European Research Studies Journal
Volume XXVII, Issue S2, 2024

pp. 1057-1094

Generative AI as Tax Attorneys: Exploring Legal Understanding Through Experiments

Submitted 18/10/24, 1st revision 10/11/24, 2nd revision 26/11/24, accepted 15/12/24

Tomasz Strąk[1]

Abstract:

Purpose: *The purpose of the research presented in this article is to assess LLM models' ability to understand the law's language and legal reasoning for tax law. The choice of the tax law was dictated by the universality of its application (the easily accessible large corpus of tests) and the fact that in the case of European Union member states (such as Poland), this law is partially harmonised. These circumstances make it possible to reduce one of the indicated barriers to applying LLM in law (the multilingual and multicultural nature of the law). The latest GPT o1 - preview model from OpenAI was used for the study. The premiere of this model took place on 12.09.2024. It is a multilingual model with a universal rather than specialised nature, which would be explicitly trained for tax law use.*
Design/Methodology/Approach: *The research used an experimental method in which selected GPT models simulated the responses of a tax law expert. The research used two GPT models by OpenAI: GPT - 4 (available 14.03.2023) and GPT o1-preview (available 12.09.2024). The method used is an extension of the Turing Test concept (Turing, 1950), in which the AI model is intended to mimic human communication by assessing the ability to think logically, be creative and understand the context. Four research experiments were conducted. The first experiment assessed the LLM's understanding of the language of law, the second and third assessed the LLM's understanding of legal language, and the fourth assessed the LLM's legal reasoning skills.*
Findings: *The obtained results of the conducted research about LLM models allow the following conclusions to be formulated for Polish tax law: 1) the quality of understanding and legal reasoning of the GPT models is such that the models help support the work of professional tax advisors; 2) the accuracy of the legal advice provided by the GPT o1 - preview model is too low for the model to be used to provide legal advice on its own; 3) the GPT o1 - preview model can predict the position of the NRAIC for a given factual situation with high probability; 4) the GPT o1 - preview model has legal reasoning skills at the level of a professional lawyer except the ability to analyse court decisions and PTRs; 5) in the case of court decisions and PTRs analysis, there was a solid hallucinatory effect in the conducted studies, which affected 50% of the analysed cases.; 6) the quality of the LLM's understanding and reasoning was significantly influenced by the size of the training set and the number of domain-specific questions asked.*
Practical Implications: *The research results have significant practical implications. They indicate the applicability of GPT models for tax attorneys and identify the main barriers to their practical application. The article also shows how to improve the accuracy of the models*

[1]*Professor of Economics, University of Szczecin, Institute of Economics and Finance, Department of Accounting, Szczecin, Poland, e-mail: tomasz.strak@usz.edu.pl;*

and significantly reduce the hallucinatory effect. The practical implementation of the presented research results may significantly affect the labour market for lawyers dealing with Polish and European tax law.
Originality/Value: *The research presents an original method of assessing the quality of legal reasoning based on the Quality of Legal Reasoning Indicator. It is also the first of its kind concerning Polish tax law. Therefore, it makes an important contribution to the development of Generative AI in the field of law, especially tax law.*

Keywords: *Generative AI, large language models (LLMs), generative pre-trained transformers (GPT), law, tax, legal datasets, judicial data, natural language processing (NLP).*

JEL Codes: *A1, C45, C99, D21, K10, K34.*

Paper type: *Research article,*

Acknowledgement: *This research was co-financed by the Minister of Science under the "Regional Excellence Initiative".*

1. Introduction

The market for legal services has changed little since the development of legal information systems (1973: the LEXIS system in the US). Until the discoveries regarding Generative AI (GenAI), academic research indicated a low probability of artificial intelligence systems being able to replace the lawyer in substantive work requiring reasoning and inference. It is worth mentioning here research by Carl Benedikt Frey and Michael A. Osborne of Oxford University (Frey and Osborne, 2013) indicating a very low probability of AI replacing the work of lawyers.

The situation in this respect changed dramatically in 2023 after ChatGPT -3 by OpenAI made its market debut in November 2022. Since then, there has been rapid development of GenAI, which McKinsey & Company defines as follows: generative artificial intelligence (AI) describes algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos (McKinsey & Company, 2023).

