

Developing a tool for resolving standard discrepancies using large language models

Lei Li, Manni Duan, and Yongheng Wang*

Zhejiang Lab, Hangzhou, China

Abstract. This paper presents a useful tool leveraging Large Language Models (LLM) to resolve discrepancies between corporate and national standards. Firstly, we introduce a Food Standard Dataset comprising 1420 Chinese national standards and 100 corporate standards. Secondly, we offer a user-friendly web application to show human-readable modification suggestions. Thirdly, we propose a technique for standard information extraction, which efficiently retrieves relevant information from complete national standards.

1 Introduction

In the context of modern industrial and technological advancement, standardization is a pivotal aspect of corporate governance [1-3]. Corporate and national standards are fundamental elements of the standardization framework, essential for streamlining production processes, enhancing product quality, and ensuring market stability. Given the complexity and volume of standards, corporates frequently need to align their internal standards with national standards to guarantee compliance and consistency.

1.1 Related work

As the scope and content of corporate and national standards continue to expand, traditional manual comparison methods are becoming increasingly inadequate. By integrating advanced machine learning technologies, corporates can significantly enhance the efficiency and precision of these alignment processes. Consequently, scholars have proposed a variety of technical approaches, which primarily encompass methods based on ontologies and knowledge graphs, methods based on Domain-Specific Languages (DSL), and methods based on Natural Language Processing (NLP).

1.1.1 Methods based on ontologies and knowledge graphs

Zhou and El-Gohary [4] propose a method that combines ontology and deep learning to align design information from Building Information Models (BIM) with energy regulations, enabling fully automated energy efficiency compliance checks. McGibney and Kumar [5]

* Corresponding author: wangyh@zhejianglab.com

introduce the WOMBRA project, an ontology-based web retrieval application that retrieves and aligns information from technical standards, facilitating quick access to relevant regulatory information. Zhang et al. [6] present a method to construct knowledge graphs from tables in Chinese power industry PDF documents, allowing the structured representation of unstructured data and the effective resolution of inconsistencies within standards. Melluso et al. [7] combine natural language processing with knowledge graphs to measure the semantic similarity of content across different standard texts, thereby enhancing standard interoperability and resolving semantic conflicts.

1.1.2 Methods based on domain-specific languages (DSL)

Dimyadi et al. [8] explored using LegalDocML and LegalRuleML to standardize normative information sharing in the AEC/FM (Architecture, Engineering, Construction, and Facility Management), offering structured representations. Bareedu et al. [9] proposed extracting modeling constraints from unstructured industrial standards documents and representing them as machine-executable rules for semantic validation. Vincini et al. [10] introduced the MOMIS data integration system, which supports data conversion and query translation through domain-specific schemas, ensuring data consistency across heterogeneous sources.

1.1.3 Methods Based on Natural Language Processing (NLP)

Gatto et al. [11] developed an automated framework that uses NLP to extract and match technical data from bid documents, identifying missing resources and inconsistencies. Wulff et al. [12] designed an openEHR-based process that employs NLP to extract and standardize data from unstructured clinical texts, improving the quality and usability of electronic health records. Bouh et al. [13] proposed a machine learning scheme to digitize paper-based medical history records and store them in a standardized format, demonstrating the potential of machine learning in unstructured data standardization.

1.2 Contributions

Existing studies of standard discrepancy resolution exhibit several limitations, which we aim to address. Our contributions are summarized as follows:

Existing studies rarely resolve the discrepancy between corporate and national standards. Thus, we introduce a **Food Standard Dataset** comprising Chinese national standards and corporate standards. Food standards involve diverse ingredients and extensive tabular data, making it challenging for discrepancy resolution.

Most studies lack human-readable guidances for resolving discrepancy. We design effective instructions for LLM to generate and summarize modification suggestions with a user-friendly web application.

Previous work seldom discusses the effects of large standard documents. To LLM-based systems, excessive data volume presents two significant challenges: increased computational costs and potential exceedance of the LLM's processing capacity. Therefore, this work introduces HTML-based Standard Information Extraction, which accurately retrieves relevant information from complete national standards.

