Article

Using hidden Markov modelling to reveal in-session stages in text-based counselling

Check for updates

Ziru Fu[1,4], Yu Cheng Hsu[1,4], Christian S. Chan[2] ✉, Joyce Liu[1] & Paul S. F. Yip[1,3] ✉

Counselling sessions have multiple stages, each with its themes and objectives. This study aimed to apply Hidden Markov Models (HMMs) to analyse counselling sessions from Open Up, an online text-based counselling platform in Hong Kong. The focus was on inferring latent stages over word distributions and identifying distinctive patterns of progression in more versus less satisfying sessions. Transcripts from 2589 sessions were categorized into more satisfying sessions ($n = 1993$) and less satisfying sessions ($n = 596$) based on post-session surveys. A message-level HMM identified five distinct stages: Rapport-building, Problem-identification, Problem-exploration, Problem-solving, and Wrap-up. Compared with less satisfying sessions, more satisfying sessions saw significantly more efficient initial rapport building (7.5% of session duration), problem introduction (20.2%), problem exploration (28.5%), elaborated solution development (46.6%), and concise conclusion (8.2%). This study offers insights for improving the efficiency and satisfaction of text-based counselling services through efficient initial engagement, thorough issue exploration, and focused problem-solving.

Counselling is a process that dynamically evolves through interactions between the counsellor and the help-seeker toward achieving the desired therapeutic outcomes, such as facilitating self-awareness, problem-solving, and adaptive behavioural changes[1,2]. Understanding the structure and stages of this process is crucial for improving counselling practices and outcomes. By identifying these stages within counselling sessions, we can gain insights into how changes occur, and address help-seeker needs more effectively[1,2]. In this study, we sought to apply Hidden Markov Models (HMMs) to identify and analyse the stages within Cantonese text-based counselling sessions, offering a scalable and interpretable approach to understanding the counselling process and enhancing session outcomes

Online text-based counselling has gained prominence with the ubiquity of digital communication tools and the growing need for accessible mental health services. This form of counselling addresses barriers to services, such as stigma, cost, and limited access[2,3]. As the number of users and providers of such services increases, effective methods of assessing counselling sessions become imperative to maintain and enhance their effectiveness and user satisfaction. Efficient and reliable assessments could also help guide the training of counsellors and highlight the areas where they excel and those needing improvement. In addition, assessing counselling sessions provides insights into the needs and preferences of help-seekers, which could potentially lead to more tailored approaches addressing specific issues and concerns. However, existing methods for evaluating session transcripts are either labour-intensive or limited by a lack of interpretability[4].

As an intersection between artificial intelligence (AI) and linguistics, NLP enables the representation, analysis, and generation of large corpora of language data[5], making it a promising tool for studying counselling transcripts. Current research on text-based counselling often focuses on either i) applying theoretically derived annotation frameworks, which involve manually examining and coding transcripts[6], or ii) using advanced NLP models like Pre-trained Language Models (PLMs) and Large Language Models (LLMs) to automatically analyse session dynamics[7].

While these approaches have helped advance the understanding of text-based counselling, they suffer from several limitations. First, many studies only focused on studying counsellor behaviours and ignored the counsellor-help seeker interaction [e.g., refs. 4,8,9]. Second, most approaches require a large set of annotated data [e.g., refs. 6,7,10]. However, manual annotation and coding are not easily scalable and adaptable to large corpora. Third, despite the capabilities of state-of-the-art NLP models, such as Pre-

[1]The Hong Kong Jockey Club Centre for Suicide Research and Prevention, Faculty of Social Sciences, The University of Hong Kong, Hong Kong SAR, China. [2]Department of Psychology and Linguistics, International Christian University, Tokyo, Japan. [3]Department of Social Work and Social Administration, Faculty of Social Sciences, The University of Hong Kong, Hong Kong SAR, China. [4]These authors contributed equally: Ziru Fu, Yu Cheng Hsu. ✉e-mail: shaunlyn@hku.hk; sfpyip@hku.hk

trained Language Models (PLMs) and Large Language Models (LLMs), they often suffer from a lack of interpretability. These models employ highly complex architectures comprising millions or even billions of parameters, making it difficult to study the contribution of individual parameters to the overall model outcomes. The transformations that occur at each layer of the model are frequently not interpretable by humans, resulting in an inherent lack of transparency. This low transparency raises concerns about applying NLP models to process data containing complex and sensitive mental health issues presented by help-seekers in professional counselling where understanding the decision-making process by such models is crucial.

Previous studies have explored counselling dynamics and processes using NLP methods. Althoff et al. [9] used anonymized text-based counselling sessions from a crisis intervention service to study the characteristics of high-quality sessions. After each session, the texter received a follow-up question ("How are you feeling now? Better, same or worse"), which the researchers used as the indicator of the session's quality. Binary labels were applied: "Better" as high-quality and "Same/Worse" as low-quality. The authors utilised HMMs to identify counselling stages in digital counselling transcripts from a crisis counselling platform in the US and identified five counselling stages (i.e., Introductions, Problem introduction, Problem exploration, Problem solving and Wrap up). The results showed that quicker transition from early stages to solution-focused stage could lead to more satisfying outcomes. However, to the best of the authors' knowledge, there is no study that has applied such method to analyse counselling dynamics in Cantonese context.

Zhang et al. [11] developed their NLP approach to study how participants oriented the flow of interactions in the context of Crisis Text Line, a crisis counselling platform that provides a free 24/7 service for those in a mental health crisis. The authors considered responses to a post-session survey question asking the texter whether the session was helpful (yes/no) and then split the sessions into high- and low-quality sessions. Their results highlighted the importance of balancing between two objectives: advancing the counselling session towards a resolution and addressing the emotional distress of the help-seeker. While their study primarily explored the balance of conversational flow (i.e., forward and backward orientation) with a latent space embedding approach, it is relatively silent on identifying distinct counselling stages, the transitions between them, and different interaction patterns between counsellors and help-seekers.

In the field of text-based counselling, understanding the dynamics between counsellors and help-seekers is crucial for improving counselling session outcomes [12]. To this end, researchers have endeavoured to propose various theoretically grounded annotation frameworks or employ NLP techniques to analyse counselling dynamics and predict outcomes. However, as mentioned above, these studies had several limitations, including the need for labour-intensive human annotation and the lack of interpretability in NLP models. For instance, Li et al. [6] designed an annotation framework to analyse how counsellors' strategies, along with help-seekers' reactions to them, could lead to better counselling outcomes. The authors used 520 sessions from Xinling, an online text-based counselling platform developed by the authors. A dozen annotators annotated tens of thousands of messages within these sessions. Help-seekers' self-reported scores in the post-session survey were used as session quality indicators. They validated the feasibility of the framework by analysing counsellor-help-seeker interaction patterns in sessions with different scores. The annotation process underscores the labour-intensive nature, thus lack of scalability, of this type of research.

Another example is Liu and colleagues [7], who utilised 169 sessions with thousands of messages from Shout, a mental health crisis text line in the UK. Each message was manually annotated into one of six session stages defined by the Shout clinical supervisors and research psychologists. Longformer NLP architecture [13] was then fine-tuned with the annotated data as a potential automatic stage classifier. Their fine-tuned model achieved

accuracy of 87.75%, which, as the authors discussed, could potentially serve as a tool to review conversations and improve counsellor training. Similarly, the manual annotation across predefined counselling stages highlighted the extensive efforts required for data preparation before the NLP application. Furthermore, their approach of applying state-of-the-art NLP models (i.e., Longformer NLP architecture) offers limited insights into how decisions were drawn from the data, revealing the challenges of low interpretability of NLP methods in current studies.