GenAI uses large language Models (LLM), which are a type of Artificial Narrow Intelligence (ANI) based on Machine Learning, including its specific variant, deep learning. Within LLMs, the most groundbreaking discovery was the GPT (Generative Pre-trained Transformers) networks, which can be defined as a family of artificial intelligence models based on a transformer architecture designed to generate, analyse, and understand natural language.

With the development of GenAI, the first GenAI Paralegal systems began to appear on the market. These systems support the lawyer in searching for the necessary information and analysing it, formulating motions, and preparing pleadings or opinions for clients.

The development of GenAI has resulted in studies on AI's impact on the labour market conducted since 2024, which indicate that legal professionals will be among the professions most affected by AI (Korgul and Swiecicki, 2024).

In order for GenAI models to be able to make a significant impact on the work of lawyers or to carry out some of the work currently carried out by lawyers themselves, they must have the ability to understand legal and jurisprudential language and the ability to reason legally (interpret the law).

The language of law is used in laws, international agreements, and regulations. Legal language is used in contracts, pleadings, legal opinions, etc. Statutory language resembles ordinary language, but sometimes, the resemblance is only superficial.

Legal reasoning is practical reasoning used in legal practice and legal science. Its characteristic features are its orientation towards legal norms (Warner, 2005) and its pragmatic nature. Legal reasoning is used in the interpretation and judicial and extrajudicial application of the law, the provision of legal advice, and other cases of attributing specific legal consequences to specific facts.

The main research problem is to answer the following questions: Are LLMs able to distinguish between the different meanings of words and phrases in a legal context, and can these models argue according to the rules of legal reasoning?

In addition, a significant problem is the occurrence of the hallucination effect in LLMs, which significantly limits the practical utility of these models in the case of the law.

These issues and the current state of knowledge in this area are presented, inter alia, in the article “To What Extent Have LLMs Reshaped the Legal Domain So Far? A Scoping Literature Review' (Padiu *et al.*, 2024), which provides a broad overview of the literature on the subject and the application solutions.

This article identifies the main main limitations of current LLMs in the legal domain: hallucination; the multilingual and multicultural nature of the law; the vast scale and complexity of legal data; the ever-changing nature of the law. The aim of the research presented in the article, in the context of the research problem posed, is to analyse whether the most developed universal LLM model currently has the ability to understand the law and the ability to argue in accordance with the principles of legal reasoning specific to Polish tax law.

The choice of the tax law was dictated by the universality of its application (the readily available large corpus of tests) and the fact that in the case of European Union member states (such as Poland), this law is partially harmonised. These circumstances make it possible to reduce one of the indicated barriers to applying LLM in law (the multilingual and multicultural nature of the law).

The latest GPT o1 - preview model from OpenAI, which premiered on 12.09.2024, was used for the study. It is a multilingual model of a universal rather than specialised nature, which would be trained explicitly for use in tax law.

In realising the research objective outlined above, the following research hypotheses were verified in the research presented:

Hypothesis 1: LLMs can learn to understand tax law at a level sufficient to pass the test exam for tax advisors.
Hypothesis 2: LLMs can provide tax law advice at the level of knowledge required for an expert in the field.
Hypothesis 3: For the given facts, LLMs can correctly predict the tax authority's position.
Hypothesis 4: LLMs can learn legal reasoning by formulating a substantively and methodically correct argument to a formulated conclusion in the field of tax law.

2. Literature Review

The research presented in this article concerns the analysis of LLM's legal reasoning and understanding (legal interpretation) of Polish tax law. In this context, the analysis of the state of knowledge concerning models of legal reasoning for ius civile and the research concerning the evaluation of the quality of LLM's legal reasoning and understanding are relevant.

Traditionally, legal language is understood as the language of law-making, i.e. the language of legal rules and norms (Wroblewski, 1948, 51-136), and legal language as the language of law application (Wroblewski, 1948, 136-183). A concept that encompasses legal and juridical language is the language of law (Pienkos 1999, 14).

For research on legal and juridical language, it is important to point to the research of A. