Received: 30 January 2023 | Revised: 11 July 2023 | Accepted: 6 September 2023

Cognitive Computation and Systems

DOI: 10.1049/ccs2.12087

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

An AI powered system to enhance self-reflection practice in coaching

Mahdi Jelodari[1] | Mohammad Hossein Amirhosseini[2] | Andrea Giraldez-Hayes[3]

[1]Keptika Ltd, London, UK

[2]Department of Computer Science and Digital Technologies, School of Architecture, Computing and Engineering, University of East London, London, UK

[3]Wellbeing and Psychological Services Clinic, Department of Professional Psychology, School of Psychology, London, UK

Correspondence

Mohammad Hossein Amirhosseini.
Email: M.H.Amirhosseini@uel.ac.uk

Abstract

Self-reflection practice in coaching can help with time management by promoting self-awareness. Through this process, a coach can identify habits, tendencies and behaviours that may be causing distraction or make them less productive. This insight can be used to make changes in behaviour and establish new habits that promote effective use of time. This can also help the coach to prioritise goals and create a clear roadmap. An AI powered system has been proposed that maps the conversion onto topics and relations that could help the coach with note-taking and progress identification throughout the session. This system enables the coach to actively self-reflect on time management and make sure the conversation follows the target framework. This will help the coach to better understand the goal setting, breakthrough moment, and client accountability. The proposed end-to-end system is capable of identifying coaching segments (Goal, Option, Reality, and Way forward) across a session with 85% accuracy. Experimental evaluation has also been conducted on the coaching dataset which includes over 1k one-to-one English coaching sessions. In regards to the novelty, there are no datasets of such nor study of this kind to enable self-reflection actively and evaluate in-session performance of the coach.

KEYWORDS

coaching, conversation map, intent classification, knowledge graph, natural language processing, personal development, transformer models, visual summary

1 | INTRODUCTION

Coaching, as a tool for helping personal development and profession, has made a significant impact on the performance of individuals and organisations [1]. It can be considered as a mechanism to support an individual's learning, self-awareness, behavioural change, wellness, growth, and career management [2]. It can also be considered as a useful leadership tool in assisting project managers, as it can help them to tackle the challenges such as staff motivation, culture development, project uncertainty, and managing employees effectively [3].

In fact, coaching is a collaborative process that empowers the clients to tap into their potential and achieve their goals. By considering the following points and continuously refining coaching approach, coaches can provide a supportive and transformative experience for their clients:

- Establish a coaching relationship
- Clarify goals and objectives
- Active listening and powerful questioning
- Cultivate self-awareness
- Goal setting and action planning
- Support and accountability
- Feedback and reflection
- Continuous learning and development

In regards to how coaching should be delivered effectively, there are different coaching models which have been used by coaches and they are discussed in the following section. In fact, a coaching model can be considered as a framework for 'how to coach'.

1.1 | Coaching models

There are various effective coaching models which can facilitate the coaching process through providing a logical structure for coaching sessions. A coaching model can help to create a framework for guiding a person through five different steps including (1) defining a desired goal, (2) understanding the current situation and where they are, (3) exploring options to see where they are headed, (4) determining possible obstacles, and (5) establishing a plan of action [4]. Coaching models are often known with their associated acronyms such as GROW, CLEAR, OSKAR, ACHIEVE, PRACTICE, OUTCOMES. Between all coaching models, the GROW model, which stands for Goal, Reality, Options, and Will, is generally accepted as the standard method for coaching [5] and it is one of the most popular and well used models [6]. Palmer and Whybrow conducted a survey in 2006 to find out what percentage of coaching psychologists use the GROW model. The results showed that this model was used by 53.2% of the coaching psychologists surveyed [7]. The GROW model is very flexible as you can jump forwards and backwards through its four elements during a session [8]. Figure 1 shows how the GROW model works.

Many other efficient coaching models have been inspired by the GROW model and their structure is very similar to the structure of this model [8]. One of them is the CLEAR coaching model which has 5 phases including Contracting, Listening, Exploring, Action, and Review [9]. The main difference between the GROW model and the CLEAR model is that the CLEAR model emphasises on the review cycle, while the GROW model is very much an 'up front' model which focuses on defining a goal and agreeing on how it will be achieved.

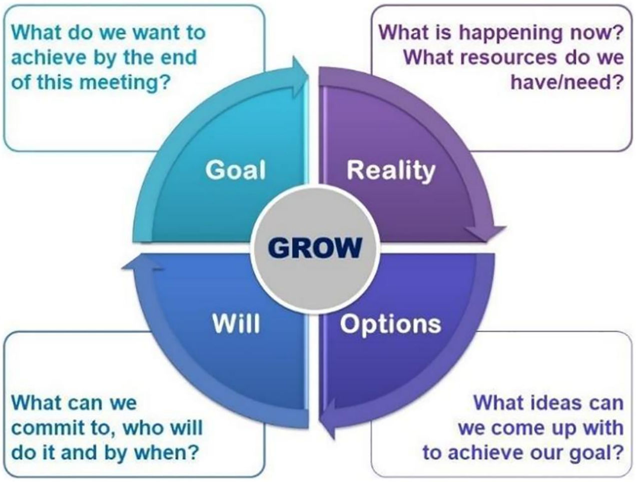

What do we want to achieve by the end of this meeting?
What is happening now? What resources do we have/need?
Goal
Reality
GROW
Will
Options
What can we commit to, who will do it and by when?
What ideas can we come up with to achieve our goal?

FIGURE 1 The GROW model coaching framework [43].

Moreover, some of the coaching models are solution-focused which present variations on the GROW model. A good example is the OSKAR model which has five elements including Outcome, Scaling, Know-how, Affirm and action; and Review [10]. In comparison with the GROW model, this model focuses on solutions and finding what works rather than analysing problems and defining a goal to be achieved.

Compared to the previous models, there are also some models which present more detailed steps. For instance, the ACHIEVE model suggests seven steps for coaching [11]. These steps include (1) Assess current situation, (2) Creative brainstorming of alternatives to current situation, (3) Hone goals, (4) Initiate options, (5) Evaluate options, (6) Valid action programme design, and (7) Encourage momentum [12]. The ACHIEVE model builds on the GROW model by adding more collaboration and conversation during the coaching process to create a more flexible coaching experience. Another example is the PRACTICE model which again suggests seven detailed steps including (1) Problem identification, (2) Realistic, relevant goals developed, (3) Alternative solutions generated, (4) Consideration of consequences, (5) Target most feasible solutions, (6) Implementation of Chosen solutions, and (7) Evaluation [13]. This model is very collaborative in nature. Compared to the GROW model which has wider applications, this model is particularly useful within a business setting where leaders work with team members to identify their unique strengths, so that they can use them to overcome challenges. This model is less effective when problems are outside the team member's control.

