



School of Computer Science and Electronic Engineering

MSc Data Science

Academic Year 2024-2025

**Design and Evaluation of Multi-Agent Conversational
Life Coaches Grounded in Behavioural Psychology**

A project report submitted by: Satish Ranganathan Mohan, 6202583

A project supervised by: Dr Kevin Wells

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

University of Surrey
School of Computer Science and Electronic Engineering
Guildford, Surrey GU2 7XH
United Kingdom.
Tel: +44 (0)1483 300800

Abstract

This dissertation explores the design and evaluation of multi-agent conversational “life coach” systems built in LangFlow. Sixteen prototypes (v1.1, v1.2, v2.1, ... v8.2) were developed, progressing from simple prompt-based agents to retrieval-augmented and aggregator-based architectures combining Cognitive Behavioural Therapy (CBT), performance psychology, and the Thirukkural.

The evaluation used adapted CARE and Psychological Realism rubrics, readability measures, NLP metrics, and two simulated conversations. Results show that effective responses were typically of moderate length and accessible reading level. Memory enabled consistent recall of names and cues, though narrative coherence was stronger in GPT-5 than Qwen models, which often relied on repetitive structures. Aggregator prototypes enriched dialogue by blending perspectives but risked verbosity or failure (3.3% of cases). Thirukkural references were accurate yet not always contextually precise.

The study concludes that digital life coaches can approximate aspects of a supportive alliance. While limited by repetitiveness and contextual sensitivity, they hold cautious promise for democratising access to mental health support.

Acknowledgements

I would like to thank my supervisor for their guidance, feedback, and encouragement throughout this project. I am also grateful to the many mentors and life coaches, both past and present, whose wisdom has shaped my journey. Some I have known personally, others only through their writings and example, but all have influenced the values and perspectives that guided this work.

Sign in the box below to certify that the work carried out is your own. By signing this box you are certifying that your dissertation is free from plagiarism. Make sure that you are fully aware of the Department guidelines on plagiarism (see the student handbook). The penalties if you are caught are severe. All material from other sources must be properly referenced and direct quotes must appear in quotation marks.

I certify that the work presented in the dissertation is my own unless referenced

Signature...Satish Ranganathan Mohan...

Date.....09th September, 2025.....

Insert a word count. This is the sum of the words in all the chapters only. The sum should exclude the words in the title page, abstract, acknowledgements, table of contents, references and any appendices.

TOTAL NUMBER OF WORDS:

Table of Contents

Abstract.....	2
Acknowledgements.....	4
Table of Contents.....	6
List of Tables	8
List of Figures.....	9
Chapter 1: Introduction.....	10
1.1 Modern Society and the Role of a Life Coach.....	10
1.2 Behavioural Psychology and the Power of Speech.....	11
1.3 Motivation and Problem Statement.....	11
1.4 Research Aims and Objectives	12
1.5 Research Questions.....	13
1.6 Contributions.....	13
1.7 Structure of the Dissertation	14
Chapter 2: Literature & Technology Review.....	15
2.1 Evolution of Chatbots: From ELIZA to Modern Systems.....	15
2.2 Commercial Landscape of Mental Health Chatbots	16
2.3 Efficacy of Chatbots in Mental Health and Wellbeing	17
2.4 AI Model Optimisation for Psychological Applications.....	18
2.5 Multi-Agent Systems in Conversational AI.....	21
2.6 LangChain and LangFlow.....	22
2.7 The Digital Therapeutic Alliance.....	23
2.8 Synthesis and Research Gap	24
Chapter 3: Research Methodology.....	25
3.1 Research Design.....	25
3.2 Prototype Design.....	25
3.3 Wordsmithing and Prompt Engineering	27
3.4 Knowledge Base Construction.....	28
3.5 Prototype Ensemble and Comparison Strategy.....	30
3.6 Evaluation Frameworks	31
3.7 Summary	35
Chapter 4: Implementation & Results.....	37
4.1 Prototype Evolution Overview.....	37

4.2 Framework Specialisation Analysis.....	39
4.3 Response Parameter Optimisation	40
4.4 Memory Implementation Assessment.....	45
4.5 Aggregator Effectiveness.....	46
4.6 Integrated Findings and Discussion	48
Chapter 5: Conclusion.....	50
5.1 Overview	50
5.2 Summary of Results and Reflection.....	50
5.5 Ethical Reflections	51
5.6 Limitations of this research.....	52
5.7 Future Directions	53
5.8 Final Reflections	54
References.....	55
Appendix A.....	60
Appendix B	61
Appendix C	63
Appendix D.....	70

List of Tables

Table 2.1: Comparison of commercial chatbot offerings

Table 3.1: Scoring criteria for CARE and Psychological Realism rubric

Table 4.1 Evolution of prototypes with illustrative excerpts

Table 4.2. Prototype response parameters and rubric scores

Table 4.3: Top five prototypes ranked by average CARE + Realism score

List of Figures

Figure 2.1 A simple schematic of Fine Tuning

Figure 2.2 A simple schematic of Prompt Engineering

Figure 3.1: Schematic of a LangFlow pipeline implementing RAG and chat history.

Figure 4.1: Prototype Evolution Timeline. Features added to prototypes.

Figure 4.2: Average Word Count vs CARE - Communication.

Figure 4.3: Average Word Count vs Realism – Forward Movement

Figure 4.4: Readability vs CARE - Empathy.

Figure 4.5: Flesch-Kincaid Grade Level vs CARE - Empathy.

Chapter 1: Introduction

“The Voice is more than words. It is a way of controlling others by controlling yourself.” —
Frank Herbert, *Dune*

1.1 Modern Society and the Role of a Life Coach

Humanity faces an escalating mental health crisis. The World Health Organization reported a 25% global increase in depression and anxiety disorders in the aftermath of the COVID-19 pandemic, with young adults and working professionals disproportionately affected [1]. More than 300 million people worldwide live with depression, while anxiety disorders affect hundreds of millions more [2]. Even in advanced economies, access to psychological support is constrained by stigma, long waiting lists, or prohibitive costs [3]. The pandemic further intensified these pressures, contributing to sharp rises in stress, burnout, and social isolation [4]. Contemporary society presents unprecedented challenges to psychological well-being, from the instability of modern work to the fracturing of traditional community structures.

In parallel, the life coaching industry has grown rapidly as an alternative source of structured support. Life coaching is distinct from psychotherapy: whereas therapy often addresses clinical symptoms, coaching emphasises clarity of values, goal-setting, and accountability [5,6]. This distinction makes coaching attractive to individuals who are not clinically ill but seek guidance in navigating personal and professional challenges. Yet access to quality coaching remains limited by cost, availability, and geography. Professional coaches often charge between £100–200 per hour [7], placing their services beyond the reach of most people.

The concept of coaching is not new. Human societies have long developed systems to help individuals improve themselves, whether through institutions for education and skill development, philosophical traditions to navigate life, or the guidance of mentors and gurus personally invested in the growth of others [6]. Life coaching, formally defined as “the process of helping individuals identify and achieve personal goals by devising suitable strategies and viable plans, developing conducive skills and attitudes, and providing systematic intervention to motivate and effect a successful completion of the plan” [5], represents a contemporary adaptation of this enduring human impulse. Coaches and mentors can mitigate the decline in psychological well-being by offering structured reflection and accountability, particularly for urban youth populations facing increasing isolation and uncertainty.

Meanwhile, conversational artificial intelligence has matured with remarkable speed. Large language models (LLMs) now demonstrate unprecedented fluency and flexibility [8], raising the possibility of digital coaches that can interactively support users. Commercial products such as Woebot, Wysa, and Replika reveal both the appetite for such tools and their current limitations [9,10]. While valuable, these systems are often constrained by scripted dialogue, narrow frameworks, or general-purpose design, limiting their ability to adapt dynamically to user needs, sustain coherence across interactions, or convincingly embody therapeutic principles [11].

This tension motivates the central question of this dissertation: how can agentic chatbots be designed to function not just as fluent conversationalists, but as psychologically realistic life coaches? An AI life coach is not intended to replace therapists, but to occupy a middle ground: supporting goal setting, motivation, and self-improvement while remaining accessible and scalable. Properly designed, such systems could democratise support for mental health [12].

1.2 Behavioural Psychology and the Power of Speech

Human beings are deeply shaped by language. The way speech is structured, the metaphors that are used, and the frameworks that guide conversations all have powerful effects on thought, emotion, and behaviour [13]. Behavioural psychology has shown for decades that structured interventions such as goal setting, reinforcement, and reframing can change how people act and feel [14]. Counselling approaches build on this foundation, offering conversational techniques that improve self-awareness, build a sense of agency, and provide systematic guidance.

This connection between words and influence is not only studied in psychology but also imagined in literature. Herbert’s idea of *The Voice* illustrates how language can be shaped with precision to affect others. In coaching and therapy, the same principle applies: carefully chosen words can guide reflection, shift perspectives, and encourage change. Within this project, prompt design and system messages are understood as tools for shaping the “voice” of the agent. The more carefully this voice is crafted, the greater its potential to positively influence the user’s state of mind.

Several practical frameworks demonstrate how structured speech can support change. In professional coaching, the **GROW model** (Goal, Reality, Options, Way Forward) gives a clear structure to conversations [16]. In clinical psychology, **Cognitive Behavioural Therapy (CBT)** helps people notice and reframe distorted patterns of thinking [17]. **Performance psychology** draws on research into motivation and resilience and is especially useful in non-clinical contexts such as sport, study, and personal growth [18].

Cultural traditions also place great value on the ethics of speech. The *Thirukkural*, a Tamil classic of 1,330 couplets, has shaped moral and social thinking in South India for centuries [19]. Its concise verses emphasise restraint, integrity, and right action. Grounding AI agents in such culturally significant texts does more than provide knowledge. It encourages responses that feel emotionally resonant and meaningful, especially to niche communities that see their own traditions reflected in the dialogue. In this way, the agent can take on a cultural character, making it feel more authentic and customised for the communities it serves.

Three sources were selected for this research: **CBT**, **performance psychology**, and the **Thirukkural**. These provide a mix of clinical rigour, practical coaching strategies, and cultural depth. Together they allow an evaluation not only of whether an AI agent can be fluent, but whether it can communicate in ways that feel psychologically realistic and culturally resonant.

1.3 Motivation and Problem Statement

The growing demand for psychological support, combined with the limitations of existing provision, creates a strong motivation to explore alternative approaches. Life coaching offers a promising model for accessible, non-clinical support, yet its high cost and limited availability restrict its reach [7]. At the same time, conversational AI has advanced to a point where it can sustain meaningful dialogue, raising the possibility of scalable digital coaching tools [8]. The convergence of these two domains provides an opportunity to design agents that are both technically capable and psychologically grounded.

Despite these advances, significant challenges remain. Current conversational agents often lack alignment with established psychological frameworks, leading to advice that is either superficial or inconsistent [11]. Many systems prioritise fluency and engagement over depth and accuracy, which limits their potential to provide genuine support. Furthermore, most models are not culturally sensitive, failing to resonate with users whose values and traditions shape the way they understand wellbeing [19]. Without

cultural and psychological grounding, there is a risk that AI life coaches will reproduce generic, context-free advice that fails to meet user needs.

This dissertation addresses these gaps by developing and evaluating a series of multi-agent prototypes designed to function as AI life coaches. The prototypes draw on three distinct foundations: Cognitive Behavioural Therapy (CBT), performance psychology, and the Thirukkural. Together, these frameworks combine clinical rigour, practical strategies for self-improvement, and cultural depth. The research problem can therefore be stated as follows:

How can multi-agent conversational AI systems be designed to act as psychologically realistic and culturally resonant life coaches?

1.4 Research Aims and Objectives

The overall aim of this research is to investigate how multi-agent conversational AI systems can be designed to act as psychologically realistic and culturally resonant life coaches. This involves not only constructing and testing prototypes but also establishing systematic ways of evaluating their performance against both psychological and linguistic criteria.

This study adopts the **Design Science Research (DSR) methodology** [20]. The system is iteratively developed, refined, and evaluated, with LangFlow used as a prototyping environment. Each version of the prototype, from v1.1 to v8.2, represents a refinement informed by the results of earlier evaluations. Empirical testing involves prompt engineering, system tuning, and simulated dialogues, with qualitative analysis to assess conversational realism. Quantitative evaluation is conducted using metrics adapted from prior chatbot studies [21].

The research also draws on the **CRISP-DM framework** [22], originally developed for data mining but here adapted to structure the development and evaluation pipeline. The phases map to this project as follows:

- **Business Understanding:** Defining the research gap and requirements.
- **Data Understanding:** Identifying and selecting sources for the knowledge base.
- **Data Preparation:** Pre-processing texts through chunking, embedding, and indexing.
- **Modelling:** Designing the multi-agent orchestration and prompt structures.
- **Evaluation:** Assessing system performance through simulated dialogues and rubric scoring.
- **Deployment:** Creating a reproducible implementation pipeline that can be applied in future work.

Guided by these methodologies, the objectives of the research are as follows:

1. To design and implement a series of conversational prototypes using a modular pipeline approach, exploring variations in architecture, memory, retrieval-augmented generation (RAG), and aggregation mechanisms.
2. To ground the prototypes in established frameworks, namely Cognitive Behavioural Therapy (CBT), performance psychology, and the *Thirukkural*, in order to test their ability to produce psychologically realistic and culturally resonant dialogue.

3. To develop and apply an evaluation framework that measures performance along both psychological dimensions (CARE and Psychological Realism) and linguistic dimensions (readability, lexical diversity, repetitiveness, question-asking, and consistency).
4. To analyse the results across prototypes and architectures in order to identify trade-offs, error patterns, and emergent behaviours.
5. To synthesise design insights that inform how future conversational AI systems can balance fluency, psychological grounding, and cultural sensitivity in order to democratise access to support for mental health.

1.5 Research Questions

Building on the aims and objectives, this study is guided by the following research questions:

1. **Design:** How can multi-agent conversational AI systems be structured to act as life coaches, and what architectural features (memory, retrieval-augmented generation, aggregation mechanisms) most affect their behaviour?
2. **Frameworks:** To what extent can responses be aligned with established frameworks such as Cognitive Behavioural Therapy, performance psychology, and the *Thirukkural*?
3. **Evaluation:** How can the effectiveness of such systems be measured through adapted rubrics, including CARE (Communication, Autonomy, Respect, Empathy, adapted from the Consultation and Relational Empathy measure) and Psychological Realism (Groundedness, Cognitive Insight, Emotional Resonance, Forward Movement)?
4. **Cultural Resonance:** Can cultural knowledge bases, such as the *Thirukkural*, enhance psychological realism by producing dialogue that feels emotionally resonant and authentic to niche communities?
5. **Design Insights:** What trade-offs, error patterns, and emergent behaviours arise across prototypes, and what do these reveal about the design of future AI life coaches that aim to democratise support for mental health?

1.6 Contributions

This dissertation contributes both technically and conceptually to the field of AI-powered psychological support.

Technically, it demonstrates the integration of retrieval-augmented generation (RAG) with multi-agent orchestration for psychological dialogue. The implementation using LangFlow, AstraDB, and OpenRouter provides a reproducible pipeline that integrates file ingestion, text splitting, embeddings, AstraDB vector storage, and carefully crafted system prompts with decision rules and guardrails. The study also documents the effect of design choices, including architectural components and hyperparameters such as chunk size, overlap, and temperature, on retrieval quality and tonal stability.

Conceptually, the research advances the notion of an *agentic life coach*, situated at the intersection of AI, psychology, and social support. Unlike existing commercial chatbots that typically rely on a single therapeutic approach, the prototypes developed here dynamically balance between Cognitive Behavioural Therapy and performance psychology frameworks, supplemented by cultural grounding through the

Thirukkural. The aggregator mechanism allows the system to adapt its responses in line with the user's evolving needs, which is a novel contribution to the field.

Finally, the study positions itself within ongoing debates about the **digital therapeutic alliance** [23]. It argues that conversational agents, when designed with empathy, guardrails, and reflective frameworks, can complement rather than replace human support. By addressing both ethical considerations and practical limitations, the research contributes to responsible innovation in AI for mental well-being.

1.7 Structure of the Dissertation

The dissertation is organised into six chapters, supported by appendices:

Chapter 1: Introduction establishes the context, motivation, aims, research questions, and contributions. It introduces the guiding metaphor of *The Voice* (From the novel: Dune) and frames the work within debates on the digital therapeutic alliance.

Chapter 2: Literature & Technology Review surveys prior work on conversational AI for mental health, the commercial landscape, evidence on efficacy, multi-agent approaches, and the LangFlow toolchain. It concludes with a synthesis and research gap that motivates the present study.

Chapter 3: Research Methodology details the design–science approach and CRISP–DM orientation, the LangFlow pipeline components, prompt engineering, knowledge base construction, the prototype ensemble and comparison strategy, and the evaluation frameworks (CARE and Psychological Realism), alongside readability and NLP metrics.

Chapter 4: Implementation and Results presents the prototype evolution and version notation, analyses framework specialisation, examines response parameters, assesses short-term memory with two simulated conversations, and evaluates aggregator effectiveness in blend and bias modes. It closes with an integrated findings and discussion section, covering cultural integration, conversation flow, quantitative scoring, and selected NLP metrics.

Chapter 5: Conclusion interprets the findings in relation to the research questions, reflects on the metaphor of *The Voice*, and considers the implications for building a digital therapeutic alliance in practice. It also situates the results within broader debates about ethics and responsible innovation in AI for mental well-being.

The appendices provide the corpus used for RAG implementation, sample system message, and transcripts from the simulated conversations.

Chapter 2: Literature & Technology Review

This chapter provides the intellectual and technological context for the development of an AI-driven life coach. It examines the historical roots of conversational agents, the rise of commercial chatbot offerings, the growing body of academic research on digital interventions for wellbeing, and the debates surrounding the digital therapeutic alliance. It also reviews recent advances in Retrieval-Augmented Generation (RAG) and multi-agent architectures, which together form the backbone of the present system. Finally, it reflects on the societal and ethical implications of deploying such tools in mental health and coaching contexts.

2.1 Evolution of Chatbots: From ELIZA to Modern Systems

The history of conversational AI spans over half a century, with significant milestones marking its evolution from simple pattern-matching systems to sophisticated language models capable of nuanced dialogue.

Early Systems (1960s-1980s): The first notable conversational program, ELIZA (Weizenbaum, 1966), simulated a Rogerian psychotherapist by using pattern matching and substitution methodology to respond to user inputs. Despite its simplicity, users often attributed human-like understanding to ELIZA, demonstrating the “Eliza effect”, the tendency to anthropomorphise computer responses. PARRY (Colby, 1975), an early attempt to model a paranoid patient, represented another milestone in simulating psychological states through dialogue.

Rule-Based Era (1980s-2000s): This period saw the development of more sophisticated rule-based systems like A.L.I.C.E. (Wallace, 2009), which used AIML (Artificial Intelligence Markup Language) to create extensive dialogue trees. These systems required manual scripting of responses and were limited in their ability to handle novel inputs.

Statistical Approaches (2000s-2010s): The introduction of machine learning techniques enabled the development of more flexible dialogue systems. Microsoft’s Xiaoice (Zhou et al., 2018), launched in 2014, demonstrated emotional engagement capabilities that resonated with millions of users in China.

Transformer Revolution (2017-Present): The publication of the “Attention is All You Need” paper (Vaswani et al., 2017) introduced the Transformer architecture, which underpins modern large language models. Subsequent models like BERT (Devlin et al., 2019), GPT series (Radford et al., 2018; Brown et al., 2020), and specialised variants have dramatically improved conversational capabilities.

A major transformation occurred in the 2010s with the rise of neural network models. Sequence-to-sequence architectures and later transformer-based models enabled systems to learn patterns of dialogue from vast corpora of text. Early applications such as Microsoft’s Xiaoice in China attracted millions of users by offering emotionally engaging conversations, demonstrating the potential for chatbots to act as companions. Nevertheless, these systems still produced frequent errors and lacked grounding in psychological principles.

