Ishigaki-IDS-Bench: A Benchmark for Generating Information Delivery Specification from BIM Information Requirements

Ryo Kanazawa
ONESTRUCTION Inc.
Tottori, Japan
r.kanazawa@onestruction.com

Koyo Hidaka
ONESTRUCTION Inc.
Tottori, Japan
k.hidaka@onestruction.com

Teppei Miyamoto
ONESTRUCTION Inc.
Tottori, Japan
t.miyamoto@onestruction.com

Takayuki Kato
ONESTRUCTION Inc.
Tottori, Japan
t.kato@onestruction.com

Tomoki Ando
ONESTRUCTION Inc.
Tottori, Japan
t.ando@onestruction.com

Chenguang Wang
AWS GenAI Innovation Center
Tokyo, Japan
wangcg@amazon.co.jp

Dayuan Jiang
AWS GenAI Innovation Center
Tokyo, Japan
jiangdy@amazon.co.jp

Naofumi Fujita
ONESTRUCTION Inc.
Tottori, Japan
n.fujita@onestruction.com

Shuhei Saitoh
ONESTRUCTION Inc.
Tottori, Japan
s.saito@onestruction.com

Atomu Kondo
ONESTRUCTION Inc.
Tottori, Japan
a.kondo@onestruction.com

Koki Arakawa
ONESTRUCTION Inc.
Tottori, Japan
k.arakawa@onestruction.com

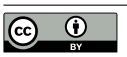
Daiho Nishioka
ONESTRUCTION Inc.
Tottori, Japan
d.nishioka@onestruction.com

Abstract

Building Information Modeling (BIM) projects increasingly use Information Delivery Specification (IDS) to formalize information requirements in a machine-checkable XML format. Because IDS conditions are grounded in the Industry Foundation Classes (IFC) vocabulary, authoring them requires expertise in IFC concepts, validation tools, and property set conventions. Existing benchmarks for structured generation do not adequately capture the additional burden of vocabulary conformance and external-validator agreement that IDS imposes. We present Ishigaki-IDS-Bench, the first publicly released benchmark for IDS generation from BIM information requirements. The benchmark contains 166 examples spanning 83 practical scenarios authored in Japanese and English by six BIM/IDS experts, each paired with a gold IDS file and metadata covering input format, turn setting, target IFC versions, and construction domain. Evaluation proceeds in two stages: (i) formal validity scored by the buildingSMART IDSAuditTool along Processability, Structure, and Content, and (ii) content fidelity scored by facet-level macro-F1 against the gold IDS. Across 10 LLMs in zero-shot, the highest Facet F1 is 65.6%, achieved by GPT-5.5, while the highest Content pass rate is only 33.1%, achieved by Claude Opus 4.5. Ishigaki-IDS-Bench is released on Hugging Face (DOI 10.57967/hf/8873) under CC BY 4.0, and the evaluation code is released on Zenodo (DOI 10.5281/zenodo.20550510) under Apache-2.0.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → *Data management systems.*



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840206

Keywords

benchmark datasets, large language models, structured generation, XML generation, Building Information Modeling, Information Delivery Specification, Industry Foundation Classes

ACM Reference Format:
Ryo Kanazawa, Koyo Hidaka, Teppei Miyamoto, Takayuki Kato, Tomoki Ando, Chenguang Wang, Dayuan Jiang, Naofumi Fujita, Shuhei Saitoh, Atomu Kondo, Koki Arakawa, and Daiho Nishioka. 2026. Ishigaki-IDS-Bench: A Benchmark for Generating Information Delivery Specification from BIM Information Requirements. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3799682.3840206

1 Introduction

Large language models (LLMs) are widely used to generate structured outputs such as JSON, SQL, and code[1, 29]. However, in practical domains, structured outputs are not sufficient simply because they are syntactically valid. They must simultaneously conform to industry-standard data formats, domain-specific vocabularies, version constraints, and audit results from external validation tools. Evaluation of structured generation that conforms to such domain standards remains less developed than evaluation of general-purpose JSON or SQL generation. This study addresses this problem using Information Delivery Specification (IDS), which describes information requirements in the architecture and construction domain in a machine-readable form.

