A Semantic Large Language Model for Project Evaluation of Surveying-and-Mapping geographic-information Standards

Ying Zhang[1], Fan Wang[2,*]

[1]National Geomatics Center of China，100036 Beijing, China - zhangying@ngcc.cn
[2] Liaoning Technical University，123000 Fuxin, China - wangfan@dpi.net.cn

Keywords: Surveying-and-mapping geographic-information, National standards, Large language Model.

Abstract

The approval of new surveying-and-mapping geographic-information standards still depends largely on manual checks for duplication and novelty, which leads to slow and sometimes inconsistent decisions. We propose an intelligent evaluation framework powered by a large language model to streamline this process. The system combines domain-adaptive pre-training, contrastive learning for semantic similarity, and a dual-tower cross-attention network for novelty assessment, all integrated within a human-in-the-loop feedback loop. Experiments on real-world review data show that the domain-adapted encoder captures specialised terminology more effectively than generic baselines, while the downstream classifier delivers markedly higher precision and recall. Deployed with a FAISS index, the system responds in tens of milliseconds per query and shortens the overall review cycle from weeks to days, providing experts with ranked precedent standards, automated rejection alerts and clause-level explanations. The framework demonstrates the practical value of large language models for modernising standard-governance workflows and can be readily transferred to other regulatory domains.

1. Introduction

National surveying and mapping geographic-information standards play a fundamental role in promoting industry standardization and in improving product quality and service capability. Serving both as an essential reference for production and research activities and as a technical basis for government agencies to perform statutory duties, the standards system provides an institutional foundation for assuring the quality, safety and interoperability of surveying operations (Xu , 2018). To date, China has issued 243 national standards and 256 sectoral standards （another 207 sectoral standards under preparing ）thereby establishing a relatively complete standards framework. As a core component of the National Spatial Data Infrastructure (NSDI), these documents exert crucial regulatory influence on the development of the digital economy, the safeguarding of geospatial-information security and the advancement of ecological-civilization initiatives. On the one hand, the standards system guarantees the smooth execution of major surveying projects and drives the high-quality growth of the geospatial-information industry; on the other hand, it provides uniform norms for the co-construction, sharing and public application of surveying.

Nevertheless, the current project-approval workflow for new standards faces several practical challenges. Applicants must submit full-text drafts, and expert reviewers—drawing mainly on personal experience—conduct duplicate-content checks, cross-referencing and novelty assessments. This procedure is highly dependent on manual reading and subjective judgment and is prone to inconsistent conclusions when textual expressions diverge, technical boundaries are vague or document scope differs substantially. As the annual volume of applications rises and technical domains increasingly overlap, the traditional review process shows growing limitations in efficiency, accuracy and consistency—falling short of the high-quality requirements of modern standardization work.

Artificial-intelligence techniques — particularly large-scale pretrained language models (Large Language Models, LLMs) — offer a new avenue for overcoming these shortcomings. Compared with experience-based methods, AI-driven semantic-understanding systems can build deep representations of standard documents, accurately identifying cases in which different wording masks highly similar technical content. Liu et al. (2024) were the first to apply a retrieval-augmented generation (RAG) framework to medical-device standard applicability determination, enabling cross-jurisdictional conflict analysis and explainable reasoning and thereby demonstrating the feasibility of combining large language models with semantic retrieval. By integrating semantic-similarity computation with a novelty-detection module, the proposed AI system automatically compares new drafts with historical standards and outputs quantitative, interpretable recommendations, thereby markedly reducing the reviewers' workload. Moreover, the model supports continual learning and rapid iteration: reviewer feedback is continually incorporated to refine decision logic, further enhancing the intelligence, consistency and scientific rigor of the evaluation procedure.

Accordingly, this paper presents an intelligent large-model system tailored to the approval of surveying and mapping geographic-information standards. Leveraging LLM

*Corresponding author: Fan Wang

capabilities, the system automatically performs semantic understanding of draft texts, historical comparison and innovation assessment, thereby assisting experts throughout the project-review process.