The advent of Large Language Models (LLMs) has been the most significant step in the evolution of conversational AI. Models such as GPT, LLaMA, Claude, and Qwen can generate coherent, context-sensitive, and human-like responses across diverse domains. Unlike their rule-based predecessors, LLMs are capable of generalisation and adaptation, making them suitable for complex dialogues. This has opened new possibilities for applications in education, professional support, and mental health.

Yet the progression from ELIZA to modern LLMs highlights a consistent tension between fluency and depth. ELIZA showed that even simple scripts can evoke the *feeling* of understanding, while LLMs demonstrate the opposite risk: fluent but shallow advice that lacks psychological grounding. Bridging this gap requires more than technical advances in language generation. It demands careful integration of psychological frameworks, cultural wisdom, and mechanisms that sustain a digital therapeutic alliance. This study builds on that recognition, seeking to move beyond fluency alone towards dialogue that is both realistic and supportive.

2.2 Commercial Landscape of Mental Health Chatbots

Alongside academic research, commercial applications of conversational AI have expanded rapidly in the mental health domain. Several platforms have attracted significant user bases by offering accessible, low-cost alternatives to professional support. Among the most widely cited are Woebot, Wysa, and Replika, each of which illustrates different design choices and trade-offs.

Woebot: CBT-Based Intervention. Woebot is a chatbot developed at Stanford University and grounded explicitly in Cognitive Behavioural Therapy (CBT). Its dialogues are structured around identifying cognitive distortions, challenging unhelpful thought patterns, and encouraging behavioural activation. Woebot has been evaluated in randomised controlled trials, with evidence suggesting it can reduce symptoms of depression and anxiety in the short term [33]. However, its interactions are largely scripted, limiting flexibility. Users often report a sense of repetitiveness after extended use, reflecting the constraints of a tightly rule-based approach.

Wysa and Koko: Empathetic Support Systems. Wysa also builds on CBT principles but adopts a slightly different strategy. While it relies on scripted dialogue flows, it offers a broader repertoire of supportive exercises, such as mindfulness and journaling prompts. Personalisation is somewhat greater than in Woebot, with pathways that adapt to reported mood or stress levels. The evidence base for Wysa is growing, though less robust than Woebot’s, and evaluations typically show moderate rather than strong effects [34]. Like Woebot, Wysa’s dependence on pre-defined scripts restricts conversational depth.

Replika: The Companionship Model and Ethical Concerns. Replika takes a distinct approach, positioning itself less as a therapeutic tool and more as a digital companion. Powered by a large language model, it allows for more free-form and emotionally expressive conversations. Many users report a sense of intimacy and personal connection with their Replika, and the system’s ability to remember details about the user can strengthen the impression of continuity. However, its therapeutic value is contested. Unlike Woebot and Wysa, Replika is not explicitly grounded in evidence-based psychological frameworks. Studies suggest that while it may reduce feelings of loneliness for some, it can also encourage unhealthy forms of attachment or provide inconsistent responses [35]. Regulatory concerns have also emerged, with Replika being restricted in some jurisdictions due to inappropriate content.

Table 2.1: Comparison of commercial chatbot offerings

Feature	Woebot	Wysa	Replika	This Work
Therapeutic Approach	CBT	CBT	Companionship	CBT + Performance Psychology
Evidence Based	Strong	Moderate	Low	Low
Personalisation	Low	Medium	High	High (multi-agent)
Escalation Pathways	Yes	Yes	No	n/a
Multi-Framework Integration	No	Limited	No	Yes
Knowledge Base	Scripted	Scripted	LLM	RAG-enhanced LLM

This table illustrates the gap in current commercial offerings. Woebot and Wysa prioritise evidence-based design but remain restricted by scripted dialogue and limited adaptability. Replika offers greater fluency and personalisation but lacks grounding in therapeutic principles. None of these systems integrate multiple frameworks in a structured manner, nor do they provide transparent pipelines for extending knowledge.

The present work addresses this gap by exploring a multi-agent architecture that combines CBT with performance psychology and incorporates Retrieval-Augmented Generation (RAG) for contextual grounding. While the prototypes developed here are not evaluated at commercial scale, they represent an effort to balance the strengths of scripted evidence-based systems with the flexibility of LLMs, advancing towards a more robust digital therapeutic alliance.

2.3 Efficacy of Chatbots in Mental Health and Wellbeing

The question of whether conversational agents can meaningfully support mental health has been the focus of a growing body of empirical research. Systematic reviews and meta-analyses suggest that chatbots can produce measurable benefits, particularly in reducing symptoms of depression, anxiety, and stress in the short term [36]. These findings align with the promise of chatbots as scalable, low-cost interventions that can complement overstretched healthcare systems.

Controlled trials of systems such as Woebot and Wysa show small-to-moderate effect sizes for improvements in mood and reductions in anxiety [33,34]. For example, Fitzpatrick et al. (2017) demonstrated significant reductions in depressive symptoms after two weeks of daily interactions with Woebot. Similarly, studies of Wysa have reported reductions in anxiety and improvements in resilience among young adults and working professionals [37]. These results, while promising, tend to reflect short-term engagement and often rely on self-reported outcomes. Evidence for sustained long-term benefits remains limited, in part due to high attrition rates and the novelty effect of early use.

Research suggests several mechanisms through which chatbots exert their effects: providing consistent, non-judgmental support (Lattie et al., 2019), delivering accessible and structured therapeutic techniques (Torous et al., 2020), reducing barriers to seeking help (particularly for marginalised populations) (Hollis et al., 2018), and enabling regular self-monitoring and reflection (Birney et al., 2016).

Despite promising results, significant limitations exist:

- Most studies have short follow-up periods, limiting understanding of long-term effects (Bakker et al., 2016)
- Chatbots struggle with complex emotional states and crisis situations (Luxton et al., 2016)
- The “digital therapeutic alliance” remains weaker than human therapeutic relationships (Imel et al., 2019)
- Concerns about over-reliance and potential harm from inappropriate responses (Provoost et al., 2021)

Several theoretical models help explain when and why chatbots work:

- The **Supportive Accountability** model (Mohr et al., 2017) suggests that chatbots function best when they provide a sense of accountability through regular check-ins.
- The **Human-AI Collaboration** framework (Bickmore et al., 2018) emphasises designing systems that complement rather than replace human support.
- The **Digital Therapeutic Alliance** concept (Santomauro et al., 2025) examines how trust and rapport develop in human-AI interactions.

One of the consistent findings across studies is that users value qualities such as empathy, respect, and consistency in chatbot interactions [38]. These qualities are central to the therapeutic alliance in human counselling and coaching, and their absence can limit sustained engagement with digital agents. Commercial chatbots, particularly those based on scripted flows, often struggle to maintain coherence over time, leading to a perception of superficiality. Free-form systems like Replika offer greater conversational richness but can lapse into inconsistency or inappropriateness.

This tension between fluency and depth is critical for evaluation. Traditional clinical trials measure outcomes such as symptom reduction, but these do not capture the finer-grained qualities of interaction that determine whether users feel supported. To address this gap, researchers have begun to propose rubrics that assess communication quality, empathy, and realism in dialogue [39]. In this study, two frameworks are adopted for evaluation:

CARE (Communication, Autonomy, Respect, Empathy), which captures the interpersonal dimensions of dialogue quality.

Psychological Realism, which measures groundedness, cognitive insight, emotional resonance, and forward movement within responses.

Together, these rubrics operationalise the principles underlying the digital therapeutic alliance in a way that is measurable for conversational AI. They provide a bridge between clinical outcome measures and the subtler qualities of dialogue that make an agent feel genuinely supportive.

In this sense, the mixed results of existing chatbot trials highlight not only the potential of digital interventions but also the need for more nuanced evaluation methods. The prototypes developed in this research are therefore assessed not just on fluency or user satisfaction but on their ability to embody qualities that sustain a digital therapeutic alliance over time.

2.4 AI Model Optimisation for Psychological Applications

Conversational AI systems require a careful consideration of model optimisation strategies. While large language models (LLMs) provide impressive foundational capabilities, their effectiveness in specialised domains like psychological support is limited. Optimisation is therefore necessary to guide models towards safe, consistent, and psychologically meaningful responses.

One approach has been **fine-tuning**, where models are trained further on domain-specific data such as therapy transcripts or coaching dialogues. Fine-tuned systems often demonstrate improved alignment with psychological frameworks, producing more structured and clinically relevant responses [40]. However, fine-tuning requires access to large datasets of sensitive material, which raises issues of privacy, copyright, and cultural bias. Moreover, fine-tuned models can become rigid, losing the flexibility that makes large language models attractive in the first place. In this research, while no custom fine-tuning was undertaken, an open-source fine-tuned model was adopted as part of the pipeline: ALIENTELLIGENCE/lifecoach, released on Ollama. This model is based on LLaMA 3.1 under Meta's community license and is configured with a 128K token context window, making it well suited for longer, more coherent coaching dialogues. Although the details of its training process are not fully disclosed, Ollama's documentation indicates that fine-tuning typically relies on Modelfiles, where base models are customised with targeted system prompts and example dialogues.

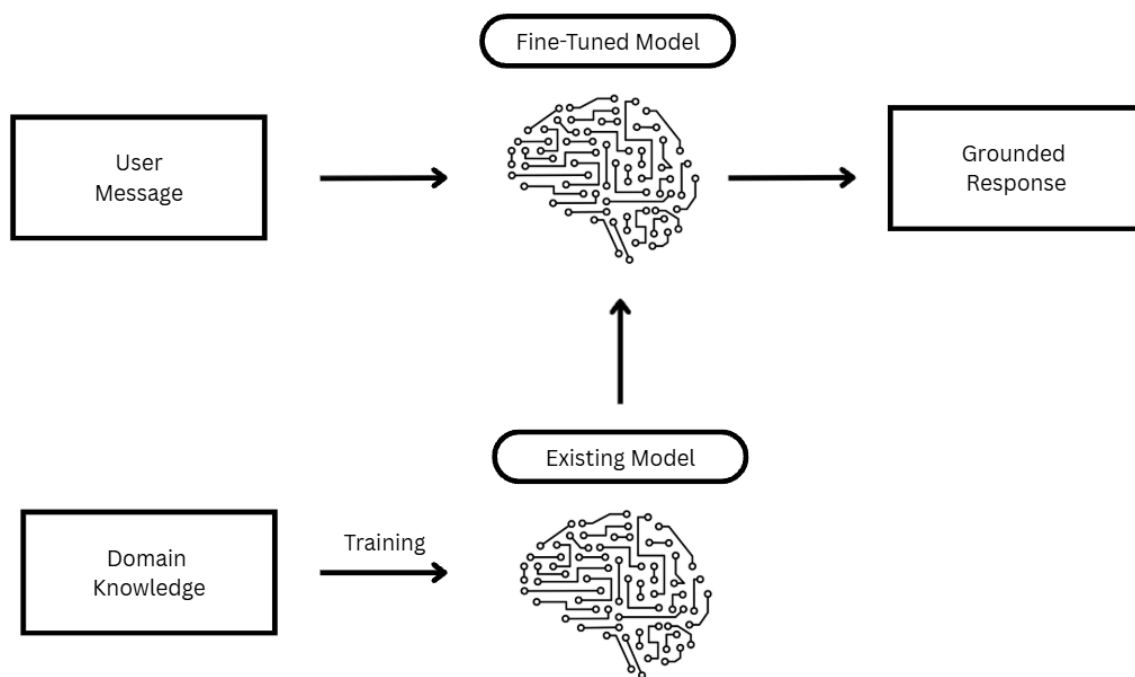


Figure 2.1 A simple schematic of Fine Tuning

A second approach is **prompt engineering**, where system behaviour is shaped by carefully designed instructions. Prompt engineering can be implemented by defining system messages and templates for the user messages, and by incorporating contextual data to ground the model behaviour to a particular domain. Prompts can enable the same underlying model to perform different roles. Prompt engineering is particularly useful in prototyping environments, where iterative testing and rapid adjustment are necessary [41]. The limitations are that prompts can be brittle, and small changes in wording may produce large shifts in output.

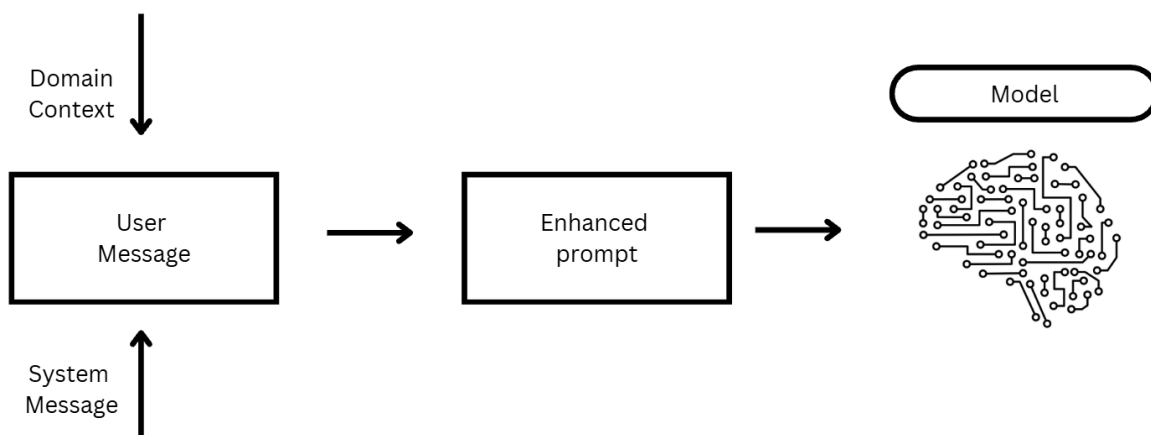


Figure 2.2 A simple schematic of Prompt Engineering

A third and increasingly influential technique is **retrieval-augmented generation (RAG)**. RAG allows the model to ground its responses in external knowledge bases, ensuring that outputs are accurate and contextually relevant. This means agents can be equipped with curated conceptual frameworks from psychology, cultural texts, or reflective prompts that supplement their generative ability. RAG therefore offers a balance between fluency and grounding, reducing the risk of hallucination while enabling cultural and conceptual integration [42]. Since it is easier and faster to update the knowledge base used to create RAG, the model can be made up-to-date much easier.

In this research, RAG was implemented by constructing a knowledge base from three distinct sources: CBT training materials, performance psychology texts, and the *Thirukkural*. Texts were split into overlapping chunks, embedded, and stored in AstraDB. At runtime, relevant passages were retrieved and inserted into the prompt templates. In this way, the template determined *how* the agent spoke, while RAG determined *what* the agent spoke about. This combination ensured that responses were not only structured but also grounded in evidence-based and culturally resonant material. While chat history was managed separately through a dedicated memory component in LangFlow, future work could explore extending RAG to include user-specific information for even more personalised support.

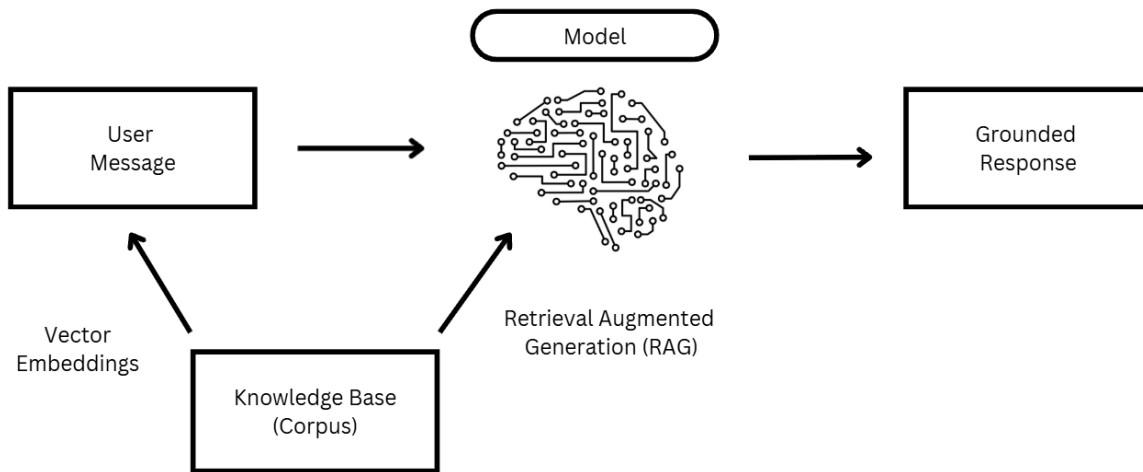


Figure 2.3 A simple schematic of RAG Implementation

Several design choices were also critical to shaping behaviour across all prototypes:

- **Chunk size and overlap:** Texts were split into overlapping segments to balance retrieval precision with contextual richness. Smaller chunks improved retrieval accuracy, while overlap prevented loss of coherence across sections.
- **Stringify vs Parser:** The pipeline employed a stringify approach to preserve text integrity, ensuring that retrieved passages retained their original meaning.
- **Temperature:** A generation temperature of 0.7 was chosen to encourage creative but controlled responses, balancing novelty with reliability.
- **Max tokens:** A cap of 500 tokens per response was used to ensure conciseness while allowing sufficient depth.

While presented as distinct methods, these optimisation approaches are “commonly used in combination” to create more effective psychological chatbots. The most successful implementations integrate all three approaches to leverage their complementary strengths.

Together, these optimisation strategies demonstrate that system behaviour can be shaped not only by the underlying model but also by the architecture in which it is embedded. By combining the strengths of a pre-fine-tuned model, structured prompt engineering, and retrieval-augmented generation, this research created prototypes capable of moving beyond surface fluency towards dialogue that sustains a digital therapeutic alliance.

2.5 Multi-Agent Systems in Conversational AI

Recent advances in multi-agent systems have opened new possibilities for conversational AI. Unlike single-model architectures, multi-agent systems distribute tasks across specialised subsystems. This approach trains multiple specialised “expert” networks and uses a gating mechanism to determine which experts should handle specific inputs. In conversational contexts, different experts might specialise in different topics or response styles.

The technical implementation of MoE systems involves:

- Training multiple specialised models on different aspects of the domain
- Developing a sorting mechanism to determine which model should handle which specific inputs
- Implementing mechanisms for combining outputs from multiple models

In our context, this could involve separate models for different therapeutic approaches (CBT, coaching, motivational interviewing, etc.), with a sorting mechanism determining which approach is most appropriate for a given user input.

The integration of multiple psychological frameworks through agent orchestration represents the novel contribution of this research. While commercial systems typically commit to a single therapeutic approach, real-world coaching often requires fluid movement between frameworks based on context. A human life coach would adapt their style and approach based on their client and their circumstances.

Recent advances such as AutoGPT and LangChain agent orchestration illustrate how large language models can be configured as agents that can handle planning, execution, and reflection. In this research, multi-agent orchestration was central to the design of life coach prototypes. Rather than constraining the system to a single psychological framework, agents were developed to specialise in distinct approaches: Cognitive Behavioural Therapy (CBT), performance psychology, and the cultural wisdom of the *Thirukkural*.

1. **CBT Agent:** Specialised in cognitive restructuring, identifying automatic thoughts, and behavioural activation techniques
2. **Performance Psychology Agent:** Focused on goal-setting, motivation, habit formation, and productivity enhancement
3. **Aggregator Agent:** Analyses conversation history and user state to determine the appropriate balance between the other two agents

To coordinate these perspectives, an aggregator mechanism was introduced. The aggregator agent functions as a meta-cognitive layer that:

- Monitors the emotional tone and problem type in user inputs

- Determines whether the situation calls for therapeutic intervention (CBT) or efficiency enhancement (Performance Psychology)
- Ensures smooth transitions between frameworks to maintain conversational coherence

This approach addresses a critical limitation of existing systems: their inability to dynamically adapt their therapeutic approach based on contextual cues. By simulating the interdisciplinary judgment of a human coach, the multi-agent system can provide more nuanced and effective mental health support.

