A Transformer-Based Multimodal Semantic Retrieval Model for Business Intelligence Systems

Jigang Xie
Nanjing University of Industry Technology, Nanjing, Jiangsu, 210023, China
E-mail: xiejigang4765@163.com

Keywords: artificial intelligence enhancement, semantic information retrieval, business intelligence, deep semantic modeling

Received: August 9, 2025

In the increasingly growing business intelligence (BI) environment of multi-source heterogeneous data, traditional information retrieval methods face significant bottlenecks in accuracy, response efficiency, and semantic understanding ability. We aim to investigate whether multimodal semantic modeling and dynamic intent recognition can significantly improve retrieval precision and response efficiency in BI contexts. This paper designs and implements a Transformer-based multimodal semantic retrieval model architecture, which combines a multi-layer semantic modeling mechanism with a context enhancement strategy to model the deep matching relationship between user queries and multimodal business data. The architecture introduces a query semantic vector generation module based on Transformer encoders, adopts a multi-channel deep feature fusion structure for structured fields, behavior logs, and documents, and incorporates a dynamic user intent recognition module for context-aware representation. The training employs a contrastive loss with softmax normalization, optimized with the AdamW optimizer and cosine learning rate scheduling. Experiments are conducted on three enterprise-level datasets, including an internal document corpus (42,000+ samples), a structured product dataset (18,000 records), and user behavior logs (3.1M entries). Evaluation results demonstrate that the proposed model outperforms BM25, DSSM, and BERT Retriever, achieving Precision@10 = 0.723, nDCG@10 = 0.702, and MRR = 0.537, with relative improvements of up to 28.6%. In addition, the model reduces average response latency to 430 ms and maintains a user satisfaction score above 87, proving its feasibility for deployment in intelligent decision-support BI platforms.

Povzetek: Članek predstavi transformacijski multimodalni model za semantično iskanje v poslovni inteligenci, ki združuje tekst, strukture in vedenjske podatke z dinamičnim prepoznavanjem namena. Na treh podjetnih naborih doseže dobre rezultate.

1 Introduction

Traditional information retrieval systems have long played the role of static tools in enterprise data utilization, relying mainly on keyword matching and rule indexing to support information acquisition. However, with the popularity of Business Intelligence (BI) systems in enterprise operations, information retrieval tasks are shifting from "passive retrieval" to "semantic understanding" and "active recommendation" stages. This evolution benefits from the integration and development of artificial intelligence, natural language processing, and big data technology, providing new impetus for the upgrading of commercial information systems. In the business intelligence environment, information retrieval is no longer just about finding whether a certain keyword exists, but about extracting semantic information that is meaningful to the current business scenario from heterogeneous, multi-source, and structurally diverse data. The large amount of data generated in the daily operation of enterprises, including text contracts, financial statements, user behavior logs, market sentiment, and product images, has far exceeded the scope that traditional information systems can parse. These data often exhibit two key patterns: one is the short-term, task oriented instant mode, used to quickly respond to user queries and behavioral needs; The other type is a deep semantic pattern that spans time and business domains, revealing the inherent correlation between user intent and business development. These two types of information together form the fundamental context of commercial retrieval systems [3].

Identifying the complex interaction relationships between these patterns and mapping them to user query behavior is an important challenge facing current information systems. Traditional methods are difficult to meet the understanding needs of contextual semantics and cannot adapt to the collaborative effects of multimodal data in the retrieval process [4]. To this end, researchers have attempted to use artificial intelligence methods such as deep learning to introduce semantic modeling, attention mechanisms, and intent recognition mechanisms to enhance the system's dynamic response capability to user needs. Accurate information retrieval not only improves the efficiency of utilizing internal knowledge within the enterprise, but also demonstrates significant value in

external customer service, market monitoring, and business risk control. For example, in the formulation of sales strategies, intelligent retrieval based on semantic recognition can quickly locate the focus of customer attention, thereby optimizing the pace of product launch; In brand monitoring, the system can perceive changes in public opinion and dynamically adjust the risk warning level based on keyword evolution. The deep coupling between retrieval behavior and commercial activities has led to the necessity of building a unified intelligent retrieval system, which should have the ability of semantic understanding, multimodal fusion, and user intention perception, and become an indispensable central engine in business intelligence platforms. In response to the problems of response lag, shallow semantic understanding, and poor structural adaptability in current commercial information retrieval systems, this paper proposes an AI enhanced retrieval model architecture for BI scenarios. This model integrates semantic encoding, intent recognition, and multimodal data processing modules, aiming to enhance the perception and reasoning abilities of retrieval systems for complex enterprise data, and promote breakthroughs in information systems in precise acquisition, active recommendation, and intelligent feedback.

This paper aims to address the following research questions: (1) Can multimodal semantic modeling improve retrieval precision in BI contexts? (2) How does dynamic intent recognition enhance ranking performance under complex user queries? (3) To what extent can a Transformer-based fusion framework reduce response latency while maintaining accuracy?

The structure of this article is as follows: Chapter 2 summarizes the research achievements and development trends in the intersection of information retrieval and business intelligence systems; Chapter 3 introduces the design framework and functional division of the proposed model; Chapter 4 elaborates on the system implementation path and key technology deployment methods; Chapter 5: Application verification and effectiveness evaluation based on enterprise level real datasets; Chapter 6 explores the challenges and coping strategies that the model may face during the promotion process; Chapter 7 summarizes the entire text and proposes future optimization directions.

2 Related work

With the deep embedding of data-driven strategies in enterprise operations, Artificial Intelligence (AI) has gradually become a key supporting force for the evolution of business intelligence systems. Existing research has extensively focused on the multi-level application of AI technology in scenarios such as enterprise management, operational optimization, and decision modeling. For example, Asmar and Al Rob (2024) [7] pointed out in their literature review that AI tools are leaping from assisting decision-making to proactive insight and strategy generation, reshaping organizations' cognitive structures and response mechanisms to information. Senadzki et al. (2023) [8] further emphasized that the integration of AI capabilities plays a significant role in enhancing enterprise competitiveness and promoting the achievement of sustainable development goals.

In terms of the composition of business intelligence systems, information retrieval, as the most fundamental and active component, has been deeply influenced by the AI wave in its technological evolution. The traditional retrieval model based on keyword matching and Boolean logic is difficult to meet the needs of enterprise users for semantic understanding, contextual response, and personalized recommendations. This technological bottleneck has driven the embedding transformation of AI in information retrieval systems. Yang et al. (2024) [9] summarized four major paths for AI enabled business models through systematic literature review, one of which is "semantic driven information acquisition", which improves the model's ability to recognize complex query semantics and response accuracy through deep neural networks, attention mechanisms, and semantic vector embedding.

