Conflict and Overlap Classification in Construction Standards Using a Large Language Model

Seong-Jin Park[1*], **Youn-Gyu Jin**[1*], **Hyun-Young Moon**[1*],
Bong-Hyuck Choi[3], **Seung-Hwan Lee**[3], **Ohjoon Kwon**[4], **Kang-Min Kim**[1,2†]
[1]Department of Artificial Intelligence [2]Department of Data Science
The Catholic University of Korea, Bucheon, Republic of Korea
[3]Korea Institute of Civil Engineering and Building Technology, Goyang, Republic of Korea
[4]Naver, Seongnam, Republic of Korea
`{sjpark,wlsdbsrb,hyunyounge03,kangmin89}@catholic.ac.kr`
`{bhchoi,seunghwanlee}@kict.re.kr, ohjoon1209@gmail.com`

Abstract

Construction standards across different countries provide technical guidelines to ensure the quality and safety of buildings and facilities, with periodic revisions to accommodate advances in construction technology. However, these standards often contain overlapping or conflicting content owing to their broad scope and interdependence, complicating the revision process and creating public inconvenience. Although current expert-driven manual approaches aim to mitigate these issues, they are time-consuming, costly, and error-prone. To address these challenges, we propose conflict and overlap classification in construction standards using a large language model (COSLLM), a framework that leverages a construction domain-adapted large language model for the semantic comparison of sentences in construction standards. COSLLM utilizes a two-step reasoning process that adaptively employs chain-of-thought reasoning for the in-depth analysis of sentences suspected of overlaps or conflicts, ensuring computational and temporal efficiency while maintaining high classification accuracy. The framework achieved an accuracy of 97.9% and a macro F1-score of 0.907 in classifying real-world sentence pairs derived from Korean construction standards as overlapping, conflicting, or neutral. Furthermore, we develop and deploy a real-time, web-based system powered by COSLLM to facilitate the efficient establishment and revision of construction standards.

1 Introduction

National construction standards provide technical guidelines for engineers, contractors, and other construction professionals to ensure the quality and safety of buildings and facilities (Vaughan and Turner, 2013). While the establishment and management of these standards vary by country, they are often grounded in legal frameworks[1] or standard codes[2]. Some countries, such as Iceland, adopt modified versions of international standards, including the Eurocodes[3], to meet local environmental requirements. Continuous advancements in civil engineering and legal systems necessitate continual revisions to construction standards. Many countries have national agencies or committees, such as the American National Standards Institute[4], the Construction Industry Council[5], and the Korea Construction Standards Center (KCSC)[6], to oversee these revisions.

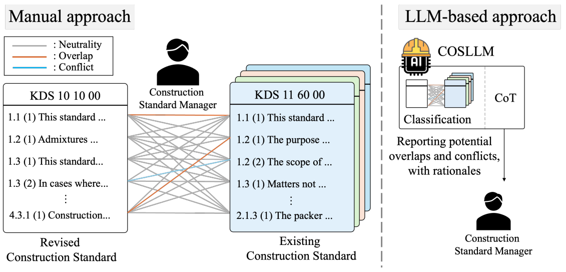

Manual approach
—— : Neutrality
—— : Overlap
—— : Conflict
Construction Standard Manager
KDS 10 10 00
1.1 (1) This standard ...
1.2 (1) Admixtures ...
1.3 (1) This standard...
1.3 (2) In cases where...
⋮
4.3.1 (1) Construction...
Revised Construction Standard
KDS 11 60 00
1.1 (1) This standard ...
1.2 (1) The purpose ...
1.2 (2) The scope of ...
1.3 (1) Matters not ...
⋮
2.1.3 (1) The packer ...
Existing Construction Standard
LLM-based approach
COSLLM
AI
Classification
CoT
Reporting potential overlaps and conflicts, with rationales
Construction Standard Manager

Figure 1: An overview of the manual and LLM-based approaches for analyzing overlapping and conflicting content in construction standards is provided. Using the proposed COSLLM, managers can review potential overlaps and conflicts identified by the LLM, along with detailed rationales, which significantly reduces manual effort.

In the process of establishing or revising standards, members of the construction standards revision committee focus on preventing overlaps and conflicts between new and existing standards (Choi, 2020). When overlapping content exists between construction standards, the revision of one stan-

[*]These authors contributed equally to this work.
[†]Corresponding author.

[1]`https://laws.e-gov.go.jp/`
[2]`https://codes.iccsafe.org/content/IBC2021P2`
[3]`https://eurocodes.jrc.ec.europa.eu`
[4]`https://www.ansi.org/`
[5]`https://www.cic.org.uk/`
[6]`https://www.kcsc.re.kr/`

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 903–917
April 30, 2025 ©2025 Association for Computational Linguistics

dard may lead to conflicts in interpretation, causing confusion among construction professionals. Such conflicts can disrupt the assurance of quality and safety during the construction of buildings and facilities. To address these challenges, some countries have adopted methods such as explicitly referencing existing standards when citing content already covered by regulations, while also conducting routine reviews to resolve overlaps and conflicts (Kim et al., 2016) (Figure 1). However, this expert-driven approach is both time-consuming and costly. Furthermore, excessive reliance on expert interpretations may result in inconsistent judgments among experts (Sun and Zhang, 2014). In South Korea, where revisions occur more frequently than in other countries, effectively resolving issues of overlap and conflict in construction standards is critical (Choi, 2020).

Recently, deep learning-based methods have been employed to classify overlaps and conflicts across various domains (Abeba and Alemneh, 2022; Malik et al., 2022). Previous study (Malik et al., 2024) has defined sentence relationship analysis as being closely aligned with natural language inference (NLI) tasks (Bowman et al., 2015), providing a foundation for analyzing sentences in construction standards. Recent studies (Lee et al., 2023; OpenAI et al., 2024; Street et al., 2024) have demonstrated that large language models (LLMs), equipped with human-level reasoning capabilities, excel at NLI tasks. In addition, the use of chain-of-thought reasoning (CoT) (Wei et al., 2022b) in LLMs enables reliable explanations of the reasoning process (Wei Jie et al., 2024), with the potential to assist construction standard managers in analyzing overlapping or conflicting sentences more effectively. Accordingly, we reframe the classification of overlaps and conflicts in construction standards as a 3-class NLI problem (including neutrality) that can be solved effectively using LLMs.