2 Food Standard Dataset

Based on the official list¹ of national food safety standards, we collected 1420 Chinese national standards from the internet. Additionally, to study the discrepancies between corporate standards and national standards, we collected 100 corporate standards from a website² searching *food*. **These raw data and the further processed data in this paper are available for public download³.**

3 Framework

Fig. 1. shows our framework for resolving standard discrepancies. In this paper, the LLM used is DeepSeek-V3⁴. Below, we will delve into key steps of our framework in detail.

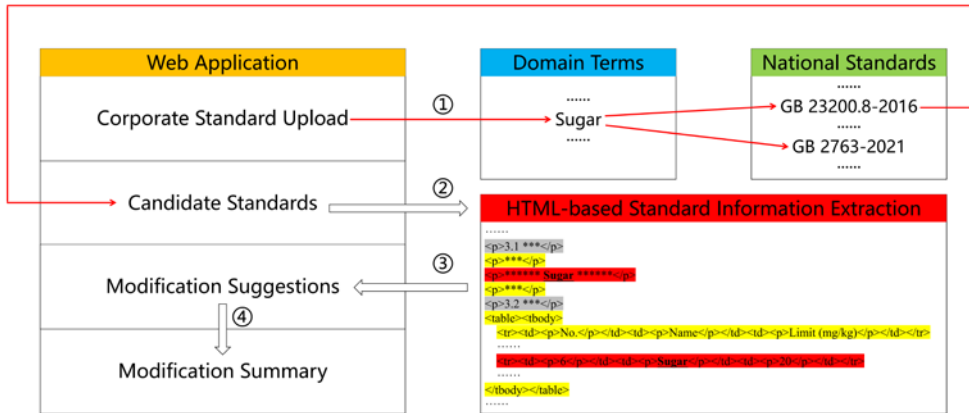


Fig. 1. Framework for resolving standard discrepancies, including four components, e.g., Web Application, Domain Terms, National Standards, and HTML-based Standard Information Extraction.

3.1 Candidate standards retrieval

When a corporate standard is given, the tool proposed in this paper aims to provide modification suggestions based on the relevant national standards. Clearly, the first task at hand is to find the candidate national standards.

The candidate standards retrieval based on sentence embedding models (i.e. Sentence Transformers⁵) and cosine similarity is a classic approach. However, we argue that sentence embedding models do not adequately address the similarity in standard texts for two main reasons: first, there are domain terminologies present in standard texts, such as "bromophos," which should be segmented into smaller tokens by the general-purpose tokenizers, thereby losing some of their semantic meaning; second, to our knowledge, there is currently a lack of paired standard texts to train sentence embedding models.

Thus, developing a sentence embedding model based on large-scale paired standard texts and a tokenizer capable of recognizing domain terminologies is essential, but we will defer this to future research, as the scope of this paper primarily focuses on releasing the

¹ <http://www.nhc.gov.cn/sps/s3594/202403/c54748d1921a4fa196aa2658aa095d37.shtml>

² <https://www.qybz.org.cn/>

³ <https://drive.google.com/file/d/1sDS-CGRf06L6de3iGY0BcJeQ7atiZVv5/>

⁴ <https://www.deepseek.com/>

⁵ <https://github.com/UKPLab/sentence-transformers>

Food Standard Dataset and developing a feasible application for resolving standard discrepancies using Large Language Models.

Therefore, in this paper, we employ a retrieval strategy based on the index of domain terminologies. We prepare domain terminologies through various resources such as textbooks, professional dictionaries, knowledge graphs, and domain experts.

When a user uploads a corporate standard to be analyzed via the Web Application, the system identifies the domain terminologies it contains and links each terminology to each national standard that includes the terminology. All selected national standards are displayed in the "Candidate Standards" area of the Web Application, allowing the user to review and refine the selection. Users can delete irrelevant national standards or manually add standards for further analysis.

