LLM-enabled Ontology Alignment to Support Complex Decision-Making in the Built Environment

Rui Jiang[1], Bige Tunçer[1], Cem Ataman[1]
Affiliation: [1]Department of the Built Environment, Eindhoven University of Technology
Email: r.jiang@tue.nl, b.tuncer@tue.nl, c.ataman@tue.nl

Abstract
Cross-domain knowledge integration is crucial for complex urban and architectural decision-making, with ontology engineering playing a foundational role. To address limited automation and low interpretability in ontology alignment, this study proposes an LLM-powered agent framework to support ontology reuse across built environment domains. The performance of four LLMs and three prompt strategies is explored in alignment tasks. Three specialised agents, for context generation, expert validation, and explanation synthesis, are proposed and integrated with retrieval and mapping agents. Together they form a closed-loop framework linking automatic alignment, expert review, and explanatory feedback. This agent-enhanced architecture improves the traceability and interpretability of alignment results and facilitates expert involvement and iterative refinement, offering a solid foundation for managing and reusing ontologies in complex urban-scale contexts.

Keywords: Large Language Modelling (LLM), ontology alignment, knowledge integration, semantic interoperability

1 Introduction

Complex decision-making in the built environment increasingly relies on multi-source and heterogeneous data across domains (Li et al., 2022; Fadhel et al., 2024). Semantic data modelling has emerged as a key strategy to explicitly represent domain knowledge in a machine-interpretable and human-comprehensible format (Gruber, 1993; Alexopoulos, 2020). Among various representation techniques, ontologies—formalized schemas of domain entities and their relationships—serve as the backbone of semantic integration. Powered by Semantic Web technologies (Berners-Lee, Hendler and Lassila, 2001), ontologies enable structured knowledge from different systems to be unified, queried, and reasoned over, bridging disciplinary boundaries.

Extensive efforts have been made to develop ontologies that capture different layers of spatial, physical, and operational semantics across the built environment. These efforts span diverse areas including infrastructure systems (Nandini and Shahi, 2019; Du et al., 2023), smart city development (Qamar et al., 2019), healthy city planning (Pietra, De Lotto and Bahshwan, 2021), urban disaster response (Dutta and Sinha, 2023), building performance optimization (Utkucu and Sacks, 2025), facility operation and maintenance (F.H et al., 2025).

While numerous ontologies have emerged across domains, their independent and purpose-specific development has led to persistent semantic heterogeneity and limited interoperability. Ontology alignment remains a critical role in enabling scalable knowledge integration and system-level semantic reasoning, especially in large-scale built environment systems like integrated digital twins and urban energy platforms. Ontology alignment, or matching, is the first step toward resolving semantic heterogeneity by generating correspondences between entities such as concepts, properties, or instances (Shvaiko and Euzenat, 2013; Algeria et al., 2015). By comparison, ontology mapping denotes the structured expression of alignment results, often stored as external documents or internal annotations.

Ontology alignment in practice remains heavily dependent on manual efforts (Elshani et al., 2025). It often requires significant cross-domain expertise to interpret conceptual differences, resolve ambiguities, and validate alignment decisions—making the process both time-consuming and difficult to scale. Moreover, the lack of

systematic validation and explanation mechanisms undermines the reliability and reuse of alignment results. Despite its critical role, alignment is rarely treated as a first-class concern in ontology engineering workflows. Mappings are often missing, informal, or undocumented, especially in domain-specific ontologies developed for isolated use (Ma et al., 2024).

These issues severely restrict semantic interoperability across domains and impede the evolution of shared knowledge infrastructures in the built environment. Recent advances in LLMs have opened new possibilities for automation in various phases of ontology engineering (Li, Garijo and Poveda-Villalón, 2025). With their powerful capabilities in semantic understanding, contextual reasoning, and cross-domain generalization, LLMs have shown promising results in a range of ontology-related tasks, including ontology learning (Babaei Giglou, D'Souza and Auer, 2023) ontology generation (Doumanas et al., 2025), terminology annotation and entity completion (Norouzi et al., 2025), ontology evaluation (Tsaneva, Vasic and Sabou, 2024) , among others. However, their potential for application in ontology engineering of the built environment remains largely unexplored and unvalidated.

To address the persistent challenges of low automation, missing alignment modules and mappings and limited interpretability in ontology alignment, this study aims to explore the potential of LLMs to enhance ontology alignment processes in the built environment domain. Specifically, it explores how different LLM types and prompting strategies perform in alignment tasks. Furthermore, this study proposes a dedicated LLM-powered agent framework to improve the interpretability, traceability, and reusability of alignment process and results.

This study presents the first empirical exploration of using LLMs to support cross-domain ontology alignment in the built environment. It fills key gaps in existing frameworks by introducing verification and documentation capabilities, and pioneers a paradigm shift in ontology alignment—from mere result generation to semantic asset construction. Furthermore, it provides an agent-based toolchain that enables ontology experts to generate interpretable alignment comments and reusable documentation.

This paper is structured as follows. Section 2 reviews key challenges in built environment ontologies, the evolution of ontology alignment methods, the integration of LLMs for semantic modelling and alignment, and recent advances in LLM-enhanced documentation generation and interpretability. Section 3 presents a systematic evaluation of four representative LLMs and five prompting strategies across six ontology alignment datasets, highlighting their strengths and limitations under varying semantic complexities. Section 4 analyses the experimental results. Section 5 introduces an enhanced agent-based framework that integrates semantic context construction, human-in-the-loop validation and explanation generation. Section 6 summarizes key findings and outlines future directions for extending ontology alignment as a traceable and interpretable semantic infrastructure.