In this paper, we propose a novel framework for automatically classifying overlaps and conflicts in construction standards, referred to as **C**onflict and **O**verlap classification in construction **S**tandards using a **L**arge **L**anguage **M**odel (COSLLM). COSLLM, built on the latest open-source LLM, is enhanced through two additional training stages. In the first stage, we adapt the LLM to the construction domain using a corpus comprising construction standards, research publications, and news articles. In the second stage, we fine-tune the model to classify sentences into overlap, conflict, or neutral categories using expert-annotated, high-quality sentence pairs from construction standards. We incorporate CoT to handle subtle semantic differences, applying it selectively through task prefixes (Hsieh et al., 2023). This strategy optimizes computational efficiency while maintaining high accuracy. Experiments on real-world construction standard data demonstrated the efficacy of COSLLM. In addition, to support the establishment and revision of construction standards using COSLLM, we develop a real-time construction standards analysis system, which has been deployed. Our main contributions are as follows:

1. We propose COSLLM, an LLM-based framework that automatically classifies overlapping and conflicting sentences, facilitating the establishment and revision of national construction standards.

2. We enhance the effectiveness of an open-source LLM by incorporating domain adaptation and selective CoT, achieving high accuracy in classifying overlaps, conflicts, and neutral relationships.

3. We demonstrate the effectiveness of COSLLM through strong performance in experiments with real-world construction standards data, achieving an accuracy of 97.9% and a macro F1-score of 0.907, highlighting its practical applicability.

4. We develop and deploy a real-time, interactive system powered by COSLLM to significantly improve efficiency and usability in construction standard management.

2 Related Work

Overlap and Conflict Classification Classifying overlaps and conflicts in textual data poses a significant challenge across various domains (Schmolze and Snyder, 1999; Gambo et al., 2024), with deep learning-based technologies are increasingly being explored to address this issue. In the medical research field, algorithms combining string matching, machine learning, and clustering techniques have been developed to automatically detect and remove duplicate data from large-scale bibliographic references across multiple databases, enhancing data quality and reducing manual effort (Hair et al., 2023). In software development, researchers have proposed (Malik et al., 2024) a transfer-learned model built on the SR-BERT architecture (Aum and Choe, 2021), which integrates Sentence-BERT (Reimers, 2019) with a bi-encoder structure. Their proposed model, fine-tuned with domain-specific data, effectively resolves ambiguities and identifies conflicts in development requirements. Building on these advancements, our study employs LLM to address overlaps and conflicts in construction standards, focusing on scalability, domain adaptation, and real-time applicability.

Large Language Model LLMs, built on the transformer (Vaswani et al., 2017) decoder-only architecture and trained with billions of parameters, excel at capturing linguistic patterns and demonstrate advanced reasoning and generation capabilities across diverse tasks (Zhao et al., 2024). Models such as GPT-4 (OpenAI et al., 2024) exhibit capabilities such as long-context understanding (Kuratov et al., 2024), showcasing abilities in in-context reasoning with few-shot (Brown et al., 2020) and zero-shot (Radford et al., 2019; Brown et al., 2020) learning. However, LLMs trained on general-purpose datasets often lack the domain-specific vocabulary and contextual understanding necessary for specialized applications (Ling et al., 2024). Previous studies (Gururangan et al., 2020; Guo and Yu, 2022; Jiang et al., 2024) have demonstrated that achieving high performance with LLMs in specialized domains requires training on tailored corpora. Consequently, fields such as law (Colombo et al., 2024) and medicine (Yang et al., 2024b) have successfully adapted LLMs to fulfill their unique requirements. To further enhance LLM capabilities for complex tasks, techniques such as CoT (Wei et al., 2022b) and plan-and-solve prompting (Wang et al., 2023) have been developed. Building on these findings, our research aims to optimize LLMs for resolving overlaps and conflicts in construction standards.

3 Method

Our framework, COSLLM, leverages LLM to classify semantic relationships between construction standard sentences. Section 3.1, describes how we adapt open-source LLM for the construction domain. Section 3.2 outlines the method for fine-tuning the LLM to classify sentence pairs. Finally, Section 3.3 introduces a real-time web-based system powered by the COSLLM to assist in establishing and revising of construction standards.

3.1 Adapting LLM to Construction Domain

Construction Domain-specific Corpus To address the limitations of general-purpose LLMs in understanding the specialized construction terminology, we curate a construction domain-specific corpus. As no open-source corpus is available, we collect full texts of construction standards, research publications, and news articles. Key sources include the Korea Construction Standards Center[7], the Korea Agency for Infrastructure Technology Advancement[8], the Korean Society of Civil Engineers[9], and construction-related news outlets such as the Civil Engineering Newspaper[10] and Construction Love[11]. Our curated corpus comprises approximately 7.42 million tokens, as measured using the Qwen2 tokenizer (Yang et al., 2024a).

Domain Adaptation Process Using the curated corpus, we fine-tune the open-source multilingual LLM Qwen2-7B-Instruct (Yang et al., 2024a) through causal language modeling. We conduct training over 10 epochs using three Nvidia A6000 GPUs, lasting approximately 3.4 days and incurring a total computational cost of 3.378e18 FLOPs. During training, the loss decreases from 2.502 to 0.581, indicating significant performance improvement. Given the improved classification performance after domain adaptation (DA) (see Section 4.4), we demonstrate that DA enhances the model's ability to comprehend the semantic relationships within the construction domain.

3.2 Two-Step Classification of Overlap, Conflict, and Neutrality Using LLMs

Rationale for the Sentence Pair Approach The ideal solution that maximizes efficiency and simplifies the system would involve an LLM trained specifically in the construction domain to fully understand the entire corpus of construction standards. Such a model can directly analyze sentences or paragraphs to identify overlaps or conflicts, eliminating the need for sentence pairs or neutrality classification. However, this approach necessitates retraining the model whenever the standards are updated, which is both resource-intensive and impractical owing to the specialized nature of construction standards and the limited user base; for instance,