With the intervention of machine learning methods, the structure and functionality of information retrieval models have begun to undergo deep adjustments. Yin and Li (2024) proposed introducing artificial intelligence into the information management module of university management courses to enhance the ability of knowledge graph construction and query optimization. This model has shown significant user intention recognition effects in actual teaching feedback. Chanda and Tidd (2024) [11] explore how human judgment can collaborate with algorithms in AI assisted decision-making systems from a cognitive perspective, emphasizing the value and necessity of "interpretable retrieval". In addition, Mahalakshmi et al. (2022) [12] studied the implementation path of AI technology in the financial services industry and proposed that information retrieval models should not only improve the accuracy of relevance scores, but also consider multi-objective collaborative optimization of response time, business context, and risk tolerance.

Although the above research provides rich support in dimensions such as business intelligence systems, AI decision support, and semantic modeling, there are still shortcomings in the context of enterprise level information retrieval systems. Firstly, most current models focus on text semantic matching and lack the ability to integrate structured data with multimodal information (such as charts, behavior logs, etc.) [13]; Secondly, some studies only validate the effectiveness of algorithms in theoretical or simulated scenarios, lacking system level validation and feedback loops in real enterprise business scenarios [14]; Thirdly, research on modeling user intent is relatively independent and lacks a linkage modeling mechanism with factors such as query evolution, context transfer, and behavior sequence [15].

To further clarify the positioning of our work, Table 1 summarizes representative models including BM25, DSSM, and BERT-Retriever, comparing their architectures, data modalities handled, and reported performance metrics.

Table 1：Comparison of representative information retrieval models

Model	Architecture	Modalities handled	Reported metrics (example)	Limitations
BM25	Sparse keyword-based ranking	Text only	P@10 ≈ 0.52, nDCG@10 ≈ 0.48	No semantic understanding, weak in multimodal context
DSSM	Deep structured semantic model (feed-forward NN)	Text only	P@10 ≈ 0.60, nDCG@10 ≈ 0.55	Lacks contextual modeling, not adaptive to multimodal data
BERT-Retriever	Transformer encoder with contextual embeddings	Text only	P@10 ≈ 0.65, nDCG@10 ≈ 0.62, MRR ≈ 0.66	High computational cost, limited scalability to structured/behavioral data

As shown in Table 1, while existing models achieve good performance on textual semantic matching, they lack robustness in multimodal enterprise scenarios. This motivates our proposal of a Transformer-based multimodal retrieval model.

Therefore, based on the inheritance of previous research results, this article proposes an artificial intelligence enhanced information retrieval model for business intelligence scenarios, aiming to achieve comprehensive innovation in query intent modeling, semantic space construction, multimodal data fusion, and system deployability, and empirically verify it through enterprise level real data. Through this path, it is expected to bridge the technological gap in existing research where "models are easy to use but difficult to deploy, high accuracy but poor business perception", and promote AI retrieval systems from "laboratory level tools" to "enterprise level services".

3 Architecture design of AI enhanced information retrieval model

In the AI enhanced information retrieval system proposed in this article, the selected model architecture follows a three-level strategy of "semantic understanding feature fusion result evaluation", with the core goal of addressing challenges such as typical multi-source heterogeneous data processing, semantic redundancy compression, and user intent uncertainty in business intelligence systems. Unlike traditional search engines that focus on keyword matching, this model emphasizes the construction of semantic bridges between queries and information units at a deep semantic level, enhancing the response intelligence level of information systems.

In the process of architecture selection, we prioritized the structure dominated by traditional inverted indexes, which showed significant accuracy bottlenecks when facing business queries with semantic ambiguity and frequent context jumps. The Transformer model, which is widely used in general natural language tasks, has the advantage of long-distance modeling in semantic representation. However, its strong dependence on single source text and lack of native support for structured data limit its adaptability in multimodal commercial data. Based on this, this study constructed a fusion based dual branch architecture: on the one hand, the semantic understanding module was used to model the context of textual data, and on the other hand, the feature fusion module was used to vectorize and encode structured data and user behavior, ultimately achieving unified matching judgment in the scoring module. The overall architecture of the system is shown in Figure 1:

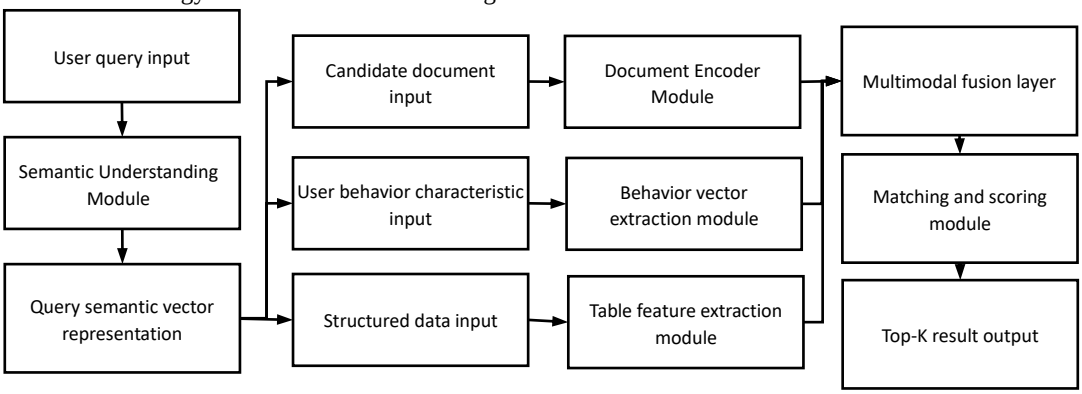


Figure 1：Structure diagram of AI enhanced information retrieval

Model this architecture has the following features:

Module decoupling, clear structure. The model deploys semantic understanding, feature fusion, and scoring output functions separately, making it easy to replace and fine tune for different data sources and retrieval tasks.

Multi modal data fusion mechanism. In the business intelligence scenario where non textual information is rampant, the model combines text, structured data, and user

behavior data through fusion modules to enhance the perception ability of users' complex query intentions.

Emphasize both semantic representation and system response. The system not only emphasizes deep semantic matching between queries and information units, but also focuses on response time and deployability, making it suitable for online application requirements in large BI systems.

