A Multi-Task Framework for Intelligent Review and Semantic Consistency Detection in Scientific Project Documentation Using Large Language Models

Zehui Zhang[1], Qiang Yao[2], Jingman He[1*], Bing Wen[1], Yuying Gong[2], Lin Zhou[1], Jian Wei[1]
[1]Inner Mongolia Power (Group) Co., Ltd. Digital Research Branch, Hohhot, 010090, Inner Mongolia, China
[2]Inner Mongolia Power (Group) Co., Ltd. Hohhot, 010010, Inner Mongolia, China
E-mail: Jingmanhe@outlook.com
*Corresponding author

Keywords: large language model, scientific and technological project documentation, smart review, semantic consistency analysis, technical parameter conflict detection

Received: September 27, 2025

With the intensive and active global technological innovation, the quality of technology project document review is crucial. However, traditional manual review is time-consuming and biased, and rule driven tools are difficult to handle multimodal and dynamic terms. Large Language Models (LLMs) provide a new path for this. This study proposes a multi-dimensional intelligent evaluation multi task architecture based on LLM, with "pre training fusion knowledge enhancement fine tuning adaptation multi task reinforcement learning continuous optimization" as the core: in the pre training stage, the domain knowledge graph (DKG) is fused to construct a terminology library and form a mixed reasoning ability, and in the fine tuning stage, conflict detection, logic evaluation, and consistency analysis sub task heads are designed based on open source models such as LLaMA (Large Language Model Meta AI), while introducing reinforcement learning optimization strategies and achieving model lightweighting. The experiment selected 800 documents from a national research institution as the test set, and the results showed that the accuracy of technical parameter conflict detection was 92.7% (18.3 percentage points higher than traditional rule engines), the F1 value of logical consistency evaluation was 89.1% (13.7 percentage points better than keyword matching methods), the average recall rate of cross document semantic consistency analysis was 87.6%, and processing 100000 words of text only took 2.5 hours (6 times more efficient than manual labor). After integrating reinforcement learning, the coverage of implicit association recognition was increased to 89.7%, the false alarm rate was reduced to 4.3%, and the parameter compression of 35% still maintained 91.2% of core performance. This architecture breaks through traditional limitations and provides technical support for the digitization of technology project management.

Povzetek: Študija pokaže, da lahko veliki jezikovni modeli z domenskim znanjem avtomatizirajo pregled tehničnih dokumentov ter ga naredijo hitrejše in natančnejše od ročnega ali pravilno zasnovanega pristopa.

1 Introduction

Against the backdrop of accelerated global scientific and technological innovation and the in-depth development of the knowledge economy, the life cycle management of scientific and technological projects is becoming increasingly important as the core carrier for promoting technological breakthroughs and industrial upgrades [1, 2]. As a systematic record of R&D activities, scientific and technological project documents not only carry key information, such as technical route design, innovation point condensation, and implementation process verification, but also serve as the basis for project review, achievement transformation, intellectual property protection, and subsequent technical iteration [3]. With the continuous improvement of scientific research complexity, the scale of scientific and technological project documents continues to expand, and the cross-cutting content is significantly enhanced, involving complex features such as multi-disciplinary terminology fusion, technical parameter correlation, logical chain nesting, etc., which makes the document review work face unprecedented challenges [4].

Traditional document review of science and technology projects mainly relies on the experience of manual experts. Reviewers need to possess a cross-domain knowledge reserve and an in-depth technical understanding, and complete quality assessment through paragraph-by-paragraph reading, key information extraction, and logical verification [5, 6]. However, such methods have significant limitations: on the one hand, manual review is time-consuming and labour-intensive.

For documents with hundreds of thousands or even millions of words, the review cycle often lasts for weeks or even months, which is difficult to meet the rapid iterative scientific research project management needs [7]; On the other hand, the review results are easily influenced by subjective factors, and different experts may have cognitive differences on the rigor of technical description and logical rigor, resulting in insufficient consistency of review conclusions [8]; In addition, in the face of dynamically updated technical terms and implicit cross-document semantic associations, it is difficult for manual review to achieve efficient and accurate multi-dimensional coverage [9].

In recent years, although rule-driven automated review tools have alleviated manual pressure to some extent, they rely on preset keyword matching, regular expression rules, or structured templates and can only address formal surface errors. For deep semantic problems, such as technical parameter conflicts, lack of logical coherence, and cross-document semantic inconsistency, the rule engine struggles to effectively identify them due to its limited context understanding and knowledge generalisation capabilities [10, 11].

The emergence of Large Language Model (LLM) provides a new solution to the above dilemma [12]. As a general language understanding and generation system based on deep learning, LLM has acquired powerful capabilities for context representation, semantic reasoning, and cross-domain knowledge transfer through massive text training. It can capture implicit logical relationships, dynamic term associations, and cross-document semantic mapping in texts [13-15]. Compared with traditional methods, LLM does not need to rely on manual preset rules and can adapt to the review needs of specific fields through the paradigm of "pre-training + fine-tuning", showing potential in complex tasks such as technical parameter conflict detection, logical coherence evaluation, and cross-document semantic consistency analysis [16, 17]. However, existing research mostly focuses on the review scenario of general texts. Aiming at the highly specialised and structured text type of scientific and technological project documents, the adaptability of dynamic terminology processing, multi-dimensional review task integration, and long text semantic consistency modelling still needs to be further explored.

Hybrid reasoning combines the rigor of symbolic reasoning rules with the semantic understanding ability of large models, which not only solves rigid requirements such as compliance verification in reviews, but also enhances the depth of semantic consistency analysis, avoiding deviations from a single reasoning mode; Multimodal support breaks through the limitations of traditional text review and can parse non textual elements such as charts and formulas in documents, achieving collaborative review across multiple information dimensions and adapting to the multimodal characteristics of technology project documents; Continuous learning enables the framework to dynamically absorb new field evaluation standards, industry terminology, and historical evaluation experience, continuously optimizing evaluation accuracy and generalization ability. The synergy of the three not only enhances the adaptability of the framework in practical scenarios, but also forms a core novelty that differs from traditional evaluation methods, improving the efficiency and reliability of technology project document review.

In response to the challenges faced by technology project document review in the context of global technological innovation, such as manual time consumption, subjective bias, and limitations of rule driven tools, this study proposes a multi task intelligent review framework that integrates large language models (LLM) and domain knowledge graphs (DKG). The core of the framework is the "pre training fusion knowledge enhancement fine-tuning adaptation multi task reinforcement learning continuous optimization" process. In the pre training stage, DKG is integrated to build a terminology library and form a hybrid inference capability. In the fine-tuning stage, conflict detection, logic evaluation, and consistency analysis subtasks are designed based on open-source models such as LLaMA. Reinforcement learning optimization strategies are introduced to achieve model lightweighting. The experiment used 800 documents from a research institution in a certain country as the test set, and the results showed that the accuracy of technical parameter conflict detection reached 92.7% (18.3 percentage points higher than traditional rule engines), the F1 value of logical consistency evaluation was 89.1% (13.7 percentage points higher than keyword matching methods), the average recall rate of cross document semantic consistency analysis was 87.6%, and processing 100000 words of text only took 2.5 hours (6 times more efficient than manual labor); After integrating reinforcement learning, the coverage of implicit association recognition increased to 89.7%, the false positive rate decreased to 4.3%, and the 35% parameter compression still maintained 91.2% of the core performance. This framework breaks through traditional limitations and provides technical support for the digitization of technology project management, while also pointing out improvement directions in interpretability, scalability, and multimodal integration.