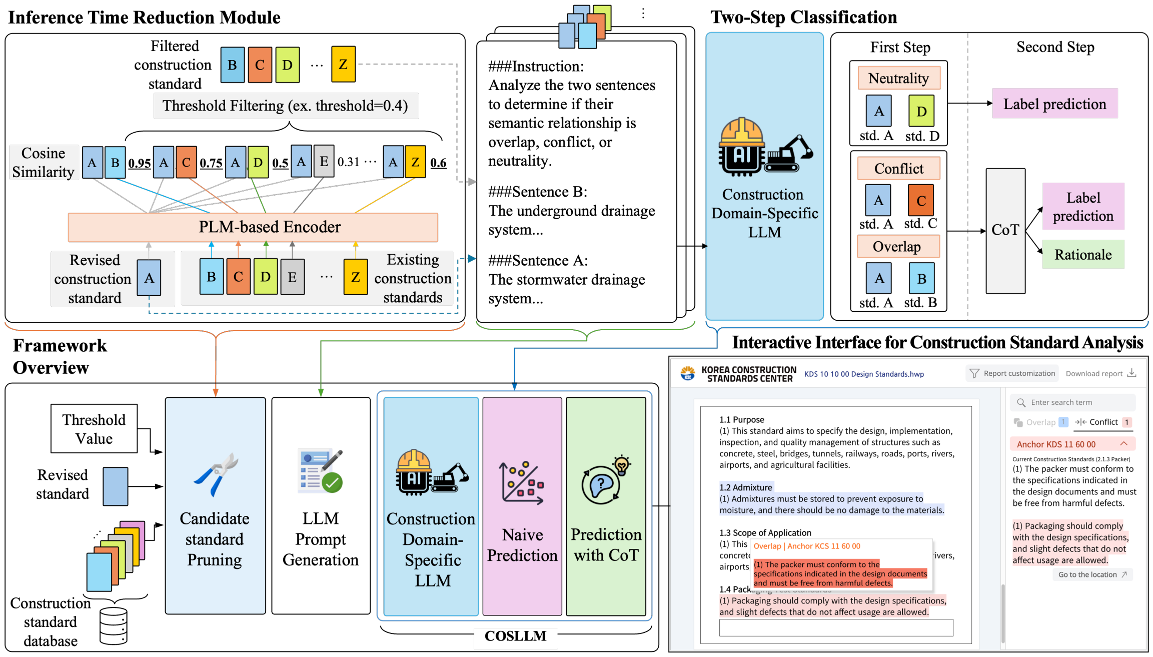
[7] https://www.kcsc.re.kr/

[8] https://www.kaia.re.kr/portal/main.do

[9] https://www.ksce.or.kr/

[10] http://www.cenews.co.kr/

[11] http://www.conslove.co.kr/



###Instruction:
Analyze the two sentences to determine if their semantic relationship is overlap, conflict, or neutrality.

###Sentence B:
The underground drainage system...

###Sentence A:
The stormwater drainage system...

1.1 Purpose
(1) This standard aims to specify the design, implementation, inspection, and quality management of structures such as concrete, steel, bridges, tunnels, railways, roads, ports, rivers, airports, and agricultural facilities.

1.2 Admixture
(1) Admixtures must be stored to prevent exposure to moisture, and there should be no damage to the materials.

1.3 Scope of Application
(1) This concrete rivers, airports

Overlap | Anchor KCS 11 60 00
(1) The packer must conform to the specifications indicated in the design documents and must be free from harmful defects.

1.4 Pack
(1) Packaging should comply with the design specifications, and slight defects that do not affect usage are allowed.

Current Construction Standards (2.1.3 Packer)
(1) The packer must conform to the specifications indicated in the design documents and must be free from harmful defects.

(1) Packaging should comply with the design specifications, and slight defects that do not affect usage are allowed.

Figure 2: Overview of COSLLM and a real-time construction standards analysis framework. Our framework leverages an inference time reduction module to efficiently filter out irrelevant sentence pairs before LLM inference. It then performs effective classification of overlapping and conflicting sentences through a two-step classification process. Finally, the results are delivered to users via an interactive interface, which highlights overlap and conflict sentences, allows result viewing, and supports downloading for optimal usability.

the KCSC currently has only 16 committee members[12]. Although a smaller LLM with fewer than 10 billion parameters is computationally efficient, we empirically observed that its limited size constrains its ability to comprehend the entire corpus, restricting its paragraph-level reasoning capabilities (see Appendix A). To balance effectiveness and efficiency, we adopt a sentence-pair approach. This approach formulates the task as a 3-class NLI problem, where the LLM predicts the semantic relationship between two input sentences.

Inference Time Reduction Module Despite the effectiveness of LLMs, our system faces efficiency challenges owing to the high computational costs of processing numerous sentence pairs, particularly when many are neutral. To mitigate this issue, we leverage the strong semantic similarity of overlapping or conflicting pairs, in contrast to neutral pairs, to pre-filter most of the neutral sentence pairs. Our inference time reduction module (ITRM) utilizes a transformer (Vaswani et al., 2017) encoder-based pre-trained language model (PLM) to compare the semantic similarity of sentence pairs. The PLM pre-embeds existing standard sentences in advance, performs real-time embedding of new sentences and compares them using cosine similarity. Sentence pairs exceeding a predefined cosine similarity threshold are sent to the LLM, significantly reducing computational costs while maintaining accuracy (Dong et al., 2024). In addition, users can adjust the threshold to balance precision and speed, tailoring the analysis to specific requirements. The average cosine similarity of sentence pairs for each class and the implementation details of ITRM are provided in Appendix B.

Leveraging an LLM for 3-Class NLI To classify the semantic relationships in construction standard sentences as a 3-class NLI task, we apply instruction tuning (IT) (Wei et al., 2022a), a technique that fine-tunes LLM by incorporating explicit task instructions, to a construction domain-adapted LLM. For each sentence pair, we create a prompt (provided in Appendix C) containing task descriptions and definitions of overlap, conflict, and neutrality relationships. We curate three example sentence pairs for each relationship to enrich the LLM's understanding of the task, which are reviewed by PhD-level experts. To enhance inference efficiency, we add class-representing tokens (`[overlap]`, `[contradict]`, and `[neutrality]`) to the LLM tokenizer and train the model to gener-

[12] https://www.kcsc.re.kr/Intro/Business

ate the appropriate token. This approach mitigates errors caused by LLM's generation instability and enhances efficiency by minimizing the number of tokens generated during inference.

Selective CoT for Efficient Inference To classify overlapping and conflicting sentences with subtle semantic differences, we employ CoT. Because CoT is time-consuming and resource-intensive (Wei et al., 2022b), we adopt a selective approach during the IT process, inspired by previous work (Hsieh et al., 2023). We add task-specific prefixes to the tokenizer, enabling the model to switch between simple inference and CoT based on the task requirements. The `[predict]` prefix allows for quick single-token prediction, while the `[rationale]` prefix activates CoT for more complex inferences. Because most sentence pairs in construction standards are neutral, COSLLM defaults to simple predictions and uses CoT only for pairs predicted as overlapping or conflicting (illustrated in the top-right section of Figure 2).