2. Related Work

2.1 Ontological Engineering Challenges and FAIR Bottlenecks in the Built Environment

Although a large number of urban and architectural ontologies have been proposed, their overall performance across the four dimensions of the FAIR principles, Findability, Accessibility, Interoperability, Reusability, which remains suboptimal. For instance, Mazimwe et al. (2021) evaluated 69 disaster management ontologies and identified significant deficiencies, particularly in interoperability and reusability. Specifically, the average FAIR score was only 19.54%, with 'findability' (or discoverability) scoring as low as 1.8%. Even urban ontologies constructed in recent years, such as city infrastructure ontologies (Du et al., 2023)the ground, roads, and buried pipes, Urban District Sustainability Assessment Ontology (Kuster, Hippolyte and Rezgui, 2020)factoring in environmental, social and economic requirements. However, the current assessment schemes are (a, and Smart City Services Ontology (Qamar et al., 2019)interconnected and intelligent. However, absence of a centralized system produces many

troubles for its stakeholders i.e. citizens and administration in order to perform their day-to-day actions. Semantic web technologies, and in particular linked open data is a tool which allows sharing and merging of data of different systems. Consequently, this paper aims to propose a Smart City Services Ontology (SCSO, generally lack standardized alignment documents and readable annotation structures. While it is widely acknowledged that publishing an ontology should be accompanied by detailed documentation of its concepts and relationships, to reduce the prevalent tendency of researchers to create new ontologies rather than reuse existing ones (Farghaly, Soman and Zhou, 2023), this best practice is rarely implemented in practice. In addition, the resulting mappings are often difficult to interpret, making it challenging for non-ontology experts to understand their semantic intent and the boundaries of their applicability (Mazimwe, Hammouda and Gidudu, 2021).

In recent years, several studies have explored automated and even semi-automated ontology alignment in the built environment (Schneider, 2019; Schlenger et al., 2025). However, no research to date has investigated the use of large language models (LLMs) for ontology alignment within the built environment.

2.2 Extending the Semantic Capabilities of LLMs in Built Environments

With the breakthrough of LLMs in natural language understanding and generation, increasing researchers are trying to introduce them into semantic modelling tasks in built environments. Feng et al. proposed the CityGPT framework, which significantly improves the LLM performance in spatial cognition, path planning, and instruction comprehension by constructing an urban task dataset and an evaluation benchmark (Feng et al., 2024). Li et al. (2025)operation and maintenance (O&M further mapped the mapping relationship between urban tasks and language modelling capabilities, pointing out that it has a high potential for interfacing with urban perception, prediction, and interactive reasoning.

In the context of building systems, Zhang et al (2025) constructed a chain of LLM-based energy simulation agents capable of directly transforming natural language descriptions into EnergyPlus model files. In addition, a control interpretation module incorporating the Shapley mechanism has been preliminarily validated to enhance the transparency of LLM in the field of semantic control of buildings (Zhang, Chen and Ford, 2024). The Shapley mechanism, originally developed in cooperative game theory, assigns credit to input features by quantifying their marginal contribution to a model's output—thus making the decision logic of black-box models more interpretable (Lundberg and Lee, 2017).

However, most of the above studies focus on mid-level tasks, such as information extraction, semantic questions and answers, or operation execution, and there is still a lack of in-depth exploration of semantic alignment at the structural level.

2.3 Evolution of Ontology Alignment Techniques and Trends in LLM Integration

Ontology alignment, as a key link to support ontology reuse, heterogeneous system integration and cross-domain semantic interoperability, has undergone three major stages in recent years. The first stage is represented by rule-based systems such as AML (Faria et al., 2013) and LogMap (Jiménez-Ruiz and Cuenca Grau, 2011), which mainly rely on manually set structural mapping rules with strong logical consistency but lack self-adaptation, which makes it difficult to apply to scenarios with variable contexts. The second stage introduces deep semantic embedding, such as BERTMap (He et al., 2022), which improves the alignment accuracy by word vector training, but requires large training sets and the generalisation ability is limited, which makes it difficult to cover the semantic migration between multiple domains. The third phase is the current key trend: generative alignment methods centred on LLMs. Such approaches achieve zero/few-shot alignment through prompt engineering, with low sample dependency, high generality, and strong semantic generalisation (Hertling and Paulheim, 2023).

Early studies on LLM-based ontology alignment primarily relied on prompt engineering to explore zero-shot and few-shot capabilities, often evaluated on benchmark datasets like Ontology Alignment Evaluation Initiative (OAEI) (Hertling and Paulheim, 2023; Norouzi, Mahdavinejad and Hitzler, 2023). These methods showed promise in terms of recall and generative flexibility, enabling the discovery of novel candidate alignments. However, they also revealed persistent limitations: low precision, unstable outputs, and poor structural consistency. Consequently, LLMs were often positioned as semi-automated tools requiring manual filtering by experts, rather than end-to-end alignment systems.

In addition, LLMs research has gradually shifted focus from early-stage prompt engineering to the development of more autonomous and modular LLM agents. As illustrated in Figure 1, the core of an LLM agent lies in its planning module, which enables the decomposition of complex tasks and iterative improvement through self-reflection. It leverages modular tools (e.g., domain-specific software) to perform both internal reasoning and external actions, and employs short-term and long-term memory to retain contextual knowledge and enhance decision-making (L. Zhang et al., 2025). In contrast to traditional prompt-based LLM usage, which is often inefficient, stateless, and hard to control when handling complex tasks such as ontology matching, Agent-OM exemplifies a paradigm shift: transforming LLMs from reactive generators into proactive decision-makers equipped with planning, memory, and tool-use capabilities. Qiang, Wang and Taylor (2025) introduced the Agent-OM framework, where LLM-based agents are equipped with planning, memory, and tool-use capabilities to modularly execute retrieval and matching workflows. The framework demonstrated competitive performance with state-of-the-art systems in simple tasks and outperformed most baselines in complex and few-shot ontology alignment scenarios, showing particular strengths in handling domain-specific and cross-domain alignment tasks. It slightly underperformed only against a few deep learning systems such as Matcha.

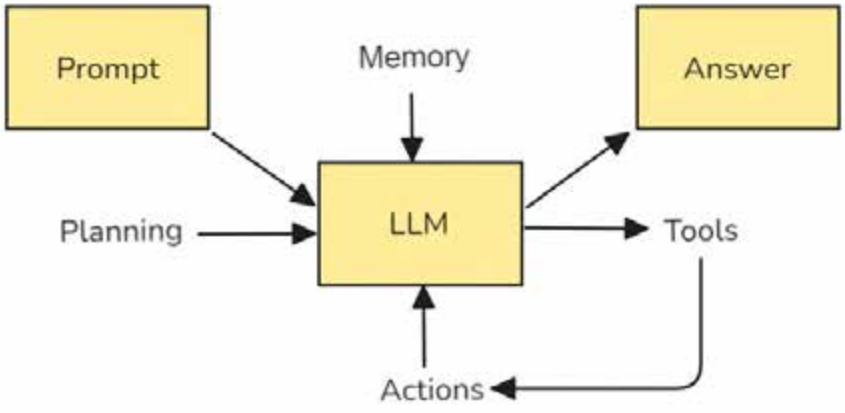

Figure 1. Architecture of an LLM-powered Agent with Prompt, Memory, and Tool Modules