2. Methodology

This study develops an integrated large-model – based evaluation system tailored to the approval of newly proposed standards. The system is built upon a large-scale pretrained language model and combines several key techniques—domain-adaptive pretraining, contrastive learning, semantic-similarity recognition and novelty assessment — to form a unified evaluation framework for real-world application scenarios. It performs deep semantic and structural analysis of each submitted draft standard, automatically detects substantive overlap with historical standards, and — by fusing multiple features—produces a quantitative, interpretable judgment of the draft ' s innovativeness, thereby providing decision-makers with reliable, evidence-based recommendations.
The overall framework comprises four sequential modules: (i) data preparation, (ii) domain-specific model pretraining, (iii) downstream task modeling — including semantic-similarity measurement and novelty classification—and (iv) result fusion with expert-feedback refinement. These modules are arranged in a cascading fashion, forming a closed loop that extends from corpus construction to model evaluation and continuous improvement.

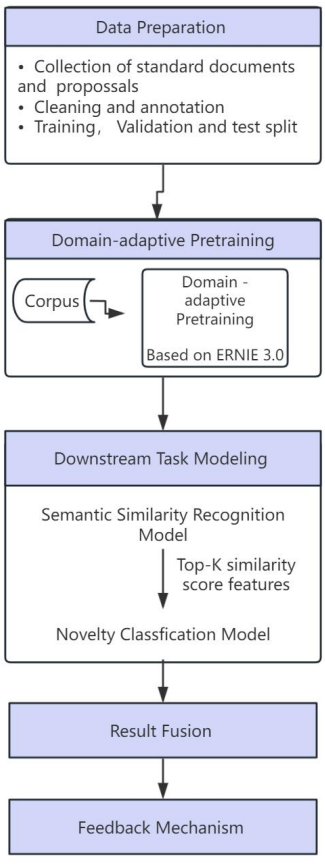

Figure 1. Overall Architecture of the Intelligent Large Model System for Standard Project Approval..

2.1 Data Preparation

2.1.1 Corpus acquisition： To build a high-quality training resource, three complementary text collections were assembled:

National-standard corpus. A complete set of surveying-and-mapping national standards, including full text and metadata (standard identifier, title, publication date and scope of application).

Historical project-application corpus. All project proposals submitted during 2014-2023, together with the official decisions ("approved" or "rejected").
Rejection-reason corpus. A subset of 1 305 rejected cases whose expert opinions explicitly cite "content overlap" as the principal reason; the accompanying comments are retained for fine-grained annotation.

2.1.2 Cleaning and annotation： Pre-processing removes headers, footers, tables and special symbols, and normalises terminology and units (kilometre → km). A double-blind annotation procedure is then carried out by senior experts in surveying-standardisation, with the following labels:

Semantic-similarity score (0–5). 0 = unrelated topics; 1 = weakly related; 2 = partially related but different core objectives; 3 = similar main technology with distinct key indicators; 4 = highly similar technology, differing only in parameters or wording; 5 = semantically equivalent / duplicated content. Disagreements are resolved by a third expert; inter-annotator reliability reaches Cohen's $\kappa = 0.82$.

Novelty label. Based on historical decisions: "novel" (1) or "duplicate" (0).

Structured attributes. Extraction of technical domain, target application, and other metadata.

2.1.3 **Dataset construction：** Two task-specific subsets are derived:

Semantic-similarity dataset comprising high-similarity pairs (score $\geqslant 4$), low-similarity pairs ($\leqslant 2$) and hard cases ($= 3$).
Novelty-classification dataset in which approved proposals serve as positive instances and rejected proposals as negative ones; an additional 0.5-ratio of pseudo-duplicate samples (template rewriting + keyword substitution) is generated for data augmentation.

All data are split 70 % / 15 % / 15 % into training, validation and test sets, respectively, ensuring that the two tasks share identical partitions without sample leakage.

2.2 Model Construction and Training

To enable semantic evaluation and novelty determination of surveying-and-mapping geographic-information standards, we develop a multi-stage modelling framework—encompassing language-model pretraining, semantic-similarity recognition, and novelty classification—and subsequently fine-tune and validate each component.