Building Information Modeling (BIM) is an information foundation for handling building and infrastructure component types, materials, performance, and management information[6], and Industry Foundation Classes (IFC) is an international standard data format for sharing BIM models across different software systems[11]. Information Delivery Specification (IDS) is a standard specification that describes in XML "which elements should have which information under which conditions" for IFC models[3]. For example, a requirement such as "all walls shall have a fire-resistance rating,

whose value is one of EI30, EI60, and EI90" must map the IFC class for walls, the information item for fire-resistance rating, and the allowed-value constraint to IDS inspection units (facets).

In practice, information requirements are often written as natural-language specifications, tabular checklists, client requirements, meeting records, and similar documents. Creating IDS therefore requires expertise in BIM, IFC, and IDS, as well as the ability to interpret input documents. LLMs may support this transformation, but there is a lack of public benchmarks that jointly evaluate the formal validity of generated IDS, conformance to the IDS standard, consistency with IFC vocabulary, and content agreement with the input document.

We therefore propose Ishigaki-IDS-Bench, a benchmark for evaluating the ability to generate IDS 1.0-compliant XML from practical documents. The benchmark contains 166 examples obtained by expanding 83 practical scenarios into Japanese and English, corresponding gold IDS, and metadata for input format, language, turn setting, target IFC versions, and construction domain.

In addition, we design a two-stage protocol for evaluating generated IDS from both formal and content perspectives. In the first stage, IDSAuditTool[2] verifies IDS extractability, schema compliance, and conformance to the IDS standard and IFC vocabulary constraints. In the second stage, content agreement with the gold IDS is measured to capture generations that are formally valid but differ from the input document.

The contributions of this work are as follows.

- We construct an IDS generation benchmark based on practical use cases. Ishigaki-IDS-Bench contains 166 expert-created and verified examples, 83 scenarios, corresponding gold IDS, and multifaceted metadata.
- We provide a two-stage evaluation protocol that combines formal-validity evaluation using IDSAuditTool with content-agreement evaluation against gold IDS.
- We report zero-shot baselines for 10 LLMs and analyze successful cases and failure tendencies in IDS generation.

2 Related Work

Schema-constrained and grammar-constrained generation have been developed as methods for improving the formal validity of structured outputs generated by LLMs. Representative examples include methods that incorporate FSMs or CFGs into decoding[10, 29], XGrammar, which co-designs the inference engine and grammar engine[5], Grammar-Aligned Decoding, which corrects distributional distortion caused by grammar constraints[23], and methods that use grammar masking for DSL generation[20]. These studies have mainly targeted outputs with general-purpose schemas or explicit syntactic constraints, such as JSON, SQL, code, and DSLs. In contrast, domain-specific XML standards such as IDS require not only syntactic validity but also consistency with IFC vocabulary, external standards, version constraints, property set conventions, and the judgments of validation tools. IDS generation should therefore be evaluated not merely as syntax-constrained generation, but as domain-standard-compliant structured generation.

For LLM evaluation in specialized domains, benchmarks have increasingly been constructed for areas that require expert knowledge, such as LawBench[7] and LeDQA[17] in the legal domain, EDINET-BENCH[25] in the financial domain, and ECKGBench[18] in the e-commerce domain. In the architecture and construction domain as well, studies have addressed building-code interpretation[8], BIM compliance checking[4, 19], construction safety datasets[22], BIM-GPT for applying LLMs to BIM information retrieval[30], Qwen-BIM specialized for BIM design tasks[16], IFC-Agent for schema-guided multi-agent reasoning over IFC[9], MCP4IFC for editing IFC through code generation[21], IFC reasoning and editing[28], and IDS/mvdXML standardization[14, 15, 24, 26, 27]. However, most of these studies focus on question answering, retrieval, building-code interpretation, compliance judgment, BIM information retrieval, IFC model understanding, design support, and model editing. There is no public benchmark that directly generates IDS from BIM information requirements and integrates input documents, gold IDS, external-validator audits, and facet-level content-agreement evaluation. Ishigaki-IDS-Bench targets domain-standard XML generation that can be verified using an external validator and gold IDS, and is positioned at the intersection of schema-constrained generation and BIM/IFC-domain evaluation.

