



Original scientific paper

AI-Driven Knowledge Transfer in Architectural Education

^{*1} Nik Ansre, ² Yara Hirsekorn, ³ Gregor Grunwald

^{1, 2, & 3} Department of Architecture, Jade University of Applied Sciences Oldenburg, Germany

¹ E-mail: nik.ansre@student.jade-hs.de, ² E-mail: jarahirsekorn@gmail.com, ³ E-mail: gregor.grunwald@jade-hs.de

¹ ORCID: <https://orcid.org/0009-0006-4080-4398>, ² ORCID: <https://orcid.org/0009-0001-7275-2450>, ³ ORCID: <https://orcid.org/0000-0003-2437-398X>

ARTICLE INFO:

Article History:

Received: 18 May 2025

Revised: 15 July 2025

Accepted: 15 July 2025

Available online: 20 July 2025

Keywords:

AI,
Architecture,
Knowledge Transfer,
Retrieval-Augmented-Generation,
Digital Education,
Model Optimization.

ABSTRACT

The use of large language models (LLMs) in architecture is still at an early stage, but is becoming increasingly important due to the need for local, domain-specific assistance systems. This study investigates how such models can be developed and extended under local hardware conditions in order to provide architecture-specific knowledge in a context-appropriate manner. For this purpose, a methodological approach was chosen that combines lightweight fine-tuning with a two-stage retrieval augmented generation system (RAG). Two model series were tested: Mini-Spyra for dialogue-oriented knowledge retrieval and IwI-Spyra for semantic analysis of structured planning data (e.g. IFC). The results show that domain-specific training in combination with dynamic knowledge integration leads to significantly more precise, comprehensible answers. The article provides a transferable model for the use of AI in teaching and planning practice - data protection-compliant, locally executable and didactically comprehensible. The study thus contributes to the development of explainable AI tools in an architectural context. This research was conducted within a "Young Researchers" project supported by the Jade University of Applied Sciences.

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International (CC BY) license.



Publisher's Note:

Journal of *Smart Design Policies* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

SMART DESIGN POLICIES (2025), 2(1), 122–139.

<https://doi.org/10.38027/smart.v2n1-8>

www.smartdpj.com

Copyright © 2025 by the author(s).

* Corresponding Author

How to cite this article: (APA Style)

Ansre, N., Hirsekorn, Y., & Grunwald, G. (2025). AI-driven knowledge transfer in architectural education.

Smart Design Policies, 2(1), 122–139. <https://doi.org/10.25034/smart-v2n1-8>

1. Introduction

1.1 Background and Context

In recent years, technological progress in the field of language models (large language models) has led to a large number of new fields of application - particularly in the context of planning, teaching and data processing in architecture. While general AI models are already established in areas such as text creation, translation and coding, they often lack technical depth, customisable semantics and the ability to reliably interpret complex, structured technical content. Especially in architecture, where digital planning processes, standards and semantic data formats such as IFC are part of everyday life, there are considerable requirements for context sensitivity, understanding of terminology and interpretative precision.

At the same time, many educational institutions and smaller planning offices are faced with the challenge of operating AI systems in compliance with data protection regulations, cost-efficiently and without cloud dependency. As a result, the focus is shifting to locally executable language models

that can be operated on in-house hardware and customised to domain-specific knowledge. The potential added value of such systems is obvious: automated standard interpretation, assisted plan analysis, semantic knowledge transfer - all embedded in dialogue-oriented systems that are also accessible to non-programming users (Hanke, 2024).

The present work is positioned at the interface between model development, architectural theory and digital planning processes. It explores the question of how specialised language models can be built, trained and supplemented by retrieval systems in order to not only correctly reproduce subject-specific knowledge from the field of architecture, but also to make it usable in the planning context - for example to support teaching, model checking or design decisions.

1.2 Research Gap and Objectives

Although there are now numerous studies on the development of generative AI models, there is still a lack of systematic approaches to the application of such systems in architecture - particularly in combination with semantically structured data formats such as IFC and in environments with limited computing power. Most existing systems are cloud-based, generically trained and not very transparent in terms of their knowledge sources. This represents a significant obstacle, particularly in the university context, where traceability, data protection and didactic clarity are essential.

A further need for research is how language models can be enriched with project or teaching-specific knowledge in a targeted manner without this having to be statically integrated into the model. Retrieval Augmented Generation (RAG) offers a promising solution here, but has hardly been implemented or documented in the field of architecture to date (Ansre et al., 2025).

This study therefore addresses the following research question:

'How can domain-specific language models be developed and extended under local hardware conditions to provide architecture-specific knowledge in a context-appropriate way and to semantically analyse structured planning data (e.g. IFC)?'

This results in three central goals:

Development, testing and evaluation of locally executable language models for the architecture-specific context.

Systematic structuring and integration of subject-specific knowledge modules as a basis for training and retrieval.

Extension of the models with a two-stage RAG system for the dynamic integration of external content.

1.3 Contribution and Structure of the Paper

This work contributes to practical AI research in the field of architecture by providing a methodological framework for the development, extension and application of subject-specific language models. For universities, it offers a realisable model of how AI systems can be used in teaching - for example to teach complex technical terms, standards or software processes. For planning practice, it enables the development of locally operated assistance systems that can specifically access project-specific knowledge and analyse structured building data such as IFC files. The study also provides a scalable framework for integrating knowledge components into AI systems and uses specific application scenarios to demonstrate the differences between various model versions - both in terms of response quality and technical efficiency. The combination of qualitative fine-tuning, structured knowledge structure and RAG-based extension represents a forward-looking model for domain-specific AI applications.

The structure of the thesis is as follows:

Chapter 2 explains the methodological approach and the technical setup, Chapter 3 presents the results of the model development and application. Chapter 4 critically discusses the findings in terms of usability, limitations and transfer potential. Chapter 5 concludes with an outlook on future developments.

2. Materials and Methods

2.1 Study Design and Setting

The aim of this study was to develop a language model tailored to the needs of architectural practice and teaching that can be executed locally, has low hardware requirements and is also capable of providing subject-specific knowledge in a precise and context-sensitive manner. The development took place as part of an interdisciplinary research project at the Jade University of Applied Sciences and was anchored in the ‘Young Researchers’ programme. In terms of methodology, a practice-orientated approach was chosen that addresses both technical and semantic-didactic challenges.

The implementation was carried out entirely in a locally containerised environment, built on the Ollama platform, which enabled the parallel execution of several model variants with dynamic resource allocation. This allowed both experimental and production-related AI applications to be tested realistically.

The methodological process of this study was divided into three overarching phases:

Model development (green): Selection of a suitable base model, establishment of the technical infrastructure, definition of the core knowledge components and subsequent fine-tuning with curated data.

Preparation of the system (blue): Segmentation and semantic embedding of the data, development of the two-stage Retrieval Augmented Generation (RAG) system and construction of several specialised model variants.

Evaluation and application (purple): Implementation of concrete application scenarios, e.g. in the area of standards analysis or IFC-supported plan review, followed by reflection, evaluation and documentation of the results.