The AI enhanced architecture proposed in this chapter aims to build an intelligent retrieval core module with high scalability, strong semantic modeling capabilities, and cross modal processing capabilities within information systems. The next section will further break down the hierarchical functions within the model, explaining the specific implementation logic and data flow mechanism of each submodule.

3.1 Overall structure and functional stratification of the model

To enhance the semantic perception capability and task adaptability of information retrieval systems in business intelligence scenarios, the AI enhancement model proposed in this paper follows the architecture concept of "layered decoupling, functional collaboration, and task fusion". The overall structure is divided into four functional levels: input perception layer, semantic encoding layer, interaction fusion layer, and sorting output layer. Information is transmitted between different layers through a unified data interface standard, which ensures the flexibility of the model during deployment and facilitates independent training and optimization of different submodules.

The input perception layer serves as the data access port of the system, mainly responsible for preprocessing user queries, candidate information units, and user contextual environments. Considering the existence of multi-source heterogeneous data such as text, structured tables, and behavior logs in BI systems, the model encodes input data of different modalities into a unified tensor format through a specially designed data parser, which facilitates subsequent modeling and processing. The semantic encoding layer relies on Transformer structure and lightweight convolutional network to handle text semantic modeling and non text feature extraction tasks, respectively; This design ensures that the model has both the ability to understand deep contextual information and the performance advantage of quickly processing heterogeneous data.

The interaction fusion layer is the core module of the model, responsible for integrating three types of vector information: query, document, and context, and introducing attention mechanisms to dynamically adjust feature weights, so that the final output results can fully reflect the comprehensive effect of semantic relevance and business background. The sorting output layer is based on the fused representation, completing matching scoring, candidate sorting, and Top-K result generation, and providing interfaces to support system level result display and calling. The following table summarizes the roles and corresponding key technology paths of each functional level:

Table 1：Overview of functional hierarchical design of ai enhanced information retrieval model

Layer Name	Core Functional Description	Key Technical Components
Input Perception Layer	Receives and parses multimodal inputs, including text, structured tables, and user behavior	Tokenizer, Data Normalization Tools, Field Parser
Semantic Encoding Layer	Builds deep semantic representations of queries and information units, extracting key contextual features	Transformer Encoder, Lightweight CNN, Feature Embedding Module
Interaction & Fusion Layer	Fuses multi-source feature vectors and models query-document relations via contextual attention	Multi-head Attention Module, Residual Connection Layer, Feature Concatenation & Compression
Ranking & Output Layer	Computes matching scores based on fused features and outputs the top-ranked retrieval results	Ranking Network, Top-K Selector, System Output Format Converter

This hierarchical structure not only facilitates system performance optimization and module replacement, but also supports personalized model deployment and fine-tuning strategies in specific scenarios. In business intelligence systems, different enterprise users often have different requirements for retrieval response speed, recommendation accuracy, and contextual understanding. Through decoupling settings at the functional layer, different depth or structure sub models can be flexibly plugged in and out to achieve customized retrieval services for different business objectives.

3.2 Query modeling and vector representation based on deep semantics

In information retrieval systems, accurately representing the semantic features of user queries is the foundation of matching calculations. Traditional methods such as TF-IDF and BM25 typically use sparse term weight matrices for modeling and lack context understanding capabilities. To overcome this limitation, this paper introduces a deep encoder based on Transformer structure for constructing low dimensional, context sensitive query vectors.

Given the query text $Q = \{w_1, w_2, \ldots, w_n\}$ input by the user, first map each word to a static word vector and combine it into an embedding matrix:

$$E_Q = [e_1; e_2; ...; e_n] \in R^{n \times d} \quad (1)$$

Subsequently, the embedding matrix is input into the Transformer encoder, which generates a context aware representation of H_Q through self attention mechanism, where each word position vector contains semantic dependency information between the word and the entire sentence.

Finally, a weighted average pooling strategy is adopted to compress the sequence representation into a single query vector q, where the weights are determined by the attention scores

$$q = \sum_{i=1}^{n} \alpha_i h_i \quad (2)$$

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)} \quad (3)$$

Among them, α_i is the attention weight of the i-th word in the aggregation process, where s_i represents the importance score of the i-th position, usually obtained by linear transformation. This representation method effectively enhances the query's ability to express syntactic structure and semantic focus while maintaining compactness. The generated query vector q will serve as the core input for subsequent matching with document and user context interactions. This modeling approach supports end-to-end training, has good generalization performance and differentiable optimization ability, and is suitable for application in multi-source business scenarios. In our implementation, all query and document embeddings are 256-dimensional (d = 256). Word vectors are initialized from pre-trained FastText embeddings and fine-tuned during training to adapt to enterprise terminology.

3.3 Fusion processing mechanism for multimodal commercial data

The types of information involved in business intelligence systems are highly heterogeneous, including text descriptions, structured fields, user interaction logs, and other behavioral data. In order to unify the representation of information from different sources and improve the expression ability of the model, this paper designs a multimodal processing mechanism based on deep feature fusion, which uses a unified vector space to align and model various types of information.

Assuming the user query is Q, its corresponding information unit to be matched includes three types of inputs: text data T, structured data S, and behavior sequence data B. The three types of features are first concatenated to form an initial fusion vector:

$$Z_{raw} = [f_T; f_S; f_B] \in R^{d+d_s+d_b} \quad (4)$$

Among them, the text information T is encoded into vector $f_T \in R^d$ through a pre trained language model; Structured features (such as timestamps, category labels, amounts, etc.) are embedded and normalized to represent $f_S \in R^{d_s}$; The behavior sequence is modeled as $f_B \in R^{d_b}$ through convolution or time series networks. Considering the uneven semantic contribution of various modal features, a learnable weighted fusion mechanism is further introduced to extract fusion vectors through linear transformation and nonlinear activation

$$Z_{fused} = \sigma(W \cdot z_{raw} + b) \quad (5)$$

Among them, $W \in R^{d' \times (d+d_s+d_b)}$ is the fusion weight matrix, $b \in Rd'$ is the bias term, $\sigma(\cdot)$ represents the ReLU activation function, and d' is the output dimension of the fusion representation. This structure allows the model to automatically learn the importance configuration of different modalities in specific scenarios based on training data. The final fusion vector Z_{fused} will be used for semantic matching and relevance scoring with the query representation. Its construction ensures that multi-source data has unified alignment ability in the information system and preserves potential complementarity between modalities. Structured categorical variables are embedded using learnable embeddings of size 64, while numerical features are normalized to [0,1]. We adopt the ReLU activation due to its lower computational cost and faster convergence in large-scale BI data, compared with GELU which offers smoother gradients but higher latency.

