43rd International CIB W78 Conference
IT in Construction
New Delhi, India

Benchmarking LLMs for Professional Reasoning in Construction

CIRB: A GPQA-Inspired Benchmark for Interpretive Reasoning in Construction

Peter N. Gade[1],*, Jonas Goul Pedersen[1], Dorthe Holmberg Lauritzen[1], Judith Fauth[2]

[1] University College of Northern Denmark (UCN), Aalborg, Denmark
[2] Technical University of Munich, Munich, Germany
* Corresponding author: pega@ucn.dk

ABSTRACT

Large Language Models (LLMs) are increasingly applied to professional reasoning tasks in the Architecture, Engineering and Construction (AEC) industry, including building permit assessment, regulatory compliance checking, and design review. However, dominant benchmarks evaluate LLMs against factual recall or general technical knowledge, not against the interpretive, institutionally grounded reasoning that characterises professional construction practice. This study proposes the Construction Interpretive Reasoning Benchmark (CIRB), a transferable methodology adapted from the GPQA Diamond framework but reoriented towards interpretive construction domains, where correctness is shaped by authoritative practice rather than stable factual answers. The methodology transforms authoritative regulatory interpretations into four-option multiple-choice items with systematically designed distractors that target recurring professional reasoning failures: excessive literalism, contextual misalignment, overgeneralization, and omission of qualifying conditions. A two-stage validation protocol (independent expert review followed by non-expert filtering) is specified to isolate a high-difficulty Diamond subset that resists surface-level retrieval; the benchmark is demonstrated here on the unfiltered corpus. The approach is demonstrated on a Danish municipal building permit case using 143 regulatory questions paired with authoritative interpretations from the Municipality of Copenhagen. Six frontier LLMs (OpenAI GPT-5.5 and GPT-5.4, Claude Opus 4.7 and Sonnet 4.6, and Gemini 3.5 Flash and 2.5 Pro) were evaluated, achieving top-1 accuracies between 89.5% and 93.7%, alongside marked differences in confidence calibration (Expected Calibration Error from 0.054 to 0.206) and a roughly 50-fold range in API cost per question. The contribution is threefold: empirical evidence that, with accuracy saturating in the low-to-mid nineties, calibration and cost rather than accuracy become the operative criteria distinguishing these models for professional deployment; a reproducible instantiation grounded in municipal regulatory practice that treats authoritative interpretations as the reference standard; and a transferable validation protocol whose two-stage expert and non-expert filtering defines the path to a high-difficulty Diamond subset in future work.

Keywords: AI benchmarking, professional reasoning, building regulations, large language models, validation protocol, calibration, GPQA, BR18.

1. INTRODUCTION

Artificial Intelligence (AI) is increasingly embedded in the Architecture, Engineering and Construction (AEC) industry, supporting tasks from design assistance and quantity take-off to safety analysis, regulatory compliance and decision support (RICS, 2025). As Large Language Models (LLMs) move from experimental tools towards operational roles in professional workflows, how their performance should be evaluated becomes critical. Progress is typically demonstrated through general-purpose benchmarks such as MMLU and GPQA, which offer limited insight into whether AI systems can perform the context-sensitive, interpretive reasoning that characterises professional

construction practice. In building permit assessment, fire safety review, accessibility evaluation and contractual interpretation, correctness is not reducible to textual recall or binary rule satisfaction; it depends on institutional interpretation, professional judgment, contextual assumptions and alignment with established administrative practice (Gade et al., 2026). Erroneous or overconfident AI outputs in these contexts affect safety, legal accountability and trust in digital systems, not merely efficiency. Recent benchmarking work in the AEC field has begun to address model competence on domain tasks. AECBench evaluates LLMs across a five-level, cognition-oriented hierarchy of AEC tasks (Liang et al., 2026); AECV-Bench targets drawing literacy in multimodal models (Kondratenko et al., 2026); and CMExamSet benchmarks LLMs on construction management certification questions (Xiong et al., 2025). These studies measure performance against textual correctness or surface-level domain knowledge without addressing how professional judgments are formed in practice, and therefore give limited guidance on whether AI systems can be trusted where rules are interpreted, exceptions negotiated and decisions justified within institutional frameworks. The GPQA Diamond framework (Rein et al., 2024) demonstrates the value of expert-authored, retrieval-resistant questions for separating genuine reasoning from superficial fluency, but was developed for scientific domains with stable ground truth and does not address domains in which correctness is shaped by institutional interpretation.