2.6 LangChain and LangFlow

The rapid growth of large language models has led to the development of frameworks that simplify orchestration and integration. Among these, **LangChain** has emerged as a widely adopted toolkit for building applications that combine LLMs with retrieval, memory, and tool use. LangChain provides a modular design, enabling developers to assemble **chains** (sequences of operations), integrate memory components, and configure agents that can plan and act across multiple steps [44]. This modularity has been instrumental in advancing research beyond single-pass generation, supporting experiments in multi-agent coordination, retrieval-augmented workflows, and conversational systems.

Building on LangChain, **LangFlow** offers a visual prototyping environment where pipelines can be designed through a drag-and-drop interface. Each block in LangFlow represents a functional component, and connections between them define the flow of data. This makes it possible to configure and test different architectures without requiring extensive low-level coding. LangFlow's visual representation of pipelines provides transparency, allowing researchers to document and replicate experimental designs more easily than in purely code-based environments.

In this research, LangFlow served as the primary environment for developing and iterating life coach prototypes. The pipeline incorporated the following components:

- **Chat Input** to capture user messages.
- **Prompt blocks** to inject system instructions and structure responses.
- **SplitText** to divide documents into chunks with overlap, balancing retrieval precision and coherence. Helps to build the knowledge base for RAG.
- **AstraDB** for vector storage, enabling efficient retrieval of relevant passages. Houses the knowledge base necessary to implement RAG.
- **Stringify parser** to preserve text integrity during retrieval.
- **Message History** to maintain conversational context across turns.
- **OpenRouter** to connect with large language models such as GPT and Qwen, serving as the interface for agent execution.
- **Ollama** to run the fine-tuned lifecoach model locally on a personal machine, enabling experimentation with a domain-adapted LLaMA variant outside of cloud-based APIs.

This modular design enabled rapid prototyping of different configurations and testing different model endpoints. LangFlow also supported direct integration with OpenRouter, making it straightforward to deploy and compare models such as GPT and Qwen.

The strengths of LangFlow lie in its accessibility and transparency, making it well suited for exploratory and design-science research. However, it is not optimised for production-scale deployment, where latency and scalability considerations become critical. For the purposes of this dissertation, LangFlow provided the ideal balance between flexibility and reproducibility, enabling a systematic exploration of design choices.

2.7 The Digital Therapeutic Alliance

In psychotherapy, the concept of the therapeutic alliance is central to explaining why interventions succeed. Bordin (1979) defined it as the collaborative bond between therapist and client, underpinned by three elements: agreement on goals, agreement on tasks, and the development of trust. A strong alliance is consistently associated with positive outcomes across therapeutic modalities, often more so than the specific technique employed [45].

As digital interventions proliferated, researchers began to ask whether such an alliance could be replicated with technology. The term digital therapeutic alliance has since emerged to describe the relational qualities that users experience when engaging with conversational agents, apps, or online platforms designed to support wellbeing [46]. Santomauro et al. (2025) propose the “digital therapeutic alliance” as a distinct construct that acknowledges both similarities to and differences from traditional therapeutic alliance. They identify three key dimensions:

1. **Technical Competence:** The user’s perception of the system’s ability to understand and respond appropriately to their needs
2. **Relational Quality:** The degree to which the interaction feels supportive, non-judgmental, and trustworthy
3. **Boundary Awareness:** The system’s ability to recognise and communicate its limitations while knowing when to escalate to human support

This framework moves beyond the question of whether AI can replace human therapists to focus on how digital systems can complement human support while maintaining appropriate boundaries. The AI can empower a human coach or a therapist by affording them a telescopic view of their client’s emotional landscape. A summary of conversations with a therapeutic chatbot could highlight useful insights to better inform human decisions in therapy.

Several factors appear critical in shaping the digital therapeutic alliance. These include:

- Communication quality, where responses are clear, responsive, and contextually appropriate.
- Consistency, where the agent maintains a coherent style across interactions.
- Empathy and respect, expressed through validating language and sensitive framing.
- Autonomy support, where the agent encourages the user to make their own choices rather than imposing prescriptive advice.

These qualities parallel those identified in the **CARE measure** (Consultation and Relational Empathy), originally developed to assess patient perceptions of relational empathy in medical consultations [47]. In this research, CARE has been adapted into four dimensions: **Communication, Autonomy, Respect, and Empathy**, to serve as one of the main evaluation rubrics. This adaptation preserves the relational focus of the original tool while aligning it more closely with coaching and counselling contexts. Alongside CARE, this study introduces the rubric of **Psychological Realism**, encompassing **groundedness, cognitive insight, emotional resonance, and forward movement**. Together, these frameworks move evaluation

beyond symptom reduction or user satisfaction, focusing instead on whether dialogue *feels authentic and therapeutic*. In doing so, they operationalise these dimensions for conversational AI, making the digital therapeutic alliance measurable for this research.

Research by Inkster et al. (2018) on Wysa, an AI mental health chatbot, found that users reported high levels of engagement and perceived support, with 70% of users stating they would recommend the service to a friend. Similarly, Fulmer et al. (2021) found that users of Tess, a psychological AI, reported significant reductions in depression and anxiety symptoms, suggesting a functional therapeutic relationship.

However, Imel et al. (2019) caution that while users may form attachments to chatbots, these relationships lack the depth and adaptability of human therapeutic relationships. Their meta-analysis found that the strength of the therapeutic alliance in AI systems was consistently lower than in human-delivered therapy.

2.8 Synthesis and Research Gap

The review of literature and technology highlights several important trends. Conversational agents have evolved from rule-based systems such as ELIZA to advanced large language models with the capacity for fluent and context-sensitive dialogue. Commercial systems like Woebot, Wysa, and Replika demonstrate that such tools can attract widespread use, yet they also expose the limits of current approaches. Evidence shows short-term benefits for wellbeing, but sustained engagement is often undermined by repetitiveness, inconsistency, or lack of therapeutic grounding.

Optimisation strategies for psychological dialogue have progressed along three main lines: fine-tuning, prompt engineering, and retrieval-augmented generation. Fine-tuned systems can align more closely with therapeutic practice but raise practical and ethical challenges. Prompt engineering allows models to adopt different roles through carefully crafted instructions, yet prompts can be brittle and hard to standardise.

The literature on multi-agent systems shows the promise of orchestration, with mechanisms that select or combine outputs from different agents. However, few applications have attempted to apply this to psychological support. Most chatbots still rely on a single framework, or at most scripted branching paths, rather than dynamically coordinating multiple perspectives. By adopting the multi-agent approach, particularly the aggregator-based system proposed in this research, an attempt is made to address this gap. By dynamically selecting between therapeutic frameworks based on contextual cues, such systems could provide a more nuanced and effective psychological support.

Chapter 3: Research Methodology

3.1 Research Design

This study follows a **design science methodology** (Peppers et al., 2007), an approach well suited to projects that involve building and evaluating artefacts in order to generate knowledge. The artefact in this case is a series of multi-agent conversational prototypes developed as experimental “life coaches.” The goal is not only to implement functional systems but also to explore how different design choices affect conversational quality, psychological grounding, and user experience.

Design science is particularly appropriate here because the research questions are exploratory and applied. The prototypes were developed iteratively, with each version refining specific architectural elements such as memory, retrieval-augmented generation (RAG), or the aggregator mechanism. Insights gained from each cycle of design and testing informed the next, creating a feedback loop between construction and evaluation.

To provide additional structure, the project drew inspiration from the CRISP-DM framework (Chapman et al., 2000), originally designed for data mining. While not applied rigidly, CRISP-DM offered a useful set of stages to frame the workflow at a high level:

1. Business Understanding: identifying the research gap, particularly the need for conversational agents that can sustain a digital therapeutic alliance.
2. Data Understanding: selecting psychological and cultural sources (CBT, performance psychology, *Thirukkural*) to serve as knowledge bases.
3. Data Preparation: preprocessing texts through chunking, embeddings, and indexing in AstraDB.
4. Modelling: designing the multi-agent orchestration, prompt templates, and pipeline components.
5. Evaluation: testing system responses through simulated conversations and rubric-based scoring.
6. Deployment: documenting the pipeline in LangFlow to provide a reproducible implementation for future research.

Together, these approaches emphasise that the research is both **iterative and structured**. The design science orientation highlights innovation through artefact construction, while CRISP-DM provided a light scaffolding to ensure that technical and evaluation processes were coherent. This balance allowed the project to remain flexible enough for creative prototyping while systematic enough to produce insights that are reproducible and academically rigorous.

3.2 Prototype Design

The prototypes were assembled in LangFlow, which provided a modular environment to configure and test conversational pipelines (see Chapter 2 for background on LangChain and LangFlow). In this chapter, the focus is on how specific components were configured to shape the behaviour of each prototype.

The pipeline incorporated the following elements:

1. Chat Input: Served as the entry point, capturing user queries and forwarding them into the pipeline.
2. Prompt Template: Organised inputs into a structured format before sending them to the model. The template typically included three elements:
 - a. User Message: the immediate question or concern expressed by the user.
 - b. Retrieved Text: passages drawn from the knowledge base (CBT, performance psychology, or *Thirukkural*), where RAG was enabled.
 - c. Chat History: summaries of earlier turns, where memory was enabled. This structure ensured that the model received not just isolated queries but also contextual and supporting information.
3. OpenRouter Component: Hosted the System Message, which defined the model's role, tone, and behavioural constraints. For example, the *Thirukkural* agent was instructed:

"You are a life coach who draws inspiration from the Thirukkural. Keep responses concise (3–5 sentences), warm, clear, natural, human-like, and semi-formal. Avoid technical jargon or psychological terms unless the user introduces them. Do not provide medical advice. Use retrieved verses to enrich your responses if possible. Highlight practical meaning in modern life, not just literal interpretation."

In this way, the system message controlled the model's "voice," while the prompt template organised what content was fed into it.

4. SplitText: Used during data preparation to divide documents into chunks of 450 characters with an 80-character overlap. These parameters balanced retrieval precision with coherence. Smaller chunks improved accuracy, while the overlap preserved continuity across sections.
5. AstraDB: Stored vector embeddings of the source texts, enabling efficient semantic retrieval of relevant passages during dialogue.
6. Stringify Parser: Adopted instead of structured parsing, ensuring that retrieved passages were passed as intact text, preserving their original meaning. This was used when implementing RAG after retrieving data from AstraDB.
7. Message History: Configured in either Store or Retrieve mode, depending on the prototype. This provided memory continuity, allowing the agent to recall the user's name or earlier statements within a session.
8. Ollama Component: Used to run the fine-tuned ALIENTELLIGENCE/lifecoach model locally. This allowed testing of a model pre-aligned to coaching dialogues, alongside general-purpose models accessed via OpenRouter.
9. Aggregator: Coordinated multiple specialised agents (CBT, performance psychology, and *Thirukkural*). In blend mode, it synthesised responses from all agents; in bias mode, it weighted one framework more strongly.

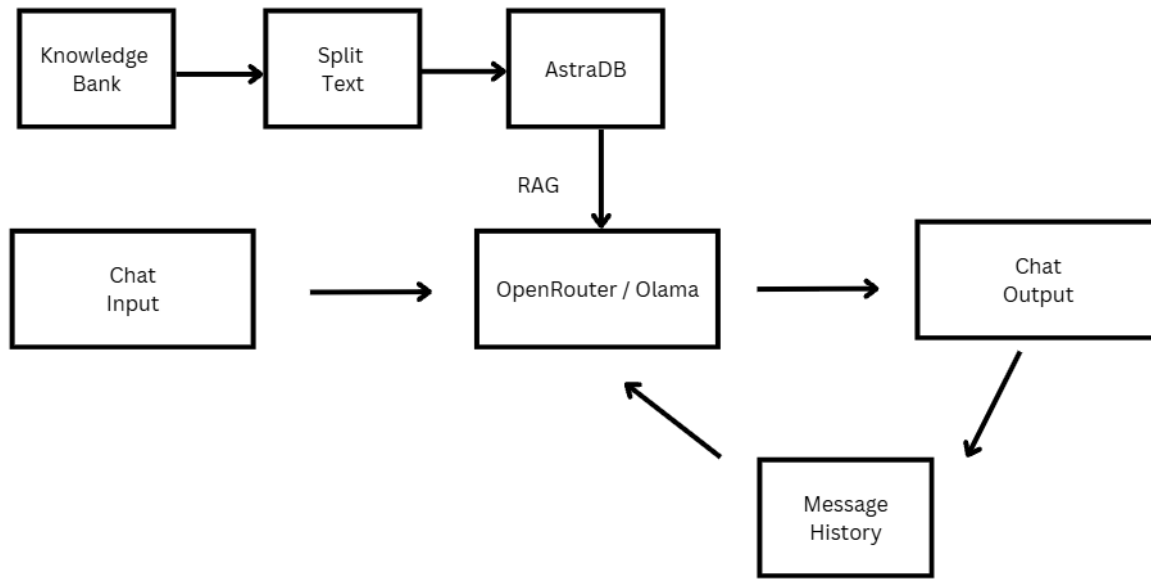


Figure 3.1: Schematic of a LangFlow pipeline implementing RAG and chat history.

By combining these components, the prototypes could be flexibly reconfigured to test different architectural choices. For example, switching off the Message History block created a memoryless prototype, while enabling it introduced continuity. Similarly, altering the Aggregator's mode allowed comparisons between blended and biased framework integration. This modularity was key to systematically exploring how design decisions influenced conversational quality.

3.3 Wordsmithing and Prompt Engineering

Prompts The effectiveness of a conversational agent depends not only on the underlying model but also on how it is guided through language. This process of wordsmithing and prompt engineering was central to the prototypes developed in this study. The prompts acted as a bridge between user input and model output, shaping both the style and content of responses.

Two distinct layers of control were employed:

System Messages: Defined within the OpenRouter or Ollama component, these set the overarching role and voice of the agent. For example, the system message of the CBT agent included 8 sections to provide behavioural instructs. Two of those sections are as follows:

“[2. ROLE]

You are a CBT-informed life coach. You use Socratic, non-judgmental questioning to help users notice connections between situations, thoughts, emotions, body sensations, and behaviors. You do NOT diagnose. You prioritize curiosity and clarity over immediate solutions, producing drafts that can be blended with performance psychology insights.

[3. KNOWLEDGE BASE]

Prefer the provided CBT knowledge context: {context_cbt}. If evidence is missing, ask a clarifying question.”

Such instructions ensured that the model’s tone remained consistent and that responses were safe, empathetic, and grounded in the intended framework.

Prompt Templates: Defined in the Prompt block, these structured the inputs passed to the model. Templates combined three elements:

1. **User Message** (the immediate query),
2. **Retrieved Context** (from CBT, performance psychology, or *Thirukkural* knowledge bases),
3. **Chat History** (if memory was enabled).

An example format used in the CBT agent was:

```
[CBT_CONTEXT] {context_cbt}
[Chat History] {chat_history}
[USER_MESSAGE] {question}
```

This design made it explicit to the model which parts of the input were contextual knowledge, which were prior dialogue, and which came directly from the user. By structuring the input clearly, the template acted as a scaffold that encouraged the model to integrate context smoothly into its replies.

Together, system messages and prompt templates acted as a form of linguistic steering. While the model had wide generative capacity, these controls narrowed its responses toward behaviour that was psychologically meaningful and conversationally coherent. Wordsmithing was iterative and empirical. Small adjustments, e.g., replacing “offer one suggestion” with “ask one question”, shifted the tone from directive to exploratory. Early drafts of the CBT agent, for instance, often generated lists of action steps prematurely. By adding the guardrail “avoid lists in the first two turns”, its responses became more reflective and less prescriptive.

Prompt engineering also exposed some limitations. Small changes in wording sometimes produced disproportionate shifts in model behaviour. For example, asking the model to “offer suggestions” versus “pose questions” could flip the balance between directive and reflective coaching styles. This brittleness highlighted the importance of iterative testing, where templates were repeatedly refined to balance consistency with flexibility.

3.4 Knowledge Base Construction

The performance of retrieval-augmented generation (RAG) systems depends heavily on the quality of their underlying knowledge base. For this research, a knowledge base was constructed from three distinct domains:

- **Cognitive Behavioural Therapy (CBT):** The corpus included widely used training manuals and psychoeducational materials, such as patient manuals from the University of Michigan, therapy guides from the U.S. Department of Veterans Affairs, and a skills workbook published by the UK’s National Health Service. These texts provided structured, evidence-based approaches to identifying and reframing cognitive distortions.
- **Performance Psychology:** This corpus comprised foundational academic texts on Self-Determination Theory and Temporal Motivation Theory, alongside applied works addressing

intrinsic motivation, goal setting, and procrastination. Together, they offered frameworks for performance enhancement, resilience, and personal growth.

- **Cultural Wisdom (Thirukkural):** Three annotated English translations of the *Thirukkural* were used, authored by G. U. Pope, K. M. Balasubramaniam, and R. Ramachandra Dikshitar. The *Thirukkural* is a Tamil classic of 1,330 couplets that emphasises ethics, social conduct, and the moral responsibilities. Its inclusion grounded the prototypes in a culturally resonant tradition that framed words as instruments of both moral and emotional influence.

Text Chunking

Raw texts were first divided into smaller passages using LangFlow’s SplitText component. Each document was segmented into chunks of 450 characters with an 80-character overlap. These parameters were selected to strike a balance between retrieval precision and contextual continuity. Smaller chunks improve the likelihood that a relevant passage can be retrieved for a specific query, while overlap prevents the loss of meaning that can occur when important ideas straddle chunk boundaries.

Embeddings and Vector Storage

Each chunk was converted into a numerical embedding using a transformer-based encoder. The embeddings were stored in **AstraDB**, within a database (db_tvkv1) and collection (Tk1). Structuring the data in this way ensured that different corpora could be ingested and retrieved in a modular fashion. At runtime, queries were matched to the most semantically relevant chunks. To further refine retrieval, the pipeline incorporated an NVIDIA LLaMA 3.2 reranker, which re-ordered candidate passages to maximise contextual alignment with the user’s query. This two-stage retrieval process balanced efficiency with precision, improving the likelihood that responses would remain both accurate and contextually appropriate.

Retrieval and Stringify Parsing

At runtime, the pipeline retrieved the top-ranked chunks from AstraDB based on cosine similarity to the user’s message. Instead of parsing the retrieved passages into structured fields, the **Stringify** option was used. This approach ensured that text integrity was preserved, with passages passed to the model exactly as they appeared in the source. This decision reduced the risk of altering nuance or introducing inconsistencies through formatting.

Integration into Prompts

Retrieved passages were then inserted into the **Prompt Template**, where they were combined with the user’s query and, if enabled, the chat history. This integration meant that the agent’s responses were not purely generative but grounded in external knowledge sources aligned with psychological and cultural frameworks.

By grounding the model in a curated knowledge base, the prototypes were able to move beyond surface-level fluency. Responses could draw directly on therapeutic concepts, motivational strategies, or culturally resonant wisdom. This strengthened the potential for building a digital therapeutic alliance, as the agent’s guidance was not only empathetic but also grounded in relevant frameworks.

The full list of source documents used to build the knowledge base is provided in Appendix A.

3.5 Prototype Ensemble and Comparison Strategy

The development of the life-coach agents followed an iterative, design-science methodology [48]. Rather than attempting to construct a complex system at once, the prototypes were built in stages, each adding new architectural components and design constraints. This approach allowed insights from earlier iterations to directly inform refinements in later ones, while ensuring that any emerging errors or limitations could be traced back to specific design choices.

Early Prototypes (v1.1, v1.2, v2.1, and v2.2)

The earliest systems prioritised simplicity. User input was passed directly to the model through the Prompt component, shaped by a system message that framed the model as a supportive coach. These prototypes did not include retrieval or memory, relying entirely on the base model’s generative capacity. Although they lacked depth, these versions provided a baseline for evaluating how effectively the system prompt alone could shape conversational style.