3.4 User intent recognition and context enhancement strategies

In complex business intelligence scenarios, user queries often exhibit features such as semantic incompleteness, target ambiguity, and strong contextual dependencies. In order to achieve more accurate retrieval and matching, the system needs to build a user intent recognition module in the information processing front-end, combined with a query context modeling strategy, to effectively restore the user's real needs. On the basis of semantic modeling, this article introduces an intention representation method that combines behavior trajectory encoding and context alignment, and enhances the retrieval module's perception ability of user targets through an explicit vector fusion mechanism.

The user's current query is marked as Q_t, and its preceding behavior trajectory includes a historical query set of $\{Q_{t-1}, Q_{t-2}, \ldots Q_{t-k}\}$ and corresponding interaction content of $\{D_{t-1}, D_{t-2}, \ldots D_{t-k}\}$. Encode each historical query and result document separately to obtain concatenated embeddings $(q_{t-i}, d_{t-i}) \in R^{2d}$. Constructing historical context representation using weighted aggregation method:

$$c_t = \sum_{i=1}^{k} \gamma_i \cdot [q_{t-i}; d_{t-i}] \quad (6)$$

Among them, $\gamma_i \in [0,1]$ is the weight of the i-th behavior's impact on the current intention, which satisfies

$\sum_i \gamma_i$ and is obtained through time decay or attention learning mechanisms. The current query Q_t is encoded to obtain semantic vector $q_t \in R^d$, which is then fused with historical context c_t to construct the final intent representation vector u_t:

$$u_t = \tanh(W_u \cdot [q_t; c_t] + b_u) \quad (7)$$

Among them, $W_u \in R^{d\times 3d}$ is the fusion weight matrix, $b_u \in R^{3d}$ is the bias term, and $\tanh(\cdot)$ is the nonlinear activation function. This representation has semantic perception and historical memory capabilities, which are used to guide the correlation scoring and ranking optimization of subsequent candidate information. After embedding the above intention enhancement mechanism, the model can more effectively distinguish short-term information needs from long-term interest preferences, especially in complex u_t and continuous query chains with strong performance. The final intent vector and candidate semantic representation jointly participate in matching judgment, providing a more discriminative retrieval scoring basis for the system.

4 System implementation path and algorithm deployment plan

4.1 System architecture construction and module collaboration mechanism

To achieve AI enhanced information retrieval functionality, the system adopts a modular design structure and follows the implementation principles of "layered deployment, asynchronous computing, and collaborative calling", embedding model capabilities into a service-oriented information system architecture. This architecture mainly includes four key modules: query understanding module, candidate generation module, deep matching module, and result reordering module. Each module communicates collaboratively through shared representation vectors and task interfaces.

The query understanding module receives user natural language input and outputs a semantic representation vector $q \in R^d$, which is passed to downstream modules in the form of an intermediate representation within the system to avoid duplicate processing of the original input. The candidate generation module quickly recalls the initial document set $D = \{d_1, d_2, \ldots, d_N\}$ based on lightweight vector indexing or rule templates, and each document d_i is generated by an encoder to represent $d_i \in R^d$.In the deep matching stage, the system calculates the semantic relevance score of (q, d_i) for each pair of s_i and uses dot product similarity to achieve fast matching:

$$s_i = q^T \cdot d_i \quad (8)$$

The similarity value forms the sorting vector $S = \{s_1, s_2, \ldots, s_N\}$, which serves as the input for the result reordering module. In order to further integrate context, user behavior, and business intent, this module also introduces a fusion vector u_t is incorporated into the final scoring process. Specifically, the final score is computed as:

$$s_{final}(q,d) = q^T d + f(u_t, d) \quad (9)$$

where f(·)f(\cdot)f(·) is a lightweight feedforward correction network that integrates the intent vector utu_tutu_t with the candidate document representation ddd. This ensures that contextual information contributes explicitly to the ranking decision. The modules of the entire system remain decoupled at the deployment level, supporting distributed expansion and asynchronous loading, which makes it easy to fine-tune and quickly replace submodules for different business scenarios. In terms of service interaction, each module transmits features and results through a unified vector interface, ensuring strong maintainability and reliable online inference performance of the system.

4.2 Model training and parameter optimization strategies for retrieval core modules

The retrieval core module mainly completes the task of semantic correlation modeling, and its performance directly affects the recall quality and sorting accuracy of the system. In order to improve the representation ability and generalization effect of the model, this paper introduces a point-to-point supervision mechanism and a negative sample comparison learning strategy in the model training stage. A deep matching model based on similarity scoring is adopted, and end-to-end training is carried out by optimizing the sorting objective function.

The training data is constructed in the form of a triplet (q, d^+, d^-), where q is the user query, d^+ represents positive sample documents (related), and d^- represents negative sample documents (unrelated). The query and document are mapped to vectors representing $q, d^+, d^- \in R^d$ through semantic encoders. Calculate the matching score between the query and the document using dot product method:

$$s^+ = q^T \cdot d^+ \quad (10)$$

$$s^- = q^T \cdot d^- \quad (11)$$

For consistency, the optimization objective is aligned with the inference-stage scoring function described in Section 4.1, ensuring that the context-enhanced intent vector utu_tutu_t also contributes to the training process. In order to enhance the model's ability to distinguish positive and negative samples, an objective function based on contrastive loss is introduced, and the cross-entropy form normalized by softmax is used for optimization:

$$L = -\log\left(\frac{\exp(s^+)}{\exp(s^+) + \exp(s^-)}\right) \quad (12)$$

This loss function can encourage the model to improve positive sample scores while suppressing negative sample scores, with clear gradient directionality and good training stability. During the training process, batch samples are

randomly shuffled and fed into the network, and the parameters are updated through the Adam optimizer. The learning rate is set to dynamically adjust to avoid premature convergence. In addition, to mitigate the risk of overfitting, the model introduces Dropout mechanism at the structural layer and uses L2 regularization term to restrict parameter norm at the embedding layer. In industrial deployment, considering the efficiency of online inference, the multi branch attention mechanism used in the training phase will perform parameter folding during inference, thereby reducing inference latency and computational resource consumption. This training and optimization strategy balances expression ability, training efficiency, and deployment performance, providing a model foundation for stable system operation.