This study argues that AI benchmarking for construction requires a distinct methodological approach, rather than evaluating models on generic accuracy or recall, benchmarks must assess whether AI systems align with authoritative, context-sensitive professional reasoning. We term this the Construction Interpretive Reasoning Benchmark (CIRB): a transferable methodology for constructing and evaluating interpretive items across AEC sub-domains, of which building permit assessment is the first instantiation. Permit assessment under the Danish Building Regulations (BR18) is used as the case, with 143 regulatory questions paired with authoritative interpretations issued by the Municipality of Copenhagen. Because these interpretations are authoritative reference points rather than purely textual solutions, benchmark accuracy is defined as alignment with accepted administrative practice. The study is guided by three research questions:

> RQ1. Can the CIRB methodology be designed and instantiated for an interpretive construction domain where correctness is defined by authoritative administrative practice rather than textual recall?
>
> RQ2. How do leading frontier LLMs perform on such a benchmark in terms of accuracy, and how reliably do their self-reported confidence values reflect that accuracy (calibration)?
>
> RQ3. What are the practical trade-offs (accuracy, calibration, and API cost) relevant to deploying these models in professional permit workflows?

The contribution is threefold: (i) it reframes AI evaluation in construction as a problem of professional and institutional alignment rather than technical correctness, and proposes a transferable methodology that adapts expert-validation principles to interpretive construction domains; (ii) it provides, to our knowledge, among the first empirical evidence in this benchmark setting that accuracy no longer separates frontier models on interpretive permit reasoning, whereas confidence calibration varies nearly four-fold (ECE 0.054-0.206) across otherwise comparable models; and (iii) it shows that API cost per question spans a roughly 50-fold range driven by reasoning-token volume rather than accuracy. Calibration and cost are therefore the operative, largely accuracy-independent criteria for selecting a model for professional deployment.

2. RELATED WORK

2.1 Saturation, reliability and contamination in general-purpose benchmarks

The Massive Multitask Language Understanding (MMLU) benchmark, spanning 57 subjects from elementary to professional level, was challenging at its introduction, when GPT-3 scored 43.9% (Hendrycks et al., 2021); by mid-2024 frontier models routinely reached approximately 88% in the standard five-shot setting and it is now widely considered saturated. An expert audit of about 5,700 MMLU questions found roughly 6.5% to contain outright errors, undermining the reliability of the score itself (Gema et al., 2025), and work on platinum benchmarks shows frontier LLMs still failing elementary tasks despite near-perfect aggregates (Vendrow et al., 2025). Saturation removes discriminative power without implying reliability, which is consequential for high-stakes deployment. Contamination compounds this: where evaluation questions appear in training corpora, reported performance is inflated. For benchmarks anchored in jurisdiction-specific regulatory interpretations the risk is substantially lower, since authoritative municipal interpretations are unlikely to form part of general training corpora.

2.2 Expert-level evaluation: GPQA, HLE and ProfBench

The GPQA (Graduate-Level Google-Proof Q&A) benchmark introduced a deliberate shift from retrieval to reasoning (Rein et al., 2024). It comprises 448 expert-authored multiple-choice questions in biology, physics and chemistry, the Diamond subset (198 questions) carrying the highest objectivity and inter-expert agreement. Questions must be Google-proof: skilled non-experts with unrestricted internet access and more than 30 minutes per question reach only 34% against 65% for PhD-level experts, and on Diamond the separation widens to 81.3% against 22.1%. The pipeline combines expert authoring, two-stage expert validation and non-expert filtering. Humanity's Last Exam extends the same logic to 2,500 questions (Phan et al., 2025), while ProfBench addresses rubric-graded professional deliverables in physics, chemistry, finance and consulting, where state-of-the-art models reach only around 66% (Wang et al., 2025). That retrieval-resistant items combined with multi-stage expert validation separate genuine reasoning from pattern matching is the principle this study transfers to interpretive construction practice.

