

# Law Meets GenAI: Using Artificial Intelligence to Derive Conceptual Models from Legal Regulations

Binh An Patrick Nguyen<sup>1</sup>, Hendrik Scholta<sup>1</sup>, David Roth-Isigkeit<sup>1</sup>, Christian Djeflal<sup>2</sup>, Friedrich Chasin<sup>3</sup>

<sup>1</sup>German University of Administrative Sciences Speyer, <sup>2</sup>Technical University of Munich,

<sup>3</sup>Offenburg University of Applied Sciences

E-Mail: {patrick.nguyen | hendrik.scholta | roth-isigkeit}@uni-speyer.de, christian.djeflal@tum.de, friedrich.chasin@hs-offenburg.de

## Abstract

*Artificial intelligence (AI) and conceptual models are both important to public organizations. AI and generative AI (GenAI) can help to cope with an increasing resource shortage, workload, and requirements, while conceptual models are essential for the design of IT systems. However, the combination of both, the creation of conceptual models using GenAI tools in public organizations, has been barely addressed in extant research. Thus, we investigate (1) how legal experts use GenAI tools when deriving conceptual models for public services from legal regulations and (2) what their experiences are in this use. In a qualitative study with 18 administrative legal experts we obtained various insights. For instance, we show that the participants either submitted strict instructions or conducted open conversations, and they followed a top-down, bottom-up or combined approach in their analysis. The GenAI tools performed better in generating text-based models (forms) than graphic-based models (process models, decision trees).*

**Keywords:** Law, Large Language Model, Prompting, Conceptual Modeling, Digital Public Service

## 1. Introduction

Generative Artificial Intelligence (GenAI) tools offer transformative opportunities for public organizations, particularly in legal contexts, to address increasing workloads and increasingly complex requirements. From automated document management to AI-assisted decision support and citizen-facing chatbots—the applications are diverse and promise substantial improvements in services and processes. Simultaneously, these technologies raise fundamental questions regarding their reliability, transparency, and legal classification.

Retrieval Augmented Generation (RAG) systems represent significant progress in reliability and

transparency through the integration of additional documents and information sources (Huang & Huang, 2024). In addition to a model's training data, a RAG tool can access additional material to incorporate further information. However, central challenges remain such as balancing innovative utilization with rule-of-law principles.

Understanding the tools' strengths and weaknesses is key to deploying these tools effectively in legal and public administration contexts (Bjelobaba et al., 2024; Pesch, 2025). In that regard, it is essential to realistically assess GenAI capabilities in comparison to human work and to recognize how human-computer interaction can create synergies that enhance both efficiency and judgment quality (Bao et al., 2023; Hemmer et al., 2024)—a complementary relationship that preserves critical human oversight while leveraging computational strengths in administrative and legal processes.

In addition to GenAI tools, conceptual models are important for public organizations, the services they offer, and their IT systems (Scholta et al., 2019). Conceptual models are used in the design of IT systems as a means for communication between domain experts and IT experts such that the IT experts can adequately realize the domain experts' requirements (Hoppenbrouwers et al., 2005). The abstraction of a system in a conceptual model is useful to create a common understanding between those experts (Nguyen & Scholta, 2024b). An abstraction is a less detailed projection of reality which only consists of details the creator deems relevant (Greca & Moreira, 2000). Such a model of reality can be further simplified and abstracted until it reaches a level that experts from the other area of expertise can understand. With ensuring common concepts to be understood by both parties, collaborative work can be done more effectively as misinterpretation can be reduced.

Although AI tools and conceptual models are both important to public organizations and their

digitalization, the use of GenAI tools for the derivation of conceptual models from legal regulations in public organizations has been barely investigated. IT experts use conceptual models to structure their data and processes; however, complexity lies in extracting the relevant information from legal regulations. Previously, this extraction was a manual task, but AI tools—and especially GenAI tools—now have the potential to support this task and improve its execution. GenAI tools can quickly analyze large amounts of text and retrieve relevant information for a certain question, providing answers almost in real-time.

However, a complete automation of the analysis of legal regulations is hardly possible and humans would still be involved to some degree because interpretations made by an AI algorithm can lead to a loss of the “intricacies of meaning and intentions inherent in [...] interactions” (Cordeiro & Cozman, 2024, p. 34) such that AI-generated output can be wrong.

To support the semi-automated analysis and generation of conceptual models from legal regulations, we address the following two research questions:

*RQ1: How do domain experts proceed when deriving conceptual models for public services from legal regulations with the help of contemporary GenAI tools?*

*RQ2: What are advantages, disadvantages, and improvement potentials of using GenAI tools when deriving conceptual models from legal regulations?*

To answer the two research questions, we performed a qualitative study with 18 domain experts (legal experts) and compared a general GenAI tool and a RAG tool. Answering these two research questions provides a better understanding of the potentials of GenAI tools in public organizations which is essential for developing and using systems that enhance efficiency while preserving reliability in administrative and legal processes.

The remainder of this paper is structured as follows. Chapter 2 presents the research background and chapter 3 is dedicated to the study’s research design. While chapter 4 presents the study’s results, chapter 5 discusses these results and chapter 6 concludes the paper.

## 2. Research background

### 2.1. AI in administrative and legal processes

AI has rapidly evolved to offer unprecedented capabilities for processing, analyzing, and generating text-based information, fundamentally changing how

administrative tasks can be approached (Sánchez-Navalón et al., 2025). At the core of GenAI tools are large language models (LLMs). A LLM is a machine learning model with many parameters that is used for natural language processing and language generation such as texts, audio or videos.