3.2 HTML-based standard information extraction

We start by employing OCR services (i.e., WPS Office) to convert the national standards from **.pdf** format into **.docx** format, and then use Mammoth⁶ to transform the **.docx** documents into HTML (HyperText Markup Language) files. Using JavaScript, we manipulate the DOM (Document Object Model) constructed from the HTML to extract the necessary information.

As shown in Fig. 1., firstly, we extract the multi-level headings highlighted in gray; these headings provide an effective skeleton for extracting other information and help clarify the scope of certain ambiguous or unclear local content. Secondly, we extract the content highlighted in red, which consists of fragments containing domain terminologies. Thirdly, we extract the sections highlighted in yellow, which are the contexts surrounding the domain terminologies. If a domain terminology appears within a regular paragraph, we include the content both above and below until we reach a heading element. If a domain terminology appears within a table, we retain table headers and the table's caption.

3.3 Generating modification suggestions

HTML-based Standard Information Extraction provides an extracted subset of the whole DOM constructed from each national standard HTML file. Then, each extracted subset is composed with the uploaded corporate standard and fed into **LLM** for generating modification suggestions. Furthermore, **LLM** summarize all generated modification suggestions into a brief report. Instructions for **LLM** is shown in Table 1., the first row is for generating and the second row is for summarizing.

Table 1. LLM Instructions for generating and summarizing modification suggestions.

Definitions of Corporate Standards and National Standards **Corporate Standards**: - Definition : Corporate standards are standards formulated and implemented by individual companies to regulate their internal production, management, services, and other activities. These standards are typically designed to meet the specific needs and goals of the Corporate, ensuring product quality, production efficiency, safety, and market competitiveness. - Characteristics : Corporate standards are generally developed based on the actual conditions of the company and market demands, making them highly targeted and flexible. These standards may cover technical requirements for products, production processes, inspection methods, packaging, labeling, and more. **National Standards**: - Definition : National standards are standards formulated and published by national

⁶ <https://github.com/mwilliamson/mammoth.js>

standardization management organizations, applicable to relevant industries and fields across the country. National standards are generally aimed at safeguarding public interests, promoting technological progress, maintaining market order, and protecting consumer rights.

- **Characteristics**: National standards are universally applicable and mandatory, typically covering a wide range of industries and sectors. The process of formulating national standards usually involves extensive consultation and expert review to ensure their scientific validity and reasonableness.

Your task is to compare and analyze the two standard documents (both represented in HTML format). Each document contains a series of attribute regulations. You need to provide modification suggestions for the first document (referred to as the "Corporate Standard") based on the content of the second document (referred to as the "National Standard").

```Corporate Standard

In this section, you will find the content of the Corporate Standard. The Corporate Standard includes a series of attributes and their corresponding regulations.

```National Standard

In this section, you will find the content of the National Standard. The National Standard also includes a series of attributes and their corresponding regulations. You should analyze and compare the content of this document with the Corporate Standard. Please note that the National Standard has been streamlined to only retain the parts relevant to the comparison with the Corporate Standard.

If you find that the provided National Standard is not applicable to the Corporate Standard, simply reply with 'IRRELEVANT'. If applicable, then consider the following modification suggestions.

Modification suggestions include the following types:

1. If the Corporate Standard includes content prohibited by the National Standard, provide a modification suggestion, i.e., the Corporate Standard should remove or rectify the prohibited content according to the National Standard.
2. If the Corporate Standard uses outdated or obsolete National Standards, provide a modification suggestion.
3. If a value in the Corporate Standard does not align with the corresponding value in the National Standard, provide a modification suggestion, i.e., the Corporate Standard should conform to the National Standard or be more stringent.
4. Matters that are included in the National Standard but not in the Corporate Standard do not require modification suggestions. If no modifications are needed, please refrain from outputting them to keep the response concise. In summary, only urgent and significant modification suggestions should be raised. Otherwise, please reply with 'NO MODIFICATION'.

