Creating an argumentation-based system for fraud detection in law enforcement with LLM support: a case study and focus group

Roos Scheffers[1,*]

[1]Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands

Abstract

This qualitative exploratory study examines how participants approach the creation of an argumentation-based system for classifying fraud in online citizen reports, comparing groups that did and did not use large language models (LLMs). The study found that building an argumentation-based system is cognitively demanding because it requires several interdependent decisions at once, such as defining legal concepts, choosing an appropriate level of abstraction, and managing the complexity of a system. LLMs were found not to be useful for directly translating legal text or examples into argumentation rules. However, they were useful as an onboarding and exploration tool. The group using an LLM was able to start faster and felt less overwhelmed initially.

Keywords

Argumentation, Focus Group, Case Study, Fraud Detection, Argumentation-based system, Expert System

1. Introduction

In this study, we conduct a qualitative exploratory investigation to study how participants construct an argumentation-based system for fraud classification, with and without access to an LLM, by making observations during a task and conducting a focus group.

Argumentation-based systems are widely regarded as well-suited to legal reasoning because they naturally represent defeasibility, conflict, and rule priorities [1]. Structured approaches such as ASPIC+ [2] have shown how legal norms and evidential reasoning can be represented computationally. Argumentation has also been used for explainable legal AI [3]. However, despite decades of theoretical development, argumentation-based systems remain uncommon in legal or forensic practice. Early legal expert systems, such as HYPO [4], already revealed the difficulty of translating legal knowledge into formal rules—often described as the knowledge acquisition bottleneck [5]. Since the power of expert systems lies in their knowledge bases, successful systems require that knowledge moves from the minds of experts into formal rules. This process can be slow and difficult, therefore creating a barrier to the implementation of these systems.

Recently, advancements in Large language models (LLMs) have renewed interest in argumentation-based systems, since they offer potential solutions to overcome the knowledge acquisition bottleneck [6]. LLMs could support the creation of expert systems in multiple ways, for example, by assisting human experts in formulating rules, synthesizing rules from documents [7], and by updating existing expert systems based on new cases [6]. In the current paper, we use the case study of creating a tool for determining fraud from citizens' reports to the Dutch police, which has been inspired by [8, 9]. Within this case study, we conduct a qualitative exploratory investigation to study how participants construct an argumentation-based system for fraud classification, with and without access to an LLM, by making observations during a task and conducting a focus group.

We found that building an argumentation-based system required participants to make several interdependent decisions simultaneously: defining legal concepts, choosing appropriate specificity of rules,

RELAF'26: Reasoning with Evidence in Law Enforcement and Forensics, Workshop at ICAIL 2026, Singapore
*Corresponding author.
r.j.scheffers@uu.nl (R. Scheffers)
0009-0008-1624-3046 (R. Scheffers)

determining sufficient evidence for decisions, and managing interactions between rules. This study found that considering these interdependent tasks simultaneously is cognitively demanding. LLMs were useful as an onboarding and exploration tool, but were not useful in the actual process of creating rules. The group using LLMs started faster and felt less overwhelmed, but the rules generated by LLMs were often too specific to their training examples. These findings suggest that better methodological and tool support is needed for effective LLM usage during rule creation and to reduce cognitive load during the system creation process. By documenting how and why participants struggled, this study aims to inform the design process of the creation of knowledge-based systems in legal and policing domains, and the potential role of LLMs in this process.

The remainder of this paper is structured as follows. Section 2 describes the method, including the participants, the tool used, and the procedure. Section 3 presents the results of the study, focusing on the processes followed by the two groups, the rule sets they produced, and the outcome of the focus group. Section 4 discusses the findings and provides recommendations for the design process of argumentation-based systems. Finally, Section 5 concludes the paper.

2. Method

This study adopts an exploratory qualitative design to investigate how participants construct an argumentation-based system for fraud classification, and how access to a large language model (LLM) influences this process. Participants worked in small groups on a system-building task, followed by a focus group discussion aimed at gathering reflections on their reasoning, strategies, and difficulties. Since time with participants was limited, the study focuses on understanding the process of rule construction rather than on creating a full system.

