Constructing Insights: Leveraging Large Language Models for Information Retrieval in Unstructured Document Collections with ConSight

Research in Progress

Moritz Diener*, Sebastian Schäfer*, Philipp Spitzer, and Michael Vössing

Karlsruhe Institute of Technology (KIT),
Institute for Information Systems (WIN), Karlsruhe, Germany
{moritz.diener, sebastian.schaefer2, philipp.spitzer, michael.voessing}@kit.edu

Abstract. Recent technological advances have led to an increase in the volume of available data that knowledge workers use in their daily routines. Current LLM-based systems face challenges in providing accurate and relevant information in an efficient manner. At the same time, because these large collections of documents are unknown to the user, many knowledge workers struggle to process the data, even when collaborating with an LLM-based system. To address this challenge, we develop a prototype "ConSight" following a Design Science Research methodology. Based on insights from the construction industry, our system integrates information retrieval with multi-modal data support to enhance information access and contextual understanding. Our contribution to the IS domain is twofold: (1) we provide insights into the interplay between LLM-based retrieval systems and human information processing in real-world scenarios, and (2) we take the first steps to derive design knowledge for LLM-based knowledge work support.

Keywords: Large Language Models, Information Retrieval, Knowledge Work, Prototype.

1 Introduction

Effective management and retrieval of data are essential across various domains, enabling professionals to access and analyze information in complex tasks (Powell & Snellman 2004, Choong & Leung 2022). Knowledge workers (KWs) need to process large collections of documents (Taherdoost & Madanchian 2023) to perform actions and conclude reasoning on complex tasks (Pakarinen & Huising 2023). Over the past decades, the development of tools supporting KWs for information management and retrieval has increased (Detlor 2010)—from early knowledge management platforms (Alavi & Leidner 2001) to artificial intelligence(AI)-based information retrieval systems (Kim et al. 2024, Krütli & Hanne 2025, Zhai 2024). Recently, large language models (LLMs) have emerged as a viable means to support KWs in information retrieval, offering the ability to access vast knowledge bases and process large amounts of data (Xu et al. 2024, Alavi et al. 2024, Lewis et al. 2020). For instance, Dell'Acqua et al. (2023) show that using LLMs for knowledge work can increase efficiency and effectiveness.

[*] Both authors contributed equally to this research.

Despite recent advances, KWs still face many challenges. Technological barriers persist as LLM-based systems focus on free-form text and neglect multi-modal content, such as structured tables, screenshots, and process diagrams (Alavi et al. 2024). Additionally, LLMs are constrained by risks of hallucination, outdated training data, limited domain-specific knowledge, and restricted reasoning capabilities (Kaddour et al. 2023). Beyond these technological constraints, real-world application challenges persist due to issues in adopting LLM-based systems (De Vreede et al. 2024). In addition to fields like law (Henderson et al. 2022) or healthcare (Kaddour et al. 2023), the construction industry is characterized by a vast collection of unstructured and semi-structured documents (e.g., contracts, technical specifications, drawings, maintenance logs, and inspection reports) (Nedeljković & Kovačević 2017). As a result, construction professionals such as civil engineers, architects, and project managers face intensive information management challenges, resulting in information overload, especially when KWs lack prior familiarity with the collection and must manually retrieve information (Nedeljković & Kovačević 2017, Arnold et al. 2023, Krütli & Hanne 2025, Duong & Lin 2022).

In this work, we take the first steps towards generating design knowledge by adopting a Design Science Research (DSR) methodology (Hevner et al. 2004) to develop a prototype that bridges the capabilities of LLMs with the practical needs of KWs. By collaborating with KWs from the **Con**struction industry, as a representative context of knowledge-intensive work, we derive design requirements (DRs) and design features (DFs) to generate in**Sight**s in vast document collections (DCs) through the prototype ("ConSight"). ConSight is designed to take unstructured DCs as input and output structured information based on the KWs context by tracking the source, timeliness, and relevance of the information for the KWs. We share first insights into our investigated problem space—knowledge work in the construction industry—through an initial design cycle and instantiate the developed DFs in a prototype.