To address these challenges, HMMs offer an efficient and interpretable alternative for analysing the counselling process. HMMs, which have been applied in diverse fields such as NLP [14], psychology [15] and bioinformatics [16], serve as a powerful approach for modelling sequential data. HMMs are probabilistic models that are used to derive a probability distribution of hidden states that are not directly observable based on a sequence of observed values. Each hidden state is characterized by a probability distribution over the observed values called *the emission probability distribution*, while the transitions among the hidden states are governed by *the transition probability distribution*.

The motivation for using HMMs in this study comes from the alignment of counselling process with HMM mechanisms, and the interpretability and efficiency of HMM. First, HMMs are particularly well-suited to capture the dynamic structure of interactions between counsellors and help-seekers. Based on various theoretical framework on counselling [e.g., refs. [17,18]], we assumed that the text-based counselling process entails specific therapeutic stages. These stages represent the hidden states in the HMM. Each stage has specific foci and goals, and thus, the words represented and used in each stage can differ. The HMM could capture these wording differences through the emission probabilities, and the transition between stages can be captured in the HMM through the transition probability.

Second, due to its simpler structure and fewer parameters compared to other deep learning techniques, the HMM could provide a transparent and interpretable framework for analysing sequential interactions in counselling. This transparency is particularly important in sensitive domains like counselling, where interpretability can guide the improvement of counsellor training and session outcomes.

Third, the HMM is quite efficient as it requires no labour-intensive human annotation process and computationally less demanding compared to deep learning models such as long short-term memory (LSTM) [19]. In summary, the HMM was applied in this study due to the natural alignment of counselling process with the HMM, the inherent interpretability, and transparency of the model and its efficiency.

Our study introduced a message-level HMM to assess in-session stages and outcomes in Cantonese text-based counselling. We examined lexicon distributions and inferred latent stages in counselling sessions. Specially, this unsupervised approach investigated different topics being discussed, time allocation, as well as stage transitions in more versus less satisfying sessions. Our findings can potentially reveal in-session stages and highlight important differences in more versus less satisfying sessions, which is potentially illuminating to enhance the effectiveness and help-seeker satisfaction of counselling practice.

Such probabilistic inference of counselling session structure has been used in recent works. For example, Althoff et al. [9] utilised HMM to study session dynamics and found that more successful counsellors tended to spend more time providing solutions. However, the study imposed a fixed order of stage transitions within session. For example, if the current message was in Stage 4 (Problem-solving), the next message must either stay in Stage 4 or progress to Stage 5 (Wrap-up), with no possibility of being categorized into any previous stages. This imposed order may not adequately capture the heterogeneity of counselling dynamics; counselling sessions may not follow a linear or fixed progression, as it depends on the theoretical approaches and work settings of the counsellor [20]. To overcome this limitation, in our study, the transitions between different stages were not constrained to a fixed order. This allowed for a more realistic representation of counselling sessions.

Methods

Data source

The data came from Open Up, a 24/7 free online text-based counselling service in Hong Kong designed to provide emotional support to Hong Kong youth between 11–35 years old. This study was approved by the Human Research Ethics Committee, The University of Hong Kong (EA230548). Users gave consent to using their anonymized text data for research purposes by accepting the privacy policy before the commencement of service. Help-providers (i.e., counsellors) gave their consent to using their text data for research purposes by signing consent forms.

In total, 83,344 counselling sessions from 1 June 2022 to 1 June 2023 were extracted. We aimed to identify a subset of counselling sessions that: i) would allow us to differentiate between more satisfying vs less satisfying sessions, and ii) included a typical range of message exchanges that reflected the overall distribution of session lengths in the dataset. Six inclusion criteria, detailed below, were applied.

First, 29,387 sessions containing at least four message exchanges between the counsellor and the help-seeker were included in the analyses and were referred to as "valid sessions." Four exchanges were used as the inclusion criteria because sessions with fewer message exchanges tended to be non-sessions; the users did not have bona fide intention to use the service.

Second, among the valid sessions, 13,340 sessions (45.4%) were of first-time visits. We only included first-time sessions in the current study because the service was primarily designed as a one-off "walk-in" service and because, the subsequent sessions, as the counsellor and the help-seeker have probably known each other, might be more heterogeneous. Therefore, analysing first-time sessions provided a more controlled environment to study fundamental interaction patterns when users initially seek help on the platform.

Third, of the 13,340 first-time visit sessions, 3850 (28.9%) were accompanied with responses in the post-session survey question, "How helpful will you rate the counselling session?" The response options were Not helpful at all, Slightly helpful, Somewhat helpful, Quite helpful, Extremely helpful. Informed by studies reviewed above [e.g., refs. 9,11], sessions with feedback "Extremely helpful" and "Quite helpful" were categorized as "more satisfying sessions", sessions with feedback "Somewhat helpful" were annotated as "somewhat satisfying sessions", and sessions with rating "Not helpful at all" and "Slightly helpful" were labelled as "less satisfying sessions".

Fourth, as "Extremely helpful", "Quite helpful", "Slightly helpful" or "Not helpful at all" feedback was employed as a session quality indicator, we excluded 1070 somewhat satisfying sessions. As our primary aim was to compare and analyse distinctive patterns in clearly more versus less satisfying sessions, these sessions were excluded because they fall in the middle of the feedback spectrum. Excluding somewhat satisfying sessions in our subsequent analyses allowed us to focus on sessions that provided more distinguishable outcomes, ensuring clearer contrasts between high and low satisfaction. In Supplementary Fig. 1, we included such sessions to train the model to demonstrate that it is necessary to exclude such sessions to ensure a more clear-cut comparison of inter-action patterns between more versus less satisfying sessions. After excluding somewhat satisfying sessions, 2780 sessions (72.7%) were included for further analysis.

Fifth, we employed a selective approach by only including 2589 sessions (93.1%) that had more than 4 message exchanges and fewer than 132 message exchanges. This upper bound was determined based on 1.5*Interquartile Range (IQR) rule, which is a standard practice to remove outliers. By focusing on sessions within the range, we concentrated on those that reflected the typical features of first-time visit sessions. Moreover, sessions that were too short (i.e., fewer than four exchanges) might not provide enough context or content to allow for a meaningful analysis, while excessively long sessions could include anomalies of typical counselling practices. The selected range ensured a balance, capturing sessions that were sufficiently detailed without being overly extensive.

Sixth, among the 2589 sessions, 1993 sessions (77.0%) were coded as "more satisfying sessions", and 596 sessions (23.0%) were coded as "less satisfying sessions". The data inclusion flow is illustrated in Fig. 1.