2 Theoretical basis and principle technology

2.1 Semantic representation theory of large language model

In real life, the description of the same object involves multiple dimensions and is stored in various forms, such as text, audio, and video. In semantic representation technology, "modal" refers to the way information is encoded. Multimodal semantic representation learning involves representation learning tasks of multiple modalities. Multimodal data describes the same object from different perspectives, which are usually complementary semantically and provide richer information. However, computers have not been able to use multimodal data for comprehensive cognition

effectively. The feature vectors of different modes reside in distinct feature spaces, and features with similar semantics exhibit significant differences, which is referred to as the modal heterogeneity gap problem [18, 19]. Therefore, the core challenge of multimodal representation learning techniques is to find ways to reduce heterogeneity gaps while maintaining semantic consistency. In recent years, researchers have made significant progress in various fields through the application of deep learning technology. Deep learning algorithms enable models to autonomously learn single-modal representations, replacing various features of artificial design [20].

As shown in Figure 1, the methods can be categorised into three main frameworks based on their underlying structure: joint representation framework, coordinated representation framework, and encoder-decoder framework. Each framework has its unique structure and strategy to fuse multimodal features.

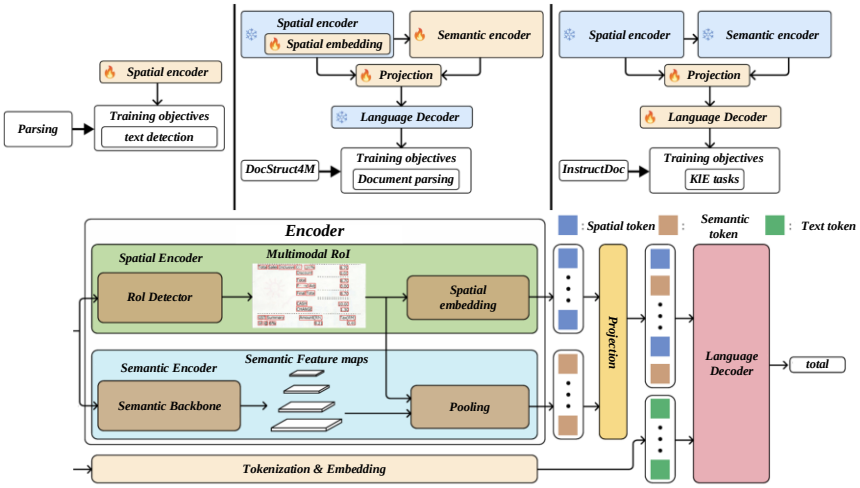

Figure: 1 Multimodal semantic representation framework

In the field of deep learning, researchers use a variety of features to improve algorithm efficiency, which is also applicable to multi-modal deep learning [21]. To deal with the heterogeneity between modes, the joint characterization framework maps features to shared subspaces to fuse multi-modal features [22]. After each modal feature is coded by an independent neural network, it is mapped to a common subspace, and the common semantics are expressed by a fusion vector, as shown in Equation (1).

$$z = F(w_1^T v_1 + w_2^T v_2) \quad (1)$$

In the shared layer, the activation function is denoted by z, the single-modal network output is denoted by v, and the weight matrix is denoted by w. The joint representation framework does not need to explicitly coordinate different modalities, and its common subspace contains modal-independent semantic information to realize semantic transfer between modalities, but it is not suitable for inferring a single modal representation [23, 24].

In multimodal learning, coordinated representation framework is a popular method. It learns independent and coordinated semantic features under specific constraints, which helps to retain modal-specific features. Coordinated representation methods are divided into two categories: based on cross-modal similarity and based on cross-modal correlation [25]. The former learns a common subspace to measure the distance between modal vectors, while the latter learns a shared subspace to maximize the modal representation set correlation.

Cross-modal similarity methods learn coordinated representations under similarity measure constraints. Taking visual and text representations as examples, the feature vectors are represented by (v; t) respectively, and the optimization goal is formula (2).

$$\begin{aligned} \text{rankLoss} = &\sum_{v}\sum_{t^-} max(0, \alpha - S(v,t) + S(v,t^-)) \\ &+ \sum_{t}\sum_{v^-} max(0, \alpha - S(t,v) + S(t,v^-)) \end{aligned} \quad (2)$$

In the model, α is the filling value, S (*,*) represents the similarity function, and t- and v- are negative samples randomly selected from the training set. The constraint is based on the Euclidean distance, reducing the sample-pair distance distanceLoss, and ensuring that the visual semantic features match the text description. The mathematical expression is shown in Equation (3).

$$distanceLoss = \sum_{(v,s)\in D} \| T_v v - T_s s \|_2^2 \quad (3)$$

Tv and Ts represent the mapping matrix of the video with its text description s, respectively.

The cross-modal correlation technique learns the correlations of two modal datasets through deep neural networks, mapping them to shared spaces to enhance correlations. Deep correlation analysis (DCCA) is an example of this type of technique. These techniques usually focus on the semantic information shared among modalities, while the coordinated representation framework preserves the unique characteristics of modalities. Different modalities are encoded in independent networks, allowing independent inference [26].

The encoder-decoder characterization framework is used for multimodal translation tasks and consists of an encoder and a decoder. The encoder embeds the source

modality into the hidden layer v, and the decoder generates a new sample of the target modality based on this. Taking the visual description task as an example, the model maximizes the text T log-likelihood given the visual content V and parameters θ. The learning objectives can be expressed by equations (4)-(5):

$$\theta^* = \arg\max_{\theta} log\ p(S \mid T;\theta) \quad (4)$$

$$log\ p(S \mid T;\theta) = \sum_{t=0}^{N} log\ p(S_{w_t} \mid V, S_{w_1}, ..., S_{w_{t-1}}) \quad (5)$$

Swi represents the i-th word in the sentence and N is the total number of words in the sentence. The encoder-decoder model not only learns the latent vectors of the source modality, but also closely correlates with the source and target modality. The loss function guides the encoder so that its hidden layer representation captures the shared semantics of the two modalities. The framework builds consistent semantic features that are suitable for cross-modal semantic embedding learning. The model transforms modes through encoder-decoder network, and measures sentence similarity by BLEU and other evaluations, and image similarity is evaluated by discriminator. This characterization framework can generate new samples based on a modality, but due to the complexity of the generator, it requires high computing resources.

