Japan Construction Cost Database: An Open Dataset for LLM-Based Cost Estimation and Fraud Detection in Residential Renovation

Toshikatsu Oga (大賀 俊勝)
ORCID: 0009-0000-9180-903X
The Horizons Co., Ltd. (The HORIZ音s株式会社), Hiratsuka, Kanagawa, Japan
contact@the-horizons-innovation.com
Dataset: https://github.com/ogasurfproject-jpg/japan-construction-cost-database

Abstract

We introduce the Japan Construction Cost Database (JCCDB), an openly available structured dataset linking residential construction plan-level pricing, contractor-size-tier margin ranges, and fraud-detection patterns for the Japanese market. The dataset comprises 87 construction plans across 7 renovation categories (屋根工事 roof construction, シロアリ駆除 termite control, 給湯器交換 water heater replacement, 窓リフォーム window renovation, 浴室リフォーム bathroom renovation, キッチンリフォーム kitchen renovation, and 電気工事 electrical work), annotated with 88 fraud-detection patterns categorized by severity. Pricing is derived from the HORIZON SHIELD (HS) Rule, a transparent cost formula grounded in 30 years of hands-on construction management experience combined with 2026 Kanto-region trade-price data. Each plan is stratified across four contractor-size tiers (sole proprietor to major national chain) with empirically derived margin ranges drawn from Ministry of Land, Infrastructure, Transport and Tourism (MLIT) industry analysis. We further describe KIRA, an LLM-based construction cost diagnostic system built on this dataset, and discuss its architecture and alignment with HS Rule pricing. The dataset is released under CC-BY 4.0 and is intended to support NLP/LLM research, consumer-protection studies, and cross-country comparative construction cost analysis.

Keywords: construction cost estimation, fraud detection, open dataset, large language models, Japan, consumer protection, renovation pricing

1. Introduction

Japan's residential renovation market is estimated to exceed \6 trillion annually [MLIT 2024], yet it remains deeply opaque to consumers. Unlike manufacturing or retail, construction prices are rarely disclosed, enabling wide variance in vendor quotations for identical work. The National Consumer Affairs Center of Japan (NCAC/PIO-NET) consistently identifies renovation-related complaints as among the highest-volume consumer dispute categories, with roof repair fraud alone accounting for approximately 30% of all renovation complaints [NCAC 2024].

Door-to-door fraud, insurance-claim manipulation, and material price inflation disproportionately harm elderly homeowners. Fraudulent contractors may quote 2–3× market rates, exploit emergency situations (water heater failure, typhoon damage), or falsify insurance claims. Despite the scale of the problem, no publicly available structured dataset links: (1) specific construction plans with itemized pricing, (2) contractor-size-appropriate margin ranges, and (3) fraud-detection patterns — simultaneously.

The emergence of large language models (LLMs) as reasoning engines opens new possibilities for consumer-facing construction cost advisory systems. However, training and evaluating such systems requires structured, ground-truth pricing data — which has not previously been available for the Japanese market.

This paper makes the following contributions:

1. We release the Japan Construction Cost Database (JCCDB), a structured, bilingual (Japanese/English), CC-BY 4.0 dataset of 87 construction plans and 88 fraud-detection patterns across 7 renovation categories.

2. We describe the HORIZON SHIELD (HS) Rule, a transparent cost formula that decomposes renovation prices into material, labor, overhead, and tax components, with contractor-size-tier stratification.

3. We describe KIRA, an LLM-based diagnostic system that operationalizes the HS Rule through a Cloudflare Worker architecture powered by Anthropic's Claude API, and discuss its alignment properties with the dataset.

2. Related Work

2.1 Construction Cost Estimation

Construction cost estimation has been studied extensively in the civil and structural engineering literature, primarily at the project level for commercial and infrastructure works [Kim et al. 2004; Sonmez 2004]. Machine learning approaches including neural networks and support vector regression have been applied to early-stage cost prediction [Chou et al. 2013; Petruseva et al. 2017]. More recent work has applied deep learning and ensemble methods to cost forecasting across building types [Pham et al. 2020; Elmousalami 2021]. However, these studies focus on large-scale construction rather than residential renovation, and do not address the consumer protection dimension of pricing transparency.

