A Demand-Side Benchmark for Consumer-Facing Construction Cost Questions: Price-Figure Span, Output Consistency, and the Case for a Verifiable Reference Layer

Toshikatsu Oga
Independent researcher, The HORIZONs, Hiratsuka, Kanagawa, Japan; ORCID 0009-0000-9180-903X
Correspondence: contact@the-horizons-innovation.com

Abstract

Consumers facing home-renovation quotes operate in a classic credence-goods market: they cannot readily verify whether a quoted price is fair, and general-purpose large language models (LLMs) are now a zero-cost place to ask. Whether LLM answers are actionable for this purpose is untested. Demand-side benchmarks exist for medical, legal, and financial advice, but not for construction costs. We present, to our knowledge, the first consumer-question benchmark for construction costs. Forty Japanese renovation-price questions were posed to frontier LLMs, with repeated-trial sets measuring output stability. A matched re-run at bare provider defaults with a current frontier model (gpt-5.5) was added to remove a settings confound present in the original configuration. Two findings are robust across models, generations, and settings: no LLM answer contained an explicit over-charge decision threshold, and repeated runs of the same question returned materially different price figures. Within-answer price spans are also wide, with a median of 10x under bare defaults. A deterministic structured engine over an open cost database is included as an existence proof that a citable reference layer is constructible. Its consistency is a design property and its accuracy is not validated here; validating it against completed real-world quotations is the next study. All questions, raw outputs, harness, and scoring code are public.

Keywords: large language models; construction cost estimation; benchmarking; consumer protection; credence goods; output consistency; open data; verifiable AI

1. Introduction

Home renovation is an archetypal credence good. A homeowner who receives a quotation for exterior painting or a bathroom replacement typically cannot judge, before the work and often even after it, whether the quoted quantities, unit prices, and overheads are appropriate [1,2]. The market conditions that theory identifies as conducive to overcharging are all present: infrequent purchases, large stakes, pronounced information asymmetry, and weak verifiability of what was actually delivered [2,3]. Field experiments confirm the mechanism in adjacent credence-goods markets, where providers systematically charge uninformed consumers more, from taxi rides to repair services [4]. The landmark laboratory test granted buyers verifiability as an experimental condition and found at best a minor effect on efficiency [3]. The literature has treated this as an open puzzle, attributing it in part to the absence of any real instrument by which buyers could verify [5]. The renovation and construction sector is among the largest household expenditures where these conditions apply, yet it has attracted almost no demand-side empirical attention [5].

Into this gap has stepped a new possible advisor. Consumers can now put price questions to general-purpose large language models (LLMs) at no cost, asking, for example, "How much should exterior painting cost for a 100 m² house?" The appeal is obvious: an LLM is free, immediate, and unaffiliated with any contractor. The reliability literature, however, gives reasons for caution. LLMs hallucinate plausible-sounding specifics [6]. The variability of their generations is itself a diagnostic signal of unreliable content [7]. Their outputs also fail to reproduce across runs even under nominally deterministic settings in hosted environments [8], a phenomenon not attributable to sampling temperature alone [9]. Whether these general concerns translate into practical unfitness for consumer price guidance is an empirical question, and one that no benchmark currently answers for construction costs.

Demand-side evaluations, meaning benchmarks built from the questions consumers actually ask and scored on properties consumers actually need, exist in neighboring high-stakes advice domains.

Physician-versus-chatbot comparisons on real patient questions established the paradigm in medicine [10], including reliability under repeated questioning [11]. Large-scale hallucination profiling has done the same for legal advice [12,13], and financial-literacy testing has probed LLMs as a source of consumer financial guidance [14]. Domain-specific benchmark construction is by now methodologically mature [15]. In Japanese, LLM evaluation has concentrated on professional licensing examinations [16], and for construction costs no consumer-question benchmark exists in any language, to our knowledge.

The construction-informatics literature, for its part, has studied cost estimation intensively but almost exclusively from the supply side. Three decades of AI-based parametric cost modeling serve estimators, contractors, and owners [17], and recent methodological reviews critique that literature's reliance on small, proprietary datasets [18]. Work on LLMs in construction likewise targets professional workflows: opportunity mapping [19,20], schedule generation [21], and natural-language interfaces to structured project data [22]. What a lay consumer gets when asking an LLM about a renovation price, the demand side of the same market, has not been measured.