The application spectrum for LLMs in legal and administrative contexts is remarkably broad (Beck, 2024). As documented in recent pilot projects across various European authorities, these models can assist in numerous critical functions (Bright et al., 2025; Vladika et al., 2024). They support processing citizen applications, analyzing legal documents, drafting standardized texts, and enhancing decision-making processes. Particularly promising are applications in automated review of applications and objections, transcription and summarization of hearings, translation of technical legal language into citizen-friendly explanations, and provision of 24/7 digital assistance services.

RAG systems offer particular advantages in legal contexts. Unlike traditional LLMs limited to their training data, RAG systems can access and incorporate current legal regulations, precedents, and organizational documents into their responses. This capability is crucial for legal work, where accessing up-to-date regulatory frameworks is essential for accurate processing and decision-making.

Despite these promising applications, significant concerns persist regarding the reliability of LLM outputs in legal contexts (Dahl et al., 2024; Wang et al., 2024). The complex, nuanced nature of legal interpretation presents challenges that current GenAI tools struggle to navigate consistently. “Hallucinations”—fabricated or incorrect “information presented as factual” (Bjelobaba et al., 2024, p. 17)—are a particularly serious risk in legal proceedings, where accuracy is paramount (Bjelobaba et al., 2024; Pesch, 2025). The architecture of LLMs can lead to variances in results that remain opaque to users, as small variations in prompts can produce completely different outcomes (Atil et al., 2024; Zhou et al., 2024). Users can not even necessarily reproduce answers by repeatedly submitting the same prompt. Such inconsistencies create tension with legal principles since transparency and legal certainty are key elements of the rule of law (Rechtsstaatsprinzip) (Schmidt-Abmann, 2004; Tamanaha, 2004; Tschohl, 2020).

However, it would be shortsighted to only criticize potential errors of LLMs, as current administrative realities—including increasing workloads and challenging demographic trends—place significant pressure on traditional “analog” administration. This raises the question of whether the standards applied to

GenAI tools exceed those applied to human administrators (Lenskjold et al., 2023).

The question of whether LLMs should function independently or as assistive tools remains central to legal implementation discussions (Enarsson et al., 2022; Urquhart et al., 2022; Weis et al., 2025). One central question is whether an augmented approach where GenAI serves as a “super tool” that supports rather than replaces human expertise is most beneficial (Cui & Yasseri, 2024).

## 2.2. AI in conceptual modeling

Conceptual models abstract the complexity of an object or scenario to ensure understanding of it to a certain target group through collaboration of trained experts in both the domain and modeling techniques (Prokop et al., 2025). Especially the domain knowledge can be spread across multiple sources and identifying, capturing, and analyzing the massive volume of information relevant to the object or scenario can be a time-intensive task before modeling starts (Franzoi et al., 2025). GenAI tools work on probabilities and not with fixed rules, hence while a properly trained and prompted GenAI tool can generate the desired model, it comes with the caveat of non-repeatable results (Fill et al., 2023; Kourani et al., 2024a).

While GenAI tools can summarize and detect patterns, meta-prompts need to tell it how to interpret the patterns and relations it discovers (Franzoi et al., 2025; Neuberger et al., 2025). Neuberger et al. (2025) found that prompting consists of the subtasks on detection, resolution, and relation of entities. Because of the probabilistic approach of LLMs, the main challenge for GenAI tools are “Long-distance Relations” (Neuberger et al., 2025, p. 42). Hereby they mean that a single word can change the meaning and context of a whole process description or in our case legal term definition, which is in line with the general challenge of understanding legal concepts that try to cover as many cases in as few words as possible (Breux & Antón, 2008; Neuberger et al., 2025). Another challenge of GenAI and conceptual modeling in the legal domain are cross references within a legal regulation and across the whole legal corpus in addition to vague legal terms that change their definition between legal subdomains (Surden, 2024).

Apart from the input source, the output (models) quality is dependent on the method used. It is possible to generate code that constitutes the model in machine-readable format (e.g., XML) (Köpke & Safan, 2025). It is, however, also possible to let the GenAI tool generate code that, as an executable program, creates the model (Kourani et al., 2024a). Especially for

domain experts who require modeling support, the visualization of the models is important, which in most cases can be achieved by turning a GenAI tool’s technical reply into a visualization or achieving the correct visualization via meta-prompts (Köpke & Safan, 2025). Meta-prompts can specify what data is to be extracted in which form and also how the output is to be displayed (Köpke & Safan, 2025). GenAI tools may be able to suggest the appropriate concept as a matching modeling element to a human domain expert (Prokop et al., 2025).

To give the GenAI tool the correct instructions, different prompting strategies can be employed (White et al., 2023). Each strategy aims at setting specific contexts for the GenAI tool to respond to like personas (e.g., a role to act in) or adding information not part of the training material (Kourani et al., 2024b; White et al., 2023). Selecting the correct strategy matching both the LLM employed as well as the domain to be modeled can influence perception and interactions (Ali et al., 2025).

## 3. Method

As part of our study, we asked the participants to perform a task and subsequently reflect on this task. The participants’ task was to derive forms, process models, and decision trees for a public service from relevant legal regulations with the help of GenAI tools. They did not receive restrictions on how to use the tool at their disposal. The participants documented their approach for deriving the models and their use of the GenAI tool. In the end, the participants wrote down their experiences with and their assessment of the tools. Afterwards, we qualitatively analyzed the participants’ reports.