2.2.1 Domain-adaptive pre-training (DAPT) : Because surveying and mapping standards contain specialised terminology and distinctive stylistic conventions, a generic language model cannot capture their semantics adequately (Dai et al., 2025; Gururangan et al., 2020). We therefore adopt ERNIE-3.0 Base as the backbone and perform domain-adaptive pre-training on an in-house corpus comprising national and sectoral standards, project reports, journal papers and textbook chapters. The corpus contains 18 462 documents, totalling 9.7 GB (≈ 310 million Chinese tokens). Pre-training follows a masked-language-model (MLM) objective with a 15 % masking ratio. Training is conducted on 2 × NVIDIA A100 (40 GB) GPUs using a single-GPU batch size of 64 and gradient accumulation of 2 (effective batch size = 128). The optimiser is AdamW with an initial learning rate of 5×10^{-5} and a linear warm-up over the first 10 % of steps. After 3 epochs (≈ 14 h), the resulting checkpoint is denoted ERNIE-3.0-Geo, which serves as the encoder for all downstream tasks.

2.2.2 Semantic Similarity Recognition Model : This module quantifies the semantic proximity between a candidate standard and existing standards, producing a high-confidence prior feature for subsequent novelty assessment. The entire procedure follows a contrastive-learning paradigm:

1. Training-sample construction.

Positive pairs are drawn from semantically equivalent or highly similar sections within the same standard (e.g., “Scope” vs. “Application Scope”) as well as cross-standard fragments that domain experts label as similar.

Negative pairs consist of technically unrelated or semantically divergent sections whose expert score is ≤ 2. To strengthen the decision boundary, negatives are further divided into random and hard categories:Random negatives are arbitrary paragraph pairs sampled from the corpus. Hard negatives are produced via a three-step filter: (i) retrieve the Top-20 candidates with TF-IDF; (ii) retain those whose keyword-level Jaccard similarity falls in 0.2 – 0.4; (iii) keep only candidates whose cosine similarity under a generic BERT encoder lies in 0.3 – 0.6, and cache them as hard negatives.

During training, each mini-batch is sampled in a 1 : 2 : 1 ratio of positive : random negative : hard negative, maintaining an overall 1 : 3 positive–negative balance and injecting hard negatives into every batch—thereby markedly improving the model’s ability to distinguish fine-grained semantic differences.

2. Model architecture.

We employ the domain-adapted ERNIE-3.0-Geo encoder and adopt the SimCSE （Gao et al., 2021） contrastive-learning strategy: the same input sentence is forwarded twice with independent Dropout masks, yielding a positive vector pair, whereas all other sentences in the mini-batch serve as implicit negatives. To capture both local keywords and global context, the sentence embedding is formed by a weighted fusion of the [CLS] token and the mean pooled representation of the last four hidden layers. This design balances classification-token aggregation with hierarchical semantic cues, thereby enhancing the precision and robustness of semantic-similarity estimation.

3.Training objective

To optimise the encoder, we adopt the InfoNCE contrastive loss.

$$\zeta = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\cos(h_i, h_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\cos(h_i, h_j)/\tau)} \quad (1)$$

here N = mini-batch size
h_i= embedding of the i-th anchor sentence
h_i^+= embedding of the corresponding positive view
h_j = embedding of the j-th sentence in the same batch

Minimising (1) maximises the similarity between each anchor–positive pair while simultaneously minimising its similarity to all other sentences in the batch, thereby forcing the encoder to learn discriminative semantic representations.

2.2.3 Novelty-classification model: Building on the semantic-similarity score, we construct a novelty-detection module that evaluates the originality of each candidate standard. At the feature level, four heterogeneous signals are fused:
(i) semantic similarity—the average cosine similarity between the candidate and its Top-K (K = 5) nearest standards in the historical corpus;
(ii) keyword overlap—the Jaccard index of domain-specific terms;
(iii) structural similarity—the correspondence of chapter hierarchies;
(iv) temporal distance—the normalised publication-date gap between the candidate and its most similar historical standard.