2 Research Approach

This study follows the DSR methodology (Hevner et al. 2004) and is guided by the BAUSTEIN framework (Schoormann et al. 2024), which offers a configurable set of activities specifically designed for DSR projects involving multiple stakeholders from practice and research. Adopting a problem-solving configuration of BAUSTEIN allows us to integrate practical problem-solving with theoretical development and to navigate this and future design cycles in the scope of the larger-scale DSR project. We conducted this research within an industry-research consortium composed of research institutes and companies from the construction industry focusing on sustainable renovation measures. Our research progress in this iteration is structured into three phases (see Figure 1).

Phase 1: Problem Space & Strategy. As part of an industry-research consortium, we identified the problem of managing vast, unstructured DCs in the construction industry (Nedeljković & Kovačević 2017). We systematically conducted three focus groups involving nine experts from three sectors in the construction domain: Construction and Facility Management, Sustainability Engineering, and Civil Engineering Research. We aimed to explore the problem space and understand the challenges in knowledge work with large, unfamiliar and unstructured DCs. In these focus groups, we conducted semi-

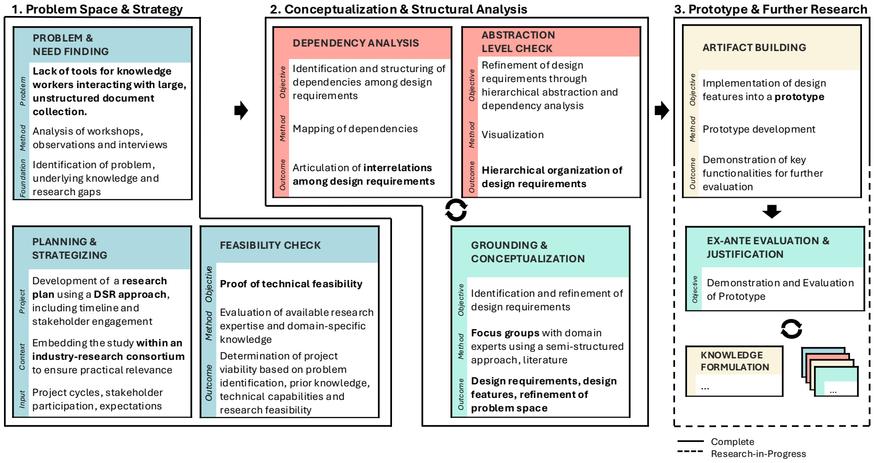

1. Problem Space & Strategy
PROBLEM & NEED FINDING
Problem
Lack of tools for knowledge workers interacting with large, unstructured document collection.
Method
Analysis of workshops, observations and interviews
Foundation
Identification of problem, underlying knowledge and research gaps
PLANNING & STRATEGIZING
Project
Development of a research plan using a DSR approach, including timeline and stakeholder engagement
Context
Embedding the study within an industry-research consortium to ensure practical relevance
Input
Project cycles, stakeholder participation, expectations
FEASIBILITY CHECK
Objective
Proof of technical feasibility
Method
Evaluation of available research expertise and domain-specific knowledge
Outcome
Determination of project viability based on problem identification, prior knowledge, technical capabilities and research feasibility
2. Conceptualization & Structural Analysis
DEPENDENCY ANALYSIS
Objective
Identification and structuring of dependencies among design requirements
Method
Mapping of dependencies
Outcome
Articulation of interrelations among design requirements
ABSTRACTION LEVEL CHECK
Objective
Refinement of design requirements through hierarchical abstraction and dependency analysis
Method
Visualization
Outcome
Hierarchical organization of design requirements
GROUNDING & CONCEPTUALIZATION
Objective
Identification and refinement of design requirements
Method
Focus groups with domain experts using a semi-structured approach, literature
Outcome
Design requirements, design features, refinement of problem space
3. Prototype & Further Research
ARTIFACT BUILDING
Objective
Implementation of design features into a prototype
Method
Prototype development
Outcome
Demonstration of key functionalities for further evaluation
EX-ANTE EVALUATION & JUSTIFICATION
Objective
Demonstration and Evaluation of Prototype
KNOWLEDGE FORMULATION
...
...
Complete
Research-in-Progress

