# The Role of Conceptual Modeling in Explainable AI: A Legal Domain Case Study

Maxim Bragilovski[†], Din Erza[†], Nir Grinberg[†] and Arnon Sturm[*,†]

Faculty of Computer and Information Science, Ben-Gurion University of the Negev, Beer Sheva, Israel

**Abstract**

The advances of recent years in generative artificial intelligence (AI) have provided ample new means to improve conceptual modeling. Yet, relatively little research has examined how AI solutions can benefit from conceptual modeling. Here, we demonstrate how conceptual modeling can support Explainable AI (XAI) rather than black-box solutions in high-stakes decision-making, thus contributing to the model's interpretability and likelihood of adoption. In particular, we reformulate a complex AI task – finding similar criminal cases – using a conceptual model that facilitates factual and interpretable AI inferences. Currently, attorneys look for similar cases manually, which is time- and resource-consuming, involving many complex comparisons, and resulting in a selection of cases that is potentially biased. Our conceptual model-based solution, in contrast, uses AI to populate values in the conceptual model from the unstructured case text, and learns what makes two cases similar from expert judgment. The findings show that our approach identifies similar cases and outperforms black-box AI solutions by 10.0% in terms of $F_1$ while delivering interpretable results based on the conceptual model.

**Keywords**

Conceptual Model, Explainable AI, Problem Solving, LLM

## 1. Introduction

Conceptual models aim to describe knowledge in a specific domain [1]. Such models may be structural or behavioral and can be used for communication, information systems design and implementation, and knowledge management. Usually, conceptual models are developed at the beginning of projects and provide a static foundation. However, the advancements in artificial intelligence (AI), machine learning (ML), deep learning (DL), and generative AI have increasingly sidelined traditional conceptual models due to their inability to adapt dynamically to the rapidly evolving needs of AI-driven systems. To address this issue, recent advancements in AI are often utilized to support the dynamic and continuous development of conceptual models.

However, it is still unclear how conceptual modeling can co-evolve with and support the development of AI-based solutions. Bork indicates opportunities for introducing conceptual modeling to AI [2]. In particular, he mentions that AI can automate various tasks or services. However, stakeholders within the domain of enterprise systems do not readily adapt to or fully understand AI methods due to their complexity and black-box characteristics. Nevertheless, conceptual modeling can make AI more accessible to non-experts. For example, it enhances transparency by embedding domain knowledge into the AI-based solutions, making its operations more interpretable and aligned with user expectations. Additionally, conceptual modeling can make AI-based solutions more understandable to users without specialized knowledge [2], and enable more effective interaction with these [3]. This applies to other domains that involve decision-making, such as law, medicine, aerospace and defense, and transportation.

ER2025: Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos, October 20-23, 2025, Poitiers, France

*Corresponding author.

†These authors contributed equally.

ezradin@post.bgu.ac.il (D. Erza); nirgrn@bgu.ac.il (N. Grinberg); sturm@bgu.ac.il (A. Sturm)

0000-0002-4778-7897 (M. Bragilovski); 0000-0002-1277-894x (N. Grinberg); 0000-0002-4021-7752 (A. Sturm)

The literature consistently highlights four recurring advantages that make conceptual models an attractive substrate for AI pipelines [4, 5]:

- Enhancing understandability: Translating tacit, expert knowledge into an explicit, shareable artifact improves collective comprehension of the domain.
- Facilitating domain expert-developer collaborations: Decomposing complex problems into modeled concepts creates a transparent workflow and a common vocabulary through which domain specialists and AI developers can communicate.
- Supporting generalization: A clean separation between domain concepts and technical implementation makes it easier to transfer the same modeling approach to other subdomains with only minor adjustments.
- Mitigating errors and bias: A structured representation constrains the AI system to focus on relevant factors, thereby reducing spurious correlations and uneven performance across cases.

This paper presents a case study that investigates the following research question: How can conceptual models (based on their advantages) support explainable AI (XAI) in the legal domain?

We do so by using a conceptual model to derive the entire process of explainable AI (XAI) in a case study from the legal domain, where transparency and interpretability are paramount to the success of the AI-based solution. Specifically, we aim to identify similar legal cases, which is a necessary part of supporting various legal arguments, particularly arguments for an appropriate punishment range in the criminal justice, also known as the sentencing boundary. This process not only requires the support to produce accurate and consistent results but also demands that its outputs be understandable and justifiable to legal professionals. By leveraging a conceptual model, we integrate domain knowledge directly into the AI solution, ensuring that the reasoning behind the suggested sentencing boundaries is sound, compliant with the law, and aligned with legal principles. By relying on a conceptual model that incorporates knowledge from the legal domain, this approach empowers attorneys – some of which may be particularly risk-averse – to understand and possibly trust the recommendations of the proposed solution, fostering both usability and accountability in high-stakes legal decisions.

To devise a solution for finding similar cases, we divide the problem into three sub-tasks: (1) Sentence-level classification, (2) Information extraction, and (3) Similarity determination. For each sub-task, we use Large Language Models (LLM) as well as standard machine-learning (ML) techniques.

Using the conceptual model as an infrastructure for the entire process provided promising results in terms of model accuracy and explainability, and outperformed an alternative approach without a conceptual model. The paper provides evidence to previous studies (e.g., [6, 7, 8]) discussing the potential of using a conceptual model to support AI-based solutions through a case study for a successful usage of a conceptual model when using AI techniques.

