



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Agent-driven big data interaction empowers personalised marketing efficiency

Xiaojun Cui

DOI: [10.1504/IJICT.2026.10076594](https://doi.org/10.1504/IJICT.2026.10076594)

Article History:

Received:	25 November 2025
Last revised:	21 December 2025
Accepted:	24 December 2025
Published online:	10 March 2026

Agent-driven big data interaction empowers personalised marketing efficiency

Xiaojun Cui

Business School,
Zhengzhou College of Finance and Economics,
Zhengzhou, 450000, China
Email: cuixiaojun2025@sina.com

Abstract: This paper proposes a cognition-aware hierarchical reinforcement learning framework to solve the decision-making inefficiency problem caused by ignoring the user's cognitive status in traditional personalised marketing systems. By formalising cognitive load theory as a reward function and constructing an agent that collaborates metacognition and executive layer, we achieve dynamic optimisation of user cognitive experience while accurately recommending. Experiments based on the real user behaviour dataset of Alibaba Mobile show that the proposed model is significantly better than the state-of-the-art baselines such as deep interest evolution network and deep reinforcement learning for news recommendation in key indicators such as normalised discounted cumulative gain at 10 and conversion rate ($p < 0.01$). Ablation experiments further verify the necessity of cognitive components and hierarchical architecture.

Keywords: personalised marketing; cognitive load theory; CLT; hierarchical reinforcement learning; agent; user experience.

Reference to this paper should be made as follows: Cui, X. (2026) 'Agent-driven big data interaction empowers personalised marketing efficiency', *Int. J. Information and Communication Technology*, Vol. 27, No. 17, pp.71–92.

Biographical notes: Xiaojun Cui is an Associate Professor at the Zhengzhou College of Finance and Economics. She received her Bachelor's in Management from the Henan Normal University (2004) and Master's in Economics from the Henan University (2012). Her research interests include digital marketing, digital economy, consumer behaviour, machine learning, and corporate customer relationships.

1 Introduction

1.1 Research background and challenges of the times

We live in a digital age defined by a flood of data and a scarcity of user attention. As a core strategy to deal with this contradiction, personalised marketing has evolved from simple demographic based segmentation to ultra-fine-grained user profiling and prediction relying on deep learning and big data analytics (Osarenkhoe and Bennani, 2007).

In recent years, especially deep neural network models, such as deep interest network (DIN) and its evolved versions (Yu et al., 2022), have achieved remarkable results in key indicators such as click-through rate (CTR) and conversion rate (CVR) prediction by capturing users' dynamically changing interests. These models represent the current frontier of recommender systems and personalised marketing. They can process serialised user behaviour data and mine complex nonlinear patterns from it.

However, most of these methods still belong to a 'static' or 'passive' prediction paradigm in nature – they are good at predicting what users may like based on historical data, but generally lack the ability to actively manage and shape the user's cognitive experience during dynamic interaction. With the aggravation of information overload, decision fatigue and cognitive overload faced by users when interacting with digital platforms have become the bottleneck that hinders the further improvement of marketing efficiency (Bonneton and Rahwan, 2020). This raises a fundamental question: where will the competitive advantage of the next generation of personalised marketing come from when all competitors have similar data and predictive technology? The core bottleneck is user decision fatigue and cognitive overload. Traditional personalised marketing systems primarily rely on predicting preferences from historical behaviour. They fundamentally lack the mechanism to perceive, quantify, or adapt to a user's real-time cognitive state. This inability to dynamically manage cognitive load during interactions leads to inefficiency, as users become overwhelmed, resulting in abandoned decisions and reduced long-term engagement, which curtails marketing effectiveness.

1.2 Theoretical limitations of existing paradigms

The limitations of current mainstream approaches are rooted in their single computational paradigm perspective. They treat the user as a signal source that produces a sequence of actions, while ignoring its nature as a complex decision maker with limited cognitive resources. This limitation stands out in two ways: First, in terms of methodology, although reinforcement learning (RL) has been introduced to model sequential decision problems, such as in some works utilising deep Q-network (DQN) or policy gradient methods (Afsar et al., 2022) for news or product recommendation, But most of these agents are designed to optimise for a single, long-term business reward (such as total clicks or total purchases). They lack explicit mechanisms to sense and optimise the cognitive state of the user in the decision-making process, such as information perplexity, choice conflict, or decision effort. Second, a more profound lack of theoretical foundation is that personalised marketing research has long been disconnected from the mature theories of human-computer interaction (HCI) and cognitive science.

Cognitive load theory (CLT) was systematically elaborated by Sweller (2011). Its core viewpoint is that human's working memory capacity is limited, and when cognitive load exceeds individual processing capacity caused by external information presentation, learning and decision-making efficiency will decline sharply. This theory and its derivative methods have been repeatedly verified to effectively improve learning efficiency in the field of educational technology and interface design (Sweller et al., 2019). However, in the field of personalised marketing, which is also full of information processing and decision-making tasks, the application of CLT has been almost completely neglected. Few existing marketing interaction systems actively evaluate the cognitive load of users and dynamically adjust the complexity, order, or amount of information presentation. This lack of interdisciplinary theory has left existing systems

with a ceiling on intelligence – they can ‘guess what you want’ but fail to ‘know what you can’, unable to provide the most complex choices when the user’s cognitive resources are most abundant, or actively simplify the process when the user’s decision-making is fatigued. The lack of this ability is the key barrier restricting the marketing efficiency from ‘precision’ to ‘intelligence’. CLT provides a key basis because it formally describes how limited working memory impacts task performance. While foundational in instructional design, its application in marketing has been minimal. The primary difference in application lies in objective: in HCI, CLT often optimises static interfaces for learning; in marketing, it must dynamically guide the content, timing, and complexity of interactive sequences to align with the user’s fluctuating cognitive capacity for decision-making.

1.3 Agent-driven interactions

To overcome these limitations, we must conceive of a new paradigm of marketing interaction. This paradigm should not merely be a more accurate predictive model, but rather an intelligent agent embedded within the marketing environment, capable of executing a sense-decide-act cycle. The objective function of this agent should be extended beyond a single business metric to encompass a multi-objective optimisation framework that incorporates the user’s cognitive state. In recent years, developments in the field of general artificial agents, especially advances in hierarchical RL, have made it technically feasible to build such complex agents (Sutton et al., 1999). HRL allows an agent to plan coarse-grained goals at a high level, (e.g., ‘stimulate exploration interest’ or ‘facilitate quick decisions’) and perform concrete actions at a low level, (e.g., ‘display a highly diverse set of items’ or ‘highlight a time-limited offer’) by introducing a hierarchy of abstract behaviours.

Combining this hierarchical architecture with cognitive theory, we are able to construct a metacognitive agent. The agent’s high-level policy is specifically responsible for inferring the user’s current cognitive state, (e.g., whether it’s in a low-load exploration phase or a high-load decision phase) and communicating this as a kind of ‘meta-action’ to the low-level policy. The bottom level strategy is based on the high-level meta-action to specifically implement personalised marketing reach. This architecture essentially builds a bridge connecting micro user cognitive theory and macro marketing interaction strategy. It makes the agent’s behaviour no longer mechanically responding to data patterns, but based on a fundamental respect for and adaptation to human information processing. As pointed out by Warner (1969) in his seminal work, wealth creation in the attention economy largely depends on the efficient allocation of limited attention. A marketing agent that can dynamically manage the user’s cognitive load is the direct practitioner under this concept. It optimises the allocation of user’s cognitive resources, and finally realises the qualitative leap of marketing efficiency. The essential architectural difference is the shift from a static, one-way prediction model to an active, closed-loop agent embedded within the environment. In objective function, traditional methods target immediate business metrics like CTR. Our paradigm employs a multi-objective reward function that explicitly integrates cognitive state optimisation alongside business goals, enabling the system to manage the entire interaction process for sustainable engagement rather than just predicting the next click.