These steps are shown schematically in Figure 1:

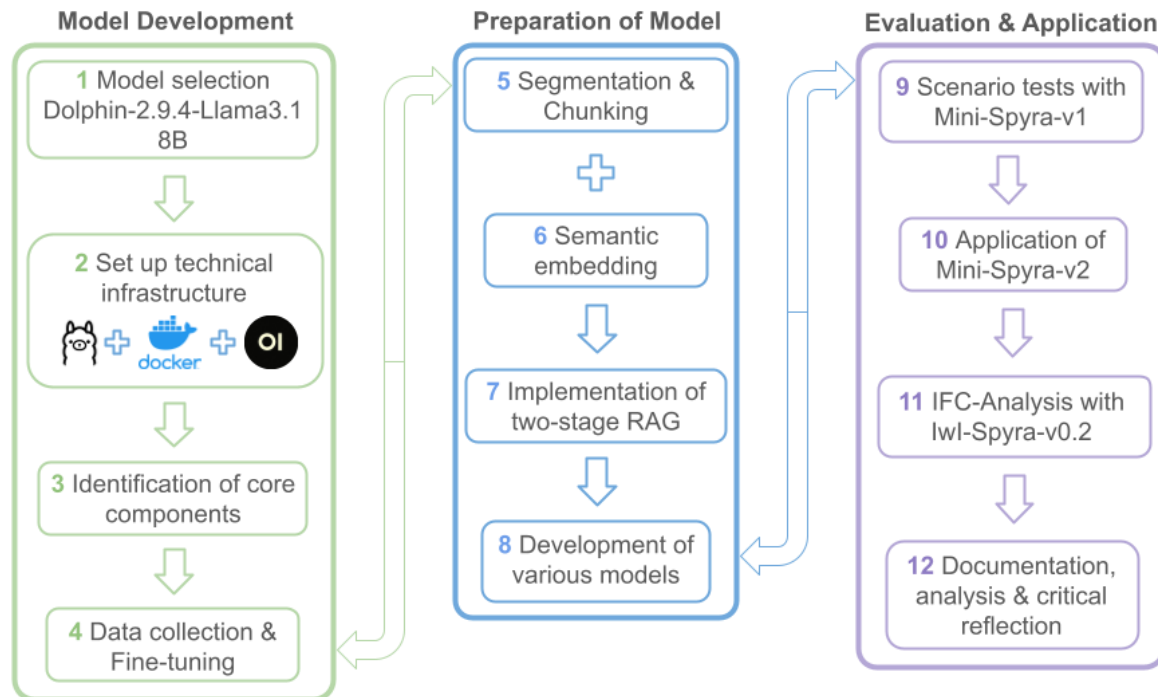


Figure 1. Methodological workflow of the study.

The knowledge areas identified in the first step formed the semantic foundation for all subsequent development phases. In order to anchor the model not only linguistically, but also domain-specifically, this content had to be recorded, organised and formalised in a structured manner - both for training and for subsequent retrieval processes.

2.2 Identification of Core Knowledge Components

The development of a domain-specific language model presupposes that relevant expertise is available not only as a set of texts, but also in a structured form. For this purpose, so-called ‘Core Knowledge Components’ (CKC) were defined: central knowledge modules that are required for processing typical tasks in the fields of architecture, BIM and digital design. These building blocks form the semantic basis for the subsequent training of the language model.

The Dolphin-2.9.4-Llama3.1 model with eight billion parameters, which is based on the Llama-3 architecture, served as the technical foundation. It was selected after several test runs and comparison with other models of different sizes on the basis of its performance, the basic knowledge available in the field of architecture/planning and its trainability on the given hardware.

| Requirement | llama3-gradient:70b | dolphin2.9-llama3:70b | dolphin2.9-llama3.1:70b | dolphin2.9.4-llama3.1:8b | qwen-2:16b |
|---|---------------------|-----------------------|---|--------------------------|------------|
| Works without hallucinations | • | • • • • | • • • • | • • • | • • • • |
| Can be run on the given Pc | • | • • • • | • • • • | • • • • | • • • • |
| Can be finetuned with the given PC | • | • | • | • • • • | • • • • |
| Knows what IFC and Python is | • | • • • | • • • | • | • • |
| Can read and analyse IFC files | • | • • • | • • • | • | • |
| Can write Python code for Rhino | • | • • • • | • • • • | • | • • |
| Can write Grasshopper code for simple geometries | • | • • • • | • • • | • | • |
| Gives fast answers | • • • | • • | • • | • • • | • • • • |
| Has good base knowledge | • • | • • • | • • • • | • • | • • • |
| Is censored | • • • • | • | • | • | • • • • |
| | | | | | |
| Requirement | qwen2:72b | deepseek-v2:16b | deepseek-v2:236b | gemma2:27b | |
| Works without hallucinations | • • | • • • | • | • | |
| Can be run on the given Pc | • • • | • • • • | • | • | |
| Can be finetuned with the given PC | • | • • • • | • | • | |
| Knows what IFC and Python is | • | • | • | • | |
| Can read and analyse IFC files | • | • | • | • • • | |
| Can write Python code for Rhino | • | • • • | • | • • | |
| Can write Grasshopper code for simple geometries | • | • | • | • | |
| Gives fast answers | • • | • | • | • | |
| Has good base knowledge | • • | • • • | • | • • | |
| Is censored | • • • • | • • • • | • • • • | • • • • | |
| | | | | | |
| Explanation of symbols | | | | | |
| • • • • = Works constantly and without errors | | | • • • = Works partially/ gives results with slight errors | | |
| • • = Works poorly/ gives few correct answers with serious errors | | | • = Doesn't work | | |

Figure 2. Testing parameters for the base-AI-model.

Initial internal tests showed that the model already had basic knowledge of CAD systems, the Rhino/Grasshopper scripting environment and Building Information Modelling (BIM). Structural information about so-called ‘Industry Foundation Classes’ (IFC) - an open standard for describing building data - could also be identified in simple queries. However, this content in the model was incomplete, partially inconsistent and not didactically organised.

In order to specifically address these gaps, the existing knowledge areas were systematically organised and prepared as structured knowledge components. These include object-based classifications, functional relationships between components, parametric dependencies in the design process and typical data formats such as JSON, which are used in many digital planning processes. The aim was to create a semantically dense, technically consistent and linguistically interpretable database on which the model could be further trained.

Didactically motivated content, such as explanatory texts, normative regulations or planning manuals, was deliberately not integrated into the model itself. Instead, it was planned to integrate these into the response process via a separate RAG system as required. In this way, the model was to remain reactive and flexible without being overloaded with static factual knowledge.