For residential renovation, published benchmarks are scarce. Industry surveys (e.g., Rishop Navi, SUUMO) provide aggregated price ranges but lack itemized breakdowns, contractor-tier stratification, or fraud-pattern annotations.

2.2 LLMs for Domain-Specific Advisory

Recent work has demonstrated that LLMs can serve as effective domain advisors when provided with structured grounding data [Lewis et al. 2020; Guu et al. 2020]. Retrieval-augmented generation (RAG) approaches have been applied to legal consultation [Trautmann et al. 2022] and medical diagnosis [Singhal et al. 2023]. The release of GPT-4 [OpenAI 2023] and subsequent instruction-tuned models has further demonstrated strong performance on structured domain advisory tasks, including cost estimation in engineering contexts [Zheng et al. 2023]. Construction cost advisory represents an underexplored application domain, particularly for consumer-facing systems in non-English languages.

2.3 Consumer Fraud Detection

Fraud detection in service markets has been studied in insurance [Baesens et al. 2015] and e-commerce [Akoglu et al. 2015]. Construction-specific fraud patterns have not, to our knowledge, been catalogued in a machine-readable, publicly available format prior to this work.

3. The Japan Construction Cost Database

3.1 Overview

JCCDB v1.0 covers seven renovation categories selected on the basis of fraud incidence (NCAC complaint volume), consumer financial exposure, and 2026 market relevance (active government subsidy programs). Table 1 summarizes the dataset.

Table 1: JCCDB v1.0 dataset summary.

Category ID	Japanese	English	Plans	Red Flags	Max Subsidy 2026
roof-construction	屋根工事	Roof Construction	8	13	—
termite-control	シロアリ駆除	Termite Control	13	13	—
water-heater-replacement	給湯器交換	Water Heater Replacement	7	15	\240,000

window-renovation	窓リフォーム	Window Renovation	8	15	\1,000,000
bathroom-reform	浴室リフォーム	Bathroom Renovation	6	15	—
kitchen-reform	キッチンリフォーム	Kitchen Renovation	6	13	—
electrical-work	電気工事	Electrical Work	39	4	—
Total			**87**	**88**	

3.2 The HS Rule Pricing Formula

Each plan's benchmark price (hs_rule_jpy) is computed using the HORIZON SHIELD Rule:

```
hs_rule_jpy = (material_trade_price × 1.2)
    + (labor_man_days × prevailing_day_rate_kanto)
    + auxiliary_costs
    + (subtotal × overhead_rate_15%)
    + (subtotal × tax_rate_10%)
```

The material markup factor of 1.2 reflects established Japanese trade conventions for materials handling, transportation, and minimal stocking costs, consistent with standard contractor procurement margins documented in MLIT construction industry analysis [MLIT 2023]. Labor day-rates are drawn from the 2023 National Confederation of Construction Worker Unions (全建設総連 Zenkensoren) Tokyo wage survey — the most recently published edition available at the time of writing, as 2024–2025 survey data had not been publicly released — ranging from \16,899/day (painter) to \35,000/day (carpenter). The 15% overhead rate and 10% consumption tax are standard parameters grounded in MLIT construction industry analysis.

Material trade prices reflect manufacturer-distributor pricing verified across major Japanese residential categories. Unit bathroom bodies (TOTO, LIXIL) are available to contractors at approximately 40% of list price (60% off); system kitchens at 50% off; window products at 50% off; and gas water heaters at 70% off. These ratios have been verified through direct supplier engagement over 30 years of CMR (Construction Management-Researcher) practice.

3.3 Contractor-Size Tier Stratification

A key methodological contribution of JCCDB is the stratification of each plan across four contractor-size tiers, with empirically derived margin ranges. Table 2 defines the tier structure.

Table 2: Contractor-size tier definitions and margin ranges (source: MLIT/CIIC).

Tier	Japanese	Description	Headcount	Margin Range
1	個人事業	Sole proprietor / local specialist	1–3	25–35%
2	中小工務店	Small-mid local contractor	5–30	25–35%
3	人気工務店	Premium specialist (brand-recognized)	30–200	30–40%
4	大手企業	Major national chain	200+	35–45%

Tier price ranges are computed as: tier_min = hs_rule_jpy × multiplier_min; tier_max = hs_rule_jpy × multiplier_max, where multipliers are (0.95, 0.98), (1.00, 1.03), (1.05, 1.10), and (1.12, 1.18) for tiers 1–4 respectively. The tier-1 range falls below the HS Rule because sole proprietors lack the corporate overhead included in the standard formula. Margin ranges are grounded in MLIT Construction Industry Management Analysis and the Construction Industry Information Management Center (CIIC) data series [MLIT 2023].