The paper is organized as follows. Section 2 introduces and analyzes the state-of-the-art of using conceptual modeling for AI tasks. Section 3 provides the necessary background of the overall task of identifying similar cases in the legal domain. Section 4 details our proposed solution for identifying similar cases using a conceptual model and providing explanations based on it. Section 5 discusses the benefits and limitations of using a conceptual model as reflected in the case study before concluding and outlining agenda for future research in Section 6.

## 2. Related Work

Research at the intersection of conceptual modeling and explainable AI now spans a wide spectrum of application domains, including medical diagnostics, recommender systems, and judicial decision-support, demonstrating the fields relevance beyond a single context. Miller

addresses the need for AI-based solutions to produce explanations for their outputs and decisions, a capability that has become crucial in these areas [9]. The ensuing literature shows a clear evolution in how conceptual models meet that need.

Design support. The first wave of work treated conceptual models as reference blueprints. Caro-Martínez et al. proposed a conceptual model for the design and implementation of recommender systems [10]. Langer et al. proposed a conceptual model that adopts stakeholders' perspectives to guide XAI research [11]. Van Den Berg et al. devise a conceptual model of categories of aspects and relationships relevant to the development of XAI [12].

Development aids. A second strand integrated these artifacts into the modeling workflow. Lukyanenko et al. wove goal diagrams, ER models, and BPMN through every CRISP-DM stage [13] in a foster-care drug monitoring project, thereby exposing data gaps and guiding feature engineering [7]. They later introduced superimposition, which projects learned feature weights onto domain concepts so practitioners see category-level explanations rather than opaque numbers [6].

Embedded explanatory layers. Moving beyond overlays, Maass et al. showed the Model Embedding Method, which embeds entire ML models inside conceptual structures to compute "concept contributions" and diagnose where observed behavior diverges from expert knowledge [8]. Maass et al.'s Conceptual Alignment method provides an iterative solution that adjusts both the conceptual model and the ML model until predictive consistency is maximized [14].

This progression from reference blueprints, to workflow aids, to tightly coupled explanatory layers shows how conceptual modeling has evolved from static documentation into an active engine for generating and validating explanations in modern XAI, however gaps remain around the standardization of evaluation metrics and the scalability of these methods, and there is limited clarity on how well they generalize across domains or constantly changing domain.

Bork [2] complemented these lines of work with a four-way taxonomy that classifies how conceptual- and AI-techniques combine: (i) combining existing techniques from both fields, (ii) combining new conceptual models with existing AI techniques, (iii) combining existing conceptual models with new AI techniques, and (iv) combining new techniques from both fields.

Building on Bork's categorization, most existing research can be neatly placed in the second category. For example, Bragilovski et al. examined multiple AI techniques to derive domain-specific conceptual models from user stories [15]. Relatively little research has investigated the potential of conceptual models to help AI systems, particularly in explainable AI. Maass's pioneering work provides compelling arguments for a paradigm shift where conceptual models transition from being primarily design tools to instruments of explanation. It underscores the importance of conceptual models in bridging the gap between the complexity of AI systems and human comprehension [16].

In the concept-based explainability literature, the term "concept" encompasses various abstractions, including symbolic concepts, unsupervised concept bases, prototypes, and textual concepts [17]. These categories serve different roles in XAI: symbolic concepts are human-defined attributes (e.g., colors or shapes), while unsupervised concepts emerge from data-driven clustering. Prototypes represent characteristic examples, and textual concepts leverage generative models like LLMs to bridge textual descriptions.

A promising new direction in concept-based models is using LLMs to come up with concept representations, eliminating the need for manual annotation [17]. These models align textual concepts generated by LLMs with latent representations of input data to produce concept scores that inform final classifications. Two key methods in this area, Language-guided Bottlenecks [18] and Label-free Concept Bottleneck Model [19], illustrate the potential of this approach to integrate interpretability into AI systems without sacrificing performance.

LaBO [18] constructs a "concept bottleneck layer" to associate importance weights with interpretable concepts, enabling users to understand the rationale behind AI predictions. LabelFree-CBM [19] takes this a step further by leveraging GPT-3 to generate concepts dynamically, eliminating the need for manual annotation while maintaining interpretability. Both models

rely on conceptual structures to organize and present their internal reasoning, validating their effectiveness through user studies. Barbiero et al. [20] propose an entropy-based explainability framework that integrates conceptual models directly into the neural network architecture to provide First-Order Logic (FOL) explanations.

Unlike earlier studies, our work emphasizes the development of textual concepts and showing their value as features for predictive models. Prior research often treated concepts as either pre-defined or easily extractable, whereas we develop concrete methods for constructing them and evaluate their contribution to the prediction against straightforward approaches (generative models). In Borks taxonomy, this aligns with class (iii): existing conceptual structures enriched with new AI techniques. By focusing on concept creation rather than assuming it, we highlight how conceptual models can both improve interpretability and enhance predictive performance in XAI.