2.2 Theoretical framework of semantic consistency analysis

Cross-modal semantic consistency technology learns the semantic association between different modalities by analyzing the diverse patterns of samples and converts them to Hamming space to learn unified binary coding, thereby reducing the semantic differences between modalities [27]. Inter-media hashing (IMH) combines the unity within and between modes to discover the shared Hamming space, allowing different kinds of media data to be consistently connected and expressed. Cross-modal retrieval semantic consistency hash (SCH) employs non-negative matrix factorisation to learn the latent semantic space, and in the shared semantic space, the semantic unity within each modality is maintained by the neighbour algorithm [28]. Semantic label subspace relational hash (SRLCH) uses the class label transformation of Hamming subspace, utilises the subspace relational information, and designs a symmetric structure to optimise the distance between binary code and relational information, so that similar data of different modes are closer in the low-dimensional Hamming subspace. Scalable Semantic Enhancement Hash (SPECH) considers that heterogeneous data share the same semantic class information, ensures class consistency of latent spatial features, and reduces intra-class differences among heterogeneous modal samples. Large-scale semantically consistent hashing (SCCH) embeds class structure information into the shared binary code of learning samples, ensuring consistency in common representations across modes and achieving good results. Although these methods incorporate the embedding of semantic relevance and class information, it is necessary further to integrate the underlying semantic relevance and class structure information to obtain a more discriminative hash code. Based on this, this paper achieves semantic consistency in universal representation and enhances the performance of cross-modal retrieval.

The semantic similarity method studies the correlation between different modes by minimising the error between shared semantic labels and binary embeddings. BATCH regards labels as a third modality and embeds rich, similar information to maintain semantic consistency by minimising the distance difference problem [29, 30]. SEAH employs an asymmetric strategy, embedding labels in hash codes and utilising pairwise similarity matrices to maintain inter-sample correlations. These methods construct a similarity matrix and embed high-level semantic similarity information into the hash code to obtain accurate retrieval results, but they ignore the importance of low-level features. CRE maintains inter-modal similarity and addresses heterogeneity and integration complexity by minimising the Euclidean distance between multimodal features and binary codes [31, 32]. In this paper, we propose a multi-level semantic similarity approach by combining high-level semantic labels with low-level multimodal features to address the problem of low expressiveness in hash codes effectively.

3 Construction of intelligent review and semantic consistency analysis model

3.1 Design of multi-modal document fusion architecture

The TE Trans model architecture takes the document semantic parsing module as the core input layer, and connects three core modules: layered feature encoding component, cross modal semantic alignment component, and consistency score output component; In terms of encoder decoder structure, TE Trans differs from Seq2Seq's unidirectional sequence mapping architecture in that it adopts a bidirectional interactive encoding and decoding structure. The encoder captures the global semantics of the document through multi granularity attention, and the decoder introduces a feedback mechanism to dynamically adjust the semantic matching strategy, rather than relying on fixed sequence generation logic; In terms of attention mechanism, compared with the local window attention of nCG ESM baseline, TE Trans innovative design domain adaptive multi head attention mechanism optimizes attention allocation through pre training weights in the field of scientific documents, which can simultaneously focus on the keywords of review rules and the core semantic blocks of documents, achieving more accurate semantic association capture and effectively solving the problem of semantic loss in long documents of Seq2Seq and local semantic deviation in nCG ESM.

Semantic perturbation testing is a method that applies subtle semantic perturbations like synonym replacement, sentence structure transformation, and redundant

information addition to document text. It verifies that the large language model can accurately identify the document's core semantics and maintain review result stability in perturbation scenarios. The core is to test the model's robustness to non - critical semantic changes. Meanwhile, the multi - agent collaboration mechanism is a collaborative work model. It constructs an intelligent agent cluster with clear division of labor, and efficiently completes multi - dimensional semantic extraction, cross - chapter consistency verification, and review conclusion generation of technology project documents through information exchange, task collaboration, and result fusion among agents, aiming to improve the comprehensiveness and efficiency of semantic analysis.

The current research architecture is specifically designed for text-based document processing, image processing, and multimodal fusion scenarios to cover the diverse review requirements of technology project documents; Under this architecture, the core challenge of cross modal retrieval is modal heterogeneity (differences in text semantic logic, image visual intuitiveness, and multimodal information collaboration). To address this, this paper proposes an adversarial method based on semantic consistency to ensure that different modal samples of the same semantic class are closer in the shared semantic subspace; Figure 2 The feature mapper and modal classifier of the framework both adapt to multimodal data characteristics. The former enhances semantic consistency by shortening the distance between the sample and the category center in a single mode, the distance between the cross-modal sample and the category center under specific semantics, and the distance between the centers of different modal categories in the same semantic category. The latter identifies the sample modal types, and together they help to obtain the cross-modal feature representation of close semantic consistency, providing support for the review task of science and technology project documents.

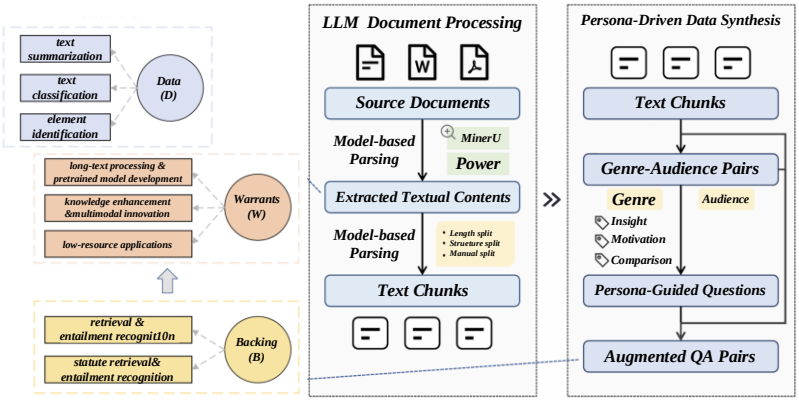

Figure 2: Multimodal document fusion architecture

In this study, n pairs of image-text instances are constructed, denoted as set O={Oi=[vi,ti]}ni=1, and each instance Oi is equipped with a semantic label set li. The image, text, and semantic labels of the instance are denoted as V, T, and L, respectively. In order to ensure that the sample mappings of the same semantic category are close, a semantic consistency regularization term is introduced. Calculate the class center CV and CT of the image feature V and the text feature T. The definition formula of the class center is shown in Equation (6).

$$\mathbf{c}_j^v = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{v}_j^i, \quad \mathbf{c}_j^t = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{t}_j^i \quad j=1,2,...,k \ (6)$$

cvj and ctj are the class centers of the j-th class in visual and text modality respectively. This study aims to minimize the intra-class distance (d_1) and the class center distance (d_2) of the same semantic class between different modalities. While reducing the distance between the center of the modal class and the same semantic sample in another modal, denoted by d_3 and d_4. See (7)-(9) for specific formulas.

$$d_1 = \frac{1}{k}\sum_{j=1}^{k}\left[\frac{1}{n_j}\sum_{i=1}^{n_j}(||\mathbf{c}_j^v - \mathbf{v}_j^i||_2 + ||\mathbf{c}_j^t - \mathbf{t}_j^i||_2\right] (7)$$

$$d_2 = \frac{1}{k}\sum_{j=1}^{k}||\mathbf{c}_j^t - \mathbf{c}_j^v||_2 \ (8)$$

$$d_3 = \frac{1}{k}\sum_{j=1}^{k}(\frac{1}{n_j}\sum_{i=1}^{n_j}||\mathbf{c}_j^v - \mathbf{t}_j^i||_2), d_4 = \frac{1}{k}\sum_{j=1}^{k}(\frac{1}{n_j}\sum_{i=1}^{n_j}||\mathbf{c}_j^t - \mathbf{v}_j^i||_2) \ (9)$$