1.4 Research orientation and core issues

Based on the above analysis, the research positioning of this paper is to propose and construct an agent-driven big data interaction framework by integrating the CLT in cognitive science and HRL in artificial intelligence, in order to achieve fundamental efficiency improvements in real-world marketing scenarios. This study aims to answer the following core research questions: First, how to objectively and quantitatively define and measure the traditionally subjective ‘cognitive load’ in a dynamic and data-driven marketing interactive environment? Second, how to design a hierarchical agent architecture so that its high-level can effectively recognise the user’s cognitive state and guide the low-level strategy to perform cognitive friendly marketing actions? Third, can this new paradigm, which integrates cognitive theory, significantly improve user interaction experience and long-term engagement while maintaining business effectiveness compared with the traditional SOTA method that only optimises business metrics? The exploration of these issues is not only to develop a new technological tool, but also to promote the paradigm shift of personalised marketing from ‘data-centric’ to ‘people-centric’, and lay the theoretical foundation for building a more sustainable and user-friendly digital business environment.

2 Relevant technologies

2.1 Evolution and limitations of personalised marketing algorithms

The core of personalised marketing is to accurately predict user preferences and behaviour intentions from massive user data. This field has moved from the early days of collaborative filtering and matrix factorisation to the classic stage represented by Ling and Yen (2001) customer relationship management (CRM) framework, and has entered a new era dominated by deep learning. Through its powerful nonlinear fitting ability, deep learning methods can automatically learn the potential characteristics of users and items, thus significantly improving the prediction accuracy. Among them, the DIN proposed by Xu et al. (2021) can adaptively learn the correlation between users ‘historical behaviours and target ads by introducing local activation units, and realise the modelling of users ‘diverse interests. Since then, the deep interest evolution network (DIEN) proposed by Lu et al. (2021) further captures the dynamics of user interests by simulating the changes of interests over time, which represents a significant progress in the field of sequence recommendation.

However, although these models perform well on static CTR prediction tasks, they are essentially fits to historical data and lack the ability to actively and sequentially interact with the environment. They treat each recommendation as an independent event or at best a point on a time series, and fail to treat it as a continuous decision process that can be optimised through policies. This paradigm limits the ability of the system to dynamically adjust and explore according to the user’s real-time feedback during the interaction process, which makes the system appear rigid and passive in the face of complex and variable user decision paths.

2.2 *The application and absence of CLT in HCI*

Since being systematised by Self and Education (2006) in his founding review, CLT has become a central framework for understanding human cognitive processing in learning and complex tasks. According to the theory, the limited capacity of working memory is the key to cognitive effectiveness, and excessive external load (caused by improper information presentation) can hinder learning and decision-making. In the field of educational technology, this theory has been widely applied and verified. For example, in the design of multimedia learning and intelligent tutoring systems, managing cognitive load by optimising information presentation has been proved to effectively improve learning outcomes (Schnotz and Kürschner, 2007). Paas and Van Merriënboer (2020) further emphasise the importance of measuring and optimising cognitive load in complex skill training.

However, there is a surprising lack of research on transferring this mature theory to the high-value application scenario of personalised marketing. Existing marketing interaction designs are mostly based on rules of thumb or A/B testing, lacking a unified theory to guide how to design ‘cognitive friendly’ recommendation sequences. Although a few studies have begun to focus on the ‘cognitive factors’ in UR, such as Bahmanzari and Odom (2015), which explored the impact of information framing on consumer decision-making efforts, most of these works remain at the descriptive and correlation level. Failure to embed cognitive load into the dynamic interaction model as a quantifiable core variable that can be optimised in real-time by intelligent systems. As a result of this lack of application of theory, current recommender systems have made progress in ‘accuracy’, but have fallen short in ‘accuracy’, that is, providing information at the most appropriate time and in the most tractable way.

2.3 *The potential and challenges of agent systems and RL in recommendation*

To overcome the limitations of traditional static models, RL is introduced into the field of recommender systems, which is modelled as a sequential decision problem. While early attempts applied contextual gambling machine algorithms to news recommendation, the rise of deep reinforcement learning (DRL) has led to more powerful solutions. The DQN proposed by Mnih et al. (2015) demonstrates the ability to learn successful policies directly from high-dimensional perceptual inputs, laying the foundation for handling complex representations of user states. Subsequently, the Asynchronous Advantage Actor-Critic (A3C) algorithm proposed by Silver et al. (2016), and the subsequent policy gradient method, provides a more stable training path for dealing with high-dimensional discrete and continuous action Spaces. In the field of recommendation, researchers have explored the use of DRL to optimise long-term user engagement. As described in their review by Afsar et al. (2022), these approaches aim to go beyond instant clicks and focus on the lifetime value of users.

However, existing RL recommendation agents face a core challenge: policy flattening. Most models learn a single, mapping policy from the original state to the original action. This structure is inefficient and difficult to interpret when facing complex tasks such as marketing interaction that require decisions on different time scales. For example, deciding whether the user currently needs to be guided to explore new products (high-level strategy) is a different level of decision than deciding which three products to show (low-level tactical), but is often confused. HRL provides a framework for solving

this problem by introducing temporal abstraction, which, as illustrated by Sutton et al. (1999) in their seminal work on temporal abstraction, allows agents to plan and make decisions at multiple levels. Despite the success of HRL in areas such as robot control, combining it with cognitive theory for building marketing agents that can actively manage the cognitive state of users is still a frontier that has not been fully explored. The existing research either focuses on the improvement of HRL algorithm purely driven by technology, or is limited to the optimisation of commercial indicators, failing to take the human cognitive model as the core basis for the high-level strategy design of agents, which is the key research gap that this paper aims to fill.

3 Methodology

This paper proposes a cognition-aware HRL framework, which aims to fundamentally improve the efficiency of personalised marketing by building a marketing agent that can dynamically infer and optimise the user's cognitive state. The core innovation of this framework lies in the organic integration of classical CLT and advanced HRL, so as to embed the understanding of human cognitive processes into the technical model.

3.1 Problem formalisation: a hierarchical decision process for marketing interactions

We model the serialised marketing interaction as a hierarchical Markov decision process (MDP), which is defined as tuples: $(\mathcal{S}, \mathcal{Z}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{Z}$ is the high-level meta-action space (cognitive state space), $\mathcal{A}$ is the low-level concrete action space (recommended action space), $\mathcal{P}$ is the state transition probability, $\mathcal{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. The complete composition of the state $s_t \in \mathcal{S}$ at time t is as follows:

- User historical behaviour embedding $\mathbf{h}_t$: the sequence of the user's L recent interaction (click, purchase, etc.) $H_t = i_1, i_2, \dots, i_L$ is converted into a sequence of vectors via an item embedding matrix $\mathbf{E} \in \mathbb{R}^{|I| \times d}$ and aggregated via an attention network (Vaswani et al., 2017) to capture the importance of different historical actions to the current state. This process can be formalised as:

$$\mathbf{e}_i = \mathbf{E} \cdot \text{one-hot}(i) \quad (1)$$

$$\mathbf{h}_t = \text{Attention}(\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \dots, \mathbf{e}_{i_L}) \quad (2)$$

where d is the embedding dimension and $|I|$ is the total number of items.