The public service that was the subject of our study was housing benefit in Germany. Housing benefit is a social security benefit that supports low-income citizens in handling their housing costs. We selected this service because it is highly relevant for citizens and the underlying requirements in German law are complex enough to lead to non-trivial conceptual models with a certain number of elements but are still manageable in a restricted timeframe.

The participants were divided into two groups based on which tool we asked them to use: the first group used a general LLM tool (ChatGPT), while the second group used an RAG tool (KIBridge). We selected ChatGPT because it is in widespread use and we selected KIBridge as RAG tool because of its benefits in configuration, control, and transparency.

As general LLM tool without embedded documents, we used ChatGPT from OpenAI with the GPT-4o model. We selected ChatGPT because it is

currently the most popular and accessible GenAI tool (Milmo, 2025). The use of the Plus license ensured that the participants were not restricted by chat volume in ChatGPT and were able to use its most powerful version. In our study, ChatGPT was able to search data in the internet.

KIBridge is a RAG tool designed for two primary user groups: knowledge base managers, who configure and maintain domain-specific data sources, and end users, who access them via dedicated chat interfaces. Acting as a wrapper around LLMs, KIBridge addresses their black-box nature across three dimensions: configuration, control, and transparency (Cappel & Chasin, 2024). Configuration is achieved by embedding organizational documents into a local vector database, enabling the generation of context-specific input tokens. Retrieval control allows managers to define how information is selected—favoring either full documents or granular chunks—and to set sensitivity thresholds for relevance. Meta-prompts further steer retrieval logic and output formatting to minimize hallucinations. For transparency, each response includes links to the specific passages used as context, along with semantic similarity scores, enhancing verifiability and user trust.

We configured KIBridge for the purposes of our study. First, the LLM we used was OpenAI’s GPT-4o model to have matching LLMs in both tools to facilitate comparable results. Second, we embedded the relevant legal regulations for housing benefit (the act and the administrative directive) as external sources for use by the RAG tool. Third, the behavior of KIBridge was guided by meta-prompts. These prompts were set to explain the relationship between the act and the administrative directive as well as prime the LLM for extracting information for the modeling purpose, e.g., “always cite a source for your answer”.

We selected 18 legal experts as the participants in our study because the underlying foundation for public services are legal regulations. These regulations contain the information that the derived conceptual models need to represent. The legal experts were post-graduate students who had already passed through a five-year study program in law equivalent to a Masters degree and are now enrolled in an additional two-year program leading to the bar exam. The students selected a seminar that was supervised by the authors. Each of the two tools was targeted by nine of the participants who were randomly assigned to the tools.

The participants’ work, their interaction with the tools, and the final analysis were subdivided into four phases. In the first phase, the participants received training in conceptual modeling. As the participants

were legal experts and had not been familiar with conceptual modeling before, they received a six-hour training in conceptual modeling with a focus on processes, forms, and decisions. They practiced creating models manually and became familiar with the syntax of the relevant modeling languages. Thereby, we ensured a baseline of knowledge and common understanding regarding conceptual modeling.

In the second phase, the participants worked individually on their task to create process models, forms, and decision trees with the support of a GenAI tool. Each participant worked on their own to produce independent results. Still, we held six meetings in which we provided feedback, answered questions of the participants, and assisted in challenges they encountered during the application of the tools.

In the third phase, each participant wrote an essay on their work. The participants documented their approach including the prompts they used, the results, their experiences with the tools as well as strengths, weaknesses, and ideas for the tools’ further developments. Each essay was about 15 pages long.

Finally in the fourth phase, two researchers coded and analyzed these essays qualitatively. Both researchers coded the essays individually and aligned their inductively created coding trees in regular meetings. They resolved conflicts in codings in these meetings and obtained a commonly agreed upon coding result in the end.

## 4. Results

The participants were asked to formulate a report, containing (1) their approach to communication with the GenAI tool, and (2) their evaluation of the specific use of the GenAI tool as a tool to derive forms, process models, and decision trees for a public service.

### 4.1. Approach (RQ1)

The participants **interpreted their task differently** (deriving forms, process models, and decision trees for a public service from relevant legal regulations with the help of GenAI tools), which is also evident in their reports: Most participants interpreted it as “create models using exclusively GenAI”. Others used GenAI either as support (e.g., “*I instructed ChatGPT to generate scenarios, that were not covered by the generated decision tree and discuss that with itself.*”) or as a pure information collection assistance (e.g., “*First [ChatGPT] was prompted to extract the legal requirements*”).

In general, the participants **followed different approaches** in the interaction with the GenAI tools as

they built emotional relationships with “*their*” GenAI tool and had the feeling that the tools “*had different abilities under different users*”. The GenAI tools’ responses to the prompts were non-repeatable, making the systems prone to either errors or uncertainty. This was reported as “*[the GenAI tool] generates new models, claiming to include the new information, but in fact the new model did not contain those changes*” and variations thereof. This led to the participants employing different approaches to deal with this caveat.

The participants’ **general approaches differed across two dimensions (tone and method)**. Tone refers to the general character, attitude or mood conveyed by a user. Method describes the logical form of the queries and prompts and indicates how the user directed the GenAI tool to generate a specific output.