2.3 AEC-specific benchmarks and automated compliance checking

Domain-specific benchmarks for AEC have emerged only recently. AECBench spans 23 tasks and 4,800 questions across a five-level, cognition-oriented framework and finds a clear decline from knowledge recall to reasoning, calculation and document generation across nine LLMs (Liang et al., 2026). AECV-Bench evaluates vision-language models on floor plans and document QA, reporting OCR-style tasks near 0.95 accuracy but symbol-centric drawing understanding at 0.40 to 0.55 (Kondratenko et al., 2026). CMExamSet benchmarks GPT-4o and Claude 3.7 on 689 construction management certification questions: both surpass the 70% pass threshold (82% and 83%) but drop sharply on multi-step and figure-referenced items (Xiong et al., 2025). Automated compliance checking (ACC) research instead transforms natural-language regulations into machine-readable rules; Fuchs et al. (2024) find complex nested and conditional clauses to remain a significant limitation. None of these evaluates LLMs against authoritative interpretive practice: CMExamSet measures certification knowledge, AECBench hierarchical domain knowledge, ACC research rule translation. Whether a model aligns with how a competent authority would actually decide a permit case is unaddressed.

2.4 Analogous evaluation in adjacent professional domains

Two adjacent fields provide methodological precedent. LegalBench, constructed with legal professionals across 162 tasks, finds that legal reasoning performance cannot be extrapolated from general scores (Guha et al., 2023), and subsequent work finds LLM legal interpretation unstable and sensitive to prompt reformulation (Purushothama et al., 2025). Medical AI evaluation has matured along similar lines: a systematic review of 39 medical LLM benchmarks, categorised as knowledge-based, practice-based or hybrid, reports a persistent gap between knowledge-based accuracy (84-90%) and

practice-based accuracy (45-69%) (Gong et al., 2025), and an analysis of 702 clinical evaluations of 85 models drawn from 168 studies finds only moderate correlation (Spearman's rho = 0.59) between benchmark scores and real-world clinical performance (Kim & Yoon, 2025). These findings reinforce the need to separate knowledge recall from professional reasoning when designing benchmarks for institutionally situated decision-making.

3. METHODOLOGY

This section describes the construction and evaluation of the BR18 benchmark, the first instantiation of CIRB, in the order in which the work was carried out: the source corpus and choice of case (3.1), the transformation of authoritative interpretations into multiple-choice items and the composition of the resulting item bank (3.2), the distractor design scheme (3.3), the two-stage validation protocol that yields the Diamond subset, specified as a transferable component of the methodology but not yet applied (3.4), and the model evaluation and calibration protocol (3.5). The underlying question-answer pairs are publicly available under a CC BY licence from Open Data DK (Københavns Kommune, 2024). The full benchmark instrument, comprising all 143 items with their four options, the designated correct option, the BR18 reference and the distractor design note, together with the Inspect AI evaluation logs, is available from the corresponding author on request.

3.1 Source corpus and choice of case

Building permit assessment is used as the primary case because it exemplifies a class of professional construction activities where AI evaluation is particularly demanding. It connects architectural design, engineering analysis, regulatory compliance and institutional decision-making within a single workflow, and permitting decisions directly affect project feasibility, legal responsibility and public safety. Prior analysis of the Danish Building Code has shown that regulatory provisions embed multiple co-existing logics, formal-legal, technical-functional, interpretive-practical and fuzzy, often within single clauses (Gade et al., 2026). Compliance judgments are therefore rarely derived through rule execution alone, and any framework assuming correctness to be reducible to textual accuracy will systematically misrepresent this reasoning. The empirical basis is a publicly available dataset of 143 regulatory question-answer pairs originating from building permit practice in the Municipality of Copenhagen (Københavns Kommune, 2024), used in full as the corpus for this study. Each answer reflects an authoritative administrative interpretation of BR18 under specific contextual assumptions. These interpretations are institutionally stabilised outcomes of professional regulatory reasoning rather than hypothetical or idealised solutions, and in the benchmark they are treated as the reference standard against which model performance is measured.

3.2 Item construction

Each question-answer pair is transformed into a four-option multiple-choice item. The correct option corresponds directly to the municipal interpretation and preserves its scope, conditions and constraints. Where the original answer relies on implicit premises, such as building category, height thresholds or use classification, these premises are made explicit in the question so that correctness depends on interpretation rather than guesswork. All items are recorded in a structured item bank capturing, for each question, an identifier, the regulatory theme, the BR18 reference, the question text, the four options, the designated correct option and a distractor design note. To remove answer-position bias the four options are shuffled deterministically using a fixed per-item seed before evaluation and the correct-option label is remapped accordingly; both orders are retained in the item metadata. Every item carries exactly four options, one correct and three distractors, so no item is answerable from format alone, and the correct key is close to uniformly distributed across the four positions in the item bank (A 42, B 33, C 32, D 36) before shuffling is applied. The 143 items span ten regulatory themes: general building code requirements (59 items), fire safety and evacuation (15), permits and administrative procedure (14), energy and thermal performance (11), water and plumbing (10), accessibility (6), acoustics (5), indoor climate and ventilation (3), structural safety (2), and a residual group of cross-cutting questions (18). All items are administered in Danish, the language of both the regulation and the

municipal interpretations; the English translations in this paper are for the reader only and are not part of the instrument.