Meanwhile, integration attempts between LLMs and traditional algorithms are also emerging. Taboada et al. (2025)but it remains challenging due to the need for large training datasets and limited vocabulary processing in machine learning approaches. Recently, methods based on Large Language Model (LLMs proposed the MILA framework, which incorporates a Prioritized Depth-First Search (PDFS) strategy to optimize LLM-driven ontology matching. This method significantly reduced the number of LLM queries required for complex alignment tasks. In multiple OAEI tracks, MILA outperformed existing unsupervised and some supervised systems without requiring fine-tuning. Sampels, Efeoglu, and Schimmler (2024) presented a hybrid approach that combines graph-based search algorithms (random walk and tree traversal) with prompt engineering. Their method extracts local

contextual information of candidate entities and verbalizes it into natural language prompts. This approach achieved an F1-score of 0.7147, outperforming zero-shot methods such as GPT and OLaLa, and highlighted the effectiveness of contextualization and verbalization in improving both precision and recall.

2.4 Documentation Generation and LLM-enhanced Interpretability

Recent advances in ontology engineering highlight a growing emphasis on human-readability and interpretability, particularly through documentation generation. While LLMs have demonstrated initial potential in generating ontology-related texts, this area remains underexplored. Only a few studies have addressed ontology publication via documentation generation (Li, Garijo and Poveda-Villalón, 2025). Existing work primarily focuses on generating Competency Questions (CQs) (Rebboud et al., 2024, 2025) and transforming ontology structures into user-facing documentation. In biomedical domains—often more mature than the built environment, researchers have developed specialized tools to enhance accessibility. For example, Giri et al. (2024)Gene Ontology (GO proposed GO2Sum, a model that converts complex lists of Gene Ontology (GO) terms into human-readable functional summaries, improving both usability and interpretability. Complementing this, Norouzi et al. (2023) introduced alignment explanation generation as a novel use of LLMs, enabling experts to provide feedback that informs model refinement through prompt tuning or few-shot updates. These approaches collectively point toward the future of ontology systems that are not only machine-readable but also meaningfully explainable to human users.

3. Performance Comparison of LLMs and Prompting Strategies in Ontology Matching

To explore the adaptability of existing LLM-powered ontology alignment in the context of architectural and urban ontologies, we designed two complementary experiments. Experiment 1 isolates the effect of model version by keeping the prompt fixed, examining whether newer or instruction-tuned models improve semantic alignment. Experiment 2 fixes the model and varies the prompt strategy, evaluating the influence of prompt design on alignment performance. The experimental protocol is presented in Table 1.

Table 1. Overview of Experimental Design

	Objective	Variable Tested
Experiment 1	Explore impact of different LLMs on ontology alignment tasks	Type of LLMs
Experiment 2	Explore impact of different prompt strategies on complex alignment tasks	Type of prompt strategies

Most existing experiments utilize large-sample OAEI datasets and primarily focus on exploring equivalence relationships between ontology entities. However, in real-world engineering applications, subclass and superclass relationships are more prevalent. Therefore, this experiment centres on non-equivalent inclusion relationships, covering the types summarized in Table 2. These relationship types are used to classify semantic mappings in ontology alignment tasks, especially when identifying non-equivalent inclusion relationships.

Table 2 Semantic Relationship Types Between Ontology Entities

Symbolic Form	Description
a ≡ b	a is equivalent to b
a ⊑ b	a is a subclass or subproperty of b

Symbolic Form	Description
a ⊒ b	b is a subclass or subproperty of a
a ⊥ b	a and b are unrelated or disjoint

Note: a refers to entity1, and b refers to entity2

3.1 The Selection of LLMs and Prompting Strategies

To systematically compare LLMs for ontology alignment, we selected four models based on two criteria: (1) demonstrated performance in recent studies on semantic alignment and reasoning, and (2) diversity across commercial and open-source paradigms. Prior work has evaluated models including GPT-3.5/4, LLaMA, BLOOMZ, FLAN-T5, and Falcon-7B (Hertling and Paulheim, 2023; Li, Garijo and Poveda-Villalón, 2025). Among these, the latest high-performing models from the GPT series have achieved state-of-the-art results in zero- and few-shot alignment tasks involving complex semantics (Macilenti, Stellato and Fiorelli, 2024; Qiang et al., 2025)Stellato and Fiorelli, 2024; Qiang {\\i{}et al.}, 2025. GPT-4.1—the most capable publicly available model in the GPT series during the study period—was selected for its semantic reasoning strength and reliable performance. Claude Sonnet 4, a commercial model from the Claude family, was selected for its balanced trade-off between semantic performance and response latency. LLaMA3.1–8B served as a widely adopted open-source baseline, while DeepSeek-R1 was chosen for its emerging benchmark strength and reasoning-oriented design. Table 3 summarizes all model configurations. The four models were evaluated under a unified prompt and task setup in Experiment 1.

Table 3. Test model version and base information

Model Name	Version	Type	Source / API Access
GPT-4.1	gpt-4.1-2025-04-14	Commercial	OpenAI
Claude Sonnet 4	claude-sonnet-4-20250514	Commercial	Anthropic
Deepseek	deepseek-ai/DeepSeek-R1	Open source	Siliconflow
llama3.1	llama3.1:8b	Open source	Local Deployment

In Experiment 2, we designed five prompt types with progressively increasing levels of contextual and semantic richness, as summarized in Table 4. This categorization was developed with reference to recent classification schemes in prompt engineering research (Macilenti, Stellato and Fiorelli, 2024; Amini et al., 2025; Dang et al., 2025), aiming to evaluate model behaviour under varying degrees of input complexity. Prompt 1 served as a minimal baseline, using only basic instructions to assess the model's default generalization ability. Prompt 2 introduced ontology-level background information, such as the ontology's name, domain, and intended purpose. Prompt 3 incorporated few-shot examples to provide more explicit behavioural guidance. Prompt 4 added syntactic and semantic attributes—including domain and range—to test the model's ability to leverage structured information. Finally, Prompt 5 further enriched the input by integrating lexical features extracted from the ontology's entity structure, namely the comment, definition, and example fields, to support fine-grained semantic reasoning and enhance robustness under potentially adversarial conditions.