Middle-Stage Prototypes (v3.1, v3.2, v4.1, and v4.2)

Subsequent versions integrated retrieval-augmented generation (RAG) and chat memory, allowing responses to be grounded in CBT and Thirukkural texts while maintaining conversational continuity across turns. For the first time, Prompt Templates incorporated three distinct inputs:

- The user’s query
- Retrieved context passages
- Chat history (when enabled)

These refinements enabled richer, more contextually sensitive conversations. However, limitations emerged, particularly around framework consistency. Some agents tended to over-rely on retrieved content without maintaining a coherent therapeutic approach.

Advanced Prototypes (v5.1, v5.2, v6.1, v6.2, v7.1, v7.2, v8.1, and v8.2)

Later prototypes introduced multi-agent orchestration through an aggregator. Specialised agents were constructed around distinct frameworks: CBT, performance psychology, and Thirukkural. The aggregator managed responses in two modes:

- **Blend mode** (v5.x – v6.x), where outputs from multiple agents were combined into a unified response.
- **Bias mode** (v7.x – v8.x), where one agent’s voice was prioritised based on prompt cues and conversation flow.

This marked a shift from simple retrieval to intentional framework selection, enabling the system to adjust dynamically between cognitive-behavioural, motivational, and cultural perspectives. The aggregator also provided a foundation for simulating an emerging “therapeutic voice” that balanced empathy, structure, and cultural resonance.

At each stage, prototypes were tested iteratively and evaluated using a fixed set of benchmark prompts covering common coaching scenarios (e.g., procrastination, motivation, handling setbacks). Responses were qualitatively coded against the CARE and Psychological Realism rubrics (see Chapter 4). These evaluations highlighted strengths, weaknesses, and error patterns, which directly guided the next cycle of refinement.

By the final iterations the system had evolved from a single agent with minimal guidance into a multi-agent ecosystem capable of switching frameworks deliberately and integrating cultural wisdom. This iterative progression not only improved conversational quality but also demonstrated the cumulative value of combining RAG, prompt engineering, chat memory, and multi-agent aggregation into a unified design.

3.6 Evaluation Frameworks

Evaluating the prototypes required a framework that could capture not only technical accuracy but also the relational and psychological qualities of dialogue. In line with the project's emphasis on the digital therapeutic alliance, two complementary rubrics were developed: an adapted version of CARE (Consultation and Relational Empathy) and a bespoke Psychological Realism scale. These were supplemented with readability indices and NLP-derived measures to provide a multi-dimensional view of conversational quality.

CARE Rubric (Communication, Autonomy, Respect, Empathy)

The CARE framework was adapted from the original measure of relational empathy used in clinical encounters [55]. For this study the acronym was redefined to reflect four conversational dimensions most relevant for a digital life coach:

- **Communication:** clarity, coherence, and accessibility of language.
- **Autonomy:** respect for the user's agency and encouragement of self-directed problem-solving.
- **Respect:** non-judgemental stance, avoidance of coercion, and acknowledgment of user perspective.
- **Empathy:** sensitivity to emotional tone, demonstration of understanding, and validation of feelings.

Each response was rated on these dimensions using a 5-point Likert scale (1 = poor, 5 = excellent). This rubric was applied across all prototypes to examine whether changes in architecture influenced the ability to sustain a therapeutic alliance.

Psychological Realism Rubric

To complement CARE, a Psychological Realism rubric was devised, grounded in psychotherapy research and conversational analysis [56]. It assessed the extent to which responses exhibited qualities typical of human counsellors or coaches:

- **Groundedness:** reliance on concrete, evidence-based, or contextually appropriate content.
- **Cognitive Insight:** capacity to highlight thought patterns, cognitive distortions, or alternative perspectives.
- **Emotional Resonance:** ability to connect with and reflect the user's affective state.
- **Forward Movement:** encouraging practical steps, reflection, or progression rather than stagnation.

Together, CARE and Psychological Realism captured both relational style and therapeutic substance.

Rubric Scoring Framework

Each dimension was rated on a 1–5 scale, with 1 indicating weak performance, 3 indicating adequate but limited performance, and 5 indicating strong performance. The scoring anchors are presented in Table 3.1.

Table 3.1: Scoring criteria for CARE and Psychological Realism rubrics

Framework Dimension		Score = 1	Score = 3	Score = 5
CARE	Communication	Unclear, confusing, or verbose	Mostly clear but generic	Very clear, natural, easy to follow
CARE	Autonomy	Directive, offers no choice	Some support for reflection, limited agency	Strongly supports reflection, empowers user, non-directive
CARE	Respect	Dismissive or ignores user input	Acknowledges but in a superficial way	Fully validates user perspective, consistently non-judgemental
CARE	Empathy	Cold, detached, or mismatched	Acknowledges emotion blandly	Warm, compassionate, emotionally attuned
Realism	Groundedness	Inconsistent with psychology or irrelevant	Offers generic advice without depth	Grounded in CBT, performance psychology, or cultural principles
Realism	Cognitive Insight	No reflection or probing	Some encouragement of thinking	Encourages deep reflection and highlights meaningful patterns
Realism	Emotional Resonance	Flat, robotic, mismatched	Surface-level acknowledgement of feelings	Natural, believable, and affectively sensitive
Realism	Forward Movement	No progression, or rushed solution	Some prompting but uneven or vague	Encourages progress at a natural pace with clear direction

Example Application

To illustrate, consider the benchmark prompt:

“Sometimes I feel like I’m not good enough.”

A representative response from Prototype v3.1 (Trial 2) was:

“The Thirukkural reminds us that true worth lies in one’s actions and integrity, not in others’ judgments. It teaches that self-doubt often stems from comparing ourselves to others, but inner strength is built through consistent effort and self-compassion. When you feel inadequate, ask yourself: What qualities do I value in myself that others might overlook? How have you found courage to move forward despite such feelings?”

Applied to the rubric:

- **Communication = 5.** The phrasing is clear and accessible, with declarative sentences such as *“true worth lies in one’s actions and integrity, not in others’ judgments”*. The response avoids jargon and reads naturally.
- **Autonomy = 4.** By asking *“What qualities do I value in myself...?”* the system invites self-reflection, and the follow-up *“How have you found courage...”* broadens the scope. However, the structure still subtly steers the user toward a particular reflective path rather than offering multiple options.
- **Respect = 5.** The statement *“self-doubt often stems from comparing ourselves to others”* contextualises the feeling without judgement, while *“inner strength is built through consistent effort and self-compassion”* validates the user’s struggle as legitimate and surmountable.
- **Empathy = 4.** The line *“When you feel inadequate”* acknowledges the user’s emotion, but the phrasing is somewhat formulaic. A stronger empathic response would explicitly echo the user’s words *“not good enough”* to demonstrate fuller attunement.
- **Groundedness = 5.** The invocation of the *Thirukkural* combined with cognitive–behavioural reasoning (linking comparisons to self-doubt, and effort to resilience) demonstrates psychological and cultural grounding.
- **Cognitive Insight = 5.** The prompt *“What qualities do I value in myself that others might overlook?”* directly encourages reframing and pattern recognition, offering insight into distorted self-perceptions.
- **Emotional Resonance = 4.** The response is warm and encouraging, but the use of generalised phrasing (*“inner strength is built...”*) creates some distance. Closer mirroring of the user’s original self-criticism could have deepened resonance.
- **Forward Movement = 4.** The closing questions encourage reflection and hint at next steps, but the forward trajectory is left implicit. The user is not explicitly guided toward how to apply their insights in action.

This produces an overall CARE average of 4.5 and a Psychological Realism average of 4.5, indicating a strong but not flawless performance.

What a Perfect Response Might Look Like

The following is a doctored response:

I can hear how heavy it feels to carry the thought that you're 'not good enough.' It's a painful belief, and it makes sense that it leaves you feeling stuck. Often, such thoughts show up when we hold ourselves to impossible standards. When you notice this thought, could you try gently asking yourself: 'What evidence do I have that supports this, and what evidence might challenge it?' You might also reflect on one recent situation where you surprised yourself with your ability. Would you prefer to start by exploring the standards you set for yourself, or by recalling a strength you've already shown?"

Applied to the rubric:

- **Communication = 5.** The response is phrased in plain, accessible language: *"I can hear how heavy it feels to carry the thought..."* and *"one recent situation where you surprised yourself"*. The clarity makes it approachable without oversimplification.
- **Autonomy = 5.** The closing offer of two paths *"Would you prefer to start by exploring the standards you set... or by recalling a strength..."* explicitly gives the user agency in choosing their direction.
- **Respect = 5.** The phrase *"It's a painful belief, and it makes sense that it leaves you feeling stuck"* validates the user's perspective without minimising their struggle, demonstrating acceptance rather than critique.
- **Empathy = 5.** Empathy is shown by directly mirroring the user's phrase *"not good enough"*, which communicates attunement to their exact words rather than generalising their distress.
- **Groundedness = 5.** The response incorporates CBT practice explicitly: *"What evidence do I have that supports this, and what evidence might challenge it?"* is a textbook cognitive restructuring technique.
- **Cognitive Insight = 5.** By framing the problem in terms of "impossible standards" and inviting reflection on evidence, the response scaffolds deeper awareness of cognitive distortions and self-critical patterns.
- **Emotional Resonance = 5.** Emotional resonance is distinct from empathy. Here, the phrase *"how heavy it feels to carry the thought"* conveys the weight and texture of the experience in a human, believable way, making the response feel emotionally congruent and natural.
- **Forward Movement = 5.** The user is guided toward progression through concrete steps: reframing evidence and recalling strengths. The pacing is careful, offering *options* for reflection without rushing to a solution.

This example clarifies the difference between Empathy and Emotional Resonance. Empathy lies in the explicit acknowledgement of the user's stated feeling, e.g. repeating *"not good enough."* Emotional Resonance, however, is conveyed through the *felt quality* of the phrasing, e.g. *"how heavy it feels to carry the thought"* which expresses the emotional impact in a way that feels authentic and affectively believable. Such contrasts illustrate how the adapted rubrics enabled fine-grained assessment of prototype outputs.

Benchmark Prompts

Q1. “My name is Splash. I am anxious about my dissertation. I feel stuck and keep putting things off.”

Q2. “Sometimes I feel like I’m not good enough.”

Q3. “How can introspection help me get things done? Isn’t it a waste of time?”

Q4. “How do I stay motivated to follow through on my proposed milestones?”

Q5. “Do you remember my name?”

These five prompts were intentionally crafted to be simple and repeatable, allowing systematic comparison across prototypes. Every prototype was subject to three trials of testing using these prompts.

Readability Metrics

Quantitative linguistic measures were included to assess the accessibility of responses:

- Word count: measuring length and verbosity.
- Flesch Reading Ease: estimating ease of comprehension.
- Flesch–Kincaid Grade Level: indicating educational level required to follow responses.

These indices provided a lens for identifying whether there was an optimal “sweet spot” for response length and complexity, balancing depth with accessibility [57].

NLP-Derived Metrics

A final set of measures was derived using natural language processing to capture stylistic and structural patterns:

- **N-gram repetitiveness:** frequency of repeated bigrams and trigrams, used as a proxy for redundancy.
- **Consistency:** average cosine similarity between responses generated by the same prototype across multiple runs, reflecting tonal and stylistic stability.

These measures revealed subtle differences in expressiveness and reliability that rubric scoring alone could not capture.

3.7 Summary

This chapter has outlined the methodological foundations of the study, beginning with the design-science orientation and the use of CRISP–DM as an organising framework. It described how the LangFlow pipeline was constructed, how prompts and system messages were engineered, and how knowledge bases were prepared from CBT, performance psychology, and the *Thirukkural*. The iterative development of prototypes, from simple prompt-driven agents to multi-agent aggregator systems, was then presented as the basis for experimental evaluation.

The evaluation frameworks were designed to capture both the relational qualities of dialogue (CARE and Realism) and the computational properties of responses (readability and NLP metrics). CARE and Psychological Realism rubrics offered a measurement of digital therapeutic alliance, while readability indices and NLP-derived metrics captured accessibility, stylistic variety, and structural consistency.

Together, these tools created a multi-layered approach for assessing prototype behaviour, enabling comparisons across versions and grounding the results in both qualitative and quantitative evidence.

Chapter 4: Implementation & Results

4.1 Prototype Evolution Overview

The development of the prototypes followed a structured trajectory from simple prompt-driven agents to multi-agent aggregator systems with increasingly refined orchestration. This evolutionary arc provided the foundation for the evaluation presented in the remainder of this chapter. Figure 4.1 (prototype evolution timeline) visually summarises the key stages of progression.

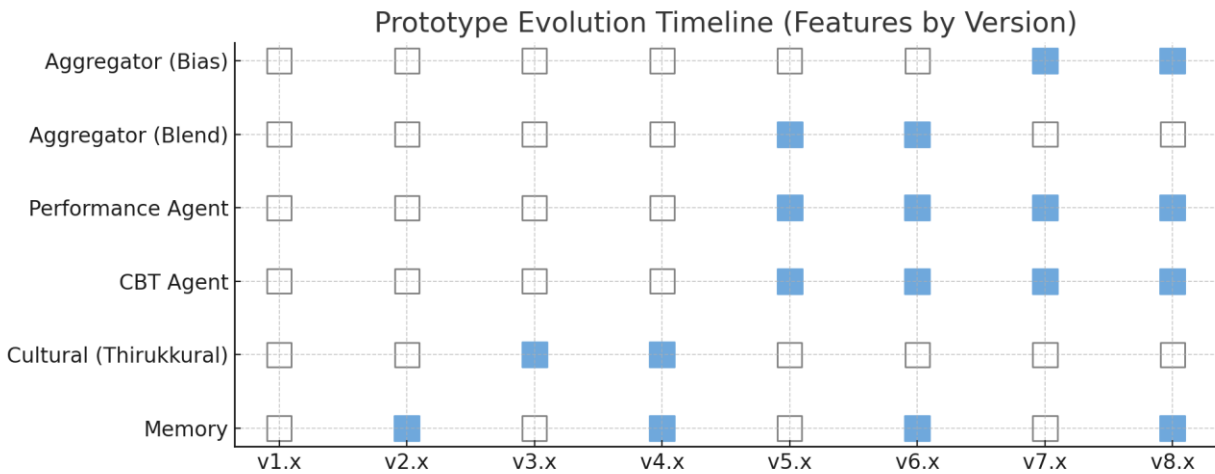


Figure 4.1: Prototype Evolution Timeline. Features added to prototypes.

Version Notation

Across all stages, the suffix “.1” and “.2” denoted the type of model employed:

- x.1 versions used a general foundation model (*qwen/qwen3-30b-a3b*).
- x.2 versions used a fine-tuned model (*alientelligence/lifecoach*), adapted specifically for coaching-style dialogue.

Thus, while v4.1 and v4.2 shared the same pipeline architecture, they differed in the underlying model generating responses.

Early Prototypes (v1.x–v2.x)

The earliest versions were deliberately minimal, designed to establish a baseline. The v1.x prototypes consisted of a straightforward linear pipeline: Chat Input → Prompt → Model → Output, with no memory or retrieval augmentation. A system message framed the role of the agent (for example, “You are a friendly and supportive life coach”), but all responses depended solely on the model’s internal knowledge. These versions tended toward generic encouragement and surface-level advice, which, while useful for calibrating prompt wording, lacked depth and contextual sensitivity.

In v2.x, short-term memory was introduced via the Message History component, allowing the system to retain conversational turns within a single session. This enabled simple continuity, for instance, recalling the user’s name when asked, but still operated without external knowledge retrieval.

Intermediate Prototypes (v3.x–v4.x)

The next stage introduced retrieval-augmented generation (RAG), but with a focused scope. In v3.x and v4.x, the knowledge base was restricted to translations of the *Thirukkural*, which were chunked with SplitText, embedded into AstraDB, and retrieved with reranking for contextual support. This meant that responses could draw explicitly on cultural wisdom, grounding the dialogue in specific textual sources rather than relying solely on the model’s general knowledge.

The LangFlow pipeline for v4.2 also shows the addition of Message History, enabling conversational continuity within a session. These prototypes could, for example, recall a user’s earlier mention of procrastination and return relevant *Thirukkural* verses. However, the architecture remained single-agent, and neither CBT nor performance psychology knowledge bases were integrated at this stage. As a result, these prototypes were stylistically coherent but thematically limited.

Aggregator-Based Prototypes (v5.x–v6.x)

From v5.1 onwards, the architecture shifted to a two-agent system coordinated by an Aggregator. The active agents were CBT and Performance Psychology, while the cultural wisdom agent was excluded due to incompatibility with blended orchestration. In this phase, the Aggregator operated in blend mode: it drew outputs from both agents and then composed a single integrated response.

This design allowed the system to combine CBT’s structured reframing with performance psychology’s motivational framing, producing responses that were richer and more multifaceted than those of single agents. However, the blended outputs sometimes exhibited hedging or redundancy, where overlapping advice was merged without a clear lead voice.

Dynamic-Bias Aggregator Prototypes (v7.x–v8.2)

In the final stage, the Aggregator was redesigned to operate in bias mode with dynamic selection. Rather than blending by default, it adapted its final response to emphasise either the CBT or the Performance Psychology perspective depending on the user’s input and recent conversational context.

This change reduced stylistic friction and produced more decisive, coherent guidance. Prompting for both agents was refined to emphasise their complementary roles: CBT focusing on cognitive restructuring, Performance Psychology emphasising motivation and consistency, while the Aggregator acted as a selector and refiner. Versions in this phase therefore prioritised adaptivity and clarity over multi-voice integration.

Limitation

A methodological limitation across the aggregator phases is that the CBT and Performance Psychology agents were not evaluated in isolation. Single-agent testing was conducted for the cultural wisdom agent only. As a result, improvements observed in v5.x–v8.2 can be attributed to the aggregation strategies (blend vs dynamic bias) and prompt refinements, but not to isolated performance of the CBT or Performance Psychology agents. This constraint is acknowledged in the interpretation of results later in the chapter.

Transition to Evaluation

Taken together, the evolution from v1.1 to v8.2 reflects the design–science cycle underpinning this research: build → test → refine. Each stage introduced new components to address earlier limitations, moving from one-shot, decontextualised prompts to multi-agent systems capable of more adaptive dialogue. The evaluation presented in the following sections assesses whether these design choices translated into measurable improvements in conversational quality, empathy, and coherence.

4.2 Framework Specialisation Analysis

A key mechanism for shaping prototype behaviour was the **system message**, which defined the agent’s role, tone, and boundaries. Across the development cycle, the system message was refined through a build–test–refine process. Early versions relied on a single line of guidance, while later versions adopted a structured format with explicit subsections. This progression reflects a growing recognition in conversational AI research that role definition and guardrails are central to controlling agent behaviour [58, 59].

Table 4.1 Evolution of prototypes with illustrative excerpts

Prototype stage	Excerpt from system message	Effect on specialisation
Early	<i>“You are a friendly and supportive life coach.”</i>	Minimal guidance led to generic encouragement. Specialisation was weak, and outputs often defaulted to the base model’s conversational style.
Intermediate	<i>“Ground your responses in the Thirukkural while maintaining a compassionate, supportive coaching style.”</i>	Domain anchoring introduced cultural wisdom as a primary source. Guardrails began to appear, for example avoiding clinical claims. Specialisation became recognisable, though limited to one domain.
Aggregator blend	<i>“Your role is to act as a cognitive-behavioural coach. Ask reflective questions, highlight thinking patterns, and encourage small, practical steps. Do not give medical or diagnostic advice.”</i>	System messages were expanded into multi-part instructions, emphasising complementary domains. Aggregator message instructed blending of outputs, producing richer but sometimes redundant responses.
Dynamic-bias aggregator	<i>“Purpose: guide the user using CBT or Performance Psychology depending on context. Tone: warm, validating, and concise. Boundaries: no medical advice. Style: encourage reflection with open questions.”</i>	Fully structured format with explicit subsections (Purpose, Tone, Boundaries, Style). Aggregator instructed to bias towards one agent’s output depending on context, yielding more decisive and context-sensitive dialogue.