## 3. The Problem Domain

The criminal law defines what actions constitute a crime and provides guiding principles for punishing criminal actions. Over the years, legal systems around the world have changed how they determine what an appropriate and reasonable sentencing decision is [21]. For example, prior to the Comprehensive Crime Control Act of 1984, federal U.S. judges had full discretion over sentencing decisions, which was heavily criticized for the large discrepancies in sentencing decisions, sometimes even for the same crime. [22]. Following the legislation, federal sentencing guidelines were developed, first imposing mandatory minimum and maximum sentences for certain crimes and circumstances, then overturned by the Supreme Court as advisory and non-binding recommendations for judges to follow. Over the years, many legal systems adopted the structuring of sentencing decisions. For example, the criminal law in Israel was amended in 2012 to require courts to specify and justify the appropriate type and range of punishment based on three aspects of the case: the social value damaged by the commission of the crime, the degree of damage to the social value, and the punishment policy used in similar cases. While the first two aspects map relatively easily to decisions, determining the similarity of a given case to previous verdicts is a laborious case-by-case kind of task that requires both breadth of coverage and depth of understanding of the nuances of each case.

Currently, lawyers look for similar cases manually. This may involve the manual collection of local spreadsheets of past cases and going through them to find similar cases. Alternatively, attorneys may use digital keyword search and sieve through the result list. A third option is to send the query to peers. All of these existing options are error-prone, rely on human recollection, are labor-intensive, and are limited in coverage.

In this paper, we focus on the domain of weapon-related verdicts to demonstrate the success of our methodological approach in a relatively simple sub-area of criminal justice, before moving to more complex offenses like homicide or fraud. Still, the study of weapon-related cases presents many significant challenges due to the diversity of weapon-related circumstances, nuances that carry weight for sentencing decisions, regulations, procedures, the language (Hebrew), a limited number of publicly available cases (due to privacy constraints), and a lack of an abundant ground truth for supervised learning. To address these challenges, we devise a conceptual model and learning procedures that can cope with these challenges as detailed next.

## 4. Conceptual Models for AI Explainability

As LLMs become more capable, it is interesting to examine how well these general-purpose models do in identifying similar cases. Indeed, we tested these capabilities using embeddings (using

fasttext[1]) and GPT-based[2] models. In both cases, the models provided unsatisfactory results that suffer from two key shortcomings: (i) they have identified surface-level lexical similarity rather than legally salient factors (e.g., weapon status, purpose, etc.); (ii) they provided little grounding for the result. So, injecting cases into such AI-based solutions without any guidance (i.e., the conceptual model) achieved poor results. We therefore introduce a conceptual layer that converts raw text into domain concepts before any similarity calculation, enabling both higher accuracy and factual explanations. Particularly, we decided to divide the problem into three sub-problems: (i) sentence classification, (ii) feature extraction, and (iii) case similarity determination, which are executed sequentially. In the following subsections, we elaborate on each of the steps.

### 4.1. The Conceptual Model

Breaking down the process into smaller sub-tasks reduces the complexity of the overall task, but still requires specifying what should be considered as a similar case. For that purpose, we consulted three district attorneys from the Ministry of Justice in Israel to develop a schema for determining case similarity. The development of such a schema required the attorneys to articulate and "formulate" how they think about case similarity. In particular, they started from a general classification and, downstream of the process, they provide additional information for each classification. This results in a hierarchical classification of the factors used to determine similarity. Based on this classification, we enrich the schema with the required information from each class, the way to obtain it, and, for relevant information types, consider their values along with their ordering (for example, in the case of weapon status, "dismantled" is less severe than "separated from ammunition"). The schema and associated information form a conceptual model that serves as an anchor for each of the stages. The conceptual model consists of the category of sentences (label), the information required to determine and explain case similarity and its related values (when applicable), and questions (or prompts) related to the required information.

Table 1 presents (part of) the specific conceptual model we used for the weapon-related domain. The first level classification consists of five categories: Offense circumstances, confession, punishment, general circumstances, and not relevant. The first category is then split into other sub-categories: weapon type, weapon status, purpose, use, held way, and more. The sub-category name further implies the required information. Other categories may include more/other information needs. We devised a prompt to extract this information using an LLM. For example, the prompt for weapon type was the following: "What is the type of the weapon? Answer from the following answers without .....". Finally, for each information item, we created after consultation with the attorneys a list of possible values, sorted in increasing order of severity. For example, a pistol is less severe than a submachine gun.

During the development of the conceptual model, the attorneys were positively surprised by the ability of this approach to systematically analyze case similarity. They usually examined cases holistically, having difficulties crystallizing the similarities and differences. Explicating (through the conceptual model) the factors affecting the similarities among cases had its own benefits, regardless of the automation proposed next.

In the case of the weapon domain, the conceptual model is quite simple, yet it provides evidence that it is quite beneficial to adopt such an approach, as we elaborate in this paper. Furthermore, we anticipate that the hierarchical approach we adopted would allow us to deal with complex conceptual models, as well.

We then generalize the conceptual model in the form of meta-model, so we can use it for other domains. The meta-model appears in Figure 1. The category class refers to the labels according to which we will classify sentences. The self-association serves for the hierarchical

[1]https://fasttext.cc/
[2]https://openai.com/

**Table 1**
The weapon-related conceptual model

| Category (label) | Sub-Category | Question/Prompt | Values |
|---|---|---|---|
| Offense Circumstances | Weapon Type | What is the type of weapon? Answer from the following answers without another word: [Pistol, ...] | 'Pistol', ' submachine gun', ... |
| | Weapon Status | I'm going to ask you a question based on the following text: Clarification – if the weapon is loaded, it means loaded in the cartridge in the insert, and if the weapon is loaded, then it means that it has a bullet in the barrel. "{text}" Is the weapon an {option}? Answer [yes, no] only. | 'dismantled', 'separated', ... |
| | Purpose | Is the purpose of using the weapon {option}? Answer with [yes, no] only. | 'wedding', 'conflict', 'self-defense'... |
| | Use | I am going to ask you a question based on the following text, and then you will answer: "{text}" Did the accused perform {option}? Answer with [yes, no] only. | ... |
| Confession | - | Did the accused plead guilty? Answer only [yes, no] only. | - |
| Punishment | Range | - | - |
| Punishment | Punishment | - | - |
| General Circumstances | - | - | - |
| Not Relevant | - | - | - |