2.3 Data collection and preparation

The creation of the training data followed an iterative process based on a systematic analysis of the model's weaknesses. At the beginning, specific test queries were formulated to identify typical conceptual, comprehension and application gaps in the basic model. These included questions about the meaning of individual IFC classes, the functionality of parametric control logic or the structure of digital design structures in Rhino/Grasshopper, as well as general questions about BIM. The results of this gap analysis served as the basis for the selection and preparation of suitable training content. The data used for the fine-tuning was manually researched and curated from various sources - including specialist literature, online tutorials, software documentation and our own research results. All content was then converted into dialogue-based question-and-answer formats with the help of GPT-4. Particular attention was paid to technical precision, linguistic clarity and consistency of

content. The final structure followed the JSONL format, which allows easy integration into common training pipelines. A typical data set consisted of a human-formulated prompt (Human) and a model-generated response (Assistant), for example to explain component relationships or design logic.

```
{"ID": "...", "Human": "...", "Assistant": "..."} 
```

The following example from the training dataset illustrates the type of questions asked and the quality of the answers generated:

```
{"ID": "146", "Human": "What is Grasshopper and how is it used in architecture?", "Assistant": "Grasshopper is a graphical algorithm editor tool that is closely integrated with Rhino 3D, a leading 3D modeling software. It allows users to create complex shapes and structures through visual programming without needing to write direct code. In architecture, Grasshopper is used to support parametric and generative design, allowing architects and designers to create complex geometric forms based on variable parameters. This is often used for creating organic shapes, facade designs, and optimizing building structures in relation to environmental factors like sunlight and wind. Grasshopper also supports iterative design and facilitates collaboration by enabling easy sharing and modification of designs."} 
```

In preparation for integration into the subsequent retrieval system, all data was also divided into sections ('chunks') of 2000 characters each, with an overlap of 300 characters. This segmentation ensured that semantic relationships were not lost and that the content could be efficiently embedded in vector spaces. The BAAI/bge-m3 model was used for semantic embedding and the BAAI/bge-reranker-v2-m3 model for re-ranking the documents in the RAG system - both open source models with proven performance in multilingual retrieval tasks.

2.4 Model Development and Integration

As part of the model development, two central model series were built and tested: Mini-Spyra and IwI-Spyra. Both models are based on different initial architectures and were customised for specific application scenarios.

Mini-Spyra was trained in two development cycles. Version 1 (Mini-Spyra-v1) was fine-tuned and quantised (Q8_0) on the basis of the curated Q&A data in order to enable model execution on locally limited hardware. Due to the low memory profile, the model could be run on a workstation with 96 GB GPU RAM in up to ten parallel instances. This version was particularly suitable for consulting tasks in which complex planning issues with higher context requirements are dealt with.

Version 2 (Mini-Spyra-v2) had a different focus: the response time of the model was significantly improved by specifically adapting the training strategy and reducing the depth of content. This version was particularly suitable for UI-based expert systems or simple dialogue queries with low latency.

The IwI-Spyra model is based on a larger architecture (Qwen/QwQ 32B) and was trained using a LoRA adapter (Low-Rank Adaptation). It was specially developed for the semantic analysis and structured interpretation of IFC files - i.e. for the automatic evaluation of complex building data. The special feature is that IwI-Spyra is able to extract geometric and functional relationships directly from the source text of IFC files and categorise them in context. This makes this model particularly suitable for dialogue-oriented analysis processes, such as the iterative evaluation of planning documents or automated model checking.

All models were run locally in a virtualised system environment (Docker). The hardware used consisted of two NVIDIA A6000 GPUs (96 GB VRAM in total), 256 GB RAM and 2 TB SSD storage for storing the knowledge data and embedding indices. The modular architecture made it possible to run several models in parallel, dynamically load new versions and test different configurations in comparison.

2.5 Implementation of Retrieval-Augmented-Generation (RAG)

A central goal of the development of the Mini-Spyra system was to go beyond static model knowledge and dynamically integrate specialised content into the model answers. A Retrieval Augmented Generation (RAG) approach was implemented for this purpose. This two-stage process was deliberately chosen because it enables a methodologically sound balance between model autonomy and content referencing.

Compared to purely generative architectures, RAG offers the advantage that knowledge content does not have to be fully anchored in the model in advance, but can be expanded and retrieved at runtime. This reduces the risk of hallucinations and increases technical precision, especially in areas with clearly defined terminology such as architecture, BIM or the interpretation of standards.

Alternative approaches such as classic fine-tuning processes with firmly integrated documents or in-context learning using prompt engineering were rejected in this use case for several reasons: Fine-tuning is memory-intensive and makes subsequent updates to the knowledge base more difficult. In-context learning, on the other hand, reaches the token limits of the models with longer text passages or complex standard references and remains limited in the depth of response.

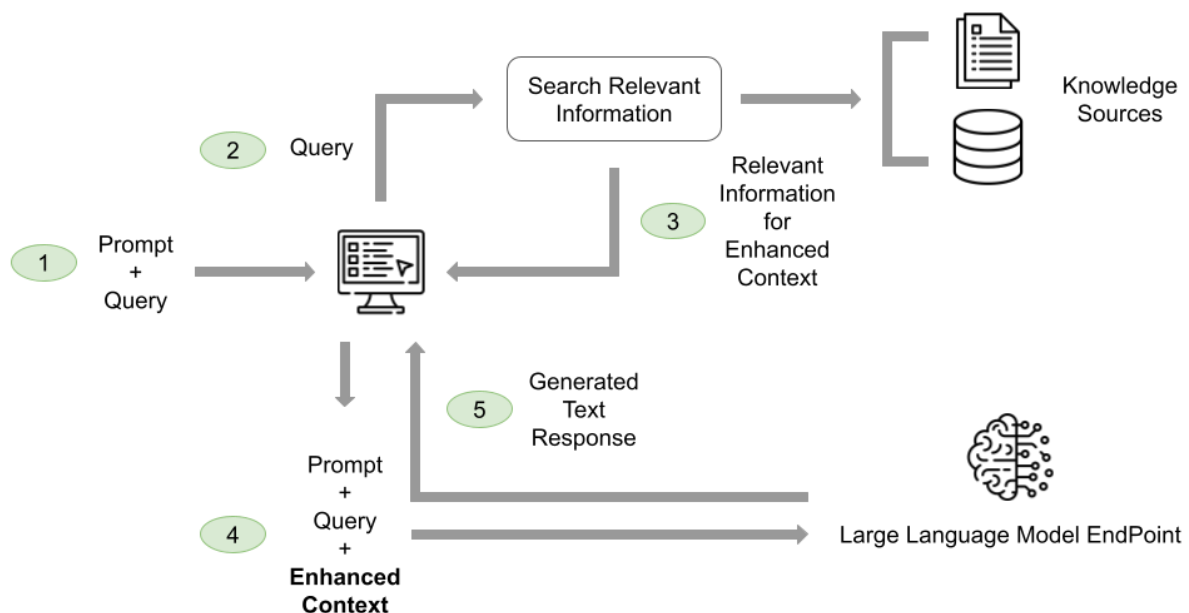


Figure 3. What is Retrieval-Augmented Generation?