Statistics of more and less satisfying session groups as well as somewhat satisfying session group are presented in Table 1. Notably, K6 score, the presence of suicidal ideation and the percentage of premature departure between the three groups were compared. Open Up implemented a pre-session survey that directly assessed distress and suicide risk in November 2021. Thereafter, help-seekers are invited to fill in a brief survey before being connected to a counsellor. The survey contained the Kessler Psychological Distress Scale (K6[21,22]), a question on suicidal ideation, and a question on the most pressing issue. The K6 comprises six questions, each providing five response options ranging from 0 ("Never") to 4 ("Always"), allowing for the evaluation of psychological distress levels of help-seekers. A larger value represented a higher psychological distress level. The question concerning suicidal ideation presented help-seekers with binary choices (yes/no) inquiring, "In the past two weeks, have you had thoughts about hurting or killing yourself?". In addition, premature departure was added as a general indicator of user dissatisfaction. Premature departure refers to the situation where the help-seeker drops out from the session without indicating that they would leave or establishing an agreement to end the session with the counsellor[23]. Analysis of Variance (ANOVA) was used to indicate whether there was a statistically significant difference in K6 distress scale between the three session groups. For pairwise comparisons of the K6 scores, Tukey's Honest Significant Difference (HSD) test was applied as a post-hoc test. Chi-square test was used to assess the presence of suicidal ideation and the percentage of premature departure between the groups. For pairwise comparisons for the categorical variables, pairwise Chi-square tests with Bonferroni correction were conducted to adjust for multiple comparisons. Python package *SciPy* (version 1.12.0; Python Software Foundation) was employed to conduct the tests.

Among the 2589 valid sessions, 1993 were "more satisfying" sessions (77.0%). Among them, 1490 had pre-session survey data (74.8%), while 446 out of the 596 "less satisfying" sessions (74.8%) containing such data. And among 1070 "somewhat satisfying" sessions, 778 sessions (72.7%) contained pre-session feedback. As illustrated in Table 1, there was a small but significant difference in psychological distress between more and less satisfying sessions (Cohend = 0.23, $p < 0.001$). There was no significant difference in suicidal ideation between more and less satisfying sessions ($OR = 0.85, p = 0.128$). The less satisfying sessions saw a significantly higher likelihood of premature departure compared with more satisfying sessions ($OR = 0.30, p < 0.001$).

Data pre-processing

Before training the HMM, the text data was pre-processed as machine-accessible vectors using the steps described below.

Word filtering. First, Python package *opencc* (version 1.1.7; Python Software Foundation) was applied to convert scattered Simplified Chinese to Traditional Chinese. Second, as there are no boundaries between words in Cantonese text (e.g., "I am depressed" looks like "Iamdepressed"), Jieba[24], a cutting-edge word segmentation dictionary, was applied to do word segmentation. Third, meaningless words or phrases (e.g., "跟住" meaning "then" and "好似" meaning "seem"), which carry little information and introduce noise in analyses were removed using a stopwords dictionary tailor-made for Open Up data. This stopwords dictionary was composed of two individual dictionaries (i.e., Jieba Stopwords Dictionary[24], PyCantonese Stopwords Dictionary[25]) using the Python package *Jieba* (version 0.42.1; Python Software Foundation). Next, we removed messages without tokenization representations after the above pre-processing steps. Across the 2589 valid sessions with 198,316 messages, 20,314 unique words were identified.

Fig. 1 | Data inclusion flow.
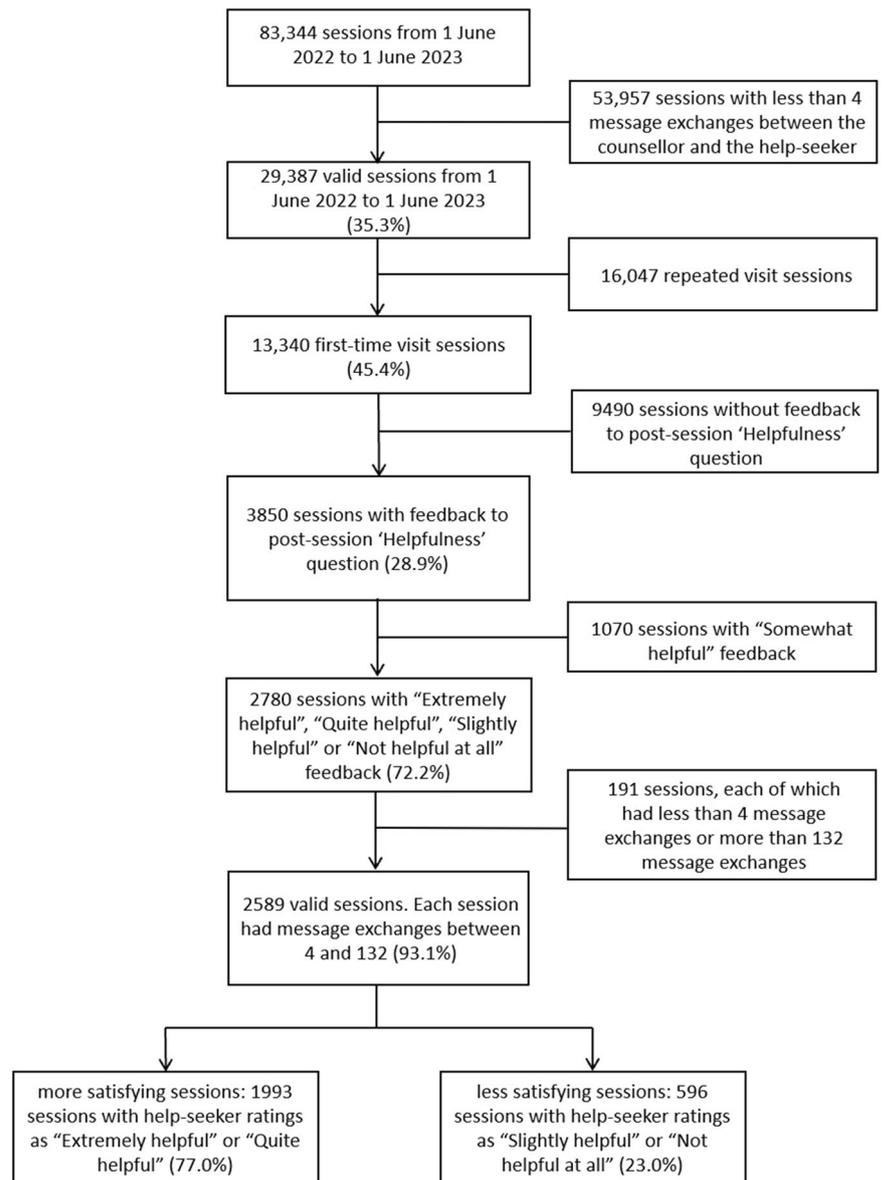
Feature extraction. To allow the HMM to effectively capture and differentiate the linguistic features of counsellors and help-seekers, to mitigate the noise introduced by unimportant words, and to reduce computational complexity, it is crucial to focus on more relevant and important words.

To select important words that can reflect typical thematic features of sessions, Python package *Scikit-learn* (version 1.3.0; Python Software Foundation) was used to calculate TF-IDF (Term Frequency-Inverse Document Frequency) score for each word. Initially, we marked the words from the counsellor and the help-seeker separately. A prefix "C_" was added to each word if it was originated from the counsellor (e.g., "hello" was tagged as "C_hello") and a prefix "H_" was added if it was from the help-seeker (e.g., "exam" was tagged as "H_exam"). This approach aimed to ensure that the subsequent HMM model training could effectively capture and differentiate the linguistic features of counsellors and help-seekers, providing insights into their respective contributions during the session. After labelling, the corpus composed of 26,017 unique words.