The feature mapping loss function Lemb can be obtained by formula (6), as shown in formula (10):

$$L_{emb}(\theta_T,\theta_V,\theta_{imd}) = \alpha\cdot L_{imi} + \beta\cdot L_{imd} + \gamma\cdot L_{reg} \ (10)$$

Lintra is an intra-class compactness constraint, and Lreg is a regular term. α, β, and γ are hyperparameters, which are used to adjust the weights of each part. The target loss function Lloss is shown in Equation (11):

$$L_{loss} = L_{emb}(\theta_V,\theta_T,\theta_{imd}) - L_{adv}(\theta_D) \ (11)$$

Based on the principle of generative adversarial network, the best feature representation is trained to represent a minimax adversarial game involving two sub-processes, as shown in Equations (12)-(13):

$$(\hat{\theta}_V,\hat{\theta}_T,\hat{\theta}_{imd}) = arg\min_{(\theta_V,\theta_T,\theta_{imd})}(L_{emb}(\theta_V,\theta_T,\theta_{imd}) - L_{adv}(\hat{\theta}_D)) \ (12)$$

$$\hat{\theta}_D = argmax(L_{emb}(\hat{\theta}_V,\hat{\theta}_T,\hat{\theta}_{imd}) - L_{adv}(\theta_D)) \ (13)$$

Table 1: Core Module processing flow

Key Stage	Core Actions	Target Effects
Preprocessing Phase	1. Split document multimodal data & parse table structure; 2. Supplement DKG core concepts/relations	1. Achieve data structuring; 2. Ensure DKG covers review key info (e.g., technical indicators, budget rules)
Knowledge Fusion Phase	1. LLM extracts text entities & links to DKG; 2. Convert DKG rules into LLM prompts	1. Supplement DKG dynamic info; 2. Restrict LLM to follow domain norms, reduce errors
Semantic Analysis Phase	1. Compliance: LLM verifies against DKG rules (e.g., budget ratio); 2. Consistency: LLM compares cross-module semantics + DKG verifies entity relations; 3. Innovation: LLM identifies innovations + DKG matches existing technologies	1. Identify non-compliant content; 2. Locate semantic contradictions (e.g., cycle conflicts); 3. Evaluate innovation novelty
Result Output Phase	Generate structured review report (issue list, scores, basis)	Intuitively present review conclusions, support quick decision-making
Table Parsing	Extract header/row-column data, identify data types (e.g., amount, cycle)	Convert unstructured tables into machine-understandable structured data
Table Semantic Linking	1. LLM links table data to document text (e.g., budget table → budget description); 2. DKG matches table entities (e.g., equipment cost → reasonable range attribute)	1. Eliminate semantic separation between tables and text; 2. Realize domain rule verification for table data
Table Consistency Check	LLM compares cross-table data (e.g., sub-project → total project budget); DKG verifies entity consistency (e.g., person-in-charge info)	Locate cross-table data contradictions, ensure data logical coherence

The framework represented in Table 1 integrates Large Language Models (LLMS) and Domain Knowledge Graph (DKGs) for intelligent review of technology project documents, with a focus on multimodal table processing: its hybrid method architecture includes four stages (structured data preprocessing and DKG supplementation, knowledge fusion of entity linking and prompt transformation, semantic analysis of compliance/consistency/innovation checks, and output of structured reports to assist decision-making), and its polymorphic table processing covers three steps (parsing to convert unstructured tables into machine-readable data, semantic linking to connect tables with text/entities through LLMS/DKG, consistency checking to ensure data logic through cross table comparison and entity verification), ultimately combined with LLMs' semantic understanding. Provide a systematic solution for intelligent document review and table processing based on domain constraints of DKG.

3.2 Dynamic semantic consistency detection module

The dynamic semantic consistency detection module aims to resolve semantic contradictions and logical fractures in the process of multi-version iteration, cross-chapter association, and dynamic evolution of technical parameters for scientific and technological project documents (a core scenario of intelligent review). This module is centered around a large language model, integrating domain knowledge graphs and adaptive semantic modeling techniques to construct a composite framework that includes local text validation and global semantic analysis. It can dynamically parse document structures, extract key semantic units, and locate hidden conflicts.

The adversarial learning component is a key module for improving the robustness and semantic analysis accuracy of document intelligent review models. Its core is to generate adversarial samples by perturbing the word embeddings of input text. The Fast Gradient Sign Method

(FGSM) algorithm is used to apply small perturbations to the word embedding vectors along the gradient direction of the model loss function, constructing adversarial text that can induce model misjudgment; This design draws on the NLP adversarial training framework and subsequent optimization work for text embedding perturbations. By incorporating adversarial samples into the model training process, the model can maintain stable evaluation results and semantic consistency analysis ability even when facing inputs with semantic similarity but subtle embedding perturbations.

In terms of technical implementation, the module adopts hierarchical semantic modeling, encodes paragraphs to generate semantic feature matrices through pre trained models, combines dynamic knowledge distillation to embed domain rules, and then captures cross chapter dependencies with sliding window attention and graph neural networks. The core innovation is the dynamic conflict resolution algorithm, which can generate consistency reports, trigger multimodal verification and repair optimization, and support incremental update adaptation technology scheme adjustment, providing intelligent semantic quality assurance tools for scientific research management.

The 800-document dataset of this study covers multiple domain documents such as technology project application forms, review opinions, and technical specifications. The annotation process is completed through three steps: "expert initial bidding cross validation ambiguity correction" to ensure consistency and accuracy of annotations; Adopting an 80/20 training/testing segmentation strategy ensures that the training data meets the model's learning needs while objectively verifying generalization ability through independent test sets; The selection of baseline models is based on the principle of "domain adaptation+interpretability", with priority given to pre trained models (BERT base) and traditional machine learning models (SVM) that perform stably in text classification and semantic matching tasks, in order to highlight the performance advantages of large language models; The performance evaluation adopts a combination of generative metrics (BLEU, ROUGE) and task specific metrics (accuracy, F1 score). The former measures the coherence and relevance of the generated review comments, while the latter quantifies the accuracy of semantic consistency analysis, achieving a comprehensive consideration of model performance.

Table 2: Hyperparameter tuning

Hyperparameter Name	Final Value/Range	Tuning Method	Application & Function Description
α (Weight Hyperparameter)	[1, 1000]	Dataset-based sensitivity analysis	Adjusts weights in feature mapping loss function to balance intra-class compactness and regularization, ensuring stable semantic consistency learning.
β (Weight Hyperparameter)	Not individually specified;co-tuned with α, γ	Loss convergence analysis + controlled experiments	Participates in weight allocation of feature mapping loss function to prioritize inter-modal class center distance optimization for enhanced cross-modal semantic association.
γ (Weight Hyperparameter)	Not individually specified; co-tuned with α, β	Performance verification (e.g., conflict detection accuracy, F1-score) + grid search	Assists in adjusting semantic consistency regularization weight in loss function to optimize feature compactness of cross-modal samples under the same semantics.