This paper measures it. We pose 40 Japanese consumer-facing renovation cost questions to frontier LLMs under provider-default settings, matched across models, and add repeated-trial sets to quantify output stability. We score properties that bear on whether an answer is actionable for a consumer: the span of price figures within an answer (median spans of 5x to 15x by model, with one answer's figures spanning 1,400x), the presence or absence of a decision threshold ("above X, seek a second opinion"), and run-to-run consistency of the figures themselves. Alongside the LLMs we place a deterministic structured cost engine built on an open Japanese construction cost database (JCCDB). Its epistemic status is stated plainly. Its run-to-run consistency is a design property declared upfront, not an empirical discovery, and its point-estimate accuracy against completed real-world projects is not validated here; that validation is the next study. The engine serves in this paper as an existence proof that a verifiable, citable reference layer for consumer cost questions can be built from open data today.

The contributions are four. First, the paper provides the first demand-side consumer benchmark for construction cost questions, released in full with questions, raw outputs, and scoring code. Second, it defines consumer-relevant metrics, namely within-answer figure span, threshold presence, and repetition consistency, that measure actionability rather than examination accuracy. Third, it offers an existence proof that a deterministic, verifiable reference layer for such questions is constructible from open data, with its limits stated. Fourth, it builds an interdisciplinary bridge between credence-goods economics, LLM evaluation, and construction informatics, with an explicit accuracy-validation roadmap. The author's commercial interest in the engine is disclosed and mitigated by the public replication package and by the fact that the paper's primary empirical contribution, the measurement of general-purpose LLMs, stands independently of the engine.

2. Related Work

2.1. Reliability and Consistency of LLM Outputs

Hallucination, meaning fluent but unfounded content, is the headline failure mode of LLMs, and its taxonomy and mitigation are the subject of a substantial survey literature [6]. A recurring finding is that variability across generations carries signal about reliability: semantic entropy over multiple sampled answers detects confabulations [7], building on the insight that meaning-level variation is what matters for uncertainty estimation [23]. Sampling multiple reasoning paths and aggregating them improves accuracy precisely because individual runs are unstable [24]. Two results are particularly relevant to a pricing benchmark. First, temperature alone does not explain output instability [9]. Second, commercial LLM APIs fail to reproduce outputs even at nominally deterministic settings in hosted environments [8]. Together these motivate our protocol choices: provider-default (matched) settings, repeated trials as a first-class measurement, and consistency reported as a consumer-facing property.

2.2. Demand-Side Benchmarks for Consumer Advice

High-stakes advice domains have developed evaluations built from real consumer questions. In medicine, chatbot answers to patient questions were compared with physician answers along quality and empathy dimensions [10], and follow-up work examined accuracy alongside reliability under repeated questioning [11]. In law, hallucination rates were profiled at scale, with harm falling disproportionately on those who cannot afford professional verification [12]; even retrieval-augmented commercial legal tools hallucinate at non-trivial rates [13]. In finance, GPT-family models were assessed both on financial-literacy instruments and on how laypeople use them for advice [14]. Methodologically, expert-informed benchmark construction is well established [15]. Japanese-language evaluation, however, has centered on professional examinations [16], which measure credentialed knowledge rather than the actionability of answers to lay questions. No comparable benchmark exists for construction costs in any language, to our knowledge; this paper supplies one.

2.3. Cost Estimation and LLMs in Construction Informatics

AI-based construction cost estimation has a three-decade history, surveyed across twenty-plus model families [17]. Its orientation is consistently supply-side: models are trained on project databases to serve estimators and owners. Recent critical reviews document methodological weaknesses such as small samples, proprietary data, and limited external validity, and they call for more transparent data foundations [18]. The arrival of LLMs has produced a fast-growing companion literature: opportunity-and-limitation maps for GPT-class models in construction [19], empirical probes of LLM performance on professional tasks such as project scheduling [21], adoption analyses for text-based generative AI [20], and LLM-driven natural-language interfaces over structured building data [22]. Closest to our territory are direct evaluations of LLM cost estimates for professional estimating: a bridge-rehabilitation bid-pricing pilot reporting consistency and reliability concerns with ChatGPT-derived estimates [25], and prompting frameworks for conceptual cost estimation benchmarked against professional cost data [26]. Our work differs on two axes. First, the questions come from consumers, not professionals. Second, our structured baseline inverts the usual pipeline: rather than training a model on closed project data, it computes deterministically over an open cost database, so that any figure it produces can be traced to a citable source.