Furthermore, the OUTCOMES model is even more complex and it has eight detailed steps including (1) Objectives for the session, (2) Understanding why the coachee wants to reach the objective, (3) Take stock, (4) Clarify, (5) Option generation, (6) Motivate to action, (7) Enthuse and encourage, and (8) Support [14]. This model is more efficient for the intake phase of coaching as it helps the coach to immediately get a clear picture of the client's current and desired situation, including options, obstacles and goals.

Based on this information about the structure of coaching models, it would be very important and crucial for the coach to be able to manage different steps and monitor the progress during a session to make sure that coaching is going forward in the right direction.

1.2 | Artificial intelligence and coaching

Artificial intelligence has rapidly transformed business and society in recent years and has brought innovations to all aspects of human life [15]. Coaching has not been exempt and the use of artificial intelligence in coaching has received a lot of attention among researchers [16]. In fact, self-help technologies which can be easily accessed through smart phones to employ coaching methods, have been challenging our understanding of the nature of coaching [17]. Research findings in this field show that during a conversational coaching process, an artificial agent can deliver positive outcomes for users [17] and it seems to be capable of effectively guiding clients through different steps in the coaching process. However, there have been some challenges and difficulties in identification of the client's problem and delivering individual feedback [15]. Researchers are confronting intrinsic complexity when they try to replicate the human coaches' skills [17].

It has been proposed that an intelligent coaching system should be able to act as an artificial entity which is able to observe, reason about, learn from, and predict an individual's behaviours [18]. This should be done over time and in context, through engaging with the user proactively in a collaborative conversation to assist planning and to encourage effective goal striving [18].

It is important to understand that artificial general intelligence (strong AI) is different from artificial narrow intelligence (narrow AI) and they have different applications. A machine using strong AI can exhibit consciousness, sentience, and the learning ability beyond what was initially intended by its developers and its intelligence can be applied in more than one specific area [19]. On the other side, narrow AI focuses on narrow and specific tasks [19]. Expert systems are a form of narrow AI and they are able to provide solutions to specific problems in a narrow area [20].

In order to design a system which is able to assist a human coach and improve productivity, strong AI would be needed. However, this field of research is in its infancy and we may not see credible examples of Strong AI in our daily life in the foreseeable future [21]. As a result, that would be highly unlikely for an intelligent system to convincingly perform all the functions that a human coach can perform any time soon and strong AI doesn't seem to be a possibility for coaching now [22]. Despite this, narrow AI in the form of expert systems can provide valuable and considerable options to facilitate the coaching process.

1.3 | Productivity tools for coach

As a form of narrow AI and expert system, conversational agents attempt to mimic human experts in a specific narrow area of expertise [23] and they can respond to users by deciding on the appropriate response given a user input [24]. Conversational agents usually interact with users via natural language in different formats such as text, voice, or both [25]. They typically receive questions and associate them with a knowledge base to offer a response [26]. Conversational agents are usually driven by scripted rules or AI technologies such as machine learning, deep learning and Natural Language Processing, Generation and Understanding [27]. Recently, the application of chatbots which are a type of conversational agents has become popular in the services industry where they are used to assist with customer queries, advice and fulfilment of orders [28].

Chatbots have also been used widely in the relevant fields to coaching such as health, well-being and therapy [23] and research has been conducted on the efficacy of chatbots which have been used to assist people with depression, promotion of physical activity, eating habits, and neurological disorders [29, 30]. It has been claimed that chatbots and AI coaching can provide a wide range of strategies and techniques to assist individuals achieve their goals for self-improvement [18, 31] and can play an important role in supporting behavioural change [32]. The possibility of anonymous interaction, especially in the context of sensitive information, is another advantage of using chatbots in coaching [33] as the client may feel less shame and be more interested in disclosing information [34].

Among all challenges for developing a realistic chatbot and an efficient AI coach, it seems that difficulty to maintain the ongoing context of a conversation is one of the most serious challenges [35]. Pattern-matching techniques are commonly used to map input to output in a conversation. However, this approach can rarely lead to purposeful and satisfying conversations [36]. Moreover, it has been realised that most of the tools for productivity in coaching are related to managing the time and tasks, and the main focus is on clients rather than the coach. According to the literature review conducted in this research, there has been no tools that allow the coach to self-reflect. Furthermore, some of the existing tools help the coach with the note taking process but a simple note taking without any analysis cannot be that much useful and efficient.

As a result and in order to fill the knowledge gap, rather than aiming the client, this study focuses on developing an AI powered system considering the GROW coaching model, to facilitate the coaching process and assist the coach through providing a conversation map and visual summary in a real-time conversation between the coach and coachee. This will help the coach with summarisation, note taking and analysis of the conversation to identify the topics/subjects that have been discussed during the session.

In the following section the methodology in terms of dataset and algorithm will be discussed. Then the results will be presented and discussed in Section 3. Following this section, experimental evaluation will be presented and discussed in Section 4, to show the efficiency of the system. Novelty in this research will be discussed in Section 5 and finally, Section 6 concludes the paper.

2 | METHODOLOGY

This section comprises 2 main parts, (1) dataset and (2) algorithm. The dataset section describes the processes of data collection, labelling, and expansion. The algorithm section describes transcription and diarisation, coach question anchoring, and real time conversation mapping.

2.1 | Dataset

Our dataset consists of +1k hours of life and executive coaching sessions collected from various sources including publicly available coaching sessions from YouTube and our coaching platform [37]. The preferred data type for this research was audio + video with HD quality. However, we did not limit ourselves to multimedia data and considered processing audio only data as well. For audio only data, the conversation map will be provided in a video format with visual annotations. Figure 2 shows the distribution of the data used in this dataset.

Moreover, Table 1 shows the YouTube Data API attributes described in terms of snippet, statistics, and content details. Our dataset follows same schema with two enhancements containing (1) 'transcription' which includes transcription in form of questions and answers labelled separately with GROW labels and (2) 'tags' which includes tagging the video class and topic.

2.1.1 | Data labelling

The sessions were pre-processed to categorise the topics discussed during the session and focus on the main topic. Overall, approximately 5% of the data were cropped to remove chit-chat and catch-up from previous sessions. We were considerate when applying this process as in few cases catch-up from the previous sessions was important for goal setting in the beginning of the session.

For topic identification across the sessions, thanks to the emerging of Transformers [38], technologies such as BERT and sentence BERT are being preferred over topic modelling techniques (e.g. Doc2Vec, LDA, etc.) by NLP researchers due to their impressive performance in both classification and detection. Next, the coach's questions were extracted and

TABLE 1 YouTube Data API attributes.