Table 4. Overview of Prompt Types Used in Experiment 2

Prompt Type	Description
Prompt 1	Minimal instructions to test default generalization
Prompt 2	Adds ontology background (name, domain, purpose)
Prompt 3	Includes few-shot examples to guide model behaviour
Prompt 4	Combines syntactic cues (e.g., labels) with semantic attributes (e.g., domain, range)
Prompt 5	Further enriches input with lexical features (comment, definition, example) for fine-grained semantic understanding

3.2 Experiment Data

This study employs two types of cross-domain ontology alignment datasets: a series of author-curated alignment modules cantered on the Occupant Feedback Ontology (OFO) (Donkers, de Vries and Yang, 2024); and a set of alignment cases derived from Building Topology Ontology (BOT). The former refers to the alignments between the Occupant Feedback Ontology (OFO) and several external ontologies, including the Building Performance Ontology (BOP), Friend of a Friend Ontology (FOAF), Smart Appliances REFerence ontology for Wearables (SAREF4WEAR), Ontology for Property Management (OPM), and the Semantic Sensor Network Ontology (SOSA/SSN) (Donkers et al., 2021). The latter refers to the alignments between the Building Topology Ontology (BOT) and both the Industry Foundation Classes (IFC) and the Brick ontology.

To introduce varied levels of alignment difficulty, we selected entity pairs from the database where the source and target names are non-equivalent and grouped them into distinct subsets according to their semantic complexity.

Table 5 provides a comparative overview of all alignment datasets used in this study. Among them, OFO-BPO represents the most challenging set, especially due to its inclusion of abstract and dynamic concepts related to communication and sensing. Similar levels of complexity are observed in the OFO-SAREF4WEAR and OFO-SOSA/SSN subsets, which also involve cross-domain and behaviour-oriented semantics. BOT-BRICK and OFO-OPM are categorized as medium-difficulty datasets. These alignments remain largely within the architectural domain and do not exhibit substantial conceptual shifts across domains, though they still involve structural and functional abstraction. Finally, the BOT-IFC dataset is considered the easiest. In this set, many target entity names contain the source entity name as a substring (e.g., IfcBuildingStorey vs. Storey), without being strictly equal. These name-based cues make the alignment more straightforward compared to the others.

Table 5. Ontology Alignment Datasets Overview

Dataset	Classes	Properties	Example Pair
OFO-BOP	0	4	monitors Feedback ↔ performs Execution
OFO-SOSA/SSN	1	2	Wearable ↔ System
OFO-SAREF-4WEAR	2	2	Result ↔ Measurement
OFO-OPM	1	0	Result ↔ Property State

BOT-IFC	5	0	Space ↔ IfcSpace
BOT-BRICK	2	3	contains Zone ↔ has Part

Note: The numbers of Classes and Properties refer to aligned pairs

3.3 Experiment Procedure

The experiment consists of three main stages: (1) environment setup and configuration, (2) data preparation and gold-standard alignment generation, and (3) ontology matching execution. All experiments were conducted in an Anaconda virtual environment. Model variants and prompt strategies were specified via the configuration file, and necessary API keys (e.g., OpenAI) were securely initialized.

This study partially adopts the retrieval module and configuration setup from Agent-OM by Qiang et al. (2025), particularly for extracting syntactic, lexical, and semantic features. To address the high semantic heterogeneity across ontologies and the limited number of reference alignments, we implemented a full-retrieval strategy that exhaustively generates all possible candidate pairs and evaluates them individually to ensure comprehensive coverage. On top of this, we customized the prompt templates and extended the explanation module to support more interpretable and task-specific alignment outputs.

Each experiment run relied on three input files: the source ontology, the target ontology, and a reference alignment file. The ontologies were retrieved from publicly accessible repositories and provided in RDF/XML format. Reference alignments, originally authored by ontology developers, were standardized into Alignment API-compliant XML files and subsequently parsed to generate gold-standard alignment pairs, from which subclass and superproperty mappings were explicitly extracted. We began by applying a baseline prompt (Prompt 1) to evaluate the ontology alignment performance of four representative LLMs. Based on the results, GPT-4.1 demonstrated the most favourable trade-off among accuracy, efficiency, and cost, and was therefore selected for downstream experiments. Subsequently, we conducted a systematic evaluation of five distinct prompt types across six benchmark datasets, allowing for a fine-grained analysis of how prompt design impacts ontology matching performance.

4. Results

This section presents the results of two ontology alignment experiments and discusses the comparative performance of different LLMs and prompting strategies.

4.1 Performance Comparison of Different Models

Across all alignment tasks, GPT-4.1 consistently outperformed other models in terms of F1 score, demonstrating superior robustness in both semantic generalization and relation directionality detection. Claude exhibited comparable precision to GPT yet suffered from reduced recall in structurally divergent pairs such as OFO-OPM and OFO-SOSA. DeepSeek-R1, while capable of identifying general subclass relations, frequently hallucinated symmetric alignments, especially in BOT-IFC and OFO-SAREF4WEAR, where it confused $\sqsubseteq$ with $\equiv$. LLaMA, though lightweight, showed competitive recall in BOT-BRICK, but struggled in abstract property mappings across OFO-centric scenarios.

Notably, precision-recall trade-offs were observed across all models. For instance, Claude exhibited high precision on BOT-BRICK but missed multiple legitimate subclass mappings, leading to low recall. Conversely, DeepSeek's aggressive recall yielded inflated false positives, lowering precision. The OFO-BOP and OFO-OPM tasks proved most challenging, as they required semantic bridging across distinct conceptual layers, such as subjective feedback and operational control, which many models failed to capture reliably. These findings indicate that

current LLMs remain sensitive to ontology lexical divergence and relation asymmetry, and highlight the need for context-augmented, agent-based refinement to achieve consistent alignment performance in multi-domain ontology ecosystems.

4.2 Comparison of Prompting Strategies

To investigate how prompting strategies affect the performance of large language models in ontology alignment tasks, we compared GPT's outputs across six ontology pairs using three prompt types: Prompt 1, Prompt 2, and Prompt 3. Figure 2 illustrates that prompt formulation significantly influences alignment quality, especially in scenarios with structural complexity or abstract semantics. For relatively straightforward alignment tasks, Prompt 1 often led to high recall, retrieving a large number of candidate alignments. However, this came at the cost of reduced precision, particularly due to overgeneration of subsumption and equivalence relations. This trend is evident in ontology pairs such as BOT-IFC, where many predicted matches reflected broad semantic overlaps but lacked precise correspondence.