2.4. Credence Goods and the Economics of Expert Markets

The economics of credence goods provides the market framing. The concept originates with Darby and Karni [1]; the canonical survey formalizes when overtreatment, undertreatment, and overcharging arise as equilibrium phenomena, with verifiability and liability as the key institutional levers [2]. The empirical record is more complicated than the theory. The landmark laboratory experiment endowed buyers with verifiability by design and found it had at best a minor effect on efficiency, against theoretical expectation [3]; this remains an open puzzle [5]. One reading of the null result is that conferred verifiability abstracts away the instrument a real buyer would need. In the experiment, verification was a rule of the game, whereas in a real market it must be acquired, that is, sought out, consulted, and brought to a negotiation. What the field evidence does show is that the buyer's informational position matters: visibly uninformed customers are charged systematically more [4], and everyday repair and construction services remain strikingly under-studied relative to healthcare and vehicle repair [5]. This literature does two jobs here. It explains why consumer-side price information matters, since asymmetry rather than scarcity of contractors is the binding constraint [4]. And it frames the open question to which a reference layer speaks: whether an acquired, instrument-based form of buyer-side information can do in the field what conferred verifiability did not do in the laboratory [3]. Building such an instrument is the precondition for asking that question empirically.

2.5. AI Advice and Consumer Protection

A normative literature is emerging on the obligations attached to AI-generated advice. Careless LLM speech has been analyzed as a distinct, cumulative societal harm with no current legal duty of truthfulness attached [27]; empirical profiling of legal hallucinations frames unreliable AI advice

explicitly as a consumer-protection problem [12,13]. Our benchmark contributes the construction-cost instance of this agenda, and our metrics of span, thresholds, and consistency are designed so that regulators or platforms could re-run them as conformance checks; the full pipeline is public.

2.6. Positioning

At the intersection of demand-side advice benchmarks, construction cost informatics, and credence-goods economics we find no prior work: no benchmark of consumer construction-cost questions, no consumer-side evaluation in construction informatics, and no credence-goods analysis that measures the new de facto advisor, the LLM. This paper occupies that empty intersection.

3. Materials and Methods

3.1. Question Set

The benchmark comprises 40 Japanese renovation-price questions phrased as a consumer would ask them, spanning six categories of residential work: exterior (Q1–8), water-related fixtures and equipment (Q9–16), interior (Q17–22), electrical and heating (Q23–28), maintenance, demolition and grounds (Q29–35), and large-scale renovation (Q36–40). The full set, with English translations, is included in the public replication package.

3.2. Systems and Settings

The original study evaluated GPT-4o (temperature 0.7, 1,200-token cap), Gemini 2.5 Flash (provider defaults), and a deterministic structured engine on the open JCCDB. To address the matched-settings concern, we re-ran the cost questions under bare provider defaults (no temperature override, no token cap, no system prompt) using the current OpenAI flagship, gpt-5.5. GPT-4o itself was excluded from the re-run because it (a) carried the non-default settings at issue and (b) was being deprecated by its provider at the time of writing; gpt-5.5 is its provider-designated successor. A current Gemini re-run (gemini-3.5-flash and gemini-2.5-flash, defaults) was begun but limited by free-tier daily quota; it is therefore reported only as a prose observation, not tabulated. The re-run harness (per-provider default calls, per-call JSON logging including token usage, resume support) and the re-run raw outputs are archived with version 2 of the replication record.

3.3. Procedure

Three passes were collected. The cross-sectional pass posed all 40 questions once per system. The broad-repetition pass posed the first twenty questions five times each. The deep-repetition pass posed four characteristically wide questions (Q31 roof-leak, Q37 condominium, Q38 30-year detached, Q39 barrier-free) five times each. The engine is deterministic by design and was not re-run for stability.