Attribute	Part	Description
Title	Snippet	The title of the video.
Description	Snippet	A description of the video.
publishedAt	Snippet	The date and time that the video was published.
channelId	Snippet	The id of the channel that uploaded the video.
channelTitle	Snippet	The title of the channel that uploaded the video.
Tags	Snippet	A list of keywords associated with the video.
viewCount	Statistics	The number of times the video has been viewed.
likeCount	Statistics	The number of likes the video has received.
dislikeCount	Statistics	The number of dislikes the video has received.
favoriteCount	Statistics	The number of times the video has been marked as a favourite.
commentCount	Statistics	The number of comments on the video.
Duration	contentDetails	The duration of the video in ISO 8601 format.
Dimension	contentDetails	The dimension of the video. Valid values: "2 days", "3 days".
Definition	contentDetails	The definition of the video. Valid values: "Standard", "high".

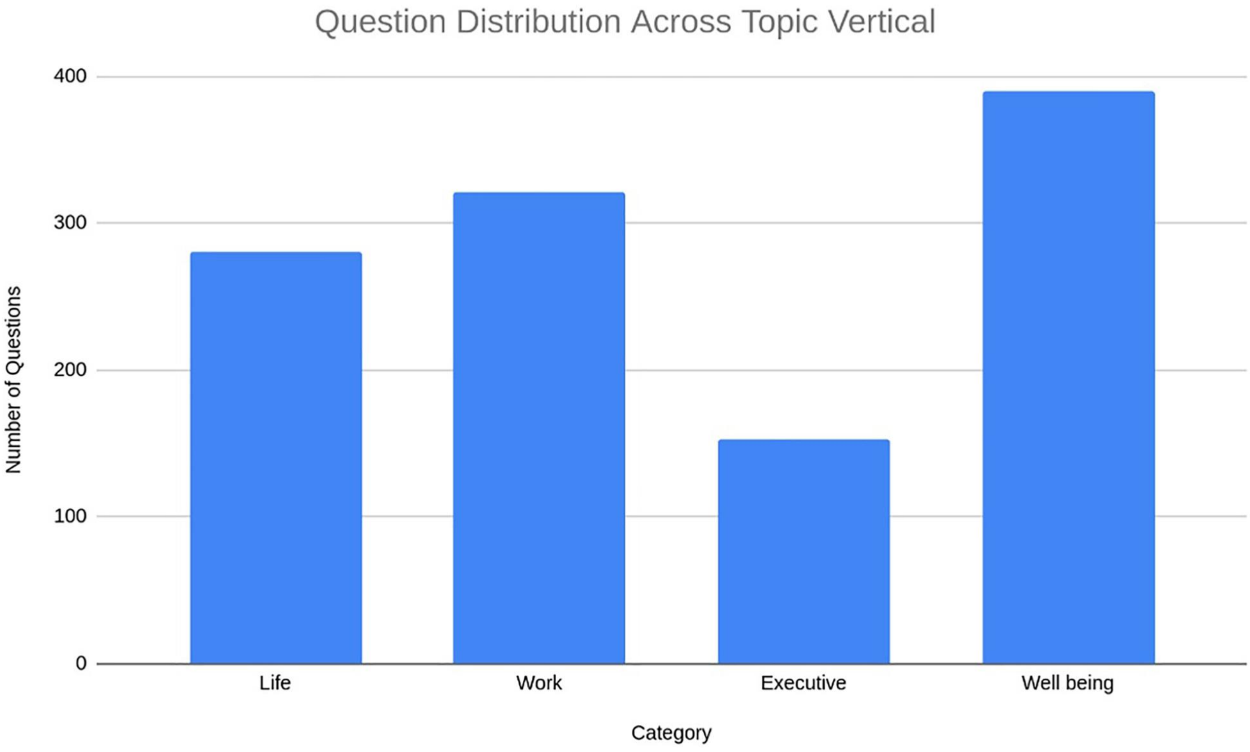


FIGURE 2 Distribution of data.

labelled for intention classification based on different categories in the GROW model including (1) goal exploration, (2) reality exploration, (3) option exploration, and (4) forward exploration. Table 2 shows examples of coach's questions provided with labels. These examples were extracted from our dataset.

2.1.2 | Data expansion

To enrich our questions dataset, intent preserving models [39] were used to retain the same semantic intent, but develop different surface forms as shown below. The intent preserving models normally encode form and meaning of the question as latent variables which allows changing the question surface without affecting its meaning. This technique allows us to apply variation in word choice, syntactic structure, and question types while preserving meaning of the coach's question.

By considering this technique our questions dataset was tripled per category, giving the classifier larger space to explore and learn from. Table 3 shows examples of paraphrased questions using intent preserving encoder-decoder.

After applying paraphrasing and class balancing on the extracted questions, Figure 3 shows the distribution of the >3k questions across the dataset based on GROW categories.

TABLE 2 Example of every question type provided with labels, extracted from our coaching dataset.

Coach's question	Label
1. "What's been working well for you since the last session?"	Goal exploration (G)
2. "What hasn't been working for you lately?"	
3. "What do you need most from me today?	
4. "I'm wondering what you would love to have happen by the end of this session?"	
5. "What specifically would you like to get out of the next 30/45/60 min?"	
6. "What's the outcome you're looking for from our session today?"	
1. "What has stopped you from doing more/moving towards your goal?"	Reality exploration (R)
2. "In a nutshell, who or what's got in the way?"	
3. "What would happen if you did nothing?"	
1. "Let's imagine you're really excited about this. What would you do?"	Option exploration (O)
2. "If you were at your best, what would you do right now?"	
3. "What could you do if you knew you couldn't fail?"	
1. "Which actions WILL you do?"	Forward exploration (W)
2. "How would you like to be held accountable for these actions?"	
3. "How might you commit to that?"	

Next section explains how the data curated in this section is used to train transformer based models to classify intent across coaching sessions.

2.2 | Algorithm

2.2.1 | Transcription and diarisation

The first and foremost step for processing the conversation data is to convert audio signals to analysable text. For this purpose, fully customisable open-source conversational AI models including automatic speech recognition and NLP models [40] were fine-tuned and leveraged to get speech to text even in presence of accent and noise.

Next, the transcription was diarised using available pretrained models [41, 42] with lowest diarisation error rate on our dataset (approx. %1.7) to be able to recognise participants in the session. Further enhancements were applied at this step to make sure any interrupts are captured during the session. Enhancements to the transcription and diarisation models are out of the scope of this work.

Due to the real-time nature of this work, models were optimised and quantised using TensortRT SDK [ref] provided by NVIDIA to be able to achieve millisecond latency when performing analysis.

2.2.2 | Coach question anchoring

There are various ways for semantically understanding and analysing flow of a coaching or a therapy session from word level to phrase or sentence level. In this work, we decided to anchor a session around the coach's questions and the client's answers to effectively identify the turns. A coach or therapist uses good questions as a door-opener phrase that invites clients' full disclosure and gradually and carefully challenges their assumptions, beliefs, and perspectives that may be contrary to their needs and goals. Question anchoring allows the system to segmentise the entire session and recognise the progress of the conversation from the context that is being discussed by the participants.