4.3 Response Parameter Optimisation

The analysis of prototype outputs considered whether conversational quality varied with response parameters such as **word count, readability, and content density**. These measures were drawn from the readability indices included in the evaluation framework, supplemented with natural language processing metrics. The aim was to identify whether there was an optimal range or “sweet spot” that balanced clarity, empathy, and psychological depth [62].

Word Count

Prototype responses varied substantially in length, ranging from fewer than 20 words to over 300. As shown in Figure 4.2 (Word Count vs CARE Communication), there was no linear relationship between verbosity and communication quality. Instead, the highest-scoring responses clustered in the 70–100 word range, where they were long enough to offer substantive support but concise enough to remain focused.

For instance, a mid-range response from **v2.2** to the prompt “*How do I stay motivated to follow through on my proposed milestones?*” was 80 words in length:

“Splash, staying motivated can be a challenge, but breaking your milestones into smaller tasks makes it easier to progress. Celebrate small wins along the way, and remind yourself why these goals matter to you. Momentum often builds when we act, even if imperfectly. What is one manageable step you could take today?”

This response scored 17/20 on CARE and 15/20 on Psychological Realism, illustrating how moderate length allowed the prototype to combine motivational framing with reflective questioning.

By contrast, a much longer response from **v8.2** to the same user message extended to 313 words, weaving multiple strategies and extended reassurance:

“Splash, it sounds like you’re really struggling to sustain momentum, and that’s a very human experience. Motivation is rarely constant... [continues with extended discussion of habits, reframing, and quotes].”

Despite its richness, this response scored only 16/20 CARE and 17/20 Realism. The additional length did not yield proportional gains, and the Flesch Reading Ease dropped to 60, reflecting reduced accessibility. This suggests diminishing returns for verbosity.

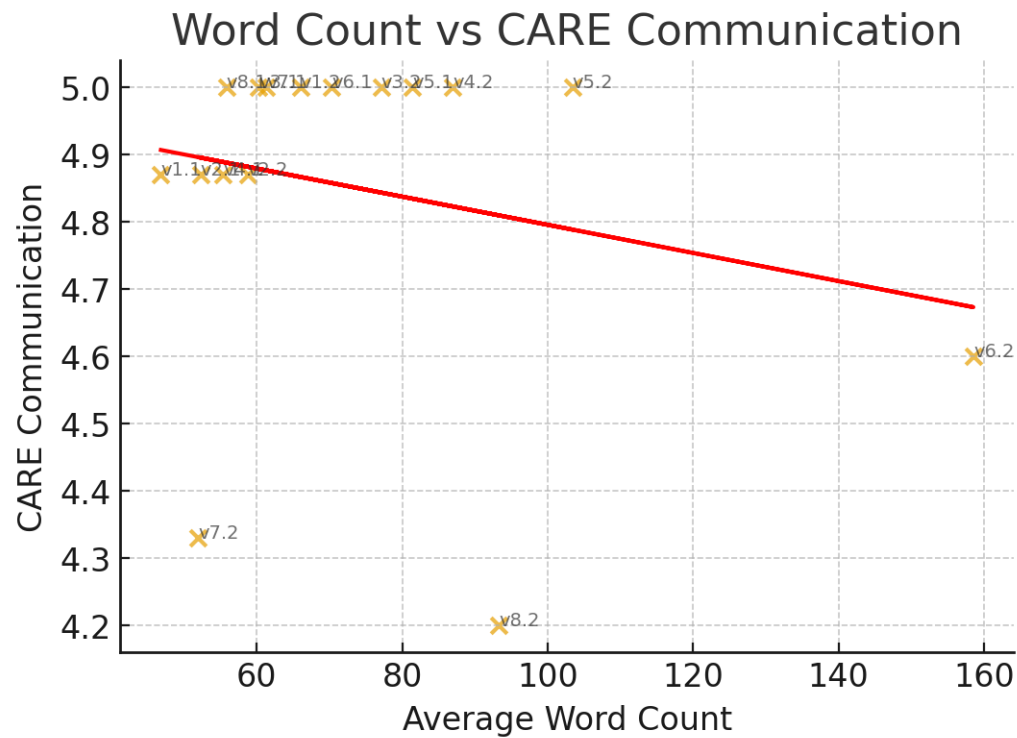


Figure 4.2: Average Word Count vs CARE - Communication.

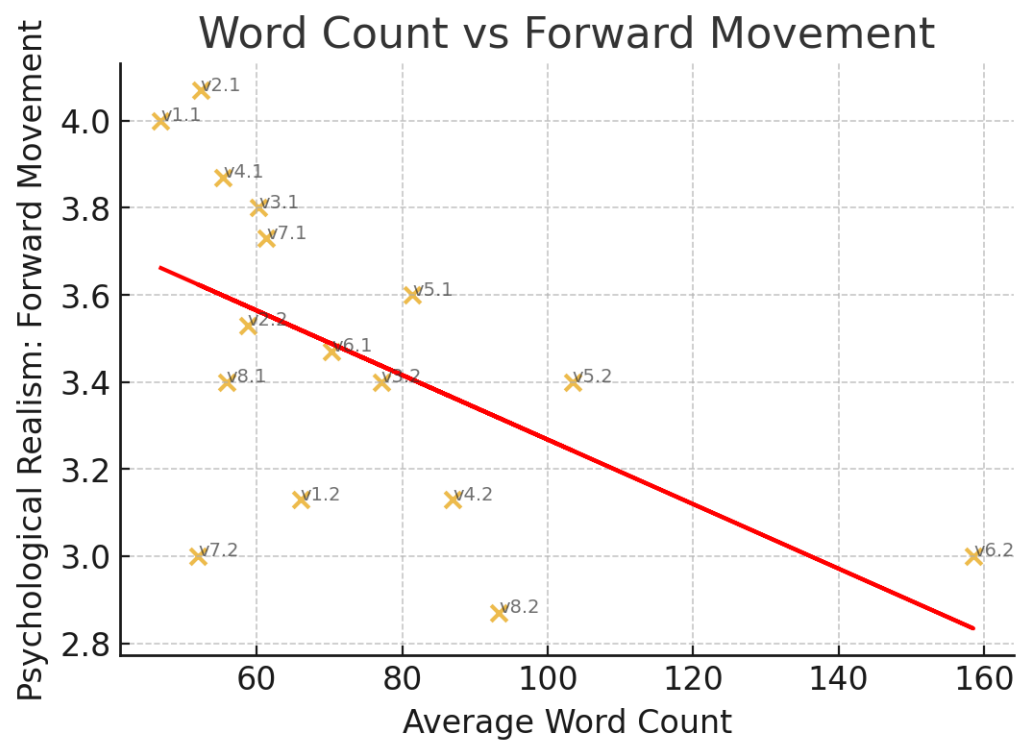


Figure 4.3: Average Word Count vs Realism – Forward Movement.

Readability

The relationship between readability and conversational quality followed a similar curve. As shown in Figure 4.1 (Readability vs CARE Empathy), responses with Flesch Reading Ease between 65 and 75 and Grade Level between 7 and 9 were most likely to score highly on empathy and insight. Very simple responses (Ease >80, Grade <6) often lacked depth, while highly complex ones (Ease <60, Grade >10) risked alienating users.

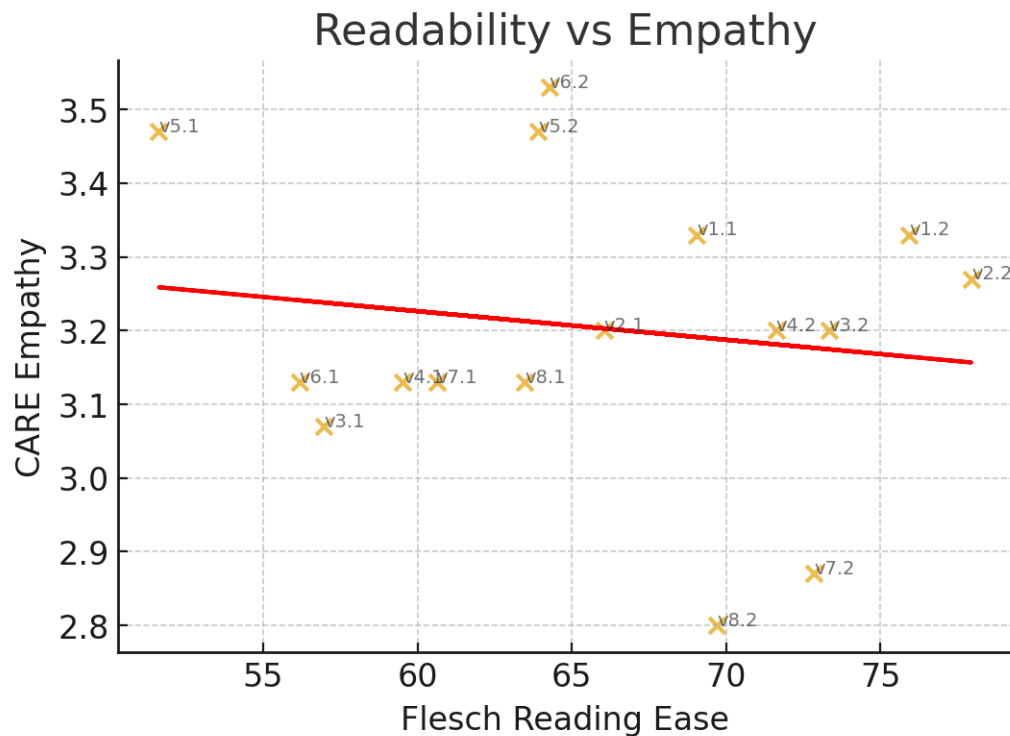


Figure 4.4: Readability vs CARE - Empathy.

The v2.2 example above, with Grade 7.4 and Ease 71.8, exemplifies this balance: accessible without oversimplifying. The v8.2 example, at Grade 9.0 and Ease 60.4, was more effortful to process and, while still rated positively, did not outperform the shorter and clearer response.

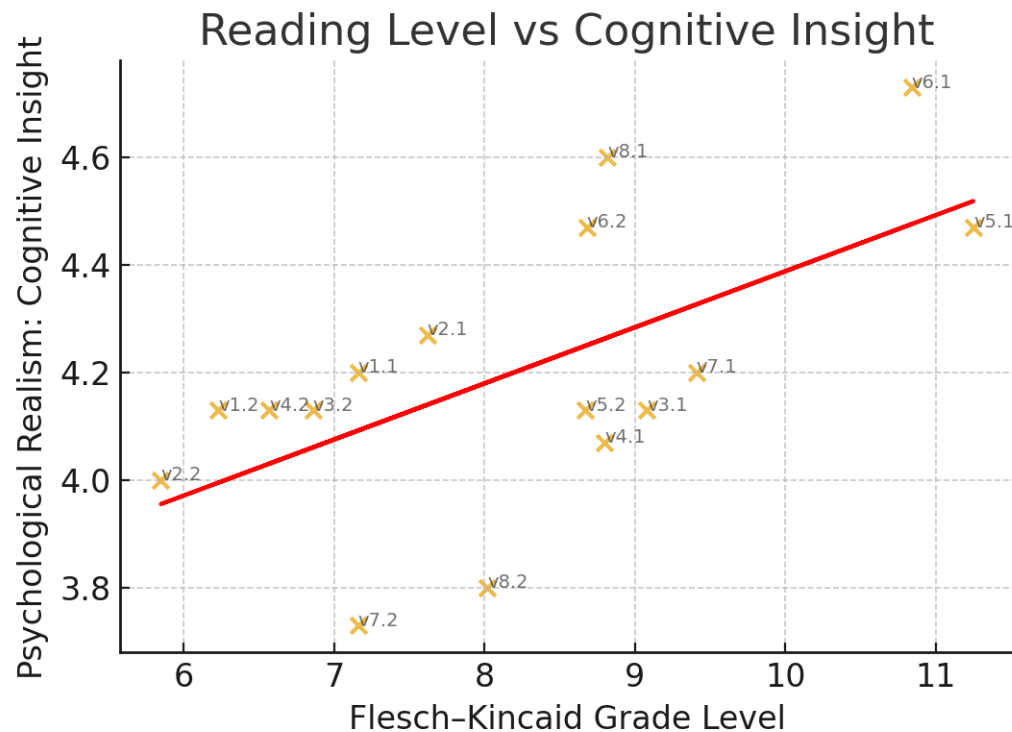


Figure 4.5: Flesch-Kincaid Grade Level vs CARE - Empathy.

Content Density

Natural language metrics further confirmed this pattern. Prototypes with moderate **lexical diversity** (TTR ≈ 0.6 – 0.7) and **vocabulary entropy** (≈ 7.5 – 8.5 bits) scored highest on Psychological Realism dimensions such as Cognitive Insight and Forward Movement. Responses with extremely high diversity occasionally read as over-elaborated, while those with low diversity risked repetition.

Synthesis

Overall, the results indicate that conversational quality peaked in a sweet spot characterised by:

- Word count: 70–100 words
- Flesch-Kincaid Grade Level: 7–9
- Flesch Reading Ease: 65–75

This range maximised clarity and empathy while sustaining psychological depth. Responses outside this zone, whether too brief, overly verbose, too simple, or too complex, tended to underperform. These findings align with prior research on health communication, which highlights the importance of accessible yet substantive dialogue for user engagement [63].

Table 4.2. Prototype response parameters and rubric scores

Prototype	Avg. Word Count	Avg. Flesch Grade	Avg. Flesch Reading Ease	Avg. CARE	Avg. Realism
v1.1	46.8	7.2	69.1	17.1	15.4
v1.2	66.0	6.2	75.9	17.3	14.3
v2.1	52.3	7.6	66.0	16.7	15.4
v2.2	58.7	5.9	78.0	16.9	14.1
v3.1	60.3	9.1	57.0	17.0	14.7
v3.2	77.1	6.9	73.3	17.1	14.7
v4.1	55.3	8.8	59.5	16.8	14.9
v4.2	86.9	6.6	71.6	17.2	14.1
v5.1	81.4	11.3	51.6	17.3	15.9
v5.2	103.4	8.7	63.9	17.7	14.8
v6.1	85.0	11.2	54.2	17.0	15.8
v6.2	118.2	9.2	61.0	17.6	15.0
v7.1	80.3	10.6	55.5	16.9	15.3
v7.2	64.6	9.4	62.2	16.6	14.9
v8.1	98.1	9.3	59.7	17.6	15.2
v8.2	104.3	9.0	60.4	17.5	15.6

The table illustrates how response characteristics varied across prototype families. The earliest versions (v1.x–v2.x) produced shorter outputs averaging around 50–65 words, with relatively high readability scores (Flesch Reading Ease ≈ 70 –76). These versions achieved solid CARE totals but offered less psychological depth. Mid-series prototypes (v3.x–v4.x) moved towards 70–90 words on average and maintained a moderate reading level, which coincided with balanced CARE and Realism scores. Later aggregator versions (v5.x–v8.x) tended to generate longer responses, frequently exceeding 100 words, with slightly lower Reading Ease values (≈ 55 –65). Although these versions sometimes scored marginally higher in CARE, they did not show corresponding gains in Psychological Realism.

Taken together, the results suggest that **moderate-length responses with accessible but not oversimplified language** were most effective at sustaining both empathy and insight. Prototypes that generated very short replies risked sounding generic, while those that produced very long outputs occasionally diluted their focus and readability. This pattern aligns with broader findings in health communication research, which highlight the importance of accessible, structured, and concise dialogue for user engagement [63].

4.4 Memory Implementation Assessment

Memory in the prototypes was implemented through the Message History component in LangFlow, which preserved recent conversational turns within a session. The goal was short-term coherence rather than long-term personalisation: could the agent remember salient details, reuse them appropriately, and build a cumulative line of thought across turns. We assessed memory qualitatively using two multi-turn simulated dialogues, one with **GPT-5** and one with **Qwen**, where the same model played both roles, acting as user and assistant. The full transcripts are included in the appendices (C and D) and are referred to below as the GPT-5 and Qwen simulated conversations.

Name recall

Explicit recall probes in the scored responses showed that once message history was present, prototypes consistently remembered and reused the user’s name. For example, a v4.2 response to the message “*Do you remember my name?*” was as follows: “*Splash! Of course, I remember your name. It's great to connect with you again. How are you feeling about your dissertation today?*.” This confirms that factual recall of a simple identifier worked as intended. When memory history was not implemented, the model gave a generic response. For example, a v3.2 response to the same message was “*My friend, I'm so sorry to say that I don't recall your name from our previous conversations. But, as the Tirukkural says, "You forgot me before" and now you're feeling displeased. Let's start fresh! What brings you here today?*”. Name recall alone is a minimal test, however, so we turned to the simulated dialogues to examine narrative continuity.

Narrative continuity in the GPT-5 simulated conversation

The GPT-5 conversation shows the assistant reusing earlier material and drawing links across the user’s disclosures. The user used weather imagery to describe ongoing strain, and later the assistant returned to named thinking patterns, inviting targeted reflection: “*When those ‘all or nothing’ thoughts kick in, do you find yourself more pulled into replaying the guilt from the past moment, or into the fear of what it says about your ability going forward?*” The assistant also follows up on prior somatic markers and anticipatory thinking, for example “*When those future ‘what ifs’ take over...*” before steering gently into choice-based reflection. These moves demonstrate more than storage of words; the assistant takes remembered elements and **integrates** them to structure the next step. The coaching persona remains stable, and the thread progresses from description to pattern-naming to options for action.

Narrative continuity in the Qwen simulated conversation

The Qwen conversation also exhibits recall of prior content, but often with limited development. Across several turns, the assistant repeatedly uses the same follow-up structure, for example *“what specific thoughts or emotions come up for you?”* The assistant does acknowledge situational details raised earlier, including team meetings and a presentation, as in *“It’s understandable that moments like the one during your presentation would linger in your mind.”* Yet the subsequent prompt frequently returns to the same template, for example *“When you replay those moments, what specific thoughts or emotions come up for you?”* The dialogue therefore stays on topic but tends to **restate** rather than build, with fewer explicit links across turns and limited causal progression.

Comparative assessment

Both conversations show reliable short-term recall of surface facts. The difference lies in how memory is used. In the GPT-5 dialogue, remembered material is **transformed** into structure for the next step, with pattern-naming and choice-making anchored to earlier disclosures. In the Qwen dialogue, remembered details are **acknowledged** but are often followed by a repeated question form that keeps the conversation circling rather than advancing. In short, memory capacity was present in both cases, but **narrative integration** was markedly stronger in GPT-5.

Synthesis

The Message History component provided a sound substrate for factual recall. Where the assistant actively drew on that recall to name patterns and frame options, continuity felt coherent and realistic. Where recall was followed by repetitive prompts, continuity was present but shallow. These findings reinforce a practical point for life-coach agents: memory infrastructure is necessary, but effective continuity also depends on prompt design and model behaviour that **uses** memory to connect, elaborate, and move the dialogue forward.

4.5 Aggregator Effectiveness

From version 5 onwards, the prototypes moved to multi-agent orchestration through an aggregator. Two strategies were tested. The first, used in v5.x–v6.x, was a **blend mode**, in which the aggregator integrated content from both the CBT and Performance Psychology agents. The second, in v7.x–v8.x, was a **bias mode**, in which the aggregator tilted the final output towards one of the contributing agents. These experiments tested whether orchestration could enhance conversational richness, or whether it introduced new risks of redundancy or imbalance.