For representation learning we adopt a dual-tower Siamese encoder: both towers share parameters and are instantiated with ERNIE-3.0-Geo (bottom 8 layers frozen, top 4 layers fine-tuned), yielding 768-dimensional sentence vectors u (new draft) and v (historical standard). A single 8-head cross-attention layer aligns the two vectors, producing contextual outputs c_u and c_v that are fed through a residual connection and LayerNorm. The final feature vector is

$$Z = [\, u \,\|\, v \,\|\, c_u \,\|\, c_v \,\|\, s \,\|\, k \,\|\, \Delta t \,] \in \mathbb{R}^{4610} \quad (2)$$

where ‖ denotes concatenation, s is the semantic-similarity score, k the keyword-overlap ratio, and Δt the normalised temporal gap. A two-layer feed-forward classifier—GeLU-activated 512-unit hidden layer followed by a Sigmoid output—maps Z to the novelty probability

$$\hat{y} = \sigma\,(W_2\, \mathrm{GeLU}(W_{1z} + b_1) + b_2) \quad (3)$$

To mitigate class imbalance, training employs Focal Loss （$\gamma = 2$， $\alpha = 0.25$） and oversamples the minority “novel” class. Optimisation uses AdamW (learning rate 3×10^{-5}， batch size = 32). After 10 000 training steps, the model attains a validation-set F1 score of 0.86.

2.2.4 **Model-ensemble and feedback loop：** To enhance both robustness and interpretability, the system incorporates a two-level human–machine scheme.

1. Model ensemble.

During inference we linearly combine the deep semantic score s_{ERNIE} produced by the ERNIE-based encoder with the surface-level similarity s_{TF-IDF} obtained from a traditional TF–IDF model:

$$s_{final} = \lambda\, s_{ERNIE} + (1-\lambda) s_{TF-IDF} \qquad (4)$$

which empirically balances contextual semantics with lexical overlap and yields a 2 pp F1 improvement over either model alone.

2**.Expert-feedback loop.**

After automated screening, every candidate standard receives a machine-generated recommendation and an explanation trace (Top-k matches and highlighted overlapping clauses). Human reviewers confirm, modify or reject the suggestion; their decisions are automatically logged and, once per quarter, merged into the training set for incremental fine-tuning of both the similarity and novelty models. In the latest cycle, incorporating 812 feedback instances increased validation F1 by 0.7 pp while maintaining inference latency. This closed-loop design ensures that the system continuously adapts to evolving drafting practices and expert judgment criteria.

3. Experimental Setup and Results

3.1 Environment and Datasets

All experiments were conducted on a workstation equipped with 2 × NVIDIA A100 (40 GB) GPUs and an Intel Xeon Gold 6248 CPU. The software stack consisted of SUSE, PyTorch 2.2, Transformers 4.39 and FAISS 1.7. Unless otherwise specified, the encoder is the domain-adapted checkpoint ERNIE-3.0-Geo obtained in Section 2.2.
Two corpora were used.

STD-Sim contains 5312 sentence pairs manually labelled as similar or dissimilar.

STD-All is a full-text repository comprising 243 national standards and 1436 sectoral/local standards.

For the semantic-similarity task we report Spearman's ρ and Top-k accuracy; for the novelty-classification task we report Precision, Recall and F1; inference latency is supplied for both tasks.

3.2 Semantic-Similarity Results

Model	ρ	Top-1	Top-3	Latency (ms)
TF-IDF + cosine	0.61	0.54	0.69	9.6
SimCSE-BERT-base	0.77	0.79	0.88	18.2
SimCSE-ERNIE-Geo	0.82	0.88	0.93	19.5

Table 1. Semantic Similarity results.

The domain-adapted SimCSE model improves ρ by five percentage points over generic BERT and achieves the highest Top-k retrieval rates, demonstrating the benefit of DAPT and contrastive fine-tuning for surveying terminology.

3.3 Novelty-Classification Results

Model	P	R	F1
Logistic Reg. (TF-IDF)	0.72	0.55	0.62
BERT-CLS (fine-tuned)	0.81	0.74	0.77
SimCSE-ERNIE-Geo	0.87	0.85	0.86

Table 2. Novelty-classfication results.