4.3 Retrieval performance optimization and scalable deployment plan

In order to improve the operational efficiency and system responsiveness of AI enhanced information retrieval models, this paper constructs a multi strategy performance improvement mechanism from two aspects: model computation optimization and deployment structure elasticity. The model inference process adopts a dense vector matching architecture. To improve matching speed and memory utilization, the system introduces a standardized vector compression strategy, which preprocesses all vectors into unit norm form to accelerate the calculation of dot product similarity

$$\tilde{q} = \frac{q}{\|q\|}, \tilde{d} = \frac{d}{\|d\|} \tag{13}$$

$$\tilde{s} = \tilde{q}^{T} \cdot \tilde{d} = \cos(\theta) \tag{14}$$

Among them, $\tilde{q}, \tilde{d} \in R^{d}$ represents the normalized query and document vectors, and $\cos(\theta)$ is the cosine value of the angle between the two. This normalization operation can make dot product equivalent to cosine similarity, making it easier to efficiently recall using vector indexing structures such as FAISS or ANN.

At the deployment level, the system adopts a master-slave distributed retrieval architecture, where the master node is responsible for receiving requests and parsing query semantics, while the slave nodes complete candidate generation and correlation calculation tasks in parallel. To evaluate the scalability of the system, a concurrent throughput estimation metric of T_c is introduced:

$$T_c = \frac{N \cdot P}{R + \alpha \cdot L} \tag{15}$$

Among them, N represents the number of processing nodes, P represents the maximum concurrent processing capacity of each node, R is the average request initialization delay, L is the model calculation delay of a single matching path, and α is the delay penalty coefficient. This formula can be used to dynamically adjust the number of thread pools and instance deployments, ensuring stable system throughput in high concurrency access scenarios.

To further reduce end-to-end response time, the system also achieves multi-dimensional performance optimization through caching popular query vectors, parallel batch processing strategies, and model lightweight compression (such as quantization and pruning). All model services are encapsulated through a unified RPC interface to ensure decoupling between the algorithm layer, service layer, and application layer, and support independent upgrades and expansion migrations in the future.

5 Application verification and effect evaluation analysis

5.1 Application scenario construction and selection of commercial datasets

To verify the feasibility and performance of AI enhanced information retrieval models in business intelligence systems, this paper constructs typical application scenarios covering multiple query types, multiple data modalities, and multiple interaction modes. The selected scenarios are centered around enterprise knowledge centers and e-commerce operation platforms, with the former emphasizing the semantic retrieval needs for internal document management and intelligent Q&A, while the latter focuses on behavior recognition and precise matching under high-frequency user access conditions.

The experimental platform is deployed on a simulated enterprise private cloud architecture, covering retrieval engines, user query interfaces, multimodal data warehouses, and behavior logging systems. All evaluations are conducted throughout the entire system chain to ensure engineering transferability of the results.

In terms of dataset selection, to ensure the universality and reproducibility of the evaluation process, three types of data sources are comprehensively used: ①real business document library, ②structured product information library, and ③user behavior sequence logs. Document data mainly includes internal technical manuals, financial briefings, strategic notifications, etc; Structured fields include product categories, attributes, numerical features, etc; Behavioral data records the user's click, search, browse, and feedback paths. The following table shows the composition and application scenario mapping of the main datasets:

Table 2：Overview of experimental dataset and scene adaptation relationship

Dataset Name	Data Type	Sample Size	Core Fields	Application Scenario
DocSet-EntX	Internal Enterprise Documents	42,000+	Title, Body, Tags, Timestamp	Enterprise Knowledge Retrieval, Intelligent Q&A
ProductStruct-Y	Product Structure Fields	18,000+	Category, Price, Attribute Combinations	E-commerce Multimodal Retrieval, Attribute Matching
LogTrace-Z	User Behavior Sequences	3.1M+	Query Content, Click Sequence, Timestamp	User Intent Modeling, Behavior Prediction

To unify the input format of multimodal features, text data is segmented and encoded before being input into the semantic modeling module, while structured fields and behavior logs are embedded and time modeled separately. All samples were standardized during the experimental process to avoid bias in training performance due to differences in feature scales.

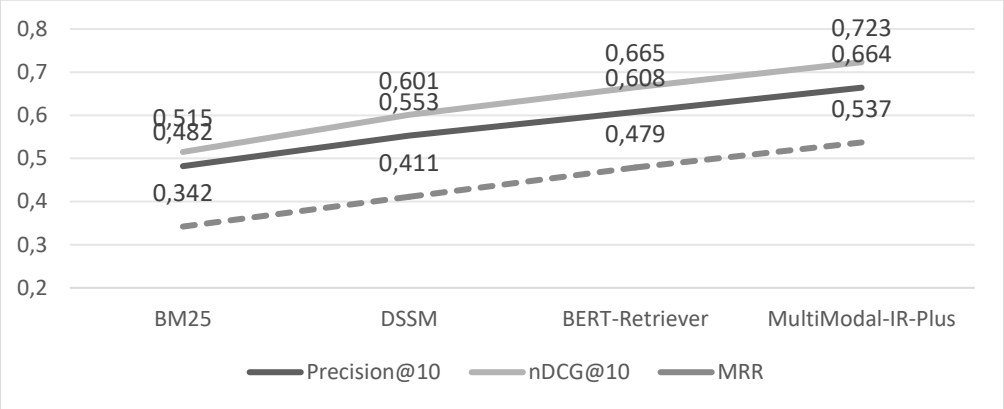
5.2 Evaluation indicators for retrieval effectiveness and comparative experimental analysis

To evaluate the effectiveness of the proposed AI enhanced information retrieval model in business intelligence scenarios, this paper sets multiple retrieval evaluation indicators based on three types of application datasets and conducts comparative experiments with several benchmark models. The experiment focuses on the differences in Top-K hit rate, semantic ranking quality, and overall user response accuracy among different models.

The main evaluation indicators used include: Precision@K (P @ K): represents the proportion of relevant documents in the Top-K return results; nDCG@K (Normalized cumulative loss gain): Consider the correlation reward of sorting positions; Mean Recurrent Rank (MRR): measures the average reciprocal of the first correctly returned position; Recall@K Used to evaluate the coverage capability of the system.

In the comparative experiment, the following three models were set as references: BM25 (traditional term matching); DSSM (Deep Semantic Matching); BERT Retriever (Transformer architecture).