Figure 1. Research approach following the BAUSTEIN framework

structured interviews covering the topics of current workflows, information needs, pain points, and requirements for a technological solution to their problems. See Table 1 for an overview of participants.

Phase 2: Conceptualization & Structural Analysis. The focus group interviews from Phase 1 were recorded, transcribed, and analyzed through systematic inductive coding following Gioia et al. (2013). We employed the methodology's three-tier progression iteratively, moving from in-vivo first-order codes through theory-augmented second-order themes to theoretical aggregate dimensions. Existing literature was primarily integrated during the final stage, resulting in eight DRs. Based on the derived design knowledge, we develop DFs as high-level technical modules with defined functionalities and concepts that are instantiated in our preliminary prototype.

Table 1. Overview of interviewed experts, their professional role, years (yrs) of experience and their interaction with DCs

Group	ID	Role	Experience	Interaction
Group 1 (40:18 min)	E1	Project Manager	8 yrs	Moderate
	E2	Head of Construction	30 yrs	Extensive
	E3	Sustainability and Due Diligence Expert	17 yrs	Extensive
	E4	Digital Process Coordinator	8 yrs	Moderate
Group 2 (41:55 min)	E5	Researcher	2 yrs	Limited
	E6	Architect	3 yrs	Moderate
Group 3 (57:13 min)	E7	Product Manager	2 yrs	Limited
	E8	Civil Engineer and Energy Consultant	3 yrs	Moderate
	E9	Senior Architect	30 yrs	Extensive

Phase 3: Prototype & Further Research. The goal of this first prototype is to explore the solution space and gain further insights into the problem domain through interactive workshops with stakeholders that inform our design knowledge evolution,

with BAUSTEIN framework configurations added for future research. In future workshops, we want to evaluate the prototype qualitatively to revisit and refine the DRs and DFs for further design iterations. With the knowledge gained, we plan to implement a fully functioning prototype. Through qualitative and quantitative evaluations, we then plan on investigating efficiency gains in the form of time savings, technology acceptance and adoption according to Venkatesh et al. (2012). Following Iivari (2015), we aim at developing design principles after iterating and evaluating our prototype.

3 Design Requirements for ConSight

Based on our approach outlined in Section 2, we derived the following eight DRs:

Multi-perspective document organization across disciplines (DR1) emerged as a primary requirement. The KWs emphasized the need for a filtering and categorization system of document content that goes beyond simple file format classification, enabling distinction between document types such as floor plans, energy certificates, CAD models, and construction regulations (Nepal & Staub-French 2016, Seidel et al. 2008). To address this, the prototype should provide search, categorization, and sorting functionalities, allowing users to filter documents based on content-related criteria and relevance.

Key information display based on user's disciplinary perspective (DR2) ensures the automated retrieval and tailored display of essential information based on the user's disciplinary needs (Nepal & Staub-French 2016). It supports topic-specific filtering based on the KW's disciplinary perspective, for instance, displaying energy-related content like energy demand (topic) for energy consultants (discipline). The prototype should automatically extract and prioritize information, ensuring that critical details (e.g., hazardous materials such as asbestos) are retrieved and presented.

Document organization across timeline (DR3) addresses the need for a temporal overview to interpret changes, resolve discrepancies, and understand the complete history of a building. Building documentation should be structured in alignment with the building's life cycle (Rezgui et al. 2013), from initial construction, through renovation periods, to its current state. The prototype should use time-related metadata, such as document creation dates, to visualize the life cycle of a building.