After word tagging, the TF-IDF (Term Frequency-Inverse Document Frequency) score for each word was calculated. Words with a lower than 3rd quartile TF-IDF score (6.44) were excluded since low TF-IDF scores usually indicate words that are frequent in the corpus but carry little meaningful information (e.g., "簡單" meaning "simple" and "對話" meaning "conversation"). These terms might not be representative of the overall thematic content of the corpus. As illustrated in Fig. 2, the threshold was set to include the top 25% of words with the highest TF-IDF scores, focusing the analysis on words that were more indicative of the themes or stages in the counselling process. The selection of the third quartile as the threshold balanced the need to retain important terms while excluding overly specific or trivial ones. For instance, using a 90th percentile as the threshold would risk removing words that are essential for identifying counselling stages, such as "你好 (hello)", "多謝 (thank) " and "聯絡 (contact)". Conversely, choosing a 50th percentile would introduce too many trivial terms, such as geographical locations such as "荃灣 (Tsuen Wan)", which are not relevant to the counselling dynamics. Therefore, the third quartile was considered valid for excluding outliers while preserving words indicative of general counselling themes.

After word selection, we removed 68,009 messages from 198,316 messages (34.3%) containing no representative words. Including these messages could potentially increase data sparsity and expose the model to more noise, which might make it more difficult to train the model. Finally,

Table 1 | Comparison of statistics between more satisfying, somewhat satisfying and less satisfying sessions

Variables	More satisfying sessions ($n = 1{,}993$)	Somewhat satisfying sessions ($n = 1{,}070$)	Less satisfying sessions ($n = 596$)	F/χ^2	p-value	Effect size ("more" vs "less")	Post-hoc
K6 score Mean (SD)[a]	11.7 (3.9)	11.9 (3.7)	12.4 (4.2)	10.00	<0.001	0.23[b]	"less" > "somewhat" > "more" [d]
With suicidal ideation n(%)[a]	485 (32.6)	238 (30.6)	160 (35.9)	4.11	0.128	0.85[c]	"less" = "somewhat" = "more" [e]
Premature departure n(%)	767 (38.5)	516 (48.2)	402 (67.4)	168.2	<0.001	0.30[c]	"less" > "somewhat" > "more" [e]
Session duration in minutes Mean (SD)	70.3 (32.4)	67.4 (31.9)	53.7 (30.2)	62.9	<0.001	0.52[b]	"more" > "somewhat" > "less" [d]
Number of message exchanges Mean (SD)	61.6 (30.7)	55.4 (28.7)	38.3 (25.7)	146.5	<0.001	0.79[b]	"more" > "somewhat" > "less" [d]
Number of message exchanges per minute Mean (SD)	1.0 (1.4)	1.1 (2.9)	0.8 (0.4)	5.7	0.003	0.21[b]	"more", "somewhat" > "less" [d]

[a] $n(moresatisfyingsessions) = 1490$, $n(somewhatsatisfyingsessions) = 778$ and $n(lesssatisfyingsessions) = 446$.
[b] Cohen's d.
[c] Odds Ratio.
[d] Tukey's Honest Significant Difference (HSD) test, the pair difference at 95% level of significance is reported.
[e] Pairwise Chi-square tests with Bonferroni Correction, the pair difference at 95% level of significance is reported.
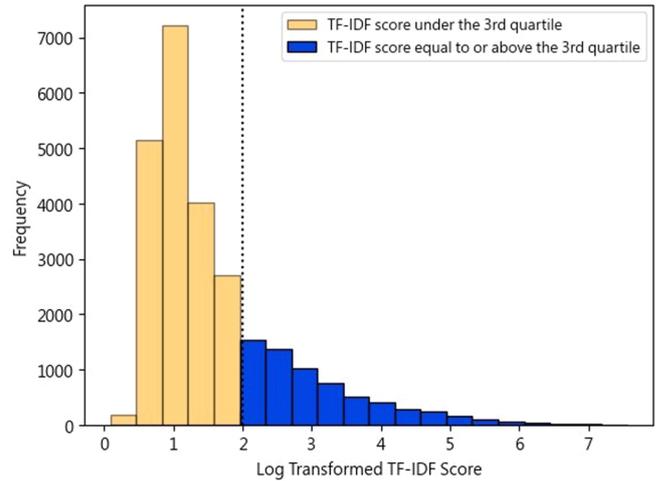

Fig. 2 | Word TF-IDF score (log-transformed) distribution.

2589 sessions with 130,307 messages composed of 6505 unique words in total were used to train the HMM model.

Session stage analysis using Hidden Markov Model (HMM)

Counselling sessions are goal-oriented, which means the counsellor and the help-seeker collaborate to aid the help-seeker in handling their emotional distress or make steps toward addressing their issues of concern. Sessions of this type often followed a common structure. Our HMM approach sought to automatically discover session stages from counselling sessions in a totally unsupervised manner. This approach maps each session into sequences of words and then infers patterns over word distributions. The inferred patterns in the form of latent session stages were then interpreted and assigned intuitive and meaningful labels.

An HMM could be represented as: $\theta = (S, O, \pi, A, B)$, where $S = \{s_1, s_2, \ldots, s_i\}$ contains a set of latent states, $O = \{o_1, o_2, \ldots, o_t\}$ is a set of observations from the first message to the t-th message, and in $\pi = \{\pi_1, \pi_2, \ldots, \pi_i\}$, $\pi_i = P(s_i)$ is the initial probability of state s_i. A is the transition probability matrix with size $i \times i$, and the element of A in the k-th row and j-th column refers to the probability transits from state k to state j. B is the emission probability matrix with the size $i \times w$, where w is the total amount of words included in this study, and the element of B in the k-th row and j-th column refers to the probability that a word j occurs in the stage k.

The HMM problem formulation implicitly makes the following assumptions: i) The transition between the hidden states is conditional to only the previous state; ii) The probability of the observed words is only conditional to the current latent state; and iii) The transition probability does not change over time (i.e. independent of time). The emission probability is estimated using a Poisson distribution, which accounts for variations in message length. The goal of training an HMM to induce a particular structure is to select the model that maximises the probability of the observed input.

A three-state HMM is illustrated as an example in Fig. 3, with stage Greetings (G), Problem-solving (P), and Wrap-up (W). The transition probability matrix A is a three-by-three matrix modelling the transition between three states. The emission probability is the word's probability of occurring given the observed state.

In our task, rather than specifying the number of session stages, the optimal number S of latent stages was learned from the observed sequences during model training, and O would be a sequence of word vectors within each message. All the counsellor and the help-seeker messages were treated as input, and each message was a distinct state. In other words, the initial hidden state was set in the following manner: the first 1/n message was seen as the 1st state, and the message from 1/n to 2/n messages was at the 2nd state, and so forth. Baum-Welch algorithm[26] was employed to approximate the parameters of HMM. This algorithm was derived from expectation
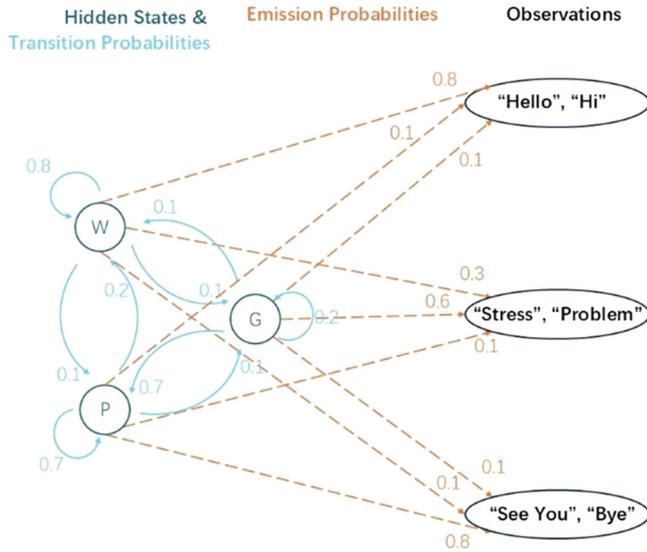
Fig. 3 | A three-state HMM for counselling process.