With respect to the **tone**, all participants either gave strict orders or had open conversations with the GenAI tool as “*with more time, working became easier and I had the feeling of humanity in the replies. The longer I worked with ChatGPT, the better the results were*”.

With respect to **method**, we distinguish a top-down and bottom-up approach. Creating a general abstract model and refining it with more specific information is the top-down approach, whereas the bottom-up approach starts with the creation of detailed model components that are integrated together into a larger model. Some participants opted for the top-down approach (e.g., “*What are the decisions to be made by an agency to process an application for housing benefits?*” followed by “*Based on this: When is an application complete and what documents need to be added to it?*”). This query starts with a general question for decisions required and then asks for documents that complement these decisions). Others used the bottom-up approach to “*extract relevant requirements for a successful application and thus identify relevant fields for a form*”. Some participants started off with the top-down approach while later switching to the bottom-up approach (e.g., “*First prompts were open and general to get an overview. Based on the answers I asked for more information on unclear or simply interesting points. [...] When creating the process model, I asked for involved departments first before asking for the typical workflow within and between those departments.*”).

Some participants applied a **more creative approach** and uploaded additional files as information, such as a picture of a model (BPMN process model or decision tree) to ChatGPT, while others asked ChatGPT for an image as an output instead of text. With the probabilistic behavior of GenAI tools this resulted in interesting, although both

syntactically and semantically wrong images. Especially the images followed general patterns (process models had squares and arrows), but the content was illegible.

All participants considered a **final fact-checking** of the GenAI tools’ answers necessary (“*Despite uploaded laws, ChatGPT seems to hallucinate more if the upload is more messages ago.*”). Some participants complained about the limited ability of the GenAI tools to find relevant information as it is “*missing the required depth of analysis to extract singular requirements. For example, KIBridge did not recognize that income has to be determined for each household member*”. Even if technically correct relevant information is found, when asked for a source either a source can be wrongly prioritized (e.g., “*From a legal point of view it is unwelcome to use administrative directives [and not the acts] as the main reference [...], as while they do have [internal] self-binding effects [...] [they] have no external legal bindings*”) or completely made up, for a human indistinguishable from a valid source.

## 4.2. Advantages, disadvantages, and improvement potentials (RQ2)

The general consensus across all participants is that the models created can serve as a **first starting point** to either have a base line (“*in my opinion ChatGPT is able to generate a base structure of a process*”) or a discussion point (“*content-wise the basic steps [of the decision tree] build on each other and were easy to follow*”).

Concerning the results, all noted that while having legally based models were a huge advantage, the **lack of depth** in analyzing the text made the GenAI prone to generalizations or missing important legal facts. Hence, in most cases the results would serve only as a starting point for modeling by hand. The participants were also creative in using GenAI tools for modeling by hand. The content was also structured by the GenAI tools, giving headlines to group information into logical blocks. This was perceived by multiple participants from both tools as the most helpful feature right before traceability of sources. KIBridge users were more confident about the answers as “*the cited sources enable a fast research and confirmation of answers*”. With structured information, models can also be created step by step. After each step, corrections can be made to the current model before proceeding with the next prompt.

Most participants expressed considerable **concerns against relying on the models** as their content can be “*incomplete or unreliable*” or “*intransparent why specific regulations were very*

*important compared to others*". Among further functional reasons were (1) different results with the same prompts and (2) prompts were either not executed at all or led to worse results in the course of time (e.g.: *"It seems like the GenAI got tired over time or worked less precise, which made working with it harder."*). As a general challenge with the concept of law, LLMs as statistical models cannot replicate the exact same answer on the exact same prompts. Also if the prompts were different *"the AI generated different answers on repeated prompts to the same topic and made inconsistent statements."*

With respect to the accuracy of the created model, the participants found that **with increasing reliance on text, the model was more precise and coherent**. Generating forms produced the most accurate results while visual descriptions like decision trees and process models were harder to generate. A participant concluded that GenAI tools would be best suited for text-based representations, especially since LLMs can simulate text generation best. Those texts include both textual descriptions of processes as a sequence of events as well as forms, especially since public organizations have a lot of similar forms hence *"it is possible that the AI had frequent access to forms already, so that generation of those is easier"*. However, it can be noted that the generated *"forms contain all relevant questions, but not as extensive as forms used by agencies would use. [...] Furthermore, the layout seems more like an unprofessional survey than an application form an agency would offer online."* Process models and decision trees did suffer from formatting issues (*"sequence flows did not connect to the tasks correctly"* and *"we could not recognize decision trees as we were taught they should look like"*), while forms in general did well in visualizations. The visualized models became more coherent, not necessarily correct, when they had to be generated in code.

With respect to the **generation of code**, the results between the GenAI tools differed. While KIBridge was able to give bpmnXML code because it was hardcoded in the meta-prompts, the same could not be said for the other model types (forms and decision trees). The other model types were only instructed through the meta-prompts *"if a decision is to be made, create a decision tree structure"* and *"if the answer is something that can be asked from a citizen, create a form or extract fields for it"*. The dependency on the data the RAG had access to was restricted to the uploaded law and administrative directive. KIBridge users had no option of uploading additional content apart from entering it into the chat field. ChatGPT was able to generate code to display models in different programming languages. In this context, however,

only some parts of the code were working (e.g., inserting *"bpmn2.0"* in a tag because the prompt specified BPMN 2.0 compliant code, whereas bpmnXML only has a reference to the current version's OMG specification URL).