Regulatory and operational compliance schedule (DR4) aims to identify documents with urgent regulatory or operational information. It emphasizes automatic retrieval and displaying of time-sensitive information, such as deadlines from safety inspections, enabling users to prioritize actions according to compliance timelines and maintain operational integrity (Khan et al. 2023, Chen et al. 2024).

Detection and resolution of data inconsistencies and gaps (DR5) emphasizes the integration and cross-referencing of multiple data sources to automatically identify and resolve discrepancies, missing elements or errors (Nepal & Staub-French 2016). By detecting issues, such as misaligned information regarding window insulation material or mismatches between floor plans and cadastral maps, the artifact should enable KWs to pinpoint inconsistencies and fill gaps, thus enhancing the overall reliability and completeness of the documentation.

Dynamic contextual detail access (DR6) mandates that the interface adapts the level of detail to the current context of the KW. It presents a high-level summary (DR1 and DR2) and then progressively reveals deeper, context-sensitive insights (Shneiderman 1996, Amershi et al. 2019), as KW information needs evolve. This adaptive approach ensures users can navigate between broad overviews and granular information.

Traceable source reference and contextual validation (DR7) ensures that KWs can transparently trace retrieved information back to its sources. It emphasizes validating LLM-generated insights by linking each piece of information to its original document context (Huang & Chang 2024). Additionally, the prototype should display specialized file formats (e.g., CAD files), eliminating the need for dedicated software licenses and facilitating the direct inspection of source materials.

Transparent and trustworthy AI interaction (DR8) fosters human-centered interaction with the LLM-based system. It mandates that the interface provides clear, visual feedback—such as processing times and confidence scores—and transparently links information to its sources (Reyes et al. 2025, Amershi et al. 2019). Moreover, the prototype should communicate its limitations by explicitly acknowledging uncertainty.

4 Design and Implementation of the Prototype

To address the DRs, we iteratively developed, evaluated, and refined five DFs instantiated as interconnected modules. Figure 2 provides an overview of the overall architecture of ConSight, while Figure 3 shows two screenshots of the dashboard module.

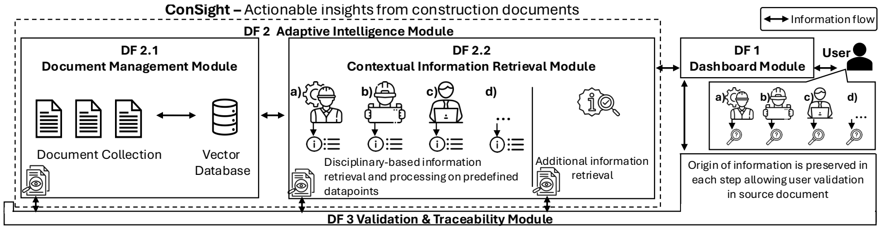

ConSight – Actionable insights from construction documents
Information flow
DF 2 Adaptive Intelligence Module
DF 2.1
Document Management Module
Document Collection
Vector Database
DF 2.2
Contextual Information Retrieval Module
a)
b)
c)
d)
...
Disciplinary-based information retrieval and processing on predefined datapoints
Additional information retrieval
DF 1
Dashboard Module
User
a)
b)
c)
d)
...
Origin of information is preserved in each step allowing user validation in source document
DF 3 Validation & Traceability Module

Figure 2. ConSight Architecture

DF1: Dashboard Module. This module provides the primary user interface, supporting KWs in navigating large DCs by leveraging outputs of DF2.2 and DF3. To avoid information overload (Arnold et al. 2023), it shows key general information, lists relevant data points and groups them by topic and discipline according to DF2.2 (DR2, DR6) while ensuring traceability for all presented information back to the source document (DR8). Documents are organized by topic coverage or along the timeline (DR1, DR3). Prior research demonstrates that dashboard-based summarization and visual interaction enhance accessibility and supports KWs (Matheus et al. 2020, Kus et al. 2022).