3.4 Fraud-Detection Patterns

Each category includes a structured catalog of fraud-detection patterns ("red flags"), encoding domain knowledge about contractor deception tactics. Each record contains: flag_code (unique ID), severity (CRITICAL / HIGH / MEDIUM), pattern_en, pattern_ja (bilingual description), reason_en (explanation), and fraud_rate_pct (where empirically available from NCAC data).

CRITICAL patterns indicate immediate refusal is warranted: these include unannounced door-to-door inspection solicitations (90–95% fraud rate for termite control [NCAC 2024]), insurance fraud facilitation, and work by unlicensed gas or electrical workers (violating Gas Business Act and Electrical Appliances and Materials Safety Act respectively). HIGH patterns signal significant risk requiring documentation and second opinions. MEDIUM patterns warrant careful price comparison.

Notable 2026 market-specific patterns include: (1) TOTO/LIXIL unit bath order suspension (April 2026) due to naphtha supply disruption, with fraudulent contractors exploiting delivery uncertainty to pressure consumers; (2) roofing membrane price surge (+40–50%, Q2 2026) used to justify inflated quotes; and (3) fraudulent subsidy claims for the government's Advanced Window Renovation 2026 program (先進的窓リノベ2026事業, max \1,000,000/household) [MOE 2026].

Note: Market-specific patterns reflect conditions as of April–May 2026 and are subject to revision in subsequent dataset versions. The quarterly update cadence described in Section 5.3 is intended to maintain currency.

3.5 Data Format and Access

JCCDB is distributed as CSV files with bilingual column headers. Each category directory contains: plans.csv (English), plans_ja.csv (Japanese), red_flags.csv, and README.md with category-specific methodology notes. A JSON Schema (schema/plan_schema.json) is provided for validation. The dataset is hosted at: https://github.com/ogasurfproject-jpg/japan-construction-cost-database and released under Creative Commons Attribution 4.0 International (CC-BY 4.0), permitting commercial use, redistribution, and derivative works with attribution.

4. KIRA: An LLM-Based Construction Cost Diagnostic System

4.1 System Architecture

KIRA (建設費インテリジェンス・リバース・アドバイザー — Construction Intelligence and Renovation Advisor) is a production LLM-based diagnostic system built on JCCDB. The system is publicly accessible as a case study of dataset operationalization; deployment URLs are provided in the dataset repository. KIRA operationalizes the HS Rule in real time, accepting natural-language renovation queries from Japanese homeowners and returning itemized price diagnostics with fraud-risk assessment.

The architecture consists of three layers: (1) a Cloudflare Worker endpoint handling routing and system-prompt injection; (2) the Anthropic Claude API (claude-haiku-4-5) as the reasoning engine; and (3) a material price database (在材DB zaisai-db, 3,350+ line items) and plan lookup module embedded in the Worker. The system prompt encodes the HS Rule formula, Kanto-region labor rates, trade-price ratios, and contractor-tier multipliers as structured context.

4.2 HS Rule Alignment

A key design challenge for LLM-based cost estimation is non-determinism: the same query may produce different price outputs across invocations. KIRA addresses this through deterministic plan-key resolution. Rather than asking the LLM to compute prices from first principles, KIRA's Worker layer resolves the query to a canonical plan_key (e.g., konsento_new_wiring for new-wiring outlet installation), then applies the HS Rule formula deterministically to generate tier-stratified prices.

This hybrid approach — LLM for intent classification and plan-key extraction, deterministic formula for pricing — achieves consistency while preserving the natural language interface. Worker-level unit tests (40/40 pass rate across all

plan categories) validate that the HS Rule is correctly applied post-LLM classification. Test scripts are maintained in the accompanying GitHub repository.

4.3 Fraud Detection Integration

KIRA integrates the red-flag catalog as a retrieval component. When a user's query or uploaded quote matches a red-flag pattern, KIRA surfaces the corresponding severity rating and defense recommendation. CRITICAL-severity flags trigger immediate warnings. HIGH and MEDIUM flags are presented as advisory notes within the diagnostic output.