3.4. Metrics

The span metric (referred to as "width" in the original release) is the largest figure stated in man-yen (10,000-yen units) anywhere in an answer divided by the smallest, rounded to one decimal place. Two properties of this measure are stated explicitly. First, it spans all man-yen figures, including headline ranges, itemized components, and unit rates alike, and is therefore monotone in answer content: a longer answer can only widen, never narrow, its span. It measures the numerical dispersion a consumer must integrate, not the narrowest range a model explicitly offers; annotating explicitly offered headline ranges is a finer measurement left to future work, for which all raw outputs are released. Second, the figure-extraction convention is inherited unchanged from the original release for comparability, including its handling of compound notations, and the scorer is public so that the convention can be audited. An answer containing fewer than two man-yen figures has no computable span; such answers are recorded as not scorable and excluded, because a missing range is not a tight range. The decision-threshold metric records the presence of an explicit over-charge ceiling marker. Run-to-run consistency is the sample standard deviation of the span across the five repeated runs, over valid runs only. The study is deliberately descriptive: we report medians, quartiles, and per-question dispersion rather than

inferential statistics, and release per-run data for finer treatment. We verified that the re-run scorer reproduces the originally published figures exactly on the original outputs.

3.5. Reproducibility

All questions, raw outputs, and the scoring script are archived at Zenodo (DOI: 10.5281/zenodo.20522846, CC BY 4.0). The matched re-run harness writes per-call records including token usage, enabling independent re-scoring.

4. Results

4.1. Cross-Sectional Within-Answer Figure Span

Table 1. Within-answer price-figure span, cross-sectional pass (40 questions, one run per model).

System (settings)	Scored	Median	Q1 / Q3	Max (qid)	Not scorable	Threshold
GPT-4o, original (temp 0.7, cap)	27/40	5.0	3.0 / 15.0	1,400 (Q37)	13	0/40
gpt-5.5, re-run (bare defaults)	39/40	10.0	4.0 / 25.0	300 (Q35)	1	0/40
Gemini 2.5 Flash, original (defaults)	33/40	15.0	7.2 / 42.5	250 (Q38)	7	0/40

Because the span statistic is monotone in answer content (Section 3.4), the uncapped re-run mechanically tends toward wider spans and higher scorability: gpt-5.5 yields a computable span on 39 of 40 questions (median 10x) where the capped GPT-4o yielded 27 (median 5x). We therefore do not read the 5x-to-10x shift as models worsening; we read it as the original token cap having truncated the figure-rich answers that bare defaults produce. The extreme values illustrate what the span captures. The 1,400x maximum arises within one answer that quotes a 10,000-yen per-square-metre unit rate alongside a 14-million-yen project total, figures a consumer must reconcile unaided. What the cross-sectional pass establishes, in every configuration, is that a consumer asking a price question receives an answer whose figures span multiples, from 5x to 15x at the median, with the integration work left to the reader (Table 1, Figure 1). A partial current-Gemini re-run at defaults (22 scorable of 22 answered for gemini-3.5-flash and 21 of 23 for gemini-2.5-flash, quota-limited and excluding the large-scale category) showed medians of 15x and 10x, directionally consistent; we do not tabulate it.

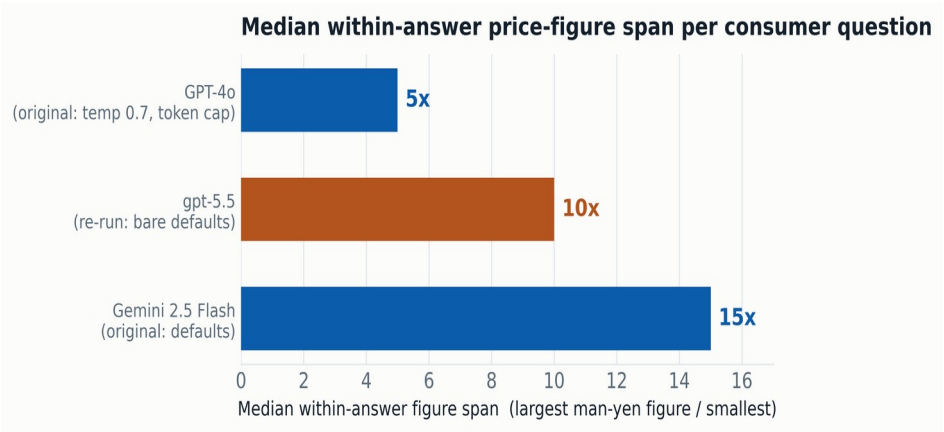

Figure 1. Median within-answer price-figure span, by model and settings (cross-sectional pass).