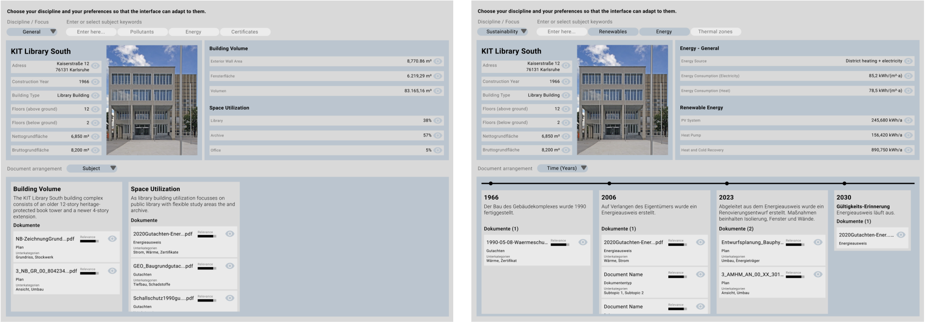
DF2: Adaptive Intelligence Module. This core module uses multiple AI components like LLMs, embedding models (Patil et al. 2023), optical character recognition (OCR) (Thorat et al. 2022) and retrieval-augmented generation (RAG) (Klesel & Wittmann 2025), which enables the prototype to extract information from documents and process them to provide the user with a structured information display.

DF2.1: Document Management Module. This module extracts information from documents such as construction documentation, energy certificates or renovation records.

OCR is applied to textual documents, while multi-modal approaches handle complex elements like diagrams (Ma et al. 2024, Xu et al. 2020). Extracted information chunks are embedded in a vector database to facilitate semantic similarity search.

DF2.2: Contextual Information Retrieval Module. Informed by co-created templates with domain experts, this module maintains discipline-specific data schemata consisting of their respective pre-defined information requests (DR1, DR2, DR4, DR6). For each of the requested data points, relevant text chunks from the vector database are retrieved (DF2.1), and matching information is output, using RAG. This enables fast and structured display of critical information without explicit prompting. Additionally, it allows custom querying to retrieve information not covered by predefined lists, ensuring flexibility and comprehensive support (Schmidt et al. 2025, Ghosh et al. 2024).

DF3: Validation and Traceability Module. This module links all extracted information to its original source, including document ID, page number, and location on page (DR5, DR7). This enables KWs to verify the origin of each data point. Beyond traceability, the module also communicates uncertainty at multiple processing stages—spanning OCR confidence, embedding similarity, and LLM generation uncertainty (Bhatt et al. 2021). While quantifying LLM uncertainty remains a challenge, promising methods are emerging (Yin 2025). Prior research emphasizes that transparency, when combined with confidence indicators, plays a crucial role in cultivating trust in AI-based systems (Reyes et al. 2025, Afroogh et al. 2024).



Choose your discipline and your preferences so that the interface can adapt to them.

Discipline / Focus: General

Enter or select subject keywords: Enter here... | Pollutants | Energy | Certificates

KIT Library South

- Adress: Kaiserstraße 12, 76131 Karlsruhe
- Construction Year: 1966
- Building Type: Library Building
- Floors (above ground): 12
- Floors (below ground): 2
- Nettogrundfläche: 6,850 m²
- Bruttogrundfläche: 8,200 m²

Building Volume

- Exterior Wall Area: 8,770.86 m²
- Fensterfläche: 6.219,29 m²
- Volumen: 83.165,16 m³

Space Utilization

- Library: 38%
- Archive: 57%
- Office: 5%

Document arrangement: Subject

Building Volume

The KIT Library South building complex consists of an older 12-story heritage-protected book tower and a newer 4-story extension.