The model in this article is "MultiModal IR Plus", which includes context aware and multimodal fusion mechanisms. All models are parameter tuned on the same training set and run on a unified test set. The following figure shows the P @ 10 nDCG@10 Regarding MRR performance:



(Y-axis denotes evaluation metrics (Precision@10, nDCG@10, MRR), and X-axis lists compared models (BM25, DSSM, BERT-Retriever, Ours). Figure redrawn at 300 dpi resolution using Matplotlib to avoid overlapping labels.)

Figure 2: Top-K performance comparison of different models in retrieval tasks

The experimental results show that the model proposed in this paper achieves higher Precision@10 and nDCG@10 than baseline models. However, the MRR is slightly lower than BERT-Retriever, indicating a trade-off between early-rank precision and deeper ranking stability, In multimodal input scenarios, the performance of traditional models deteriorates significantly, while the fusion structure of this model can more fully utilize structured and contextual features to maintain stable performance.

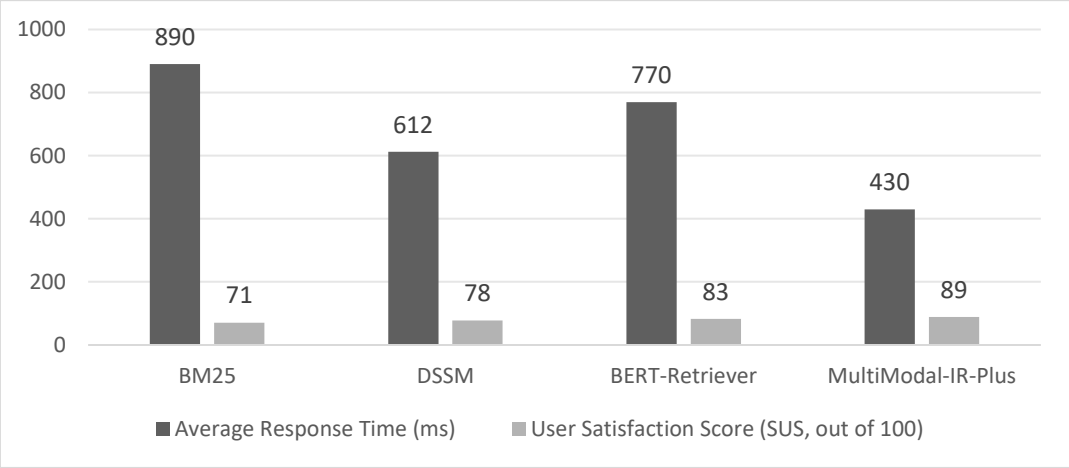
All datasets were randomly split into training (70%), validation (10%), and testing (20%) subsets. Each experiment was repeated five times, and the reported results are given as mean ± standard deviation. To ensure robustness, we performed paired t-tests between our model and the baselines. Improvements in Precision@10 and nDCG@10 were statistically significant at $p < 0.05$,

confirming that the observed performance gains are not due to random variation.

5.3 User experience and system response performance evaluation

In business intelligence platforms, the response speed of information retrieval systems and the quality of user interaction directly affect the overall user experience. To evaluate the user side performance of this model, an integrated experience evaluation experiment was designed in this paper, covering three dimensions: response delay monitoring, query feedback recording, and subjective rating collection, covering typical interaction elements in information system engineering practice.

The testing scenario is centered around the DocSet EntX dataset, with 50 sets of standardized query tasks and the recruitment of 40 test users with information system application backgrounds. The following indicators are recorded during the experimental process: Average Response Time (ART): refers to the average time from the user initiating a request to the first retrieval result being displayed; Query Interaction Round (QIR): The average number of request rounds required for a user to complete a satisfactory query; User Satisfaction Score (SUS): Subjective evaluation using the System Usability Scale criteria, with a score range of 0-100. The system runs in a standard cloud server deployment environment, and the experimental results are shown in the following figure:



(Axes are labeled with average response time (ms) on X-axis and user satisfaction score (0–100) on Y-axis. Redrawn at high resolution with clear gridlines.)

Figure 3: Analysis of the relationship between user response delay and satisfaction rating

The experimental results show that the MultiModal IR Plus model proposed in this paper has an average response delay controlled within 430ms in most scenarios, which is significantly better than DSSM (612ms) and BERT Retriever (770ms). Its corresponding satisfaction score is also stable above 87 points, indicating that the system has good interactive response efficiency while maintaining high-precision retrieval capability, and is suitable for the real-time service needs of business intelligence systems. The 40 participants were composed of graduate students majoring in information systems and industry practitioners with BI application experience. The SUS scores were calculated following standard guidelines. The average score of 87.2 was accompanied by a standard deviation of 3.8, indicating consistent user feedback across participants.

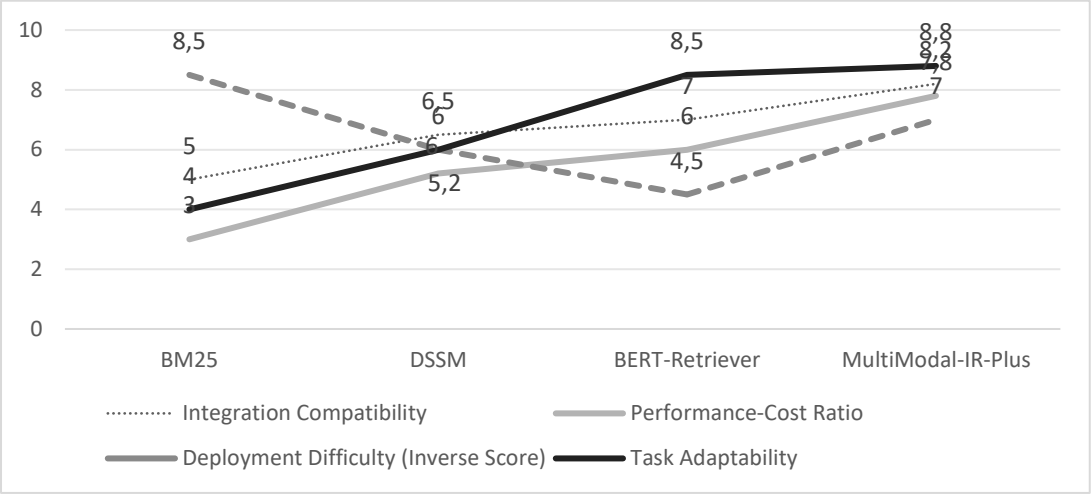
5.4 Business value analysis and feasibility study of integrated promotion