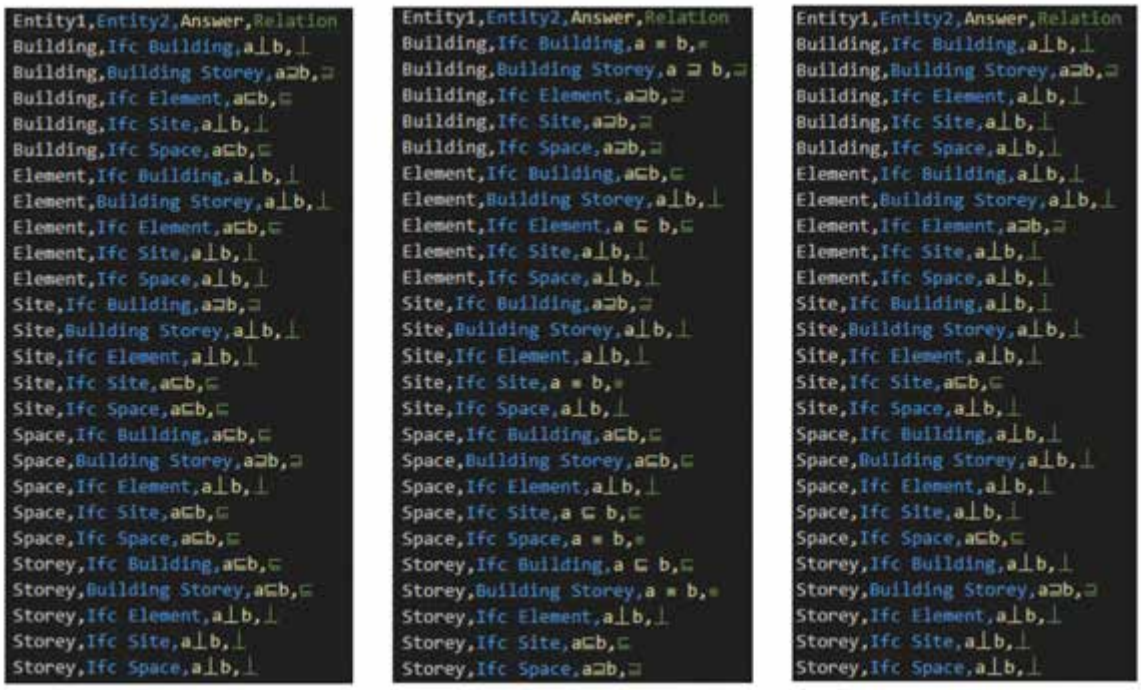

Figure 2. Alignment Outputs under Three Prompting Strategies (Prompt 1 → Prompt 2 → Prompt 3, from left to right)

In contrast, Few-shot prompting improved precision, especially for subclass and part-of relationships. This indicates that GPT was able to effectively learn from curated examples, producing more selective and accurate matches. Nonetheless, minor directional errors were observed in some subsumption cases, suggesting sensitivity to example framing. When Context-enhanced prompts were used, the output pattern resembled that of the Simple prompts: high recall with a large number of inferred equivalence and subsumption relationships. This suggests that while additional ontology context enriches the semantic field, it may also amplify GPT's tendency to generate broader relational assertions, potentially at the expense of specificity.

Overall, Few-shot prompting yielded better balance in precision-recall trade-off for structurally simple align-

ments, while Context-enhanced prompts were more effective in hierarchical or nested ontologies, such as BOT-BRICK and OFO-OPM, where semantic context is essential for interpreting complex relations. These findings highlight the nuanced effects of prompt design and support the inclusion of tailored prompting strategies in ontology alignment systems.

Prompt 1, Prompt 4, and Prompt 5 were compared to evaluate the impact of different types of semantic information on ontology alignment performance. Among them, Prompt 4, which integrates structural and contextual information (e.g., domain, range, hierarchical roles), consistently achieved the highest balanced F1-score, demonstrating its necessity for reliable prediction. In terms of prompt design, Prompt 4 significantly outperformed Prompt 5, particularly in alignment tasks of moderate difficulty within the built environment domain. While Prompt 5 introduces additional annotations and ontology-specific details (e.g., comments, URIs), this extra information occasionally misleads the model, causing it to infer inclusion or equivalence relationships even when entities are only loosely or indirectly related.

As illustrated in Figure 3, entity pairs such as contains Element and is Location Of were correctly judged as unrelated under Prompt 4. However, under Prompt 5, GPT tended to generate false positive relations, interpreting such pairs as semantically connected due to the presence of shared terms in the textual context. Moreover, in moderate alignment tasks, entity pairs often exhibit partial or ambiguous overlap, making it easy for the model to avoid predicting "unrelated." This behaviour inflates recall but comes at the cost of reduced precision, particularly when weak semantic signals are amplified by prompt structure.

Figure 3. Alignment Outputs under Three Prompting Strategies (Prompt 1 → Prompt 4 → Prompt 5, from left to right)

5. Discussions and extended framework for enhanced ontology alignment

5.1 Discussions

The adoption of agent-based design in ontology alignment is not merely beneficial but increasingly necessary. As the complexity of alignment tasks and the scale of ontological data continue to grow, agent architectures offer critical advantages—most notably in modularity, scalability, integration with external reasoning tools, contextual management, and expert-in-the-loop support.

The Agent-OM framework proposed by Qiang, Wang, and Taylor (2025) provides an important starting point by introducing LLM agents for retrieval and matching. However, it lacks a dedicated mechanism for generating alignment context, which limits its ability to capture precise domain semantics and background knowledge. In addition, the capabilities of large language models remain constrained, particularly when addressing abstract or structurally complex concepts. Despite ongoing efforts to enhance automation, human-in-the-loop verification remains indispensable, especially in cross-domain ontology alignment tasks within the built environment. Current approaches still lack robust mechanisms for error correction, result validation, and for ensuring interpretability and traceability throughout the alignment process. As Doumanas et al. (2024) emphasize, while LLMs can significantly accelerate ontology development, expert oversight remains essential to guarantee alignment quality, transparency, and reliability. Bridging this gap calls for standardized benchmarks and hybrid workflows that combine automated alignment with structured expert feedback (Li, Garijo and Poveda-Villalón, 2025).