Table 2 | BIC under different number of stages

Number of stages	BIC
3	2979.10
4	2892.29
5	2805.82
6	2888.57
7	2884.62

maximization (EM), beginning with randomly initialized parameters and then iterating until convergence.

Python package *hmmlearn* (version 0.3.0; Python Software Foundation) was employed to implement our HMM. We ran the model with multiple initialised transition probabilities to avoid getting stuck in local optima. The initial transition probability matrix, which produced the best model that converged, is included in Supplementary Table 1. The number of hidden stages S was allowed to range from three to seven, and Bayesian Information Criterion (BIC), a measure of how likely the observed sequences would be under a proposed model, was applied to obtain the best-fit S. Smaller values of BIC mean better model fit over observations.

After training the HMM, we assigned a thematic title to each stage based on the keywords exhibiting the highest emission probabilities. In addition, we compared the session progression patterns of more satisfying and less satisfying sessions, focusing on the time allocated to each stage and the transition probability matrix. To determine the statistical differences in time allocations across different stages in more satisfying and less satisfying sessions, Shapiro-Wilk Test was first applied using Python package *SciPy* (version 1.12.0; Python Software Foundation) to assess if time allocated in each stage was normally distributed. Subsequently, Welch's t-test was implemented using Python package *SciPy* (version 1.12.0; Python Software Foundation) given its capacity to handle potentially unequal variances and sample sizes[27]. Additionally, Cohen's d was calculated using Python package *Numpy* (version 1.23.5; Python Software Foundation) as a measure of effect size to quantify the magnitude of the differences in time allocation in more satisfying and less satisfying sessions.

Results
Session stage topic analysis
We first determined the optimal number of stages by BIC. As shown in Table 2, $S = 5$ produced the best BIC fit for the current corpus.

Keywords with the highest emission probability in each stage are shown in Fig. 4. Based on the emission probability, we found the HMM stages aligned with the general counsellor practice in Open Up, as well as that described in the general counselling literature.

Stage 1 was labelled as the Rapport-building stage. It consisted of basic introductions, such as keywords "輔導員 (counsellor)", "稱呼 (address)", "你好 (hello)" "當值 (on duty)" and "請問 (may I ask)" in dominant messages like "你好, 我是今日當值輔導員 (Hello, I am the counsellor on duty today)" and "請問如何稱呼你 (May I ask how to address you)".

Stage 2 was entitled as the Problem-identification stage. This was the only stage where emission probability was dominated by the help-seeker keywords. In this stage, the help-seeker introduced and briefly described the issues they were facing, which helped the counsellor to get an overview of the problem and to determine the counselling strategies. For example, keywords like "朋友 (friend)", "問題 (problem)", "情緒 (emotion)", "之前 (before)" and "工作 (work)" were often contained in messages sent by help-seekers implying their presenting issues.

Stage 3 was referred to as the Problem-exploration stage. In this stage, both the counsellor and the help-seeker dived deeper into the problem. They might discuss the help-seeker's feelings, thoughts, and reactions related to the presenting issue. Other people involved in the issue might also be discussed. Goal-setting was also included in this stage. Keyword "情緒 (emotion)" implied the counsellor's focus on discussing the help-seeker's feelings in this stage. Keywords such as "明白 (understand)" and "擔心 (worry)" indicated that the counsellor frequently expressed empathy like "I understand you are sad right now" or "I worry about you". Keyword "情況 (situation)" and "影響 (affect)" suggested a comprehensive assessment of the help-seeker's current circumstances, the various factors impacting their emotional states and possible outcomes of their reactions.

Stage 4 was the Problem-solving stage. This stage involved the counsellor and the help-seeker collaboratively developing actionable strategies that were supportive and informative, such as counsellors referring help-seekers to offline social services. The counsellor might also guide the help-seeker to find solutions on their own without providing explicit solutions. Keywords like "社工 (social worker)", "服務 (service)", "學校 (school)", "聯絡 (contact)" and "轉介 (referral)" were markers of counsellors' emphasis of their effort to help establish connections with external resources.

Stage 5 was the Wrap-up stage. This stage was the concluding stage of the session. The counsellor and the help-seeker reviewed the agreed-upon action steps, re-confirmed any follow-up sessions or referral services, and bid farewell to each other. Keyword "希望 (hope)" frequently appeared in phrases such as "希望今次既對話會令你覺得舒服" ("I hope this conversation has made you feel comfortable"), indicating that the counsellor aimed to conclude the session on a positive note. Keyword "休息 (rest)" in messages like "俾自己好好休息下 (Have a good rest)" suggested that the counsellor often encouraged the help-seeker to take time for self-care after the session, reinforcing the supportive and caring nature of the interaction. Keyword "謝謝 (thanks)" sent by the help-seeker implied the help-seeker's appreciation for the received support. Keyword "意見 (comment)" indicated that the counsellor invited the help-seeker to fill in a post-session survey and leave their suggestions. Keyword "傾到 (talk up to here)" indicated that the counsellor proposed to end the session.