3.3 Interactive Interface for Construction Standard Analysis

Overview We develop a real-time web-based interactive system powered by COSLLM to prevent overlaps and conflicts during the establishment or revision of construction standards. This system allows users to compare new construction standards with existing ones and resolve any overlaps and conflicts before release. Users can upload drafts as PDFs or texts, select relevant sections of existing standards, and initiate analysis. The system highlights overlapping or conflicting sentences in the draft, links them to corresponding standard codes, and allows users to download a detailed report (illustrated in the bottom-right section of Figure 2). The CoT results of COSLLM are provided to users, enhancing the convenience of managers during the semantic analysis process. The inference server is implemented using Nvidia Triton (NVIDIA Corporation), with additional modules for real-time construction standard updates. The detailed interfaces of the system are presented in Appendix D.

Real-time Data Collection To ensure accurate comparisons with the latest standards, we develop a real-time data collection system. This system utilizes dynamic crawling techniques to extract the content and structure of current construction standards from the KCSC website, maintaining reliability even with database changes. Built with Selenium[13], the system enables administrators to effortlessly update the standards database.

4 Experiment

4.1 Dataset

We collected 81 overlap instances and 45 conflict instances from Korean construction standards, identified by PhD-level experts. While this dataset provides a solid foundation, its limited size and diversity hinder the model's ability to generalize effectively (Feng et al., 2021). In addition, the vast volume of construction standards makes manual data collection impractical. To address these challenges, we adopted a data augmentation approach proposed in prior research (Yoo et al., 2021), using GPT-4 to generate additional instances for each class. In this process, a real sentence from construction standards was input into GPT-4, accompanied by a carefully crafted prompt and examples, to generate overlapping or conflicting sentences. The augmented data were then reviewed and validated by PhD-level experts, expanding the dataset to 304 instances. Since the majority of sentence relationships in practice are neutral, we included 1,265 neutral sentence pairs derived from actual construction standards. The final dataset comprises 1,569 instances: 144 overlap cases, 160 conflict cases, and 1,265 neutral sentence pairs.

4.2 Evaluation Metrics

We evaluated the classification performance on the overlap, conflict, and neutrality dataset using accuracy and macro-F1 scores. Macro-F1 calculates the F1-score for each class individually and averages them, making it a robust metric for addressing class imbalance (Yang, 1999).

4.3 Baselines

In this study, we evaluate the performance of COSLLM by comparing it with both PLMs and LLMs. PLMs have demonstrated strong performance in classification tasks (Soyalp et al., 2021) and NLI tasks (Liu et al., 2019). We compared COSLLM with PLMs specifically optimized for the Korean language, including BGE-M3-Korean (Chen et al., 2024), KLUE-RoBERTa-large (Park et al., 2021), and KoSimCSE-RoBERTa (Gao et al., 2021). For LLMs, we evaluated Polyglot-Ko-5.8B (Ko et al., 2023), a Korean-trained model, and Qwen2-7B-Instruct. PLM baselines are trained to

[13] `https://selenium-python.readthedocs.io/`

		Incl. Augmented		Excl. Augmented	
	Model	Accuracy	Macro-F1	Accuracy	Macro-F1
PLM	BGE-M3-Korean	0.898	0.736	0.950	0.815
	KLUE-RoBERTa-large	0.936	0.824	0.950	0.739
	KoSimCSE-RoBERTa	0.955	0.874	0.972	0.881
LLM	polyglot-ko-5.8b	0.943	0.853	0.957	0.760
	Qwen2-7B-Instruct	0.955	0.882	0.957	0.714
	COSLLM (Ours)	**0.981**	**0.962**	**0.979**	**0.907**

Table 1: Experimental results on classifying overlap, contradict, and neutrality. Incl. Augmented refers to test sets with augmented instances, while Excl. Augmented includes only real-world data. **Boldfaced** indicates the best results.

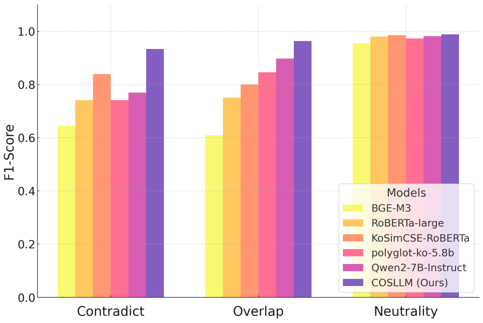


Figure 3: Class-wise F1-scores for classifying overlap, contradiction, and neutrality.

predict the class using the [CLS] token, with two sentences separated by the [SEP] token. LLM baselines are trained using IT with CoT. More implementation details are described in Appendix E.

4.4 Experimental Results

Table 1 presents the experimental results comparing the performance of baseline models and COSLLM. Our proposed method, COSLLM, consistently outperforms the baselines, achieving an accuracy of 98.1% and a macro-F1 score of 0.962. Although the augmented data were reviewed by experts, we also tested a setting where the augmented data were excluded from the test set to better simulate real-world conditions. Even under this condition, COSLLM demonstrated superior performance with an accuracy of 97.9% and a macro-F1 score of 0.907, outperforming all baselines. Figure 3 illustrates the class-wise F1-scores for each baseline model and COSLLM. While baseline models struggle to classify overlap and conflict sentences, COSLLM demonstrates strong performance across all classes (details are provided in Appendix F).

5 Analysis

Effectiveness of DA and IT Table 2 presents the results comparing models with and without DA and IT. The model without DA shows a slight performance decline, achieving an accuracy of 95.5% and a macro-F1 score of 0.882, which highlights the importance of DA. In contrast, the model without both DA and IT, tested using few-shot prompting with one example per class (otherwise same as IT prompt), exhibits a significant performance drop, with an accuracy of 80.9% and a macro-F1 score of 0.298, further emphasizing the critical role of IT.

Method	Accuracy	Macro F1
COSLLM (Ours)	**0.981**	**0.962**
- DA	0.955	0.882
- DA & IT	0.809	0.298

Table 2: Experimental results on ablations of DA and IT. **Boldfaced** indicates the best results.

Method	Accuracy	Macro F1	Inference Time (sec)
COSLLM (CoT)	**0.981**	**0.962**	1,364
COSLLM (Selective CoT)	**0.981**	**0.962**	496
+ ITRM	0.961	0.918	323
- CoT	0.949	0.862	**198**

Table 3: Experimental results on ablations of Selective CoT and ITRM. **Boldfaced** indicates the best results.

Efficacy of Selective CoT and ITRM Table 3 demonstrates the efficacy of CoT. Applying CoT to every sentence pair, including neutral ones, results in the longest inference time. In contrast, Selective CoT matches the performance of full CoT while significantly optimizing inference time, making it the most efficient and effective option for real-time applications. The approach without CoT achieves the fastest inference but delivers the lowest performance. A detailed example of CoT-based sentence analysis is provided in Appendix G.