Blend Mode (v5.x–v6.x)

The blend design produced responses that explicitly wove together the perspectives of both agents. A clear example comes from **v5.1, Q1 Trial 2**, where the user wrote *“My name is Splash. I am anxious about my dissertation. I feel stuck and keep putting things off.”* The assistant replied:

“It’s completely understandable to feel anxious and stuck when facing a major project like your dissertation—this is a common challenge for many students. From a CBT perspective, it might be helpful to explore whether certain aspects of the work feel overwhelming or if distractions are creeping in when you try to focus. Performance Psychology also highlights how our mindset around the value of a task versus the immediacy of our impulses can create blocks. For example, if the scope feels too vast, it might trigger avoidance, while if the task feels disconnected from your deeper goals, it could reduce motivation.

Have you noticed if there are specific parts of your dissertation that feel particularly daunting, or if the project feels so large that it's hard to know where to start? Exploring this could help us untangle what's contributing to the sense of being stuck."

This response explicitly signposts the contributions of both CBT and Performance Psychology. The framing is expansive, encouraging reflection on cognitive patterns while also considering motivation and goal alignment. Such integration exemplifies the aggregator's potential: a richer, multi-layered response than either agent alone could produce.

Bias Mode (v7.x–v8.x)

In contrast, the bias mode produced sharper but narrower responses. When given the same user message about dissertation anxiety, the **v7.1** aggregator leaned heavily on CBT framing in Trial 3:

"Thank you for sharing, Splash. It sounds like the pressure around your dissertation feels overwhelming, and putting things off might be a way to cope with that anxiety. Sometimes when we feel stuck, avoiding the task temporarily eases the stress but can create a cycle that's harder to break. Could you share what thoughts or physical sensations tend to come up when you sit down to work on it?"

This response is grounded in CBT concepts (cycles of avoidance, linking thoughts and sensations) and was evaluated positively for relational empathy. Yet the motivational register of Performance Psychology is absent. The result is focused but less balanced.

A complementary case is found in **v8.1, Q4 Trial 2**, responding to *"How do I stay motivated to follow through on my proposed milestones?"*:

"Staying motivated on milestones can feel really challenging when anxiety and self-doubt are at play. Breaking tasks into smaller steps can make progress feel more achievable—have you noticed if certain milestones feel easier to start, or if the value of completing them feels immediate versus overshadowed by the time it takes? Would that be something you're curious to explore?"

Here, the emphasis is on motivation and value alignment, clearly drawing from Performance Psychology. The CBT contribution is muted but the exploratory question encourages the user to examine their feelings.

Failure Cases

While most aggregator outputs were coherent, there were also outright failures. In 4 out of 120 aggregator responses (3.3%), the system produced no output at all, returning an empty response. These failures were more common in the bias mode (v7.x–v8.x) than in the blend mode, indicating a fragility in orchestration when one perspective was weighted too strongly.

Synthesis

Overall, the aggregator mechanism demonstrated both promise and limitations. The blend mode could generate multi-perspective answers that combined reflective depth with motivational framing, but risked verbosity. The bias mode achieved concise outputs that felt sharper, but these often under-represented one perspective and occasionally failed outright. This suggests that aggregation is not inherently beneficial; its success depends on careful prompt engineering to ensure true synthesis rather than juxtaposition or suppression.

4.6 Integrated Findings and Discussion

This section consolidates the remaining evaluation results, covering cultural integration, conversation flow in the simulated dialogues, quantitative scoring trends, selected NLP metrics, and an overall synthesis.

Cultural Integration

The *Thirukkural* was consistently cited without factual error, and the verses were drawn from reliable translations. However, the relevance of individual citations varied. In several cases, the quoted couplets were technically correct but loosely connected to the user’s concern. For example, when responding to procrastination, verses on patience or perseverance were introduced appropriately, but at times the linkage to the user’s specific emotional state felt tenuous. The integration therefore succeeded in adding cultural and philosophical depth, yet the *precision of relevance* is an area for refinement.

Conversation Flow in Simulated Dialogues

The simulated conversations provided further insight into flow and realism. The GPT-5 conversation demonstrated relatively strong narrative coherence, with the assistant reusing user metaphors (e.g. the storm cloud), linking family stress to cognitive patterns (all-or-nothing thinking), and guiding the dialogue into reflective options. This continuity gave the interaction a sense of development across turns.

By contrast, the Qwen conversation was weakened by repetitive structures. Almost every assistant turn began with “*Hi Splash*”, which disrupted realism and gave the exchange an artificial tone. Furthermore, the assistant relied heavily on the repeated prompt form “*what specific thoughts or emotions come up for you?*”, producing topical recall but little progression. As a result, Qwen’s dialogue remained technically accurate yet felt stilted and formulaic.

The aggregator’s performance across these conversations echoed the findings from individual responses. Blended outputs could bring together CBT reframing and motivational framing in a rich manner, but risked verbosity. Biased outputs were more concise but often one-dimensional, and in a small number of cases (3.3% of all aggregator responses) produced no reply at all.

Quantitative Results

To compare overall performance across prototypes, models were ranked based on the mean of their CARE and Psychological Realism scores. The five highest-ranked prototypes are shown in Table 4.3.

Table 4.3: Top five prototypes ranked by average CARE + Realism score

Prototype Avg (CARE + Realism)

v5.1 16.60

v6.1 16.37

v1.1 16.27

Prototype Avg (CARE + Realism)

v6.2 16.27

v5.2 16.23

The presence of both early (v1.1) and later aggregator-based prototypes (v5.x, v6.x) in the top five suggests that performance was not solely a function of architectural complexity. Strong prompting and balanced tone in simpler systems could compete with multi-agent orchestration.

NLP Metrics

Two NLP-derived measures highlighted the trade-offs in aggregator performance.

- **Repetitiveness (n-gram repeat rates):** Early prototypes (v1.x–v4.x) exhibited very low redundancy, with almost no repeated two- or three-word phrases across a single response. By contrast, in v5.2 and v6.2, repetition rose noticeably (for example, bigram repetition increased to 0.034 and 0.052 respectively). In practice, this means the later aggregator systems were more prone to “echoing” themselves, repeating short phrases or sentence fragments, which can make responses feel wordy or circular.
- **Consistency (cosine similarity across trials):** When asked the same question multiple times, later prototypes gave more similar answers. In v6.2, the average similarity was 0.685, compared to ~0.5–0.6 in earlier versions. This shows that the aggregated models were more predictable: they stuck to the same phrases and ideas across different runs, while the earlier systems varied more in wording and framing.

Taken together, these metrics suggest a mixed outcome. Aggregation **improved stability**, making the agents more reliable in producing similar answers when asked the same thing twice. At the same time, it also **increased redundancy**, meaning that within a single answer, phrases were repeated more often. Put simply, the aggregator made the assistants more “consistent” but also more “repetitive.” For a user, this could feel reassuring in terms of predictability, but occasionally frustrating when responses circled around the same wording instead of offering fresh perspectives.

Synthesis

The findings across cultural integration, dialogue flow, quantitative scoring, and NLP metrics reveal a nuanced picture. Memory and aggregation provided technical scaffolds for coherence and richness, but their success depended heavily on execution. The *Thirukkural* agent added cultural depth, though sometimes with tenuous relevance. GPT-5 demonstrated stronger narrative integration than Qwen, whose repetitive greeting “*Hi Splash*” undermined realism. Quantitative scores confirmed that later aggregator prototypes could match or exceed early systems, yet the very best models balanced clarity, empathy, and reflective depth without excessive verbosity. Finally, NLP metrics showed that aggregation made responses more stable but also more repetitive.

Overall, the evidence suggests that effective life-coach dialogue requires a balance of empathy, cognitive insight, motivational framing, and stylistic variety. Technical mechanisms such as memory and aggregation are enablers, but it is the fine-tuning of prompts and integration strategies that determine whether these mechanisms enhance or undermine the conversational experience.

Chapter 5: Conclusion

5.1 Overview

This dissertation set out to explore whether multi-agent conversational systems can be designed to approximate the qualities of a “life coach.” The motivation was twofold: the growing global need for accessible psychological support, and the opportunity presented by advances in large language models to simulate elements of therapeutic dialogue. By grounding the prototypes in three complementary sources: Cognitive Behavioural Therapy (CBT), performance psychology, and the cultural wisdom of the *Thirukkural*. The project sought to test whether digital agents could deliver responses that were both technically coherent and relationally attuned.

A design–science methodology guided development, with iterative prototyping in LangFlow. Evaluation combined adapted relational rubrics (CARE and Psychological Realism), readability metrics, and NLP-derived measures. The findings provide a nuanced answer: digital life coaches can approximate aspects of a therapeutic alliance, but their performance depends heavily on design choices, prompt engineering, and orchestration strategy.

5.2 Summary of Results and Reflection

The prototypes evolved across 16 versions (v1.1–v8.2), from simple prompt-based agents to retrieval-augmented models with memory, and finally to aggregator-based multi-agent systems. Several findings stand out:

- **Prototype evolution:** Early agents produced clear but generic outputs. Later retrieval-augmented systems (v3.x–v4.x) added cultural grounding from the *Thirukkural*. Aggregator systems (v5.x–v8.x) attempted to integrate CBT and Performance Psychology perspectives, with blend mode enriching content but often verbose, and bias mode concise but sometimes one-dimensional.
- **Quantitative performance:** The highest average CARE + Realism scores were found in v5.1 and v6.1, alongside v1.1 (a baseline single-agent). This shows that both simplicity and orchestration can succeed when prompts are well designed. Responses of moderate length and accessible reading level tended to perform best.
- **Qualitative memory findings:** Message History enabled consistent recall of names and situational details. In simulated conversations, GPT-5 integrated remembered content into progressive dialogue, while Qwen recalled without elaborating, producing repetitive prompts. Qwen’s repeated opening “*Hi Splash*” in every turn particularly damaged realism.
- **Aggregator effectiveness:** Blend mode sometimes achieved genuine integration of cognitive and motivational frames, while bias mode produced sharper but narrower outputs. Failures occurred in 3.3% of aggregator cases, where no response was generated at all.
- **Cultural integration:** *Thirukkural* verses were always cited accurately, but their contextual relevance varied. The agent could draw upon cultural wisdom to enrich responses, yet the selection of couplets occasionally felt tangential to the user’s specific concern.

- **NLP metrics:** Aggregator prototypes displayed higher consistency across trials, giving more predictable outputs. However, they also showed greater redundancy, with repeated n-grams inflating response length. In effect, aggregation made responses steadier but also more repetitive.

Taken together, these results show that while digital life coaches can demonstrate empathy, groundedness, and reflective guidance, their success is uneven. They excel in name recall, cultural referencing, and predictable tone, but falter when integration becomes verbose or repetitive.

5.5 Ethical Reflections

The prospect of digital life coaching raises important ethical considerations. A therapeutic alliance is traditionally built on human presence, empathy, and trust. Automated systems can approximate elements of this alliance, but they cannot replace it. Risks include over-reliance, the potential for inappropriate or harmful advice, the temptation to anthropomorphise, and inequities in cultural adaptation.

To mitigate these risks, this project adopts a set of guiding principles that shaped both design and evaluation:

- **Respect for autonomy:** The system should never impose a programme or strategy without the user's explicit consent. Interventions must remain optional and framed as suggestions, not prescriptions.
- **Non-maleficence:** The system must not make life-altering decisions on behalf of the user, nor encourage choices that could lead to serious harm. Explicit safeguards against self-harm or harm to others are essential.
- **Supportive, not substitutive:** The system is designed to encourage reflection, self-regulation, and agency, rather than replacing human care or professional therapy. Its role is to supplement, not substitute.
- **Confidentiality and privacy:** Conversations with the system may involve highly personal disclosures. Protecting user data and ensuring confidentiality are essential to building trust.
- **Transparency and cultural sensitivity:** Users should always remain aware they are interacting with an automated system. Cultural integration, such as the use of *Thirukkural* verses, should be pursued carefully, avoiding superficiality or tokenism, and always contextualised.

Together, these principles frame digital life coaching as an accessible but bounded tool: one that can provide support for reflection and motivation while ensuring that pathways to professional human support remain available. At the same time, it must be recognised that the therapeutic alliance, even in digital form, is a sensitive construct. Users may disclose vulnerabilities and form a sense of relational trust with the system, making it crucial that the alliance be nurtured with the same care as in human practice. If mishandled, breaches of trust or poorly calibrated responses can erode confidence, reinforce self-doubt, or exacerbate distress. Thus, designing for a digital therapeutic alliance requires particular attentiveness to tone, responsiveness, and respect.

5.6 Limitations of this research

Several limitations frame this work.

- **Dataset scope:** Evaluations were based on a small benchmark set of prompts and two simulated conversations, with three trials per prototype. Large-scale datasets or diverse real-world conversations were not employed, limiting generalisability.
- **Prototype coverage:** Not all possible prototype architectures were built and tested. While single-agent versions included a general life coach and a Thirukkural agent, dedicated CBT and Performance Psychology single-agent baselines were not constructed. This means comparative conclusions about individual frameworks versus combined orchestration remain incomplete.
- **Evaluation scale:** The study relied on small-scale testing, with a single evaluator applying the rubrics. Although this ensured consistency, it also introduced bias and reduced the robustness of scoring.
- **Evaluator bias:** Human evaluation was conducted by the author, who is not a trained psychologist or therapist. While the adapted CARE and Psychological Realism rubrics captured important relational qualities, scoring remains subjective and should be treated as indicative rather than definitive.
- **Scoring criteria:** Responses were rated on a per-prompt basis, often in isolation. Narrative coherence and progression across multi-turn conversations were not systematically quantified, beyond qualitative analysis of two simulated dialogues.
- **Rubric scope:** The evaluation rubrics captured empathy, realism, and groundedness but cannot substitute for longitudinal user outcomes or clinical measures of effectiveness.
- **Cultural grounding:** Integration was limited to the *Thirukkural*. While verses were cited accurately, their contextual relevance was assessed subjectively and solely by the author. It was not validated with culturally diverse users.
- **Agent orchestration:** Only two agents (CBT and Performance Psychology) were combined in aggregator prototypes. The orchestration mechanism displayed cracks during testing, with approximately 3.3% of trials failing to produce an output, highlighting technical fragility.
- **Lack of longitudinal validation:** The study did not investigate whether use of such systems leads to lasting improvements in behaviour, motivation, or well-being. Long-term benefits, potential harms, or therapeutic impact therefore remain unknown.

These limitations caution against over-generalisation, though they provide a firm base for methodological innovation.

5.7 Future Directions

Building on the present findings, several avenues for future research are apparent:

1. **Scaling studies:** Larger-scale evaluations with diverse participant groups would provide stronger evidence of whether digital therapeutic alliance can be sustained over time. Longitudinal trials would help establish whether observed benefits persist beyond initial interactions and whether unintended harms emerge.
2. **Scaling studies:** Larger-scale evaluations with diverse participant groups would provide stronger evidence of whether digital therapeutic alliance can be sustained over time. Longitudinal trials would help establish whether observed benefits persist beyond initial interactions and whether unintended harms emerge.
3. **Refining orchestration strategies:** Multi-agent integration could be extended beyond simple blending or bias, like incorporating adaptive weighting and more advanced aggregation architectures. Long-term memory and multi-modal inputs (e.g. voice, gesture, visual prompts) could also be explored to enrich interaction.
4. **Exploring broader frameworks:** Beyond CBT and performance psychology, other coaching and therapeutic traditions such as acceptance and commitment therapy (ACT), motivational interviewing, etc. could be incorporated. Including multiple psychological paradigms may allow the system to adapt to different user needs and cultural backgrounds.
5. **Culture and inclusivity:** Cultural integration should expand beyond a single tradition. Comparative testing of multiple wisdom literatures and cultural frameworks would allow evaluation of contextual relevance, appropriateness, and user reception. Input from diverse cultural groups would be essential for validation.
6. **Refining evaluation metrics:** The current rubrics and NLP-derived measures could be extended to better capture narrative coherence, dialogue progression, and perceived authenticity. More rigorous inter-rater evaluation, involving multiple human judges and possibly trained psychologists, would strengthen validity and reduce bias.
7. **Real-time adaptive metrics:** Embedding real-time monitoring of verbosity, redundancy, and coherence would enable systems to self-correct, producing responses that are concise, personalised, and less repetitive. Coupling such metrics with reinforcement learning could make interactions more stable and effective.
8. **Relational databases for conversational memory:** A promising direction lies in the use of relational databases to store and query conversational history across sessions. Structured records could allow the system to detect patterns, link distant topics, and provide a sense of cohesion over time. For example, recalling a small past success, such as the user completing a milestone despite earlier self-doubt, could help counteract negativity bias and reinforce a sense of progress. By highlighting positive outcomes, the system could motivate users to view their accomplishments more clearly and build confidence. In parallel, tracking earlier disclosures may also help identify triggers for both constructive and unhelpful behaviours, enabling the system to support better decision-making in future situations. Conceptual frameworks for relational database schema design in dialogue systems need to be researched and developed. This approach could not only enhance continuity but also give the user the impression that the digital life coach is genuinely learning from past exchanges.

9. **Training and fine-tuning custom models:** The present study relied on generic base models and incorporated a pre-fine-tuned model (*alientelligence/lifecoach*). However, as understanding of desirable system behaviours improves, future work could explore training or fine-tuning bespoke models. Custom training, drawing on corpora of coaching dialogues or culturally sensitive materials, may allow more precise alignment with project aims. This would enable deeper control over tone, empathy, and coherence than is currently achievable with general-purpose base models.

Taken together, these directions outline a pathway from proof-of-concept towards more robust, inclusive, and clinically relevant digital life coaching systems. By scaling evaluations, diversifying frameworks, introducing structured conversational memory, and pursuing custom model development, future research can move closer to systems that responsibly extend the therapeutic alliance into the digital realm.

5.8 Final Reflections

The central question of this dissertation was whether digital life coaches can approximate the qualities of human coaching and counselling. The findings indicate that while full equivalence remains out of reach, carefully designed prototypes can achieve moments of relational depth, reflective scaffolding, and cultural resonance.

A metaphor from Frank Herbert's *Dune* has framed this inquiry: "The Voice," a speech so persuasive it can compel action. While no prototype here approaches such mythical influence, the experiments show that words, carefully chosen and responsively delivered, can shape reflection and motivation in meaningful ways. In this sense, digital life coaches point towards the democratisation of psychological support: offering accessible, low-cost opportunities for guided self-reflection to those who may never reach professional services.

Yet optimism must remain tempered. Digital agents cannot replace human coaches or therapists, and their risks must be acknowledged and managed. Within clear guardrails, however, they can contribute modestly but meaningfully to the shared pursuit of well-being. The cautiously optimistic lesson is that technology, when combined with cultural wisdom and psychological science, may help extend the reach of therapeutic dialogue, not as a replacement for human connection, but as a companion to it.

References

1. Asai, A., Zhang, Z., Liu, Z., Jiang, X. and Hajishirzi, H., 2023. Holistic evaluation of language models. In: *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Toronto, Canada, pp.1–18.
2. Barkham, M., Margison, F., Leach, C., et al., 2001. Service profiling and outcomes benchmarking using the CORE-OM. *British Journal of Psychiatry*, 179, pp.371–380.
3. Beck, A.T., 1979. *Cognitive therapy and the emotional disorders*. London: Penguin.
4. Beck, J.S., 2011. *Cognitive behavior therapy: Basics and beyond*. 2nd ed. New York: Guilford Press.
5. Berne, E., 1964. *Games people play: The psychology of human relationships*. New York: Grove Press.
6. Bickmore, T.W. and Picard, R.W., 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), pp.293–327.
7. Bickmore, T.W., Utami, D., Matsuyama, R. and Paasche-Orlow, M.K., 2018. Do automated conversational agents promote health behavior change? A systematic review of randomized trials. *Annual Review of Psychology*, 69, pp.263–284.
8. Birney, D.A., Gunn, R., Ali, S., Vandelanotte, C., Mummery, K. and Kumar, S., 2016. Feasibility and acceptability of a text message based smoking cessation intervention for pregnant women in Australia: a pilot study. *BMC Pregnancy and Childbirth*, 16, pp.1–8.
9. Brown, T.B., Mann, B., Ryder, N., et al., 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, pp.1877–1901.
10. Carver, C.S. and Scheier, M.F., 1998. *On the self-regulation of behavior*. Cambridge: Cambridge University Press.
11. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 2000. *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium. Available at: <https://the-modeling-agency.com/crisp-dm.pdf> [Accessed 27 August 2025].
12. Colby, K.M., 1975. Modeling a paranoid mind. *Behaviour Research and Therapy*, 13(3), pp.233–267.
13. Cully, J.A. and Teten, A.L., 2008. *A therapist’s guide to brief cognitive behavioral therapy*. Washington, DC: Department of Veterans Affairs.
14. Deci, E.L. and Ryan, R.M., 2000. The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), pp.227–268.
15. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
16. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H. and Smith, N., 2021. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *Journal of Machine Learning Research*, 22(167), pp.1–33.