- Context feature vector $\mathbf{c}_t$: includes real-time information such as time, device, and network environment.
- User static profile $\mathbf{d}$: anonymised demographic characteristics.
- Last moment cognitive state $\mathbf{z}_{t-1}$: provides memory for high-level strategies.

Thus, the full state is denoted by:

$$s_t = \text{Concat}(\mathbf{h}_t, \mathbf{c}_t, \mathbf{d}, z_{t-1}) \quad (3)$$

The state space comprises four integrated elements: user behaviour embedding, context features, static portraits, and historical cognitive states. These components work together by providing a layered representation: behaviour and context capture the immediate scenario, the static portrait offers persistent user traits, and the cognitive history tracks the evolving mental resource level, enabling a holistic and temporally-aware understanding of the user. The action space is hierarchical. The high-level policy π_{high} selects an abstract cognitive state label z_t from the meta-action space $\mathcal{Z} = \{z^1, z^2, \dots, z^K\}$, such as z^1 for ‘explore cognitive state’, z^2 for ‘evaluate cognitive state’, and z^3 for ‘decision cognitive state’. The underlying strategy π_{low} selects a specific commodity subset $a_t \subset \mathcal{A}$ from the total commodity set for recommendation according to z_t and s_t , where $|a_t| = N$ is the number of commodities recommended each time. This hierarchical structure implements temporal abstraction and allows high-level policies to be planned at a more macro level.

3.2 Agent architectures: the synergy of metacognition and execution

Our agent adopts a hierarchical architecture under the actor-critic framework, consisting of a high-level metacognitive policy network and a low-level executive Q-network.

- High-level metacognitive actor network (π_{high}): this network acts as a cognitive state recogniser whose function is to map the observed user state s_t onto a probability distribution over cognitive states. We employ a long short-term memory (LSTM) (Graves, 2012) to capture the temporal evolution of the user’s cognitive state. The update and output of the network are as follows:

$$\mathbf{m}_t = \text{LSTM}(s_t, \mathbf{m}_{t-1}; \phi_m) \quad (4)$$

$$\mathbf{o}_t = \text{MLP}_{\text{high}}(\mathbf{m}_t; \phi_o) \quad (5)$$

$$\pi_{\text{high}}(z_t | s_t) = \text{Softmax}(\mathbf{o}_t) \quad (6)$$

where $\mathbf{m}_t$ is the hidden state of LSTM, which serves as the system’s internal memory for the user’s cognitive history. ϕ_m and ϕ_o are high-level network parameters; MLP_{high} is a multilayer perceptron. The goal of the strategy is to learn which cognitive state to guide the user into under which context maximises the long-term benefit.

- Underlying executive critic network (Q_{low}): the underlying network is a DQN (Mnih et al., 2015) to evaluate the long-term value of performing a specific recommended action a_t given a state s_t and a cognitive state z_t . Its architecture is designed as follows:

$$\mathbf{x}_t = \text{Concat}(s_t, \text{Embedding}(z_t)) \quad (7)$$

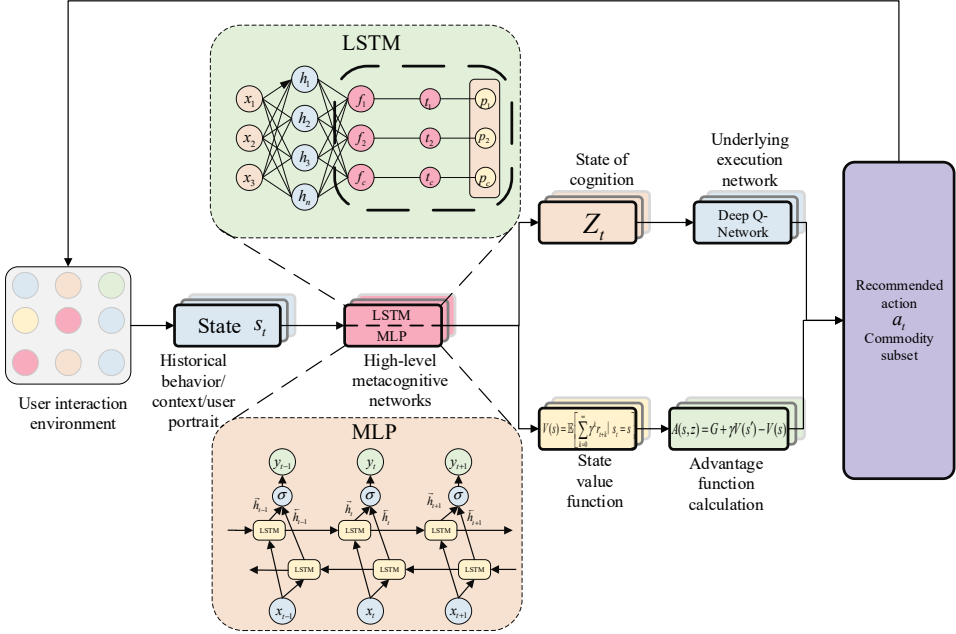
$$Q_{\text{low}}(s_t, z_t, a; \theta) = \text{MLP}_{\text{low}}(\mathbf{x}_t; \theta)_a \quad (8)$$

where Embedding transforms the discrete meta-action z_t into a dense vector, and the output layer dimension of MLP_{low} is $|\mathcal{A}|$, that is, each candidate action (item) has a corresponding q-value. The underlying policy selects actions in a greedy way:

$$a_t = \arg \max_{a \in \mathcal{A}} Q_{\text{low}}(s_t, z_t, a; \theta) \quad (9)$$

This design ensures that the low-level concrete recommended actions are always constrained and guided by the high-level abstract cognitive policies.

Figure 1 CA-HRL agent architecture (see online version for colours)



As shown in Figure 1, our proposed agent architecture forms a complete perception-decision-action closed loop. The most significant feature of the architecture is the double-loop decision-making mechanism. The inner loop is composed of high-level metacognitive network, which is responsible for continuously inferring and guiding the user's cognitive state. The outer loop consists of the underlying execution network, which generates specific recommended actions under the constraints of high-level policies. This hierarchical design enables agents to tackle decision problems at different time scales simultaneously – the high-level network focuses on cognitive state transitions across multiple interaction steps, while the low-level network focuses on optimal action selection within a single-step interaction. The metacognitive network performs high-level inference, using structures like LSTM to analyse the state and assign a cognitive mode label. The execution network then performs the concrete task of item recommendation based on this label. Their collaboration creates temporal abstraction: the meta-network sets sustained goals over multiple steps, while the execution network handles granular actions, allowing for coordinated decision-making across different time scales.

The data flow in the architecture clearly shows how the components work together. The state information s_t generated by the user's interaction with the environment is fed

into both high-level and low-level networks. The cognitive state z_t output by the high-level network is passed to the low-level network as context prior, which significantly reduces the space for low-level action search and improves the learning efficiency. At the same time, the reward signals from environmental feedback are synthesised through a carefully designed multi-objective reward function, which simultaneously optimises business value and cognitive experience. In particular, it is worth noting that the state-value function $V(s)$ provides an important benchmark signal for the optimisation of high-level policies, which enables metacognitive decision making to accurately evaluate the long-term value of different cognitive state-guided policies.