3.3 Distractor design

Distractor quality is the primary lever that converts a multiple-choice test into a diagnostic instrument. The three incorrect options for each item are constructed as plausible but incorrect alternatives, following four categories derived from misinterpretations that municipal evaluators identify as characteristic of non-expert regulatory analysis and corroborated by the embedded-logic analysis of BR18 (Gade et al., 2026) and by distractor-design research in clinical benchmarking (Yang et al., 2025). Because each item carries exactly three distractors, no single item instantiates all four categories; the design note recorded for each item states which category each option realises, enabling later analysis of which failure modes a given model is most susceptible to. A known artefact of the format is that, because an authoritative interpretation usually carries its qualifying conditions with it, the correct option is often the longest, so option length is a potential surface cue. Measured across the item bank, the correct option is the longest of the four in 65% of items (56% on the 114 items whose stored option text is not truncated), against 25% under chance, so a length-only heuristic would score well above guessing but far below the 89.5% to 93.7% the models achieve. Normalising option length is therefore a defined task for the expert validation stage in Section 3.4, alongside the removal of items solvable without regulatory knowledge, and the cue is revisited as a limitation in Section 5. A worked item shows how the categories operate in a single question (theme: permits and administrative procedure; BR18 §47(2); administered in Danish, given here in English only).

The question asks whether a demolition permit from the municipality is required to demolish a garage that is integrated into a single-family house. The authoritative option states that a permit is required, because the exemption in §47(2) covers only garages and carports that are not integrated into the primary building, together with sheds, greenhouses, covered terraces and similar outbuildings erected in connection with the dwelling. The three distractors are: that no permit is required for demolishing a garage on a single-family property (context-shifted interpretation); that a permit is always required before any demolition work (overgeneralization); and that a permit is required, but without the integration condition on which the case turns (omission of qualifying conditions). The four categories, each illustrated with a characteristic BR18 example, are: literal-text interpretation (e.g., stating a single fixed door width without recognising that BR18 distinguishes general access doors, accessible routes, and operable installations); context-shifted interpretation (e.g., applying a residential exemption to a non-residential permit, or renovation rules to new-build); overgeneralization (e.g., asserting that demolition permits are always required, ignoring the exemption for non-integrated garages, carports and storage sheds); and omission of qualifying conditions (e.g., affirming an accessibility requirement without stating the dimensional limit, the building category, or accepted alternative solutions).

3.4 Planned validation and Diamond-subset selection

The full CIRB methodology specifies a two-stage validation process, inspired by the GPQA Diamond framework, to isolate a Diamond subset of items with the strongest expert agreement and the greatest resistance to non-expert solution (Rein et al., 2024): independent review by two domain experts in building regulation and permitting, with inter-rater agreement quantified by Cohen's kappa, followed by a non-expert filtering round that removes items solvable through superficial retrieval or phrasing artefacts. This validation has not yet been carried out. The results below therefore use the full 143-item corpus as issued by the Municipality of Copenhagen, treating the authoritative municipal interpretations directly as the reference standard, with Diamond-subset construction planned for subsequent work.

3.5 Model evaluation and calibration protocol

Models are evaluated using the Inspect AI evaluation framework. Each question is presented independently, including all contextual information and the fixed set of four shuffled options, and models select a single best answer without access to external tools. Temperature is set to zero in all runs; for reasoning models whose APIs do not honour a temperature parameter, provider defaults for the deployed reasoning mode are used, and the exact model identifiers, including provider snapshot

versions, API access dates and pricing date are recorded with the item metadata. Accuracy is the proportion of items for which the selected option matches the remapped reference key. Because every model answers the same items, pairwise accuracy differences are assessed with McNemar's test on the discordant pairs. Calibration is assessed in a separate run in which each model returns both an answer and a self-reported confidence value on a 0-100 scale, interpreted as its estimated probability of being correct; these values are grouped into bins and compared with the observed proportion correct. Calibration is summarised by Expected Calibration Error (ECE), the sample-weighted average absolute gap between confidence and accuracy across bins (Guo et al., 2017), and shown as reliability diagrams with Wilson confidence intervals on the per-bin accuracy (Wilson, 1927). API cost per question is computed from sample-level token usage in the Inspect logs using official list pricing.