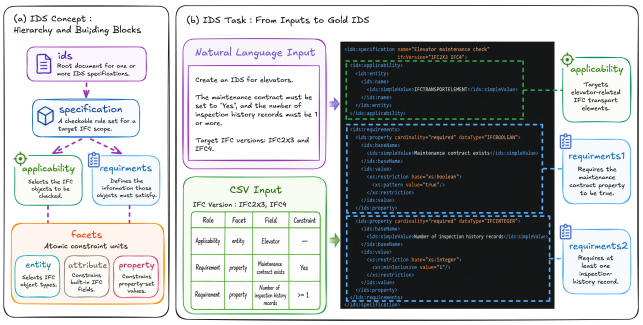


Figure 1: IDS hierarchy and task definition in Ishigaki-IDS-Bench. The left panel shows the main units that constitute IDS, and the right panel shows the correspondence used to create gold IDS from natural-language or CSV inputs.

3 Ishigaki-IDS-Bench

Task and scope. Ishigaki-IDS-Bench targets the task of generating complete IDS files that conform to IDS 1.0 from information requirements. Each input includes a requirement written in CSV or natural language, an output file name, and one or more target IFC versions. The model outputs the corresponding IDS using only the requirements explicitly stated in the input.

Figure 1 shows the IDS structure and representative input–output pairs. IDS consists of one or more `specification` elements, each with `applicability` for the target and `requirements` for required conditions. These conditions are represented as facets. We evaluate `entity`, `attribute`, and `property` facets, which represent IFC element types, basic IFC-schema fields, and practical additional information. Thus, the task is not XML formatting but a structured generation task that maps practical information requirements to IDS target and requirement conditions. Because this study focuses on target specification and information-item specification as

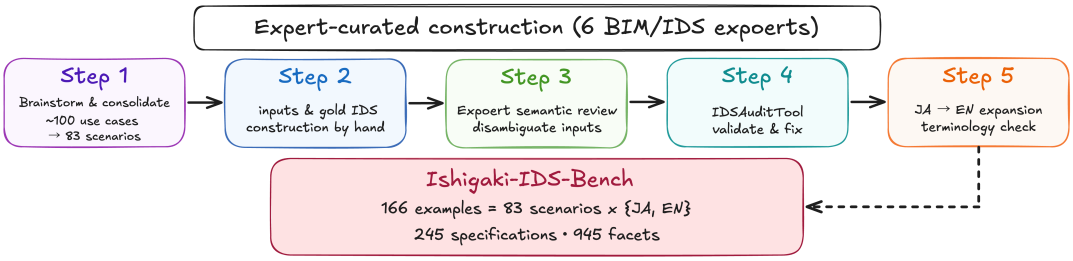


Figure 2: Construction pipeline of Ishigaki-IDS-Bench. Six BIM/IDS experts consolidated approximately 100 candidates into 83 scenarios, conducted gold IDS authoring, semantic review, IDSAuditTool validation, and JA/EN expansion. The artifacts include 166 examples, 245 specifications, and 945 facets.

Table 1: Metadata distribution over 166 Ishigaki-IDS-Bench examples; IFC-version counts exceed 166 because an example may target multiple versions.

Axis	Distribution
Input format	CSV=58, Natural language=108
Language	JA=83, EN=83
Turn setting	Single-turn=122, Multi-turn=44
Target IFC versions	IFC2X3=78, IFC4=54, IFC4X3=70
Domain	Architecture=48, Structural=42, MEP=40, General=36

the core of practical information requirements, `classification`, `material`, and `partOf` are excluded from evaluation.

Taxonomy and statistics. Table 1 summarizes the five-axis taxonomy over 166 examples. The benchmark expands 83 scenarios into Japanese and English examples. An example is one input context paired with one gold IDS. For the 44 multi-turn examples, the model receives a fixed history containing the initial user request, dataset-provided assistant IDS drafts, and subsequent user revision requests. The model is required to produce a complete IDS reflecting the final revision request, while the final assistant IDS is withheld as the gold reference. A scenario is the base use case that groups corresponding examples representing the same information requirement. Each example is annotated with input format, language, turn setting, target IFC versions, and construction domain, enabling analysis of performance differences between natural-language and CSV inputs, Japanese and English inputs, single-turn and multi-turn settings, target IFC versions, and construction domains.

