Using Large Language Models and Law-Based Rules for the Analysis of VAT Chain-Transaction Cases in Austrian Tax Law

Marina Luketina[1,*], Lukas Knogler[2] and Christoph G. Schuetz[2]

[1]University of Applied Sciences Upper Austria, Wehrgrabengasse 1, 4400 Steyr, Austria
[2]Johannes Kepler University Linz, Altenberger Str. 69, 4040 Linz, Austria

Abstract

In tax advisory practice, case descriptions are typically not structured in a machine-readable format, with clients describing their situation in natural language. Large language models excel at natural-language understanding. However, for legal reasoning, including tax law, the propensity of LLMs to hallucinate presents a considerable challenge. Rule-based systems, on the other hand, offer verifiably correct reasoning given the correct input. Therefore, in this paper, we propose a hybrid approach to support tax advisors with analyzing tax cases, combining a rule-based system with large language models. We focus on the analysis of chain-transaction cases in value-added tax (VAT) law, where the law states a clear set of rules for regular chain-transaction cases. We employ a large language model (LLM) for the construction of structured representations of natural-language VAT case descriptions and law-based rules for the identification of the movable supply, which determines tax liabilities. Human tax advisors can obtain a graphical visualization of the structured representation to verify the correctness of the LLM's output while the law-based rules return reliable decisions.

Keywords

Neuro-symbolic artificial intelligence, Knowledge graphs, Decision support systems, Tax management, Value-added tax.

1. Introduction

Tax advisors typically receive natural-language case descriptions from clients, which are not uniformly structured. Before providing a legally founded analysis of the client's situation, a tax advisor must first make sense of the situation by structuring the case description. Ultimately, a legally-founded decision is grounded in logic [1]. Thus, the tax advisor's task consists of language understanding and logical reasoning. In this paper, we focus on the chain-transaction cases in value-added-tax (VAT) law. In cross-border chain transactions between multiple enterprises, VAT liabilities are determined according to a set of rules. The *movable supply* in a chain transaction determines the tax liabilities.

A large language model (LLM) is a type of machine-learning model with billions of parameters trained on vast amounts of text data to predict the probability of the next word (or token) given a sequence of previous words, with applications in many domains, e.g., in the legal field [2, 3, 4, 5, 6], including applications in tax advisory [7, 8, 9, 10] and summarization of legal documents [11, 12, 13]. Nevertheless, the well-known problem of hallucination limits the use of LLMs in sensitive areas such as the legal field, since the necessary reliable and comprehensible decisions cannot be ensured [14, 15, 16, 17]. LLMs show promising results in the extraction of knowledge from unstructured natural language texts [18, 19]. In this regard, LLMs are becoming an increasingly popular tool for transforming natural language texts into knowledge graphs [20, 18, 21, 22, 23], which serve as a framework for structuring knowledge in a comprehensible format streamlined for certain applications [24]. In this paper, we therefore investigate the capabilities of the integration of LLMs as knowledge extractors to build a

1st Workshop on Bridging Hybrid (Artificial) Intelligence and the Semantic Web (HAIBRIDGE 2025), co-located with the 24th International Semantic Web Conference (ISWC 2025)

*Corresponding author.
marina.luketina@fh-steyr.at (M. Luketina); knogler@dke.uni-linz.ac.at (L. Knogler); schuetz@dke.uni-linz.ac.at (C. G. Schuetz)
0009-0007-7440-6002 (M. Luketina); 0000-0002-0955-8647 (C. G. Schuetz)

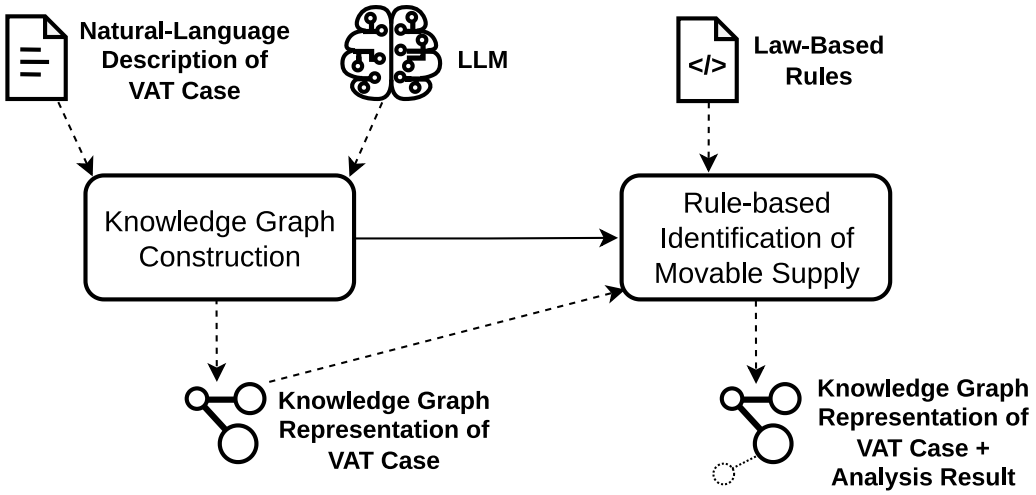

Figure 1: Method for analyzing a VAT chain-transaction based on a textual, natural-language description of the case, using LLMs for construction of a KG representation and law-based rules for identification of the movable supply

knowledge-graph representation of legal cases, which can be used as the basis for legal reasoning. To analyze a VAT case, the natural-language description of the case is translated into a knowledge graph, and rules can be executed to derive additional information. Tax advisors can obtain a visualization of the extracted knowledge graph to easily verify the correct representation of a case.

We employ the design-science research methodology by Wieringa [25]. Following Wieringa's template, we formulate the design problem as follows.