As shown in Table 3, applying ITRM to pre-filter neutral sentences resulted in a slight performance decrease but reduced inference time by approximately 35% compared to the original time. The performance-time trade-off can be adjusted by modifying the threshold, which we made configurable within the framework. In scenarios requiring both rapid analysis and slower but highly accurate analysis, ITRM effectively balances these demands.

6 Conclusion

In this study, we propose COSLLM, which addresses the challenges of overlapping and conflicting content in construction standards by leveraging a domain-adapted LLM with CoT. The COSLLM achieves high accuracy and efficiency, consistently outperforming baselines. The COSLLM-powered construction standards analysis framework facilitates the effective establishment and revision of construction standards.

Limitations

Our methodology introduces a framework for automatically classifying overlapping and conflicting sections in construction standards, along with a novel system for addressing the challenges during the establishment and revision process. However, there are certain limitations. First, collecting a sufficiently large dataset of genuine overlapping or conflicting sentences proved challenging. As discussed throughout the paper, the vast volume of construction standards and the substantial time required for expert analysis posed significant obstacles. Second, our analysis focused exclusively on Korean construction standards, limiting the generalizability of our findings. Nonetheless, we believe the methodology is broadly applicable to other languages, as it is not heavily language-dependent. With adequate corpora and sentence-pair data from construction standards in other languages, our approach could be adapted for diverse linguistic contexts. Third, there is a potential risk that incorrect analysis by our framework could lead to the establishment or revision of flawed construction standards. However, our framework is not intended to replace human decision-making but to serve as an auxiliary tool that simplifies and supports experts' work. Since the final decisions are made by well-trained and experienced professionals, we believe this risk is unlikely to pose significant practical issues.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was supported by the Basic Research Program through a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1C1C1010317) and the 2025 Korea Construction Standards Center Operation – Digital Construction Standards Development Research.

References

Getasew Abeba and Esubalew Alemneh. 2022. Identification of nonfunctional requirement conflicts: Machine learning approach. In *Advances of Science and Technology: 9th EAI International Conference, ICAST 2021, Hybrid Event, Bahir Dar, Ethiopia, August 27–29, 2021, Proceedings, Part I*, pages 435–445. Springer.

Sungmin Aum and Seon Choe. 2021. srbert: automatic article classification model for systematic review using bert. *Systematic reviews*, 10:1–8.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Bong-Hyuk Choi. 2020. Current status and development directions of national construction standards. Technical report, SSY Engineering. Technical Report.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. Saullm-7b: A pioneering large language model for law. *Preprint*, arXiv:2403.03883.

Jiancheng Dong, Lei Jiang, Wei Jin, and Lu Cheng. 2024. Threshold filtering packing for supervised fine-tuning: Training related samples within packs. *Preprint*, arXiv:2408.09327.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Ishaya Gambo, Rhodes Massenon, Roseline Oluwaseun Ogundokun, Saurabh Agarwal, and Wooguil Pak. 2024. Identifying and resolving conflict in mobile application features through contradictory feedback analysis. *Heliyon*, 10(17):e36729.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Xu Guo and Han Yu. 2022. On the domain adaptation and generalization of pretrained language models: A survey. *Preprint*, arXiv:2211.03154.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kaitlyn Hair, Zsanett Bahor, Malcolm Macleod, Jing Liao, and Emily S Sena. 2023. The automated systematic search deduplicator (asysd): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *BMC biology*, 21(1):189.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. Improving domain adaptation through extended-text reading comprehension. *Preprint*, arXiv:2401.07284.

Seok Kim, Tae-Song Kim, and Hwan-Pyo Park. 2016. Development of korean code system for construction specifications and design standards. *KSCE Journal of Civil Engineering*, 20(5):1605–1612.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *Preprint*, arXiv:2306.02254.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *Preprint*, arXiv:2305.18703.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Garima Malik, Mucahit Cevik, Devang Parikh, and Ayse Basar. 2022. Identifying the requirement conflicts in srs documents using transformer-based sentence embeddings. *arXiv preprint arXiv:2206.13690*.

Garima Malik, Savas Yildirim, Mucahit Cevik, Ayse Bener, and Devang Parikh. 2024. Transfer learning for conflict and duplicate detection in software requirement pairs. *Preprint*, arXiv:2301.03709.

NVIDIA Corporation. Triton Inference Server: An Optimized Cloud and Edge Inferencing Solution.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer

McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. *Preprint*, arXiv:2105.09680.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

J.G. Schmolze and W. Snyder. 1999. Detecting redundancy among production rules using term rewrite semantics. *Knowledge-Based Systems*, 12(1):3–11.

Gokhan Soyalp, Artun Alar, Kaan Ozkanli, and Beytullah Yildiz. 2021. Improving text classification with transformer. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 707–712.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *Preprint*, arXiv:2405.18870.

Zhi Sun and Shoujian Zhang. 2014. Complex system modeling on establishment of construction standard system. *Structural Survey*, 32(1):5–13.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ellen Vaughan and Jim Turner. 2013. The value and impact of building codes. *Environmental and Energy Study Institute White Paper*, 20:501–517.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Guoxing Yang, Xiaohong Liu, Jianyu Shi, Zan Wang, and Guangyu Wang. 2024b. Tcm-gpt: Efficient pretraining of large language models for domain adaptation in traditional chinese medicine. *Computer Methods and Programs in Biomedicine Update*, 6:100158.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1–2):69–90.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Appendix

A Paragraph-level Reasoning with Domain-adapted LLM

This section presents experimental results evaluating whether a domain-adapted LLM with fewer than 10 billion parameters effectively understands current construction standards. We conducted experiments using the domain-adapted Qwen2-7B-Instruct model by inputting prompts querying specific content from current construction standards and comparing the generated responses with the actual standard content.

Table 4 presents the results of querying the content of KDS 27 17 00 from the current construction standards, which the domain-adapted Qwen2-7B-Instruct model encountered during training. The model generated outputs entirely different from the actual content of the construction standards, suggesting that it does not retain the current standards accurately. The original text was in Korean and has been translated into English.

B Implementation Details of ITRM

We implemented ITRM using KLUE-RoBERTa-Large, a PLM specialized in the Korean language. By measuring the cosine similarity of sentence pairs from the collected construction standards dataset, we observed that neutral pairs showed an average cosine similarity of 0.7554, while overlapping pairs averaged 0.9218 and conflicting pairs 0.8852. Based on these findings, we hypothesized that leveraging PLM embeddings could effectively pre-filter neutral sentence pairs.