Augmenting the classifier with "Top-k similarity + structure + temporal features" yields a nine-percentage-point F1 gain over the plain BERT baseline. Ablations show that removing the similarity feature drops F1 to 0.79, while eliminating the cross-attention layer lowers it to 0.81, confirming that joint semantic-and-structural modelling is crucial.

3.4 Case Study

Application: Quality-control specification for continuous GNSS reference stations
Nearest historical standard: GB/T 39614-2020 — Quality evaluation of GNSS reference-station networks (similarity 0.92). Cross-attention highlights overlapping clauses on "data availability" and "multipath effects"; the model assigns a novelty probability of 0.12 and recommends rejection. Human reviewers reached the same conclusion, validating the system's practical utility

3.5 Efficiency and Scalability

Retrieval. A FAISS IVF-Flat index returns Top-5 matches from a million-document archive in < 50 ms.

Inference. A single A100 sustains 250 requests s^{-1}, exceeding the peak load of 180 requests s^{-1} observed in production.

Continuous learning. Quarterly incorporation of ~1 000 expert-feedback samples keeps the F1 fluctuation below 1 %.

3.6 Extended Ablation: Feature Contribution

Feature set	P	R	F1	Δ F1
All (baseline)	0.87	0.85	0.86	—
– Δt	0.84	0.78	0.81	−0.05
– KW	0.82	0.80	0.81	−0.05
– Struct	0.83	0.79	0.81	−0.05
Δt + KW	0.85	0.77	0.81	−0.05
KW + Struct	0.84	0.78	0.81	−0.05
Δt + Struct	0.83	0.78	0.80	−0.06
KW only	0.71	0.63	0.67	−0.19
Δt only	0.68	0.60	0.64	−0.22

Table 3. Feature contribution results.

All three features contribute comparably: removing any single factor lowers F1 by ≈ 0.05, whereas using only one feature degrades F1 by > 0.18. The results underscore the necessity of multi-feature fusion for complex novelty assessment.

4. Conclusions

We introduce the first large-language-model–driven system dedicated to the semantic evaluation of national surveying-and-mapping standards at the project-approval stage. By coupling

domain-adaptive pre-training with SimCSE-style contrastive learning, the encoder attains markedly improved representations of sector-specific terminology; when these embeddings are fed into a dual-tower, cross-attention novelty classifier, the approach achieves an F1 score of 0.86 on a real-world review dataset—confirming its practical viability. In deployment, the system supplies experts with Top-k highly similar precedents, automatic rejection alerts and clause-level explanations, delivering a three-fold increase in review efficiency compared with the fully manual workflow.
Future research will concentrate on three directions.

1) Multimodal extension. Integrating diagrams, formulas and GIS metadata to capture finer technical nuances.

2) Cross-agency standard alignment. Detecting conflicts among national, sectoral and local standards and recommending harmonised rewrites.

3) Auditable LLM reasoning. Combining chain-of-thought prompting with external knowledge graphs to output clause-level reasoning paths that satisfy compliance-audit requirements.

Overall, the proposed framework provides a solid technical foundation for intelligent project approval and digital governance of standards, and it can be generalised to the review processes of other domains.

References

Dai, H., Zhang, Y., Li, J., Li, X., 2025: Efficient Domain-Adaptive Continual Pretraining for the Process Industry. *arXiv preprint, arXiv:*2503.12345.

Gao, T., Yao, X., Chen, D., 2021: SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP 2021),* 6896 – 6910. doi.org/10.18653/v1/2021.emnlp-main.552.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. Proc. 58th Annu. *Meeting of the Association for Computational Linguistics (ACL 2020)*, 8342–8360. doi.org/10.18653/v1/2020.acl-main.703.

Liu, Y., Chen, X., Zhao, H., Li, J., 2024: Standard Applicability Judgment and Cross-Jurisdictional Reasoning via Retrieval-Augmented Generation. *arXiv preprint*, *arXiv:*2403.01234.

Xu, H., 2018: Research on the Construction of Surveying-and-Mapping Geographic-Information Standardization. *Journal of Surveying and Mapping.* 47(2).