- Increase the efficiency of tax consultants *(problem context)*
- by designing an LLM-based method *(artifact)*
- that *(requirements)*
    1. identifies movable and unmovable supplies in chain transactions
    2. by processing natural-language case descriptions
    3. to conduct a logically founded and explainable analysis,
- enabling the allocation of correct tax liabilities under Austrian tax law *(stakeholder goals)*.

From this design problem, we derive the following research question.

> *How can LLMs be used to extract information from natural-language text to form the basis for logically founded and explainable decisions regarding chain transactions in Austrian tax law?*

To answer the research question, we implemented the LLM-based method for analyzing VAT chain-transaction cases illustrated in Fig. 1; the implementation is available in an online repository [26]. An LLM serves to construct a structured knowledge graph presentation from the natural-language description of a VAT chain-transaction case. More specifically, the LLM outputs Cypher statements to build the knowledge graph for storage in a Neo4j database. Law-based rules executed over the contents of the knowledge graph then identify the movable supply, providing the basis for a decision in the VAT chain-transaction case.

We use different datasets for development of the prototype and evaluation of the performance (see Fig. 2); data from the experimental evaluation is available in an online repository [27]. We use example cases from Kollmann's textbook on chain transactions [28] for prompt engineering. To evaluate the performance of the developed prototype, we conducted experiments with example cases from other textbooks [29, 30] as well as real-world cases and exam questions from a university course. We measure the prototype's performance in terms of the correctness of the obtained knowledge

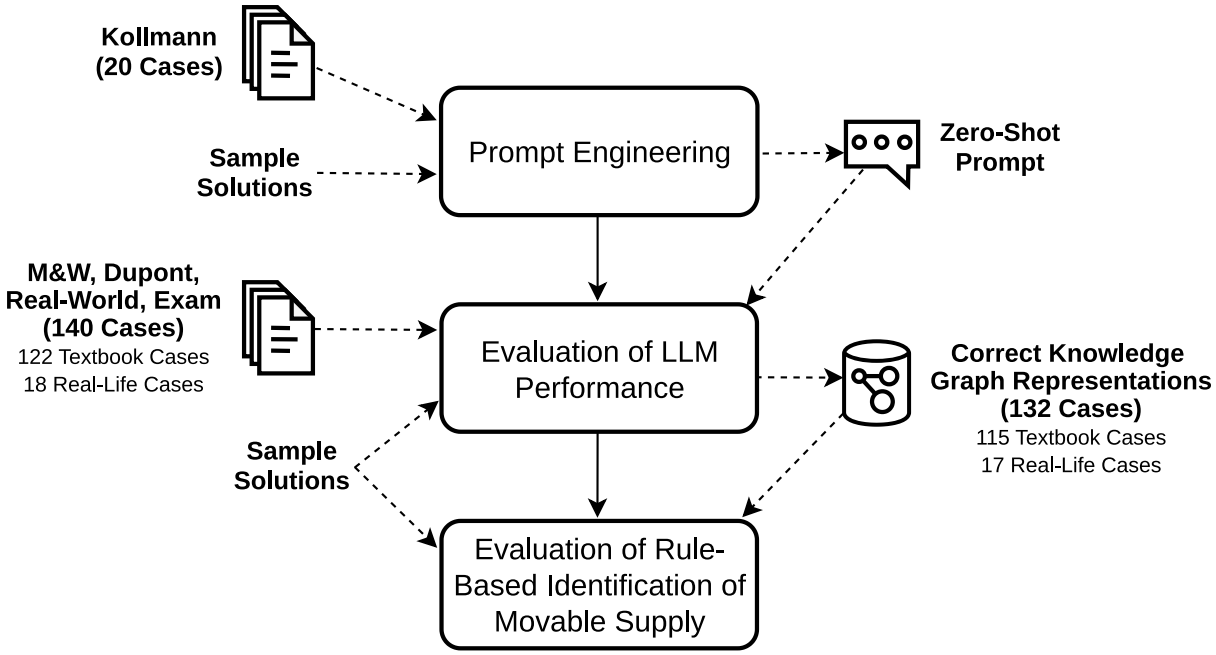

Figure 2: Overview of research methodology

graph representations from the natural-language descriptions of VAT chain-transaction cases and the subsequent identification of the movable supply. The results of these experiments suggest good performance of LLMs for knowledge graph construction, with an overall accuracy of 94.74 %. On correctly constructed knowledge graphs, the law-based rules work as expected and never fail to correctly identify the movable supply.

The hybrid approach of using subsymbolic AI for language understanding and symbolic AI for actual legal reasoning has clear advantages in terms of explainability and understandability for a human tax advisor using the AI system for decision-making. A human tax advisor can manually review the graphical representation of the chain transaction to verify the accuracy of the output, and law-based rules provide reliable logical reasoning. A purely subsymbolic AI approach, e.g., using LLMs both for language understanding and legal reasoning [31, 32], lacks the reliable logical reasoning; an LLM does not really perform logical reasoning. Furthermore, a human tax advisor cannot easily review the LLM "reasoning" output without conducting both an analysis of the case and the legal reasoning.

The remainder of this paper is organized as follows. In Sect. 2, we provide a brief introduction to the subject of chain transactions in VAT law. In Sect. 3, we present an implementation of the proposed method. In Sect. 4, we describe the experimental evaluation of the implemented prototype of the proposed method. In Sect. 5, we review related work. In Sect. 6, we conclude the paper with a summary and an outlook on future work.

2. Background: Chain Transactions in VAT Law

VAT is a tax on consumption that is added to goods and services [33], which in Europe is harmonized by EU regulation, implemented into national legislation by EU Member States. A chain transaction is a series of at least two transactions for the same goods that are transported from the first supplier to the last customer in a single transportation operation [29]; a chain transaction involves at least three enterprises. To allocate the correct VAT liabilities in case of cross-border chain transactions, the Austrian Value-Added Tax Act ("Umsatzsteuergesetz") 1994 [34], like similar legal provisions in other countries, defines step by step how to determine taxation of a regular chain transaction.