For our experiments, we set the threshold at 0.797. As a result, 66.6% of neutral pairs were pre-filtered, along with 3.4% of overlapping pairs and 11.8% of conflicting pairs.

C Prompt for COSLLM

Table 5 presents the prompts used for instruction tuning COSLLM. Same prompts used to train baseline LLMs.

D System Interfaces

To facilitate convenient management of construction standards, we developed a web-based interactive system. Figure 4 illustrates the system's main interface (a), the screen for setting analysis parameters (b), the process of directly inserting target content (c), and uploading documents to initiate analysis (d). Figure 5 illustrates the screens for selecting the target construction standards for analysis (a), reporting the analysis progress (b), viewing the analysis results (c), and the analysis results presented in an Excel report format (d).

E Implementation Details of COSLLM and Baselines

We implemented the models and baselines using PyTorch (Paszke et al., 2017) and HuggingFace[14]. We conducted a series of experiments with various hyperparameters to enhance the accuracy and F1 scores of the models. For the PLMs, we conducted experiments with batch sizes [8, 16, 32], learning rates [1e-4 to 1e-6], and epochs [20, 30, 40]. For the LLMs, batch sizes [8, 16, 32], learning rates [5e-4 to 5e-5], and epochs [20, 30] were used. Prediction loss alpha and explanation loss alpha for selctive CoT ranged from 0.1 to 1.0. All of the LLMs are trained to utilize CoT while predicting. All experiments were conducted on three NVIDIA A6000 GPUs.

F Detailed Class-wise Performance

Table 6 presents detailed class-wise performance for classifying overlap, contradict, and neutrality.

G CoT Reasoning Examples

Table 7 presents the CoT reasoning results for one pair each of overlapping and conflicting sentence pairs. The original is in Korean but translated into English.

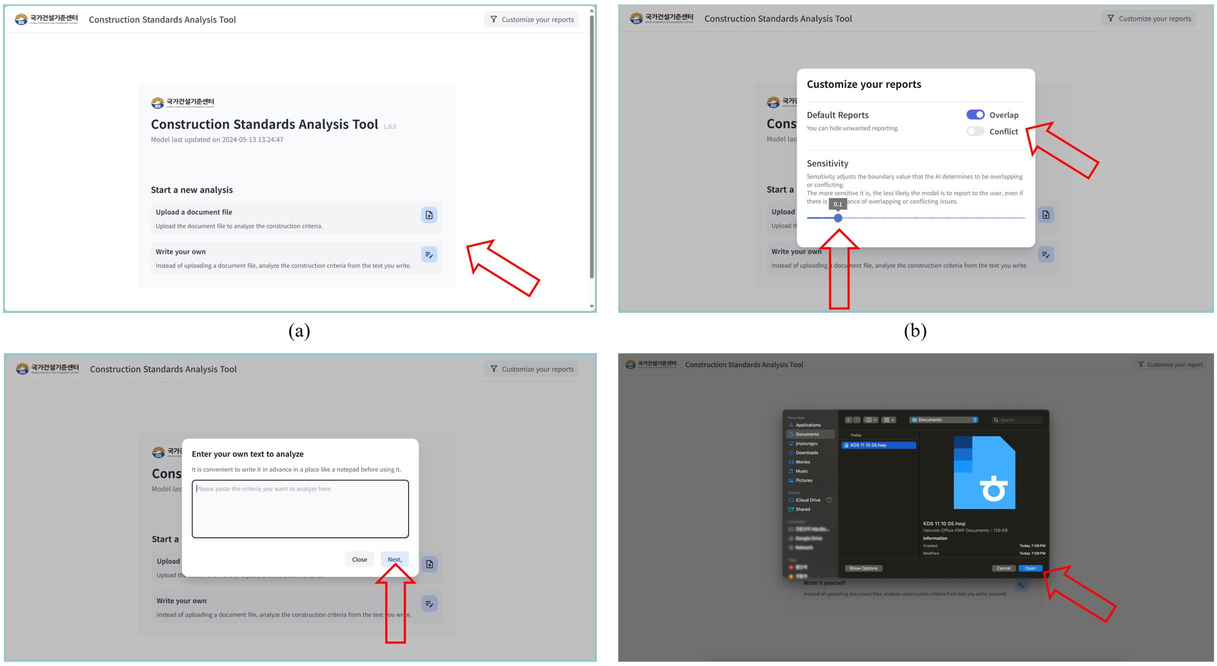
[14] https://huggingface.co/

<table>
<tr><th colspan="2">Prompt</th></tr>
<tr><td colspan="2">### Command: As an expert in the field of civil engineering, you should provide answers to questions related to construction standards, which are documents that present standards for the design, construction, and maintenance of structures.
Question: In Tunnel Seismic Design KDS 27 17 00, Section 4.2 provides seismic analysis methods for tunnel seismic design (1), (2), (3), (4), and (5). Could you tell me each item individually?</td></tr>
<tr><td>Generated Response</td><td>The above sentence asks for information regarding seismic analysis methods for tunnel seismic design. It requires finding the respective methods presented in the construction standard ’Tunnel Seismic Design’ KDS 27 17 00. This standard provides detailed procedures and content, so it is necessary to refer to the relevant parts to explain each method.
For example,
(1) could be the Simple Seismic Analysis,
(2) could be Static Seismic Analysis,
(3) could be Dynamic Seismic Analysis,
(4) could be Seismic Vulnerability Analysis, and
(5) could include Seismic Performance Estimation.
It is important to find and respond with these details to answer the question. If additional information such as exact definitions or application methods for each method is needed, the detailed content of the construction standard should also be provided. Additionally, when answering, consider the specific context in which the question is asked. For instance, by providing information on how each seismic analysis method is applied, depending on tunnel length, ground conditions, and seismic performance goals. This approach will help provide the questioner with more useful information.</td></tr>
<tr><td>Correct Response</td><td>4.2 Seismic Analysis Methods for Tunnel Seismic Design
(1) The seismic analysis methods for tunnels may include the Response Displacement Method, Dynamic Analysis Method, and Pseudo-Static Analysis Method.
(2) For tunnels, when the interaction between the ground and the structure is significant and displacement must be carefully considered, the Response Displacement Method can be applied.
(3) In the dynamic analysis method for Seismic Grade I, the input seismic waves can apply artificially synthesized ground motion time-history curves inferred by considering the seismic damage history of the area where the tunnel is built, ground characteristics, and the natural period of the structure. The generation of these artificially synthesized ground motion time-history curves should follow KDS 17 10 00 (4.2.1.4(10)).
(4) The Pseudo-Static Analysis Method is locally applied to the portal structures of retaining wall-type tunnel entrances, simplifying the dynamic time-history characteristics into static loads to make the design simpler and more conservative.
(5) During seismic design, analysis methods that consider the non-linear behavior characteristics of the ground can be used.</td></tr>
</table>