To do this, first, coach's questions are identified across the session. This requires the transcriber and dirization models to be fine-tuned properly (as discussed in Section 2.2.1) to

TABLE 3 Example of every question paraphrased using intent preserving encoder-decoder.

Paraphrased questions
How might you commit to that?
What would be your level of commitment?
How much of your time would you dedicate to it?
How would you commit to that?
Your level of commitment?

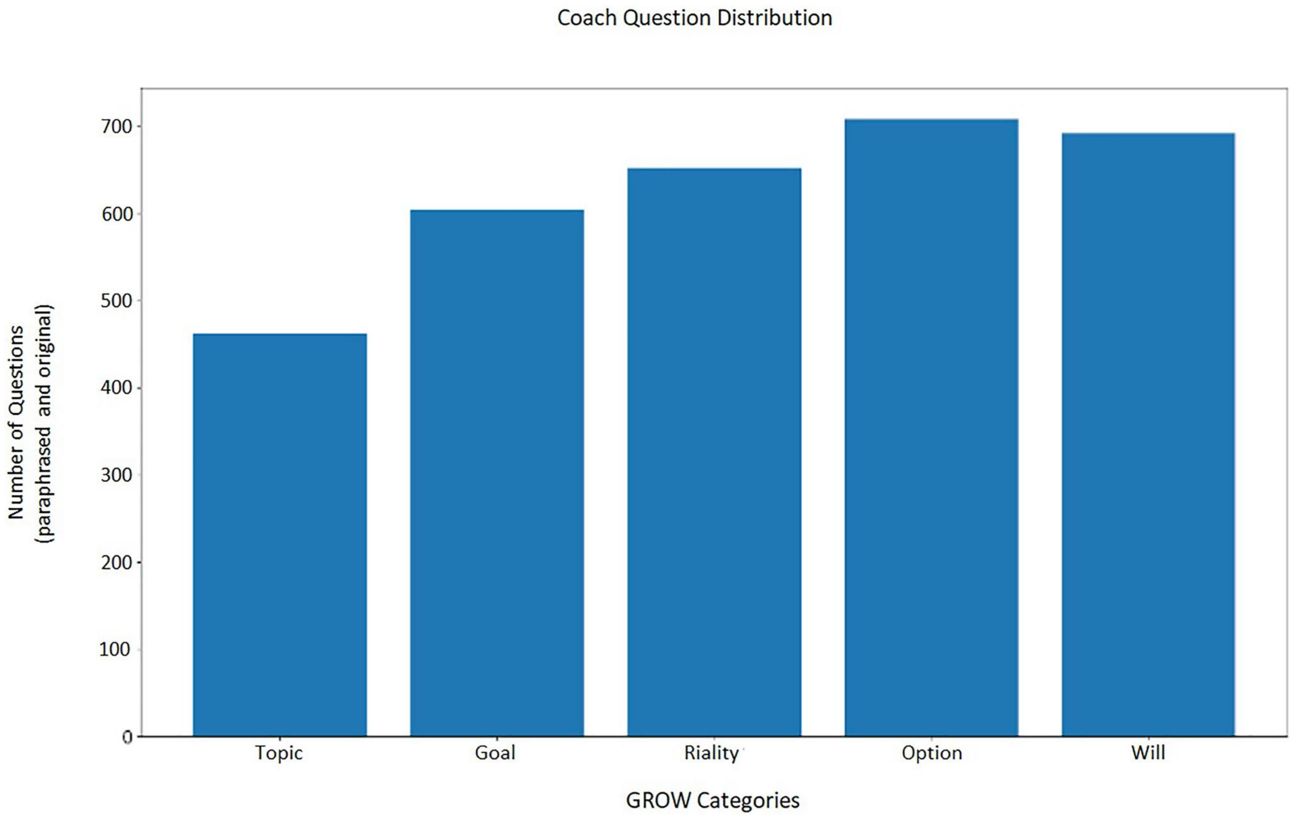


FIGURE 3 Distribution of the >3 k questions across the dataset based on GROW categories.

recognise coach's questions from client's questions which are normally shorter and are of a clarifying nature (e.g. "could you repeat that again?" and "does it make sense?").

Next, the coach's questions were analysed based on their intent. Due to the sophisticated nature of the questions, including them being compound, an advanced model was required for intent classification. An attention-based model, BERT [43], was exploited for this purpose.

BERT and other Transformer encoder architectures have been wildly successful on a variety of tasks in NLP. They compute vector-space representations of natural language that are suitable for use in deep learning models. The BERT family of models uses the Transformer encoder architecture to process each token of input text in the full context of all tokens before and after, hence the name: Bidirectional Encoder Representations from Transformers. BERT models are usually pre-trained on a large corpus of text, then fine-tuned for specific tasks.

In this regard, "Bert-base-uncased" pre-trained model was used for both tokenising the input text and creating vector embeddings and also as a base model for our fine tuning step. The original "Bert-base-uncased" model was trained on unlabelled English Wikipedia and BookCorpus with overall 3,300M words. It has 12 layers of Transformer encoder, 12 attention heads, 768 hidden size, and 110M parameters. For comparison, the largest BERT model consists of 24 attention layers. GPT-2 [44] has 12 attention layers and GPT-3 [45] has 96 attention layers. Given the API latency of the large language models such as GPT3 from OpenAI we decided to look into a smaller expert model whose response time is shorter and can be deployed locally.

The following summaries why we chose BERT architecture for this task:

1. Pre-training Objective: BERT is trained using a masked language modelling objective, where it learns to predict missing words within a sentence. This objective helps BERT to gain a deeper understanding of the context and meaning of individual words within a sentence, making it well-suited for tasks requiring fine-grained semantic understanding, such as text classification. In contrast, GPT-2 is trained using an autoregressive language modelling objective, focusing on generating coherent and contextually relevant text, which makes it more suitable for tasks like text generation.
2. Bidirectional Context: BERT incorporates bidirectional context by leveraging both left and right contexts during training. This allows BERT to capture a more comprehensive understanding of the relationship between words within a sentence, enabling it to grasp dependencies and nuances that exist across the entire input sequence. Such bidirectional modelling is crucial for accurate text classification, where the context and interactions between different words are essential for making correct predictions. GPT-2, on the other hand, only considers the left context during training, which might limit its ability to capture certain dependencies effectively.

