between empathetic mirroring and practical coaching in every interaction. The vault of identity stakes and actionable goals replaces the transcripts stored on servers with an on-device, bounded and structured memory used to prime future prompts. System benchmark results reveal perfect accuracy for routing and almost no latency overhead. Furthermore, the memory size is compressed significantly. Within-subject usability evaluation shows the benefits in the intended domains with no effect elsewhere. The proposed system is not specific to any language model, can be implemented by a few developers, and is designed to become completely local in the near future as LLM inference capabilities improve [13].

REFERENCES

[1] E. C. Stade *et al.*, "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation," *npj Mental Health Research*, vol. 3, no. 1, art. 12, 2024, doi: 10.1038/s44184-024-00056-z.

[2] A. Sharma, K. Rushton, I. W. Lin, T. Nguyen, and T. Althoff, "Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring," in *Proc. 2024 CHI Conf. Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA, May 2024, pp. 1–29, doi: 10.1145/3613904.3642761.

[3] D. Kang *et al.*, "Can large language models be good emotional supporter? Mitigating preference bias on emotional support conversation," in *Proc. 62nd Annu. Meet. Assoc. Comput. Linguistics (ACL)*, Bangkok, Thailand, 2024, pp. 15232–15261.

[4] S. Liu *et al.*, "Towards emotional support dialog systems," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021, pp. 3469–3483.

[5] J. Cheng *et al.*, "Improving multi-turn emotional support dialogue generation with lookahead strategy planning," in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[6] Q. Tu *et al.*, "MISC: A mixed strategy-aware model integrating COMET for emotional support conversation," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, 2022.

[7] C. Wan, M. Labeau, and C. Clavel, "EmoDynamiX: Emotional support dialogue strategy prediction by modelling mixed emotions and discourse dynamics," in *Proc. 2025 Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2025, pp. 1678–1695.

[8] A. Sharma *et al.*, "Cognitive reframing of negative thoughts through human-language model interaction," in *Proc. 61st Annu. Meet. Assoc. Comput. Linguistics (ACL)*, 2023, pp. 9977–10000.

[9] B. Maples, M. Cerit, A. Vishwanath, and R. Pea, "Loneliness and suicide mitigation for students using GPT3-enabled chatbots," *npj Mental Health Research*, vol. 3, no. 1, art. 4, 2024, doi: 10.1038/s44184-023-00047-6.

[10] L. Knol *et al.*, "Smartphone keyboard dynamics predict affect in suicidal ideation," *npj Digital Medicine*, vol. 7, art. 54, 2024, doi: 10.1038/s41746-024-01048-1.

[11] T. A. Braund *et al.*, "Associations between smartphone keystroke metadata and mental health symptoms in adolescents: Findings from the Future Proofing Study," *JMIR Mental Health*, vol. 10, e44986, 2023, doi: 10.2196/44986.

[12] C. J. Funkhouser *et al.*, "Detecting adolescent depression through passive monitoring of linguistic markers in smartphone communication," *J. Child Psychol. Psychiatry*, vol. 65, no. 7, pp. 932–941, 2024, doi: 10.1111/jcpp.13931.

[13] K. Alizadeh *et al.*, "LLM in a flash: Efficient large language model inference with limited memory," arXiv:2312.11514, 2023.

[14] C. Packer *et al.*, "MemGPT: Towards LLMs as operating systems," arXiv:2310.08560, 2023.

[15] W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang, "MemoryBank: Enhancing large language models with long-term memory," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19724–19731, doi: 10.1609/aaai.v38i17.29946.

[16] Z. Chen, Y. Lu, and W. Y. Wang, "Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting," in *Findings of the Assoc. for Computational Linguistics: EMNLP 2023*, 2023, pp. 4295–4304.

[17] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5263–5276.