4.2. Span Distribution

Table 2. Span distribution, scored questions only.

System	>2x	>10x	>50x	scored n
GPT-4o (original)	23	8	2	27
gpt-5.5 (re-run)	36	17	4	39

System	>2x	>10x	>50x	scored n
Gemini 2.5 Flash (original)	31	18	7	33

Under bare defaults, 36 of gpt-5.5's 39 scorable answers span more than 2x, and 17 span more than 10x (Table 2).

4.3. Decision Thresholds

No LLM answer, in any configuration, whether original or re-run, contained an explicit over-charge ceiling marker (GPT-4o 0/40, gpt-5.5 0/40, Gemini 2.5 Flash 0/40; likewise 0 in both partial Gemini re-runs). This result does not depend on the span metric or on answer length. A numeric threshold, meaning an amount above which a received quote warrants challenge, is one directly actionable element an answer can carry, and it is absent from general-purpose LLM answers regardless of model generation or settings. Whether its absence reflects capability or deliberate provider caution in giving definitive financial advice, the consumer-facing outcome is the same. The structured engine's released outputs include such a marker where its reference data supports one (3 of 40 questions), reported here as a structural difference in answer format, not a performance ranking.

4.4. Deep Repetition: Four Wide Questions, Five Runs

Table 3. Median / sample standard deviation of the span across five identical runs.

Question	GPT-4o (orig)	Gemini 2.5 (orig)	gpt-5.5 (re-run)
Q31 roof-leak repair	35.0 / 41.6	50.0 / 47.5	200.0 / 77.8
Q37 condominium renovation	150.0 / 741.1	200.0 / 100.0	100.0 / 69.5
Q38 30-yr detached renovation	15.0 / 22.4	13.3 / 60.6	43.8 / 8.7
Q39 barrier-free renovation	100.0 / 76.0	166.7 / 101.1	100.0 / 75.8

The instability survives the matched re-run intact: gpt-5.5, asked the same question five times, returns answers whose figure span varies with sample standard deviations up to 78 (roof-leak repair; Table 3, Figure 2). Re-running at clean defaults did not stabilize the answers: the figures a consumer receives depend materially on which run they happen to get.

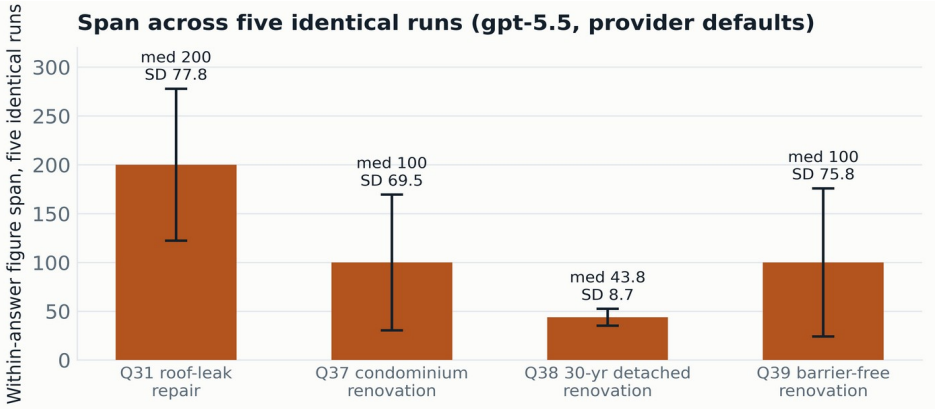

Figure 2. Within-answer figure span across five identical runs (gpt-5.5, bare defaults); bars show the median, whiskers the sample standard deviation.

4.5. Broad Repetition: Twenty Questions, Five Runs

Table 4. Broad repetition summary.

	GPT-4o (orig)	Gemini 2.5 (orig)	gpt-5.5 (re-run)
Questions with computable SD	14/20	19/20	19/20
SD range across questions	0.29–34.2	0.0–97.6	0.25–26.2

Instability is the norm, not the exception: on 19 of 20 questions gpt-5.5's figure span varies across repeats (Table 4).