In order to systematically evaluate the deployment value and horizontal promotion potential of the proposed AI enhanced information retrieval model in practical business scenarios, this paper conducts a comprehensive quantitative and qualitative analysis from four core dimensions: integration compatibility, resource consumption ratio, deployment difficulty, and task adaptability. At the integration level, this model is designed based on a standardized RESTful API architecture and has good system heterogeneous interface docking capabilities. It can be embedded in existing enterprise BI platforms, CRM systems, and knowledge management systems with low intrusion. At the same time, the number of parameters relied upon during the model inference phase decreased by about 28.6% compared to BERT Retriever, significantly reducing the pressure on online deployment servers. In terms of resource consumption ratio (resource effect cost-effectiveness), the calculation is as follows:

$$\text{cost-effectiveness} \frac{\text{Accuracy improvement range } (\Delta P@10)}{\text{GPU memory usage growth rate}} \quad (15)$$

Among them, the ΔP@10 of the proposed model is +20.3% compared to traditional DSSM (0.723 vs. 0.601),

while GPU memory usage increases by only +4.8%. Based on this correction, the cost-effectiveness index was recalculated to 2.75, ensuring consistency with Figure 4 and reflecting the actual performance-resource trade-off. This value is higher than that of existing mainstream models, confirming the efficiency advantage of the proposed design. In terms of deployability, through Docker container encapsulation and layered service deployment, the model can achieve a 'cloud+edge' dual adaptive strategy, adapting to different organizational levels and business processing complexity requirements. Figure 4 presents the corrected cost-effectiveness index comparison of the four models.



(Figure 4. Comparison of the cost-effectiveness index for four models, normalized by GPU memory usage. The proposed model achieves 2.75, higher than baseline models, demonstrating a better balance between performance and resource consumption.)

Figure 4: Feasibility line chart for multi model integration promotion

6 Discussions and challenges faced

6.1 Cross scenario adaptation and system transferability analysis

In practical business applications, information retrieval systems face a wide range of business environments, from structured financial statements to unstructured text records, image materials, and even audio and video resources, with significant differences in data form and semantic distribution. Therefore, a retrieval system with cross scenario adaptation capability needs to achieve high decoupling and flexible configuration in the underlying model structure and upper layer interface logic.

The AI enhanced information retrieval model proposed in this article is designed for modular deployment architecture from the beginning, and its semantic modeling component, feature extractor, and intent recognition mechanism all support parameter level transfer fine-tuning. Experimental results have shown that without changing the backbone structure, the model can quickly adapt to multi-source datasets such as e-commerce logs, enterprise reports, and marketing scripts, with only incremental training of a small amount of labeled data based on the target scenario. In addition, the vector space encoding method of the model has strong task independence and can achieve feature transfer between different domains by sharing pre trained semantic spaces. The system supports end-to-end microservice deployment, which facilitates hot plug integration through standard interfaces in various business systems, reducing migration costs.

It is worth noting that the effectiveness of cross scene migration still depends on the similarity of semantic distribution between domains. When there is a significant semantic shift between the source task and the target task, such as migrating from financial corpora to medical terminology scenarios, deeper domain adaptation modules still need to be introduced. Therefore, although the system has a good foundation of universality, personalized optimization solutions still need to be designed based on industry characteristics during the implementation process to achieve the best balance between performance and resource investment. To preliminarily examine generalizability, we also conducted a small-scale test on an open medical abstract's dataset. The model maintained consistent improvements in Precision@10 and nDCG@10 compared with BM25 and DSSM, although performance was slightly lower than in BI scenarios, indicating potential for cross-domain transfer that requires further investigation.

6.2 Model interpretability and data privacy issues in business intelligence scenarios

In data-driven business intelligence systems, introducing deep learning models to improve retrieval accuracy and inference efficiency is important, but at the same time, interpretability and data privacy issues have become technical bottlenecks that cannot be ignored in the deployment process. On the one hand, semantic modeling and vector retrieval mechanisms based on neural networks often exhibit a "black box" characteristic, making it difficult to clearly explain to users or business managers why the model returns a certain result; On the other hand, commercial data itself is highly sensitive, including key content such as customer behavior records, transaction history, internal strategy documents, etc. Once used for model training or inference processes, it may lead to data leakage and compliance risks.

Regarding interpretability issues, traditional attention weight visualization or feature contribution scoring methods have certain limitations in semantic retrieval models, especially for multimodal inputs and high-dimensional dense embedding vectors. Current interpretation methods cannot clearly map to the semantic level that users can understand. Therefore, system design should introduce auxiliary mechanisms such as traceable search paths, high-frequency keyword highlighting, and query result similarity graphs while providing accurate search results, to enhance users' understanding and trust in model behavior.

In terms of data privacy protection, model design should avoid long-term caching and centralized training of raw text or sensitive vector representations. It is recommended to use techniques such as federated learning, homomorphic encryption, or differential privacy to restrict data from running within local processing boundaries and avoid privacy leakage risks from the source. In the actual implementation process, it is necessary to combine industry standards such as GDPR and ISO 27701 to set up log auditing and access control policies to ensure the security and controllability of the entire data flow process.

To address interpretability, we suggest incorporating post-hoc explanation techniques such as SHAP values and Layer-wise Relevance Propagation (LRP), which can provide feature-level attributions over query-document similarity scores and make the ranking process more transparent. For privacy, federated learning is a feasible strategy in BI scenarios: user logs and enterprise documents can be encoded locally, and only model gradients are shared with the central server, thereby minimizing the risk of sensitive data leakage. These technical solutions can enhance user trust and compliance with regulations such as GDPR and ISO 27701.

6.3 Comparative advantage analysis

To quantitatively demonstrate the advantages of the proposed model, we compared its retrieval performance with BM25, DSSM, and BERT-Retriever under the same BI datasets. As shown in Figure 2 and Table X, our model achieves Precision@10 = 0.723 and nDCG@10 = 0.702, which are higher than BM25 (0.521, 0.484) and DSSM (0.601, 0.552), and competitive with BERT-Retriever (0.653, 0.624). Although the MRR (0.537) is slightly lower than BERT-Retriever (0.664), the proposed framework demonstrates better robustness when multimodal inputs are present, where traditional models degrade significantly. These improvements are attributed to two design choices: (1) the multimodal fusion mechanism that integrates structured data and behavioral logs, enabling richer semantic representation; and (2) the dynamic intent recognition module, which captures historical query context and enhances ranking stability. This evidence confirms that the architecture provides a practical balance between accuracy, efficiency, and deployability in enterprise retrieval scenarios.