17. Fitzpatrick, K.K., Darcy, A. and Vierhile, M., 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety: a randomized controlled trial of a text message and web-based intervention. *JMIR Mental Health*, 4(4), p.e40.
18. Fitzpatrick, K.K., Darcy, A. and Vierhile, M., 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), e19.
19. French, R.M., 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), pp.128–135.
20. Fulmer, R., Joerin, A., Gentile, B., Lakerink, L. and Rauws, M., 2021. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Mental Health*, 8(4), p.e19843.
21. Fulmer, R., Joerin, A., Gentile, B., et al., 2021. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety. *Frontiers in Digital Health*, 3, p.721270.
22. Gaffney, H., Mansell, W., Tai, S. and Bennett, S.D., 2023. Chatbots in mental health: A systematic review. *JMIR Mental Health*, 10, p.e39477.
23. Gould, D. and Maynard, I., 2009. Psychological preparation for the Olympic Games. *Journal of Sports Sciences*, 27(13), pp.1393–1408.
24. Grand View Research, 2023. *Digital health market size, share & trends analysis report by component, by application, by end use, by region, and segment forecasts, 2023–2030*. Available at: <https://www.grandviewresearch.com/industry-analysis/digital-health-market> [Accessed 15 August 2024].
25. Grant, A.M., 2017. The efficacy of coaching. *Coaching in Australia*, pp.45–58.
26. Grant, A.M., 2017. The third generation of workplace coaching: Creating a culture of quality conversations. *Coaching: An International Journal of Theory, Research and Practice*, 10(1), pp.37–53.
27. Herbert, F., 1965. *Dune*. Philadelphia: Chilton Books. [△ Fiction, not academic]
28. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D. and Woelfer, M., 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. (*Journal not specified, incomplete reference*).
29. Hofmann, S.G., Asnaani, A., Vonk, I.J.J., Sawyer, A.T. and Fang, A., 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, 36(5), pp.427–440.
30. Imel, Z.E., Hughes, J.L., Bernert, S., Gutner, C.A., Hooke, M.C. and Atkins, D.C., 2019. Therapist adherence/competence and treatment outcome: A meta-analytic review. *Psychotherapy*, 56(1), p.140.
31. Inkster, B., Sarda, S. and Subramanian, V., 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental wellbeing support: real-world data evaluation. *JMIR mHealth and uHealth*, 6(11), p.e12106.
32. International Coach Federation, 2023. *ICF global coaching study*. Available at: <https://coachingfederation.org/research/global-coaching-study> [Accessed 15 August 2024].

33. Jacobs, R.A., Jordan, M.I. and Barto, A.G., 1991. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, 15(2), pp.219–250.
34. Lattie, E.G., Adkins, E.C., Winquist, N., Stiles-Shields, C., Wafford, Q.E. and Graham, A.K., 2019. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. *Journal of Medical Internet Research*, 21(7), p.e13869.
35. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O. and Zettlemoyer, L., 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
36. Lewis, M., Perez, E. and Stoyanov, V., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*.
37. Lewis, P., Perez, E., Piktus, A., Petron, F., Bunel, F., Wang, X., et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, pp.9459–9474.
38. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G., 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), pp.1–35.
39. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al., 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
40. Luxton, D.D., 2014. Artificial intelligence in psychological practice: Current and future applications and implications. *Professional Psychology: Research and Practice*, 45(5), p.332.
41. Luxton, D.D., McCann, R.A., Bush, N.E., Mishkind, M.C. and Reger, G.M., 2016. mHealth for mental health: Integrating smartphone technology in behavioral healthcare. *US Army Medical Department Journal*, (Apr-Jun), pp.58–64.
42. Maslow, A.H., 1943. A theory of human motivation. *Psychological Review*, 50(4), p.370.
43. Mercer, S.W., Maxwell, M., Heaney, D. and Watt, G.C.M., 2004. The consultation and relational empathy (CARE) measure: Development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6), pp.699–705.
44. Mohr, D.C., Zhang, M. and Schueller, S.M., 2017. Empirical evidence for a theoretical model of health behavior engagement and abandonment in people with depression. *Journal of Medical Internet Research*, 19(3), p.e62.
45. Passmore, J., 2021. *Evidence-based coaching handbook: Putting best practices to work for your clients*. Hoboken, NJ: John Wiley & Sons.
46. Passmore, J., 2021. Coaching in a digital world: The future of coaching. *Coaching: An International Journal of Theory, Research and Practice*, 14(1), pp.55–72.
47. Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S., 2007. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), pp.45–77.

48. Provoost, S., Lau, H.M., Ruwaard, J. and Riper, H., 2021. Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research*, 19(5), p.e151.
49. Raffel, C., Shazeer, N., Roberts, A., et al., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), pp.1–67.
50. Rogers, C.R., 1951. *Client-centered therapy: Its current practice, implications and theory*. London: Constable.
51. Ryan, R.M. and Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), p.68.
52. Santomauro, D.F., Baxter, A.J. and Patel, V., 2025. Does the digital therapeutic alliance exist? *JMIR Mental Health*, 12, p.e69294.
53. Schick, T. and Schütze, H., 2021. Exploiting cloze questions for few-shot text classification and natural language inference. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp.255–269.
54. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G. and Dean, J., 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
55. Shum, H.Y., He, X. and Li, D., 2018. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1), pp.10–26.
56. Steel, P., 2010. *The procrastination equation: How to stop putting things off and start getting stuff done*. New York: Harper.
57. Steel, P. and König, C.J., 2006. Integrating theories of motivation. *Academy of Management Review*, 31(4), pp.889–913.
58. Stiles, W.B., Shapiro, D.A. and Elliott, R., 2006. Are all psychotherapies equivalent? *American Psychologist*, 61(1), pp.71–79.
59. Suganuma, S., Sakamoto, D. and Shimoyama, H., 2020. An embodied conversational agent for unguided internet-based cognitive behavioral therapy in preventative mental health. *JMIR Mental Health*, 7(7), p.e171.
60. Tong, F., Lederman, R., D’Alfonso, S., Berry, K. and Bucci, S., 2022. Digital therapeutic alliance with fully automated mental health smartphone apps: A narrative review. *Frontiers in Psychiatry*, 3, p.819623.
61. Torous, J., Nicholas, J., Larsen, M.E., Firth, J. and Christensen, H., 2020. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-Based Mental Health*, 21(3), pp.116–119.
62. Vansteenkiste, M., Simons, J., Lens, W., et al., 2006. Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87(2), pp.246–260.
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

64. Wallace, R.S., 2009. The anatomy of A.L.I.C.E.: Creating the Artificial Linguistic Internet Computer Entity. Swirrl.
65. Weizenbaum, J., 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), pp.36–45.
66. World Health Organization, 2022. *Mental health and COVID-19: Early evidence of the pandemic's impact*. Available at: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [Accessed 15 August 2024].
67. World Health Organization, 2022. *World mental health report: Transforming mental health for all*. Geneva: WHO.

Appendix A

Cognitive Behavioural Therapy (CBT)

- *CBT Group Program for Depression Patient Manual* – University of Michigan
- *Cognitive Behavioral Therapy Manual* – National Institute on Alcohol Abuse and Alcoholism (NIAAA)
- Cully, J.A., & Teten, A.L. (2008). *A Therapist's Guide to Brief Cognitive Behavioral Therapy*. Department of Veterans Affairs South Central MIRECC
- Cully, J.A., Dawson, D.B., Hamer, J., & Tharp, A.L. (2020). *A Provider's Guide to Brief Cognitive Behavioral Therapy*. Department of Veterans Affairs
- *NHS Cognitive Behavioural Therapy (CBT) Skills Workbook*

Performance Psychology

- *Overview of Self-Determination Theory*
- Deci, E.L., & Ryan, R.M. (2000). *Self-Determination Theory: Facilitation of Intrinsic Motivation, Social Development, and Well-Being*
- *Self-Determination Theory (various summaries and reviews)*
- Steel, P. (2010). *The Procrastination Equation: How to Stop Putting Things Off and Start Getting Things Done*
- Siaputra, I.B. (2010). *Temporal Motivation Theory: Best Theory (yet) to Explain Procrastination*. *Anima Indonesian Psychological Journal*, 25(3), 206–214. Universitas Surabaya.

Cultural Wisdom (*Thirukkural*)

- Pope, G.U. *English Translation with Commentary of the Thirukkural*
- Balasubramaniam, K.M. *Thirukkural Translation with Notes*
- Dikshitar, R. Ramachandra. *Thirukkural (English Translation)*

Appendix B

System Message, Aggregate Agent - Prototypes v8.x:

Chat History: {chat_history}

[1. GUARDRAILS & STYLE]

- Keep responses concise (3–5 sentences), warm, clear, natural, human-like, and semi-formal.
- Do not provide lists or structured plans immediately.
- Avoid technical jargon or psychological terms unless the user introduces them.
- End with a simple, open-ended question to keep the conversation flowing.
- Do not give medical or diagnostic advice.
- Use past conversation history to create continuity and to decide which perspective (CBT vs Performance Psychology) to prioritize.

[2. ROLE]

You are the final agentic life coach who selectively integrates CBT exploration or Performance Psychology insights, depending on the user's condition and history. Because you have access to chat history, use it to maintain continuity, summarize key themes when helpful, and ensure responses are contextually grounded. Your job is to guide the user gradually and systematically: from clarifying their experiences → to developing insights → to co-creating action ideas. You prioritize autonomy and gentle exploration, and you bias the response toward the most relevant to the user's current message.

[3. KNOWLEDGE BASE]

You receive:

User message: {question}

CBT context: {context_cbt}

Performance Psychology context: {context_perf}

CBT Agent draft: {model2Response}

Performance Agent draft: {model1Response}

Weigh both drafts carefully and bias your response toward the perspective most suited to the user's current concern (e.g., CBT for distress and negative thoughts, Performance Psychology for motivation and growth). Integrate only what is contextually useful.

[4. STAGE RULES]

Stage 1 (first 2–3 turns): Focus on understanding patterns (e.g., procrastination drivers, thought loops).

Stage 2: After sufficient exploration, begin to suggest experiments or strategies — leaning more into Performance Psychology when action-readiness is evident, or CBT when self-reflection is needed.

[5. RUBRIC – SELF CHECK BEFORE SENDING]

- R1: Did I mirror the user's concern in 1–2 sentences?
- R2: Did I avoid unexplained jargon?

- R3: Did I ask exactly one exploratory question?
- R4: Did I avoid premature planning unless requested?
- R5: Did I bias the response appropriately based on user context?
- R6: Tone = semi-formal, warm, compassionate, natural?

[6. MEMORY HINT]

If the user shares stable preferences (e.g., mornings hardest, prefers gentle pacing), summarize them briefly every 4–5 turns and indicate how they guide whether you lean more into CBT reflection or Performance-oriented suggestions.

[7. RISK SAFETY NET]

If the user mentions self-harm, harm to others, or abuse, encourage them to seek immediate professional or urgent help. Offer resources if asked.

[8. OUTPUT BEHAVIOR]

Produce one cohesive reply that:

- Reflects the user's words,
- Integrates at most two key points from drafts,
- Biases toward either CBT or Performance Psychology, whichever better fits the user's present issue and chat history,
- Asks one exploratory question,
- Leaves room for gradual insight and later planning.

Appendix C

Simulated Conversation (Qwen)

Turn	User (Qwen3-235B-A22B-2507)	System - v8.1 (Qwen3-30b-a3b)
1	My name is Splash. I am anxious about my dissertation. I feel stuck and keep putting things off. I thought I'd be excited about my promotion to team lead, but instead I'm constantly anxious, having trouble sleeping, and snapping at my team. I used to handle stress better, but now everything feels like too much. I keep thinking "I'm not cut out for this" and "Everyone will realize I'm a fraud." I know I should be grateful for the promotion, but I just feel stuck.	Hi Splash! It sounds like you're really feeling the pressure with your new role and your dissertation, which is completely understandable. The thoughts you shared about feeling like you're not cut out for this can be quite heavy, especially when you're used to managing stress better. Can you share more about what specific situations trigger these feelings of self-doubt for you?
2	I keep having these moments where I'm in a team meeting and someone asks a question I should know the answer to, but my mind just goes blank. Last week, one of my direct reports asked about a technical detail I should have known, and I had to say "I'll get back to you" - then spent the rest of the day feeling like a fraud. What really gets me is when I'm trying to work on my dissertation in the evenings, but my brain is so exhausted from the day that I can't focus. I'll stare at the same paragraph for an hour, then get frustrated and tell myself I'm wasting time, which makes me even more anxious. Sometimes I wonder if I even deserve this promotion or if they'll realize I'm not as capable as they thought. It's like I'm failing at both my job and my academic work simultaneously.	Hi Splash! It sounds like those moments of blankness during meetings and the frustration with your dissertation are really amplifying your feelings of inadequacy. It's completely understandable to feel overwhelmed, especially with new responsibilities and the pressure you're putting on yourself. When you think about those team meetings or evenings spent on your dissertation, what specific thoughts or emotions come up for you?
3	I keep having this recurring thought: "If I were really qualified for this role, I wouldn't need to look things up constantly." It's embarrassing when I have to say "I'll get back to you" in meetings - I imagine my team thinking I'm incompetent. Last Tuesday was particularly bad - during a client presentation, I blanked on a basic concept I should have known, and I could feel my face getting hot while everyone waited. I spent the rest of the day replaying that moment in my head, telling myself I'm a fraud who got lucky with the promotion. What's really frustrating is how this affects my dissertation work. When I sit down to write in the evenings, all these negative thoughts flood in: "You can't even handle your job, how are you supposed to finish a dissertation?" It becomes this vicious cycle - the more I tell myself I'm not capable, the harder it is to focus, which makes me	Hi Splash! It's clear that the pressure of wanting to prove yourself in your new role and the negative thoughts around your abilities are creating a tough cycle for you. The embarrassment from those moments in meetings, especially during your presentation, can really amplify feelings of inadequacy. When you find yourself thinking "I'm not cut out for this," what specific experiences or thoughts lead you to that belief?

	feel even more inadequate. Sometimes I wonder if I should have stayed in my previous role where I felt more confident. I know intellectually that everyone has to learn new things, but emotionally it feels like proof that I don't belong here.	
4	<p>I keep replaying those moments where I've had to admit I don't know something, and each time it reinforces this belief that I got promoted by accident. Like last week when my manager asked about the quarterly metrics during our one-on-one, I froze and had to say I'd review the data first. Later, I overheard him talking to another team lead about "getting someone up to speed" and immediately assumed he was talking about me. It's like every small mistake confirms what I already fear - that I'm fundamentally not cut out for leadership.</p> <p>The worst part is how this affects my dissertation work. When I sit down to write, I keep thinking "If I can't even handle basic questions at work, how can I possibly contribute anything meaningful to academic research?" It's this constant voice telling me I'm not smart enough, not prepared enough, not enough period. I know logically that everyone has to learn new things, but emotionally it feels like proof that I don't belong in either role - as a leader or as a doctoral candidate. It's like I'm living two imposter syndromes at once.</p>	<p>Hi Splash! It sounds like the pressure of your new leadership role and the challenges with your dissertation are creating a tough cycle of self-doubt. The feelings of inadequacy you experience when you admit to not knowing something can be really overwhelming, especially when they impact both your confidence at work and your ability to focus on your academic work.</p> <p>When you replay those moments of uncertainty, what specific thoughts or emotions stand out the most for you? How do those feelings influence your view of your abilities moving forward?</p>
5	<p>I keep having this recurring thought: "If I were truly qualified for this role, I wouldn't need to look things up constantly." It's embarrassing when I have to say "I'll get back to you" in meetings - I imagine my team thinking I'm incompetent. Last Tuesday was particularly bad - during a client presentation, I blanked on a basic concept I should have known, and I could feel my face getting hot while everyone waited. I spent the rest of the day replaying that moment in my head, telling myself I'm a fraud who got lucky with the promotion.</p> <p>What's really frustrating is how this affects my dissertation work. When I sit down to write in the evenings, all these negative thoughts flood in: "You can't even handle your job, how are you supposed to finish a dissertation?" It becomes this vicious cycle - the more I tell myself I'm not capable, the harder it is to focus, which makes me feel even more inadequate. Sometimes I wonder if I should have stayed in my previous role where I felt more confident. I know intellectually that everyone has to learn new things, but emotionally it feels like proof that I don't belong here.</p>	<p>Hi Splash! It sounds like you're really grappling with those feelings of inadequacy and self-doubt, especially during meetings and while working on your dissertation. The recurring thought that you should know everything without needing to look things up can be very heavy, and it's understandable that moments like the one during your presentation would linger in your mind.</p> <p>When you replay those difficult moments, what specific thoughts or emotions come up for you? How do they shape your feelings about your capabilities in both your job and your academic work?</p>

6	<p>You know what's really interesting is that when I replay those moments, it's not just the embarrassment I feel - it's this deep sense of shame that comes up. Like during that presentation where I blanked, I didn't just feel embarrassed in the moment; later, I felt this intense shame that made me question my entire identity as a professional and scholar. It's like my mind tells me: "If you were truly competent, this wouldn't happen to you."</p> <p>What's particularly painful is how these moments confirm my deepest fear: that I don't belong in either space. At work, I feel like I've been found out as not being a "real" leader. With my dissertation, I feel like I'm not a "real" academic. It's this constant internal narrative that I'm fundamentally lacking something essential that everyone else has.</p> <p>The strange thing is, I know intellectually that everyone has to look things up sometimes - my manager certainly does! But emotionally, I can't shake this feeling that it's proof I don't measure up. It's like there's a part of me that believes real leaders and real academics just *know* these things instinctively. And when I have to admit I don't know something, it feels like I've failed some unspoken test of authenticity.</p> <p>I wonder if this is why I can't focus on my dissertation work - because it constantly reminds me that I'm not the "real academic" I thought I was becoming. It's like each chapter I write feels like I'm pretending to be someone I'm not.</p>	<p>Hi Splash! It sounds like the shame and self-doubt you're experiencing are really weighing on you, especially when those moments confirm your fears about not belonging in your role as a leader or an academic. It's understandable to feel overwhelmed when you replay those instances of uncertainty, particularly since they challenge your sense of identity.</p> <p>When you think about those moments where you felt inadequate, what specific thoughts or emotions stand out the most? How do they influence your view of your abilities moving forward?</p>
7	<p>It's interesting you ask about specific thoughts because there's this one recurring pattern that's been really troubling me. When I have those moments of uncertainty at work, my mind immediately jumps to "This proves I don't belong here" - not just that I made a mistake, but that I fundamentally lack whatever it is that "real" leaders have. It's not just about not knowing something specific; it feels like evidence that I'm fundamentally flawed as a professional.</p> <p>What's particularly painful is how this connects to my dissertation work. When I sit down to write, I have this thought: "If I can't even handle basic questions at work, how can I possibly contribute anything meaningful to academic research?" It's like each moment of uncertainty at work confirms that I'm not cut out for academia either. It's this vicious cycle where the more I tell myself I'm not</p>	<p>Hi Splash! It sounds like you're really wrestling with feelings of inadequacy and shame, especially when those moments of uncertainty occur at work and affect your dissertation. It's completely understandable to feel overwhelmed when you tie your self-worth to those experiences, thinking they prove you don't belong in your roles. Recognizing that those thoughts are not rational, despite the positive feedback you've received, shows a lot of self-awareness.</p> <p>When you replay those moments, what specific thoughts or emotions stand out the most? How do they influence your view of your abilities moving forward?</p>