The *movable supply* in a cross-border chain transaction determines the VAT liabilities, i.e., which countries levy the tax. The movable supply is the delivery connected to the actual movement, i.e.,

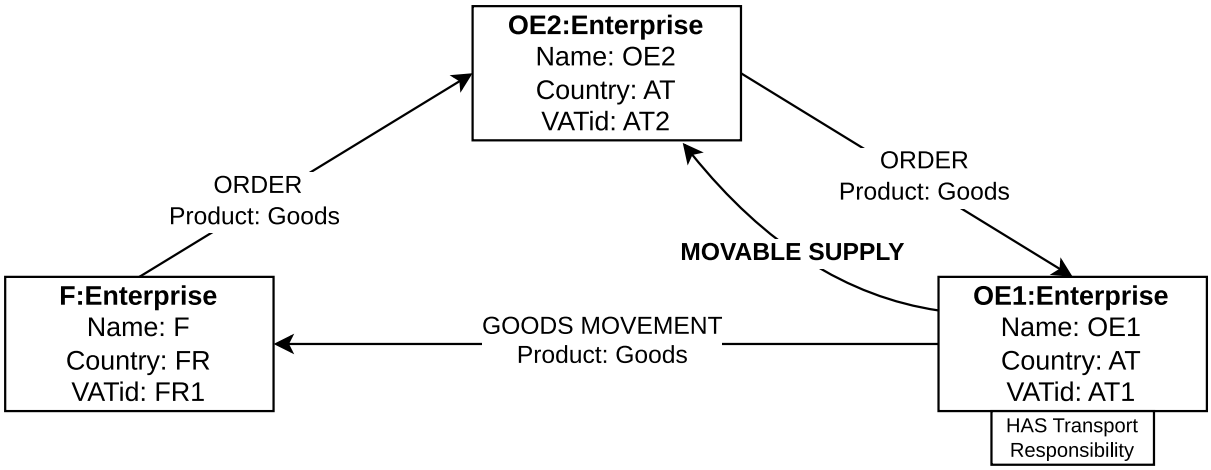

Figure 3: Graph representation of a chain transaction. Rectangles represent entities, which have an identifier and an entity type. Arrows represent the relationships between entities. Each relationship has a relationship type (*ORDER*, *GOODS MOVEMENT*, or *MOVABLE SUPPLY*). The entities and the relationships are characterized by attributes. The *HAS* relationship of the *OE1* enterprise to a *Transport responsibility* entity is indicated using an attached box.

transportation, of goods. In general, for most VAT chain-transaction cases, the responsibility for transportation of the goods indicates the movable supply. The party responsible for transportation can be the first party, the last party, or an intermediary in the chain transaction. If the first or the last party in a VAT chain transaction has the responsibility for transportation, the movable supply is the supply of the respective party. If the party responsible for transportation is an intermediary, the indicated VAT identification number, i.e., "an individual identification number for companies that are registered for VAT" [35], influences the decision regarding the movable supply. Depending on the VAT identification number, the movable supply is either that of the intermediary or that of the party before the intermediary. The responsibility for transportation and, consequently, the movable supply then determines the allocation of tax liabilities.

The following example, adapted from the textbook by Mayr and Weinzierl [29], describes VAT chain-transaction case and the corresponding identification of the movable supply (see Fig. 3 for a visualization).

Case Description. The enterprise *F* from France orders goods from the enterprise *OE2* in Salzburg, Austria. The latter does not have the goods in stock and buys those goods from the enterprise *OE1* in Klagenfurt, Austria. *OE1* has the goods transported directly from Austria to *F* in France on *OE1*'s own account by a carrier. *OE2* presents itself to *OE1* with an Austrian VAT number.

Solution. The first supplier in this example is *OE1*. The last customer is *F* and the intermediary is *OE2*. According to the case description, *OE1* is responsible for the transportation of the goods; a supplier is not automatically responsible for transportation. Therefore, *OE1*'s supply is the movable supply (*OE1* $\rightarrow$ *OE2*). In the graph representation (Fig. 3), the *movable supply* relationship points in the opposite direction of the corresponding *order* relationship.

3. Implementation

In this section, we describe the implementation of the proposed method, including the LLM-based knowledge graph construction and the rule-based identification of the movable supply. The implementation is available in an online repository [26].

Listing 1: Condensed version of the prompt used for knowledge graph construction (translated from German)

```
Identity: Tax consultant for Austrian VAT law specialized on chain transactions
Response rules: representation as knowledge graph; capture the entire chain
    transaction; no changes or additions allowed
Detailed instructions: identification of enterprises, transactions, and transport
    responsibility; identification of available property values for Name, Country
    and VATid (if VATid is not stated, generate a VATid using the country of the
    respective enterprise and a consecutive number); definition of allowed
    vocabulary, assignment of transport responsibility and direction of relations;
    double check query statements with your explanation
Structured output: Include Cypher statements, argumentation, and a summary.
```

3.1. Knowledge Graph Construction from Case Description

We use the Neo4j Aura graph database for knowledge graph management. To construct a knowledge graph from natural-language case descriptions, an LLM was instructed to generate Cypher statements based on the information contained in the textual description. We used GPT-4.1 as LLM due to its ability to follow instructions and its long-context understanding [36]. The LLM was accessed via the LangSmith framework published by LangChain, which provides comprehensive monitoring and prompt engineering possibilities as well as data management for sample data [37]. Our model configuration of GPT-4.1 was set to a *temperature* value of 0 and a *top_p* value of 1. This allowed the model to provide deterministic results and reduce the risk of hallucination [38].