Finally, although the original Agent-OM architecture supports multi-layer parsing and semantic-syntactic feature extraction, the direct feedback from LLMs tends to be verbose and lacks clarity. It does not generate concise, high-quality explanatory documentation necessary for human interpretation and downstream reuse. Our preliminary experiments using JSON-formatted outputs confirmed the difficulty of interpreting LLM errors in the absence of structured rationale. These findings highlight the need for enhanced tool integration for structured reasoning, tighter expert–LLM interaction loops, and more interpretable output formats.

5.2 An LLM-powered agent framework for enhanced ontology alignment

To address the above, we propose an LLM-powered agent framework for enhanced ontology alignment. The overall architecture and workflow are presented in Figure 4. Building on the architecture of Agent-OM, our extended framework integrates Retrieval and Matching Agents into a centralized planning loop and introduces three additional agents: the Context Agent, which constructs structured semantic context; the Expert Review Agent, which enables multi-level human validation for complex relations such as containment and subclassing; and the Explanation Agent, which generates interpretable alignment rationales and enhanced OWL/XML annotations.

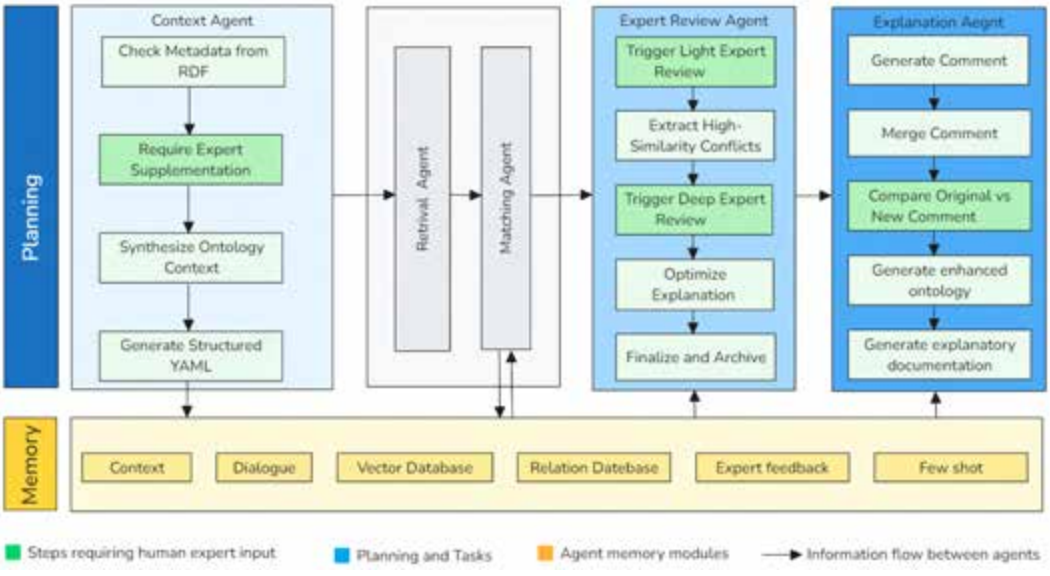

Figure 4. Explainable ontology alignment agent architecture

Note: The system consists of four agent modules: a Context Agent that synthesizes ontology background knowledge, a Retrieval Agent and Matching Agent for candidate generation, an Expert Review Agent for human-in-the-loop validation, and an Explanation Agent that produces revised comments and final documentation.

The Context Agent constructs a complete semantic context to support ontology alignment. It begins by extracting RDF metadata from the source ontology and, if necessary, performing supplementary web searches to enrich missing descriptions. When critical context fields are absent, expert-driven supplementation is triggered. Experts provide additional inputs, including the expected application domain of the target ontology, representative use cases for both ontologies, and the rationale for alignment—that is, why specific components of the source ontology should be aligned to parts of the target ontology. The resulting metadata is stored both in the central database and in the Context Agent's internal memory module, enabling reuse across future alignment sessions and improving contextual continuity.

The Expert Review Agent supports interactive human-in-the-loop validation, including not only equivalence but also complex structures such as containment and subclassing. It conducts both light and deep review stages: the light review focuses on quickly confirming high confidence matches and identifying potential misclassifications, while the deep review targets ambiguous or structurally complex entity pairs. Experts annotate each decision with structured justifications. These justifications are further refined by the LLM to ensure clarity, consistency, and future reusability. Finally, all verified results are archived for use by the Explanation Agent to generate comprehensive alignment rationales. Expert feedback is also stored as potential few-shot exemplars to guide future alignment tasks.

The Explanation Agent transforms validated alignment rationales into structured, human-readable documentation. It first processes the verified results from the Expert Review Agent to generate semantic alignment justifications, which are then merged with existing ontology comments. Experts may review and confirm whether the new comments should overwrite the originals. Finalized annotations are written back into the ontology, resulting in an enhanced OWL/XML version with improved interpretability and traceability. In parallel, the agent compiles an external explanation document, serving as both a human-readable rationale archive and a resource for ontology reuse and future alignment transparency.

5.3. Limitations and future work

Several limitations in the current study warrant further investigation. First, although multiple LLMs were evaluated, the analysis did not encompass some of the most recent models—such as Claude 4 and LLama 4—whose advanced architectures may offer advantages in particular alignment scenarios. Second, the evaluation datasets used in this study primarily basic containment relations. However, ontologies in the built environment often embody more complex semantic structures, including hierarchical constraints, compositional dependencies, and semantic overlaps. Expanding benchmark coverage to include such relation types is necessary to fully validate the utility of the experimental results. Third, additional algorithmic and empirical efforts are needed to determine optimal few-shot configurations and prompt engineering strategies that can improve LLM-based ontological reasoning. The interaction between prompt design, context granularity, and relation types remains insufficiently explored, yet is likely to significantly impact the robustness and generalizability of alignment performance across domains. Furthermore, based on the present framework, we are developing an LLM-powered agent system that operationalizes enhanced ontology alignment through modular planning, expert feedback integration, and explainable output generation.