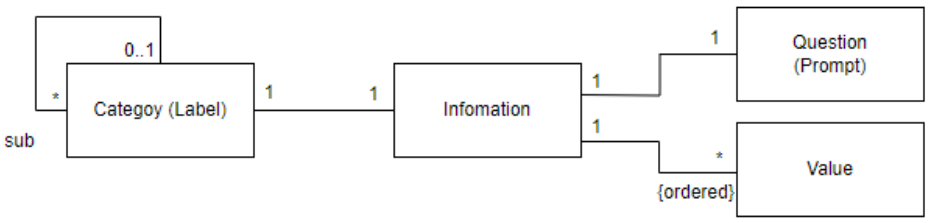


**Figure 1:** The conceptual meta-model

category we propose (see the example in Table 1, the columns of category and sub-category). The information class refers to the element that we would like to extract from the sentences. The question class is used for the means of extracting the element, and the value class holds the ordered list of possible values (for the sake of determining the similarity).

In the following, we describe the three stages that utilize the conceptual model. As mentioned before, the execution of the various stages without the conceptual model achieved poor (random) results.

## 4.2. Sentence Classification

To address the challenge of sentence classification, we adopted a two-step hierarchical approach to maximize the quality of the results. The classification followed the conceptual model that appears in Table 1. For brevity, we elaborate here only on one thread of the classification (offense circumstances), while in practice, we had two levels of classification processes. The first process focused on associating each sentence with one or more of the five predefined categories (labels). For each label, we trained a binary classifier to predict whether a sentence is associated with it.

Since sentences often carry multiple labels simultaneously, we adopted a multi-label classification approach.

Before training the classifiers, we manually tagged 3,200 sentences drawn from 137 criminal judgments (cases). The cases were retrieved from Nevo[3] under the following two criteria: (1) The decision was handed down in 2018 or later; (2) the most serious offence in each case was the illegal purchase, sale, use, or possession of a weapon. We then assessed intercoder reliability. Measured using Cohen's Kappa, we found a high (Kappa $>$ 0.8) inter-coder reliability, indicating that people can consistently classify the sentences. After examining several transformer-based models as well as larger language models, we decided to train the classifiers using SetFit [4] [23] with dictaBert[5]. For the second level, we applied the same technique. However, each second-stage classifier operated only on sentences that passed the classification of the first-stage classifier.

For instance, consider the sentence: "The defendant was charged with trafficking in M16 weapons," which describes the type of weapon (M16) and how it was used (trafficked). To classify this sentence, we first utilized the five classifiers from the first level, which indicated correctly that the sentence describes the Offense Circumstances. Then, second-stage classifiers within the Offense Circumstances sub-category are applied to identify that the sentence contains information about Weapon Type and Use.

We used the following methodology to train and test the classifiers: Using the sentences from the 137 cases we classified, we divided those into groups of 12 cases (resulting in approximately 280 sentences each); we then trained the classifiers on each group and tested the results with respect to all sentences from the other groups.

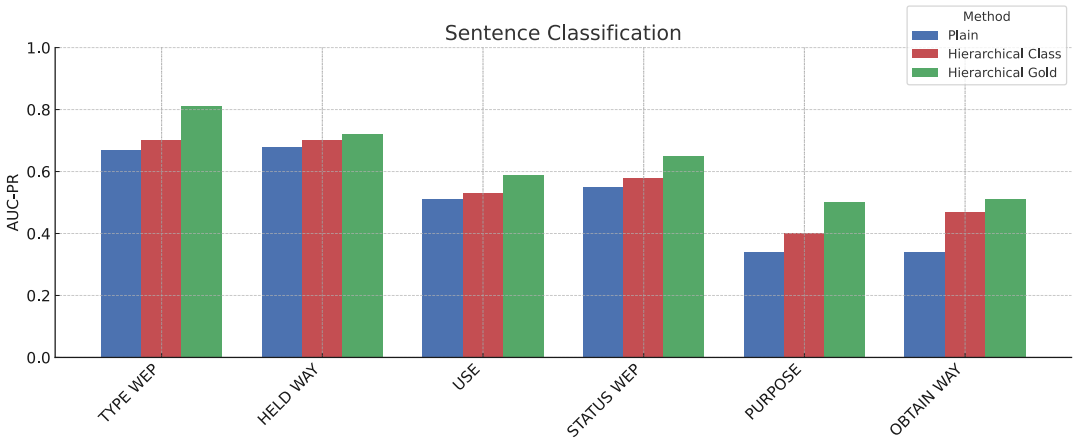


**Figure 2:** Results of the sentence classification stage

The results are presented in Figure 2. The graph shows the PR-AUC[6] of the categories of the conceptual model in Table 1. The results indicate that the strategy of having multi-level classification achieves the best results[7]. The "Plain" bins show the results when classifying the sentences only by the second-level classifier (without applying first-stage classifiers). The "Hierarchical Class" bins show the results when classifying the sentences with the classifier of both levels. The "Hierarchical Gold" bins show the results of classifying sentences by the second-stage classifiers when applying only to manually coded sentences of the first stage. We also observed

[3] https://www.nevo.co.il/

[4] SetFit is a framework that contrastively fine-tunes any BERT-style sentence encoder, enabling accurate few-shot classification with minimal compute.