Dokumente

- NB-ZeichnungGrund...pdf — Plan — Unterkategorien: Grundriss, Stockwerk
- 3_NB_GR_00_804234...pdf — Plan — Unterkategorien: Ansicht, Umbau

Space Utilization

As library building utilization focusses on public library with flexible study areas the and archive.

Dokumente

- 2020Gutachten-Ener...pdf — Energieausweis — Unterkategorien: Strom, Wärme, Zertifikate
- GEO_Baugrundgutac...pdf — Gutachten — Unterkategorien: Tiefbau, Schadstoffe
- Schallschutz1990gu....pdf — Gutachten

Choose your discipline and your preferences so that the interface can adapt to them.

Discipline / Focus: Sustainability

Enter or select subject keywords: Enter here... | Renewables | Energy | Thermal zones

KIT Library South

- Adress: Kaiserstraße 12, 76131 Karlsruhe
- Construction Year: 1966
- Building Type: Library Building
- Floors (above ground): 12
- Floors (below ground): 2
- Nettogrundfläche: 6,850 m²
- Bruttogrundfläche: 8,200 m²

Energy - General

- Energy Source: District heating + electricity
- Energy Consumption (Electricity): 85,2 kWh/(m²·a)
- Energy Consumption (Heat): 78,5 kWh/(m²·a)

Renewable Energy

- PV System: 245,680 kWh/a
- Heat Pump: 156,420 kWh/a
- Heat and Cold Recovery: 890,750 kWh/a

Document arrangement: Time (Years)

1966

Der Bau des Gebäudekomplexes wurde 1990 fertiggestellt.

Dokumente (1)

- 1990-05-08-Waermeschu... — Gutachten — Unterkategorien: Wärme, Zertifikat

2006

Auf Verlangen des Eigentümers wurde ein Energieausweis erstellt.

Dokumente (1)

- 2020Gutachten-Ener...pdf — Energieausweis — Unterkategorien: Wärme, Strom
- Document Name — Dokumententyp — Unterkategorien: Subtopic 1, Subtopic 2
- Document Name

2023

Abgeleitet aus dem Energieausweis wurde ein Renovierungsentwurf erstellt. Maßnahmen beinhalten Isolierung, Fenster und Wände.

Dokumente (2)

- Entwurfsplanung_Bauphy... — Plan — Unterkategorien: Umbau, Energieträger
- 3_AMHM_AN_00_XX_301... — Plan — Unterkategorien: Ansicht, Umbau

2030

Gültigkeits-Erinnerung

Energieausweis läuft aus.

Dokumente (1)

- 2020Gutachten-Ener..... — Energieausweis

(a) General information arranged by topic

(b) Specific information arranged by year

Figure 3. Screenshots of ConSight's dashboard module

5 Conclusion and Outlook

This paper introduces ConSight, an LLM-based prototype to support KWs in the construction industry with information retrieval and contextual understanding of large, unstructured DCs. Our prototype implements DFs that address DRs derived from our qualitative analysis. As a next step, we conclude the first design cycle through qualitative evaluation and plan to employ quantitative methods to assess the appropriateness and effectiveness of the derived design knowledge in the form of design principles. This research-in-progress provides initial insights into how LLM-based information retrieval systems can effectively support KWs in structuring information from unstructured DCs, contributing to the advancement of AI-assisted support tools for KWs.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. (2024), 'Trust in ai: Progress, challenges, and future directions', *Humanities and Social Sciences Communications* **11**(1), 1568.

Alavi, M. & Leidner, D. E. (2001), 'Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues', *MIS Quarterly* **25**(1), 107–136.

Alavi, M., Leidner, D. & Mousavi, R. (2024), 'A Knowledge Management Perspective of Generative Artificial Intelligence', *Journal of the Association for Information Systems* **25**(1), 1–12.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R. & Horvitz, E. (2019), Guidelines for human-ai interaction, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', CHI '19, Association for Computing Machinery, New York, NY, USA, p. 1–13.