Further **challenges came up for those unfamiliar with technical code**. One noted, that they were unable to follow the conversion instructions given until another participant helped them along step-by-step: *"Short statements like 'Click here' ... 'save that this way' ... 'then open this here'. It probably also helped that they were able to see my screen, which I probably wouldn't have allowed ChatGPT to do."*

In general, a major challenge was getting the **GenAI tool to understand what the human user wanted it to do**. The misinterpretation of prompts led to different functional outcomes: (1) ignoring the prompt, (2) ignoring past answers (general inconsistencies), (3) claiming abilities that it does not have, (4) failing to answer, (5) long consideration times with either no answer or repeated answer, (6) doing other things than prompted. The participants were able to overcome some of the functional challenges by changing their prompting strategy or restarting the chat.

However, one workaround to be highlighted was the usage of the **German word "Doch"**, a word capable of multiple meanings at once, among them correction, reminder of a fact or assertion of dominance (Zeevat & Karagjosova, 2009) (freely translated as "yes, you can" if a GenAI tool replies with being unable to do something). In this context, it shows that GenAI tools as a simulation of conversation focus more on matching the prompt than giving factual correct answers. Because of the corrective assertion properties of the word "Doch", the GenAI tools were convinced of being capable of performing the desired request although they had previously neglected it and finally often complied with the initial prompt.

**Misunderstandings arose when participants tried to provide a more accurate definition of legal terms** than suggested by the GenAI tool. Here, *"the extensions were not edited in, but attached to the end at random as a new subtree"*.

Ultimately, participants provided **suggestions** on how to improve the use of GenAI tools in the legal and administrative domain. These suggestions can be clustered into three groups: user experience, content, and technology.

**With respect to user experience, there were several suggestions:** (1) Visual indicators: While navigation in a chat environment happened intuitively, there were no indicators where a user could edit a GenAI's text to enter it as a part of a new prompt; (2)

Traceability of information: Indicators were missing to derive the source of an information; (3) More GenAI training: Getting the correct data often required precise prompting which in the opinion of many participants also required training, some even suggested having the GenAI tool teach the user how to properly prompt it as one user of KIBridge “*asked ChatGPT how to prompt it*”; (4) Saving function: Models should also be able to be saved as a milestone to be reused in case further prompting resets progress, or if new information is to be included.

**Content improvements** included (1) the visualization of models, especially decision trees and process models and (2) a higher consistency of answers, especially in law. For GenAI tools to be able to interact with legal documents correctly, it was also suggested to adapt legal regulations for GenAI uses or develop a GenAI tool that is specialized on legal documents. The participants mentioned (3) the admission of lacking knowledge: Another improvement suggested by the participants is to make the GenAI tool admit that it is unable to answer properly and refer to the probable location of the information. Especially with legal terms, GenAI tends to either make up the information or attribute correct information to the wrong source.

**Technical improvements** concerned system stability and processing speed: (1) In particular with ChatGPT, some functionalities like image generation bottlenecked during the study; (2) KIBridge suffered from getting stuck in loops. As prompts always regarded the last few prompts and replies, it could get stuck in a loop when it was “*unable to answer in the given context*”. A suggestion here is to add a functionality to reset the regarded answers without having to start a new chat and loose progress made before.

## 5. Discussion

Our study confirms and challenges existing literature. The literature on the use of GenAI tools in administrative and legal processes supports our findings concerning the lack of legal precision, especially when legal concepts with room for interpretation are involved (Wang et al., 2024). This room for interpretation mirrors the long-distance relations between concepts (Neuberger et al., 2025).

With ChatGPT’s function to simulate RAG behavior with data uploads, the difference between ChatGPT and KIBridge was less prominent, although the direct sourcing of data in KIBridge did give more accurate results if prompted properly. In general, the differences between the general LLM tool and the RAG tool were negligible in our study as most

challenges arose from the GenAI tools’ capability to create the requested models and to understand what the user requested or intended by their prompt.

Open and loosely defined terms in legal regulations (see Marmor, 2018) hamper GenAI tools as they have difficulties in identifying the proper context in order to define the term for a certain scenario. For instance, in the housing benefit rules there are multiple references to terms defined in other legal regulations in the social domain, where the same term, depending on the context, can be defined differently. The possibility exists that the lack of the complete legal corpus is a reason for incomplete responses, but that at the same time additional information may confuse the GenAI tool. Therefore, the context of a term is crucial and the previously submitted prompts might help the GenAI tool in determining this context. As recognized in the literature, long-term memory is a problem with GenAI tools (Zhong et al., 2024) and the participants in our study questioned how long the perceived chat memory of the GenAI tools would last, especially after discovering that if the GenAI tool has not been used for some time, the context gets lost.

The participants received no dedicated training in prompting regarding information extraction and thus approached the chatbot intuitively and by a trial-and-error manner. Some of the intuitive prompting strategies employed by the users mirror established prompting patterns like the Persona Pattern (GenAI tool is asked to pretend to be in a certain role) or the Reflection Pattern (GenAI tool is asked to evaluate its own answer) (White et al., 2023), however, most participants had natural and spontaneous conversations with their GenAI tool. Literature suggests applying established patterns to use GenAI tools (White et al., 2023), however, current research does not touch upon comparability when participants combine approaches/patterns.

Still, the participants differentiated between a top-down and a bottom-up approach and switched according to their situation and their satisfaction with the GenAI tool’s performance. This highlights that intuitively and in a learning-by-doing manner legal domain experts can adapt to the use of GenAI tools and how to interpret the results for their purpose. However, there is no clear pattern as to when would be the sweet spot to switch from one approach to the other to maximize the quality of results.