[5] https://huggingface.co/dicta-il/dictabert

[6] We also used precision, recall, F1-score, however, PR-AUC is particularly well-suited in this context because it focuses on the positive, often rare, class while balancing precision and recall.

[7] Although the hierarchical architecture outperforms the flat baseline in our current dataset, supplementary experiments with a much larger set of newly-tagged examples showed that the performance gap narrows considerably as the amount of labeled data grows.

that low PR-AUC values are associated with categories that have a limited number of sentences, which limits the learning.

> Quantitative Finding 1. Hierarchical approach consistently outperforms other approaches across categories. This provides a clear indication that conceptual modeling (with respect to the categories) can improve the performance of an AI-based solution in enhancing explainability.

### 4.3. Information Extraction

Once the sentences are classified, we distill each one into the case details required by the conceptual model, capturing key facets such as weapon type and offence circumstances. Every labeled sentence becomes a structured attribute that fuels downstream tasks like case-similarity analysis. For each second-level category, the model specifies (i) the exact piece of information to extract and (ii) the guiding question that locates it in the text.

To set the ground truth and validate that task, based on the conceptual model, we manually extracted the related information from the 137 cases.

We experiment with several information extraction techniques, including regular expressions which served as a baseline, and LLMs. In particular, we tested three LLMs that can handle text in the Hebrew language: dicta2.0[8], C4AI[9], and Claude[10].

Based on the prompt engineering for each category (and related information), we executed the prompt to retrieve the related information. For example, for the Weapon Type information, the following prompt was used: "What is the type of weapon? Answer from the following answers without another word [Pistol, submachine gun, improvised submachine gun, Molotov cocktails, explosive device, grenade, assault rifle, stun/gas grenade, LAW missile, Matador missile, hunting rifle, sniper rifle, improvised explosive device, rifle impromptu storm]" Applying the prompt to the sentence: "The defendant was charged with trafficking in M16 weapons", the result is a submachine gun. Note that the result may have a single item, yet in many cases, the results will contain a list of values. Injecting all sentences into their related prompts, following the conceptual model, results in a feature vector as appears in Table 2.

**Table 2**
An example of a feature vector.

| **Case ID** | ME12 |
|---|---|
| **Weapon Type** | Improvised Submachine Gun |
| **Weapon Status** | Separated from ammunition |
| **Purpose** | Conflict |
| **Use** | Shooting |
| **Held Way** | on his body, in the car, hidden |
| **Obtained Way** | - |

The results of extracting the features appear in Table 3. The results are calculated using Dice coefficient, which measures the overlap between sets of values. Where the sets, in our case, comprise relevant values, as appears, for example, for the "Held Way" category in Table 2. A dice coefficient higher than 0.6 is considered at least of medium quality. From the results it can be seen that the larger LLMs provided superior results, and specifically that dicta2.0 provided the best results.

[8]https://huggingface.co/dicta-il/dictalm2.0
[9]https://huggingface.co/CohereForAI/
[10]https://www.anthropic.com/api

Quantitative Finding 2. By identifying the key features and detailing how to capture them, the conceptual model becomes a valuable artifact for extracting relevant information, thereby markedly enhancing explainability.

**Table 3**
Results of the feature extraction by method (Dice coefficient).

| Method | Weap. Type | Weap. Stat. | Purpose | Use | Held Way | Obt. Way | Avg. |
|---|---|---|---|---|---|---|---|
| Dicta2.0 | 0.75 | 0.40 | 0.67 | 0.94 | 0.71 | 0.70 | 0.70 |
| Claude 3.0 | 0.72 | 0.43 | 0.73 | 0.82 | 0.59 | 0.70 | 0.67 |
| C4AI | 0.77 | 0.36 | 0.55 | 0.63 | 0.40 | 0.49 | 0.53 |
| RepEx | 0.13 | 0.41 | 0.57 | 0.53 | 0.03 | 0.52 | 0.37 |

## 4.4. Case Similarity

The previous stage populated feature vectors for cases as demonstrated in Table 2. Following the vectors created for the 137 manually-labeled cases, we sampled 156 pairs, which we manually tagged for similarity on a 1-5 scale, where a value of 1 represents non-similar cases, 3 represents somewhat similar cases, and 5 represents highly similar cases. We then converted the similarity scale into a binary variable where pairs with a score of 3 or above were considered similar, so as to decide whether to present them to the attorneys as similar cases.

We tried two different approaches for identifying similar cases. First, we tried a supervised learning approach using a Random Forest model. The model received vectors of dice coefficients capturing the similarity between the feature vectors of the two cases. Table 4 shows examples of highly similar (M15, M10) and non-similar (ME21, ME11) pairs of cases. We trained the Random Forest model using leave-one-out and tested its performance on held-out pairs. To turn model scores into binary classifications, we use a threshold of 0.76 (of the probability of the pair being similar) that maximizes performance. Analyzing the result using SHAP, we found out that the information that mostly affected the indication of non-similar cases is related to the weapon type, the weapon status, and the held way.