The advantage of this architecture design lies in its unification of theoretical self-consistency and engineering realisability. From a theoretical point of view, it transforms the core idea of CLT – managing rather than increasing the user’s cognitive resource consumption – into a computable learning goal. From an engineering point of view, the hierarchical architecture naturally supports parallel training and distributed deployment, which provides technical feasibility for large-scale online applications. In the following sections, we elaborate on the multi-objective reward function design that drives the learning of this agent.

3.3 Reward function design: balancing business effectiveness and cognitive experience

The reward function is the key to guide the agent to learn the desired behaviour. We design a multi-objective reward function that aims to balance immediate business conversion with long-term user experience (UX). The immediate reward r_t at each step consists of a three-part linear weighting:

$$r_t = \lambda_1 \cdot r_{\text{business}}(s_t, a_t) + \lambda_2 \cdot r_{\text{engagement}}(s_t, a_t) - \lambda_3 \cdot r_{\text{overload}}(s_t, a_t) \quad (10)$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ is a hyperparameter that adjusts the relative importance of each objective.

- Business reward (r_{business}): a direct measure of the business value of a marketing action. We define it as a step function based on immediate user feedback:

$$r_{\text{business}} = \begin{cases} r_{\text{purchase}} & \text{if purchase occurs} \\ r_{\text{click}} & \text{if click occurs} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $r_{\text{purchase}} > r_{\text{click}} > 0$.

- Engagement rewards ($r_{\text{engagement}}$): this is used to encourage actions that lead to deeper engagement, even if they do not directly translate. We use session depth, the number of consecutive interactions a user has, as a proxy metric:

$$r_{\text{engagement}} = \eta \cdot \log(1 + \text{SessionDepth}(t))$$

where η is a scaling factor and $\text{SessionDepth}(t)$ is the current session length up to time t .

- Cognitive overload penalty (r_{overload}): this is the theoretical core of this study, which aims to quantify and minimise the cognitive load endured by users during their decision-making process. Based on the CLT proposed by Medin (2001), we model it as a function of information complexity and decision pressure:

$$r_{\text{overload}} = \alpha \cdot \text{InfoComplexity}(a_t) + \beta \cdot \text{DecisionConflict}(s_t, a_t)$$

- Information complexity: measures the information density and diversity of the recommendation list a_t itself. Let's use the entropy of the features in the list of recommendations:

$$\text{InfoComplexity}(a_t) = - \sum_{f \in \mathcal{F}} P(f|a_t) \log P(f|a_t) \quad (12)$$

where $\mathcal{F}$ is the set of product features, (e.g., category, price range), and $P(f|a_t)$ is the empirical probability that the feature f appears in the recommendation list a_t . Too high entropy means that the list content is cluttered, which may increase the cognitive burden of users.

- Decision conflict: measure how well the new recommendation a_t matches the current cognitive state s_t of the user. We use the KL divergence between the recommendation list and the user's historical interests as a measure:

$$\text{DecisionConflict}(s_t, a_t) = D_{KL}(P(\cdot|a_t), \parallel, P(\cdot|h_t)) \quad (13)$$

where $P(\cdot|h_t)$ is the interest distribution inferred from the user history $\mathbf{h}_t$. Too high KL divergence means that the recommendation is seriously inconsistent with user expectations, which may lead to decision confusion. The coefficients α and β control the strength of the penalties for these two parts. The penalty quantifies cognitive load through two computable metrics: information complexity, measured by the entropy (disorder) of the recommendation list, and decision conflict, measured by the KL divergence between the recommendation distribution and the user's interest profile. This design directly reflects CLT's core idea by operationalising the abstract concept of 'extraneous cognitive load' into a tangible, minimisable component within the optimisation objective.

3.4 The training algorithm: optimisation on two timescales

We adopt a dual-timescale asynchronous optimisation algorithm based on experience replay. The experience generated by the agent's interaction with the environment, including high-level experience (s_t, z_t, G_t, s_{t+K}) and low-level experience $(s_t, z_t, a_t, r_t, s_{t+1})$,

is stored in separate replay buffers $\mathcal{D}_{\text{high}}$ and $\mathcal{D}_{\text{low}}$ where $G_t = \sum_{k=0}^{K-1} \gamma^k r_{t+k}$ is the

discounted cumulative reward within K steps from time t , which is used as the delayed reward of the high-level policy.

- Underlying network update: the parameters θ of the underlying critic network Q_{low} are learned by minimising the temporal difference (TD) error. The loss function is:

$$\mathcal{L}_{\text{low}}(\theta) = \mathbb{E}_{(s, z, a, r, s') \sim \mathcal{D}_{\text{low}}} \left[(y_{\text{low}} - Q_{\text{low}}(s, z, a; \theta))^2 \right] \quad (14)$$

where the target value y_{low} is computed by the target network (parameter θ^-):

$$y_{\text{low}} = r + \gamma \max_{a'} Q_{\text{low}}(s', z, a'; \theta^-) \quad (15)$$

The parameter θ is updated via gradient descent:

$$\theta \leftarrow \theta - v_{\text{low}} \nabla_{\theta} \mathcal{L}_{\text{low}}(\theta) \quad (16)$$

where v_{low} is the underlying learning rate.

- **High-level network update:** the parameters ϕ of the high-level metacognitive actor network π_{high} are optimised using a policy gradient approach. The objective is to maximise the expected cumulative reward $J(\phi) = \mathbb{E}_{\pi_{\text{high}}} \left[\sum_t \gamma^t G_t \right]$. We use the value function from the underlying critic as a baseline to reduce the variance. The gradient of the high-level policy is estimated as follows:

$$\nabla_{\phi} J(\phi) \approx \mathbb{E}_{(s, z, G, s') \sim \mathcal{D}_{\text{high}}} \left[\nabla_{\phi} \log \pi_{\text{high}}(z | s; \phi) \cdot (G + \gamma V(s') - V(s)) \right] \quad (17)$$

where $V(s) = \mathbb{E}_{a \sim \pi_{\text{low}}} [Q_{\text{low}}(s, z, a)]$ is the state value function, again approximated by a critic network. The parameter update is: $\phi \leftarrow \phi + v_{\text{high}} \nabla_{\phi} J(\phi)$, where v_{high} is the high-level learning rate, and usually $v_{\text{high}} < v_{\text{low}}$, reflecting the slower update of the high-level policy.

Through the asynchronous update of the whole training process, the high-level strategy can learn a stable cognitive state division, while the bottom strategy can quickly adapt and perform specific recommendation tasks. The two strategies cooperate to achieve the joint optimisation of marketing efficiency and UX.

The above theoretical framework is implemented by the following algorithm, which completely outlines the training life cycle of our proposed cognitive-aware hierarchical reinforcement learning (CA-HRL) model.