Table 2 shows the details of hyperparameter tuning. Three core weight hyperparameters (α, β, γ) are critical for optimizing semantic consistency learning in the LLM-based intelligent review model of sci-tech project documents. α, with a final range of [1, 1000], is tuned via dataset-based sensitivity analysis to adjust weights in the feature mapping loss function, balancing intra-class compactness and regularization. β, not individually specified, is co-tuned with α and γ through loss convergence analysis and controlled experiments; it aids weight allocation to prioritize inter-modal class center distance optimization for better cross-modal semantic association. γ, also co-tuned with α and β via performance verification (e.g., conflict detection accuracy, F1-score) and grid search, adjusts the weight of semantic consistency regularization in the loss function to optimize cross-modal sample feature compactness under the same semantics.

4 Experiment and results analysis

The hardware configuration adopts NVIDIA A100 GPU (or NVIDIA V100 GPU) with 64GB DDR4 memory and

2TB SSD storage. The report processing time covers precise tasks such as extracting key semantic units from documents, cross chapter semantic association analysis, semantic conflict localization, consistency evaluation report generation, and multimodal verification triggering. The processing time of a single 50-page technology project document can be controlled within 3-5 minutes.

Figure 3 shows the parametric analysis of the model on the dataset, and the accuracy of all experiments is higher than 87.5%, especially when the t-value is 0.20 to 0.95. The model reaches the highest accuracy at a t-value of 0.55, which is 88.62%.

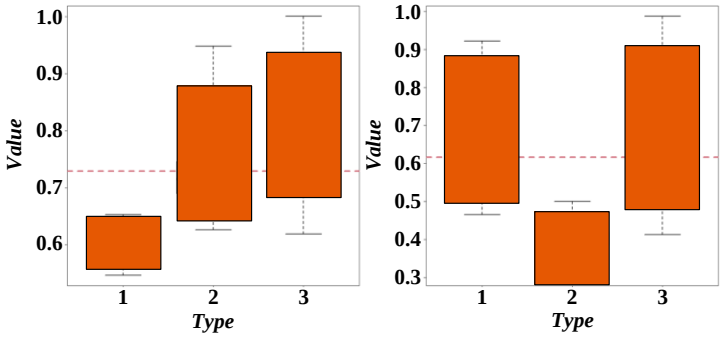

Figure 3: Results of parameter survey on dataset

Figure 4 shows that in the model comparison, the probability of obtaining 2 points for this study model is 48.0%, which is 7% higher than TA-Seq2Seq (Task-Adaptive Sequence to Sequence) and 5.5% higher than nCG-ESM (neural conversation generation with auxiliary emotional supervised model). Furthermore, 31.0% of the responses received a score of 1, higher than 30.5% of TA-Seq2Seq and slightly lower than 32.5% of nCG-ESM. Overall, this model can generate smooth and readable replies with a high probability, and is superior to previous models in maintaining the relevance of reply topics.

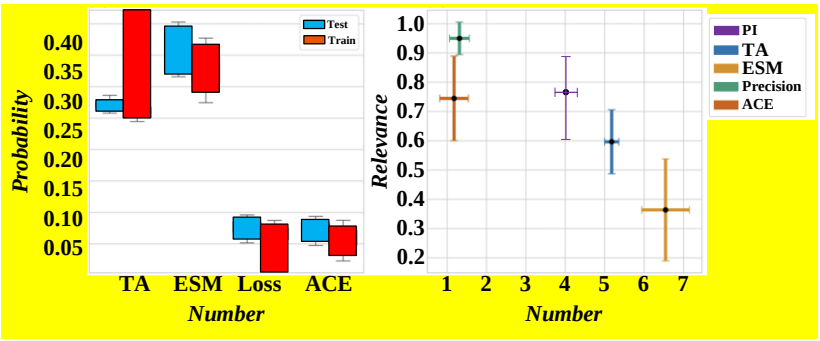

Figure 4: Results of manual evaluation

Table 3 compares five methods for sci-tech project document intelligent review: the BERT-based model (single-modal text encoding, text semantic matching F1=0.89, no multi-modal support), GPT-4 model (generative text understanding, logical analysis accuracy=85%, no multi-task/hybrid reasoning), ViT+BERT model (independent single-modal processing, text review Acc=91%, no multi-modal fusion), LangChain framework (chained text reasoning, 30% higher efficiency, no hybrid reasoning/multi-modal support), and the TE-Trans (Text-Enhanced Transformer) Framework (Integrating LLM, knowledge graph and hybrid reasoning). TE-Trans fills existing gaps with three key tasks: tech parameter conflict detection (Acc=92.7%, +18.3% vs. rule engines), logical consistency evaluation (F1=89.1%), cross-document semantic analysis (recall=87.6%); it is 6x more efficient than manual review and retains 91.2% performance with 35% parameter compression.

Table 3: Model evaluation comparison

Method Name	Core Mode	Key Tasks	Core Evaluation Metrics	Common Datasets	Performance Highlights
BERT-based Review Model	Single-modal Text Encoding	Text Compliance Check, Keyword Extraction	F1-score, Precision	Sci-Tech Project Application Text Dataset	Text semantic matching F1 = 0.89
GPT-4 Single-document Analysis Model	Generative Text Understanding	Document Logical Coherence Analysis, Opinion Extraction	BLEU, ROUGE-L	General Academic Document Dataset	Logical analysis accuracy = 0.85
ViT+BERT Simple Fusion Model	Single-modal Independent Processing	Text Content Review, Image Format Check	Accuracy, F1-score	Image-Text Mixed Project Document Dataset	Text review Acc = 0.91, only image format check
LangChain Rule-driven Framework	Chained Text Reasoning	Multi-step Text Review, Simple Semantic Alignment	Review Efficiency, Recall Rate	Enterprise Internal Project Document Dataset	Review efficiency increased by 30%
TE-Trans Intelligent Review Framework	LLM + Knowledge Graph + Hybrid Reasoning	Tech Parameter Conflict Detection, Logical Consistency Evaluation, Cross-document Semantic Consistency Analysis	Accuracy, F1-score, Recall Rate	800 Tech Reports/Implementation Plans from National Research Institution	Conflict detection Acc = 92.7% (+18.3% vs. rule engine), logical evaluation F1 = 89.1%, cross-document recall = 87.6%, efficiency 6x higher than manual review

The model performance is compared on the test data set through five automatic evaluation indicators, and the detailed results are shown in Table 4. The proposed TE-Trans model outperformed the nCG-ESM model on the three indices of BLEU with improvements of 0.0005, 0.0329, and 0.0107, respectively. The Dinstinct-1 index was 0.0068 higher than the TA-Seq2Seq model, and the Dinstinct-2 index was 0.1891 higher than the nCG-ESM model, showing the best overall effect of the TE-Trans model.

Table 4: Results of automatic evaluation

Baseline	BLEU-1	BLEU-2	BLEU-3	Distinct-1	Distinct-2
Seq2Seq	0.7722	0.1485	0.0560	0.0241	0.1711
Transformer	0.7971	0.1215	0.0578	0.0344	0.3767
TA-Seq2Seq	0.8325	0.1374	0.0471	0.1624	0.4094
nCG-ESM	0.9678	0.2522	0.0843	0.1531	0.4919
TE-Trans	0.9683	0.2858	0.0952	0.1693	0.6848

Figure 5 shows a comparison of the two methods of combining word representations. The difference between the two is small, but the splicing form is slightly superior in some dimensions. This may be because splicing retains the original information, avoids information mixing, reduces loss and improves information utilization efficiency.