**Table 4**
Examples of Case Similarities

(a) Highly Similar Cases

| Case ID | M15 | M10 | Dice Coef. |
|---|---|---|---|
| Weapon Type | Pistol | Pistol | 1.00 |
| Weapon Status | OK | Separated from ammunition | 0.00 |
| Purpose | Self-defense | Self-defense | 1.00 |
| Use | No | No | 1.00 |
| Held Way | In the car, Hidden | In the car | 0.90 |
| Obtained Way | - | - | 1.00 |

(b) Non-Similar Cases

| Case ID | ME21 | ME11 | Dice Coef. |
|---|---|---|---|
| Weapon Type | Imp. Submachine Gun | Pistol | 0.24 |
| Weapon Status | With a bullet in the barrel | Separated from Ammu. | 0.62 |
| Purpose | Conflict | Self-defense | 0.40 |
| Use | Shooting | - | 0.00 |
| Held Way | On his body, In the car | At home, In the car | 0.81 |
| Obtained Way | - | - | 1.00 |

The second approach for identifying case similarity involved Zero-shot learning with an existing LLM (Claude). To formulate the case similar task to the LLM, we created a prompt that includes

To determine their similarity, you are tasked with comparing two law cases based on their property vectors. Your goal is to analyze the characteristics of each case, identify similarities and differences, and provide a reasoned assessment of how similar the cases are in general. The purpose of similarity is to help the lawyer determine whether the cases are similar and whether punishment can be based on them. Here are the property vectors for the two cases:

| Category | Details |
|---|---|
| **Weapon Type** | Case 1: *"pistol"*<br>Case 2: *"improvised submachine gun"* |
| **Weapon Status** | Case 1: *"functional"*<br>Case 2: *"separated from ammunition"* |
| **Holding Location** | Case 1: Only *"on body"*<br>Case 2: *"in vehicle"*, *"hidden"*, *"on body"* |

Based on your analysis, determine if the cases are similar or not. Present your response in the following format:

```
<analysis>
[Your detailed analysis of similarities and differences]
</analysis>
<similarity_score>
[Your similarity score is 0 or 1]
</similarity_score>
<justification>
[A brief justification for your similarity score]
</justification>
```

**Figure 3:** The prompt used with Claude to determine case similarity.

<analysis>
1. **Purpose**: Both cases involve a *"conflict"*, indicating a similar motivational context.
2. **Weapon Use**: Both cases involve *"shooting"*, showing active weapon usage.
3. **Holding Method**: Both cases include *"on his body"* as a weapon carrying method.

**Differences:**

| Category | Details |
|---|---|
| **Weapon Type** | Case 1: *"pistol"*<br>Case 2: *"improvised submachine gun"* |
| **Weapon Status** | Case 1: *"functional"*<br>Case 2: *"separated from ammunition"* |
| **Holding Location** | Case 1: Only *"on body"*<br>Case 2: *"in vehicle"*, *"hidden"*, *"on body"* |

</analysis>
<similarity_score> 1 </similarity_score>
<justification>
Despite some nuanced differences in weapon type and specific offense categorization, the cases are fundamentally similar. They share critical characteristics: a conflict-driven purpose, active weapon use, similar weapon-carrying methods, and core weapon-related offenses. The substantial overlap in key legal and contextual properties strongly supports treating these cases as similar for legal assessment and potential punishment considerations.
</justification>

**Figure 4:** An example of a generated result from Claude

a short description of the task and the feature values of the two cases. The prompt asked the model to determine the similarity of the two cases as well as to provide detailed justification for feature-level similarities and dissimilarities and for pair-level determination. Figure 3 shows an example of this prompt, and Figure 4 is an example of its response. As shown by previous work [24], asking LLMs to provide step-by-step justification for their responses improves accuracy.

Interestingly, the zero-shot approach, involving no training, yields superior results over the supervised approach. For the zero-shot approach, both the precision and the recall reached 0.68. For the Random Forest, the precision reached 0.7 and the recall reached 0.57. The results indicate that the zero-shot approach (using Claude) facilitated the retrieval of more relevant cases with negligible lower precision. This means that the zero-shot approach saves the attorneys time in allocating more similar cases.

The explanations from the RF execution are derived by following the rules derived from the model results. The explanations from Claude explicitly appear in its output. The results and the explanation are directed by the conceptual model.

Quantitative Finding 3. Deploying a zeroshot language model to judge case similarity surpassed the supervised baseline while supplying clear, feature-level rationales. These explanations enhance understandability, and fostering domain experts-developers collaboration.

**Pipeline Error-Propagation Analysis.** As a final step, we assessed the robustness of the full pipeline and examined how errors cascade from one stage to the next. For each sub-task we used the best-performing model identified in earlier experiments: SetFit with DictaBERT and hierarchical decoding for sentence classification, Dicta 2.0 for feature extraction, and Claude for case similarity.

Figure 5 reports case-similarity results under four feature configurations: (1) Oracle upper bound (orange) - the similarity model receives gold-standard features (i.e., the ground-truth output of the feature-extraction stage); (2) Perfect sentences (blue) - features are extracted from gold-standard sentences; (3) Full pipeline (purple) - features are extracted from the sentences predicted by the classification model; (4) FastText baseline (green) - similarity is computed directly from raw FastText document embeddings (without the conceptual model).

The figure highlights the sequential dependency of the pipeline: performance ($F_1$ and accuracy) systematically declines whenever upstream predictions replace gold data. Precisionrecall curves further reveal that FastText attains high recall but very low precision, effectively labelling most case pairs as similar. Overall, FastText records the poorest precision and accuracy, whereas our structured pipeline yields substantial gainsabout 20% in accuracy and roughly 10% in $F_1$when moving from the FastText baseline to features produced by the sentence-classification stage. These results underscore the value of the modular approach for improving AI performance on the case-similarity task.

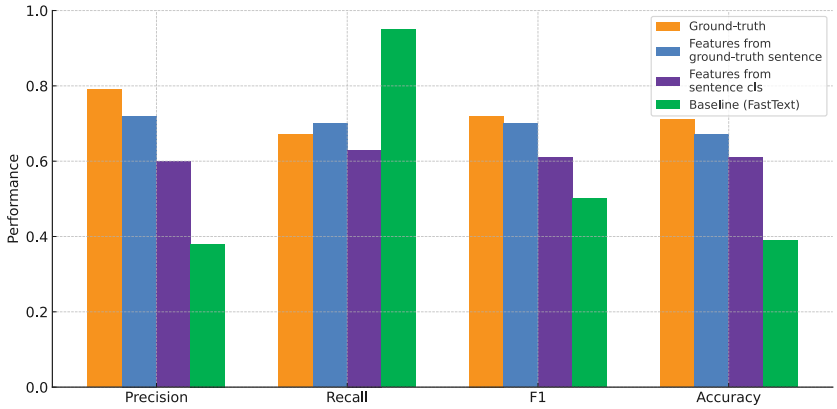


**Figure 5:** Case-similarity performance with four alternative feature sources.

## 5. Discussion

First, we analyze our results through the lens of the four key advantages of conceptual modeling: enhanced understandability, facilitated domain experts-developers collaboration, improved generalization, and reduced errors and bias from a qualitative perspective. Second, we unpack the practical challenges that emerged during implementation.

### 5.1. Qualitative Perspective

Qualitative Finding 1. The conceptual model enhanced domain understandability as it rendered the legal context behind each verdict explicit and shareable.

This insight emerged from a series of conversations with practicing attorneys. As they co-constructed the model, tacit legal knowledge precedents, statutory nuances, and case-specific factors were externalized into a coherent schema. The participants remarked that the weapon-offense model distilled scattered principles and fine-grained distinctions into a single, dependable reference, preserving insights that would otherwise have remained undocumented or overlooked.

> Qualitative Finding 2. The conceptual model facilitated domain experts-developers collaboration as it decomposed the problem into transparent, manageable components and provided a shared vocabulary for attorneys.

By translating legal concepts into a structured schema, the model served as a bridge between attorneys and the AI pipeline. Attorneys, despite their limited technical background, reported that they could (i) inspect intermediate outputs, (ii) follow how each model component shaped the final recommendation, and (iii) discuss issues with the development team in familiar legal terms. This level of transparency nurtured trust, improved usability, and, in their words, enabled them to apply AI-generated insights more confidently in practice.

> Qualitative Finding 3. The conceptual model supported generalization as its structured framework could be transferred to adjacent legal domains with minimal effort.

Although our case study refers to weapon-related offenses, the participating attorneys quickly repurposed the model for drug-related cases by adjusting only a handful of categories and required information. Their experience suggests that the schema's clear separation between domain concepts and technical implementation makes it readily adaptable to other branches of law such as civil or administrative proceedings, affirming the models versatility and broad applicability. To verify this portability, we co-created a dedicated drug-offense schema with the attorneys; the entire adaptation took only a fraction of the time required to build the original model, confirming the approachs versatility and efficiency.

> Qualitative Finding 4. The conceptual model mitigated errors and bias as its structured pathway guided the AI toward legally relevant evidence and away from spurious patterns.

By explicitly encoding the legal attributes that drive sentencing similarity, the schema prevents the AI from seizing on spurious lexical quirks or term-frequency artifacts. Embedding-only methods, which weight every token equally, often align cases on superficial stylistic features-such as a judge's prose, rather than on substantive facts. In contrast, our approach reduces false positives by focusing on features highlighted by attorneys. This focused reliance helps shield the model from biases that can emerge from case-specific variables, such as demographic factors.

### 5.2. Practical Challenges

Nevertheless, using the conceptual model as the core asset introduces various challenges:

- Evolving Domain Knowledge - One significant challenge lies in maintaining the relevance of conceptual models in dynamic domains. For instance, discussions with attorneys highlighted the continuously evolving nature of legal systems, driven by factors such as new legislation, shifts in judicial practices, or changes in court systems. These developments necessitate regular updates to the conceptual models to maintain their accuracy and applicability. Without such updates, static models risk becoming outdated, compromising both their effectiveness and the trust users place in them.
- Cost and Resource Intensiveness - Developing and refining the "ultimate" conceptual model is a resource-intensive process that demands time and effort. It often involves multiple

iterations and extensive collaboration to bridge knowledge gaps among domain experts. This iterative procedure requires substantial investment from both experts and technical teams, highlighting the need for efficient resource allocation and management.

- Bias in Model Development - The development of a conceptual model is inherently shaped by the subjective perspectives of domain experts, such as attorneys, who may bring biases rooted in their individual experiences or interpretations. This challenge became particularly evident in our work when attorneys were tasked with tagging sentences or ranking the importance of concepts for similarity calculations. Such biases, if left unchecked, can inadvertently be embedded into the model, potentially distorting AI interpretations and influencing decision-making processes in unintended ways. To mitigate these risks, it is crucial to involve a diverse group of experts with varied perspectives and to adopt fairness-aware methodologies.
- Abstraction Limitations - By design, conceptual models abstract complex domain knowledge into manageable concepts. However, this abstraction can overlook critical details, which might lead to oversimplified or inaccurate AI decisions.

## 6. Summary

This paper highlights the pivotal role of conceptual models in enhancing XAI, particularly within the domain of identifying similarities in weapon-related legal cases. By structuring domain knowledge, the conceptual model not only facilitated AI techniques in delivering clear and actionable explanations but also guided the technical processes for extracting and organizing critical information. This approach showcased key benefits, including enhanced understandability, improved collaboration between domain experts and AI developers, and effective error mitigation, making it a useful approach for addressing complex, high-stakes domains.

In the future, we plan to test the support of conceptual models across other domains, such as civil or administrative law, and test their scalability in dynamic environments. Additionally, we plan to refine AI techniques to better integrate with and leverage the strengths of conceptual models.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly to grammar and spelling check, paraphrase and reword, and improve writing style.

## References


[1] A. Olivé, Conceptual modeling of information systems, Springer Science & Business Media, 2007.

[2] D. Bork, Conceptual modeling and artificial intelligence: Challenges and opportunities for enterprise engineering, in: Advances in Enterprise Engineering XV, Springer International Publishing, Cham, 2022, pp. 3–9.

[3] Q. Yang, J. Suh, N.-C. Chen, G. Ramos, Grounding interactive machine learning tool design in how non-experts actually build models, in: Proceedings of the 2018 designing interactive systems conference, 2018, pp. 573–584.

[4] Y. Wand, R. Weber, Research commentary: information systems and conceptual modelinga research agenda, Information systems research 13 (2002) 363–376.

[5] C. Woo, The role of conceptual modeling in managing and changing the business, in: International conference on conceptual modeling, Springer, 2011, pp. 1–12.

[6] R. Lukyanenko, A. Castellanos, V. C. Storey, A. Castillo, M. C. Tremblay, J. Parsons, Superimposition: augmenting machine learning outputs with conceptual models for explainable ai, in: Advances in Conceptual Modeling: ER 2020 Workshops, Vienna, Austria, November 3–6, 2020,, Springer, 2020, pp. 26–34.
[7] R. Lukyanenko, A. Castellanos, J. Parsons, M. Chiarini Tremblay, V. C. Storey, Using conceptual modeling to support machine learning, in: Information Systems Engineering in Responsible Information Systems: CAiSE Forum 2019, Rome, Italy, June 3–7, 2019, Proceedings 31, Springer, 2019, pp. 170–181.
[8] W. Maass, A. Castellanos, M. Tremblay, R. Lukyanenko, V. C. Storey, Ai explainability: Embedding conceptual models (2022).
[9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[10] M. Caro-Martínez, G. Jiménez-Díaz, J. A. Recio-García, Conceptual modeling of explainable recommender systems: An ontological formalization to guide their design and development, J. Artif. Int. Res. 71 (2021) 557589. doi:10.1613/jair.1.12789.
[11] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)? a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473. doi:https://doi.org/10.1016/j.artint.2021.103473.
[12] M. Van Den Berg, O. Kuiper, Y. Van Der Haas, J. Gerlings, D. Sent, S. Leijnen, A conceptual model for implementing explainable ai by design: Results of an empirical study, in: HHAI 2023: Augmenting Human Intellect, IOS Press, 2023, pp. 60–73.
[13] J. Jackson, Data mining; a conceptual overview, Communications of the Association for Information Systems 8 (2002) 19.
[14] W. Maass, A. Castellanos, M. C. Tremblay, R. Lukyanenko, V. C. Storey, J. S. Almeida, Conceptual alignment method., in: AMCIS, 2023.
[15] M. Bragilovski, A. T. Van Can, F. Dalpiaz, A. Sturm, Deriving domain models from user stories: Human vs. machines, in: 2024 IEEE 32nd International Requirements Engineering Conference (RE), IEEE, 2024, pp. 31–42.
[16] H.-G. Fill, J. Cabot, W. Maass, M. Van Sinderen, Ai-driven software engineering–the role of conceptual modeling, Enterprise Modelling and Information Systems Architectures (EMISAJ) 19 (2024).
[17] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based explainable artificial intelligence: A survey, arXiv preprint arXiv:2312.12936 (2023).
[18] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, M. Yatskar, Language in a bottle: Language model guided concept bottlenecks for interpretable image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19187–19197.
[19] T. Oikarinen, S. Das, L. M. Nguyen, T.-W. Weng, Label-free concept bottleneck models, arXiv preprint arXiv:2304.06129 (2023).
[20] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, S. Melacci, Entropy-based logic explanations of neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 6046–6054.
[21] J. F. Pfaff, The evolution of sentencing policy: An analytical history of the role of politics and public opinion, Journal of Legal Studies 44 (2015) 37–78.
[22] D. Weisburd, A. Petrosino, G. Mason, Design sensitivity in criminal justice experiments, Crime and Justice 17 (1993) 337–379. URL: http://www.jstor.org/stable/1147554.
[23] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).
[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.