We note that the general principle demonstrated in this paper is agnostic to the choice of underlying database technology. The graph nature of VAT chain transactions makes a graph database such as Neo4j Aura a convenient choice. An alternative to using Neo4j Aura as a database would be to use an RDF database. Although the Resource Description Framework (RDF) [39] is also typically considered a graph-based representation of data, RDF focuses more on expressing statements consisting of subject, predicate, and object, rather than relationships between objects. When using an RDF database, SPARQL statements [40] or Datalog statements [41] would then have to be generated analogously to the Cypher statements presented in this paper, which would also be possible. An RDF database would have the advantage of better support for ontology languages, e.g., OWL, and often provide the corresponding built-in inference mechanisms for ontological reasoning, which would probably have been of limited utility in the presented use case, though.

Regarding prompt design, we adopted the principles mentioned in the GPT-4.1 prompting guide [42] as well as best-practice approaches regarding query generation with ChatGPT from the experiments by Meyer et al. [21]. We used a zero-shot approach, i.e., an approach without task-specific training, because its positive performance has already been demonstrated in previous work by Sciannameo et al. [43] on knowledge extraction using LLMs and work by Carta et al. [44] on knowledge graph construction using LLMs. Furthermore, the use of a zero-shot approach ensures the ability of the application to process various case settings [22], which was essential due to the diversity of VAT cases. The cases used for prompt engineering were taken from course material on chain transactions by Kollmann [28], a recognized VAT expert. A total of 20 cases were selected, which cover various real-world scenarios and contain all the necessary information to solve and represent a valid chain transaction.

Listing 1 shows a condensed version of a prompt, translated to English from the original German version. The full prompt can be viewed in the online repository [26]. The prompt was developed through an iterative process involving 35 experiments using the Kollmann cases, with the prompt refined after each iteration to achieve an accuracy of 100 % on the Kollmann cases. The prompt is structured as follows (see Listing 1). Initially, an identity is assigned to the model and a number of high-level response rules are specified. Then, detailed instructions with example statements are declared to define a specific behavior for recurring patterns in the text. In addition, we opted for a structured output schema to ensure that the output contains the Cypher query to insert the knowledge graph into

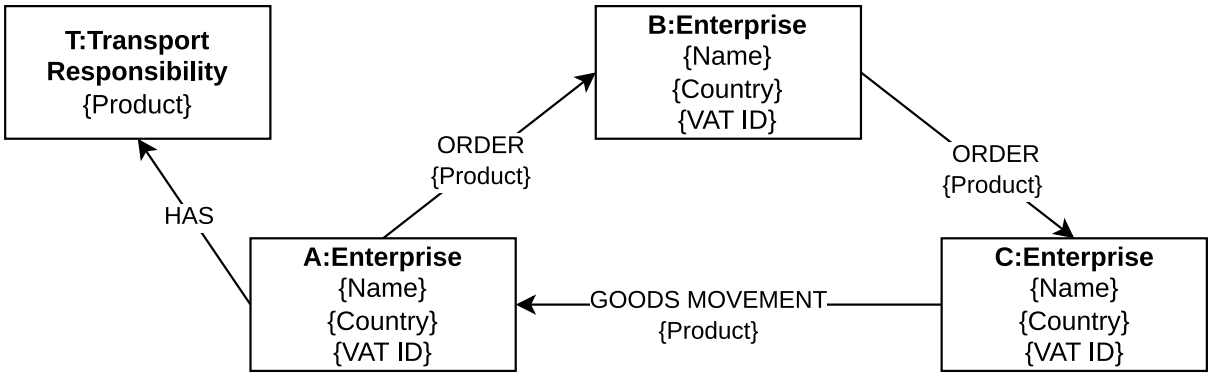

Figure 4: Illustration of the schema of the knowledge graph representation of VAT chain-transaction case descriptions

the graph database, an explanation of how the query statements were derived, and a summary of the overall case. Explanation and summary are important for a human tax advisor to be able to validate the results of the extraction.

The underlying schema of the knowledge graph was defined by a restricted vocabulary in the prompt. The knowledge graph's content is illustrated in Fig. 4. The knowledge graph consists of two types of nodes, namely *Enterprise* nodes and *Transport Responsibility* nodes. *Enterprise* nodes have *Name*, *Country*, and *VATid*properties. Furthermore, *Enterprise* nodes in the extracted knowledge graph are connected by two types of relationships, namely *ORDER* and *GOODS MOVEMENT*. The relationships are directed and have a *Product* property. The *Transport Responsibility* node, which has a Product property, is connected to an *Enterprise* node by the *HAS* relationship, which does not have any property. Beside the case description as basis for the knowledge graph construction, a tax law specialist was engaged in the development of the knowledge graph and validated it in accordance with tax law provisions and juridical methods.

3.2. Rule-Based Identification of Movable Supply

The identification of the movable supply is performed using a combination of Cypher queries and programmatic control structures (if–then–else statements) with logical expressions in Python. The implementation includes a check whether the chain transaction is valid. The required information for evaluating the rules is extracted using a set of Cypher queries, which are executed over the previously constructed knowledge graph stored in the graph database. In particular, the queries retrieve the first and last enterprise in the chain transaction, the country of dispatch of the goods, and the transport responsibility as well as the numbers of goods traded, orders, enterprises, movement of goods, and transport responsibilities. Furthermore, using the outputs from the previous queries, other queries retrieved the first and last supply, the intermediary supplies, and the supply of the enterprise before the intermediary.

We derived the rules from §3(15) of the Austrian Value-Added Tax Act (Umsatzsteuergesetz 1994) [34], which we therefore refer to as law-based rules. We chose this law-based process since in standard chain-transaction cases the law offers a clear step-by-step guide on how such a case must be solved. The first set of rules uses the information from the queries to check if the chain transaction is valid or not (Listing 2). The second set of rules, when applied on a valid chain transaction, uses information about the first, last, and intermediary enterprises in the chain transaction to determine the movable supply (Listing 3).