RAG combines two processes: Firstly, a semantically matching text passage is searched from a locally stored knowledge database ('AUFLADEN Knowledge') for the user enquiry. This vector search is based on a pre-trained embedding model (BAAI/bge-m3), which converts texts into numerical representations and thus enables a comparison of content regardless of the specific choice of words. The system uses both a semantic and a classic search method (BM25) and selects the most relevant text excerpts.

In a second step, a so-called re-ranking process is applied. The content initially found is compared again with the original query in order to assess its actual relevance in context. The BAAI/bge-reranker-v2-m3 cross-encoder is used for this purpose, which analyses the query and the text section together. The model then returns a weighting of the search results so that only the most relevant parts of the text are included in the final response of the language model.

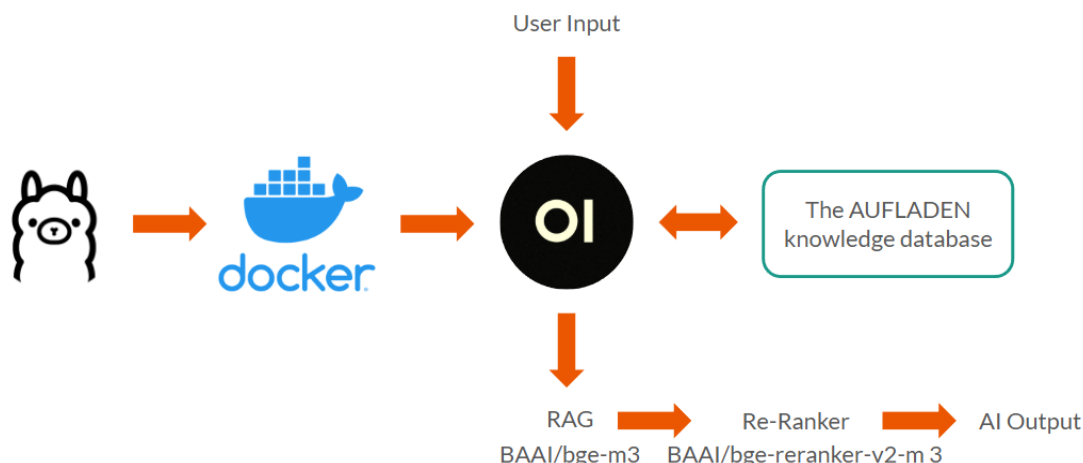


Figure 4. Setting up the infrastructure.

This system architecture allows content such as standards, didactic explanations, tutorials or specific subject definitions to be dynamically integrated into the dialogue process without anchoring them in the language model itself. A major advantage is traceability: the text sources used can be output transparently by the system, which is particularly valuable in didactic contexts or for examination-relevant applications.

The entire architecture can be run locally and has been configured in such a way that both performance and flexibility are guaranteed. Thanks to the modular organisation of ‘AUFLADEN’ knowledge, new content can be integrated quickly and outdated information can be replaced in a targeted manner. The result is a system that not only provides static knowledge, but can also react to changes in teaching and practice in a lively manner.

3. Results

3.1 Application of Domain-Specific Language Models

The developed language models Mini-Spyra-v1, Mini-Spyra-v2 and IwI-Spyra were tested in real application scenarios, which were taken from typical task fields in architecture, BIM modelling and plan analysis. The aim was to evaluate the domain-specific suitability, semantic precision and application relevance of the models under realistic conditions - especially with regard to specialised contexts with a high degree of information density (cf. Hanke et al., 2023; Heins & Grunwald, 2024). The test environment followed a structured prompt design based on practical application scenarios. The models were used specifically for queries that required semantic understanding, structural categorisation and technical accuracy. Mini-Spyra-v1 was tested in document-based knowledge contexts and worked on the basis of the ‘AUFLADEN’ database. The model used a two-stage retrieval process with RAG architecture (cf. Chen et al., 2024) to provide targeted and contextualised answers. For example, it provided a structured explanation to the question about ISO 12006-3:

User:

bim_glossar.pdf

Tell me what ISO 12006-3 is about?

Mini-Spyra-v1:

ISO 12006-3 is about the standard for building information modeling (BIM), which provides methods and guidelines for creating comprehensive BIM data from traditional construction plans, sections, and elevations. It defines how to create detailed BIM models from these drawings. This ensures consistency in terminology and definition across different software platforms. The standard also

covers relationships between objects, properties, metadata, and spatial dimensions. It is widely used in the building industry for creating precise and accurate BIM data from traditional blueprints.

Semantically neighbouring standards such as ISO 23386 were also clearly delineated thematically and correctly explained in terms of content:

User:

bim_glossar.pdf
and ISO 23386?

Mini-Spyra-0.3.4.3:

ISO 23386 is focused on the concept of 'Building Information Modeling' (BIM) within the construction industry. It provides a conceptual framework to define and organize BIM concepts, including the components that make up a building or facility. The standard includes relationships between these concepts and defines what constitutes valid BIM data. This ensures consistency in terminology and understanding across different software platforms. Overall, ISO 23386 is essential for effective communication during construction projects because it provides a common language for discussing complex BIM concepts. It is widely used in the construction industry as an integral component of comprehensive BIM implementation processes.

Mini-Spyra-v2, an optimised variant with a focus on response speed, was used for the interpretation of legal regulations, e.g. §19 of the German Building Code (Baugesetzbuch, 2024).

Building Code *) (BauGB) § 19 Division of plots of land (Federal Ministry of Justice, n.d.)

(1) The division of a plot of land is the declaration made by the owner to the land registry or otherwise made recognisable that a part of the plot of land is to be written off in the land register and entered as an independent plot of land or as a plot of land together with other plots of land or with parts of other plots of land.

(2) The division of a plot of land within the scope of a development plan may not give rise to any circumstances that contradict the provisions of the development plan.

In a specific case study on Section 19 BauGB, the model was able to recognise context, correctly interpret the legal source and generate a comprehensible, structured response. This shows the potential of AI-supported systems in planning-related advice.

User:

BauGB
Collection

Pls tell me what § 19 Teilung von Grundstücken means?

Mini-Spyra-v2:

§ 19 Teilung von Grundstücken (Art. 5 Abs. 1 Satz 2, Art. 10 Nr. 8 und Art. 12 Nr. 3, § 16) - This section outlines the process for dividing a property into several parts. It includes rules for the creation of legal entities to hold specific shares in the divided property and ensures that all parties are represented fairly when new ownership structures are established.