5. Discussion

5.1 Research Applications

JCCDB enables several previously impossible research directions. For NLP/LLM research, it provides a bilingual ground-truth pricing corpus for fine-tuning or evaluating models on Japanese construction cost estimation. The fraud-pattern catalog with severity labels supports training fraud-detection classifiers. For consumer protection research, the tier-stratified pricing structure allows empirical study of markup dispersion across contractor sizes. For comparative construction economics, the HS Rule formula and Kanto-region labor rates provide a replicable benchmark that could be adapted to other regional or national markets.

5.2 Limitations

Several limitations should be noted. First, the dataset reflects 2026-04 Kanto-region market conditions; labor rates and material prices differ by ±10–20% in other Japanese regions. Second, hs_rule_jpy values are normative benchmarks derived from expert judgment, not empirical transaction records. Third, the electrical-work category (39 plans) is more granular than other categories (6–13 plans), reflecting the breadth of electrical services. Fourth, commercial and industrial construction are out of scope. Fifth, fraud-rate percentages are available only for a subset of red-flag patterns with NCAC empirical support; others reflect practitioner judgment.

Sixth, JCCDB is constructed by a single practitioner-researcher whose commercial service (HORIZON SHIELD) operationalizes this dataset. While this provides domain-depth advantages — specifically, access to supplier pricing and fraud-pattern data unavailable through desk research — it introduces potential confirmation bias in both pricing benchmarks and fraud-pattern selection. Independent validation by third-party researchers and practitioners is strongly encouraged, and the CC-BY 4.0 license is intended to facilitate such review.

Seventh, the dataset is descriptive rather than inferential: no statistical hypothesis testing is conducted on tier-margin distributions or red-flag inter-rater reliability. Such validation is left to future work as additional data and independent annotators become available.

5.3 Update Cadence

The maintainer commits to quarterly updates incorporating: material price shifts exceeding 5% category-wide, new or expired subsidy programs, new fraud patterns from NCAC complaint data, and methodological refinements based on community feedback. The GitHub repository's issue tracker is the primary channel for corrections and contributions.

6. Conclusion

We have introduced JCCDB, an openly available structured dataset linking construction plan-level pricing, contractor-tier margin ranges, and fraud-detection patterns for the Japanese residential renovation market. The dataset encodes 30 years of hands-on CMR expertise in a machine-readable format suitable for LLM training, consumer protection research, and comparative market analysis. We have described the HS Rule as a transparent, replicable pricing formula, and KIRA as a production LLM system that operationalizes this formula for real-world consumer advisory. We release the dataset under

CC-BY 4.0 and invite the research community to build upon, critique, and extend this foundation.

Conflict of Interest Declaration

The author is the founder and Chief Executive Officer of The Horizons Co., Ltd. (The HORIZ音s株式会社), which operates HORIZON SHIELD, a commercial construction cost diagnostic service that operationalizes the JCCDB dataset described in this paper. The dataset itself is released under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license to enable independent validation and reuse by third-party researchers. The author declares no other competing financial or non-financial interests. No external funding was received for the construction of this dataset or the preparation of this manuscript.

AI Usage Disclosure

In accordance with the preprint server's AI usage policy, the author discloses the following uses of AI and large language model (LLM) tools in the preparation of this manuscript and the construction of the associated dataset:

1. Research subject (not authorship assistance): The KIRA system described in Section 4 is itself an LLM-based diagnostic system built on the Anthropic Claude API. This system constitutes the research artifact under study, not a tool used to write this paper.

2. Pre-writing assistance: AI tools were used for literature search organization and idea structuring during manuscript preparation. All cited references were independently verified by the author for existence, accurate authorship, and faithful representation in the in-text citations.

3. Copy-editing and formatting: AI tools were used for English-language editing (grammar, clarity, idiom) and PDF formatting assistance. The substantive content, claims, and conclusions remain authored by the human author.