3. Task-Specific Fine-Tuning: BERT's architecture lends itself well to fine-tuning on specific tasks, including text classification. After pre-training on a large corpus of text, BERT can be further trained on task-specific labelled data, allowing it to adapt to the specific classification objective. This fine-tuning process enables BERT to learn task-specific patterns and improve performance on text classification tasks. GPT-2, while also capable of fine-tuning, may not be as optimised for text classification tasks out of the box.
4. Sentence-Level Understanding: BERT excels in understanding the semantics and relationships between sentences, which is particularly useful for tasks like sentiment analysis or document classification, where the context and connections between multiple sentences are important. By leveraging attention mechanisms and contextual embeddings, BERT can capture the nuances of different sentences, leading to more accurate text classification results. GPT-2, being focused on generating text, might not have the same level of sentence-level understanding [46].
5. Model Size and Real-time Performance: BERT typically has a smaller model size compared to GPT-2, making it more suitable for real-time applications, especially when computational resources or memory constraints are a concern. The smaller model size of BERT allows for faster inference times, which is crucial for scenarios where real-time predictions are required. GPT-2, with its larger model size, may be more resource-intensive and might pose challenges for efficient deployment in real-time systems.

The structure of our dataset for the training phase was provided in CSV format:

<ID>,<Question>,<Intent>

With 5 different intents (TOPIC, GOAL, REALITY, OPTION, and WILL/FORWARD) based on the extended GROW model discussed in section 1.1.

The maximum length of questions accepted by the model is 512 tokens. By enabling Truncate, the model is able to truncate the input tokens and process longer questions when required.

The output of the model is a vector of size 768 for every input token. A linear classifier receives this vector and proposes the probability of the input token belonging to one the five target classes based on which loss is computed per batch and back propagated in the training process. In this work, categorical cross entropy was used as our loss function and AdamW from Pytorch framework was used as our optimiser with learning rate 1-e5.

2.2.3 | Real-time conversation mapping

For the real-time conversation mapping various optimization techniques were used (e.g., quantisation) to reduce model footprint so that 30 milliseconds browser-to-browser and server-to-browser is achievable. For web-to-web communications in real-time WebRTC protocol and the provided open-source API was used [47].

To be able to annotate frames with questions and answers for the coach in real-time a GPU-based agent was implemented to run the models in real-time and annotate the frames in sequence visible to the coach. For better performance and enhanced experience two GPUs (AWS V100 T - 16 GB RAM) one for transcription/diarisation and one for classification and mapping were used.

Overall, visualisation of our proposed conversation mapping system for coaches is adoptable for therapists, mentors, and teachers in the future.

3 | RESULTS AND DISCUSSION

3.1 | Evaluating the accuracy of the proposed conversation mapping system

As explained in Section 2.2, a pre-trained BERT model "*Bert-base-uncased*" was selected for the task of coach question intent classification. The dataset described in Section 2.1 was used for fine-tuning the classifier. Using the parameters explained in Section 2.2, the model initially achieved approx. 70% accuracy on the validation set which is 20% of the entire data. Figure 4 shows the training performance of the model on this dataset.

Regarding the validation loss increasing during training, cross-entropy loss for classification was selected. In this function bad predictions are penalised much more strongly than good predictions are rewarded. For instance, no matter if several sentences are classified correctly, one misclassified sentence significantly increases the loss.

In binary classification, where the number of classes M equals 2, cross entropy can be calculated as:

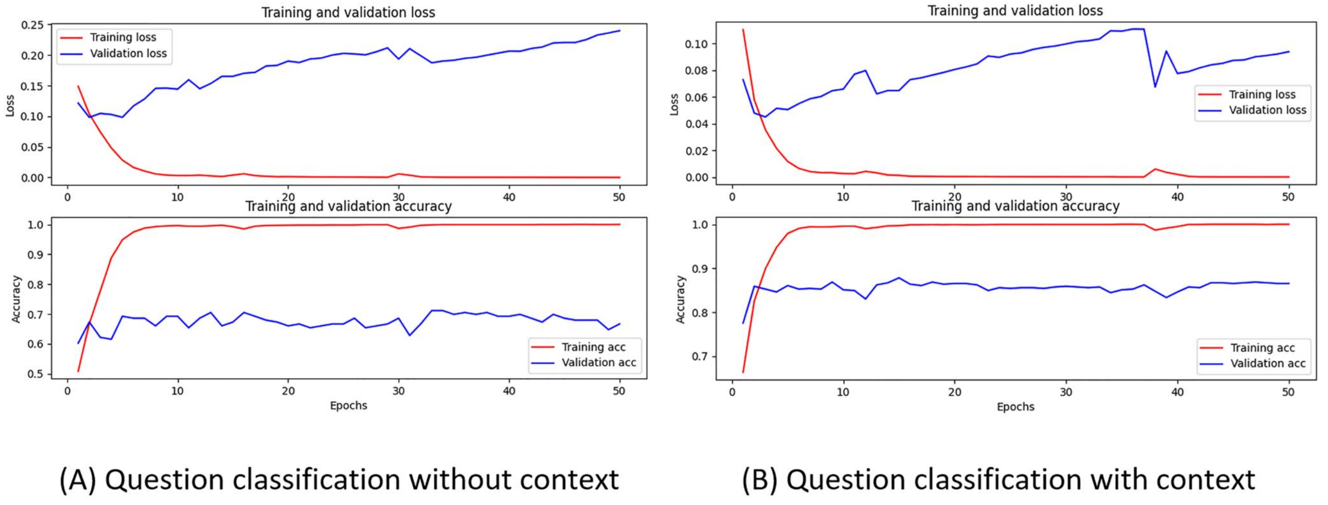
$$-\left(y\,log\,log(p) + (1-y)log\,log(1-p)\right) \tag{1}$$

If $M > 2$ (i.e. multiclass classification), we calculate a separate loss for each class label per observation and sum the result.

$$-\sum_{c=1}^{M} y_{o,c}\,log\,log\left(p_{o,c}\right) \tag{2}$$

For training a Tesla GPU of 16 GB RAM was used with 16 batch size and 50 epochs using AdamW optimiser and with learning rate of 1-e5. To better understand the performance of the intent classifier on the validation set, a confusion matrix was computed across the classes and visualised using a heatmap in Figure 5.

