AI-Assisted Quality Diagnosis of Lists of Environmental Permits in China: Evidence from Beijing-Tianjin-Hebei

Siyan Cun
School of Environment, Tsinghua University, Beijing, China
Contact: Siyan Cun, cunsy24@mails.tsinghua.edu.cn

Abstract. Lists of environmental permits (LEPs) are the list-based instruments through which China's integrated environmental zoning-based regulation (EZR) is translated into unit-level entry conditions, industrial controls, and management requirements. As LEPs move from compilation to implementation and dynamic updating, regulators need to know not only whether a clause exists, but whether it is specific enough, whether it corresponds to real regulated objects, and whether it differentiates among units with different local conditions. We develop an AI-assisted, evidence-traceable framework for diagnosing LEP quality. In the Beijing-Tianjin-Hebei region, we benchmark domestic large language models against an expert-adjudicated sample and apply the selected model to 2,838 Critical Regulatory Units (CRUs) and Protection in-Priority Units (PPUs). The diagnosis indicates that the main problems lie in insufficiently detailed implementation, selected industry-text mismatches, and remaining template-like wording, rather than a wholesale absence of control items. The framework offers a practical screening tool for targeted review, dynamic LEP updating, and transparent environmental assessment.

Keywords: integrated environmental zoning-based regulation; list of environmental permits; large language model; policy text analysis; quality diagnosis

1. Introduction

Environmental assessment often turns on the quality of policy texts: what objects they regulate, what actions they require, and how clearly they define conditions for implementation. In China, integrated environmental zoning-based regulation (EZR) links ecological protection, pollution prevention, resource constraints, and project access control within a spatially differentiated governance system [1,2]. Its spatial basis is the Integrated Ecological and Environmental Unit (IEEU), and its list-based carrier is the list of environmental permits for human activities (LEP). The LEP organizes differentiated behavioral rules through four dimensions--spatial distribution constraint, environmental risk prevention, emission control, and efficiency management--so that zoning objectives can be translated into unit-level source prevention.

An LEP is not simply a long policy document. It combines policy rules, spatial boundaries, and real regulated objects. A clause can be complete in wording but irrelevant to the industries actually present in a unit; it can target the right object but remain too general for implementation; or it can repeat similar wording for units with different industrial or ecological conditions. Natural language processing has already been used to extract policy keywords and classify regulatory measures in LEPs [3], and the broader policy-text literature has clarified both the value and the risks of automated content analysis [4-6]. For dynamic LEP updating, however, the key question is more operational: does the text match the object and the place?

We therefore build a progressive diagnostic framework for LEP quality. The framework first identifies required textual elements, then checks recognized objects against real-world evidence, and finally compares textual similarity with cross-unit reality differences. The large language model (LLM) extracts structured judgments and original textual evidence; permitting and spatial data then help verify whether clauses align with the management context. The model is used as a screening and evidence-organization tool, not as a substitute for expert judgment.

2. Materials and Methods

The case covers Beijing-Tianjin-Hebei, a region with intensive industry, ecological protection pressure, and cross-boundary governance needs. The policy corpus contains 3,157 raw unit records. After text de-duplication, it includes 3,137 unique text units, 9,628 non-empty regulatory text blocks, and about 1.065 million Chinese characters. Spatial processing linked 3,041 units with external evidence, including 2025 pollutant discharge permits, enterprise points of interest for auxiliary checking, industrial land, land cover, river density, net ecosystem productivity, and water-quality station information.

Before full-scale assessment, we tested whether the model could follow expert judgments on an ordered policy-text task. The benchmark contains 110 samples from 89 units, covering both CRUs and PPUs. Two annotators independently assigned one of four ordered labels: not involved, generally mentioned, effectively referenced, and detailed implementation. Their agreement was high (quadratic weighted Kappa = 0.979), and three disagreements were adjudicated. Among the domestic LLMs tested with the same prompt and output format, GLM-5.1 in non-reasoning mode showed the best task fit, with a quadratic weighted Kappa of 0.602, exact agreement of 57.3%,

and adjacent-grade agreement of 94.55% against the adjudicated benchmark. These results justify batch screening and evidence organization, but not automatic policy adjudication.

We organize the diagnosis in three linked layers. Completeness asks whether required textual elements are present and operational. For CRUs, these elements include industry-specific requirements, production process or treatment links, quantitative indicators, emission control, and efficiency management requirements. For PPUs, they include protected objects, prohibited activities, allowed or restricted activities, exit requirements, and environmental risk prevention or ecological restoration requirements. Spatial adaptation then asks whether recognized controls for steel, thermal power, petrochemical, and building-material industries correspond to licensed production subjects in the same unit. Differentiation adequacy compares text similarity with industrial or ecological reality differences, screening for cases where different units receive similar wording despite different conditions.