Please summarize the following suggestions for modifications to the Corporate standards, merging any repetitive points and organizing the content in a clear and logical manner. Ensure that the final summary is concise, highlights the key points, and maintains clarity and coherence.

Specific requirements:

1. Merge repetitive content: Identify and combine all repeated modification suggestions to ensure that each topic appears only once.
2. Organize the structure: Arrange the summarized content in a logical order, ensuring it flows from general to specific or from important to secondary points.
3. Clear conclusions: While summarizing, focus on suggestions that are clear, actionable, and provide substantial input. Disregard redundant content that is correct but lacks actionable value.

4 Conclusion

In this paper, we demonstrate a tool for resolving standard discrepancies. First, we release the **Food Standard Dataset**, which, to our knowledge, is the first large-scale dataset available for researching the discrepancies between national standards and corporate standards. Second, after extensive tests, we have successfully designed an instruction set that can drive LLMs to generate and summarize modification suggestions for corporate standards based on national standards. We find that this instruction set has good applicability and can effectively work with most models exceeding 30 billion parameters.

In contrast, models with fewer than 30 billion parameters lack sufficient intelligence to follow our instructions. Third, this paper is one of the few LLM applications that utilize HTML as its underlying data. Another concurrent effort is **HtmlRAG** [14], which uses cleaned HTML as static data for the LLM. In contrast, our tool dynamically manipulates the DOM tree generated from HTML before delivering it to the LLM. We believe that HTML may become the foundational data for next-generation LLM models for three significant advantages: (1) it can be viewed as structured data, allowing for precise editing through programming; (2) it serves as text, enabling analysis by LLMs; and (3) it can be rendered by browsers, providing excellent visualization and interactivity.

This work has been supported by the National Key R&D Program of China (Grant No. 2022YFF0608000).

References

1. J. A. Cabral de Barros, *Pharm. Epidemiol. Drug Saf.* **9**, 281--287 (2000)
2. D. Fei, *J. Contemp. China* **33**, 465--485 (2024)
3. J. Mangers, C. Oberhausen, M. Minoufekr, P. Plapper, *Shaping the Future Through Standardization*, 1--26 (2020)
4. P. Zhou, N. M. El-Gohary, *Adv. Eng. Informatics* **48**, 101239 (2021)
5. L. J. McGibbney, B. Kumar, *Proc. Int. Conf. CIB W78*, 64 (2011)
6. R. Zhang, C. Wang, S. Bi, Q. Fu, X. Li, T. Gan, *Proc. Int. Conf. Computing, Networks and Internet of Things*, 525--530 (2023)
7. N. Melluso, I. Grangel-González, G. Fantoni, *Comput. Ind.* **140**, 103676 (2022)
8. J. Dimyadi, G. Governatori, R. Amor, *Proc. Joint Conf. Computing in Construction* **1**, 637--644 (2017)
9. Y. S. Bareedu, T. Frühwirth, C. Niedermeier, M. Sabou, G. Steindl, A. S. Thuluva, S. Tsaneva, N. T. Ozkaya, *Semantic Web* **15**, 517--554 (2024)
10. M. Vincini, D. Beneventano, S. Bergamaschi, *J. Univers. Comput. Sci.* **19**, 1986--2012 (2013)
11. C. Gatto, M. Gholamzadehmir, M. Zampogna, A. Pavan, et al., *Int. Conf. Construction Applications of Virtual Reality*, 841--851 (2023)
12. A. Wulff, M. Mast, M. Hassler, S. Montag, M. Marschollek, T. Jack, *Methods Inf. Med.* **59**, e64--e78 (2020)
13. M. M. Bouh, F. Hossain, A. Ahmed, *Proc. 9th Int. Conf. Inf. Communication Technol. Ageing Well e-Health*, 230--236 (2023)
14. J. Tan, Z. Dou, W. Wang, M. Wang, W. Chen, J.-R. Wen, *HtmlRAG: HTML is Better Than Plain Text for Modeling Retrieved Knowledge in RAG Systems*, arXiv:2411.02959 (2024). Available at: <https://arxiv.org/abs/2411.02959>