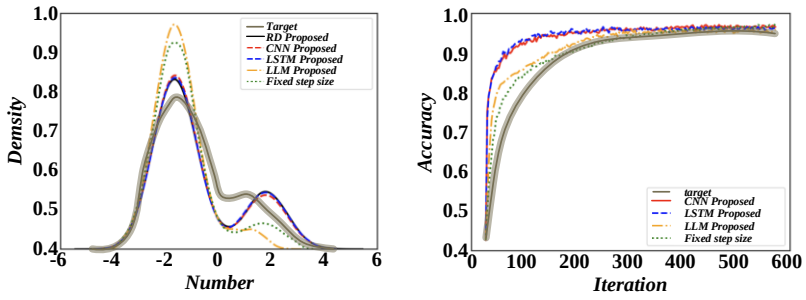

Figure 5: Word embedding method test

Figure 6 shows the execution time of different algorithms for entity expansion under the same conditions. Through effective data preprocessing, the running time of these algorithms is lower than DMA algorithm. However, the running time of SCEE algorithm is slightly higher than that of TSPMA algorithm because it needs to calculate pattern matching and semantic similarity.

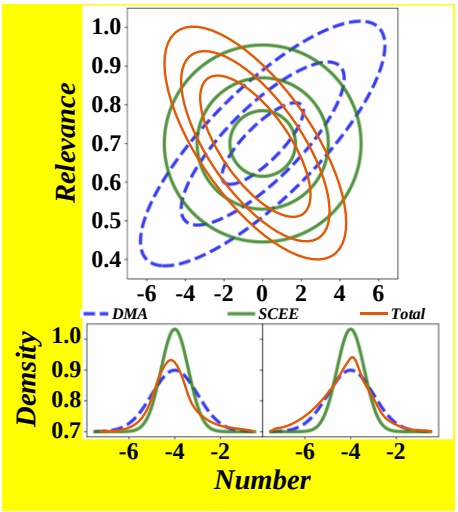

Figure 6: Comparison of running time under different coverage thresholds

According to the analysis in Table 5, the first two pre-training techniques were used to train the word vector, and the effect was comparable to that of BERT. But the third and fourth techniques have improved the effect and showed improvement. Therefore, this paper selects the third technique as the word vector training method and applies it to the model calculation.

Table 5: Evaluation of generation results of different pre-training methods

Baseline	Distinct-1	Distinct-2
TE-TransB	0.0260	0.3468
TE-Transl	0.0344	0.3766
TE-Trans2	0.0297	0.3760
TE-Trans3 (selected)	0.1693	0.6848
TE-Trans4	0.1652	0.6869

Figure 7 shows that the algorithm takes less time than the other three stand-alone algorithms when executing entity expansion tasks, and the running time decreases as the number of parallel tasks increases. This shows that the parallel processing entity expansion algorithm can significantly reduce the running time and improve the computational efficiency.

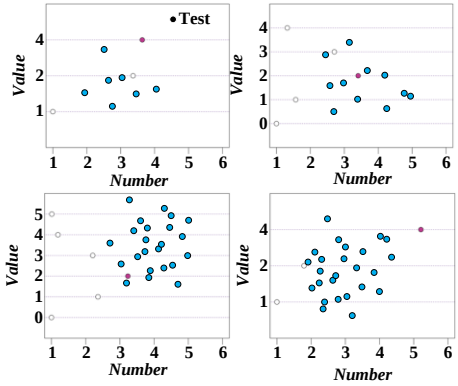

Figure 7: Comparison of running time of different algorithms

Record the embedding loss versus adversarial loss for the first 500 training sessions, as shown in Figure 8. In training, the embedding loss decreases monotonically and converges steadily; Adversarial loss fluctuates greatly at the initial stage and is stable later, which is in line with the expectation of adversarial learning theory.

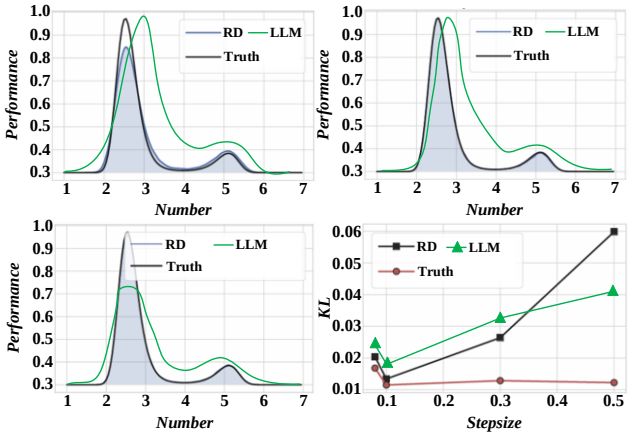

Figure 8: Variation curves of embedding loss and adversarial loss of training process on dataset

The influence of the number of anchors on the retrieval accuracy was evaluated experimentally, and the results of both tasks were analyzed. As shown in Figure 9, the retrieval accuracy improves as the number of anchor points increases from 100 to 500; When the number of anchor points is increased from 500 to 1000, the performance is stable, and both tasks achieve the best results at 1000 anchor points.

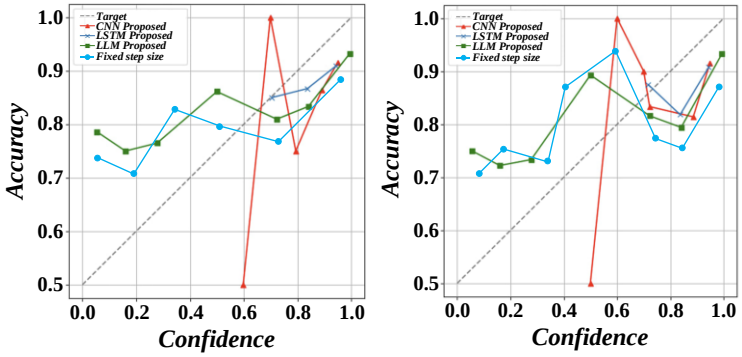

Figure 9: Effect of the number of anchor points on the dataset

Figure 10 shows that α is stable in the [1, 1000] interval, and the fluctuation increases slightly when μ values exceed 1, but remains stable in the [10, 1000] interval. Different values have an impact on the search effect, but have little impact on the overall effect, which shows that this method is robust to parameter changes and can maintain a wide range of stability performance.

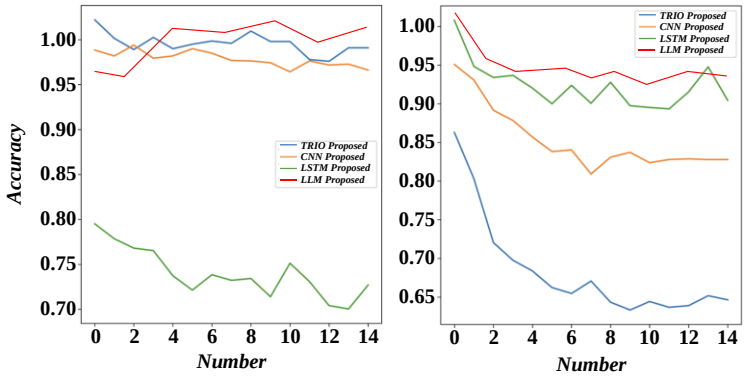

Figure 10: Parameter sensitivity curve on data set

Figure 11 shows that after normalization, the objective function shows a monotonic decreasing trend with the increase of iteration times. Usually, the target value is stable after 20 iterations, which verifies the effectiveness of the optimization algorithm.