4. RESULTS

Results are reported for accuracy (RQ2), confidence calibration (RQ2) and the accuracy-cost trade-off (RQ3). RQ1 is addressed by the 143-item instrument described in Section 3, which uses the municipal question-answer pairs directly as the reference standard pending the validation stage.

4.1 Benchmark accuracy

Six frontier LLMs were evaluated on the full 143-item benchmark: OpenAI GPT-5.5 and GPT-5.4, Anthropic Claude Opus 4.7 and Claude Sonnet 4.6, and Google Gemini 3.5 Flash and Gemini 2.5 Pro. All runs used identical prompts, identical shuffled item ordering, and temperature 0. Top-1 accuracy ranged from 89.5% to 93.7% (Fig. 1). Gemini 3.5 Flash was highest at 93.7%, followed by GPT-5.5 (93.0%), Gemini 2.5 Pro and Claude Opus 4.7 (both 92.3%), and GPT-5.4 (91.6%); Claude Sonnet 4.6 was lowest at 89.5%. Wilson 95% intervals overlap substantially across the field, so with 143 items the models are not cleanly separable on accuracy alone. Pairwise McNemar tests confirm this formally: no difference between any two of the six models reaches significance, and even the widest gap in the field, Gemini 3.5 Flash against Claude Sonnet 4.6, yields $p = 0.07$ on 8 discordant items. The headline observation is that the current generation of frontier models clusters tightly in the low-to-mid nineties on this interpretive permit benchmark, a range that confirms the benchmark is demanding while remaining within the reach of current frontier systems.

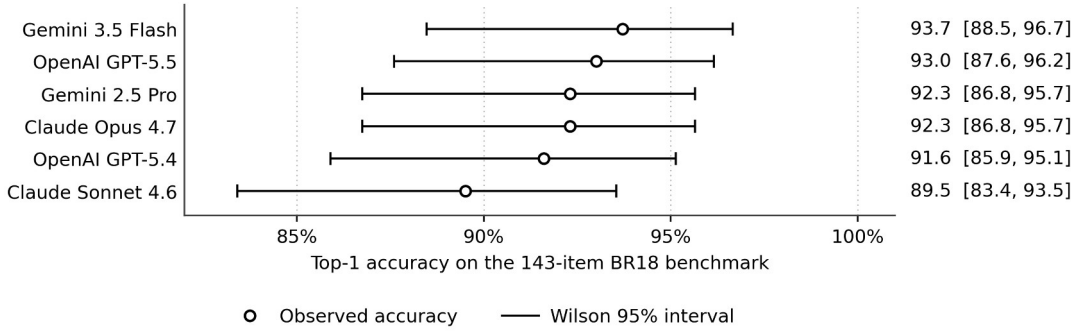

Figure 1: Top-1 accuracy on the 143-item BR18 benchmark with Wilson 95% confidence intervals.

Even the strongest model answers roughly one in sixteen permit-style questions incorrectly, and the lowest scoring model roughly one in nine, error rates that remain material in unsupervised regulatory practice. Aggregate accuracy alone therefore does not establish deployment readiness, which motivates the calibration and cost analyses that follow.

4.2 Calibration

Under the confidence-eliciting prompt, accuracy shifted slightly for several models relative to the plain benchmark, both upward and downward but within plausible error margins, consistent with the known interaction between prompt structure and model behaviour. Calibration quality, summarised by ECE in Table 1 and visualised in Fig. 2, varied substantially; with $n = 143$ and accuracies in the nineties, the ECE values are read here as an indicative ordering rather than as precise point estimates. The OpenAI models were well calibrated (GPT-5.4 ECE 0.054, GPT-5.5 ECE 0.059): their reliability

curves track the diagonal and their confidence is spread across a usable range. The Gemini models also achieved low ECE (2.5 Pro 0.058, 3.5 Flash 0.063) but concentrated almost all predictions in the upper confidence bins, which limits the practical informativeness of their confidence signals. The Claude models were the least well calibrated: Sonnet 4.6 (ECE 0.206) and Opus 4.7 (ECE 0.122) spread their confidence widely but were correct at or near 100% in those bins, indicating systematic under-confidence. For human-in-the-loop permit workflows, a model whose low-confidence flags reliably indicate likely errors is more valuable than a marginally more accurate model that signals high confidence uniformly.