Besides the textual representation of the movable supply, our implementation generates a visualization of the chain transaction (see Fig. 3), using the Graphviz library [45], the output of which is also stored in the knowledge graph. Thus, the implementation generates a machine-readable output on the one hand and a visual representation comprehensible for humans on the other hand.

Listing 2: Validation of Chain Transaction

```
IF number of ORDER relationships is >= 2 AND
   number of ENTERPRISE nodes >= 3 AND
   the GOODS MOVEMENT edge starts at the first enterprise AND
   the GOODS MOVEMENT edge ends at the last enterprise  AND
   all product properties of edges have the same value AND
   there exists only one GOODS MOVEMENT edge AND
   there is only one Transport Responsibility node
THEN
    Return "Valid Chain Transaction"
ELSE
    Return "Invalid Chain Transaction"
```

Listing 3: Identification of Movable Supply Type

```
IF first_enterprise HAS Transport Responsibility THEN
  movable_supply := "First Supply"
ELSE IF last_enterprise HAS Transport responsibility THEN
  movable_supply := "Last Supply"
ELSE IF intermediary HAS Transport responsibility THEN
  IF country of VATid of intermediary =
     country of first_enterprise THEN
    movable_supply := "Intermediary's Supply"
  ELSE
    movable_supply := "Before Intermediary Supply"
```

4. Experimental Evaluation

In this section, we describe the datasets and the setup for experiments with the implementation of the proposed method. We then present and discuss the results of the experiments.

4.1. Datasets and Experimental Setup

To evaluate the proposed method and the prototype implementation, we selected a total of 167 VAT chain-transaction cases from four different sources. Most cases were obtained from two practically-oriented textbooks, namely, 72 cases from a Mayr and Weinzierl [29] and 67 cases from Dupont [30]. Another 24 cases were selected from real-world cases, which were collected by a tax consultant. The remaining four examples were taken from an exam on tax management at Johannes Kepler University Linz. The dataset is available in an online repository [27].

With regard to the length of the case descriptions from the different datasets, measured in terms of token size, which we obtained from LangSmith, we noticed the following differences (see Table 1). Relative to the "training set" by Kollmann, which we used for prompt engineering, the Mayr and Weinzierl (M&W) as well as the Dupont datasets were similar to the Kollmann dataset in terms of mean token size. However, the higher standard deviation and maximum values for M&W and Dupont indicate that those dataset also contained some cases with longer descriptions. The longest M&W case description was 250 tokens longer, and the longest Dupont case was 177 tokens longer than the mean of the Kollmann dataset. The M&W and Dupont cases were more similar to real-life cases than those in the Kollmann dataset. The real-world cases were more varied in terms of token size. In general, the real-world cases were even longer than the M&W and Dupont cases. The cases from the tax management exam were similar to the real-world cases in terms of extensiveness and structure.

In summary, the real-world cases and the exam cases had longer descriptions and either contained more information which was not necessary to solve the case or the case was more extensive in general. Consequently, they were more complex and difficult to understand, which made them harder to solve. The M&W, Dupont and the exam datasets already provided sample solutions containing the movable

Table 1
Summary statistics of datasets, including the number of cases per dataset as well as the maximum and minimum token size of cases along with standard deviation of token size per dataset

Dataset	Cases	Max	Min	Mean	SD
Kollmann	20	1 475	1 314	1 402,7	38,4
M&W	72	1 725	1 285	1 409,6	89,7
Dupont	67	1 653	1 284	1 421,4	82,8
Real-World	24	1 936	1 277	1 540,6	190,8
Exam	4	2 230	1 460	1 736,0	338,8

supply, which was necessary for evaluating the performance of the implementation. The real-world cases were solved by tax consultants, who manually identified the movable supply so that the performance of the implementation could be evaluated.

Each case was manually checked to ensure that the case was complete and not an exceptional case. Incomplete and exceptional cases were excluded from further consideration. This restriction was necessary because assumptions would have been necessary without complete information and this was fundamentally limited by the configuration of the LLM. In addition, the use of law-based rules also prevented the application from interpreting the law or making assumptions. A total of 15 cases from M&W had to be excluded because they were exceptional cases. Seven of these excluded cases involved incomplete transportation of goods (e.g., due to a car accident), and six cases fell under an exemption for electronic platform transactions, which are handled differently to normal chain transactions. One case involved a credulity assumption, which leads to different outcome than a regular case, and one case contained contradictory information and was therefore undecidable. Moreover, two exceptional cases were excluded from the Dupont dataset because they contained a shared transport responsibility, which is also handled differently than regular chain transactions. From the real-world cases, nine cases were excluded. Of these excluded cases, four cases were excluded for containing more than one transaction, which could not be represented in the current implementation, while three cases were excluded due to missing or unclear information that cannot be resolved by a human and, therefore, also cannot be resolved by the implementation. One real-world case was excluded due to the use of a consignment warehouse, which required special handling, and one case was excluded due to shared transportation responsibility. Eventually, we had to exclude also one exam case because it involved more than one transaction. Ultimately, 57 cases of M&W, 65 of Dupont, 15 real-world cases, and 3 exam cases were suitable for the experimental consideration. The datasets are available in an online repository [27].

Regarding real-world applicability, the exclusion of certain cases in the evaluation means the following. The use of an AI-based decision support system for VAT cases is most useful for routine cases. Exceptional and more complicated cases have to be reviewed from a human tax expert. Such cases would not be solved by the implemented prototype.