Table 4: Generated results from the domain-adapted model for queries related to construction standards

Prompt
Instruction: As a civil engineering expert, your task is to analyze sentences extracted from construction standards, which are documents that provide guidelines for the design, construction, and maintenance of structures. Your job is to determine whether the sentences are semantically overlapping, conflicting, or unrelated. Semantic overlaps and conflicts between sentences in construction standards require analysis and judgment based on meaning, rather than just identifying similar words or tones. Overlapping sentences describe the same content under the same section, while conflicting sentences describe different content under the same section. In particular, conflicts may include cases where the same content is described with different values (e.g., numerical discrepancies) or referenced with different construction standard codes. Sentences that are neither overlapping nor conflicting are considered unrelated, meaning they address entirely different topics. The data provided to you are formatted as follows. Sentences from construction standards appear after <|sentence1|> and <|sentence2|>. The label after <|pred|> indicates whether the relationship is semantic overlap, conflict, or none: <|overlap|>, <|contradict|>, or <|none|>. The explanation for the judgment follows <|expl|>. Based on this structure, carefully review the two sentences and provide the correct semantic judgment (overlap, conflict, or none) along with an explanation. An example is as follows: [Overlap Examples] [Conflict Examples] [Neutrality Examples] Now, based on the given construction standard sentences, provide the appropriate semantic classification (overlap, conflict, or none) and explain your reasoning. [Data] ### Response:

Table 5: Prompts used for instruction tuning COSLLM



Model last updated on 2024-05-13 13:24:47

Upload the document file to analyze the construction criteria.

Instead of uploading a document file, analyze the construction criteria from the text you write.

(a)

You can hide unwanted reporting.

Sensitivity adjusts the boundary value that the AI determines to be overlapping or conflicting.
The more sensitive it is, the less likely the model is to report to the user, even if there is ance of overlapping or conflicting issues.

(b)

It is convenient to write it in advance in a place like a notepad before using it.

Please paste the criteria you want to analyze here.

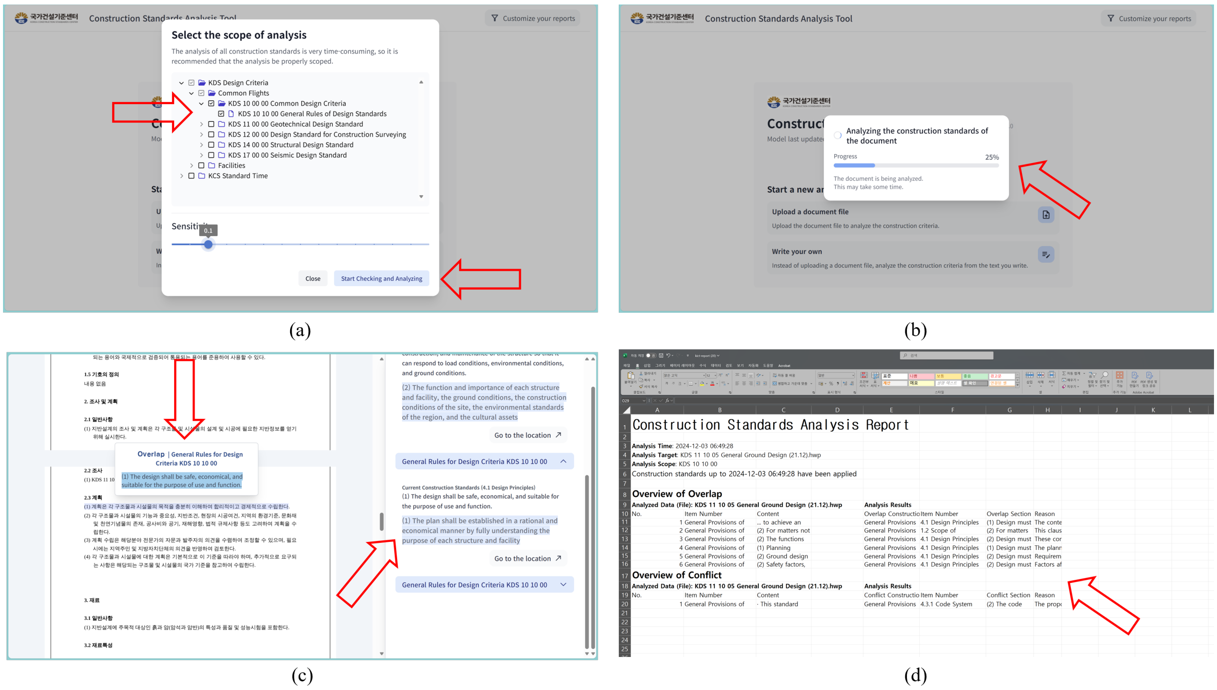
(c)

KDS 11 10 05.hwp
Hancom Office HWP Documents - 129 KB
Created Today 7:08 PM
Modified Today 7:08 PM

Instead of uploading document files, analyze construction criteria from text you write yourself.

(d)

Figure 4: Interfaces of the Interactive Interface for Construction Standard Analysis. (a) Main Interface: Users can select a document or input newly established or revised sentences in text form to initiate analysis. (b) Analysis Parameter Settings: Users can selectively analyze overlaps and conflicts or configure the cosine similarity threshold for ITRM. (c) Text Input Screen: Users can input target sentences for analysis in text form. (d) File Upload Screen: Users can upload .hwp or .pdf files to start the analysis.



국가건설기준센터 Construction Standards Analysis Tool

Customize your reports

Select the scope of analysis

The analysis of all construction standards is very time-consuming, so it is recommended that the analysis be properly scoped.

- ☑ KDS Design Criteria
  - ☑ Common Flights
    - ☑ KDS 10 00 00 Common Design Criteria
      - ☑ KDS 10 10 00 General Rules of Design Standards
    - □ KDS 11 00 00 Geotechnical Design Standard
    - □ KDS 12 00 00 Design Standard for Construction Surveying
    - □ KDS 14 00 00 Structural Design Standard
    - □ KDS 17 00 00 Seismic Design Standard
  - □ Facilities
- □ KCS Standard Time

Sensitivity: 0.1

Close | Start Checking and Analyzing

(a)

국가건설기준센터 Construction Standards Analysis Tool

Customize your reports

Analyzing the construction standards of the document

Progress 25%

The document is being analyzed.
This may take some time.

Start a new an

Upload a document file

Upload the document file to analyze the construction criteria.