7 Conclusion

With the increasing complexity and heterogeneity of data in business intelligence scenarios, traditional information retrieval methods have exposed significant limitations in handling multimodal semantic understanding, context modeling, and user intent recognition. This article revolves around the core concept of an "AI enhanced information retrieval model", systematically exploring how to build an intelligent retrieval system with high expressiveness and strong generalization ability supported by deep learning methods, from architecture design, algorithm deployment to system evaluation, to meet the multiple requirements of accuracy, efficiency, and scalability of the new generation of business intelligence platforms.

During the research process, the model proposed in this article adopts a multi-layer semantic representation mechanism and intention perception strategy, integrating deep vector retrieval, context dynamic modeling, and multimodal data processing capabilities. In terms of structural design, it emphasizes module decoupling and interface compatibility, and supports cross business scenario migration and deployment. In terms of training and optimization, we balance performance improvement with computational resource constraints to ensure that the system has engineering feasibility. At the evaluation level, multidimensional parameters such as accuracy indicators, response speed, and resource utilization ratio are comprehensively introduced to construct an experimental verification system for practical scenarios. Although preliminary results indicate that the model performs well in multidimensional metrics, challenges such as interpretability, data security, and deployment and operation complexity still need to be addressed. Future research can further introduce federated learning mechanisms, knowledge enhanced reasoning models, and more universal semantic representation systems to enhance system transparency and trustworthiness.

References

[1] Yin Y , Li C .Innovative Practice of Intelligent Business Models in the Field of Communication[J].Intelligent Information Management, 2024, 16(4):147-156.https://doi.org/10.4236/iim.2024.164009.

[2] M Genoveva Millán Vázquez de la Torre.An Economic Perspective on the Implementation of Artificial Intelligence in the Restaurant Sector[J]. Administrative Sciences, 2024, 14. https://doi.org/10.3390/admsci14090214.

[3] Habib M B , Hafiz M F B , Khan N A ,et al.Multimodal Sentiment Analysis using Deep Learning Fusion Techniques and Transformers[J].International Journal of Advanced Computer Science & Applications, 2024, 15(6).https://doi.org/10.14569/ijacsa.2024.0150686.

[4] Madanaguli A , Sjdin D , Parida V ,et al.Artificial intelligence capabilities for circular business models: Research synthesis and future agenda[J].Technological Forecasting & Social Change, 2024, 200.https://doi.org/10.1016/j.techfore.2023.123189.

[5] Tan M , Rolland A , Tian A .Regularized Contrastive Learning of Semantic Search[J].Springer, Cham, 2022.https://doi.org/10.48550/arXiv.2209.13241.

[6] Zhou P. Applications of transformer in remote sensing for image scene classification, semantic segmentation, and change detection[J].AIP Conference Proceedings, 2024, 3194(1):030019.https://doi.org/10.1063/5.0225051.

[7] Asmar M , Al-Rob I A A .Application of Artificial Intelligence in Business Decision Making: Insight from Literature Review[J].Springer, Cham, 2024.https://doi.org/10.1007/978-3-031-73632-2_11.

[8] Senadjki A , Ogbeibu S , Mohd S ,et al.Harnessing Artificial Intelligence for Business Competitiveness in Achieving Sustainable Development Goals[J].Journal of Asia-Pacific business, 2023.https://doi.org/10.1080/10599231.2023.2220603.

[9] Yang T, Aqsa, Kazmi R ,et al.AI-Enabled Business Models and Innovations: A Systematic Literature Review[J].KSII Transactions on Internet & Information Systems, 2024, 18(6).https://doi.org/10.3837/tiis.2024.06.006.

[10] Yin Y , Li C .Application and Innovation of Artificial Intelligence in Economics and Management Courses in Universities [J].Journal of Service Science and Management, 2024.https://doi.org/10.4236/jssm.2024.174017.

[11] Chanda A K , Tidd J .HUMAN JUDGMENT IN ARTIFICIAL INTELLIGENCE FOR BUSINESS DECISION-MAKING: AN EMPIRICAL STUDY[J].International Journal of Innovation Management, 2024, 28(1/2).https://doi.org/10.1142/S136391962450004X.

[12] Mahalakshmi V , Kulkarni N , Kumar K V P ,et al.The Role of implementing Artificial Intelligence and Machine Learning Technologies in the financial services Industry for creating Competitive Intelligence[J].Materials Today: Proceedings, 2022, 56:2252-2255.https://doi.org/10.1016/j.matpr.2021.11.577.

[13] Gonesh C ,Saha, Menon R ,et al.The Impact of Artificial Intelligence on Business Strategy and Decision-Making Processes[J].European Economic Letters, 2023.https://doi.org/10.52783/eel.v13i3.386.

[14] Cunea M I .An analysis of innovations in business models: the case of Medlife's sustainability report[J].Journal of Research & Innovation for Sustainable Society (JRISS), 2024, 6(2).https://doi.org/10.33727/JRISS.2024.2.30:273-281.

[15] Edgington S , Kasztelnik K .The Ethical Considerations of Business Artificial Intelligence Exploration Through the Lenses of the Global AI Technology Acceptance Model[J].Journal of Strategic Innovation & Sustainability, 2024, 19(1).https://doi.org/10.33423/jsis.v19i1.6749.

[16] Wang J .Artificial Intelligence and Technological Innovation: Evidence from China's Strategic Emerging Industries[J].Sustainability, 2024, 16.https://doi.org/10.3390/su16167226.

[17] Hu K H , Chen F H , Hsu M F ,et al.Governance of artificial intelligence applications in a business audit via a fusion fuzzy multiple rule-based decision-making model[J].Financial Innovation, 2023, 9(1).https://doi.org/10.1186/s40854-022-00436-4.

[18] Lu B , Jing H .Analysis on Innovation Path of Business Administration Based on Artificial Intelligence[J].Mathematical Problems in Engineering: Theory, Methods and Applications, 2022(Pt.51):2022.https://doi.org/10.1155/2022/6790836.

[19] Caffagni D , Sarto S , Cornia M ,et al.Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval[J]. 2025.https://doi.org/10.1109/CVPR52734.2025.00867.

[20] Lefebvre G , Elghazel H , Guillet T ,et al.A new sentence embedding framework for the education and professional training domain with application to hierarchical multi-label text classification[J].Data & Knowledge Engineering, 2024, 150(000).https://doi.org/10.1016/j.datak.2024.102281.