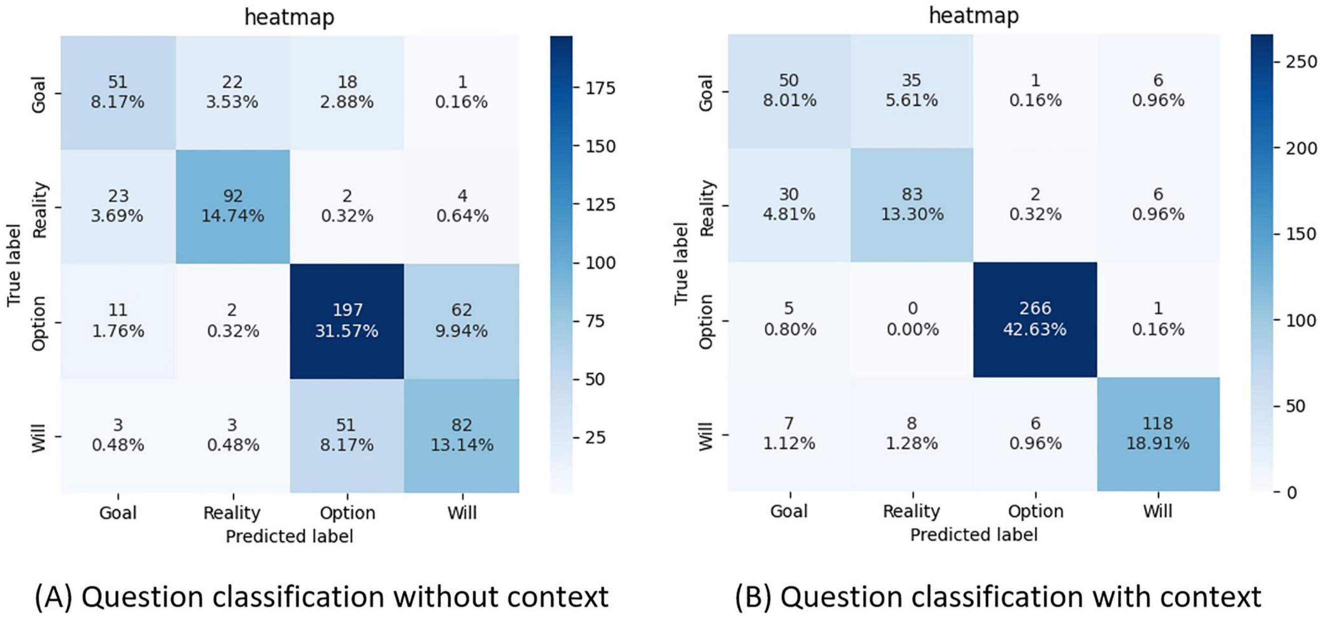
As indicated in Figure 5, there could be two major segments in the coaching conversation including (1) *Goal* and *Reality* exploration, and (2) *Option* and *Will* exploration, where the latter is more intuitive because questions corresponding to *Reality* and *Will* classes share similar context.



(A) Question classification without context

(B) Question classification with context

FIGURE 4 Training and validation performance of the intent classification model.



(A) Question classification without context

(B) Question classification with context

FIGURE 5 Confusion matrix for the classification model for the target classes.

This is also approved by computing cosine similarity which is calculated using the Distributed Representations of Sentences and Documents technique [9]. The average similarity between these two classes is approx. 0.7 whilst this value is much lower when comparing *GOAL* and *WILL* questions' similarity. Similar observation is valid for the *WILL* and *OPTION* classes where normally the coach's questions share similar intent when exploring options or commitment that the client is willing to take as the next step. As an example consider these two questions "*If you were at your best, what would you do?*" and "*What would you like to do as the next step?*". The former is exploring options and allows the client to propose as many options as possible whilst the latter is exploring the next step and looking into a commitment. Note that questions are abstracted for this discussion. Normally, there is conversation context carried by every question (e.g. *"depending on how your friend would react to this matter... how do you see yourself committing to the option you just mentioned?"*).

To better distinguish between such questions and be more effective in identifying the context of the conversation, it would be possible to consider the client's abstract answers along with the coach's questions to reduce the confusion between the classes.

In this regard, another experiment was conducted and we considered coaches' questions and client's answers to provide the model extra context to be able to classify the category in a more effective way. Figure 4b shows the training and validation profile of the second experiment. As it's shown in the graph we managed to achieve 85% accuracy (approx. +15% higher) on the validation set which consists of 20% of the entire dataset by including context. The overall performance of the system on our audio/video +1k hours labelled dataset is presented in Table 4.

Table 4 shows performance of the Question Intent classification on a different vertical where the session topic is considered as a variable across the dataset. It is observable that due to the complexity of the *well-being* and *life* topics, Question Intent classification model performs poorly compared to the other two topics. This is mainly because those topics are more likely to be compound, or answers are being complicated and normally tend to touch on different areas. However, topics such as leadership under the Executive category, have a more objective nature and it is easier to set a target for them.

4 | EXPERIMENTAL EVALUATION

4.1 | The first experiment

As an overall report on the performance of the system, several feedbacks have been collected on 19 sessions performed in real-time via the platform, 14 coaches found the application useful and would use it again. 8 out of 14 found the mapping *'accurate'*, 3 found the mapping *'not accurate'* and 3 found it *'can be improved'* Two of the sessions were removed due to low audio quality. Based on the feedback provided, our plan is to leverage self-supervised techniques including reinforcement learning to enhance both classification and mapping models over time with more data.

4.2 | The second experiment

This experiment has been designed to test the effectiveness of our feedback system. In this experiment, 30 coaching students were recruited with a diverse range of experience and qualifications. Then, they were divided into two groups including a control group and an experimental group. The control group conducted coaching sessions using their usual methods, while the experimental group conducted coaching sessions using the system provided. In order to comply with ethics and regulations we avoided using real client data in this experiment.

The students were assigned coaching videos from our dataset to watch in a 3-way coaching setting and both groups were asked to score the session, score the coach's performance, score the client's performance, and take notes of the session. Each student was asked to review a minimum of 10 coaching sessions, each 50 min to an hour.

TABLE 4 The overall performance of the proposed model across the category vertical in the dataset.

Category	Model	Performance
Life	Question intent classification	78.3%
Work	Question intent classification	92.1%
Executive	Question intent classification	87.7%
Well being	Question intent classification	79%

Accordingly, real-time feedback was provided to the experimental group on their coaching sessions, and data on the clients' progress and satisfaction with the coaching sessions was collected. The results between the control group and the experimental group were compared and the data was analysed to determine the effectiveness of the feedback system provided.

In the first step of our analysis, the length and structure of the notes were compared. The results show that the experimental group had 42% shorter notes, better structured notes and more specific notes as they could pay uninterrupted attention to the conversation flow and gestures. To be able to compare the notes, Cornell's note-taking system was used as ground truth and each note was compared against them. The group with visual feedback had better understanding of each segment in the conversation and had better time for structuring their notes.

In the second step of our analysis, the coach performance scores were compared. In regards to the coach's time management, the results show that the experimental group were more confident with their rating by 73%. This indicates that the provided feedback allowed the students to checkpoint the conversation easily and rely on the system for more effective time assessing the session.

4.3 | The third experiment

This experiment was designed to take into account the social impact of the coaching sessions based on the post-session analysis we provide to the coach. The aim was to observe the correlation between predicted factors in the conversation and the community feedback. This could be exploited towards self-development and improvement over time.

In this experiment we intend to know if there is a significant association between the conversation quality predicted using our proposed system and the number of views, likes and comments the video has received, aka community feedback. Therefore, a study was designed to run a chi-square test on the data we have collected. Table 5 shows a sample of the data collected using Youtube API.