Dataset construction. Because public IDS examples are limited, we prioritized expert-judged practical representativeness over mechanical IFC-schema coverage. Six BIM/IDS experts, five with buildingSMART-related experience and five with IFC/IDS practice, consolidated about 100 candidates into 83 scenarios and manually authored input information requirements and gold IDS through the procedure shown in Figure 2. Quality was checked through semantic review, IDSAuditTool validation, facet-scorer reruns, and JA/EN terminology checks; two additional ML experts reviewed the task definition, metadata, evaluation procedure, and reproducibility. Because the gold IDS files were created through expert authoring and standards-compliance validation rather than independent parallel annotation by multiple annotators, pairwise IAA was not computed. The scale of 166 examples results from prioritizing expert-validated, standards-compliant samples over synthetic augmentation.

4 Evaluation Protocol

Evaluation on Ishigaki-IDS-Bench consists of two stages: formal validity evaluation using IDSAuditTool (ids-tool v1.0.96) and facet-level content-matching evaluation against gold IDS. In the first stage, Processability reports whether an IDS candidate can be supplied to the audit tool, Structure reports XML and IDS XSD validity, and Content reports compliance with the IDS standard and IFC vocabulary constraints. In the second stage, we use a facet scorer that compares gold IDS and generated IDS. The comparison targets are IFC-version labels, entity, attribute, property, and the dataType/cardinality of property; exact matching is computed over a multiset of comparison blocks consisting of applicability/requirements, facet type, and constraint content. For IDS files containing multiple specifications, specifications in the gold IDS and generated IDS are aligned, and only the comparison representation is canonicalized for XML namespaces, constraint child elements, and simpleValue ordering. Precision, recall, and F1 are macro-averaged over examples, and LLM-as-judge is not used because direct comparison against gold IDS is possible. Because practical requirements may omit property set names, we consistently apply the non-IDS-standard internal convention explicitly stated in the zero-shot prompt to all models, and publish the corresponding slice and convention together with the evaluation scripts.

5 Baseline Evaluation

Experimental setup. We evaluated the 10 LLMs in Table 2 in a zero-shot setting in May 2026. All models receive the same system prompt and the example-specific context containing the information requirements and target IFC versions. No separate single-turn or multi-turn examples, IDS outputs, or gold IDS are added as demonstrations. Prior IDS drafts in a multi-turn example are part of that example's fixed context, not demonstrations. If no IDS candidate could be extracted from the output or XML parsing failed, the output was treated as invalid. We did not perform multiple-sample selection, reranking, post-generation repair, regeneration, or manual correction.

Table 2: Zero-shot baseline results on Ishigaki-IDS-Bench. Model groups are divided based on whether their weights are publicly available. Processability, Structure, and Content denote first-stage audit rates, and Facet F1, Recall, and Precision denote second-stage macro averages over 166 examples.

	Stage 1			Stage 2 (Facet scorer)		
Model	Processability	Structure	Content	Facet F1	Recall	Precision
Closed proprietary LLMs						
GPT-5.5 (xhigh)	94.6	30.1	27.7	**65.6**	**65.8**	**65.6**
Claude Opus 4.5	**99.4**	**45.8**	**33.1**	42.2	45.5	40.5
Gemini 3.1 Pro	97.0	37.3	25.9	52.2	52.6	52.2
Open-weight LLMs						
gpt-oss-120b (xhigh)	98.8	13.9	12.0	21.6	21.4	21.9
Kimi K2.6	**99.4**	41.0	22.3	48.4	48.9	48.4
DeepSeek V4 Pro	97.0	19.3	16.3	40.0	36.3	49.1
Qwen3.5-397B-A17B	98.2	20.5	19.3	26.8	26.2	28.5
Qwen3-32B	37.3	13.9	10.2	22.3	21.7	23.5
Qwen3-14B	96.4	18.1	13.9	21.3	21.2	21.6
Qwen3-8B	26.5	20.5	15.7	20.2	20.0	20.5

Note. Decoding settings were fixed for each model family before evaluation and were not tuned on Ishigaki-IDS-Bench. Qwen-family models were evaluated with `temperature=0.6`, `top_p=0.95`, `top_k=20`, `min_p=0.0`, and other models with `temperature=0.0`, `top_p=1.0`; the output length limit for all models was 15,000 tokens. The `(xhigh)` notation indicates a setting in which reasoning effort was fixed to `xhigh`. Truncation, structure-constrained decoding, repair, and regeneration were not used.