4.6. Matched Re-Run: The Finding Is Robust to Settings and Model Generation

A natural objection to the original study is that it confounded model with configuration: GPT-4o ran with a raised temperature and a token cap while Gemini ran at defaults. The matched re-run removes that confound. Its result is not that the numbers are unchanged, since the span statistic is monotone in answer content and uncapped answers tabulate wider (Section 4.1); it is that the consumer-facing properties of concern survive the correction. Two of the three properties do not depend on the span definition at all: no configuration produced a decision threshold (Section 4.3), and repeated runs remained materially unstable (Sections 4.4 and 4.5). The third property, wide within-answer spans, persists at bare defaults (median 10x) with the mechanical caveat stated in Section 4.1. We therefore report the actionability gap as a property of current general-purpose LLMs on these questions rather than an artifact of one model's settings; a complete same-generation Gemini pass, quota-limited here, is a straightforward extension.

5. Discussion

5.1. Implications for Construction Informatics

The construction-informatics literature has invested three decades in supply-side cost models [17,18], yet the first place a consumer now encounters a "cost estimate" is none of these systems. It is a general-purpose LLM. Our results indicate that this de facto front door leaves the consumer's core problem unsolved: within-answer figure spans of 5x to 15x at the median across configurations, no decision thresholds in any configuration, and figures that change across runs. For the field, this reframes an old topic. Cost-estimation research has optimized accuracy for professionals holding project data; the open problem exposed here is reference infrastructure for non-professionals holding nothing but a quotation. The critical reviews' call for transparent, open data foundations [18] acquires a consumer-protection rationale in addition to a methodological one: a reference layer can only be citable if its data are open.

5.2. A Credence-Goods Reading of the Results

Credence-goods theory names verifiability as an institutional lever [2], yet the landmark experiment found that verifiability conferred by experimental fiat moved outcomes at best marginally [3], while field evidence consistently shows that the buyer's informational position changes what they are charged [4]. Read through this lens, the two artifacts examined here occupy different institutional roles. An LLM answer, being figure-dense, threshold-free, and unstable, functions as another opinion: it may reassure or alarm, but a consumer cannot take it into a negotiation, and a contractor has no reason to concede to it. A deterministic figure traceable to a public database is different in kind, since it is checkable by both parties, an acquired instrument rather than a conferred condition. Whether such an instrument shifts market outcomes in the field, the question the conferred-verifiability experiments could not answer [3], becomes empirically askable once the instrument exists. This is why we frame the engine as an existence proof of a reference layer rather than as a better oracle: the economically relevant property is not that its numbers are narrower, but that they are contestable.

5.3. What the Existence Proof Does and Does Not Establish

The epistemic status of the structured engine is worth stating carefully. Its run-to-run consistency is a consequence of deterministic design and is therefore declared upfront, not presented as an empirical discovery. Its accuracy against realized project outcomes is not validated in this study; a consistent reference can still be a consistently imperfect one. What the existence proof establishes is narrower but decisive for the research agenda: the informational ingredients required for a verifiable consumer reference layer, namely an open cost database, explicit quantity assumptions, and reproducible computation, already exist and can be assembled today. Whether the resulting figures are right is the validation question we define as the next study (Section 7).

5.4. Reproducibility in a Moving Ecosystem

Two reproducibility hazards deserve emphasis because they are structural. First, hosted LLM APIs do not guarantee output reproducibility even under nominally deterministic settings [8], and our repeated-trial results quantify the consumer-facing consequence. Second, the models themselves retire: between the original experiments and the matched re-run reported here, one of the two originally tested models (GPT-4o) was deprecated by its provider, so that exact replication of a published LLM benchmark can become impossible within months. Both hazards argue for the design adopted here, namely provider-default matched settings, dated snapshots, and full release of questions, raw outputs, and scoring code. Both also underline, by contrast, the value of reference layers whose behavior does not drift silently.