2.1. Participants

The study involved five participants, all affiliated with an AI and Law research group. Participants had an interest in applying AI techniques in the legal domain, and some had prior familiarity with computational argumentation. However, they were not domain experts in criminal law or fraud. Their backgrounds varied in terms of seniority and expertise. Participation took place during a research visit and was entirely voluntary. All participants provided informed consent prior to the study.

2.2. Tool

During the study, participants used a tool that allowed them to build a dialogue system using argumentation rules.[1] In this study, we use a simplified version of ASPIC$^+$ [10], following [11] and [8], which has been shown to be suitable for applications within the Dutch police. The simplification is intended to make the system more accessible for non-expert users, such as police employees, while retaining the core mechanisms needed to model legal reasoning. This simplified version of ASPIC+ consists of three main elements. First, it uses defeasible rules, which capture general default inferences from premises to conclusion. These rules are applied to elements in the knowledge base to construct arguments. For example, given premises such as ‘paid for a product’ and ‘product not delivered’, a defeasible rule can draw the conclusion that a case is potentially fraudulent. Second, the tool includes rebuttal attacks, in which arguments attack arguments that have contradictory conclusions. For instance, a rule stating that a ‘trusted seller’ generally implies ‘not fraud’ can produce an argument that directly conflicts with the earlier conclusion of potential fraud. Third, the tool allows for preferences between rules, which are used to resolve such conflicts. When two arguments attack each other, a preference ordering determines which argument wins. For example, if the rule associated with ‘trusted seller’ is preferred over the rule linking ‘paid for a product’ and ‘product not delivered’ to fraud, the tool will resolve the conflict in favor of the conclusion that the case is not fraud.

[1]The tool used in this study is available at: https://git.science.uu.nl/D.Odekerken/demo-argumentation-based-inquiry/-/tree/d524e5f30af3fc5d9d3663b70d1a08aed5d43115/

Participants did not interact directly with ASPIC+. Instead, in the tool, participants could create IF-THEN rules. They could use NOT and AND as basic operators in the rules' antecedents and consequents. Rules written by participants were translated into argumentation rules by the system. To decide on conflicts between rules, participants could also specify preferences between the rules. To test whether their rules worked as intended, the system provided a chat interface where participants were asked about the presence of premises, and when enough information was provided, the system provided a decision as output.

2.3. Data

To give participants an idea of complaints the Dutch police receives about online fraud, participants were given posts from an online forum that a police employee identified as relevant to online fraud. The forum is a Dutch forum called Radar.[2] Radar is a Dutch online platform and programme on the Public Broadcast Network that advocates for better treatment of consumers by businesses and government, and takes a critical look at the service provided to consumers. Part of their online platform is a forum where people can post about their experiences with companies and organisations. This platform contains many posts that are similar to fraud complaints that the Dutch police receives.

The participants received an overview of nine kinds of online fraud and one or two examples from the forum for each kind of fraud. Messages on the forum were originally in Dutch, but were translated into English and anonymized for this study. The kinds of fraud are purchase and sales fraud, payment request fraud, boiler room fraud, dating fraud, WhatsApp fraud, acquisition fraud, subscription fraud, identity fraud, and telephone helpdesk fraud. In Example 2.1 an example of a (translated) message posted to the forum on acquisition fraud is shown. Acquisition fraud is a form of fraud in which orders are solicited without any intention of fulfilling them. This also includes offers that look like invoices.[3]

Example 2.1. An example of a message concerning acquisition fraud posted to the Radar forum, translated from Dutch.

"An acquaintance, female, (18 years old) received an email today from Toeslagenportal b.v. (Benefit portal LLC) with an urgent request to transfer €45 to them for their invoice dated 24 March 2021 for apparently provided intermediary services for applying for a benefit. They are threatening to take legal action if payment is not made within 7 days. She has never contacted this company. I read some reviews about the company, and I can see that they are scammers/thieves. These are very recent reviews."

2.4. Procedure

The study consisted of three phases: introduction, system construction, and focus group. First, an introduction to the Dutch law on fraud, 'Oplichting' article 326 of the Dutch Criminal Code. The English translation of this article is:

> He who, with the intention of unlawfully benefiting himself or another, either by assuming a false name or false identity, or by cunning tricks, or by a web of fabrications, induces someone to surrender any property, to render any service, to make any information available, to incur any debt, or to cancel any debt, shall be punished as guilty of fraud with imprisonment of up to four years or a fifth-category fine (WvSr article 326).