The participants also did not receive specific training in machine readable code representing visual models (e.g., what BPMNxml is to BPMN 2.0). While the literature often suggests to generate models as code or in another technical textual representation (Köpke & Safan, 2025; Kourani et al., 2024a), the lack of IT

expertise among the majority of the administrative legal experts in this study proved to be a major hindrance because they could not check the code created by the GenAI tool.

While the majority of the literature focusses on how to optimize the prompts for correct answers (Kinderen & Winter, 2024; Neuberger et al., 2025; White et al., 2023), there is a lack of concern on how to react when the prompts do not result in a helpful answer (also an incorrect answer can be helpful). The automatic reaction of the legal domain experts to check the source document (i.e., the underlying legal regulation) is an expectable but biased reaction as in this study, housing benefit has an existing law with already existing forms and processes are at least implied by the decisions to be made and facts to be checked. What remains unresearched is when laws and other legal regulations do not specify the tasks explicitly or implicitly, relying either on past data to discover a pattern (e.g., a specific legal term always matching an administrative action) or on the legal domain expert to specify this information in the conversation with the GenAI tool.

We extend guidelines for the design of AI systems in the public sector (e.g., Androutsopoulou et al., 2019; Hemesath & Tepe, 2024) by making recommendations for GenAI tools that support the derivation of conceptual models from legal regulations. Such tools should provide references to the source of a piece of information to increase a user's confidence in the tool and its results. When providing such references, GenAI tools should consider prioritization of sources (e.g., act over administrative directive) which may be provided by the user. GenAI tools should also indicate the degree of reliability of their answers. Moreover, they should enable the storage of intermediate model results for later use.

## 6. Conclusion and outlook

In this paper, we investigated the use of GenAI tools by legal experts in the derivation of conceptual models for public services from legal regulations. First, we analyzed the approaches the legal experts used. Second, we analyzed their experiences and the resulting advantages, disadvantages, and improvement potentials. The participants in our study were split into two groups: one group used a general LLM tool (ChatGPT), the other group used a RAG tool (KIBridge).

Limitations concern both the study target as well as the tools used. The legal regulation used was a housing benefit application. Housing benefits as such have existed for a long time, with multiple forms to apply for it existing online and thus being part of the

data used to train GPT 4o. Another limitation is the limited comparability of the resulting models as participants understood the task of the qualitative study differently (create conceptual models using the GenAI tool exclusively vs. create conceptual models using the GenAI tool as assistance).

These concerns can be addressed in future research by giving more precise tasks and following up on the actual applicability of the resulting models in the public service design process. Especially precision and transparency of the LLMs employed can be investigated further, particularly with LLMs trained specifically on legal data, as legal documents form a subset of the general language both in application and interpretation. Furthermore, it can be investigated how a changing legal corpus in case law differs from the more sturdy and consistent civil law when employing GenAI tools.

Moreover, our meta-prompts for the RAG tool might have been too restrictive and might have limited the legal experts too much in their prompts and the resulting answers. Future work can investigate suitable balances in meta-prompts that guide the tool and user but are not too restrictive.

GenAI tools can support the derivation of conceptual models from legal regulations (e.g., Ghanavati, 2023; Kulkarni et al., 2023; Nguyen & Scholta, 2024a; van Engers & van Doesburg, 2016), but improvements and further research are necessary. This study is a starting point for further investigations into this topic in the future.

## Acknowledgment

This project received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – FOR 5393, Project No. 462287308 (SCHO 1965/1-1).

## References

- Ali, S. J., Reinhartz-Berger, I., & Bork, D. (2025). How are LLMs Used for Conceptual Modeling? An Exploratory Study on Interaction Behavior and User Perception. In W. Maass, H. Han, H. Yasar, & N. Multari (Eds.), *Lecture Notes in Computer Science: Vol. 15238, Conceptual Modeling* (pp. 257–275). Springer Nature.
- Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, 36(2), 358–367.
- Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2024). *Non-*

- Determinism of "Deterministic" LLM Settings.*  
<http://arxiv.org/pdf/2408.04667>
- Bao, Y., Gong, W., & Yang, K. (2023). A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems*, 11(9), 442.
- Beck, B. (2024). KI - zu den Auswirkungen und Chancen der „fünften industriellen Revolution“ für die Justiz. *Juris - Die Monatszeitschrift*, 11(4), 209–213.
- Bjelobaba, S., Waddington, L., Perkins, M., Foltýnek, T., Bhattacharyya, S., & Weber-Wulff, D. (2024, December 13). *Research Integrity and GenAI: A Systematic Analysis of Ethical Challenges Across Research Phases.* <http://arxiv.org/pdf/2412.10134>
- Breaux, T. D., & Antón, A. I. (2008). Analyzing Regulatory Rules for Privacy and Security Requirements. *IEEE Transactions on Software Engineering*, 34(1), 5–20.
- Bright, J., Enock, F., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2025). Generative AI is already widespread in the public sector: evidence from a survey of UK public sector professionals. *Digital Government: Research and Practice*, 6(1), 1–13.
- Cappel, J., & Chasin, F. (2024). Bridging Enterprise Knowledge Management and Natural Language Processing - Integration Framework and a Prototype. In M. Mandviwalla, M. Söllner, & T. Tuunanen (Eds.), *Lecture Notes in Computer Science: Vol. 14621. Design Science Research for a Resilient Future* (pp. 278–294). Springer Nature.
- Cordeiro, V. D., & Cozman, F. (2024). Artificial Intelligence and Everyday Knowledge. In H. S. Dunn, M. Ragnedda, M. L. Ruiu, & L. Robinson (Eds.), *The Palgrave Handbook of Everyday Digital Life* (pp. 23–35). Springer Nature.
- Cui, H., & Yasserli, T. (2024). Ai-enhanced collective intelligence. *Patterns*, 5(11).
- Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1), 64–93.
- Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, 31(1), 123–153.
- Fill, H.-G., Fettke, P., & Köpke, J. (2023). Conceptual Modeling and Large Language Models: Impressions From First Experiments With ChatGPT. *Enterprise Modelling and Information Systems Architectures*, 2023(18).
- Franzoi, S., Delwaulle, M., Dyong, J., Schaffner, J., Burger, M., & vom Brocke, J. (2025). Using Large Language Models to Generate Process Knowledge from Enterprise Content. In K. Gdowska, M. T. Gómez-López, & J.-R. Rehse (Eds.), *Lecture Notes in Business Information Processing: Vol. 534, Business Process Management Workshops* (pp. 247–258). Springer Nature.
- Ghanavati, S. (2023). *Legal-URN framework for legal compliance of business processes* [PhD thesis]. University of Ottawa, Ottawa.
- Greca, I. M., & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education*, 22(1), 1–11.
- Hemesath, S., & Tepe, M. (2024). Public value positions and design preferences toward AI-based chatbots in e-government. Evidence from a conjoint experiment with citizens and municipal front desk officers. *Government Information Quarterly*, 41(4).
- Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024, March 21). *Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence.* <http://arxiv.org/pdf/2404.00029>
- Hoppenbrouwers, S. J. B. A., Proper, H. A., & van der Weide, T. P. (2005). A Fundamental View on the Process of Conceptual Modeling. In L. Delcambre, C. Kop, H. C. Mayr, J. Mylopoulos, & O. Pastor (Eds.), *Lecture Notes in Computer Science: Vol. 3716, Conceptual Modeling - ER 2005* (pp. 128–143). Springer Nature.
- Huang, Y., & Huang, J. (2024, April 17). *A Survey on Retrieval-Augmented Text Generation for Large Language Models.* <http://arxiv.org/pdf/2404.10981>
- Kinderen, S. de, & Winter, K. (2024). Towards Taming Large Language Models with Prompt Templates for Legal GRL Modeling. In H. van der Aa, D. Bork, R. Schmidt, & A. Sturm (Eds.), *Lecture Notes in Business Information Processing: Vol. 511, Enterprise, Business-Process and Information Systems Modeling* (pp. 213–228). Springer Nature.
- Köpke, J., & Safan, A. (2025). Efficient LLM-Based Conversational Process Modeling. In K. Gdowska, M. T. Gómez-López, & J.-R. Rehse (Eds.), *Lecture Notes in Business Information Processing: Vol. 534, Business Process Management Workshops* (pp. 259–270). Springer Nature.
- Kourani, H., Berti, A., Schuster, D., & van der Aalst, W. M. P. (2024a). Process Modeling with Large Language Models. In H. van der Aa, D. Bork, R. Schmidt, & A. Sturm (Eds.), *Lecture Notes in Business Information Processing: Vol. 511, Enterprise, Business-Process and Information Systems Modeling* (pp. 229–244). Springer Nature.
- Kourani, H., Berti, A., Schuster, D., & van der Aalst, W. M. P. (2024b, November 17). *Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis.* <http://arxiv.org/pdf/2412.00023>
- Kulkarni, A., Ramanathan, C., & Venugopal, V. E. (2023). Ontology Mediated Document Retrieval for Exploratory Big Data Analytics. In *Proceedings of the IEEE 17th International Conference on Semantic Computing (ICSC)* (pp. 100–103). IEEE.
- Lenskjold, A., Nybing, J. U., Trampedach, C., Galsgaard, A., Brejnebol, M. W., Raaschou, H., Rose, M. H., & Boesen, M. (2023). Should artificial intelligence have lower acceptable error rates than humans? *BJR Open*, 5(1).