Table 1. Progressive diagnostic framework for lists of environmental permits

Layer	Core question	Main evidence	Output for review
Completeness	Are required policy elements written and sufficiently operational?	LEP text; unit type; expert rules	Candidate clauses lacking objects, indicators, or implementation detail
Spatial adaptation	Do listed industry controls correspond to real regulated objects?	LLM-recognized industry controls; 2025 discharge permits	Coverage gap, over-coverage, coverage consistency, or no evidence
Differentiation adequacy	Do different real conditions receive differentiated textual provisions?	Text embeddings; permit and ecological features	High-priority unit pairs for manual comparison

3. Results

The completeness assessment covered 2,838 CRUs and PPUs and produced 12,856 unit-rule judgments. The results do not point to a general absence of regulatory topics. Detailed implementation and effective reference together account for a large share of the records. The more relevant weakness is that some items remain at the level of reference or general wording. In CRUs, quantitative requirements are the weakest element. In PPUs, clauses on allowed or restricted development activities and exit arrangements are less specific than prohibitions or general protection clauses.

The spatial adaptation analysis focused on 1,506 CRUs and four industries. After excluding 24 units without readable LEP text and model-uncertain unit-industry pairs, 5,639 comparable unit-industry combinations remained. The building-material sector showed the most prominent coverage gap: 30.11% of valid combinations had licensed production subjects but no clearly identified corresponding control requirement in the LEP text. The gap rates were 7.29% for steel, 9.09% for thermal power, and 5.93% for petrochemicals. Over-coverage was more visible for petrochemicals and thermal power, at 13.73% and 12.47%, respectively. These categories mark priorities for review rather than direct proof of policy error, because preventive controls, document style, and reference chains may also explain apparent mismatches.

The differentiation analysis shows that many LEP texts are broadly consistent with real differences, but some template-like patterns remain. In the expanded valid sample of 1,472 CRUs with permit-based industry evidence, the median insufficient-differentiation tendency was 0.184 and the 95th percentile reached 0.689. For 1,291 PPUs, the median was 0.186 and the 95th percentile was 0.677. High values identify unit pairs where similar textual provisions coincide with large industrial or ecological differences, such as differences in licensed-industry presence, forest or cropland share, net ecosystem productivity, unit area, and impervious surface proportion.

4. Discussion and Policy Implications

The study moves automated policy-text analysis from classification toward evidence-linked quality diagnosis. Keyword extraction, topic modelling, and supervised classification can reveal policy themes [4,7], but LEP refinement requires a more concrete unit of review: which unit, which industry or ecological object, which clause, and which piece of evidence should be checked? LLMs are useful for this task because they combine semantic extraction, rule-based judgment, and original evidence return in one workflow. At the same time, recent work on LLM-based information extraction and policy coding shows that task-specific validation and safeguards remain necessary [8-11].

For environmental assessment practice, the framework supports three decisions. First, it helps planners and regulators locate clauses that lack quantitative indicators, clear objects, or implementation conditions. Second, it

connects provisions with external evidence, especially discharge permits, to identify industry controls that need confirmation. Third, it screens for insufficient differentiation among units, helping experts detect excessive reliance on template wording. In this workflow, the model narrows the search space, organizes evidence, and records traceable reasons; experts still make the final judgment.

Several limits remain. The 110-sample benchmark is useful for task adaptation and prompt comparison, but it should be expanded with more "not involved" cases and independent regional samples. Permit absence is an operational indicator; it does not prove that no industrial activity exists. The differentiation tendency is a screening metric, not a causal estimate of environmental performance. Future work should add ecological-space data, environmental-quality outcomes, and technology-cost parameters so that LEP quality diagnosis can be linked to bounded scenario analysis of expected environmental and economic effects.

5. Conclusions

Overall, the framework identifies three refinement needs in Beijing-Tianjin-Hebei: clauses that are present but insufficiently detailed, industry requirements that do not fully match licensed production subjects, and similar text used for units with different real conditions. The results show how LLMs can strengthen environmental governance without replacing experts: they turn large, heterogeneous policy documents into structured, traceable, and reviewable evidence. This provides a practical route for dynamic LEP updating and more transparent environmental assessment.

References

[1] General Office of the CPC Central Committee and General Office of the State Council. Opinions on strengthening integrated environmental zoning-based regulation. People's Daily, 2024-03-18.

[2] Wang Z, Li W, Li Y, Qin C, Lv C, Liu Y. The "three lines one permit" policy: an integrated environmental regulation in China. Resources, Conservation and Recycling. 2020;163:105101.

[3] Wei Z, Wang Z, Gong M, et al. Policy content analysis of lists of environmental permits based on natural language processing (NLP). Journal of Environmental Engineering Technology. 2025;15(1):1-10.

[4] Grimmer J, Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Political Analysis. 2013;21(3):267-297.

[5] Linder F, Desmarais BA, Burgess M, Giraudy A. Text as policy: measuring policy similarity through bill text reuse. Policy Studies Journal. 2020;48(2):546-574.

[6] Scott TA, Marantz N, Ulibarri N. Use of boilerplate language in regulatory documents: evidence from environmental impact statements. Journal of Public Administration Research and Theory. 2022;32(3):576-590.

[7] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003;3:993-1022.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

[9] Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Computing Surveys. 2023;56(2):1-40.

[10] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Computing Surveys. 2023;55(9):1-35.

[11] Luitel D, Hassani S, Sabetzadeh M. Improving requirements completeness: automated assistance through large language models. Requirements Engineering. 2024;29(1):73-95.

[12] Larosa F, Hoyas S, Conejero JA, et al. Large language models in climate and sustainability policy: limits and opportunities. Environmental Research Letters. 2025;20(7):074032.

[13] Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27(4):857-871.