Session stage progression analysis
We compared the time distribution (in minutes) between more satisfying sessions and less satisfying sessions. In order to explore how counsellors and help-seekers allocated their time through the stages, Viterbi algorithm[28] was applied to assign each session the most likely sequence of stages according to our HMM. We then computed the average duration percentage of each stage for both more satisfying and less satisfying sessions. As Shapiro-Wilk test results demonstrated that time allocated in each stage were all normally
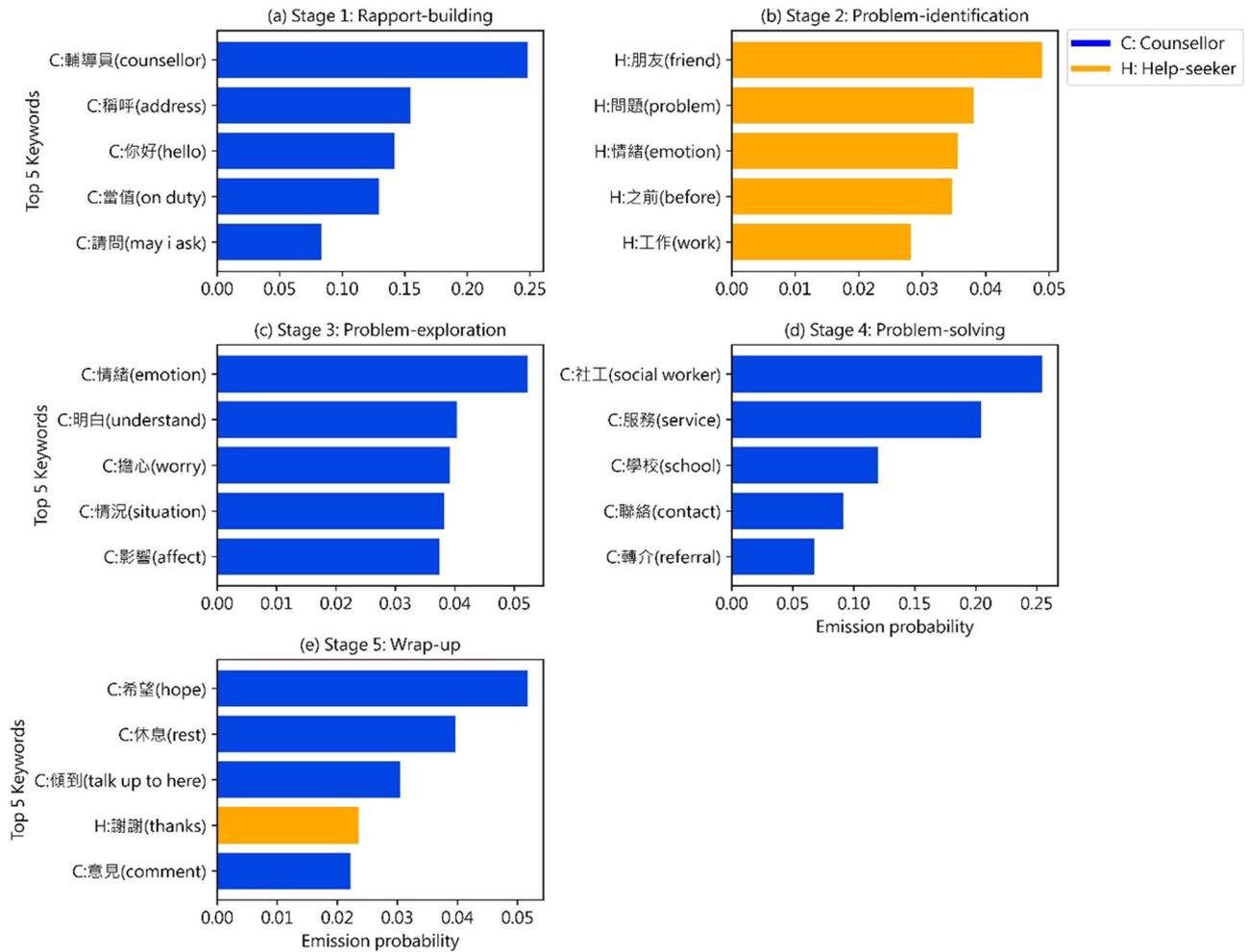
Fig. 4 | Top five keywords with the highest emission probabilities in each five stages. a Stage 1: rapport-building; **b** Stage 2: problem-identification; **c** Stage 3: problem-exploration; **d** Stage 4: problem-solving; **e** Stage 5: wrap-up.

distributed (see Supplementary Table 2), Welch's t-test was then performed to compare the differences in the two session groups.

As illustrated in Fig. 5, Stage 1 in both more satisfying (7.5% of session time) and less satisfying sessions (8.3% of session time) was quite efficient (Cohen $d = 0.07$, $p = 0.189$).

A relatively small portion of time was spent on Stage 2 (20.2% of session time) among more satisfying sessions. In contrast, less satisfying sessions had a larger portion of time in Stage 2 (40.7%, Cohen $d = 1.18$, $p<0.001$).

Among the more satisfying sessions, the vast majority of time was allocated to Stage 3 (28.5% of session time) and Stage 4 (46.6%). Less satisfying sessions were marked by a significantly larger allocation of time to Stage 3 (43.4%, Cohen $d = 0.79$, $p<0.001$), in contrast to a significantly less amount of time spent in Stage 4 (16.6%, Cohen $d = 1.56$, $p<0.001$).

Stage 5 in both more satisfying (8.2% of session time) and less satisfying sessions (6.3%) was relatively brief (Cohen $d = 0.11$, $p = 0.074$), suggesting a similarly concise wrap-up of the session.

The transition matrices (see Supplementary Table 3) depicted the probability of transitioning from one stage to another. In total, three transition probability matrices were trained, using all counselling sessions that met inclusion criteria, only more satisfying sessions, and only less satisfying sessions, respectively. In the transition matrices (see Supplementary Table 3), the hidden state name in each row represents the current stage and the hidden state columns refer to the next stages. The values in the table are the transition probabilities moving from the current stage to the next stage. For example, in the matrix for "All sessions", "transiting from Stage 1 to Stage 2"
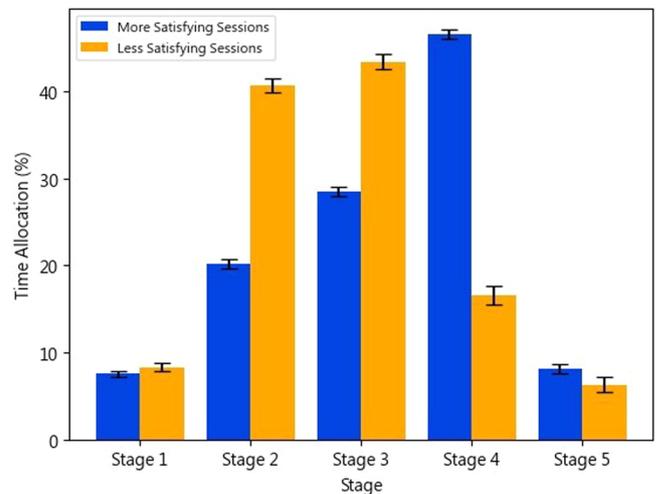

Fig. 5 | Average stage time duration percentage of more satisfying and less satisfying sessions.

has a probability of 0.457, illustrating that if the current message is in Stage 1, there was a probability of 0.457 for the next message to advance to Stage 2.

To provide a structured comparison, transition probability matrix of each stage from more satisfying and less satisfying counselling sessions were analysed and interpreted in Table 3.

Table 3 | Analysis and interpretation of transition matrices

Hidden state	Transition matrix from more satisfying sessions	Transition matrix from less satisfying sessions
Stage 1: Rapport-building	Has a probability of 5.3% of remaining in Stage 1, indicating a relatively more efficient start.	Has a probability of 21.4% of staying in Stage 1 and 57.0% transiting to Stage 2, indicating a relatively less efficient start.
Stage 2: Problem-identification	Has a high likelihood of progressing to Stage 4 (45.5%), demonstrating a tendency of quickly engaging in solution development after identifying the problem.	Often remains in the same stage (49.0%) or moves to Stage 3 (43.6%), indicating a thorough, albeit perhaps prolonged problem identification and exploration.
Stage 3: Problem-exploration	Exhibits a strong orientation towards solution development with a probability of 52.1% moving to Stage 4, reflecting efficient problem exploration.	Displays a quite high probability of reverting to Stage 2 (52.4%) and staying in Stage 3 (39.9%), pointing to possible unclarities of the problem and difficulties in conceiving any actionable solutions.
Stage 4: Problem-solving	Demonstrates a focus on developing solutions with a high likelihood (47.5%) of staying in Stage 4 and a need to frequently reassess the problem with a probability of 31.7% revisiting Stage 3.	Has a probability of 46.2% of staying in Stage 4 and a notable chance of revisiting Stage 2 (37.6%), suggesting efforts to maintain solution-focused discussion and a need to repeatedly clarify the problem.
Stage 5: Wrap-up	Always stays, marking the end of sessions	

Discussion

This study employed a Hidden Markov Model to analyse Cantonese text-based counselling sessions, revealing five distinct stages and uncovering important distinctions between more and less satisfying sessions in terms of the interaction patterns between the counsellor and the help-seeker. Our HMM approach offers concise stochastic modelling of complex interactions in counselling transcripts, enabling a deeper understanding of the structure and dynamics inherent in counselling sessions.