- Marmor, A. (2018). Varieties of Vagueness in the Law. In G. Bongiovanni, G. Postema, A. Rotolo, G. Sartor, C. Valentini, & D. Walton (Eds.), *Handbook of Legal Reasoning and Argumentation* (pp. 561–580). Springer Nature.
- Milmo, D. (2025, February 1). DeepSeek, ChatGPT, Grok ... which is the best AI assistant? We put them to the test. *The Guardian*.  
<https://www.theguardian.com/technology/2025/feb/01/deepseek-chatgpt-grok-gemini-claude-meta-ai-which-is-the-best-ai-assistant-we-put-them-to-the-test>
- Neuberger, J., Ackermann, L., van der Aa, H., & Jablonski, S. (2025). A Universal Prompting Strategy for Extracting Process Model Information from Natural Language Text Using Large Language Models. In W. Maass, H. Han, H. Yasar, & N. Multari (Eds.), *Lecture Notes in Computer Science: Vol. 15238, Conceptual Modeling* (pp. 38–55). Springer Nature.
- Nguyen, B. A. P., & Scholta, H. (2024a). From Text to Model to Execution: A Literature Review on Methods for Creating Conceptual Models from Legal Regulations. In *ECIS 2024 Proceedings*. Association for Information Systems.
- Nguyen, B. A. P., & Scholta, H. (2024b). A Method for the Collaborative and Semi-automated Generation of Conceptual Models from Legal Regulations in Public Organizations. In M. R. Johannessen, C. Csáki, L. Danneels, S. Hofmann, T. Lampoltshammer, P. Parycek, G. Schwabe, E. Tambouris, & J. Ubacht (Eds.), *Lecture Notes in Computer Science: Vol. 14891, Electronic Participation* (pp. 194–208). Springer Nature.
- Pesch, P. J. (2025). Potentials and Challenges of Large Language Models (LLMs) in the Context of Administrative Decision-Making. *European Journal of Risk Regulation*, 16(1), 76–95.
- Prokop, D., Stenclák, Š., Škoda, P., Klímeck, J., & Nečáský, M. (2025). Enhancing Domain Modeling with Pre-trained Large Language Models: An Automated Assistant for Domain Modelers. In W. Maass, H. Han, H. Yasar, & N. Multari (Eds.), *Lecture Notes in Computer Science: Vol. 15238, Conceptual Modeling* (pp. 235–253). Springer Nature.
- Sánchez-Navalón, H., Monserrat, C., Garigliotti, D., & Ferri, C. (2025). Evaluating Performance and Trustworthiness of RAG Systems for Generating Administrative Text. In V. Julian, D. Camacho, H. Yin, J. M. Alberola, V. B. Nogueira, P. Novais, & A. Tallón-Ballesteros (Eds.), *Lecture Notes in Computer Science: Vol. 15346, Intelligent Data Engineering and Automated Learning – IDEAL 2024* (pp. 410–421). Springer Nature.
- Schmidt-Aßmann, E. (2004). Der Rechtsstaat. In J. Isensee & P. Kirchhof (Eds.), *Handbuch des Staatsrechts der Bundesrepublik Deutschland: Band II, Verfassungsstaat* (3rd, pp. 542–612). C. F. Müller.
- Scholta, H., Niemann, M., Delfmann, P., Räckers, M., & Becker, J. (2019). Semi-automatic inductive construction of reference process models that represent best practices in public administrations: A method. *Information Systems*, 84, 63–87.
- Surden, H. (2024). ChatGPT, Large Language Models, and Law. *Fordham Law Review*, 92(5), 1941–1972.  
<https://ir.lawnet.fordham.edu/flr/vol92/iss5/9>
- Tamanaha, B. Z. (2004). *On the Rule of Law: History, Politics, Theory*. Cambridge University Press.
- Tschohl, C. (2020). Zum Verhältnis von Recht und Technik: Rechtsstaatlichkeit durch Technikgestaltung. In W. Hötendorfer, C. Tschohl, & F. Kummer (Eds.), *International trends in legal informatics: Festschrift für Erich Schweighofer: Liber amicorum* (pp. 439–452). Editions Weblaw.
- Urquhart, L. D., McGarry, G., & Crabtree, A. (2022). *Legal Provocations for HCI in the Design and Development of Trustworthy Autonomous Systems*.  
<http://arxiv.org/pdf/2206.07506>
- van Engers, T. M., & van Doesburg, R. (2016). Modeling the interpretation of sources of norms. In C. Granja, R. Oberhauser, L. Stanchev, D. Malzahn, & International Conference on Information, Process, and Knowledge Management (Eds.), *Proceedings of the EKNOW 2016* (pp. 41–50). IARIA.
- Vladika, J., Meisenbacher, S., Preis, M., Klymenko, A., & Matthes, F. (2024). Towards A Structured Overview of Use Cases for Natural Language Processing in the Legal Domain: A German Perspective. *AMCIS 2024 Proceedings*.  
[https://aisel.aisnet.org/amcis2024/sig\\_osra/sig\\_osra/1](https://aisel.aisnet.org/amcis2024/sig_osra/sig_osra/1)
- Wang, J., Zhao, H., Yang, Z., Shu, P., Chen, J., Sun, H., Liang, R., Li, S., Shi, P., Ma, L., Liu, Z [Zongjia], Liu, Z [Zhengliang], Zhong, T., Zhang, Y., Ma, C., Zhang, X., Zhang, T., Ding, T., Ren, Y., . . . Zhang, S. (2024, November 15). *Legal Evaluations and Challenges of Large Language Models*.  
<http://arxiv.org/pdf/2411.10137>
- Weis, S., Montsch, C., Delpechithrage, T., & Nguyen, B. A. P. (2025). From Regulation to Implementation: Understanding the Impact of the EU AI Act on Public Sector Institutions in Germany. In S. Hofmann, L. Danneels, R. Dobbe, A.-S. Novak, P. Parycek, G. Schwabe, V. Spitzer, & J. Ubacht (Eds.), *Lecture Notes in Computer Science: Vol. 15978, Electronic Participation* (pp. 87–101). Springer Nature.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023, February 21). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. <http://arxiv.org/pdf/2302.11382>
- Zeevat, H., & Karagjosova, E. (2009). History and grammaticalisation of "doch"/"toch". *ZAS Papers in Linguistics*, 51, 135–152.
- Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2024). Memorybank: Enhancing Large Language Models with Long-Term Memory. In M. J. Wooldridge, J. Dy, & S. Natarajan (Eds.), *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (pp. 19724–19731). AAAI Press.
- Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 634(8032), 61–68.