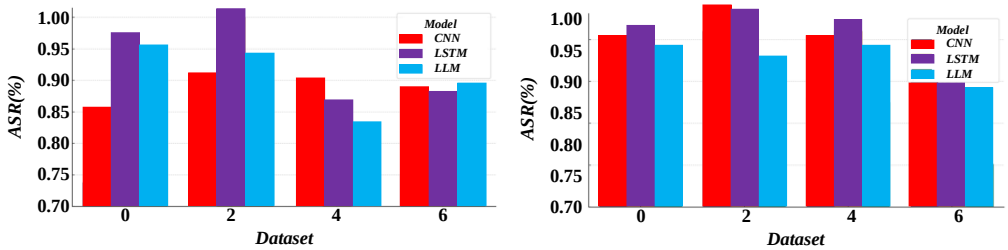

Figure 11: Experimental results of objective function on different datasets

5 Discussion

The TE Trans intelligent evaluation framework proposed in this study demonstrates advantages over baseline methods in multiple dimensions: in the core evaluation task, technical parameter conflict detection Acc=92.7% (compared to rule engine+18.3%), logical consistency evaluation F1=89.1% (higher than GPT-4's 85%), cross document semantic analysis recall=87.6%, and parameter compression of 35% still maintains 91.2% text evaluation accuracy; In terms of generation quality, BLEU-2 is higher than nCG-ESM by 0.0329 and Distinctit-2 is higher than nCG-ESM by 0.1891. The probability of obtaining a score of 2 through manual evaluation is 48.0% (higher than TA-Seq2Seq by 7% and nCG-ESM by 5.5%); In terms of efficiency, it is 6 times higher than manual labor, and the performance is stable when $\alpha \in [11000]$ and $\mu \in [101000]$ are used. This advantage stems from the "LLM+Knowledge Graph+Hybrid Reasoning" architecture: integrating LLM context understanding and knowledge graph structured reasoning capabilities, constructing an "entity relationship" network to strengthen cross document multimodal knowledge association, and adapting to multiple review tasks through modular design. However, the framework has limitations: in terms of interpretability, the reasoning process lacks textual explanations, review opinions have no traceability

identification, and the adversarial training mechanism is opaque; In terms of scalability, large-scale document processing faces the bottleneck of knowledge graph maintenance, and cross domain adaptation requires the reconstruction of entities and rules. Multimodal fusion only stays at simple collaboration, and in the future, it needs to be optimized through inference chain visualization, domain adaptive knowledge graphs, and multimodal unified embedding models to promote technology implementation.

There are three main types of core errors in the model: firstly, inconsistent classification of errors, manifested in repeated classification and judgment of evaluation dimensions such as compliance of technical indicators and rationality of R&D plans in documents, such as technical defects sometimes classified as "key issues" and sometimes classified as "secondary issues". The root cause is that the model does not fully learn the semantic boundaries of evaluation criteria, and the fuzzy labeling of evaluation classification labels in the training data will further exacerbate this problem; The second is the omission of contradictions, that is, the failure to identify semantic conflicts within documents or information contradictions between documents. This is not only related to the model's insufficient ability to understand the context of long texts, but also due to the dense use of professional terminology and contradictions hidden in the detailed logic of technical documents, making it difficult for the model to capture deep correlations; The third is semantic understanding bias, manifested as misreading the meaning of technical parameters and misinterpreting the intent of evaluation requirements. This is mainly due to insufficient coverage of professional language materials in the field of science and technology in the training data, and the model's weak ability to disambiguate ambiguous expressions in the field, making it susceptible to interference from general semantic knowledge.

6 Conclusion

With the rapid development of artificial intelligence technology, large language models have shown significant advantages in complex text processing tasks, especially in the fields of intelligent review and semantic consistency analysis of science and technology project structured reports. Multi-dimensional experiments have verified its application value. The multi-task framework constructed in this study integrates domain knowledge graphs and large language models, and optimises the model's perception of the dynamic evolution of technical terms through pre-training and fine-tuning strategies.

(1) Breakthroughs have been made in core tasks such as technical parameter conflict detection, logical coherence evaluation and cross-web semantic consistency analysis. In the experiment, 800 technical structured reports accumulated by a national scientific research institution in the past three years were selected as the test set. The accuracy rate of the model in the technical parameter conflict detection task reached 92.7%, which was 18.3 percentage points higher than that of the traditional rule engine. It effectively identified the hidden contradictions caused by parameter unit confusion or version iteration in the document. In the logical coherence evaluation task, the F1 value generated by the model is 89.1%, which is significantly better than the traditional method based on keyword matching (75.4%). Especially in the matching degree analysis of experimental design, rationality verification, and conclusion derivation, the model can accurately locate logical faults;

(2) The average recall rate of the cross-web semantic consistency analysis module reaches 87.6%, and it only takes 2.5 hours to process a 100,000-word-long text. The efficiency is 6 times higher than manual review, and the model parameter scale is compressed to the original scale through dynamic knowledge distillation technology. After 35%, it still maintains a core performance of 91.2%.

(3) The research further verifies the optimisation effect of the multi-agent collaboration mechanism on review efficiency. Experimental data show that after integrating a domain knowledge base and a reinforcement learning strategy, the system's recognition coverage rate of implicit associations between web page chapters increases from 68.4% to 89.7%, and the false alarm rate decreases to 4.3%.

Fundings

Supported by Inner Mongolia Power (Group) Co., Ltd. Science and Technology Project ——Research and Application of Intelligent Review Technology Assistant Based on "AI+RPA"（Project Number: 2025-3-3）

References

[1] J. M. Carrillo de Gea, C. Ebert, M. Hosni, A. Vizcaino, J. Nicolas, and J. L Fernandez-Aleman, "Requirements Engineering Tools:An Evaluation, " leee Software, vol. 38, no. 3, pp. 17-24, 2021. doi:10. 1109/ms. 2021. 3058394.

[2] W. De Coster, and R. Rademakers, "NanoPack2: population-scale evaluation of long-read sequencing data, " Bioinformatics, vol. 39, no. 5, pp. btad311, 2023. doi:10. 1093/bioinformatics/btad311.

[3] N. K. Ghalenoei, M. B. Jelodar, D. Paes, and M. Sutrisna, "Challenges of offsite construction and BIM implementation: providing a framework for integration in New Zealand, " Smart and Sustainable Built Environment, vol. 13, no. 4, pp. 780-808, 2024. doi:10. 1108/sasbe-07-2022-0139.

[4] L. Assis, A. C. Rodrigues, A. Vivas, C. G. Pitangui, C. M. Silva, and F. A. Dorca, "Relationship Between Learning Styles and Learning Objects: A Systematic Literature Review, " International Journal of Distance Education Technologies, vol. 20, no. 1, pp., 2022. doi:10. 4018/ijdet. 296698.