BauGB.pdf

The IwI-Spyra model was designed in versions 0.1 and 0.2 for the interpretation of real IFC files. In version 0.1, it was already able to recognise structural elements such as walls and windows, classify them semantically and relate them to their spatial relationships:

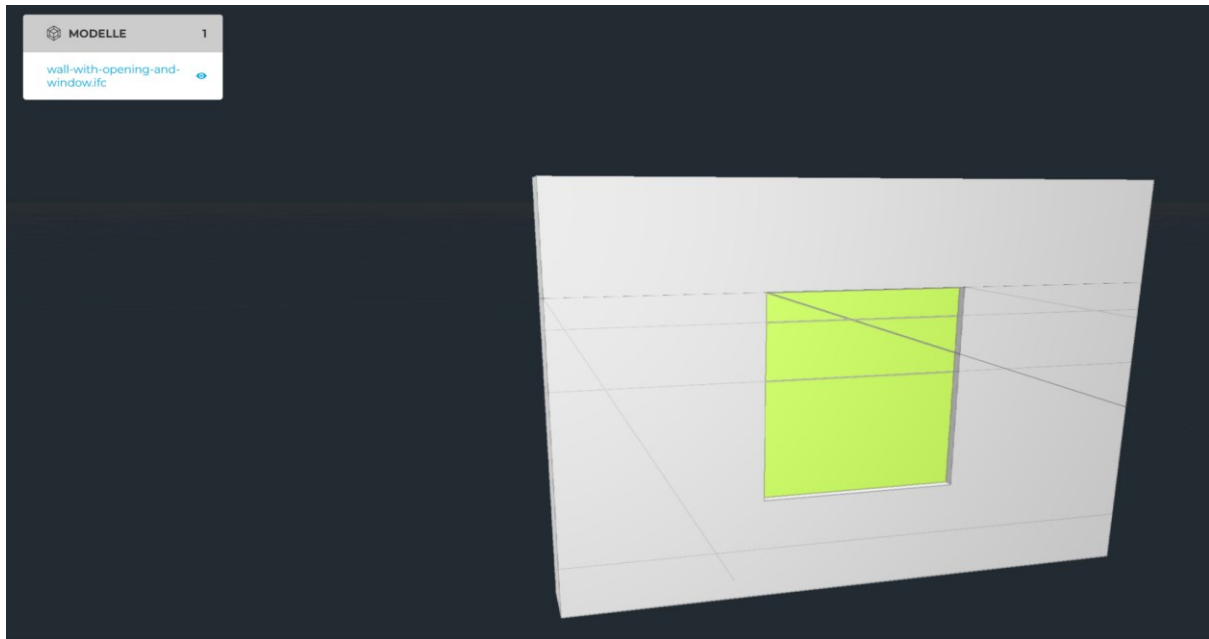


Figure 5. Wall with opening - IFC File.

User:

wall-with-opening-window.ifc

Please explain to me what can be seen in the file.

IwI-Spyra-V0.1: (Thought for a minute)

The file is an IFC model (IFC4) and contains the following main elements:

Overall project structure: *project, building site, building and storey with spatial hierarchy.*

Wall: *With material description (layering of one material), geometric representation (3000mm x 2000mm) and properties such as heat transfer value (0.24 W/(m²K)), external division.*

Window: *Inserted in the wall as an opening, with geometric shape (1000 x 200 mm), material components (glass, wood) and properties such as flow rate (0.3 m³/s).*

Units of measurement: *Length in millimetres, angle in degrees, temperature in °C.*

Spatial relationships: *Placement in the coordinate system, aggregation of building parts and spatial assignment.*

The model follows the model view definition "ReferenceView_V1.2" and contains geometric representation forms for axes (cutting) and bodies (3D extrusion).

wall-with-opening-and-window.ifc

The enhanced version 0.2 was able to perform precise geometric calculations, such as the volume of a wall element based on extrusion data. Material-related information was also correctly extracted from the IFC MaterialLayerSet:

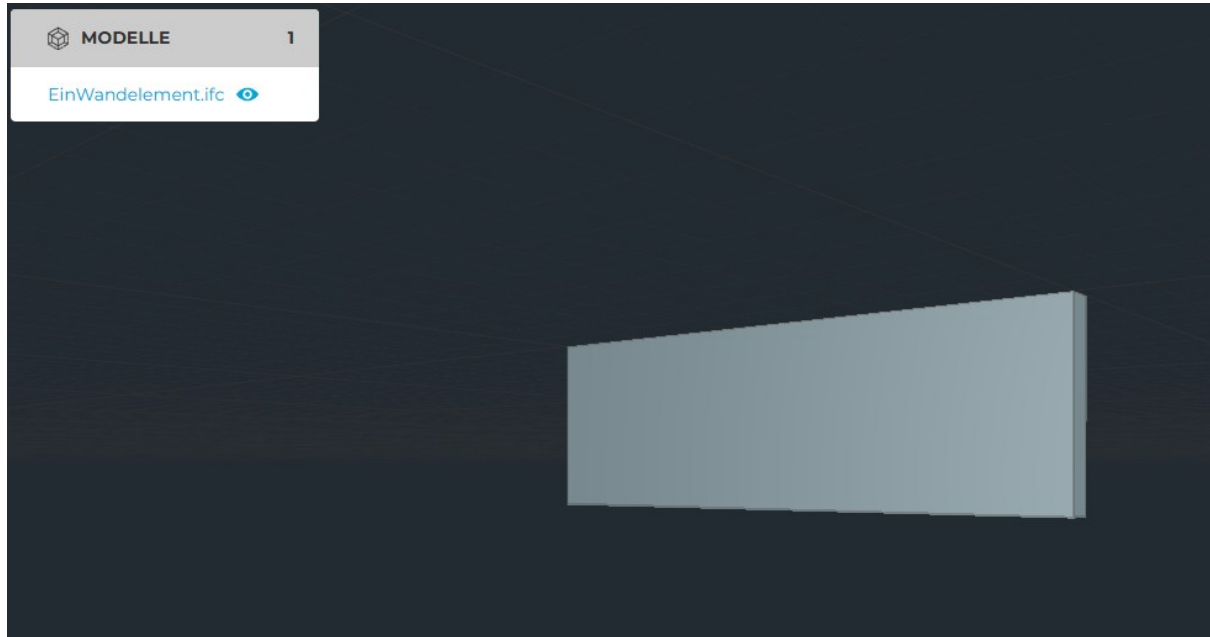


Figure 6. Wall Element- IFC File.

User:

EinWandelement.ifc

This file contains a wall element. Could you please calculate the volume of this element?

Iwl-Spyra-V0.2: (Thought for a minute)

The volume of the wall element is 13.8 cubic meters.

It was calculated from: Profile area (Rectangular area): $10\text{ m} \times 0.46\text{ m} = 4.6\text{ m}^2$ Extrusion height: 3 m Volume: $4.6\text{ m}^2 \times 3\text{ m} = 13.8\text{ m}^3$ The values come from the dimensions of the wall profiles and the extrusion length in the IFC data set.

EinWandelement.ifc

User:

What is the wall made of?

Iwl-Spyra-V0.2: (Thought for 17 seconds)

The wall consists of two material layers Brickwork with a thickness of 30 cm Hard thin screed with a thickness of 16 cm The layer breakdown is: Brickwork (30 cm) + Hard thin screed (16 cm) = Total thickness 46 cm. The material information comes from the defined layer sets (MaterialLayerSet) in the IFC data set.