Table 1: Accuracy, calibration and runtime summary across the six models (n = 143). ECE from the calibration run. Average tokens and seconds per question are combined means across the benchmark and calibration runs (n = 143 each).

Model	Benchmark acc. (%)	Calibration acc. (%)	ECE	Avg tokens/q	Avg sec/q
OpenAI GPT-5.5	93.0	93.7	0.059	528	8.86
Claude Opus 4.7	92.3	93.7	0.122	453	3.00
Gemini 3.5 Flash	93.7	93.0	0.063	1036	16.34
Claude Sonnet 4.6	89.5	90.9	0.206	368	2.85
OpenAI GPT-5.4	91.6	91.6	0.054	279	1.31
Gemini 2.5 Pro	92.3	92.3	0.058	1890	21.09

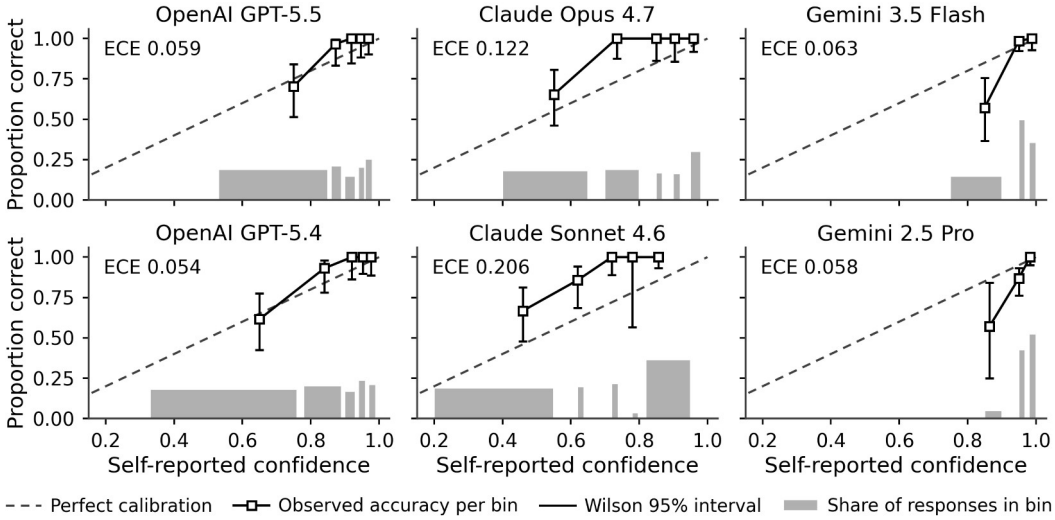

Figure 2: Reliability diagrams for the six models on the calibration run (n = 143). The dashed diagonal is perfect calibration, the marked line the observed proportion correct per confidence bin with Wilson 95% intervals, and the grey bars the share of responses in each bin (relative height, not on the accuracy scale). Points above the diagonal indicate under-confidence, below it over-confidence.

4.3 Accuracy, cost and efficiency trade-offs

Because the models cluster tightly on accuracy, cost and efficiency become primary selection criteria. Table 2 reports estimated API cost per question on the benchmark run. Cost per question spans a roughly 50-fold range, from about 0.0003 USD (GPT-5.4) to 0.0153 USD (Gemini 2.5 Pro), driven primarily by reasoning- and thinking-token volume rather than by accuracy. The spread is not an artefact of that single outlier: excluding Gemini 2.5 Pro, the range is still roughly 37-fold. The Pareto-efficient choice for cost-sensitive deployment is GPT-5.4, which delivers 91.6% accuracy at the lowest cost and fastest runtime. The most accurate model, Gemini 3.5 Flash (93.7%), emits the second-largest token volume and is in fact the second most expensive model at 0.0111 USD per question (nearly 40x GPT-5.4's cost), while Gemini 2.5 Pro imposes the highest cost despite mid-field accuracy (92.3%)

because it emitted very large numbers of billed thinking tokens. At the query volumes a municipality actually processes, these differences are economically significant and invisible to accuracy-only benchmarks.