Results and findings. Table 2 shows the results. Even the model with the highest Facet F1 score, GPT-5.5, reaches only 65.6% Facet F1 and a 27.7% Content pass rate, indicating that XML generation that complies with the IDS standard and IFC vocabulary constraints remains difficult even when some input requirements are recovered as facets. High Processability with low Content is common, exposing a gap between audit-tool processability and standards-compliant content generation.

In Stage 2, GPT-5.5 and Gemini 3.1 Pro lead facet-level agreement, while Kimi K2.6 lags in Content and Facet F1 despite high Processability. Within the Qwen family, larger models show broadly higher Facet F1 but non-monotonic audit pass rates, indicating that scale alone does not ensure robust IDS processability or standards-compliant content generation.

For GPT-5.5, the 44 multi-turn examples reached Content=79.5% and Facet F1=84.8%, compared with 9.0% and 58.6% on the 122 single-turn examples. This pattern suggests that history-conditioned IDS revision may be more tractable than generation from scratch in this benchmark. CSV inputs achieved higher Facet F1 than natural-language inputs (77.3% vs. 59.2%).

Qualitatively, we observed three failure tendencies. First, many outputs can be extracted and audited as IDS candidates but still violate the IDS XSD, IDS standard, or IFC vocabulary constraints, appearing as lower Structure or Content scores. Second, `entity` and `attribute` facets are relatively easy to recover, whereas errors are common in `property` facets involving property set names, cardinality, and value constraints. Third, some natural-language errors involved ambiguous mappings between requirement targets and value constraints, leading to facet omissions or overgeneration.

6 Availability, Ethics, and Maintenance

The dataset and gold IDS, with per-example IDs and five-axis metadata, are released on Hugging Face under CC BY 4.0 (dataset DOI 10.57967/hf/8873) [12]; the evaluation scripts and zero-shot prompt are released on GitHub/Zenodo under Apache-2.0 (fixed-release DOI 10.5281/zenodo.20550510) [13]. The dataset excludes personal information, real IFC files, confidential project data, and copyrighted real-project documents; releases will be versioned and maintained on Hugging Face, Zenodo, and GitHub.

7 Limitations

Ishigaki-IDS-Bench has scope limitations. This study focuses on the `entity`, `attribute`, and `property` facets, and leaves `classification`, `material`, and `partOf` for future extension. Although the inputs are based on practical use cases, end-to-end performance on noisy real document collections is not directly measured because confidential documents and real IFC models are not redistributed. The handling of unspecified property sets is an internal evaluation convention rather than part of the IDS standard. The 166 examples are intended for diagnosis rather than large-scale training, and few-shot prompting, retrieval-augmented prompting, fine-tuning, and grammar-constrained decoding remain subjects for future comparison.

8 Conclusion

Ishigaki-IDS-Bench is an expert-created and verified public benchmark for IDS 1.0-compliant XML generation from BIM information requirements. Across 10 zero-shot baselines, the highest Facet F1 (65.6%) and the highest Content pass rate (33.1%) were achieved by different models, showing that IDS generation requires separate evaluation of domain vocabulary, standards compliance, and facet-level content consistency, not only syntactic well-formedness.

Acknowledgments

This work was conducted as part of the GENIAC (Generative AI Accelerator Challenge) Project, which aims to strengthen Japan's capability to develop generative AI and is promoted by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO).

GenAI Usage Disclosure

Generative AI tools were used as part of the research subject and experimental procedure. Specifically, zero-shot generation results were obtained from each LLM and used as baseline evaluation targets on Ishigaki-IDS-Bench. The input examples, metadata, gold IDS, evaluation design, metric computation, and result interpretation for Ishigaki-IDS-Bench were created and checked by the authors and experts. LLMs were not used to construct the benchmark examples or gold IDS files. During manuscript writing, generative AI tools including ChatGPT and DeepL were used for translation, language polishing, terminology consistency checks, and LaTeX formatting assistance. The authors checked and revised the outputs and take responsibility for all claims, experimental design, analysis, conclusions, and the final manuscript.