Next, we grouped the number of views into 3 major categories. Table 6 is a contingency table that shows the number of videos in each category for conversation quality and number of views.

The expected frequencies for each cell in Table 6 is computed based on the total number of videos and the row and column totals, and the results are presented in Table 7.

Expected values computed using the following equation:

$$\frac{(row\ total \times column\ total)}{grand\ total} \tag{3}$$

Finally the *chi-squared* formula was applied to the observed and the calculated expected values to check the relation between two variables:

TABLE 5 Sample of the data collected using YouTube API.

Video ID	Conversation time management	Views	Likes	Comments	Duration	Publish date
6EKseAbVcpo	Good	14,219	214	27	'PT31M29S'	2022-10-22
Vdwya5j3D8k	Good	20,305	223	10	'PT1H37M25S'	2018-10-04
WyVC3c5pUS0	Poor	15,110	122	18	PT30M49S	2018-11-13
WoP9LIHFK6k	Poor	13,407	115	25	'PT30M49S'	2018-11-27
cPHY0C8Poqk	Poor	17,071	144	40	'PT28M20S'	2018-12-05

TABLE 6 Number of videos in each category for conversation quality and number of views.

	Good management quality	Poor management quality	Total
<10000 views	30	22	52
10000–2,0000 views	28	10	38
>20000 views	11	1	12
Total	69	33	102

TABLE 7 The expected frequencies.

	Good management quality	Poor management quality	Total
<10000 views	35.18 [0.76]	16.82 [1.59]	52
10000–2,0000 views	25.71 [0.20]	12.29 [0.43]	38
>20000 views	8.12 [1.02]	3.88 [2.14]	12
Total	69	33	102

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

In this equation, x^2 is the *chi-squared*, O_i is representing the observed value, and E_i is representing the expected value. In this experiment, the *chi-square* score was 6.1507.

Furthermore, the degree of freedom was computed using the following equation:

$$(number\ of\ rows - 1) \times (number\ of\ columns - 1) \tag{5}$$

The significance level of 0.05 was 4.605.

The calculated *chi-square* statistic (6.1507) was larger than the critical value (4.605), thus we can reject the null hypothesis and we can conclude that there is a significant association between conversation quality and number of views of the videos.

Our plan is to extend this experiment and take into account other factors such as video length, and other conversation features such as speak ratio and cadence to provide more robust feedback to the coach towards self-development and reflection. We also plan to expand our experiments to group coaching in the future [48].

5 | CONCLUSION

This paper presented a real-time conversation mapping system based on the recent advancements in the area of conversational AI that allows the coach to fully focus on the conversation, hence maximising engagement and minimising distractions by note taking. Moreover, the provided information could be used as feedback for managing time/session more effectively and avoiding timeouts or dead end conversation circles.

We demonstrated that the transformer based question intent classifier trained on our curated coaching dataset with >3k questions is capable of achieving 85% accuracy in identifying the session progress and mapping the conversation to the target coaching model (GROW in this study). Based on the feedback collected from the coaches on performance of this approach, approximately 75% find the approach useful and would want to use it again across their sessions.

We plan to release our coaching dataset and make it available for the community. Additionally, the platform will be available publicly for coaches to run their sessions using the proposed AI system. This will allow us to gather more feedback for the future.

In regards to the novelty of this research, to the best of our knowledge, this work is the first attempt in applying data and AI into the self-reflection practice in coaching and therapy. We are aiming to extend this work and incorporate different coaching models based on which the coach can identify the session progress and use it as real-time feedback on the performance of the session. Meaning, if there are options or realities that are not explored enough, the coach could raise reminders without being distracted by note taking. This work is unique in the sense that it is the first in its type that targets the coach instead of the client while other systems mostly target the client.

ACKNOWLEDGEMENT

We would like to thank Dr Joel Digirolamo from the international coaching federation for his helpful advice on reflective coaching and outcome measurements. We would also like to thank Saber Jelodari and Mousoumi Chudhury for helping with data labeling, review and dataset topic tagging.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The coaching dataset created in this research can be accessed upon request.