Algorithm 1 Cognitive-aware hierarchical RL training

Input: Dataset $\mathcal{D}$, learning rates $\alpha_{\text{high}}, \alpha_{\text{low}}$, discount γ , update rate τ , high-level interval K

Output: Trained networks $\pi_{\text{high}}, Q_{\text{low}}$

- 1: **Initialise** network parameters $\theta_{\pi}, \theta_Q, \theta_{Q'} \leftarrow \theta_Q$, replay buffers $\mathcal{B}_{\text{high}}, \mathcal{B}_{\text{low}}$
- 2: **for** episode = 1 to M **do**
- 3: Sample user session from $\mathcal{D}$, initialise state s_0 , LSTM state h_0
- 4: **for** $t = 0$ to $T - 1$ **do**
- 5: // **High-level meta-cognitive decision (every c steps)**
- 6: **If** $t \bmod c = 0$ **then**
- 7: Sample meta-action $m_t \sim \pi_{\text{high}}(s_t, h_t)$
- 8: **else**
- 9: $m_t \leftarrow m_{t-1}$ // Reuse previous meta-action
- 10: **end if**
- 11: // **Low-level execution decision**
- 12: Select action $a_t = \arg\max_a Q_{\text{low}}(s_t, m_t, a)$

```

13:      // Environment interaction
14:      Execute  $a_t$ , observe reward  $r_t$  and next state  $s_{t+1}$ 
15:      Compute multi-objective reward:  $r_t = \lambda_b \cdot r_{\text{business}} + \lambda_e \cdot r_{\text{engagement}} - \lambda_c \cdot r_{\text{cognitive}}$ 
16:      // Store experiences
17:      Store  $(s_t, m_t, a_t, r_t, s_{t+1})$  in  $\mathcal{B}_{\text{low}}$ 
18:      if  $(t + 1) \bmod c = 0$  then
19:          Compute  $c$ -step return:  $R_t = \sum_{i=0}^{c-1} \gamma^i r_{t+i}$ 
20:          Store  $(s_{t-c+1}, m_{t-c+1}, R_t, s_{t+1})$  in  $\mathcal{B}_{\text{high}}$ 
21:      end if
22:      // Update LSTM hidden state
23:       $h_{t+1} = \text{LSTM}(s_t, h_t)$ 
24:      // Low-level network update (fast timescale)
25:      if  $|\mathcal{B}_{\text{low}}| > \text{batch\_size}$  then
26:          Sample batch from  $\mathcal{B}_{\text{low}}$ 
27:          Compute target:  $y = r + \gamma \cdot \max_{a'} Q'(s_{t+1}, m_t, a')$ 
28:          Update  $\theta_Q$  to minimise  $(y - Q(s_t, m_t, a_t))^2$ 
29:      end if
30:      // High-level network update (slow timescale)
31:      if  $(t \bmod c = 0)$  and  $|\mathcal{B}_{\text{high}}| > \text{batch\_size}$  then
32:          Sample batch from  $\mathcal{B}_{\text{high}}$ 
33:          Compute advantage:  $A = R_t + \gamma^c \cdot V(s_{t+1}) - V(s_t)$ 
34:          Update  $\theta_\pi$  using policy gradient:  $\nabla J \approx A \cdot \log \pi_{\text{high}}(m_t | s_t)$ 
35:      end if
36:      // Target network update
37:      Update  $\theta_{Q'} \leftarrow \tau \cdot \theta_Q + (1 - \tau) \cdot \theta_{Q'}$ 
38:  end for
39:  end for

```

In the above algorithms, the high-level decision interval K is a key hyperparameter, which balances the stability and agility of metacognitive strategies. In the underlying network update, the action value function is approximated by TD learning, and the target network is used to provide a stable learning objective, which is the key technology to solve the problem of training instability of nonlinear function approximators (Mnih et al., 2015). In high-level network updates, the dominance function $A(s, z) = G + \gamma^K V(s') - V(s)$ is used to evaluate the relative dominance of meta-actions, reducing the variance of policy gradients. Finally, the soft update mechanism further ensures the stability of the training process. Through this asynchronous update of the double time scale, the whole process ensures that the high-level strategy and the bottom strategy can co-evolve, and finally forms an agent that can have insight into the user's cognitive state and perform precision marketing. This algorithm ensures coordination by employing different update frequencies: the high-level policy updates slowly, providing stable, long-term guidance

on cognitive state management. The low-level policy updates rapidly to adapt to immediate feedback. This asynchronous mechanism, coupled with techniques like experience replay, prevents destabilising interference between policy levels, which is crucial for training convergence in complex hierarchical models.

Table 1 Model hyperparameter system configuration and sensitivity analysis

<i>Parameter categories</i>	<i>Parameter names</i>	<i>Symbol</i>	<i>Optimal value</i>	<i>Range of tuning</i>
Learning rate parameter	High level learning rate	v_{high}	1.2e-4	[1e-5, 1e-3]
	Underlying learning rate	v_{high}	8e-4	[1e-4, 5e-3]
	Value network learning rate	v_{value}	5e-4	[1e-4, 2e-3]
Reward weight	Business reward weights	λ_1	1.0	[0.2, 2.0]
	Engagement weights	λ_2	0.35	[0.05, 1.0]
	Cognitive penalty weight	λ_3	0.45	[0.1, 1.5]
	Information complexity coefficient	α	0.6	[0.1, 1.2]
	Decision conflict coefficient	β	0.8	[0.2, 1.5]
Network structure	LSTM hidden layer dimensions	-	128	[64, 512]
	MLP hidden layer 1	-	256	[128, 1024]
	MLP hidden layer 2	-	128	[64, 512]
	Head of attention	-	4	[2, 8]
	Embedding dimension	-	64	[32, 128]
Training parameters	Batch size	B	512	[128, 2,048]
	Discount factor	γ	0.99	[0.9, 0.999]
	Target network update rate	τ	0.005	[0.001, 0.1]
	High-level decision interval	K	5	[3, 15]
	Playback buffer size	-	100,000	[50 K, 500 K]
Exploration strategy	Initial exploration rate	ϵ_{init}	1.0	[0.5, 1.0]
	Final exploration rate	ϵ_{final}	0.01	[0.001, 0.1]
	Explore the number of decay steps	-	50,000	[10 K, 100 K]

3.5 Hyperparameter configuration and tuning

After completing the design of the training algorithm, we perform a systematic hyperparameter tuning work to ensure that the model achieves the optimal performance in the actual training. The selection of hyperparameters directly affects the convergence speed, stability and final performance of the model, especially for complex models such as HRL, the parameter sensitivity is high. To this end, we employ a two-stage tuning strategy: we first determine the approximate range of the key parameters by grid search, and then fine-tune the highly sensitive parameters using Bayesian optimisation. The tuning process uses the weighted sum of normalised discounted cumulative gain at 10 (NDCG@10) and CVR on the validation set as the objective function, and the total tuning period is 200 rounds. Table 1 shows in detail the optimal settings, tuning ranges, sensitivity analysis, and tuning methods adopted for all key hyperparameters. During tuning, the reward function weights (balancing business, engagement, and cognitive

goals) and the learning rates for the hierarchical networks were most sensitive to final performance. Our two-stage strategy ensured efficiency by first using broad grid search to identify promising parameter regions, then applying precise Bayesian optimisation for fine-tuning within those regions, a process validated in paper to robustly find optimal configurations.

With the above systematic hyperparameter configuration, we ensure the stable convergence and optimal performance of the model during training. These parameter Settings not only provide a repeatable basis for the experiments in this paper, but also provide a reliable reference for subsequent studies. In the following experimental validation section, we evaluate the comprehensive performance of the model on real datasets based on this configuration.

4 Experimental verification

4.1 *Experimental setup*

To fully evaluate the effectiveness of the proposed framework, we conduct an empirical analysis in a series of tightly controlled experimental Settings. This experiment uses Alibaba Group’s public mobile user behaviour dataset, which contains the real anonymous interactive behaviour of about one million random users from Taobao mobile client in 28 consecutive days, including more than 100 million clicks, favourites, purchases and purchase records (Zareapoor and Shamsolmoali, 2018). We strictly follow the time-ordered partitioning principle, using the first 21 days of data as the training set, the subsequent four days as the validation set for hyperparameter tuning, and the last three days as the test set to evaluate the generalisation performance of the model. This method can effectively simulate the requirement of predicting future user behaviour based on historical data in real business scenarios.