References

[1] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2024. Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 3658–3673. https://proceedings.mlr.press/v235/beurer-kellner24a.html

[2] buildingSMART International. 2024. IDS Audit Tool. https://github.com/buildingSMART/IDS-Audit-tool

[3] buildingSMART International. 2024. Information Delivery Specification (IDS). https://www.buildingsmart.org/standards/bsi-standards/information-delivery-specification-ids/

[4] Nanjiang Chen, Xuhui Lin, Hai Jiang, and Yi An. 2024. Automated Building Information Modeling Compliance Check through a Large Language Model Combined with Deep Learning and Ontology. *Buildings* 14, 7 (2024), 1983. doi:10.3390/buildings14071983

[5] Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. 2025. XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models. arXiv:2411.15100 [cs.CL] doi:10.48550/arXiv.2411.15100

[6] Chuck Eastman, Paul Teicholz, Rafael Sacks, and Kathleen Liston. 2008. *BIM Handbook: A Guide to Building Information Modeling for Owners, Managers, Designers, Engineers and Contractors*. John Wiley & Sons.

[7] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. arXiv:2309.16289 [cs.CL] doi:10.48550/arXiv.2309.16289

[8] Stefan Fuchs, Michael Witbrock, Johannes Dimyadi, and Robert Amor. 2024. Using Large Language Models for the Interpretation of Building Regulations. arXiv:2407.21060 [cs.CL] doi:10.48550/arXiv.2407.21060

[9] Yan Gao, Fuji Hu, Chengzhang Chai, Yiwei Weng, and Haijiang Li. 2026. Multi-agent Framework for Schema-guided Reasoning and Tool-augmented Interaction with IFC Models. *Automation in Construction* 186 (2026), 106888. doi:10.1016/j.autcon.2026.106888

[10] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 10932–10952. doi:10.18653/v1/2023.emnlp-main.674

[11] International Organization for Standardization. 2024. ISO 16739-1:2024: Industry Foundation Classes (IFC) for Data Sharing in the Construction and Facility Management Industries—Part 1: Data Schema. https://www.iso.org/standard/84123.html

[12] Ryo Kanazawa, Koyo Hidaka, Teppei Miyamoto, Takayuki Kato, Tomoki Ando, Chenguang Wang, Dayuan Jiang, Naofumi Fujita, Shuhei Saitoh, Atomu Kondo, Koki Arakawa, and Daiho Nishioka. 2026. Ishigaki-IDS-Bench. doi:10.57967/hf/8873 Hugging Face dataset. Accessed: 2026-05-21.

[13] Ryo Kanazawa, Koyo Hidaka, Teppei Miyamoto, Takayuki Kato, Tomoki Ando, Chenguang Wang, Dayuan Jiang, Naofumi Fujita, Shuhei Saitoh, Atomu Kondo, Koki Arakawa, and Daiho Nishioka. 2026. Ishigaki-IDS-Bench: Evaluation Code and Reproducibility Repository. doi:10.5281/zenodo.20550510 GitHub evaluation-code release v1.0.1. Apache License 2.0. Accessed: 2026-06-05.

[14] Magdalena Kładź and Andrzej Szymon Borkowski. 2025. IDS Standard and bSDD Service as Tools for Automating Information Exchange and Verification in Projects Implemented in the BIM Methodology. *Buildings* 15, 3 (2025), 378. doi:10.3390/buildings15030378

[15] Yong-Cheol Lee, Moeid Shariatfar, Pedram Ghannad, Jiansong Zhang, and Jin-Kook Lee. 2020. Generation of Entity-Based Integrated Model View Definition Modules for the Development of New BIM Data Exchange Standards. *Journal of Computing in Civil Engineering* 34, 3 (2020), 04020011. doi:10.1061/(ASCE)CP.1943-5487.0000888

[16] Jia-Rui Lin, Yun-Hong Cai, Xiang-Rui Ni, Shaojie Zhou, and Peng Pan. 2026. Qwen-BIM: Developing Large Language Model for BIM-based Design with Domain-specific Benchmark and Dataset. arXiv:2602.20812 [cs.AI] doi:10.48550/arXiv.2602.20812