4.2. Results

Table 2 summarizes the performance of the LLM for knowledge graph construction from the natural-language case descriptions. For the Kollmann dataset, which was used for prompt engineering, the LLM performs with 100 % accuracy, which is not surprising since we optimized the prompt based on these cases. For our "test sets" M&W and Dupont the accuracy for knowledge graph construction is 92.98 % and 95.38 %, respectively. From the M&W dataset, four cases were not represented correctly, even though they included a valid chain transaction and the description included the required information. From the Dupont dataset, three cases were not represented correctly. Focusing on the real-world cases, the implementation performed at 93.33 % accuracy, with one incorrectly represented case. The implementation was able to obtain correct representations for all three of the considered exam questions.

Table 3 summarizes the performance of the rules in correctly identifying the movable supply in the correctly represented VAT chain-transaction cases. Only those cases were considered where the LLM

Table 2
Accuracy of knowledge graph representations obtained from the natural-language case descriptions

	Kollmann	M&W	Dupont	Real-World	Exam
Total cases	20	72	67	24	4
Invalid cases	-	15	2	9	1
Considered cases	20	57	65	15	3
Correct	20	53	62	14	3
Wrong	-	4	3	1	-
(Accuracy)	*(100 %)*	*(92.98 %)*	*(95.38 %)*	*(93.33 %)*	*(100 %)*

Table 3
Accuracy of rule-based identification of movable supply from correct knowledge graph representations of VAT chain-transaction cases

	Kollmann	M&W	Dupont	Real-World	Exam
Considered cases	20	53	62	14	3
Correct	20	53	62	14	3
Wrong	-	-	-	-	-
(Accuracy)	*(100 %)*	*(100 %)*	*(100 %)*	*(100 %)*	*(100 %)*

produced a correct knowledge-graph representation. The results demonstrate the proper implementation of the rules, which will provide the correct output when using correct knowledge-graph representations as input.

5. Related Work

In the following, we review existing literature with respect to the contribution in this paper. We present a novel LLM application in the field of tax law. The proposed approach aims to improve the reliability and explainability of LLM applications by incorporating KGs and rule-based decision-making to mitigate the problem of hallucination in sensitive domains and to provide explainable and reproducible results from LLMs.

5.1. LLM Applications in Tax Law

Our work presents a novel LLM application in the legal domain, the proposed method allowing to analyze VAT chain-transaction cases in Austrian tax law. Related LLM applications are designed for various specific tasks, which can be categorized into applications to summarize legal documents [12, 11], extract knowledge from legal documents [46], provide legal advice [10, 7, 9, 8], perform metamorphic tax-software tests [6], or solve legal exams [47, 48]. The AMELIA system [12] and OntoVAT [46] focus specifically on VAT law. AMELIA incorporates LLMs for argument mining on decisions on cases in Italian VAT la. OntoVAT uses ontologies for knowledge extraction and LLMs for generating numeric embeddings. Other related work [31, 32] employs LLMs to analyze VAT cases, investigating the potential of LLMs to identify the place of delivery from textual descriptions of VAT textbook cases as well as a general analysis of real-world VAT cases. Retrieval-augmented generation and fine-tuning are employed to improve performance of LLMs. Crucially, unlike this paper, the related work employs LLM for both natural language processing and legal reasoning. This approach differs from our approach: In our case, the LLM output is a structured representation of the natural-language case description and the law-based rules serve for actual legal reasoning.

5.2. Using LLMs for Knowledge Graph Construction

Our work provides insights into the possibilities of KG construction for solving complex tasks using LLMs. The KG produced by the output of the LLM serves as the basis for actual legal decision-making. The work of [21] already illustrated the possibilities of KG construction via simple prompt instructions. Furthermore, [22] emphasized the use of zero-shot approaches to ensure applicability across various scenarios. Related work demonstrated that zero-shot approaches can efficiently extract knowledge graphs [44] and that they can outperform one-shot approaches for extraction tasks with GPT-4 in terms of accurately extracted entities [23]. Other work [23] points out that LLMs do have limitations in KG construction tasks, which can effectively be addressed by prompt engineering, though. Therefore, we improved the prompt by providing the LLM with instructions on how to process recurring textual patterns and formulations to increase the accuracy of the output. Nevertheless, hallucinations limit the potential of LLMs for KG construction [23]. Moreover, hallucination is a frequently mentioned problem that leads to random and untraceable results [8, 15]. In particular, LLM applications in the legal field are severely restricted by hallucination, since comprehensible and explainable results are required in this critical domain [14, 15]. Although there are techniques to detect hallucinations [16], the corresponding results still need to be compared with related literature to ensure their factuality and plausibility [17]. In our work, we aimed to mitigate hallucinations by removing randomization with a temperature setting of 0, which was also proposed by Beutel et al. [38]. In addition, we used structured knowledge to create a comprehensible and reliable basis for decision-making as well as to reduce the risk of ambiguous and non-reproducible results.

6. Summary and Future Work

Tax advisors receive natural-language case descriptions from clients, which must be interpreted before rendering a decision on tax liabilities. In this paper, we proposed a method and implementation for using LLMs for structuring descriptions of VAT chain-transaction cases and apply rules derived from VAT law to produce decisions compliant with the law. As opposed to purely LLM-based approaches, the presented method is more reliable and explainable, the validity of the results can be easier checked by human tax advisors and used as evidence for purposes of tax compliance and communication with financial authorities.

We evaluated the approach using realistic use cases from textbooks and real-world practice. Improvements could be made regarding the handling of natural-language descriptions that actually contain more than one case. Future work will also investigate the applicability of the approach, where LLMs serve to extract structured representations from natural-language case descriptions and legal-based rules serve for actual decision-making, to other areas of tax law.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull to check spelling and grammar, as well as for suggestions for rewording and paraphrasing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Kelsen, Law and logic, in: Essays in Legal and Moral Philosophy, Springer, 1973, pp. 228–253. doi:10.1007/978-94-010-2653-6_10.