4. Data construction: The JCCDB dataset (87 construction plans, 88 fraud-detection patterns across 7 categories) was constructed manually by the author based on 30 years of Construction Management-Researcher (CMR) practice, supplier engagement, and Ministry of Land, Infrastructure, Transport and Tourism (MLIT) industry analysis. No AI-generated data, synthetic plans, or fabricated red-flag patterns are included in the dataset. Pricing values (hs_rule_jpy) were computed deterministically using the HS Rule formula described in Section 3.2.

5. Author responsibility: The author has thoroughly reviewed all AI-assisted content and retains full responsibility for the accuracy of factual claims, references, methodological descriptions, and conclusions of this manuscript. No portion of this paper consists of verbatim AI-generated text presented as the author's own writing.

Data Availability

All data described in this paper are openly available at https://github.com/ogasurfproject-jpg/japan-construction-cost-database under the CC-BY 4.0 license. The repository includes CSV files, JSON Schema for validation, and category-specific README files documenting methodology.

References

[Akoglu et al. 2015] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph-based anomaly detection and description: a survey. Data Mining and Knowledge Discovery, 29(3), 626–688.

[Baesens et al. 2015] Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques. Wiley.

[Chou et al. 2013] Chou, J. S., Pham, A. D., & Wang, H. (2013). Bidding strategy to support decision-making by integrating fuzzy AHP and regression-based simulation. Automation in Construction, 35, 517–527.

[Elmousalami 2021] Elmousalami, H. H. (2021). Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative study. IEEE Transactions on Engineering Management, 68(1), 183–196.

[Guu et al. 2020] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). Retrieval augmented language model pre-training. ICML 2020.

[Kim et al. 2004] Kim, G. H., Yoon, J. E., An, S. H., Cho, H. H., & Kang, K. I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. Building and Environment, 39(11), 1333–1340.

[Lewis et al. 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 2020.

[METI 2026] Ministry of Economy, Trade and Industry, Agency for Natural Resources and Energy. "給湯省エネ2026事業 Water Heater Energy Saving 2026 Program." https://kyutou-shoene2026.meti.go.jp/, 2026.

[MLIT 2023] Ministry of Land, Infrastructure, Transport and Tourism, Japan. "Construction Industry Business Management Analysis." Construction Industry Information Management Center (CIIC), 2023. https://www.mlit.go.jp/totikensangyo/const/totikensangyo_const_fr2_000031.html

[MLIT 2024] Ministry of Land, Infrastructure, Transport and Tourism, Japan. "住宅リフォームの市場規模 (Residential Renovation Market Scale)." Construction Industry Analysis Division, 2024. https://www.mlit.go.jp/jutakukentiku/house/jutakukentiku_house_tk3_000059.html

[MOE 2026] Ministry of the Environment, Japan. "先進的窓リノベ2026事業 Advanced Window Renovation 2026 Program." https://window-renovation2026.env.go.jp/, 2026.

[NCAC 2024] National Consumer Affairs Center of Japan. "消費者安全法に基づく事故情報及び消費生活相談情報 — 住宅リフォーム関連 (Consumer Safety Act: Accident and Consultation Statistics — Renovation-related)." PIO-NET Consumer Consultation Statistics, 2024. https://www.kokusen.go.jp/soudan_topics/data/reform.html

[OpenAI 2023] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[Pham et al. 2020] Pham, A. D., Ngo, N. T., Ha Thi, T. N., Ngo, T. D., & Pham, N. D. (2020). Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. Journal of Cleaner Production, 260, 121082.

[Petruseva et al. 2017] Petruseva, S., Sherrod, P., Panchovska, V., & Sherrod, P. (2017). Predicting the cost performance of construction projects using a neural network model. Tehni č ki vjesnik, 24, 195–202.

[Singhal et al. 2023] Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

[Sonmez 2004] Sonmez, R. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. Canadian Journal of Civil Engineering, 31(4), 677–683.

[Trautmann et al. 2022] Trautmann, D., Petrova, A., & Schilder, F. (2022). Legal prompt engineering for multilingual legal judgement prediction. arXiv preprint arXiv:2212.02199.

[Zheng et al. 2023] Zheng, L., Chiang, W. L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-bench and Chatbot Arena. NeurIPS 2023.

[Zenkensoren 2023] 全建設総連東京都連 National Confederation of Construction Worker Unions, Tokyo Metropolitan Federation. "2023年賃金調査 (2023 Wage Survey)." Tokyo, 2023. https://www.zenkensoren.org/