ORCID

Mohammad Hossein Amirhosseini https://orcid.org/0000-0002-3404-084X

REFERENCES

1. Yordanova, S.D., Dineva, S.Y.: Emotion regulation at work: employee and leader perspectives. In: Advancing Interpersonal Emotion Regulation and Social Regulation, pp. 113–149. IGI Global (2022)
2. Segers, J., Inceoglu, I.: Exploring supportive and developmental career management through business strategies and coaching. Hum. Resour. Manag. 51(1), 99–120 (2012). https://doi.org/10.1002/hrm.20432
3. Berg, M.E., Karlsen, J.T.: Mental models in project management coaching. EMJ - Engineering Management Journal 19(3), 3–13 (2007). https://doi.org/10.1080/10429247.2007.11431736
4. The peak performance centre: Coaching Model. [Online]. Available: (2021). https://thepeakperformancecenter.com/development-series/skill-builder/interpersonal/coaching-for-performance/coaching-model/
5. Martinez, D.: The mentor program for undergraduate students at stamford international university. The National Conference on Management and Higher Education 64 (2014)
6. Edgerton, N., Palmer, S.: SPACE: a psychological model for use within cognitive behavioural coaching, therapy and stress management. Coach. Psychol. 2(2) (2005). https://www.researchgate.net/profile/Stephen-Palmer-6/publication/322509343_SPACE_A_psychological_model_for_use_within_cognitive_behavioural_coaching_therapy_and_stress_management/links/5cb4a9234585156cd79ad21d/SPACE-A-psychological-model-for-use-within-cognitive-behavioural-coaching-therapy-and-stress-management.pdf
7. Palmer, S.: An international perspective on the development of coaching psychology: from Socrates to the present and where do we go from here? In: Keynote Paper Given on the 26 May 2010, at the 1st International Congress of Coaching Psychology, South Africa (2010)
8. Bresser, F., Wilson, C.: Excellence in Coaching, the Industry Guide, 2nd ed. Association for Coaching, London (2010)
9. Hawkins, P., Smith, N.: Coaching, Mentoring and Organisational Consultancy: Supervision and Development. Open University Press, London (2007)
10. Jackson, P.Z., McKergow, M.: The Solutions Focus: The SIMPLE Way to Positive Change. Nicholas Brealey, London (2002)
11. Dembkowski, S., Eldridge, F.: Beyond GROW: a new coaching model. The International Journal of Mentoring and Coaching 1(1) (2003)
12. Grant, A.M.: Is it time to REGROW the GROW model? Issues related to teaching coaching session structures. Coach. Psychol. 7(No. 2), 98–106 (2011). https://doi.org/10.53841/bpstcp.2011.7.2.98
13. Palmer, S.: PRACTICE: "A model suitable for coaching, counseling, psychotherapy and stress management". Coach. Psychol. 3(2), 71–77 (2007). https://doi.org/10.53841/bpstcp.2007.3.2.71
14. Mackintosh, A.: Growing on GROW – a more specific coaching model for busy managers. OUTCOMES (2005). [Online]. https://ezinearticles.com/?Growing-On-G.R.O.W---A-More-Specific-Coaching-Model-For-Busy-Managers&id=27766
15. Grabmann, C., Schermuly, C.C.: Coaching with artificial intelligence: concepts and capabilities. Hum. Resour. Dev. Rev. 20(1), 106–126 (2020). https://journals.sagepub.com/doi/abs/10.1177/1534484320982891?journalCode=hrda
16. Leadership, Oxford: Using Artificial Intelligence to Enhance Coaching – A New Study (2023). https://www.oxfordleadership.com/using-artificial-intelligence-to-enhance-coaching-a-new-study/
17. Brown, R.p., et al.: 2The impact of coaching on emerging leaders. International Journal of Evidence Based Coaching Coaching and Mentoring 19(2) (2021)
18. Kamphorst, B.A.: E-coaching systems: what they are, and what they aren't. Personal Ubiquitous Comput. 21(4), 625–632 (2017). https://doi.org/10.1007/s00779-017-1020-6
19. Siau, K.L., Yang, Y.: Impact of artificial intelligence, robotics, and machine learning on sales and marketing. In: Midwest United States Association for Information Systems 12th Annual Conference, 18-19 May 2017, Springfield, Illinois (2017). http://aisel.aisnet.org/mwais2017/48
20. Chen, Y., et al.: Constructing a nutrition diagnosis expert system. Expert Syst. Appl. 39(2), 2132–2156 (2012). https://doi.org/10.1016/j.eswa.2011.07.069
21. Panetta, K.: Widespread Artificial Intelligence, Biohacking, New Platforms and Immersive Experiences Dominate This Year's Gartner Hype Cycle (2018). https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/
22. Terblanche, N.: A design framework to create Artificial Intelligence coaches. International Journal of Evidence Based Coaching and Mentoring 18(2) (2020). https://radar.brookes.ac.uk/radar/items/312d40ec-ccdf-431c-a062-2aa862166ac4/1/
23. da Silva, L., et al.: Conversational agents in healthcare: a systematic review. J. Am. Med. Inf. Assoc. 25(9), 1248–1258 (2018). https://doi.org/10.1093/jamia/ocy072
24. Saenz, J., et al.: The usability analysis of chatbot technologies for internal personnel communications. In: Industrial and Systems Engineering Conference, Pittsburgh, Pennsylvania, USA, pp. 1375–1380 (2017). http://toc.proceedings.com/36171webtoc.pdf
25. Chung, K., Park, R.C.: Chatbot-based healthcare service with a knowledge base for cloud computing. Cluster Comput. 22(1), S1925–S1937 (2019). https://doi.org/10.1007/s10586-018-2334-5
26. Fryer, L., Carpenter, R.: Bots as language learning tools. Lang. Learn. Technol. 10(3), 8–14 (2006)
27. Neff, G., Nagy, P.: Talking to bots: symbiotic agency and the case of Tay. Int. J. Commun. 10, 4915–4931 (2016)
28. Araujo, T.: Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput. Hum. Behav. 85(August), 183–189 (2018). https://doi.org/10.1016/j.chb.2018.03.051
29. Bickmore, T.W., et al.: A randomized controlled trial of an automated exercise coach for older adults. J. Am. Geriatr. Soc. 61(10), 1676–1683 (2013). https://doi.org/10.1111/jgs.12449
30. Watson, A., et al.: An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. J. Med. Internet Res. 14(1), e1 (2012). https://doi.org/10.2196/jmir.1629
31. Kaptein, M., et al.: Personalizing persuasive technologies: explicit and implicit personalization using persuasion profiles. Int. J. Hum. Comput. Stud. 77, 38–51 (2015). https://doi.org/10.1016/j.ijhcs.2015.01.004
32. Kocielnik, R., et al.: Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent, pp. 881–894. DIS '18: Designing Interactive Systems Conference, Hong Kong (2018)
33. Pereira, J., Diaz, O.: 'Using health chatbots for behavior change: a mapping study. J. Med. Syst. 43(5), 135 (2019). https://doi.org/10.1007/s10916-019-1237-1
34. Lucas, G.M., et al.: It's only a computer: virtual humans increase willingness to disclose. Comput. Hum. Behav. 37, 94–100 (2014). https://doi.org/10.1016/j.chb.2014.04.043
35. Bradeško, L., Mladenić, D.: A survey of chatbot systems through a Loebner Prize competition. In: Proceedings of the Slovenian Language Technologies Society, Eighth Conference of Language Technologies, Ljubljana, Slovenia, pp. 34–37 (2012). http://nl.ijs.si/isjt12/proceedings/isjt2012_06.pdf
36. Terblanche, N.: International Journal of Evidence Based Coaching and Mentoring 18(2), 152–165 (2020). https://doi.org/10.24384/b7gs-3h05
37. Keptika company: Keptika Coaching Platform (2022). www.keptika.com
38. Vaswani, A., et al.: Attention Is All You Need (2017). https://arxiv.org/abs/1706.03762

39. Hosking, T., Lapata, M.: Factorising Meaning and Form for Intent-Preserving Paraphrasing. Available: (2021). https://arxiv.org/pdf/2105.15053.pdf
40. Kuchaiev, O., et al.: Nemo: A Toolkit for Building Ai Applications Using Neural Modules (2019). https://arxiv.org/pdf/1909.09577.pdf
41. Rao Koluguri, N., Park, T., Ginsburg, B.: TitaNet: Neural Model for Speaker Representation with 1D Depth-wise Separable Convolutions and Global Context (2022). https://arxiv.org/abs/2110.04410
42. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: Proceedings of AAAI 31 (2018). Available: [1612.03975] ConceptNet 5.5: An Open Multilingual Graph of General Knowledge (arxiv.org)
43. Culture at work: The GROW Model. [Online]. Available: (2020). https://www.coachingcultureatwork.com/the-grow-model/
44. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI blog 1(8), 9 (2019)
45. Brown, T., et al.: Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020)
46. Kojima, T., et al.: Large language models are zero-shot reasoners. Adv. Neural Inf. Process. Syst. 35, 22199–22213 (2022)
47. WebRTC: Real-time Communication for the Web (2022). https://webrtc.org/
48. Nacif, A.P., et al.: Online group coaching: the experience of postgraduate students during the COVID-19 pandemic. Coaching 16(2), 1–15 (2023). https://doi.org/10.1080/17521882.2023.2205598

How to cite this article: Jelodari, M., Amirhosseini, M. H., Giraldez-Hayes, A.: An AI powered system to enhance self-reflection practice in coaching. Cogn. Comput. Syst. 1–12 (2023). https://doi.org/10.1049/ccs2.12087