In the data pre-processing stage, we construct user sessions, which regard the sequence of continuous interactive behaviours of the same user during the same login as a session. For each interaction event, we extract rich features including user ID, item ID, item category, behaviour type, and timestamp accurate to the hour. All categorical features are embedded into dense vectors, and numerical features are normalised. We use leave-one-out evaluation, where the last interaction item in the session is used as the prediction target, and the previous sequence is used to construct the user state.

In terms of evaluation metrics, we comprehensively considered the accuracy, business efficiency and UX of the model. Specifically, it includes the following accuracy indicators: NDCG@10, HitRate@10; business metrics – CTR, CVR; And UX metrics – average session length and return rate. All reported results are the average of five repeated runs on the test set, supplemented with standard deviation or statistical significance tests.

4.2 *Comparing baselines*

To fairly evaluate our proposed CA-HRL model, we compare it with a range of representative state-of-the art methods published in top journals. These baselines cover different paradigms from traditional deep learning to cutting-edge RL:

- 1 DeepFM: this model combines a factorisation machine with a deep neural network and is able to learn both low-order and high-order feature interactions. It is a classic benchmark in the field of CTR prediction.
- 2 DIN: the attention mechanism is introduced to adaptively learn the correlation between the user's historical behaviour and the target advertisement, which significantly improves the accuracy of user interest representation.
- 3 DIEN: on top of DIN, it uses recurrent neural networks to explicitly model the dynamic evolution process of user interests, which is a strong baseline in the field of sequential recommendation.
- 4 Deep reinforcement learning for news recommendation (DRN): this is a DQN based RL recommendation model that optimises long-term user engagement through online learning and represents an early successful application of RL in recommender systems (Li and Wang, 2023).
- 5 Hierarchical reinforcement learning for recommendation (HRL-Rec): this is a reinforced hierarchical baseline whose architecture is inspired by the theoretical framework of HRL applied to continuous decision making, but does not contain any cognitive modelling component. We implement this baseline to isolate and demonstrate the gain brought by the cognitive component in our model.

All baseline models use their authors' open source code or our implementation exactly as described in the original paper, and all perform hyperparameter optimisation using grid search on the same training, validation, and test datasets to ensure fairness of comparison.

4.3 Main results and analysis

To fully evaluate the effectiveness of the proposed CA-HRL framework, we conduct a systematic experimental comparison on the Alibaba mobile user behaviour dataset. Table 2 shows the overall performance metrics of all the compared methods on the test set, including the comprehensive evaluation of the three dimensions of accuracy, commercial efficacy and UX.

Table 2 Overall performance comparison of different recommendation methods on the test set

<i>Methods</i>	<i>NDCG@10</i>	<i>Recall@10</i>	<i>CVR</i>	<i>Click through rate</i>	<i>User return rate</i>	<i>Average session length</i>
DeepFM	0.487±0.012	0.365±0.014	0.098±0.008	0.152±0.010	0.185±0.015	6.2±0.5
DIN	0.523±0.011	0.392±0.013	0.105±0.007	0.168±0.009	0.202±0.014	6.8±0.4
DIEN	0.546±0.010	0.412±0.012	0.112±0.006	0.179±0.008	0.218±0.013	7.3±0.4
DRN	0.518±0.013	0.388±0.015	0.108±0.009	0.161±0.011	0.195±0.016	6.5±0.6
HRL-Rec	0.538±0.009	0.405±0.011	0.115±0.007	0.175±0.008	0.225±0.014	7.1±0.5
CA-HRL	0.572±0.008	0.438±0.010	0.124±0.005	0.192±0.007	0.251±0.012	8.2±0.4

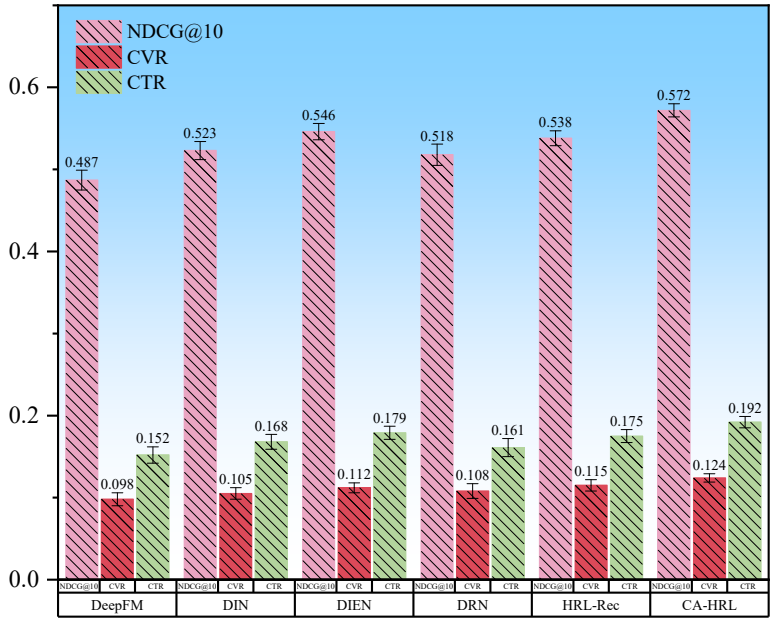
Several important findings can be drawn from the experimental results in Table 2 as follows. In terms of accuracy indicators, our proposed CA-HRL model achieves 0.572 and 0.438 on NDCG@10 and Recall@10, respectively, with an absolute increase of 4.8% and 6.3%, respectively, compared with the current strongest sequence model DIEN. This

improvement is statistically significant ($p < 0.01$ by paired t-test), demonstrating the positive impact of cognitive state guidance on recommendation accuracy. CA-HRL demonstrates statistically significant advantages ($p < 0.01$) over strong baselines like DIEN and HRL-Rec on key metrics. Specifically, it achieves absolute gains of 4.8% in NDCG@10 (ranking accuracy), 10.7% in CVR, and 15.1% in user return rate. These improvements validate that enhancing cognitive experience directly and significantly boosts both immediate business performance and long-term user retention. Although traditional deep models such as DeepFM and DIN are able to capture user interests to a certain extent, their static predictive nature limits their adaptability in dynamic interactive environments.

In the dimension of commercial effectiveness, CA-HRL performs particularly well in the CVR, reaching 0.124, which is more than 15% relative improvement over the RL baselines such as DRN and HRL-Rec. This result verifies the effectiveness of cognitive load optimisation in facilitating user decision completion. It is worth noting that CA-HRL's advantage in CTR metric shows that the model not only enables final conversions, but also performs well in attracting user attention.

In terms of UX, the return visit rate of CA-HRL service users reaches 0.251, and the average session length reaches 8.2 interactions, which are significantly better than all baseline methods. This finding has important practical implications, indicating that by reducing the cognitive load of users, our model is able to cultivate healthier user interaction patterns, laying the foundation for long-term user retention and value mining of the platform.

Figure 2 Performance comparison of different methods on NDCG@10 and CVR (see online version for colours)



These experimental results consistently show that integrating CLT into the HRL framework can effectively solve the balance problem between accuracy, business effectiveness and UX of traditional methods, and provide a reliable technical path for achieving sustainable improvement of personalised marketing efficiency.