Write your own

Instead of uploading a document file, analyze the construction criteria from the text you write.

(b)

되는 용어와 국제적으로 검증되어 통용되는 용어를 준용하여 사용할 수 있다.

1.5 기호의 정의

내용 없음

2. 조사 및 계획

2.1 일반사항

(1) 지반설계의 조사 및 계획은 각 구조물 및 시설물의 설계 및 시공에 필요한 지반정보를 얻기 위해 실시한다.

Overlap | General Rules for Design Criteria KDS 10 10 00

(1) The design shall be safe, economical, and suitable for the purpose of use and function.

2.2 조사

(1) KDS 11 10

2.3 계획

(1) 계획은 각 구조물과 시설물의 목적을 충분히 이해하여 합리적이고 경제적으로 수립한다.

(2) 각 구조물과 시설물의 기능과 중요성, 지반조건, 현장의 시공여건, 지역의 환경기준, 문화재 및 천연기념물의 존재, 공사비와 공기, 재해영향, 법적 규제사항 등도 고려하여 계획을 수립한다.

(3) 계획 수립은 해당분야 전문가의 자문과 발주자의 의견을 수렴하여 조정할 수 있으며, 필요시에는 지역주민 및 지방자치단체의 의견을 반영하여 검토한다.

(4) 각 구조물과 시설물에 대한 계획은 기본적으로 이 기준을 따라야 하며, 추가적으로 요구되는 사항은 해당되는 구조물 및 시설물의 국가 기준을 참고하여 수립한다.

3. 재료

3.1 일반사항

(1) 지반설계에 주목적 대상인 흙과 암(암석과 암반)의 특성과 품질 및 성능시험을 포함한다.

3.2 재료특성

can respond to load conditions, environmental conditions, and ground conditions.

(2) The function and importance of each structure and facility, the ground conditions, the construction conditions of the site, the environmental standards of the region, and the cultural assets

Go to the location ↗

General Rules for Design Criteria KDS 10 10 00

Current Construction Standards (4.1 Design Principles)
(1) The design shall be safe, economical, and suitable for the purpose of use and function.

(1) The plan shall be established in a rational and economical manner by fully understanding the purpose of each structure and facility

Go to the location ↗

General Rules for Design Criteria KDS 10 10 00

(c)

Construction Standards Analysis Report

Analysis Time: 2024-12-03 06:49:28
Analysis Target: KDS 11 10 05 General Ground Design (21.12).hwp
Analysis Scope: KDS 10 10 00
Construction standards up to 2024-12-03 06:49:28 have been applied

Overview of Overlap

Analyzed Data (File): KDS 11 10 05 General Ground Design (21.12).hwp			**Analysis Results**			
No.	Item Number	Content	Overlap Constructio	Item Number	Overlap Section	Reason
1	General Provisions of	... to achieve an	General Provisions	4.1 Design Principles	(1) Design must	The conte
2	General Provisions of	(2) For matters not	General Provisions	1.2 Scope of	(2) For matters	This claus
3	General Provisions of	(2) The functions	General Provisions	4.1 Design Principles	(2) Design must	These cor
4	General Provisions of	(1) Planning	General Provisions	4.1 Design Principles	(1) Design must	The plann
5	General Provisions of	(2) Ground design	General Provisions	4.1 Design Principles	(2) Design must	Requirem
6	General Provisions of	(2) Safety factors,	General Provisions	4.1 Design Principles	(2) Design must	Factors af

Overview of Conflict

Analyzed Data (File): KDS 11 10 05 General Ground Design (21.12).hwp			**Analysis Results**			
No.	Item Number	Content	Conflict Constructio	Item Number	Conflict Section	Reason
1	General Provisions of	· This standard	General Provisions	4.3.1 Code System	(2) The code	The propo

(d)

Figure 5: Interfaces of the Interactive Interface for Construction Standard Analysis. (a) Target Construction Standards Selection Screen: Allows experts to select only the relevant standards from the current construction standards for analysis. (b) Analysis Progress Screen: Displays the progress of LLM inference on the analysis server in real-time. (c) Analysis Results Screen: Highlights results directly on the uploaded document, allowing users to immediately view overlapping or conflicting construction standards. A comprehensive list and summary are available on the right panel for an at-a-glance overview. (d) Analysis Results in Excel Report Format: Provides a downloadable construction standards analysis report in MS Excel format.

Method	Model	Class	Accuracy	F1-Score
Encoder	BGE-M3	contradict	0.625	0.645
		neutrality	0.976	0.954
		overlap	0.500	0.609
	RoBERTa-large	contradict	0.625	0.741
		neutrality	0.984	0.980
		overlap	0.857	0.750
	KoSimCSE-RoBERTa	contradict	0.813	0.839
		neutrality	1.000	0.985
		overlap	0.714	0.800
Decoder	polyglot-ko-5.8b	contradict	0.625	0.741
		neutrality	1.000	0.973
		overlap	0.786	0.846
	Qwen2-7B-Instruct	contradict	0.625	0.769
		neutrality	1.000	0.981
		overlap	0.929	0.897
	COSLLM (Ours)	contradict	0.875	0.933
		neutrality	1.000	0.988
		overlap	0.929	0.963

Table 6: Class-wise performance

Predict	Result	
Overlap	sentence1	When determining the scale of the drainage exclusion facility, the safety of past flood records must also be reviewed.
	sentence2	In studies for flood prevention, the design of drainage exclusion facilities must ensure sufficient safety by referring to past flood data.
	CoT	In sentence 1, it is mentioned that the **safety of past flood records must also be reviewed** during the process of determining the scale of the drainage exclusion facility. In sentence 2, **the importance of ensuring sufficient safety by referring to past flood data** in the design of drainage exclusion facilities for flood prevention is emphasized. Accordingly, both sentences approach the design of drainage exclusion facilities through safety reviews, containing similar key content, and can be considered to have an overlap relationship.
Conflict	sentence1	If a retest is required based on the results of the inspection test, the manufacturer shall bear the costs for the retest, inspection, and any time-related expenses for the retest and inspection.
	sentence2	In the case where defects are found in the inspection test and a retest is necessary, all retest costs shall be borne by the client, and the manufacturer's responsibility is limited to repairing or replacing the product.
	CoT	The statement that **the manufacturer must bear all costs when a retest is required through inspection testing** exceeds the manufacturer's role. While the client can request a retest if defects are found in the inspection test, **the responsibility for covering all costs is limited to the manufacturer.** Therefore, the two sentences provide conflicting instructions regarding the responsibility for retest costs and the authority of the client.

Table 7: CoT reasoning examples