We identified five stages characterized by keywords within counselling sessions, which were rapport-building, problem-identification, problem-exploration, problem-solving, and wrap-up. These stages are consistent with prior frameworks[17,18]. A notable finding is that more satisfying sessions tended to spend more time in the problem-solving stage. As demonstrated in the results concerning time allocations among stages, in Stages 1 and 2, in more satisfying sessions the counsellor and the help-seeker quickly built rapport and moved beyond the initial introduction of the presenting problems. This efficiency allowed more time for later stages. In contrast, the counsellor and the help-seeker in less satisfying sessions spent more time in the problem-identification stage. It is possible that help-seekers' significantly higher psychological distress levels in less satisfying sessions, as reported in Table 1, might make them more reluctant or less able to disclose and articulate their presenting problems. The counsellor might thus spend more time helping them explore their emotional states and identifying their core concerns. In Stages 3 and 4, among more satisfying sessions, the counsellor and the help-seeker devoted relatively more time on working to understand the presenting problem and conceiving actionable strategies for problem-solving. Contrarily, in less satisfying sessions, there was a relative lack of identification and development of a clear and actionable solution. This might also have been contributed by baseline differences in level of psychological distress.

We also analysed the transition between stages. The transition matrix suggests that a more satisfying session usually had a relatively swift process of identifying and exploring problems and focused on problem-solving. As revealed by the three transition probability matrices (see Supplementary Table 3), more satisfying sessions exhibited a quicker transition from initial greetings and problem identification to problem exploration and demonstrated more efficient progress toward actionable strategies and solutions. In contrast, less satisfying sessions were characterized by prolonged problem identification and exploration, difficulties in effectively transitioning to solution development and the need to return to the problem-identification stage. The presenting issues of help-seekers in less satisfying sessions were likely more complex than those in more satisfying sessions. Additionally, it is possible that these help-seekers were dealing with multiple issues rather than just one, which can pose challenges for the counsellors to prioritise strategically the session goals and emphases. As a result, counsellors might be compelled to return to the problem-identification stage, indicating that the help-seekers either struggled to clearly articulate their issues or that their problems were multifaceted and required additional exploration.

The findings of distinct patterns of time allocation as well as stage transition probability could inform good practice of counselling and indicate help-seeker satisfaction. This findings could also be informative for training frontline counsellors to enhance service effectiveness and help-seeker satisfaction. For example, our findings can provide practical guidance for counsellors to manage session flow, particularly in moving through rapport-building and problem-identification in a timely manner and focusing on problem-solving. The identified patterns could be incorporated into training, helping both voluntary and professional counsellors be mindful of the within-session evolving foci and refine their techniques for managing session dynamics. Moreover, our findings could guide the development of digital tools such as AI models that monitor time allocation as well as stage transition and provide real-time feedback, assisting counsellors in making moment-to-moment decisions. Furthermore, our study provides insights to the overall structure of counselling sessions and the study of change processes. While existing frameworks often focused on the impacts of specific aspects of counselling (e.g., expressing empathy, providing solutions) on the counselling outcomes, this study contributes to the literature by examining the relationship between stage progression and help-seeker satisfaction, which allows counsellors and service providers to better understand how the overall structure of a counselling session, not just an individual component, could influence help-seeker experience. By quantifying such relationship, the study provides evidence-based suggestions for improving text-based counselling services.

Furthermore, the study's focus on assessing counselling transcripts by employing NLP probabilistic model filled a notable gap in existing literature, which primarily concentrated on analysing the counsellor's behaviour, required a large set of annotated data to validate theoretical frameworks, or applied non-interpretable black-box NLP models. By examining the sequential nature of text-based counselling sessions, our HMM model presents an efficient, reliable, and interpretable machine learning approach to capture and quantify the complex interactions constituting more satisfying counselling. Our method can be extended beyond the scope of counselling to analyse user behaviours in different types of text-based records, including user communications on social media, tutor-student interactions on tutoring systems, for example.

This study has several limitations. First, the dataset had a higher proportion of more satisfying sessions compared to less satisfying ones. There is a possibility of response bias, as help-seekers who had a more satisfying experience might be more likely to respond to the post-session survey compared with those who had a less satisfying experience. This response bias and data imbalance could potentially cause the HMM to produce biased results, leading to a model that more accurately represented the characteristics of more satisfying sessions. The underrepresentation of less satisfying sessions might skew the understanding of their specific dynamics. Future studies could consider employing stratified sampling techniques or targeted recruitment strategies to ensure a more balanced distribution of session types and minimize response bias.

Second, it should be noted that help-seekers' psychological distress levels, as measured by K6 score in the pre-session survey, might have contributed to session satisfaction. As exhibited in Table 1, help-seekers in less satisfying sessions had significantly higher pre-session K6 scores compared to those in more satisfying sessions. This suggests that the dynamics observed in more satisfying sessions might not solely reflect the quality of the session itself but could also be influenced by the psychological distress of the help-seeker. In addition, the quality of the session may also be impacted by the help-seeker's level of psychological distress. In other words, rather than merely on the differences between more versus less satisfying sessions, this study results possibly reflect differences between help-seekers of lower versus higher initial psychological distress levels. Although this study was not designed to assess the causal relationship between psychological distress and session satisfaction, it is crucial to acknowledge this limitation. Further exploration of the potential causal relationship between help-seekers' psychological distress levels and session satisfaction is warranted.

Third, the decision to set the third quartile as the threshold to remove words with low TF-IDF scores was based on our assessment; it was not grounded in a more systematic method. This limitation may affect the robustness of our word selection process, as some important words could have been excluded, while some trivial words may still remain. Future research could explore selecting a TF-IDF threshold for removing words based on a more data-driven approach.

Fourth, unlike state-of-the-art NLP models such as pre-trained language models (e.g., BERT, GPT), which can dynamically generate embeddings for unseen words based on context, the need to retrain the HMM with the introduction of new words posed challenges for real-time application. This limitation affects the model's adaptability and responsiveness to an evolving volume of counselling sessions and linguistic variations. Despite this limitation, HMMs remain an effective choice in tasks where interpretability and simplicity are prioritized, as HMMs provide transparency into the sequential structure of the data. Future research could explore different configurations of HMMs or integrate HMMs with deep learning models. For example, as we chose EM algorithm to estimate parameters, an interval estimation of the HMM parameters could be done in the future using alternative estimation methods, such as Markov Chain Monte Carlo (MCMC) techniques. In addition, hybrid models that integrate the strengths of HMMs with more advanced NLP techniques could potentially enhance real-time applicability while maintaining interpretability.

Fifth, another limitation of this study is the potential influence of issue complexity on the differences observed in time allocation and stage transition across counselling stages. While our findings suggest that more satisfying sessions saw higher efficiency in moving through earlier stages and allocated more time for problem-solving, we acknowledge that this claimed "efficiency" might be influenced by the complexity of the issues discussed during the sessions. Simpler problems presented by the help-seeker might require less time to disclose, allowing for quicker transitions to later stages, while more complex issues might take longer to articulate and explore, potentially prolonging the earlier stages of counselling. In addition, help-seekers might present more than one issue, which could add to the complexity of the counselling process. However, in our study, help-seekers reported their primary concern in the pre-session survey by responding to the question, "Which issue concerns you the most?" They were allowed to select only one issue from a list of 14 predefined categories (e.g., family relationship, school, work). This limited our understanding on the complexity of the presenting problems, making it difficult to accurately quantify or compare issue complexity between more and less satisfying sessions. Without a more detailed understanding of the full scope of the help-seeker's concerns, it is challenging to determine whether the observed differences in session progression were a result of the counselling process itself or merely reflective of the complexity of the issues being addressed.