EinWandelement.ifc

The chosen test strategy makes it possible to identify the strengths and limitations of individual model architectures and to further develop them in a targeted and application-specific manner. The combination of generative language processing with structured knowledge integration represents a methodologically sound alternative to purely generative or retrieval-based systems (cf. Zhou et al., 2025; Reinemann & Watanabe, 2024). The results show that adaptive fine-tuning and structured prompting can also be used to realise powerful, subject-specific assistance systems at a local hardware level.

3.2 Evaluation of Functional Capabilities

The analysed results can be assigned to three main functions:

1. contextual knowledge processing

Mini-Spyra-v1 demonstrated the ability to extract content from extensive collections of knowledge in a contextualised manner and convert it into linguistically comprehensible, thematically focused responses. This applies to both explicit knowledge (e.g. from standard works) and implicit knowledge from heterogeneous text collections. The two-stage RAG system increased response reliability and significantly reduced hallucinations.

2. generation of action-related instructions

Another area of application was the automatic derivation of step-by-step instructions. Mini-Spyra was able to generate precise and comprehensible action steps for enquiries about software usage (e.g. *'How do I create a new layer in Photoshop?'*). The answers contained both functional explanations and optional alternatives - a key feature for didactic application scenarios.

3. analysis of structured building data (IFC)

IwI-Spyra stood out for its ability to analyse IFC files. The model was able to recognise not only structural objects (walls, windows), but also their geometric characteristics, material layers and semantic relationships. In version 0.2, calculations (e.g. volumes) based on extrusion profiles were also possible. Multi-layer material definitions were also correctly identified and named - a step towards automated model checking in the BIM process.

3.3 Summary of Key Findings

The tests and application scenarios carried out show that the chosen methodological approach - consisting of locally executed, specialised language models, enriched by a Retrieval Augmented Generation (RAG) system - is fundamentally suitable for structuring complex specialist knowledge from the field of architecture and digital planning in a machine-readable way, preparing it didactically and making it available in a context-appropriate manner.

Compared to general basic models such as Dolphin-Llama3 , Mini-Spyra-v1 in particular showed a significant improvement in response quality in architecture-specific contexts. The ability to correctly categorise technical terms such as `IfcWallStandardCase`, `MaterialLayerSet` or parametric control, to explain them in a linguistically coherent manner and to translate them into instructions suitable for use represents a clear advance over general models. The central objective - the reduction of semantic gaps and technical vagueness - was demonstrably achieved here.

The use of the RAG system also proved to be a powerful addition: the answers from Mini-Spyra-v1 gained significantly in depth, referenceability and comprehensibility thanks to the structured integration of external knowledge sources. This is particularly important in teaching or for verification documentation in the planning process, as the source and context of the answer play a role in addition to the answer itself.

Mini-Spyra v2, on the other hand, proved to be an efficient tool for reactive systems with low latency requirements - for example in graphical user interfaces or as an input amplifier for parameter inputs. The reduction of content depth in favour of speed makes this model particularly interesting where formal correctness is sufficient but no content discourse is necessary.

Particularly noteworthy is IwI-Spyra-v0.2, which was the first model in the series to be able not only to semantically analyse IFC files, but also to correctly extract and linguistically classify complex relationships such as component relationships, material hierarchies and geometric dependencies. This

step marks a decisive threshold in further development: from purely text-based knowledge generation to structured data interpretation. This opens up new possibilities for automated inspection processes, dialogue-supported model analyses and intelligent assistance systems in BIM-based workflows.

To summarise, the combination of a fine-tuned language model, structured knowledge component development and retrieval integration provides a robust foundation for AI-based tools in architectural practice. The results show that these models are not only capable of correctly reproducing technical knowledge, but also of interpreting this knowledge within a typical architectural framework - for example when dealing with material layering, geometry extraction or normative classifications. The next logical step is to scale up this approach for multimodal applications and the further automation of planning logics.

4. Discussion and Policy level implication

The results presented in this thesis show that the development and application of locally executed, domain-specific language models in an architecture-specific context is not only technically feasible, but also functionally profitable. With regard to the research question formulated at the beginning - how domain-specific knowledge can be provided efficiently under hardware constraints and structured data formats such as IFC can be analysed - the study provides concrete indications of key success factors, system-related limitations and future development directions.

4.1 Findings in relation to the research question

Firstly, it was shown that even models with a manageable number of parameters (e.g. Mini-Spyra-v1 with 8B) are capable of providing highly specialised knowledge in a correct, context-related and linguistically comprehensible manner - provided that the training data used is precisely structured and tailored to the target application. The use of so-called 'core knowledge components' proved to be methodologically viable: The targeted selection and curation of architecture-specific terms, processes and structures laid the semantic foundation for subsequent knowledge processing.

The integration of a retrieval augmented generation system made a decisive contribution to the flexibilisation and contextualisation of the model answers. Particularly in the case of normative, explanatory or didactically sensitive content, the system was able to correctly allocate relevant sources, selectively integrate them and make them usable for the response process. This is an aspect that is of central importance in the educational and examination environment.

With IwI-Spyra v0.2, it also became clear that structured, non-natural language data such as IFC files can be interpreted semantically using specialised models, including complex geometry analyses, material evaluations and component relationships. A central objective of the research question - the automatic analysis of structured planning data - was thus successfully realised in a first step.

4.2 Implications for teaching, practice and model development

The results suggest that domain-specific AI models in architecture can address three areas in particular:

Didactic support: comprehensible, understandable and adaptable explanation formats make it easier to convey content such as standard terms, software functions or model structures. Mini-Spyra-v1 offers a powerful basis for providing students with context-sensitive specialist knowledge, for example - without cloud constraints or technical barriers.

Planning support and standards-based consulting: Mini-Spyra-v2 shows potential for fast, formally correct feedback in UI systems or digital design tools - for example for parameter checks, plausibility

analyses or the interpretation of regulations. The combination with RAG enables situational contextualisation.

Automated data analysis: Particularly relevant for the future is the further development of models such as IwI-Spyra, which can read, interpret and evaluate not only texts but also structured formats such as IFC. This opens up prospects for AI-based model checking, dialogue-based BIM analysis and automated design support.

At the same time, the separation of model core (language competence) and external knowledge (RAG) not only makes sense from a technical point of view, but also allows for better maintainability, transparency and updatability - aspects that are of crucial importance in dynamic fields of application such as architecture.

4.3 Limitations of the study

Despite the promising results, the work is not without its limitations. In particular, the following points need to be critically considered:

Data depth and breadth: the quality of the models depends heavily on the selection and preparation of the training data. Even if great care was taken with the CKCs, the database remains limited - both in terms of the variety of topics and multilingual content or special cases of planning practice.

Hardware dependency: Even if all models can be executed locally, the fine-tuning of the models in particular, but also the use of models such as IwI-Spyra, requires considerable computing resources. This may limit the immediate applicability for smaller offices or educational institutions with limited equipment.

Limitations due to the prompt format: The dialogical structure of the training data allows a realistic application, but in some cases leads to redundant or linguistically stilted output, even if this is rather low - especially with Mini-Spyra v2, which has been optimised for speed.