[17] Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. LeDQA: A Chinese Legal Case Document-based Question Answering Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. ACM, 5385–5389. doi:10.1145/3627673.3679154

[18] Langming Liu, Haibin Chen, Yuhao Wang, Yujin Yuan, Shilei Liu, Wenbo Su, Xiangyu Zhao, and Bo Zheng. 2025. ECKGBench: Benchmarking Large Language Models in E-commerce Leveraging Knowledge Graph. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*. ACM, 6461–6465. doi:10.1145/3746252.3761613

[19] Soumya Madireddy, Lu Gao, Zia Ud Din, Kinam Kim, Ahmed Senouci, Zhe Han, and Yunpeng Zhang. 2025. Large Language Model-Driven Code Compliance Checking in Building Information Modeling. *Electronics* 14, 11 (2025), 2146. doi:10.3390/electronics14112146

[20] Lukas Netz, Jan Reimer, and Bernhard Rumpe. 2024. Using Grammar Masking to Ensure Syntactic Validity in LLM-based Modeling Tasks. In *Proceedings of the 27th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*. ACM, 115–122. doi:10.1145/3652620.3687805

[21] Bharathi Kannan Nithyanantham, Tobias Sesterhenn, Ashwin Nedungadi, Sergio Peral Garijo, Janis Zenkner, Christian Bartelt, and Stefan Lüdtke. 2025. MCP4IFC: IFC-Based Building Design Using Large Language Models. arXiv:2511.05533 [cs.CL] doi:10.48550/arXiv.2511.05533

[22] Zhenhui Ou, Dawei Li, Zhen Tan, Wenlin Li, Huan Liu, and Siyuan Song. 2025. Building Safer Sites: A Large-Scale Multi-Level Dataset for Construction Safety Research. arXiv:2508.09203 [cs.LG] doi:10.48550/arXiv.2508.09203

[23] Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. 2024. Grammar-Aligned Decoding. In *Advances in Neural Information Processing Systems*, Vol. 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/2bdc2267c3d7d01523e2e17ac0a754f3-Abstract-Conference.html

[24] Seungwoo Son, Ghang Lee, Jeaeun Jung, Jungdae Kim, and Kahyun Jeon. 2022. Automated Generation of a Model View Definition from an Information Delivery Manual Using idmXSD and buildingSMART Data Dictionary. *Advanced Engineering Informatics* 54 (2022), 101731. doi:10.1016/j.aei.2022.101731

[25] Issa Sugiura, Takashi Ishida, Taro Makino, Chieko Tazuke, Takanori Nakagawa, Kosuke Nakago, and David Ha. 2025. EDINET-Bench: Evaluating LLMs on Complex Financial Tasks using Japanese Financial Statements. arXiv:2506.08762 [q-fin.ST] doi:10.48550/arXiv.2506.08762

[26] Artur Tomczak, Claudio Benghi, Léon van Berlo, and Eilif Hjelseth. 2024. Requiring Circularity Data in BIM with Information Delivery Specification. *Journal of Circular Economy* (2024). doi:10.55845/REJY5239

[27] Artur Tomczak, Léon van Berlo, Thomas Krijnen, André Borrmann, and Marzia Bolpagni. 2022. A Review of Methods to Specify Information Requirements in Digital Construction Projects. In *Proceedings of the 39th International Conference of CIB W78*. Melbourne, Australia. doi:10.1088/1755-1315/1101/9/092024

[28] Yu Hsiu Tung, Samaneh Rezvani, Fatemeh Asgharzadeh, Maurijn Neumann, and Timo Hartmann. 2025. IFC Whisperer: Querying Building Information Models with Large Language Models and IFC-based Knowledge Graphs. *Journal of Physics: Conference Series* 3140, 16 (2025), 162007. doi:10.1088/1742-6596/3140/16/162007

[29] Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. arXiv:2307.09702 [cs.CL] doi:10.48550/arXiv.2307.09702

[30] Junwen Zheng and Martin Fischer. 2023. BIM-GPT: A Prompt-Based Virtual Assistant Framework for BIM Information Retrieval. arXiv:2304.09333 [cs.CL] doi:10.48550/arXiv.2304.09333