Sixth, the findings were obtained from a model trained on a specific dataset from a particular online text-based Cantonese counselling service, which limits the generalizability of our findings. Moreover, due to the low response rate to the post-session survey and anonymity in Open Up, there was little demographic information about help-seekers. This limited our understanding of the help-seeker's background, potentially impacting the analysis and interpretation of counselling sessions. Future research could expand the model to include a broader range of linguistic and cultural contexts and enhance its applicability in diverse counselling settings. By continuing to refine and adapt NLP techniques to the counselling context, there is significant potential to improve the effectiveness of online counselling services and support mental health initiatives.

This study utilised an HMM approach in analysing Cantonese counselling transcripts from an online text-based counselling service, providing insights into counselling session structures and dynamics that can be illuminating to enhancing counselling effectiveness and improving user satisfaction. Notably, the differing patterns between more and less satisfying sessions in terms of session progressions offered valuable guidance for counsellor training and development, emphasizing more efficient rapport building and problem identification, thorough problem exploration, focused solution development, and concise wrap-up. This research not only helps to fill a gap in the existing literature on counselling session assessment but also paves the way for future explorations of utilising efficient, reliable, and interpretable probabilistic NLP techniques in counselling.

Data availability
The code of this study is available on GitHub: https://github.com/iharrisonfu/hmm-counseling.git. Given its sensitivity, the raw transcript data are not publicly available but are available from the corresponding author on reasonable request.

Code availability
The code of this study is available on GitHub: https://github.com/iharrisonfu/hmm-counseling.git.

References
1. Michael, L. Implications of outcome research for psychotherapy integration. In *Handbook of Psychotherapy Integration* (eds Norcross, J. C. & Goldfried, M. R.) 94–129 (Basic Books, New York, NY, 1992).
2. Berrouiguet, S., Baca-García, E., Brandt, S., Walter, M. & Courtet, P. Fundamentals for Future Mobile-Health (mHealth): A Systematic Review of Mobile Phone and Web-Based Text Messaging in Mental Health. *J. Med. Internet Res.* **18**, e135 (2016).
3. Yip, P. S. F. et al. The Opportunities and Challenges of the First Three Years of Open Up, an Online Text-Based Counselling Service for Youth and Young Adults. *Int. J. Environ. Res. Public. Health* **18**, 13194 (2021).
4. Pérez-Rosas, V., Wu, X., Resnicow, K., & Mihalcea, R. What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019) pp. 926–935.
5. Malgaroli, M., Hull, T. D., Zech, J. M. & Althoff, T. 'Natural language processing for mental health interventions: a systematic review and research framework'. *Transl. Psychiatry* **13**, 309 (2023).
6. Li, A. et al. Understanding Client Reactions in Online Mental Health Counseling. arXiv, http://arxiv.org/abs/2306.15334 (2023).
7. Liu, Z. et al. Listening to Mental Health Crisis Needs at Scale: Using Natural Language Processing to Understand and Evaluate a Mental Health Crisis Text Messaging Service. *Front. Digit. Health* **3**, 779091 (2021).
8. Sharma, A., Miner, A., Atkins, D. & Althoff, T. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2020) 5263–5276.

9. Althoff, T., Clark, K. & Leskovec, J. 'Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health'. *Trans. Assoc. Comput. Linguist.* **4**, 463–476 (2016).

10. Liu, S. et al. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: *Long Papers*) (eds Zong, C., Xia, F., Li, W. & Navigli, R.) (Association for Computational Linguistics, 2021) 3469–3483. https://doi.org/10.18653/v1/2021.acl-long.269.

11. Zhang, J. & Danescu-Niculescu-Mizil, C. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020) pp. 5276–5289.

12. Lee, Y., Goldwasser, D. & Reese, L. S. Towards Understanding Counseling Conversations: Domain Knowledge and Large Language Models. *arXiv*, http://arxiv.org/abs/2402.14200. (2024) Accessed: Feb. 28, 2024.

13. Beltagy, I. Peters, M. E. & Cohan, A. Longformer: The Long-Document Transformer. Available: https://arxiv.org/abs/2004.05150v2. (2020) Accessed: Jan 05, 2024.

14. Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, Almutiri, T. & Nadeem, F. Markov Models Applications in Natural Language Processing: A Survey. *Int. J. Inf. Technol. Comput. Sci*. **14**, 1–16 (2022).

15. Catarino, A. et al. Refining our understanding of depressive states and state transitions in response to cognitive behavioural therapy using latent Markov modelling. *Psychol. Med*. **52**, 332–341 (2022).

16. Mor, B., Garhwal, S. & Kumar, A. 'A Systematic Review of Hidden Markov Models and Their Applications'. *Arch. Comput. Methods Eng.* **28**, 1429–1448 (2021).

17. Hill, C. E. *Helping skills: Facilitating, exploration, insight, and action*, 3rd ed. (American Psychological Association, 2009), pp. xix, 430.

18. Hackney, H.L. & Bernard, J. M. *Professional Counseling: A Process Guide to Helping*, 8th ed. (Pearson, 2016).

19. Bidwai, S., Mayannavar, S. & Wali, U. V. Performance Comparison of Markov Chain and LSTM Models for Spectrum Prediction in GSM Bands. In *Machine Learning for Predictive Analysis*, (eds Joshi, A., Khosravy, M. & Gupta, N.) 289–298 (Springer, 2021).

20. Sangganjanavanich, V. F. & Reynolds, C. *Introduction to Professional Counseling* (SAGE Publications, 2013).

21. Kessler, R. C. et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol. Med*. **32**, 959–976 (2002).

22. Kessler, R. C. et al. Screening for Serious Mental Illness in the General Population. *Arch. Gen. Psychiatry* **60**, 184 (2003).

23. Xu, Y. et al. Detecting premature departure in online text-based counseling using logic-based pattern matching. *Internet Interv*. **26**, 100486 (2021).

24. Sun J. Jieba. GitHub URL: https://github.com/fxsjy/jieba (2021).

25. Lee, J. L., Chen, L., Lam, C., Lau, C. M. & Tsui, T.-H. PyCantonese: Cantonese Linguistics and NLP in Python. In *Proceedings of LREC* (2022).

26. Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Stat.* **41**, 164–171 (1970).

27. Moser, B. K. & Stevens, G. R. Homogeneity of Variance in the Two-Sample Means Test. *Am. Stat.* **2024**, 19–21 (1992).

28. Forney, G. D. The viterbi algorithm. *Proc. IEEE* **61**, 268–278 (1973).

Acknowledgements

Author contributions

FU Z. was responsible for the conceptualization and data management work. FU Z. and Hsu Y.C. pre-processed the data, trained the HMM, conducted the statistical analysis, and interpreted the results. Chan C.S., Liu J. and Yip P.S.F. arranged the resources for this study. FU Z. and Hsu Y.C. wrote the original draft. FU Z., Hsu Y.C., Liu J., Chan C.S. and Yip P.S.F. reviewed and edited the manuscript. This study was supervised by Chan C.S. and Yip P.S.F. All authors had full access to the data in this study and had final responsibility for the decision to submit for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44184-024-00103-9.

Correspondence and requests for materials should be addressed to Christian S. Chan or Paul S. F. Yip.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.