6. Conclusion

This study investigates the potential of LLMs to support ontology alignment across built environment domains and proposes an agent-based enhancement to address observed limitations. We first conducted systematic experiments comparing four LLMs and five prompting strategies across ontology alignment tasks. The results show that GPT-4.1 achieves a favorable balance between performance and cost, and that prompting strategies significantly influence alignment quality depending on the semantic complexity and structural characteristics of ontology pairs. However, the experiments also reveal that relying solely on prompt engineering is insufficient—particularly for handling complex, ambiguous, or hierarchical relations. In response, we propose an enhanced LLM-powered agent framework that builds on the initial pipeline and introduces three additional agents: a Context Agent for semantic grounding, an Expert Review Agent for human-in-the-loop validation, and an Explanation Agent for generating interpretable alignment rationales. These enhancements improve the interpretability, traceability, and reusability of the alignment process and outputs. Rather than treating alignment as a static, one-time task, the proposed framework supports iterative expert involvement and positions ontology alignment as an evolving semantic asset in complex, multi-domain settings. Rather than treating alignment as a static, one-time task, the proposed framework supports iterative expert involvement and positions ontology alignment as an evolving semantic asset in complex, multi-domain settings.

Acknowledgements

This publication is partially supported by the project Multi-Ontology AI based Facility Digital Twin (MOAF-DiT) with project number estar24226 E! 5902 of the research programme Eurostars, which is (partly) financed by the Rijksdienst voor Ondernemend Nederland (Netherlands Enterprise Agency). Eurostars is part of the European Partnership on Innovative SMEs. The partnership is co-funded by the European Union through Horizon Europe.

References

Alexopoulos, P. (2020) Semantic modeling for data. O'Reilly Media.

Algeria et al. (2015) 'Ontology-alignment techniques: Survey and analysis', International Journal of Modern Education and Computer Science, 7(11), pp. 67–78. Available at: https://doi.org/10.5815/ijmecs.2015.11.08.

Amini, Reihaneh et al. (2025) 'Towards complex ontology alignment using large language models', in S. Tiwari et al. (eds) Knowledge Graphs and Semantic Web. Cham: Springer Nature Switzerland, pp. 17–31. Available at: https://doi.org/10.1007/978-3-031-81221-7_2.

Babaei Giglou, H., D'Souza, J. and Auer, S. (2023) 'LLMs4OL: Large language models for ontology learning', in T.R. Payne et al. (eds) The Semantic Web – ISWC 2023. Cham: Springer Nature Switzerland, pp. 408–427. Available at: https://doi.org/10.1007/978-3-031-47240-4_22.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'Web semantic', Scientific American, 284(5), pp. 34–43.

Dang, P. et al. (2025) 'Semantic-driven parametric 3D geographic scene modeling: Integrating knowledge graphs and large language models', Environmental Modelling & Software, 188, p. 106399. Available at: https://doi.org/10.1016/j.envsoft.2025.106399.

Donkers, A., de Vries, B. and Yang, D. (2024) 'Creating occupant-centered digital twins using the occupant feedback ontology implemented in a smartwatch app', Semantic Web, 15(2), pp. 259–284. Available at: https://doi.org/10.3233/SW-223254.

Donkers, A.J. et al. (2021) 'Real-time building performance monitoring using semantic digital twins', in 9th Linked Data in Architecture and Construction Workshop, LDAC 2021. CEUR-WS. org, pp. 55–66.

Doumanas, D. et al. (2024) 'From human- to LLM-centered collaborative ontology engineering', Applied Ontology, 19(4), pp. 334–367. Available at: https://doi.org/10.1177/15705838241305067.

Doumanas, D. et al. (2025) 'Fine-tuning large language models for ontology engineering: A comparative analysis of GPT-4 and mistral', Applied Sciences, 15(4), p. 2146. Available at: https://doi.org/10.3390/app15042146.

Du, H. et al. (2023) 'City infrastructure ontologies', Computers, Environment and Urban Systems, 104, p. 101991. Available at: https://doi.org/10.1016/j.compenvurbsys.2023.101991.

Dutta, B. and Sinha, P.K. (2023) 'An ontological data model to support urban flood disaster response', Journal of Information Science, p. 01655515231167297. Available at: https://doi.org/10.1177/01655515231167297.

Elshani, D. et al. (2025) 'AEC co-design workflow for cross-domain querying and reasoning using semantic web technologies', Automation in Construction, 176, p. 106226. Available at: https://doi.org/10.1016/j.autcon.2025.106226.

Fadhel, M.A. et al. (2024) 'Comprehensive systematic review of information fusion methods in smart cities and urban environments', Information Fusion, 107, p. 102317. Available at: https://doi.org/10.1016/j.inffus.2024.102317.

Farghaly, K., Soman, R.K. and Zhou, S.A. (2023) 'The evolution of ontology in AEC: A two-decade synthesis, application domains, and future directions', Journal of Industrial Information Integration, 36, p. 100519. Available at: https://doi.org/10.1016/j.jii.2023.100519.

Faria, D. et al. (2013) 'The AgreementMakerLight ontology matching system', in R. Meersman et al. (eds) On the Move to Meaningful Internet Systems: OTM 2013 Conferences. Berlin, Heidelberg: Springer, pp. 527–541. Available at: https://doi.org/10.1007/978-3-642-41030-7_38.

Feng, J. et al. (2024) 'CityGPT: Empowering urban spatial cognition of large language models'. Available at: https://doi.org/10.48550/arXiv.2406.13948.

F.H, A. et al. (2025) 'BIM ontology for information management (BIM-OIM)', Journal of Building Engineering, 107, p. 112762. Available at: https://doi.org/10.1016/j.jobe.2025.112762.

Giri, S.J., Ibtehaz, N. and Kihara, D. (2024) 'GO2Sum: Generating human-readable functional summary of proteins from GO terms', npj Systems Biology and Applications, 10(1), pp. 1–12. Available at: https://doi.org/10.1038/s41540-024-00358-0.

Gruber, T.R. (1993) 'A translation approach to portable ontology specifications', Knowledge Acquisition, 5(2), pp. 199–220. Available at: https://doi.org/10.1006/knac.1993.1008.

He, Y. et al. (2022) 'BERTMap: A BERT-based ontology alignment system', Proceedings of the AAAI Conference on Artificial Intelligence, 36(5), pp. 5684–5691. Available at: https://doi.org/10.1609/aaai.v36i5.20510.

Hertling, S. and Paulheim, H. (2023) 'OLaLa: Ontology matching with large language models', in Proceedings of the 12th Knowledge Capture Conference 2023. New York, NY, USA: Association for Computing Machinery (K-CAP '23), pp. 131–139. Available at: https://doi.org/10.1145/3587259.3627571.

Jiménez-Ruiz, E. and Cuenca Grau, B. (2011) 'LogMap: Logic-based and scalable ontology matching', in L. Aroyo et al. (eds) The Semantic Web – ISWC 2011. Berlin, Heidelberg: Springer, pp. 273–288. Available at: https://doi.org/10.1007/978-3-642-25073-6_18.