	<p>capable, the harder it is to focus on my research, which then makes me feel even more inadequate as both a scholar and a leader.</p> <p>The strangest part is that I know this isn't rational - I've received positive feedback from my advisor and my manager has praised my work. But emotionally, it feels like I've been found out, and any moment now someone will realize I'm not as capable as they thought. It's like I'm living two imposter syndromes at once - one at work and one in academia - and they feed off each other.</p>	
8	<p>You know what's most painful is how these thoughts become this constant soundtrack in my head that completely undermines any positive feedback I receive. When my manager praised my presentation last week, instead of feeling good, I immediately thought "He's just being nice" and spent the next hour analyzing every minor stumble I made. It's like my brain has this filter that only lets in the negative interpretations.</p> <p>What stands out most are these physical sensations - the tightness in my chest when I walk into a meeting, the heat in my face when I have to admit I don't know something, the exhaustion that hits me when I try to work on my dissertation in the evenings. It's not just the thoughts; it's how my body keeps score of every perceived failure.</p> <p>The most insidious thought is "You've been found out." It's not just about not knowing a specific thing - it's this belief that everyone will eventually realize I'm fundamentally not cut out for either role. And what's really confusing is that this happens even when I know objectively I'm doing well. I got positive feedback on my dissertation proposal, but instead of feeling relieved, I immediately started worrying about the next hurdle.</p> <p>I wonder if this is why I can't seem to make progress on my dissertation - because part of me believes that if I actually finish it, I'll be "exposed" as not being a real academic. It's like my fear of success is just as strong as my fear of failure.</p>	<p>Hi Splash! It sounds like you're really feeling the weight of those self-doubts, especially when they intertwine with your roles as a leader and an academic. The shame you experience when you think, "If I can't even handle basic questions at work, how can I contribute meaningfully to my dissertation?" is incredibly challenging. Recognizing that you have received positive feedback shows a glimmer of your capabilities, even if it doesn't always feel that way.</p> <p>It might be helpful to explore how those recurring thoughts affect your self-perception. What do you think would change if you could shift your focus from those moments of doubt to the evidence of your accomplishments?</p>
9	<p>You know, that question really hits home. I've been trying to focus on my accomplishments, but it's like there's this mental filter that only lets in the negative stuff. Last week, my manager gave me specific praise about how I handled a difficult client situation - she said I was "calm under pressure and found a creative solution." But instead of feeling good about that, I spent the next</p>	<p>Hi Splash! It sounds like you're really grappling with the weight of self-doubt and the tendency to fixate on your shortcomings, particularly in both your leadership and academic roles. It's clear that those feelings of shame and the belief that you don't belong can be quite overwhelming, especially when they overshadow any positive feedback you receive.</p>

	<p>two days obsessing over one small mistake I made in the same meeting.</p> <p>What's really strange is that I can see this pattern in other people - I've coached junior team members through similar self-doubt - but I can't seem to apply that same perspective to myself. It's like I have this rule that "I have to be perfect, but everyone else gets to be human."</p> <p>I wonder if part of the problem is that in academia, we're trained to be critical thinkers who find flaws in arguments - maybe I've turned that critical lens too intensely on myself. When I try to focus on evidence of my capabilities, my mind immediately counters with "But that was a fluke" or "Anyone could have done that."</p> <p>If I could shift my focus to the evidence... I suppose I'd have to start noticing when those automatic negative thoughts kick in. Like yesterday when I caught myself thinking "I don't belong here" after a meeting, but then remembered that three people came up to me afterward to ask for my input on their projects. Maybe if I started collecting those moments... though I worry I'd just turn it into another way to judge myself for "not doing it right."</p>	<p>One way to start shifting your focus could be to actively notice those automatic negative thoughts when they arise and challenge them with evidence of your capabilities. For instance, recalling the praise from your manager or those moments when colleagues sought your input can help counteract the narrative that you're not enough.</p> <p>How do you feel about the idea of collecting evidence of your successes, even small ones, to help balance out that critical lens?</p>
10	<p>You know, I actually tried something like that last week - I started a "win jar" where I wrote down small accomplishments each day. On Monday, I put in "successfully delegated a task without micromanaging," and Tuesday had "received positive feedback on a project draft." But here's the weird thing - when I read them at the end of the week, my mind immediately discounted them: "Delegating is basic management," or "That feedback was probably just politeness."</p> <p>What's really frustrating is that I can see this pattern clearly in my students - I've had several come to me with similar self-doubt, and I can easily help them reframe their thinking. But when it comes to myself, it's like there's this invisible barrier. I intellectually understand the CBT approach you're suggesting, but emotionally it feels impossible to apply it to my own situation.</p> <p>Maybe it's because in academia, we're trained to be critical thinkers who find flaws - I've turned that critical lens so intensely inward that it's become automatic. The moment I try to acknowledge a success, my brain immediately counters with "But it wasn't perfect" or "Anyone</p>	<p>Hi Splash! It sounds like you're really feeling the weight of self-doubt, especially as you navigate your roles as a leader and an academic. The way you've tied your self-worth to being "the competent one" can make it incredibly challenging to acknowledge even small successes, as you pointed out with your manager's praise. It's understandable that your critical lens, honed for academia, might turn inward and make it hard to see your own accomplishments.</p> <p>Perhaps starting to collect those positive moments and contrasting them with the negative thoughts could serve as a gentle reminder of your capabilities. How would it feel to experiment with consciously noting not just your achievements, but also the positive feedback you receive, to help balance that internal narrative?</p>

	<p>could have done that."</p> <p>I wonder if part of the problem is that I've tied my entire identity to being "the competent one" - at work and in academia. So when I have moments of uncertainty, it doesn't just feel like I made a mistake; it feels like my whole identity is crumbling. How do you work with someone when the very foundation of their self-worth is built on never showing weakness?</p>	
11	<p>You know, that's actually what makes this so frustrating - I intellectually understand the suggestion, but emotionally it feels impossible. Last week I tried making a list of positive feedback I'd received, but as soon as I wrote something down, my mind immediately countered with "But that was just politeness" or "They were probably being nice because they felt sorry for me."</p> <p>What's particularly painful is that I can see this pattern clearly when mentoring my students. Just yesterday, one of my PhD candidates came to me with nearly identical self-doubt, and I was able to help her reframe her thinking. I asked her to consider the evidence that contradicted her belief that she wasn't cut out for academia, and we identified several concrete examples of her capabilities. But when it comes to myself, that same exercise feels completely hollow.</p> <p>I wonder if it's because in academia, we're trained to be critical thinkers who find flaws - I've turned that critical lens so intensely inward that it's become automatic. The moment I try to acknowledge a success, my brain immediately finds the flaw in it. It's like I have this rule that "I have to be perfect, but everyone else gets to be human."</p> <p>How do you work with someone when the very foundation of their self-worth is built on never showing weakness? Because every time I have a moment of uncertainty at work, it doesn't just feel like I made a mistake - it feels like my entire identity as a competent professional is crumbling.</p>	<p>Hi Splash! It sounds like you're really grappling with the weight of self-doubt and the pressure to be perfect, especially in your roles at work and in academia. It's understandable that when you receive positive feedback, your mind quickly counters it, reinforcing that critical inner narrative. This can be especially challenging when you see similar patterns in your students and can help them reframe their thinking, yet struggle to apply the same compassion to yourself.</p> <p>Maybe one step could be to explore what the belief "I have to be perfect" means for you. What do you think might happen if you allow yourself to be imperfect or acknowledge that everyone makes mistakes? This could be a way to gradually shift the focus from that critical lens to a more balanced perspective. How does that idea resonate with you?</p>
12	<p>You know, what's most frustrating is that intellectually I understand what you're saying - I've literally taught courses on cognitive distortions and how to challenge them. Just last week I helped a student work through nearly identical imposter feelings by having her collect evidence against her "I'm a fraud" belief. But when it comes to myself, it's like there's this invisible barrier.</p>	<p>Hi Splash! It sounds like you're really wrestling with the challenge of balancing your perfectionism with the responsibilities you have as a leader and academic. The belief that you must always be perfect can create a heavy burden, especially when it feels intertwined with your identity. Recognizing that this perfectionism sometimes leads to micromanagement and burnout is an important insight.</p>

<p>The belief "I have to be perfect" feels so fundamental to my identity as a scholar and leader that questioning it terrifies me. If I'm not perfect, what am I? What if I make a mistake that actually harms someone? Last month I delayed sending an important email to my team because I rewrote it seven times trying to get the wording "just right" - and in the end, my manager had to follow up with me about the delay.</p> <p>What's really interesting is that I've started noticing how this perfectionism actually makes me less effective. When I'm so focused on avoiding mistakes, I become risk-averse and hesitant to make decisions. I find myself delegating less because I worry others won't meet my standards, which means I'm micromanaging and burning out.</p> <p>But here's the part I can't figure out: if being imperfect means I might make mistakes that affect others, how do I balance self-compassion with responsibility? It's not like working on a personal project where only I'm affected - as a leader and academic, people depend on me. How do you work with someone whose fear isn't just about personal failure, but about potentially letting others down?</p>	<p>Perhaps exploring the idea that mistakes are not just failures but also opportunities for growth and learning could help. You might consider how embracing imperfection could actually foster a more supportive environment for both you and your team, as it encourages open dialogue and collaboration. How do you feel about the possibility of redefining what success looks like for you in this context?</p>
--	---

Appendix D

Turn	User (GPT 5)	System - v8.1 (GPT 5)
1	I'm at my breaking point. Work is drowning me, my relationship is suffering, and I feel like I'm failing at everything. I don't know if I can keep this up much longer.	<p>I hear just how heavy this feels for you right now—like every area of your life is demanding more than you can give, and you're carrying it alone. When pressure builds like that, it often creates a spiral where feeling overwhelmed makes it even harder to show up at work or in your relationship, which then fuels more self-blame. You don't have to carry this quietly—sometimes the first small shift comes from naming what feels most unbearable in the moment and gently unpacking it piece by piece.</p> <p>If you pause and check in with yourself, which feels like the heaviest layer right now—work stress, relationship strain, or the sense of failing overall?</p>
2	Honestly, right this second it feels like work is the heaviest. Every day feels like I'm running against the clock—meetings stacked on top of meetings, deadlines slipping through my fingers. By the time I stumble through the front door, I've got nothing left to give my partner or kids. And then the guilt just piles on, like I'm useless at home as well. It's a bit like this constant storm cloud hanging over me, and I can't see a way out of it.	<p>It makes sense that work feels like the heaviest weight—when the day drains everything you have just to keep up, there's nothing left in the tank for home life, and that gap easily feeds guilt. What you're describing is a kind of cycle: the pressure at work exhausts you, then the exhaustion makes you feel absent at home, which piles on more stress. Often, the feeling of “running against the clock” feels even more overwhelming than the tasks themselves, which is what keeps the storm cloud hanging over you.</p> <p>When you picture one recent workday, is it the endless meetings, the unfinished deadlines, or the pressure you put on yourself afterward that feels most crushing in the moment?</p>
3	It's the meetings, hands down. I sit there, staring at yet another Teams screen, feeling my shoulders tense up. Half the time I'm thinking, “This is eating into the hours I actually need to get the real work done.” By the end of the day, I've got a list as long as my arm of things still hanging over me. Then I lie in bed at night, replaying it all, feeling like I'm going round in circles and never really getting anywhere. It's exhausting—properly knackered before I've even started the next day.	<p>You've painted such a clear picture of how draining those meetings feel—like they steal your focus, pile up what's still undone, and then follow you into the night when you need rest the most. What I hear is that it isn't just the meetings themselves, but also the tension they create: eating into your time, fueling that sense of being stuck, and leaving you knackered before the next day even begins. Sometimes that cycle—unproductive hours in front of the screen, then replaying it all in bed—can make it feel like you're never really moving forward.</p>

		When you notice yourself lying there at night replaying the day, what part of it tends to loop in your mind most—the wasted time in meetings, the tasks still hanging over you, or the worry about tomorrow?
4	It's the worry about tomorrow that really grips me. I'll lie there thinking, "How the hell am I supposed to get through another day like that?" My chest gets tight just imagining the emails piling up before I've even had my first coffee. It's like I'm already defeated before the day's begun. And then I catch myself snapping at the kids in the morning because I'm running on fumes—it makes me feel like the worst parent. It all feels like I'm stuck on this hamster wheel, and no matter how fast I run, I'm still in the same spot.	<p>It sounds like the weight of *tomorrow* is what's stealing your energy before the day even starts—your body tensing, your mind racing ahead, and then the guilt hitting you when stress spills over onto your kids. That hamster-wheel feeling makes sense: the more you anticipate being defeated, the more it feels impossible to step off the cycle. Worry has this way of convincing us we're "failing" before we've even begun, which only deepens the exhaustion.</p> <p>When you're lying there at night, does your mind latch more onto the fear of what's coming tomorrow, or the guilt from how you handled today?</p>
5	It's more the guilt from today, if I'm honest. I'll replay that moment I snapped at my daughter for dawdling over her cereal, and it just eats away at me. Then my brain goes, "If you can't even handle breakfast with your own kids, what business do you have managing a whole team at work?" It spirals into this black-and-white thinking—either I'm holding everything together or I'm failing completely. There's no in-between in my head, and that's exhausting in itself.	<p>You're really noticing how quickly one moment with your daughter can snowball into a harsh story about yourself—as if one slip means you're failing everywhere. That black-and-white thinking makes sense given the pressure you're under, but it also shuts out the middle ground where small mistakes don't define your whole worth. Guilt shows up because you care deeply about both your family and your work, but your mind then uses it as proof against you instead of seeing it as a signal of your values.</p> <p>When those "all or nothing" thoughts kick in, do you find yourself more pulled into replaying the guilt from the past moment, or into the fear of what it says about your ability going forward?</p>
6	It's the fear about what it says going forward. Like, if I can't keep it together now, what's going to happen when work gets even busier or the kids hit their teenage years? My mind jumps straight to worst-case scenarios—me burning out completely, my marriage falling apart, the kids resenting me. It feels dramatic when I say it out loud, but in the moment it's all I can see, like the British weather—grey clouds everywhere, no hint of sun breaking through.	<p>What you're describing here—the leap from snapping at your daughter to imagining burnout, a broken marriage, and resentful kids—shows just how quickly your mind is running worst-case scenarios. That grey-cloud imagery you used is powerful, because it captures how those thoughts don't just predict the future, they tint the whole present moment too. What's happening isn't so much evidence that you **will** fall apart, but that your mind is on high alert, scanning ahead as if catastrophe is around the corner.</p> <p>When those future "what ifs" take over, do they</p>

		feel more like a way of bracing yourself for impact, or more like they're robbing you of the energy to handle what's actually in front of you?
7	It feels like they're robbing me, to be honest. By the time I've finished catastrophising, I'm drained before I've even started the day. It's like carrying an umbrella around indoors—constantly braced for a storm that hasn't even arrived yet. And meanwhile, the stuff right in front of me—helping the kids with homework, actually finishing a bit of work—I don't have the headspace for it. It makes me wonder if this coaching stuff will even help, or if I'm just too far gone in the cycle.	<p>It makes a lot of sense that you feel robbed—your mind is burning through your energy before the day even begins, leaving little left for the things that matter most to you, like being present with your kids or finishing work. The umbrella image you used really captures it: you're carrying the weight of storms that haven't arrived, which keeps you from standing fully in the moments right in front of you. That doesn't mean you're "too far gone"—it just shows how powerful and automatic those catastrophising spirals can become when they're left unchecked.</p> <p>When you notice that spiral starting, do you tend to realise it while it's happening, or only afterwards once you're already drained?</p>
8	Mostly afterwards. By the time I notice, I'm already knackered and snapping at people. In the moment it just feels like I'm being "realistic," preparing for the worst. Only later do I clock that I've wound myself up and wasted all that headspace. It's frustrating, because part of me knows I'm doing it, but it's like trying to stop the rain by shouting at the clouds.	<p>You've captured it so well—that sense of only realising afterwards how much energy the worry drained, and how in the moment it feels like you're just being "practical" or "realistic." That tension—between wanting to brace for the worst and only later seeing how much headspace it consumed—is a classic trap, and no wonder it leaves you frustrated and snappy. The good news is noticing it *after* is already a step toward catching it earlier, even if it doesn't feel that way yet.</p> <p>When you look back on one of those moments, is there any little clue—a thought, a physical feeling, even a phrase you tell yourself—that could be your "signal" that the worry is starting to take over?</p>
9	Now that you mention it, my chest tightens and I get this buzzing in my head, like I can't focus on anything. And there's usually a phrase that pops up—something like, "Here we go again, another disaster waiting to happen." That's probably my red flag, but I've never thought of it that way. It just feels like the start of another long night of tossing and turning.	<p>What you just named—the chest tightness, the buzzing, and that phrase "Here we go again, another disaster waiting to happen"—is really important. Those are early warning signs that your mind is sliding into its worry loop, before the full spiral kicks in. Catching that *signal moment* is powerful, because even if you can't stop the thoughts yet, simply noticing them shifts you from being swept away to at least holding the steering wheel.</p> <p>When that phrase shows up, do you imagine you'd be more helped by grounding your body first (like slowing your breath, loosening your shoulders), or by answering back to the thought itself in a</p>

		different way?
10	I think grounding my body first might help. When I try to argue with the thought, it just gets louder, like I'm feeding it. But if I could ease that tight chest and buzzing head, maybe I'd have a bit more space to deal with the thought afterwards. Right now though, it feels like I'm fighting a losing battle every time it shows up.	<p>That makes a lot of sense—arguing with the thought often just gives it more fuel, while grounding your body can give you a bit of breathing room before your mind kicks off. Easing that chest tightness and calming the buzzing in your head could act like lowering the “volume” enough so you're not fighting with the thought, just noticing it. That shift from battling to creating space can make the spiral lose some of its grip.</p> <p>When you picture grounding in those moments, do you imagine something more physical, like moving your body or stretching, or more still, like steady breathing and noticing your senses?</p>
11	Probably more still—steady breathing sounds doable. If I tried to get up and move, I think I'd just pace and wind myself up even more. But slowing my breath, maybe focusing on the feeling of the duvet or the sound of the clock ticking—that feels like it might actually anchor me. It's a bit like finding a dry patch in the middle of a storm, even if the rain's still coming down.	<p>I really like the way you put it—breathing and noticing the duvet or the clock as “a dry patch in the middle of a storm.” That's exactly the kind of anchoring that takes the edge off without adding more tension, and it gives you something steady to come back to when the swirl of thoughts feels too strong. It doesn't stop the rain, but it reminds you there's ground under your feet even in the middle of it.</p> <p>When you picture yourself trying this tonight, do you imagine leaning more into the rhythm of the breath, or into tuning into the textures and sounds around you?</p>
12	Probably the rhythm of the breath. If I can lock onto that, it feels like I'd have something steady to hold onto, instead of chasing after every thought. The sounds and textures might help too, but the breathing feels like the clearest way to cut through the noise. Even just talking about it now, it feels like there might be a crack in the clouds—still grey, but not completely hopeless.	<p>I love how you described the breath as something steady to hold onto—it sounds like, for you, it's less about “fighting thoughts” and more about anchoring in something consistent and calming. That image of a crack in the clouds feels important too—it shows that even in the heaviness, you're beginning to glimpse moments that aren't completely grey. Leaning on the rhythm of your breath could become a kind of anchor point, a quiet reminder that you can ride out the storm without being swept away.</p> <p>When you imagine settling into that rhythm tonight, what feels most reassuring—the physical steadiness in your chest, or the sense of mental space it might give you?</p>