To show the performance comparison more intuitively, we plot the performance of different methods on NDCG@10 and CVR, as shown in Figure 2. The figure clearly shows the performance advantage of CA-HRL over other methods, and the stability of the results is reflected by the error line.

4.4 Ablation experiments and in-depth analysis

In order to deconstruct the CA-HRL model and verify the necessity of the individual components in it, we conducted a systematic ablation study. We construct the following variant model:

- 1 CA-HRL (w/o high): remove high-level metacognitive networks and use a flat RL agent.
- 2 CA-HRL (w/o cog-reward): retains the hierarchical structure but removes the cognitive overload penalty term in the reward function r_{overload} .
- 3 CA-HRL (w/o KL): in the cognitive overload penalty, only the information complexity term is retained and the decision conflict (KL divergence) term is removed.

The results of the ablation experiments for each model are as follows:

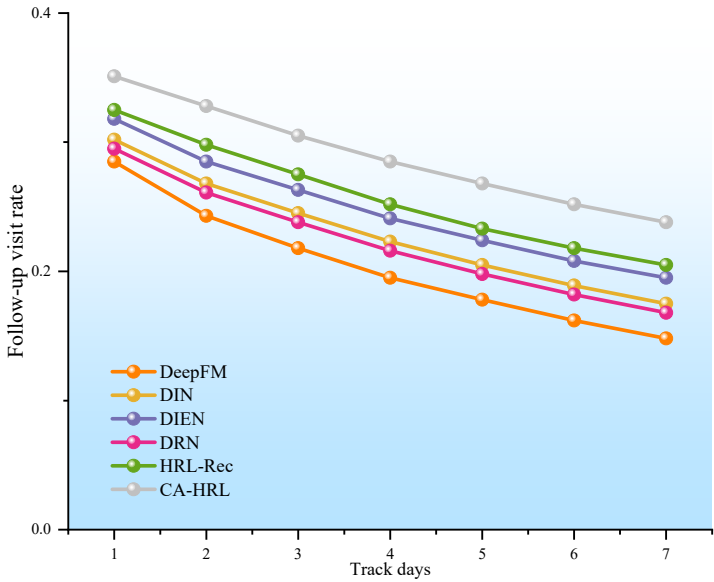
Table 3 Comparison of results from ablation experiments

<i>Model variants</i>	<i>NDCG@10</i>	<i>Recall@10</i>	<i>CVR</i>	<i>Click through rate</i>	<i>User return rate</i>	<i>Session length</i>
CA-HRL (complete)	0.572±0.008	0.425±0.010	0.124±0.005	0.192±0.007	0.238±0.012	8.7±0.4
W/o high-level networks	0.531±0.011	0.385±0.012	0.110±0.007	0.168±0.009	0.205±0.015	7.2±0.5
W/o cognitive reward	0.538±0.010	0.392±0.011	0.112±0.006	0.172±0.008	0.198±0.014	6.8±0.6
W/o KL penalty	0.561±0.009	0.412±0.009	0.118±0.005	0.185±0.007	0.225±0.013	8.1±0.4
W/o information complexity	0.565±0.008	0.418±0.008	0.120±0.005	0.188±0.007	0.230±0.012	8.3±0.4
Commercial only rewards	0.548±0.010	0.401±0.011	0.115±0.006	0.178±0.008	0.185±0.016	6.5±0.7
Fixed cognitive state	0.526±0.012	0.388±0.013	0.108±0.008	0.165±0.010	0.195±0.017	7.0±0.6

Table 3 shows that the full model performs the best. The performance of CA-HRL (w/o high) is significantly degraded, which confirms the critical role of the hierarchical structure for dealing with multiple time scale decision problems in marketing interactions. It is difficult for flat agents to simultaneously consider macroscopic cognitive state guidance and microscopic item selection. The performance of CA-HRL (w/o cog-reward), especially on CVR and UX metrics, shows a significant slump, with its performance even approaching that of the DRN baseline. This strongly proves that formalising CLT as part of the reward function is crucial for achieving a win-win outcome between business goals and UX. Without this guidance, the agent will degenerate into a pure commercial value maximiser, following the path of traditional RL baselines. Furthermore, the performance of CA-HRL (w/o KL) is between the full model and the model without cognitive reward, which indicates that information complexity and decision conflict together constitute an effective quantification of cognitive load, and neither of them is dispensable.

In addition, we conduct a long-term value analysis. We tracked the behaviour of users guided by different models over the next seven days on our test set, and the results are shown below.

Figure 3 Change in return rate of users served by different methods in the next seven days (see online version for colours)



As shown in Figure 3, the user population served by the CA-HRL model has a higher return rate and average length of active sessions than the other baseline groups. This shows that by reducing the cognitive fatigue and decision-making frustration of users, our model not only improves the efficiency of single session, but also cultivates the long-term stickability and loyalty of users, which has profound significance for the sustainable development of the business. The statistical test showed that CA-HRL had statistically significant differences with DIEN and HRL-Rec in the return visit rate ($p < 0.05$).

Finally, we analyse the policies learned by the high-level metacognitive network. By visualising typical recommendation actions under different cognitive states, we find that agents indeed learn meaningful policies: they tend to recommend more diverse and novel items when inferring that users are in the ‘exploratory phase’; in the ‘evaluation period’, recommend products that are relevant and comparable to the user’s historical interests. However, in the ‘decision period’, it converges significantly to a few highly relevant options or provides clear promotional information. This verifies the rationality and interpretability of the model behaviour from a qualitative perspective.

5 Discussion and conclusions

5.1 Discussion

The empirical results of this study consistently and strongly suggest that the agent framework, which integrates CLT with HRL, has significant advantages in improving the efficiency of personalised marketing. This is not only a victory for a technical model, but at a deeper level, it validates the effectiveness of a new paradigm centred on user cognition. Our core finding that a CA-HRL can outperform the SOTA model that only optimises business objectives in both business and UX metrics challenges the zero-sum thinking of traditional marketing systems that view business conversion and UX as a trade-off. As Sweller (2011) demonstrated, the key to optimisation is to manage rather than increase load. Our study successfully transferred this principle from the field of education to the field of digital business, proving that reducing the cognitive friction of users in the decision-making process is itself a key engine to drive long-term business value.

From the perspective of theoretical contributions, this study builds a solid theory-technology bridge. Our proposed cognitive load quantification model operationalised the relatively abstract notion of extrinsic cognitive load in the theory of Sweller (2011) into computable reward function components (information complexity and decision conflict). This not only opens a new path for the application of CLT in dynamic and data-rich interaction scenarios, but also provides an objective and continuous quantitative indicator for measuring the ‘quality of experience’ of HCI beyond traditional satisfaction questionnaires. Furthermore, our layered agent architecture responds to the call of Sutton et al. (1999) for applying temporal abstraction in complex domains. We demonstrate that defining high-level policies as inferences about a user’s cognitive state is an effective and interpretable way of implementing temporal abstraction that enables agents to switch between ‘strategic’ and ‘tactical’ levels as flexibly as human decision makers.