[5] G. Bathla, P. Singh, R. K. Singh, E. Cambria, and R. Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning, " Neural Computing& Applications, vol. 34, no. 22, pp. 20213-20229, 2022. doi:10. 1007/s00521-022-07531-8.

[6] D. Castellanos-Cardenas, N. L. Posada, A. Orozco-Duque, L. M. Sepulveda-Cano, F. Castrillon, O. E. Camacho, and R. E. Vasquez, "A Review on Data-Driven Model-Free Sliding Mode Control, "Algorithms, vol. 17, no. 12, pp., 2024. doi:10. 3390/a17120543.

[7] M. Nassif, and M. P. Robillard, "Identifying Concepts in Software Projects, " leee Transactions on Software Engineering, vol. 49, no. 7, pp. 3660-3674, 2023. doi:10. 1109/tse. 2023. 3265855.

[8] J. Oh, S. Hong, B. Choi, Y. Ham, and H. Kim, "Integrating text parsing and object detection for automated monitoring of finishing works in construction projects, " Automation in Construction, vol. 174, no., pp., 2025. doi:10. 1016/j. autcon. 2025. 106139.

[9] K. Pavelka Jr, K. Pavelka, and L Beloch, "A Reconstruction of the Shrine of the Prophet Nahum: An Analysis of 3D Documentation Methods and Data Transfer Technology for Virtual and Augmented Realities, " Applied Sciences-Basel, vol. 15, no. 2, pp., 2025. doi:10. 3390/app15021000.

[10] W. -L Tsai, "A cooperative mechanism for managing multimedia project documentation, "Multimedia Tools and Applications, vol. 81, no. 24, pp. 35069-35082, 2022. doi:10. 1007/s11042-021-10521-y.

[11] M. Shanahan,"Talking about Large Language Models," Communications of the Acm, vol.67,no.2,pp.68-79,2024.doi:10.1145/3624724.

[12] L. Zhou, "Trustworthy digital twinning data platform for power infrastructure construction projects using blockchain and semantic web, " Frontiers in Built Environment, vol. 10, no., pp., 2024. doi: 10. 3389/fbuil. 2024. 1440513.

[13] M. -H. Chao, A. J. C. Trappey, and C. -T. Wu, "Emerging Technologies of Natural Language-Enabled Chatbots: A Review and Trend Forecast Using Intelligent Ontology Extraction and Patent Analytics, "Complexity, vol. 2021, no., pp., 2021. doi:10. 1155/2021/5511866.

[14] L. A. M. Hernandez, A. L. S. Orozco, and L. J. G. Villalba, "Analysis of Digital Information in Storage Devices Using Supervised and Unsupervised Natural Language Processing Techniques, "Future Internet, vol. 15, no. 5, pp., 2023. doi:10. 3390/fi15050155.

[15] Boulkroune, A., S. Hamel, F. Zouari, A. Boukabou & A. Ibeas, "Output - Feedback Controller Based Projective Lag - Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities," Mathematical Problems in Engineering, vol. 2017, no. 1, pp. 8045803, 2017

[16] Boulkroune, A., F. Zouari & A. Boubellouta, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Journal of Vibration and Control, vol., pp. 10775463251320258, 2025

[17] Li, Y., S. Tong & T. Li, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," Hammamet, pp. 483-494, 2013.

[18] Nazir, R. N. Mir, and S. Qureshi, "Idea plagiarism detection with recurrent neural networks and vector space model, " International Journal of Intelligent Computing and Cybernetics, vol. 14, no. 3, pp. 321-332, 2021. doi:10. 1108/ijicc-11-2020-0178.

[19] Y. Wu, F. Liu, L. Zheng, X. Wu, and C. Lai, "CSR-SVM: Compositional semantic representation for intelligent identification of engineering change documents based on SVM, " Advanced Engineering Informatics, vol. 57, no., pp., 2023. doi:10. 1016/j. aei. 2023. 102050.

[20] Y. Xue, "Towards automated writing evaluation: A comprehensive review with bibliometric, scientific, and meta-analytic approaches, " Education and Information Technologies, vol. 29, no. 15, pp. 19553-19594, 2024. doi:10. 1007/s10639-024-12596-0.

[21] X. Yang, Z. Wang, Q. Wang, K. Wei, K. Zhang, and J. Shi, "Large language models for automated Q&A involving legal documents: a survey on algorithms, frameworks and applications, " International Journal of Web Information Systems, vol. 20, no. 4, pp. 413-435, 2024. doi:10. 1108/ijwis-12-2023-0256.

[22] G. Yin, F. Chen, Y. Dong, and G. Li, "Attentive convolutional neural network with the representation of document and sentence for rating prediction, " Applied Intelligence, vol. 52, no. 8, pp. 9556-9573, 2022. doi:10. 1007/s10489-021-03045-3.

[23] Y. Zhang, C. Zhao, W. Liao, W. Zhou, and M. Yuan, "Asymmetrical Attention Networks Fused Autoencoder for Debiased Recommendation, " Acm Transactions on Intelligent Systems and Technology, vol. 14, no. 6, pp., 2023. doi:10. 1145/3596498.

[24] H. Alostad, "Large Language Models as Kuwaiti Annotators, " Big Data and Cognitive Computing, vol. 9, no. 2, pp., 2025. doi:10. 3390/bdcc9020033.

[25] D. Banks, C. Bosone, B. Carpenter, T. Shah, and C. Shi, "Large Language Models: Trust andRegulation, " Harvard Data Science Review, vol. 6, no. 3, pp., 2024. doi:10. 1162/99608f92. 1be2ab6e.

[26] A. Bates, R. Vavricka, S. Carleton, R. Shao, and C. Pan, "Unified modeling language code generation from diagram images using multimodal large language models, " Machine Learning with Applications, vol. 20, no., pp., 2025. doi:10. 1016/j. mlwa. 2025. 100660.

[27] S. Batsakis, I. Tachmazidis, M. Mantle, N. Papadakis, and G. Antoniou, "Model Checking Using Large Language Models-Evaluation and Future Directions, " Electronics, vol. 14, no. 2, pp., 2025. doi: 10. 3390/electronics14020401.

[28] S. Gao, W. Liu, J. Zhu, X. Dong, and J. Dong, "Business Process Modeling Notation-Large Language Model: Transforming Business Processing Modeling Notation Models Into Smart Contracts Using Large Language Models, " leee Software, vol. 42, no. 4, pp. 50-57, 2025. doi:10. 1109/ms. 2025. 3553293.

[29] Y. Guo, W. Qiu, G. Leroy, S. Wang, and T. Cohen, "Retrieval augmentation of large language models for lay language generation, " Journal of Biomedical Informatics, vol. 149, no., pp., 2024. doi:10. 1016/j. jbi. 2023. 104580.

[30] Rigatos, G., M. Abbaszadeh, B. Sari, P. Siano, G. Cuccurullo & F. Zouari, "Nonlinear optimal control

for a gas compressor driven by an induction motor," Results in Control and Optimization, vol. 11, pp. 100226, 2023

[31] Zouari, F., K. B. Saad & M. Benrejeb, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012

[32] Zouari, F., K. B. Saad & M. Benrejeb, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," In: 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13). IEEE, pp. 1-6, 2013.