Table 2: Estimated API cost on the 143-item benchmark run, using official provider list pricing. Gemini output cost includes billed thinking tokens.

Model	Acc. (%)	Total tokens	Total cost (USD)	Cost/q (USD)
OpenAI GPT-5.5	93.0	59,283	0.549	0.0038
Claude Opus 4.7	92.3	46,976	0.266	0.0019
Gemini 3.5 Flash	93.7	122,172	1.585	0.0111
Claude Sonnet 4.6	89.5	34,084	0.118	0.0008
OpenAI GPT-5.4	91.6	28,250	0.042	0.0003
Gemini 2.5 Pro	92.3	244,129	2.186	0.0153

4.4 Failure pattern analysis

Item-level inspection of incorrect responses reveals two recurring patterns. The first is omission of qualifying conditions: items whose correct interpretation includes a dimensional limit, a building-category condition or an alternative-solution clause are answered correctly in outline, but the chosen distractor drops the qualifier. The second is excessive literalism: items where the regulatory text, read in isolation, would support a stricter or more general answer than the municipal interpretation. Both align with the legal and medical literature, in which LLMs favour textually salient interpretations over institutionally authoritative ones, and with the embedded-logic analysis of BR18 (Gade et al., 2026). Errors are also strongly concentrated: 120 of the 143 items (84%) are answered correctly by all six models, while just 10 items account for 51 of the 68 total error instances (75%), and 4 items defeat every model. That most of the corpus is solved by the entire field while a small shared subset drives most failures is direct evidence for the value of the Diamond-subset filtering specified in Section 3.4. Frequencies for the two remaining categories, contextual misalignment and overgeneralization, are not reported here. A complete distractor-category analysis, and one across a Diamond subset once validation is complete, is planned as subsequent work.

5. DISCUSSION AND CONCLUSIONS

The results show that CIRB can be designed and instantiated for an interpretive construction domain using institutionally grounded ground truth, even prior to the full expert-validation stage, and that frontier LLMs achieve high accuracy on it while differing markedly in calibration and cost. Three implications follow, each pointing to a change in how construction-AI benchmarks should be reported. First, accuracy has become a weak discriminator at the frontier: with six models within four percentage points, accuracy-only leaderboards no longer separate systems meaningfully, echoing the saturation seen for MMLU and GPQA. Evaluation intended to inform deployment should treat calibration and cost as first-class metrics rather than annexes to an accuracy score. Second, calibration carries operational meaning: a model whose low confidence reliably flags likely error supports efficient human review, whereas a uniformly confident model offers no triage signal. Under-confidence and over-confidence are both failures, but an under-confident system invites unnecessary review while an over-confident one risks unsupervised acceptance of incorrect outputs. That systems of near-identical accuracy differ this sharply in fitness for human-in-the-loop use is precisely what an accuracy-only benchmark conceals. Third, cost is a selection criterion in its own right: a roughly 50-fold spread per question, driven by thinking-token volume rather than accuracy, is decisive at the query volumes a municipality actually processes. Several limitations apply.

The benchmark currently covers a single jurisdiction (Denmark) and is administered in Danish, which restricts generalisation to other regulatory regimes; the corpus is publicly available and no contamination check has been performed, so training-data exposure cannot be ruled out; Diamond-

subset filtering is specified but not yet applied; the corpus has not yet been normalised for option length, and the correct option is the longest in 65% of items against 25% under chance, so part of the measured accuracy may reflect a surface cue rather than regulatory reasoning; calibration relies on self-reported confidence rather than token probabilities; and multiple-choice evaluation cannot fully capture the open-ended reasoning that defines genuine permit assessment. The methodology should therefore be viewed as a rigorous instrument for measuring a necessary precondition of professional competence, not a sufficient one. Subject to these caveats, the construction, validation and evaluation steps are designed to be portable, and applying them to a further interpretive sub-domain would test the transferability the methodology is intended to provide. Planned work will complete the two-stage expert and non-expert validation to release a BR18 Diamond subset, conduct a full distractor-category analysis of model failure modes, and apply the methodology to a second interpretive sub-domain (fire-safety performance review) to test cross-domain transferability. Taken together, these results point toward AI benchmarking that is honest about the interpretive nature of construction practice, reports calibration and cost as first-class metrics alongside accuracy, and is directly informative for integrating LLMs into professional permit workflows.