Arnold, M., Goldschmitt, M. & Rigotti, T. (2023), 'Dealing with information overload: a comprehensive review', *Frontiers in Psychology* **14**, 1122200.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A. & Xiang, A. (2021), Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', AIES '21, Association for Computing Machinery, New York, NY, USA, p. 401–413.

Chen, N., Lin, X., Jiang, H. & An, Y. (2024), 'Automated building information modeling compliance check through a large language model combined with deep learning and ontology', *Buildings* **14**(7), 1983.

Choong, K. & Leung, P. (2022), 'A critical review of the precursors of the knowledge economy and their contemporary research: Implications for the computerized new economy', *Journal of the Knowledge Economy* **13**, 1573–1610.

De Vreede, T., Singh, V. K., De Vreede, G.-J. & Spector, P. (2024), The effect of is engagement on generative ai adoption, *in* 'Hawaii International Conference on System Sciences 2024 (HICSS-57)', p. 3.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023), Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality, Technical Report 24-013, Harvard Business School Technology & Operations Mgt. Unit Working Paper, The Wharton School Research Paper. Available at SSRN: `https://ssrn.com/abstract=4573321`.

Detlor, B. (2010), 'Information management', *International Journal of Information Management* **30**(2), 103–108.

Duong, H. D. & Lin, J. J. (2022), 'Reality model-based facility management framework for existing building', *Frontiers in Built Environment* **8**, 815672.

Ghosh, M., Mukherjee, S., Ganguly, A. & et al. (2024), ‘Alpapico: Extraction of pico frames from clinical trial documents using llms’, *Methods* **226**, 78–88.

Gioia, D. A., Corley, K. G. & Hamilton, A. L. (2013), ‘Seeking qualitative rigor in inductive research: Notes on the gioia methodology’, *Organizational Research Methods* **16**(1), 15–31.

Henderson, P., Krass, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D. & Ho, D. E. (2022), Pile of law: learning responsible data filtering from the law and a 256gb open-source legal dataset, *in* ‘Proceedings of the 36th International Conference on Neural Information Processing Systems’, NIPS ’22, Curran Associates Inc., Red Hook, NY, USA.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004), ‘Design science in information systems research’, *MIS quarterly* pp. 75–105.

Huang, J. & Chang, K. (2024), A key to building responsible and accountable large language models, *in* K. Duh, H. Gomez & S. Bethard, eds, ‘Findings of the Association for Computational Linguistics: NAACL 2024’, Association for Computational Linguistics, Mexico City, Mexico, pp. 464–473.

Iivari, J. (2015), ‘Distinguishing and contrasting two strategies for design science research’, *European Journal of Information Systems* **24**(1), 107–115.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R. & McHardy, R. (2023), ‘Challenges and applications of large language models’, *arXiv preprint arXiv:2307.10169* .

Khan, N., Zaidi, S. F. A., Yang, J., Park, C. & Lee, D. (2023), ‘Construction work-stage-based rule compliance monitoring framework using computer vision (cv) technology’, *Buildings* **13**(8), 2093.

Kim, J., Chung, S. & Chi, S. (2024), ‘Cross-lingual information retrieval from multilingual construction documents using pretrained language models’, *Journal of Construction Engineering and Management* **150**(6), 04024041.

Klesel, M. & Wittmann, H. F. (2025), ‘Retrieval-augmented generation (rag)’, *Business & Information Systems Engineering* pp. 1–11.

Krütli, D. & Hanne, T. (2025), ‘Augmenting llms to securely retrieve information for construction and facility management’, *Information* **16**(2), 76.