Kuster, C., Hippolyte, J.-L. and Rezgui, Y. (2020) 'The UDSA ontology: An ontology to support real time urban sustainability assessment', Advances in Engineering Software, 140, p. 102731. Available at: https://doi.org/10.1016/j.advengsoft.2019.102731.

Li, J., Garijo, D. and Poveda-Villalón, M. (2025) 'Large language models for ontology engineering: A systematic literature review'.

Li, X. et al. (2022) 'The six dimensions of built environment on urban vitality: Fusion evidence from multi-source data', Cities, 121, p. 103482. Available at: https://doi.org/10.1016/j.cities.2021.103482.

Li, Y. et al. (2025) 'A large language model-based building operation and maintenance information query', Energy and Buildings, 334, p. 115515. Available at: https://doi.org/10.1016/j.enbuild.2025.115515.

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', in Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc. (NIPS'17), pp. 4768–4777.

Ma, R. et al. (2024) 'An ontology-driven method for urban building energy modeling', Sustainable Cities and Society, 106, p. 105394. Available at: https://doi.org/10.1016/j.scs.2024.105394.

Macilenti, G., Stellato, A. and Fiorelli, M. (2024) 'Prompting is not all you need evaluating GPT-4 performance on a real-world ontology alignment use case', Procedia Computer Science, 246, pp. 1289–1298. Available at: https://doi.org/10.1016/j.procs.2024.09.557.

Mazimwe, A., Hammouda, I. and Gidudu, A. (2021) 'Implementation of FAIR principles for ontologies in the disaster domain: A systematic literature review', ISPRS International Journal of Geo-Information, 10(5), p. 324. Available at: https://doi.org/10.3390/ijgi10050324.

Nandini, D. and Shahi, G.K. (2019) 'An ontology for transportation system', in Kalpa Publications in Computing. Selected Student Contributions and Workshop Papers of LuxLogAI 2018, EasyChair, pp. 32–25. Available at: https://doi.org/10.29007/qt2m.

Norouzi S.S. et al. (2025) 'Ontology population using LLMs', in Handbook on Neurosymbolic AI and Knowledge Graphs. IOS Press, pp. 421–438. Available at: https://ebooks.iospress.nl/doi/10.3233/FAIA250217 (Accessed: 23 May 2025).

Norouzi, S.S., Mahdavinejad, M.S. and Hitzler, P. (2023) 'Conversational ontology alignment with ChatGPT'. Available at: https://doi.org/10.48550/arXiv.2308.09217.

Pietra, C., De Lotto, R. and Bahshwan, R. (2021) 'Approaching healthy city ontology: First-level classes definition using BFO', Sustainability, 13(24), p. 13844. Available at: https://doi.org/10.3390/su132413844.

Qamar, T. et al. (2019) 'Smart city services ontology (SCSO): Semantic modeling of smart city applications', in 2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC). 2019 Seventh

International Conference on Digital Information Processing and Communications (ICDIPC), pp. 52–56. Available at: https://doi.org/10.1109/ICDIPC.2019.8723785.

Qiang, Z. et al. (2025) 'OAEI-LLM: A benchmark dataset for understanding large language model hallucinations in ontology matching'. Available at: https://doi.org/10.48550/arXiv.2409.14038.

Qiang, Z., Wang, W. and Taylor, K. (2025) 'Agent-OM: Leveraging LLM agents for ontology matching'. Available at: https://doi.org/10.48550/arXiv.2312.00326.

Rebboud, Y. et al. (2024) 'Benchmarking LLM-based ontology conceptualization: A proposal', in ISWC 2024, 23rd International Semantic Web Conference. Baltimore, United States. Available at: https://hal.science/hal-04757816 (Accessed: 29 May 2025).

Rebboud, Y. et al. (2025) 'Can LLMs generate competency questions?', in A. Meroño Peñuela et al. (eds) The Semantic Web: ESWC 2024 Satellite Events. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), pp. 71–80. Available at: https://doi.org/10.1007/978-3-031-78952-6_7.

Sampels J., Efeoglu S. and Schimmler S. (2024) 'Exploring prompt generation utilizing graph search algorithms for ontology matching', in Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI. IOS Press, pp. 2–19. Available at: https://ebooks.iospress.nl/doi/10.3233/SSW240003 (Accessed: 5 May 2025).

Schlenger, J. et al. (2025) 'Reference architecture and ontology framework for digital twin construction', Automation in Construction, 174, p. 106111. Available at: https://doi.org/10.1016/j.autcon.2025.106111.

Schneider, G. (2019) 'Automated ontology matching in the architecture, engineering and construction domain - A case study', in. Linked Data in Architecture and Construction Workshop (LDAC) 2019. Available at: https://publica.fraunhofer.de/handle/publica/406940 (Accessed: 5 June 2025).

Shvaiko, P. and Euzenat, J. (2013) 'Ontology matching: State of the art and future challenges', IEEE Transactions on Knowledge and Data Engineering, 25(1), pp. 158–176. Available at: https://doi.org/10.1109/TKDE.2011.253.

Taboada, M. et al. (2025) 'Ontology matching with large language models and prioritized depth-first search'. Available at: https://doi.org/10.48550/arXiv.2501.11441.

Tsaneva, S., Vasic, S. and Sabou, M. (2024) 'Llm-driven ontology evaluation: Verifying ontology restrictions with chatgpt', The Semantic Web: ESWC Satellite Events, 2024.

Utkucu, D. and Sacks, R. (2025) 'Ontology for holistic building performance modeling and analysis', Automation in Construction, 175, p. 106197. Available at: https://doi.org/10.1016/j.autcon.2025.106197.

Zhang, L. et al. (2025) 'Automatic building energy model development and debugging using large language models agentic workflow', Energy and Buildings, 327, p. 115116. Available at: https://doi.org/10.1016/j.enbuild.2024.115116.

Zhang, L., Chen, Z. and Ford, V. (2024) 'Advancing building energy modeling with large language models: Exploration and case studies', Energy and Buildings, 323, p. 114788. Available at: https://doi.org/10.1016/j.enbuild.2024.114788.

Zhang, Y. et al. (2025) 'Data-driven building load prediction and large language models: Comprehensive overview', Energy and Buildings, 326, p. 115001. Available at: https://doi.org/10.1016/j.enbuild.2024.115001.