6. Limitations

Four limitations bound our claims. (i) Accuracy is unvalidated. This study measures actionability properties of LLM answers and demonstrates the constructibility of a deterministic reference layer; it does not establish that the layer's point estimates match realized market prices. This is the principal limitation and the subject of the planned follow-up. (ii) Scope of the question set. The benchmark covers 40 consumer questions in Japanese, centered on common residential renovation categories; other trades, regions, and languages remain untested. (iii) Model coverage is a snapshot. We test frontier models available at the time of the experiments under default settings; results are dated by design, and the public pipeline exists precisely so that others can re-run the benchmark as models evolve. The current-Gemini re-run is a partial sample owing to API quota. (iv) Scoring granularity and construct. The span metric aggregates all man-yen figures in an answer, including itemized components, and its extraction convention, inherited unchanged from the original release for comparability, misses some compound notations (e.g., a lower bound written without the 万円 suffix); it therefore measures the dispersion of figures presented, not the narrowest range a model explicitly offers, and headline-range annotation is left to future work on the released outputs. The metrics also do not capture answer qualities such as empathy or explanation quality that demand-side studies elsewhere examine [10,11]. (v) The engine appears as a structural reference only. Its released outputs are curated consumer-facing text, not raw traces; for this reason, and given the author's interest in it, it is not tabulated as a comparator in the span results, and its internal computation is not part of the replication package.

7. Conclusion

We introduced, to our knowledge, the first consumer-question benchmark for construction costs, released in full, and used it to measure what a consumer actually receives when asking frontier LLMs about renovation prices: answers whose figures span multiples (median 5x to 15x by model), no decision thresholds in any configuration, and figures that vary materially across repeated runs. The latter two properties are independent of how the span is defined. A matched re-run with the current OpenAI flagship at bare defaults confirms the pattern is not an artifact of the original settings. We also demonstrated by construction that a deterministic, verifiable reference layer over an open cost database is feasible today, and situated its significance in credence-goods economics: verifiability, not another opinion, is what changes the consumer's position. A natural next step, and the subject of a planned follow-up study, is validating the engine's estimates against completed real-world renovation projects and accepted quotations, closing the accuracy question that a benchmark of question-answering behavior alone cannot address. The benchmark, outputs, and scoring pipeline are public; we invite the community to re-run, extend, and contest them.

Author Contributions: The single author (T.O.) is solely responsible for the conceptualization, methodology, software, formal analysis, data curation, visualization, and the writing of the original draft and its revision. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All benchmark questions, the original raw model outputs, the matched re-run raw outputs, the re-run harness, and the scoring scripts are openly available at Zenodo under CC BY 4.0 (concept DOI: 10.5281/zenodo.20522846, which always resolves to the latest version; the re-run materials were added in version 2). The underlying Japanese construction cost database (JCCDB) is published separately as open data.

Conflicts of Interest: The author develops and commercially operates the structured cost engine referenced in this study. To mitigate this conflict, the engine is not included as a comparator in the benchmark results (Section 4): it appears only as a structural reference and existence proof (Sections 4.3 and 5.3), its consistency is framed as a declared design property rather than a measured result, and the paper's primary contribution, the measurement of general-purpose LLMs, is independent of it. All LLM evaluation materials, raw outputs, and scoring code are public and independently re-runnable.

AI Use Disclosure: AI assistants were used for language editing and literature-search support under the author's full supervision; all claims, analyses, and final wording are the author's.

References

1. Darby, M. R.; Karni, E. Free Competition and the Optimal Amount of Fraud. J. Law Econ. 1973, 16, 67–88. doi:10.1086/466756.

2. Dulleck, U.; Kerschbamer, R. On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods. J. Econ. Lit. 2006, 44, 5–42. doi:10.1257/002205106776162717.

3. Dulleck, U.; Kerschbamer, R.; Sutter, M. The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition. Am. Econ. Rev. 2011, 101, 526–555. doi:10.1257/aer.101.2.526.

4. Balafoutas, L.; Beck, A.; Kerschbamer, R.; Sutter, M. What Drives Taxi Drivers? A Field Experiment on Fraud in a Market for Credence Goods. Rev. Econ. Stud. 2013, 80, 876–891. doi:10.1093/restud/rds049.

5. Balafoutas, L.; Kerschbamer, R. Credence Goods in the Literature: What the Past Fifteen Years Have Taught Us about Fraud, Incentives, and the Role of Institutions. J. Behav. Exp. Finance 2020, 26, 100285. doi:10.1016/j.jbef.2020.100285.

6. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans. Inf. Syst. 2025, 43, 42. doi:10.1145/3703155.

7. Farquhar, S.; Kossen, J.; Kuhn, L.; Gal, Y. Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature 2024, 630, 625–630. doi:10.1038/s41586-024-07421-0.

8. Atıl, B.; Aykent, S.; Chittams, A.; Fu, L.; Passonneau, R. J.; et al. Non-Determinism of "Deterministic" LLM System Settings in Hosted Environments. In Proc. 5th Workshop Eval4NLP 2025, 135–148. doi:10.18653/v1/2025.eval4nlp-1.12.

9. Renze, M. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In Findings of EMNLP 2024, 7346–7356. doi:10.18653/v1/2024.findings-emnlp.432.

10. Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; Zhu, Z.; et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern. Med. 2023, 183, 589–596. doi:10.1001/jamainternmed.2023.1838.

11. Goodman, R. S.; Patrinely, J. R.; Stone, C. A.; et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw. Open 2023, 6, e2336483. doi:10.1001/jamanetworkopen.2023.36483.

12. Dahl, M.; Magesh, V.; Suzgun, M.; Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. J. Legal Anal. 2024, 16, 64–93. doi:10.1093/jla/laae003.

13. Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; Ho, D. E. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. J. Empir. Legal Stud. 2025, 22, 216–242. doi:10.1111/jels.12413.

14. Niszczota, P.; Abbas, S. GPT Has Become Financially Literate: Insights from Financial Literacy Tests of GPT and a Preliminary Test of How People Use It as a Source of Advice. Finance Res. Lett. 2023, 58, 104333. doi:10.1016/j.frl.2023.104333.

15. Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In NeurIPS 2023 Datasets and Benchmarks. arXiv:2308.11462.

16. Kasai, J.; Kasai, Y.; Sakaguchi, K.; Yamada, Y.; Radev, D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv 2023, arXiv:2303.18027.

17. Elmousalami, H. H. Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review. J. Constr. Eng. Manage. 2020, 146, 03119008. doi:10.1061/(ASCE)CO.1943-7862.0001678.

18. Jafari, M.; Mousavi, E. A Critical Review of the Data-Driven Cost Estimation Approach in Construction Research. J. Constr. Eng. Manage. 2026, 152, 03125002. doi:10.1061/JCEMD4.COENG-16840.

19. Saka, A.; Taiwo, R.; Saka, N.; Salami, B. A.; Ajayi, S.; Akande, K.; Kazemi, H. GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation. Dev. Built Environ. 2024, 17, 100300. doi:10.1016/j.dibe.2023.100300.

20. Ghimire, P.; Kim, K.; Acharya, M. Opportunities and Challenges of Generative AI in Construction Industry: Focusing on Adoption of Text-Based Models. Buildings 2024, 14, 220. doi:10.3390/buildings14010220.

21. Prieto, S. A.; Mengiste, E. T.; García de Soto, B. Investigating the Use of ChatGPT for the Scheduling of Construction Projects. Buildings 2023, 13, 857. doi:10.3390/buildings13040857.

22. Zheng, J.; Fischer, M. Dynamic Prompt-Based Virtual Assistant Framework for BIM Information Search. Autom. Constr. 2023, 155, 105067. doi:10.1016/j.autcon.2023.105067.

23. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In Proc. ICLR 2023. arXiv:2302.09664.

24. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In Proc. ICLR 2023. arXiv:2203.11171.

25. Ghasemi, A.; Dai, F. Can ChatGPT Assist in Cost Analysis and Bid Pricing in Construction Estimating? A Pilot Study Using a Bridge Rehabilitation Project. Smart Constr. 2024, 1(2). doi:10.55092/sc20240009.

26. Ghimire, P.; Kim, K.; Stentz, T.; Roy, T. Modular Chain-of-Thought (CoT) for LLM-Based Conceptual Construction Cost Estimation. Buildings 2026, 16, 396. doi:10.3390/buildings16020396.

27. Wachter, S.; Mittelstadt, B.; Russell, C. Do Large Language Models Have a Legal Duty to Tell the Truth? R. Soc. Open Sci. 2024, 11, 240197. doi:10.1098/rsos.240197.