Kus, K., Poehler, L., Kajüter, P., Arlinghaus, T. & Teuteberg, F. (2022), Vaccination dashboard development during covid-19: A design science research approach, *in* ‘Proceedings of the 17th International Conference on Wirtschaftsinformatik (WI)’, Nürnberg, Germany.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S. & Kiela, D. (2020), Retrieval-augmented generation for knowledge-intensive nlp tasks, *in* ‘Advances in Neural Information Processing Systems’, Vol. 33, pp. 9459–9474.

Ma, X., Lin, S.-C., Li, M., Chen, W. & Lin, J. (2024), Unifying multimodal retrieval via document screenshot embedding, *in* ‘Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)’, pp. 6492–6505.

Matheus, R., Janssen, M. & Maheshwari, D. (2020), ‘Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities’, *Government Information Quarterly* **37**(3), 101284.

Nedeljković, D. & Kovačević, M. (2017), ‘Building a construction project key-phrase network from unstructured text documents’, *Journal of Computing in Civil Engineering* **31**(6), 04017058.

Nepal, M. & Staub-French, S. (2016), ‘Supporting knowledge-intensive construction management tasks in bim’, *Journal of Information technology in Construction* **21**, 13–38.

Pakarinen, P. & Huising, R. (2023), ‘Relational expertise: What machines can’t know’, *Journal of Management Studies* .

Patil, R., Boit, S., Gudivada, V. & Nandigam, J. (2023), ‘A survey of text representation and embedding techniques in nlp’, *IEEE Access* **11**, 36120–36146.

Powell, W. W. & Snellman, K. (2004), ‘The knowledge economy’, *Annual Review of Sociology* **30**(1), 199–220.

Reyes, J., Batmaz, A. U. & Kersten-Oertel, M. (2025), ‘Trusting ai: does uncertainty visualization affect decision-making?’, *Frontiers in Computer Science* **7**.

Rezgui, Y., Beach, T. & Rana, O. (2013), ‘A governance approach for bim management across lifecycle and supply chains using mixed-modes of information delivery’, *Journal of Civil Engineering and Management* **19**(2), 239–258.

Schmidt, L., Olorisade, B., Thomas, J. & et al. (2025), ‘Data extraction methods for systematic review (semi)automation: Update of a living systematic review’, *F1000Research* **14**, 664.

Schoormann, T., Möller, F., Chandra Kruse, L. & Otto, B. (2024), ‘BAUSTEIN—A design tool for configuring and representing design research’, *Information Systems Journal* **34**(6), 1871–1901.

Seidel, S., Mueller-Wienbergen, F., Michael & Becker, J. (2008), A conceptual framework for information retrieval to support creativity in business processes, *in* ‘ECIS 2008 Proceedings’, p. 251.

Shneiderman, B. (1996), The eyes have it: a task by data type taxonomy for information visualizations, *in* ‘Proceedings 1996 IEEE Symposium on Visual Languages’, pp. 336–343.

Taherdoost, H. & Madanchian, M. (2023), ‘Artificial intelligence and knowledge management: Impacts, benefits, and implementation’, *Computers* **12**(4), 72.

Thorat, C., Bhat, A., Sawant, P., Bartakke, I. & Shirsath, S. (2022), ‘A detailed review on text extraction using optical character recognition’, *ICT Analysis and Applications* pp. 719–728.

Venkatesh, V., Thong, J. Y. L. & Xu, X. (2012), ‘Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology’, *MIS Quarterly* **36**(1), 157–178.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y. & Chen, E. (2024), ‘Large language models for generative information extraction: A survey’, *Frontiers of Computer Science* **18**(6), 186357.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F. & Zhou, M. (2020), Layoutlm: Pre-training of text and layout for document image understanding, *in* ‘Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)’, pp. 1192–1200.

Yin, M. (2025), ‘Bridging the gap between machine confidence and human perceptions’, *Nature Machine Intelligence* **7**(3), 330–331.

Zhai, C. (2024), Large language models and future of information retrieval: Opportunities and challenges, *in* ‘Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval’, SIGIR ’24, Association for Computing Machinery, New York, NY, USA, p. 481–490.