The second presentation was an introduction to computational argumentation. This included the elements of the simplified version of ASPIC+ used in the tool: defeasible rules, rebuttal attacks, and preferences between rules.

After the two presentations, participants were divided into two groups, given access to the tools, and provided with the posts from the Radar forum. They were instructed to build an argumentation-based system consisting of a knowledge base and rules to classify specific kinds of fraud in the Radar forum

[2] https://radar-forum.avrotros.nl/

[3] Based on https://www.politie.nl/informatie/wat-is-acquisitiefraude-spookfacturen.html

data. The first group was explicitly asked not to use LLMs to assist them in the task. The second group was allowed to use LLMs. They were not given any instructions on what kind of LLM or what kinds of prompts to use. Both groups were allowed to look things up online during the task (such as the online fraud article and the police website). During the study, several police experts were present and available for questions and assistance. Participants worked in groups for about 75 minutes. This was not enough time to complete a full system, and this was intentional. The goal of the study is to investigate difficulties in the creation of an argumentation-based system and whether LLMs can aid this process. Therefore, participants were given an intentionally difficult task.

Following the task, a 30-minute focus group discussion was conducted. Participants first described their approaches and experiences, after which they engaged in a joint discussion comparing their strategies, challenges, and use of tools. This discussion provided information to understand participants' reasoning processes and experienced challenges during the task.

3. Results

The results are presented in four parts: 1) the process of the group without LLM support, 2) the process of the group with LLM support, 3) a comparison of the rule sets produced by both groups, and 4) insights from the group discussion. In this section, quotes are used for illustration. Participants 1, 2, and 3 (P1, P2, P3) were in group 1 (no LLM), and participants 4 and 5 (P4, P5) were in group 2 (with LLM). All quotes are from the group discussion, but are also used to illustrate the process of individual groups. Note that participants had only received a brief introduction to the legal definition of fraud. As a result, their interpretations of legal concepts are exploratory and might (often) not be correct.

3.1. Group 1 (No LLM)

This group consisted of three members. They started by reading and dissecting the legal definition of fraud and identified specific elements needed for fraud. Then they considered the examples to look for these specific elements of fraud and what information citizens gave that could point to fraud. Members of the group took turns reading out quotes from the provided examples and attempted to reconcile this with both the legal definition and their own knowledge of fraud.

The provided tool creates an argumentation-based system using IF-THEN rules entered by the participants. Based on these rules, attacks are created. For this group, these attacks were not intuitive to understand based on the created rules, leading to discussion between the group members. They started drawing out the arguments and attacks to establish a shared interpretation of the system they were creating. This gave them something concrete to discuss and made the task go smoother afterwards. During the remainder of the task, they often referred to this shared representation.

The group identified many different elements relevant to fraud, but found it difficult to know how much evidence is enough to determine fraud. Especially since some elements might correlate with fraud, but not be a direct cause or proof of fraud. An example of this is the following discussion between the three group members:

> "P3: Yeah, for example, you even have like if something is 'too good to be true' it doesn't mean like, it's not really strong something, but it's also not like, it does have some kind of merit like...
> P2: It's correlated. [...] But it's not a causal indicator.
> P3: Yeah [...]
> P1: So that was kind of a thing where we were like, it would be nice if there were some numbers attached in some way, or at least something more than just preferences."

In the discussion, a deal being 'too good to be true' makes it more likely for the deal to be fraudulent, but it is not sufficient evidence to establish that it is fraudulent for sure. The provided tool only used rules and preferences between rules. The intuition of the group was to use weights and probabilities, since this was not possible, they struggled with reconciling their intuitions with the provided system.

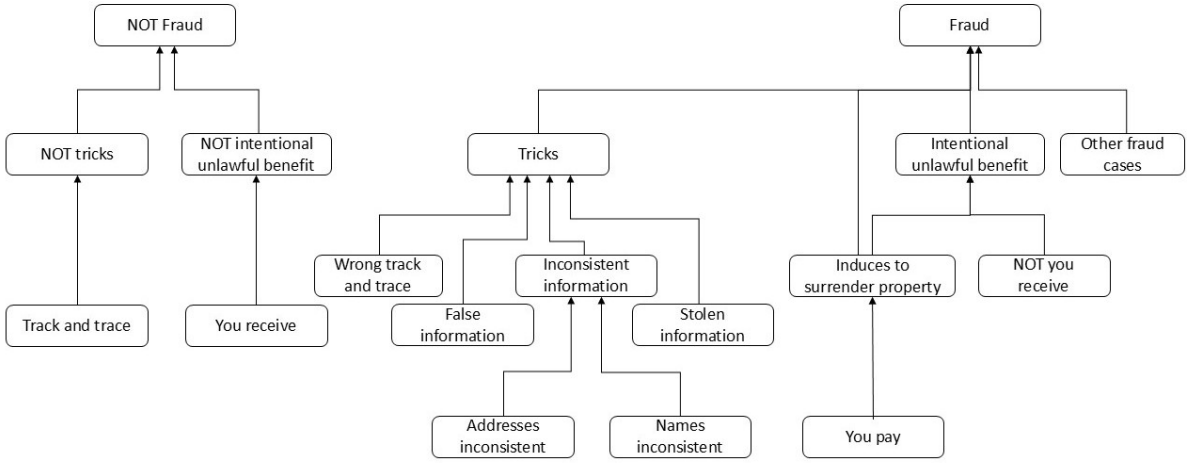

Figure 1: Visualization of rules created based on the rules created by group 1. Arrows represent rules from premises to conclusions, arrowed are joined to represent ANDs used by participants. For example, the rule (r1) IF intentional unlawful benefit AND tricks AND induces to surrender property is represented by the rule in the top right (the left rule for fraud), where tricks, induces to surrender property, and intentional unlawful benefit are premises, and fraud is the conclusion.

3.2. Group 2 (LLM)

This consisted of two members, and immediately started by asking the LLM[4] to generate rules to identify fraud. With the goal of creating a minimal viable product, to then troubleshoot. They found that the Dutch law fraud article was not specific enough for the LLM to generate rules that could be applied to the provided examples. So they continued by giving the LLM example cases and asking it to provide reasons for and against fraud in these specific cases. These rules were deemed more appropriate by the group, but were very specific to the example cases and did not generalise well.

The group attempted to use the LLM to create visualizations of their rule set, but were not able to generate any visualizations that they were satisfied with. After realizing the LLM-generated rules were too specific, but having gained insight into the task and the examples, the group changed their strategy. Starting with the legal article, they identified two main elements (intentional unlawful benefit and something removed from a person). Then they worked from there to specify what this could look like. For example, using a fake name is a form of deception.

> "P5: No, but I think that visualization with ChatGPT was not good, but for the start it was faster.
> P4: Yeah
> P5: For me at least. The rules. Then we know how to do that, and the next cases become easier for us without ChatGPT.
> P4: Yeah, yeah, I think I would also agree that by doing this incorrectly really quickly, you learn a lot."

Overall, as described in the quote above, the LLM did not directly provide useful rules, but it was very helpful in helping them get started quicker and learn from its mistakes.

3.3. Created rules

The full set of rules and preferences defined by the two groups can be found in Appendix A. Group 1 created 14 rules that argue for both fraud and non-fraud. Group 2 created 13 rules to determine when a case is fraud.

[4]They used ChatGPT in browser on December 18th 2025.

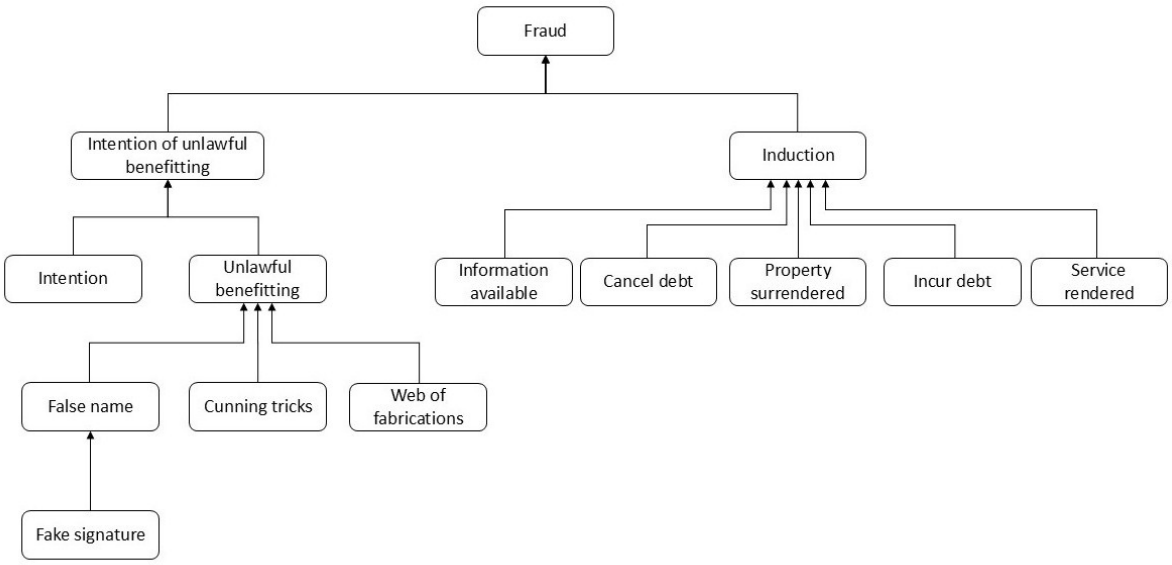

Figure 2: Visualization of rules created based on the rules created by group 2.

A visualization representing the rules created by Group 1 can be found in Figure 1. Group 1's main rule for fraud consisted of three premises: 'intentional unlawful benefit', 'induces to surrender property', and 'tricks'. Out of these three premises, tricks received the most attention during the session, and this can be seen in the many different rules made for and against tricks. One example is that an inconsistent name means information is inconsistent, and inconsistent information is a trick. The other two premises in the main rule for fraud received less attention and are argued for and against based on whether products were paid for and received.

A visualization representing the rules created by Group 2 can be found in Figure 2. Their main rule for fraud consists of two premises: 'intention of unlawful benefitting' and 'induction'. The former is further split into 'intention' and 'unlawful benefitting', which is then split into 'benefitting' and 'unlawful'. Both 'unlawful' and 'induction' have several rules for them. For example, cancelling debt can be a form of induction. And a fake signature is a form of fake name, which is unlawful.

A main difference between the rules created by the two groups is in the interpretation of the Dutch legal article on fraud. Group 1 separated unlawful benefitting and tricks as two premises needed for fraud, while Group 2 formalized tricks as a type of unlawful benefitting. Based on the legal article, three elements are needed for fraud: 1) the intention of unlawfully benefiting, 2) inducing someone to surrender any property, to render any service, to make any information available, to incur any debt, or to cancel any debt, and 3) using a false name or false identity, cunning tricks, or a web of fabrications to do so. Therefore, Group 1 was more successful in capturing these three aspects of the legal article. However, Group 1 did categorize all elements of part 3 of the legal article above as tricks, even though a false name and a web of fabrications are separate from tricks. Group 2 did capture this distinction, but formalized it as premises of 'unlawful', which is not in line with the article. Group 2's rule set is very complete, in the sense that they have captured all legal terms in the article. This makes sense since they focused more on the legal article, whereas Group 1 focused more on the examples. Since the rule set by Group 1 also includes information from the examples, not all elements from their rule set can be tied directly to the legal fraud article.

3.4. Group discussion

In the group discussion, first each group described their process to the other group and the researcher. These descriptions have been combined with observations by the researcher to create the summaries of

each group's process above.

After each group described their process, the researcher invited them to compare and discuss their process and output. Part of the discussion concerned the definitions of 'trick' and 'unlawful benefit', and the interaction between these concepts. For participants, it was difficult to reconcile these legal concepts with the real-world examples and to understand the intricacies of the concepts. For example, trickery may be lawful due to consent or contractual contexts. Not fully understanding these definitions and their legal context made them difficult to model in the study, an example of this is provided below.

> "P4: Yeah. So we would say intention of unlawful benefiting is further described by things like, uh, false name, cunning tricks. It's all part of being unlawful. So this could maybe be an example of of of such a way.
> P1: Benefit seems very tight. This is also something we struggled with,
> P4: Yeah
> P1: Like the benefit is also like the consequence of surrendering the property."

Across both groups, a recurring theme was the question of how specific rules should be. Rules that are too specific do not generalize well, and rules that are too general are not operationalizable. Additionally, the rules in the system are defeasible, meaning exceptions to the rules can be formulated. But many kinds of exceptions to rules are possible, and it was difficult, especially for Group 1, to determine the appropriate level of specificity. The quote below illustrates this difficulty. In particular, it shows how participants repeatedly reconsidered whether exceptions should be made explicit, while also recognizing that adding more detail risked making the system more complex.

> "P1: And then we kind of also got stuck on the difference between, like what can be supported and like if something is not supported, does that mean that it's false or does it mean that you don't have any information? And I think for some at some point. We were like, oh, we have to put that in explicitly, and then we're we were like, oh, but maybe not. And then at another point, they were like, well, no, but we do.
>
> P1: And this thing with receiving. I don't. I think the more rules we had, the more complicated it got and the more. More points of like, at some point it just becomes so vague that you don't really know what's going on. "

The same quote also relates to a third difficulty experienced by participants, the complexity of the argumentation-based system increases quickly when new rules are added. The interaction between rules and exceptions to rules can quickly increase the complexity, which, paired with other uncertainties about legal definitions and the right level of rule specificity, makes it difficult to maintain an overview of the task. Participants from both groups report being initially overwhelmed. Participants in the LLM group mention that they believe the LLM support made this more manageable.

Participants agreed that design choices depend heavily on the intended users of the system(such as citizens versus experts) and that LLMs are useful for rapid exploration and learning, but less suitable as the sole basis for legal reasoning.

4. Discussion

This study explored how participants construct an argumentation-based system for classifying fraud from citizen reports, and whether an LLM could facilitate this process. The results suggest that an important challenge in building such systems lies in the need to make multiple simultaneous design decisions at different levels of abstraction. Participants had to simultaneously interpret legal concepts, determine appropriate rule specificity, relate abstract legal text to specific examples, and manage the increasing complexity of the system due to interactions between rules. This combination made the task cognitively complex and overwhelming to participants.

This cognitive load is created by the interaction between three main factors: uncertainty about legal interpretation, difficulty in selecting the appropriate level of abstraction, and the growing complexity of the rule system as new elements are added. These factors reinforce each other. For example, uncertainty about legal concepts makes it harder to decide how specific rules should be, while increasing system complexity makes it more difficult to maintain an overview and evaluate these decisions. This helps explain why participants struggled to feel confident in their progress, even when they successfully identified relevant elements of fraud.

The comparison between groups based on observations and a focus group provides insight into the role of LLMs in the creation of an argumentation-based system. The group with access to an LLM was able to start more quickly and reported feeling less overwhelmed in the initial phase. This suggests that LLMs may help reduce the barrier to entry to a difficult task by supporting early exploration and orientation within the task. However, the rules generated with LLM support were often specific to example cases, making them not generalizable. In contrast, the group without LLM support produced a rule set that more closely aligned with the legal article on fraud, although neither group interpreted the legal article fully correctly. Taken together, these observations show that LLMs are more effective as tools for onboarding and exploration than for directly generating rules in this context.

4.1. Recommendations

These findings point to three key needs that should be addressed in future research and in practice when developing expert systems for the policing domain.

First, visualizations were beneficial to participants, both for them to understand the system and for them to understand each other. When rules were written out on paper, participants were able to analyze and communicate about them more effectively. However, the LLM was not useful in providing visualizations for rules. This suggests that integrated visualization in the provided tools would be helpful in reducing cognitive load by making dependencies and conflicts between rules more explicit.

Second, LLMs appear to be most useful in a supporting role during the early stages of the process. In this study, they helped participants to engage with the task quickly and to explore the data, but were less effective for producing usable rules. A more promising application may be to use LLMs for ongoing feedback, for example, by prompting reflection on design choices or highlighting potential inconsistencies, rather than relying on them for direct rule generation.

The third recommendation, based on the results of this study, would be to separate interacting with the data and with legal articles into separate phases. Participants struggled in part because they had to work with legal texts and real-world examples simultaneously, requiring shifting between abstraction levels. Structuring the process into distinct phases—such as initial data exploration followed by formalization, or vice versa—may help participants to focus on one type of reasoning at a time.

4.2. Limitations and Future Work

This study is exploratory in nature and has several limitations. First, the number of participants and groups was small, making it difficult to draw strong conclusions about differences between conditions. Observed differences between the LLM and non-LLM groups may be influenced by variation in participants rather than the presence of LLM support.

Second, participants were AI and Law researchers rather than legal practitioners, and their understanding of the legal domain was based on a brief introduction. This likely contributed to the difficulties in interpreting legal concepts and limited the generalizability of the findings to professional settings.

Third, the task was intentionally underspecified with respect to the intended use and users of the system. While this helped to surface design challenges, it also meant that participants lacked guidance on appropriate levels of rule specificity, which is an important factor in knowledge engineering. Additionally, the use of LLMs was unspecified and left up to participants. Therefore, only one LLM was used in a specific way, while different models with different prompts would have potentially provided different results.

Future work could address these limitations by involving larger and more diverse groups, including legal experts and practitioners, and by more clearly defining the intended application context. For a similarly small-scale study, it would be beneficial to decompose the task into smaller subtasks with specific instructions on how to use the LLM and to compare them one-on-one between LLM and no LLM conditions. This could provide more granular information about the use of LLMs in the process.

It would also be valuable to systematically investigate the role of LLMs in collaborative settings, for example, as tools for facilitating communication between experts with different backgrounds. Finally, implementing and evaluating the proposed recommendations, such as visualization support, LLM-based onboarding and feedback, and distinct data exploration and legal interpretation phases, would provide further insight into how the cognitive demands of building argumentation-based systems can be reduced.

5. Conclusion

In this paper, we presented an exploratory case study and focus group investigating how participants construct an argumentation-based system for classifying fraud in citizen reports, with and without access to LLM support. The results show that an important hurdle in building such systems is the cognitive complexity during the design process caused by the need to make multiple interdependent decisions simultaneously, including interpreting legal concepts, selecting appropriate levels of rule specificity, and managing complexity because of interaction between rules.

The comparison between groups suggests that LLMs can play a useful role in initial exploration and onboarding, helping participants to engage with the task more quickly and with less initial uncertainty. However, LLM-generated rules were often overly specific and did not generalize well. This indicates that LLMs are currently better suited for exploration and onboarding than to directly producing formal rule sets. This conclusion should be interpreted in light of the specific conditions of our study: participants were AI and Law researchers with only a brief introduction to the legal domain, rather than legal practitioners, and the task was intentionally underspecified.

Based on these observations, three recommendations for the development of argumentation-based systems in practice were provided: providing better visualizations to manage system complexity, using LLMs for feedback and exploration rather than rule generation, and structuring the design process into separate phases for data exploration and legal interpretation.

Acknowledgments

Thanks to Daphne for the use of the tool and for providing motivation for this study, to AnneMarie for providing valuable feedback, to Floris, Daphne, Elize, and Michiel for assisting during the session, and the members of the Hybrid Law group for their participation and enthusiasm on this topic.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.3, Claude Sonnet 4.5, and Grammarly in order to: Grammar and spelling check, improve writing style, paraphrase and reword, and for peer review simulation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Bench-Capon, H. Prakken, G. Sartor, Argumentation in Legal Reasoning, in: G. Simari, I. Rahwan (Eds.), Argumentation in Artificial Intelligence, Springer US, Boston, MA, 2009, pp. 363–382. URL: https://doi.org/10.1007/978-0-387-98197-0_18. doi:10.1007/978-0-387-98197-0_18.

[2] H. Prakken, An abstract framework for argumentation with structured arguments, Argument & Computation 1 (2010) 93–124. URL: http://content.iospress.com/doi/10.1080/19462160903564592. doi:10.1080/19462160903564592.
[3] K. Atkinson, T. Bench-Capon, D. Bollegala, Explanation in AI and law: Past, present and future, Artificial Intelligence 289 (2020) 103387. URL: https://www.sciencedirect.com/science/article/pii/S0004370220301375. doi:10.1016/j.artint.2020.103387.
[4] K. D. Ashley, Modelling legal argument: Reasoning with cases and hypotheticals. (1989). URL: https://elibrary.ru/item.asp?id=7547687.
[5] E. A. Feigenbaum, Knowledge Engineering: The Applied Side of Artificial Intelligence., Technical Report, 1980. URL: https://apps.dtic.mil/sti/html/tr/ADA092574/.
[6] M. Billi, G. Pisano, M. Sanchi, Fighting the Knowledge Representation Bottleneck with Large Language Models, in: Legal Knowledge and Information Systems, IOS Press, 2024, pp. 14–24. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA241230. doi:10.3233/FAIA241230.
[7] M. Schinckus, A. Simonofski, N. Bono Rosselló, Large Language Models for Process Knowledge Acquisition, Business & Information Systems Engineering 68 (2026) 7–33. URL: https://doi.org/10.1007/s12599-025-00976-w. doi:10.1007/s12599-025-00976-w.
[8] D. Odekerken, F. Bex, A. Borg, B. Testerink, Approximating stability for applied argument-based inquiry, Intelligent Systems with Applications 16 (2022) 200110. URL: https://www.sciencedirect.com/science/article/pii/S2667305322000485. doi:10.1016/j.iswa.2022.200110.
[9] D. Odekerken, F. Bex, Towards Transparent Human-in-the-Loop Classification of Fraudulent Web Shops, in: S. Villata, J. Harašta, P. Křemen (Eds.), Legal Knowledge and Information Systems, volume 334, IOS Press, 2020, pp. 239–242. URL: http://ebooks.iospress.nl/doi/10.3233/FAIA200873. doi:10.3233/FAIA200873.
[10] S. Modgil, H. Prakken, A general account of argumentation with preferences, Artificial Intelligence 195 (2013) 361–397. URL: https://www.sciencedirect.com/science/article/pii/S0004370212001361, publisher: Elsevier.
[11] D. Odekerken, T. Lehtonen, J. P. Wallner, M. Järvisalo, Argumentative Reasoning in ASPIC+ under Incomplete Information, Journal of Artificial Intelligence Research 83 (2025). URL: https://www.jair.org/index.php/jair/article/view/18404. doi:10.1613/jair.1.18404.

A. Rules written by the two groups

A.1. Group 1 (no LLM)

```
Rules:

  R1:   IF Intentional_unlawful_benefit AND tricks AND induces_to_surrender_property
        THEN fraud
  R2:   IF track_and_trace
        THEN NOT tricks
  R3:   IF wrong_track_and_trace
        THEN tricks
  R4:   IF false_information
        THEN tricks
  R5:   IF inconsistent_information
        THEN tricks
  R6:   IF names_inconsistent
        THEN inconsistent_information
  R7:   IF addresses_inconsistent
        THEN inconsistent_information
  R8:   IF stolen_information
        THEN tricks
  R9:   IF other_fraud_cases
        THEN fraud
  R10:  IF NOT Intentional_unlawful_benefit
        THEN NOT fraud
  R11:  IF NOT tricks
        THEN NOT fraud
  R12:  IF you_pay
        THEN induces_to_surrender_property
  R13:  IF you_receive
        THEN NOT Intentional_unlawful_benefit
  R14:  IF NOT receive AND induces_to_surrender_property
        THEN intentional_unlawful_benefit

  Preferences:

  R1 < R10
  R1 < R11
  R2 < R3
  R2 < R4
  R2 < R5
  R2 = R8
  R9 < R10
  R9 < R11
```

A.2. Group 2 (LLM)

```
Rules:
```

```
R1:   IF intention_of_unlawful_benefitting AND induction
      THEN fraud
R2:   IF intention AND unlawful_benefiting
      THEN intention_of_unlawful_benefitting
R3:   IF benefiting AND unlawful
      THEN unlawful_benefiting
R4:   IF false_name
      THEN unlawful
R5:   IF cunning_tricks
      THEN unlawful
R6:   IF web_of_fabrications
      THEN unlawful
R7:   IF fake_signature
      THEN false_name
R8:   IF property_surrendered
      THEN induction
R9:   IF service_rendered
      THEN induction
R11:  IF information_available
      THEN induction
R12:  IF incur_debt
      THEN induction
R13:  IF cancel_debt
      THEN induction
```

Preferences:

None