REFERENCES

Fuchs, S., Witbrock, M., Dimyadi, J., & Amor, R., (2024). Using large language models for the interpretation of building regulations. Journal of Engineering, Project, and Production Management, 14(4): p. 1-12. DOI: 10.32738/JEPPM-2024-0035

Gade, P. N., Fauth, J., Muneeb, M., & Holmberg Lauritzen, D., (2026). Beyond binary compliance: Mapping hybrid rule logics for logic-aware digitalisation of building codes. In Fischer, S., Fauth, J., Schranz, C., & Urban, H. (eds.), Proceedings of the Digital Building Permit Conference 2025 (DBP25), Lecture Notes in Civil Engineering. Vienna, Austria, December 2-4. Cham: Springer, [Online], Available: https://link.springer.com/book/9783032182258 [7 September 2026].

Gema, A. P., et al., (2025). Are we done with MMLU? In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2025), Volume 1: Long Papers. Albuquerque, NM, USA, April 29 - May 4: p. 5069-5096. DOI: 10.18653/v1/2025.naacl-long.262

Gong, E. J., Bang, C. S., Lee, J. J., & Baik, G. H., (2025). Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. Journal of Medical Internet Research, 27: e84120. DOI: 10.2196/84120

Guha, N., et al., (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), Datasets and Benchmarks Track. New Orleans, LA, USA, December 10-16: p. 44123-44279.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q., (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017). Sydney, Australia, August 6-11: p. 1321-1330.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J., (2021). Measuring massive multitask language understanding. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021). Virtual, May 3-7.

Kim, S. & Yoon, H.-J., (2025). Questioning our questions: How well do medical QA benchmarks evaluate clinical capabilities of language models? In Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025). Vienna, Austria, August 1: p. 274-296. DOI: 10.18653/v1/2025.bionlp-1.24

Københavns Kommune, (2024). Spørgsmål/svar til Bygningsreglementet og tilhørende vejledninger [Dataset]. Copenhagen: Open Data DK, licence CC BY, [Online], Available: https://www.opendata.dk/city-of-copenhagen/eksempel-spm-svar-bygningsreglementet [25 August 2026].

Kondratenko, A., Birhane, M., Hsain, H. E., & Maciocci, G., (2026). AECV-Bench: Benchmarking multimodal models on architectural and engineering drawings understanding. arXiv preprint arXiv:2601.04819.

Liang, C., et al., (2026). AECBench: A hierarchical benchmark for knowledge evaluation of large language models in the AEC field. Advanced Engineering Informatics, 71: 104314. DOI: 10.1016/j.aei.2026.104314

Phan, L., et al., (2025). Humanity's last exam. arXiv preprint arXiv:2501.14249.

Purushothama, A., Min, J., Waldon, B., & Schneider, N., (2025). Not ready for the bench: LLM legal interpretation is unstable and uncalibrated to human judgments. In Proceedings of the Natural Legal Language Processing Workshop 2025 (NLLP 2025). Suzhou, China, November 8: p. 317. DOI: 10.18653/v1/2025.nllp-1.22

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R., (2024). GPQA: A graduate-level Google-proof Q&A benchmark. In Proceedings of the First Conference on Language Modeling (COLM 2024). Philadelphia, PA, USA, October 7-9.

RICS, (2025). RICS Artificial Intelligence in Construction Report 2025. London: Royal Institution of Chartered Surveyors, [Online], Available: https://www.rics.org/news-insights/artificial-intelligence-in-construction-report [25 August 2026].

Vendrow, J., Vendrow, E., Beery, S., & Mądry, A., (2025). Do large language model benchmarks test reliability? arXiv preprint arXiv:2502.03461.

Wang, Z., et al., (2025). ProfBench: Multi-domain rubrics requiring professional knowledge to answer and judge. arXiv preprint arXiv:2510.18941.

Wilson, E. B., (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158): p. 209-212.

Xiong, R., Wang, Y., Gunhan, S., Zhu, Y., & Berryman, C., (2025). Can AI master construction management (CM)? Benchmarking state-of-the-art large language models on CM certification exams. arXiv preprint arXiv:2504.08779.

Yang, R., Deng, W., Chen, M., Zhou, Y., & Li, X., (2025). Enhancing clinical multiple-choice questions benchmarks with knowledge graph guided distractor generation. arXiv preprint arXiv:2506.00612.