Generalisability: The concrete use cases tested are based on specific, curated scenarios. Generalisation to any planning, design or analysis context is not easily possible and requires further scaling and validation.

4.4 Follow-up questions and further development

The study raises several follow-up questions that should be addressed in future work:

How can multimodal content (e.g. plans, sketches, models) be integrated?

How can a continuous feedback loop be established between user, model and knowledge base?

Which quality metrics are suitable for evaluating AI responses in normative contexts?

How can such systems be integrated into didactic settings - for example as tutors, assistants or examination tools?

Another logical step would be to combine text-based analysis capabilities with visual plan interpretation, i.e. an expansion towards multimodal AI systems that can process 3D models and plan layouts simultaneously, for example. Integration into existing software solutions, such as CAD/BIM programmes or learning platforms, would also be conceivable.

5. Conclusion

This study explored the development, extension, and integration of specialised language models into practical architectural workflows under local hardware constraints. The central research question guiding this investigation was: *How can domain-specific language models be developed and extended under local hardware conditions in order to provide architecture-specific knowledge in a context-appropriate manner and semantically analyse structured planning data (e.g. IFC)?* The systems created—namely Mini-Spyra (versions 1 and 2) and IwI-Spyra—demonstrated that it is both technically feasible and contextually advantageous to customise large language models (LLMs) for architectural planning and education. By implementing targeted fine-tuning processes, structured knowledge modules, and retrieval-augmented generation (RAG) systems, significant improvements in semantic precision, technical accuracy, and the overall explainability of the models were achieved. The application scenarios, ranging from the interpretation of standards and software documentation to the semantic analysis of complex Industry Foundation Classes (IFC) files, further underscored the models' ability not only to reproduce information but also to contextualise and articulate it in linguistically clear and pedagogically meaningful ways. Notably, IwI-Spyra-v0.2 was capable of precisely identifying material layers, geometric elements, and semantic relationships within IFC data—a substantial advancement toward automated model checking in architectural practice.

Moreover, the research established that advanced LLMs can be effectively deployed and enhanced under local hardware conditions. Leveraging open-source architectures such as LLaMA 3 and Qwen/QwQ, combined with efficient quantisation techniques and RAG strategies, the study illustrates the viability of decentralised, data-sovereign AI infrastructures—an especially critical consideration in academic and institutional environments.

Scientific contribution and implications

The study offers a concrete methodological framework for developing, training, and applying domain-specific LLMs tailored to the architectural domain. It provides a compelling example of how AI-driven tools can be integrated into the digital transformation of architectural education and professional planning—serving both as didactic aids and as technical assistance systems. For research institutions, the study introduces a replicable model for building specialised AI systems in a transparent, modular, and locally operable manner, eliminating reliance on opaque cloud-based solutions. In practical terms, this enables AI-supported plan analysis, documentation, and rule-based validation while retaining control over both the data and the underlying infrastructure.

Limitations and future work

Despite the promising outcomes, this research should be regarded as a proof of concept. Several components—such as the selection of knowledge sources and the definition of query strategies—still rely on manual procedures. Additionally, the experimental scope was limited to selected use cases and illustrative queries, without yet incorporating comprehensive external evaluations. Feedback from real-world user groups, including students and professional planners, remains an essential next step.

Future research should prioritise several key areas: (1) the scaling and automation of model training and data preparation, potentially using active learning or semi-supervised techniques; (2) multimodal expansion of the models to accommodate visual and CAD-based inputs; (3) structured user testing to refine didactic integration and practical planning support; and (4) enhancing source attribution within the AI outputs to ensure transparency and interpretability. Overall, the findings affirm that domain-specific AI systems are not only technically viable but also practically impactful—provided they are developed with well-defined goals, robust knowledge structures, and transparent, locally operated architectures. The intersection of domain specialisation, local deployment, and dynamic RAG-based augmentation presents a sustainable trajectory for the next generation of AI-supported architectural planning tools.

Acknowledgements

The authors would like to extend their appreciation to the reviewers of the Journal of Smart Design Policies for their insightful and constructive comments, which have significantly enhanced the quality of this article. Furthermore, AI tools such as Chat-GPT 4o and Gemini 2.5 Pro were employed to refine grammar and enhance overall fluency.

Funding

This research received no external funding.

Conflicts of Interest

The Authors declare that there is no conflict of interest.

Data availability statement

The data that support the findings of this study are available from the corresponding author, N.A., upon reasonable request.

Institutional Review Board Statement

Not applicable.

CRedit author statement:

Conceptualization: N.A., Y.H., G.G.; Data curation: N.A., Y.H.; Formal analysis: N.A., Y.H.; Investigation: N.A., Y.H.; Methodology: N.A., Y.H.; Project administration: G.G.; Software: N.A.; Supervision: G.G.; Validation: G.G.; Visualization: Y.H.; Writing - original draft: N.A., Y.H.; Writing - review & editing: N.A., Y.H., G.G.. All authors have read and agreed to the published version of the manuscript.

References

- Ansre, N., Hirsekorn, Y., & Grunwald, G. (2025). Wissensmanagement 2.0 – die AUFLADEN KI für eine neue Dimension der Wissensvermittlung. In T. Luhmann & T. Sieberth (Eds.), *Photogrammetrie – Laserscanning – Optische 3D-Messtechnik: Beiträge der Oldenburger 3D-Tage und des BIMtags 2025* (pp. 282–291). Wichmann Verlag.
- Baugesetzbuch: Mit Immobilienwertermittlungsverordnung, Baunutzungsverordnung, Planzeichenverordnung, Raumordnungsgesetz, Raumordnungsverordnung: Textausgabe mit ausführlichem Sachverzeichnis und einer Einführung.* (2024). C. H. Beck.
- Calcagno, G., Alves, S., & Grunwald, G. (2024). Assessing the quality of an innovative learning path for BIM education: The DIGITAL DECATHLON. *Journal of Civil Construction and Environmental Engineering*, 9(5), 143–150. <https://doi.org/10.11648/j.jccee.20240905.11>
- Calcagno, G., Bertelli, M., & Grunwald, G. (2024). The Digital Decathlon: A journey in building information modelling education. *Journal of Mediterranean Cities*, 4(1). https://doi.org/10.38027/mediterranean-cities_vol4no1_3
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (ArXiv Preprint No. arXiv:2402.03216). *arXiv*. <https://doi.org/10.48550/arXiv.2402.03216>
- Damen, T., Sebastian, R., MacDonald, M., Soetanto, D., Hartmann, T., Di Giulio, R., Bonsma, P., & Luig, K. (2015). The application of BIM as collaborative design technology for collective self-organised housing. *International Journal of 3-D Information Modeling*, 4(1), 1–18. <https://doi.org/10.4018/ij3dim.2015010101>
- Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., ... Muennighoff, N. (2025). MMTEB: Massive multilingual text embedding benchmark (ArXiv Preprint No. arXiv:2502.13595). *arXiv*. <https://doi.org/10.48550/arXiv.2502.13595>

- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024, 31 July). The Llama 3 herd of models (ArXiv Preprint No. arXiv:2407.21783). *arXiv*. <https://arxiv.org/abs/2407.21783>
- Groenmeyer, L., & Grunwald, G. (2024). Integration of AI and LLMs for enhanced BIM analysis: A study on IFC model interpretation. In *AI-Driven Architecture: Pioneering the Digital Frontier* (1st ed., pp. 117–144). Alanya University Publication. <https://doi.org/10.38027/ai-driven-6>
- Grunwald, G., & Heins, C. (2023). BIM Game: A testing ground for specifying, modelling, evaluating and visualising information in IFC formats. In *Lecture Notes in Civil Engineering* (pp. 677–688). https://doi.org/10.1007/978-981-99-4049-3_52
- Gregor, G., Sharina, A., Matteo, B., Gisella, C., Ireneusz, C., Emilia, D., ... Zuber, L. (2025). Building BIM competence: Learning in the DIGITAL DECATHLON. In *Innovative Renewable Energy* (pp. 981–997). https://doi.org/10.1007/978-3-031-82323-7_77
- Hajirasouli, A., & Banihashemi, S. (2022). Augmented reality in architecture and construction education: State of the field and opportunities. *International Journal of Educational Technology in Higher Education*, 19(1), Article 61. <https://doi.org/10.1186/s41239-022-00343-9>
- Hanke, T. (2024). AUFLADEN – The web portal for self-study in the field of digital planning and construction. In T. Luhmann & T. Sieberth (Eds.), *Photogrammetry, Laser Scanning, Optical 3D Metrology – Contributions to the Oldenburg 3D Days and the BIM Day 2024*. VDE Verlag.
- Hanke, T., & Grunwald, G. (2024). Artificial intelligence as a design tool in the architectural design process. In *AI-Driven Architecture: Pioneering the Digital Frontier* (1st ed., pp. 53–71). Alanya University Publication. <https://doi.org/10.38027/ai-driven-3>
- Hanke, T., Kawasaki, J. Y., & Grunwald, G. (2023). Manufacturing processes of complex shapes and structures using 3D printing and augmented reality. *Proceedings of the International Conference of Contemporary Affairs in Architecture and Urbanism (ICCAUA)*, 6(1), 62–66. <https://doi.org/10.38027/iccaua2023en0134>
- Heins, C., & Grunwald, G. (2024). BIM and IPA – Excerpt of an automated assessment system for an autodidactic teaching concept. In *Proceedings of the 41st International Symposium on Automation and Robotics in Construction (ISARC 2024)* (pp. 830–837). <https://doi.org/10.22260/ISARC2024/0001>
- Heins, C., Grunwald, G., & Helmus, M. (2021). Gamification and BIM – The didactic guidance of decentralised interactions of a real-life BIM business game for higher education. In *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC 2021)* (pp. 932–939). <https://doi.org/10.22260/ISARC2021/0126>
- Kawasaki, J., & Grunwald, G. (2023). Enhanced learning and teaching through AR: Case studies on design-build projects. *JSEE Annual Conference International Session Proceedings, 2023*, 18–19. https://doi.org/10.20549/jseer.2023.0_18
- Kawasaki, J. Y., Hirsekorn, Y., Ansre, N., & Grunwald, G. (2024). AUFLADEN LAB – Parametric AI-supported design. In T. Luhmann & T. Sieberth (Eds.), *Photogrammetry, Laser Scanning, Optical 3D Metrology – Contributions to the Oldenburg 3D Days and the BIM Day 2024*. VDE Verlag.
- Leite, R. M. C., Winkler, I., & Alves, L. R. G. (2022). Visual management and gamification: An innovation for disseminating information about production to construction professionals. *Applied Sciences*, 12(11), 5682. <https://doi.org/10.3390/app12115682>
- Michail, A., Clematide, S., & Sennrich, R. (2025, 12 February). Examining multilingual embedding models cross-lingually through LLM-generated adversarial examples (ArXiv Preprint No. arXiv:2502.08638). *arXiv*. <https://doi.org/10.48550/arXiv.2502.08638>
- Özener, O. Ö. (2023). Context-based learning for BIM: Simulative role-playing games for strategic business implementations. *Smart and Sustainable Built Environment*, 13(4), 908–933. <https://doi.org/10.1108/SASBE-08-2022-0184>

- Patil, A., & Jadon, A. (2025). Advancing reasoning in large language models: Promising methods and approaches (ArXiv Preprint No. arXiv:2502.03671). *arXiv*. <https://doi.org/10.48550/arXiv.2502.03671>
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... Qiu, Z. (2024, 19 December). QWen 2.5 technical report (ArXiv Preprint No. arXiv:2412.15115). *arXiv*. <https://doi.org/10.48550/arXiv.2412.15115>
- Reinmann, G., & Watanabe, A. (2024). KI in der universitären Lehre. In *De Gruyter eBooks* (pp. 29–46). <https://doi.org/10.1515/9783111351490-004>
- Tasnia, T., & Grunwald, G. (2024). Evaluate the outcome of the digital learning platform AUFLADEN. In T. Luhmann & T. Sieberth (Eds.), *Photogrammetry, Laser Scanning, Optical 3D Metrology – Contributions to the Oldenburg 3D Days and the BIM Day 2024*. VDE Verlag.
- Wang, P., Liu, T., Wang, C., Wang, Y., Yan, S., Jia, C., ... Yu, Y. (2025, 10 June). A survey on large language models for mathematical reasoning (ArXiv Preprint No. arXiv:2506.08446). *arXiv*. <https://doi.org/10.48550/arXiv.2506.08446>
- Xu, F., Hao, Q., Zong, Z., Wang, J., Zhang, Y., Wang, J., ... Li, Y. (2025). Towards large reasoning models: A survey of reinforced reasoning with large language models (ArXiv Preprint No. arXiv:2501.09686). *arXiv*. <https://doi.org/10.48550/arXiv.2501.09686>
- Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., ... Zhou, T. (2024). A survey on knowledge distillation of large language models (ArXiv Preprint No. arXiv:2402.13116). *arXiv*. <https://doi.org/10.48550/arXiv.2402.13116>
- Yalçın, Z. Ö. (2024). Integrating gamification with BIM for enhancing participatory design. *Journal of Computational Design*, 5(2), 317–344. <https://doi.org/10.53710/jcode.1505309>
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, 11(12), nwae403. <https://doi.org/10.1093/nsr/nwae403>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. (2023). A survey of large language models (ArXiv Preprint No. arXiv:2303.18223). *arXiv*. <https://doi.org/10.48550/arXiv.2303.18223>
- Zheng, C., Zhang, Z., Zhang, B., Lin, R., Lu, K., Yu, B., ... Lin, J. (2024). ProcessBench: Identifying process errors in mathematical reasoning (ArXiv Preprint No. arXiv:2412.06559). *arXiv*. <https://doi.org/10.48550/arXiv.2412.06559>