[2] J. Shi, Q. Guo, Y. Liao, Y. Wang, S. Chen, S. Liang, Legal-LM: Knowledge Graph Enhanced Large Language Models for Law Consulting, in: D.-S. Huang, Z. Si, C. Zhang (Eds.), Advanced Intelligent Computing Technology and Applications, volume 14878, Springer Nature Singapore, Singapore,

2024, pp. 175–186. doi:10.1007/978-981-97-5672-8_15, series Title: Lecture Notes in Computer Science.

[3] J. Shi, Q. Guo, Y. Liao, S. Liang, LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework, in: D.-S. Huang, Z. Si, W. Chen (Eds.), Advanced Intelligent Computing Technology and Applications, volume 14880, Springer Nature Singapore, Singapore, 2024, pp. 25–37. doi:10.1007/978-981-97-5678-0_3, series Title: Lecture Notes in Computer Science.

[4] D. Shu, H. Zhao, X. Liu, D. Demeter, M. Du, Y. Zhang, LawLLM: Law Large Language Model for the US Legal System, 2024. doi:10.1145/3627673.3680020, arXiv:2407.21065 [cs].

[5] S. Ghosh, D. Verma, B. Ganesan, P. Bindal, V. Kumar, V. Bhatnagar, Human centered ai for indian legal text analytics, 2024. arXiv:2403.10944.

[6] D. Srinivas, R. Das, S. Tizpaz-Niari, A. Trivedi, M. L. Pacheco, On the potential and limitations of few-shot in-context learning to generate metamorphic specifications for tax preparation software, 2023.

[7] M. Kaoutar, B. Chaima, B. Omar, B. Outmane, Unlocking the Potential of Large Language Models in Legal Discourse: Challenges, Solutions, and Future Directions, in: 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), 2024. doi:10.1109/ICDS62089.2024.10756345.

[8] T. Seabrooke, E. Schneiders, L. Dowthwaite, J. Krook, N. Leesakul, J. Clos, H. Maior, J. Fischer, A Survey of Lay People's Willingness to Generate Legal Advice using Large Language Models (LLMs), in: Proceedings of the Second International Symposium on Trustworthy Autonomous Systems, 2024. doi:10.1145/3686038.3686043.

[9] J. Nay, D. Karamardian, S. Lawsky, W. Tao, M. Bhat, R. Jain, A. Lee, J. Choi, J. Kasai, Large language models as tax attorneys: a case study in legal capabilities emergence, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 382 (2024). doi:10.1098/rsta.2023.0159.

[10] J. Presa, C. Camilo Junior, S. Oliveira, Evaluating Large Language Models for Tax Law Reasoning, in: Intelligent Systems, volume 15412 of *LNAI*, 2025, pp. 460–474. doi:10.1007/978-3-031-79029-4_32.

[11] A. Fidelangeli, F. Galli, A. Loreggia, G. Pisano, R. Rovatti, P. Santin, G. Sartor, The Summarization of Italian Tax-Law Decisions: The Case of the PRODIGIT Project, IEEE Access 13 (2025) 38833–38855. doi:10.1109/ACCESS.2025.3545419.

[12] G. Grundler, A. Galassi, P. Santin, A. Fidelangeli, F. Galli, E. Palmieri, F. Lagioia, G. Sartor, P. Torroni, AMELIA - Argument Mining Evaluation on Legal documents in ItAlian: A CALAMITA Challenge, in: CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, 2024. URL: https://ceur-ws.org/Vol-3878/124_calamita_long.pdf.

[13] A. Deroy, K. Ghosh, S. Ghosh, How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?, 2023. doi:10.48550/arXiv.2306.01248, arXiv:2306.01248 [cs].

[14] R. El Hamdani, T. Bonald, F. D. Malliaros, N. Holzenberger, F. Suchanek, The Factuality of Large Language Models in the Legal Domain, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, ACM, Boise ID USA, 2024, pp. 3741–3746. doi:10.1145/3627673.3679961.

[15] M. Dahl, V. Magesh, M. Suzgun, D. E. Ho, Large legal fictions: Profiling legal hallucinations in large language models, Journal of Legal Analysis 16 (2024) 64–93. doi:10.1093/jla/laae003.

[16] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, Y. Xiao, Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, ACM, Birmingham United Kingdom, 2023, pp. 245–255. doi:10.1145/3583780.3614905.

[17] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, Hallucination-free? assessing the reliability of leading ai legal research tools, Journal of Empirical Legal Studies 22 (2025) 216–242. doi:https://doi.org/10.1111/jels.12413.

arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jels.12413.
[18] P. Bellan, M. Dragoni, C. Ghidini, Process Knowledge Extraction and Knowledge Graph Construction Through Prompting: A Quantitative Analysis, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1634–1641. doi:10.1145/3605098.3635957.
[19] A. G. Regino, J. Cesar dos Reis, Generating e-commerce related knowledge graph from text: Open challenges and early results using llms, 2024.
[20] D. N. Ribeiro, K. Forbus, Combining analogy with language models for knowledge extraction, in: 3rd Conference on Automated Knowledge Base Construction, 2021. doi:10.24432/C5KK5X.
[21] L.-P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT, in: C. Zinke-Wehlmann, J. Friedrich (Eds.), First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow, Springer Fachmedien Wiesbaden, Wiesbaden, 2024, pp. 103–115. doi:10.1007/978-3-658-43705-3_8, series Title: Informatik aktuell.
[22] Y. Lairgi, L. Moncla, R. Cazabet, K. Benabdeslem, P. Cléau, iText2KG: Incremental Knowledge Graphs Construction Using Large Language Models, 2024. doi:10.48550/ARXIV.2409.03284, version Number: 1.
[23] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities, World Wide Web 27 (2024) 58. doi:10.1007/s11280-024-01297-w.
[24] M. A. Haque, M. Kamal, R. George, K. D. Gupta, Utilizing structural metrics from knowledge graphs to enhance the robustness quantification of large language models, International Journal of Data Science and Analytics (2024). doi:10.1007/s41060-024-00643-5.
[25] R. J. Wieringa, Design Science Methodology for Information Systems and Software Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. doi:10.1007/978-3-662-43839-8.
[26] L. Knogler, VAT chain transactions solver - the application for the research paper: Using large language models and knowledge graphs for deciding VAT chain-transaction cases in Austrian tax law, 2025. doi:10.5281/zenodo.16112381.
[27] L. Knogler, Experimental results for the research paper: Using large language models and knowledge graphs for deciding VAT chain-transaction cases in Austrian tax law, 2025. doi:10.5281/zenodo.16112863.
[28] G. Kollmann, Reihen- und Dreiecksgeschäfte, Umsatzsteuer aktuell, 2024. URL: http://www.gerh-kollmann.at/Reihengeschaefte%20Kollmann%20August%202024.pdf.
[29] M. Mayr, C. Weinzierl, SWK-Spezial Reihen- und Dreiecksgeschäfte, SWK-Spezial, 3. edition 2023 ed., Linde Verlag Ges.m.b.H, Wien, 2023.
[30] F. Dupont, Umsatzsteuer: Reihen-und Dreiecksgeschäfte in Beispielen; Praxishandbuch; mit praxisgerechten Erläuterungen, Grafiken und Beispielen, Kitzler, 2012.
[31] M. Luketina, C. G. Schuetz, T. Wageneder, An Experimental Evaluation of the Capability of Large Language Models to Reason About Value-Added Tax Cases in Austrian Tax Law, Proceedings of the 2024 Pre-ICIS SIGDSA Symposium (2024). URL: https://aisel.aisnet.org/sigdsa2024/15.
[32] A. Benkel, Using Large Language Models for Legal Decision Making in Austrian Value-Added Tax Law: an Experimental Investigation of Retrieval-Augmented Generation and Fine-Tuning, 2025.
[33] European Commission, VAT rules and rates: standard, special & reduced rates, 2025. URL: https://europa.eu/youreurope/business/taxation/vat/vat-rules-rates/index_en.htm.
[34] Umsatzsteuergesetz 1994, 2023. URL: https://ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10004873&FassungVom=2025-06-05, accessed: 2025-06-04.
[35] Austrian Federal Ministry of Finance, Business Service Portal: VAT Identification Number, 2025. https://www.usp.gv.at/en/themen/steuern-finanzen/umsatzsteuer-ueberblick/weitere-informationen-zur-umsatzsteuer/umsaetze-mit-auslandsbezug/umsatzsteuer-identifikationsnummer.html, accessed 2025-08-01.
[36] OpenAI, Introducing GPT-4.1 in the API, 2025. URL: https://openai.com/index/gpt-4-1/.
[37] LangSmith, Get started with LangSmith | LangSmith, 2025. URL: https://docs.smith.langchain.com/.

[38] G. Beutel, E. Geerits, J. T. Kielstein, Artificial hallucination: GPT on LSD?, Critical Care 27 (2023) 148. doi:10.1186/s13054-023-04425-6.
[39] G. Kellogg, O. Hartig, P.-A. Champin, A. Seaborne, RDF 1.2 Concepts and Abstract Data Model – W3C Working Draft 07 October 2025, 2025. URL: https://www.w3.org/TR/2025/WD-rdf12-concepts-20251007/.
[40] O. Hartig, A. Seaborne, R. Taelman, G. Williams, T. Pellissier Tanon, SPARQL 1.2 Query Language – W3C Working Draft 25 September 2025, 2025. URL: https://www.w3.org/TR/2025/WD-rdf12-concepts-20251007/.
[41] T. Eiter, G. Gottlob, H. Mannila, Disjunctive datalog, ACM Transactions on Database Systems 22 (1997) 364–418. doi:10.1145/261124.261126.
[42] N. MacCallum, J. Lee, GPT-4.1 Prompting Guide | OpenAI Cookbook, 2025. URL: https://cookbook.openai.com/examples/gpt4-1_prompting_guide, accessed: 2025-05-22.
[43] V. Sciannameo, D. J. Pagliari, S. Urru, P. Grimaldi, H. Ocagli, S. Ahsani-Nasab, R. I. Comoretto, D. Gregori, P. Berchialla, Information extraction from medical case reports using OpenAI InstructGPT, Computer Methods and Programs in Biomedicine 255 (2024) 108326. doi:10.1016/j.cmpb.2024.108326.
[44] S. Carta, A. Giuliani, L. Piano, A. S. Podda, L. Pompianu, S. G. Tiddia, Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction, 2023. doi:10.48550/ARXIV.2307.01128, version Number: 1.
[45] Graphviz, About Graphviz, 2024. URL: https://graphviz.org/about/, accessed: 2025-05-27.
[46] D. Liga, A. Fidelangeli, R. Markovich, Using Ontological Knowledge and Large Language Model Vector Similarities to Extract Relevant Concepts in VAT-Related Legal Judgments, in: New Frontiers in Artificial Intelligence, volume 14644 of *LNAI*, 2024, pp. 115–131. doi:10.1007/978-3-031-60511-6_8.
[47] K. Inoshita, Assessing GPT's Legal Knowledge in Japanese Real Estate Transactions Exam, in: 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2024, pp. 149–155. doi:10.1109/3ICT64318.2024.10824669.
[48] L. Xu, C. Hu, H. Zhang, J. Zhai, W. Tang, Y. Li, Z. Peng, Q. Chen, S. Sun, A. Ji, Y. Sun, Z. Liu, S. Wen, L. Bin, Surpassing Human Counterparts: A Breakthrough Achievement of Large Language Models in Professional Tax Qualification Examinations in China, in: 2024 IEEE Conference on Artificial Intelligence (CAI), 2024, pp. 1365–1370. doi:10.1109/CAI59869.2024.00243.