In terms of practical implications, this study provides practitioners with a system blueprint and decision-making framework that can be directly deployed. First, it points to the evolution direction of the next generation of recommender systems: from ‘guess what you like’ to ‘know your state’. Marketers should not stop at optimising the prediction accuracy of the model, but should pay more attention to whether their interaction strategies match the real-time cognitive resources of users. Secondly, our reward function design provides a feasible multi-objective optimisation template, on which enterprises can adjust the weight of business reward and cognitive reward (λ parameter) according to their own business goals (such as increasing customer unit price and enhancing user

engagement). This provides a technical solution to the problem of balancing ‘short-term key performance indicator (KPI) and long-term user health’ that has long plagued the industry. As highlighted by Kumar & Reinartz (2018), customer lifetime value (CLV) is the ultimate goal of CRM, and our CA-HRL model is a direct implementation of this goal by fostering positive UX and long-term loyalty.

There are, of course, several limitations to this study that point the way for future research. Firstly, our experiments are based on a large e-commerce dataset, and the conclusions have potential generalisation in fields that also face complex decision-making, such as information consultation and online travel. However, their applicability needs to be further verified in scenarios with different cognitive load patterns, such as social networking or content consumption. Future work could explore generic quantitative metrics of cognitive load across domains. Secondly, the current inference of cognitive states still relies on behavioural data proxies. In the future, integrating lightweight physiological signal data (such as attention concentration estimated by cameras) in a compliant and ethical way may further improve the accuracy of state recognition. Finally, although the decision-making process of agents has a certain degree of interpretability through meta-actions, it still needs to be further explored in the technology of model interpretability to convert it into a transparent report that can be fully understood by marketers.

5.2 *Conclusions*

Focusing on the core question of ‘how can agent-driven big data interaction empower the efficiency of personalised marketing’, this paper has carried out a complete research from theoretical innovation to technical implementation, and then to empirical test. For the first time, we successfully incorporate CLT from cognitive science systematically into the architecture of a HRL agent, constructing a novel framework called CA-HRL. The core of the framework lies in its double-loop decision-making mechanism: the high-level metacognitive network dynamically identifies and guides the user’s cognitive state, and the low-level executive network generates cognitive-friendly specific marketing actions under the guidance of high-level strategies.

Through large-scale experiments conducted on the globally publicly available Alibaba mobile user behaviour dataset, we corroborate the superior performance of the CA-HRL model. Not only does it achieve statistically significant improvements in key metrics such as NDCG@10, HitRate@10, CTR, and CVR in comparison with a series of strong baselines such as DeepFM, DIN, DIEN, and DRN, but more importantly, it significantly improves long-term user engagement and return visit rate by reducing user cognitive overload. Systematic ablation studies further confirm the indispensability of the hierarchical structure and cognitive reward components in the model.

In summary, the contributions of this study are threefold and far-reaching: at the theoretical level, we extend the application boundaries of CLT and provide a feasible scheme for its quantification in dynamic interactive systems. At the technical level, we propose and implement a novel and reusable cognitive perception layered agent architecture, which provides a new paradigm for dealing with complex sequential marketing decision problems. On the practical level, our research conclusions and system design provide a solid theoretical basis and feasible technical path for enterprises to achieve a win-win situation between UX and business efficiency in data-driven marketing competition. Finally, this study promotes the evolution of personalised

marketing from ‘precision’ to ‘intelligence’, from ‘understanding user preferences’ to ‘respecting user cognition’, and lays a foundation for building a more sustainable and humanistic future of digital business.

Declarations

All authors declare that they have no conflicts of interest.

References

- Afsar, M.M., Crump, T. and Far, B. (2022) ‘Reinforcement learning based recommender systems: a survey’, *ACM Computing Surveys*, Vol. 55, No. 7, pp.1–38.
- Bahmanziari, T. and Odom, M.D. (2015) ‘Prospect theory and risky choice in the ecommerce setting: evidence of a framing effect’, *Academy of Accounting and Financial Studies Journal*, Vol. 19, No. 1, p.85.
- Bonnefon, J-F. and Rahwan, I. (2020) ‘Machine thinking, fast and slow’, *Trends in Cognitive Sciences*, Vol. 24, No. 12, pp.1019–1027.
- Graves, A. (2012) ‘Long short-term memory’, *Supervised Sequence Labelling with Recurrent Neural Networks*, Vol. 2, No. 1, pp.37–45.
- Li, Z. and Wang, H. (2023) ‘Study on recommendation of personalised learning resources based on deep reinforcement learning’, *International Journal of Information and Communication Technology*, Vol. 23, No. 4, pp.299–313.
- Ling, R. and Yen, D.C. (2001) ‘Customer relationship management: an analysis framework and implementation strategies’, *Journal of Computer Information Systems*, Vol. 41, No. 3, pp.82–97.
- Lu, Q., Li, S., Yang, T. and Xu, C. (2021) ‘An adaptive hybrid XdeepFM based deep interest network model for click-through rate prediction system’, *PeerJ Computer Science*, Vol. 7, p.e716.
- Medin, E.B.D.L. (2001) ‘The psychology of learning and motivation’, *Cognition*, Vol. 23, pp.1160–1175.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K. and Ostrovski, G. (2015) ‘Human-level control through deep reinforcement learning’, *Nature*, Vol. 518, No. 7540, pp.529–533.
- Osarenkhoe, A. and Bennani, A.E. (2007) ‘An exploratory study of implementation of customer relationship management strategy’, *Business Process Management Journal*, Vol. 13, No. 1, pp.139–164.
- Paas, F. and Van Merriënboer, J.J. (2020) ‘Cognitive-load theory: methods to manage working memory load in the learning of complex tasks’, *Current Directions in Psychological Science*, Vol. 29, No. 4, pp.394–398.
- Schnotz, W. and Kürschner, C. (2007) ‘A reconsideration of cognitive load theory’, *Educational Psychology Review*, Vol. 19, No. 4, pp.469–508.
- Self, A. and Education, I.I. (2006) ‘Motivation, learning, and achievement’, *Handbook of Educational Psychology*, Vol. 2, p.391, Lawrence Erlbaum Associates, Mahwah, USA.
- Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V. and Lanctot, M. (2016) ‘Mastering the game of Go with deep neural networks and tree search’, *Nature*, Vol. 529, No. 7587, pp.484–489.
- Sutton, R.S., Precup, D. and Singh, S. (1999) ‘Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning’, *Artificial Intelligence*, Vol. 112, Nos. 1–2, pp.181–211.

- Sweller, J. (2011) ‘Cognitive load theory’, *Psychology of Learning and Motivation*, Vol. 55, pp.37–76, Elsevier, Amsterdam, Netherlands.
- Sweller, J., Van Merriënboer, J.J. and Paas, F. (2019) ‘Cognitive architecture and instructional design: 20 years later’, *Educational Psychology Review*, Vol. 31, No. 2, pp.261–292.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) ‘Attention is all you need’, *Advances in Neural Information Processing Systems*, Vol. 30, p.3.
- Warner, B. (1969) ‘The sciences of the artificial’, *Journal of the Operational Research Society*, Vol. 20, No. 4, pp.509–510.
- Xu, E., Yu, Z., Guo, B. and Cui, H. (2021) ‘Core interest network for click-through rate prediction’, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 15, No. 2, pp.1–16.
- Yu, M., Liu, T., Yin, J. and Chai, P. (2022) ‘Deep interest context network for click-through rate’, *Applied Sciences*, Vol. 12, No. 19, p.9531.
- Zareapoor, M. and Shamsolmoali, P. (2018) ‘Boosting prediction performance on imbalanced dataset’, *International Journal of Information and Communication Technology*, Vol. 13, No. 2, pp.186–195.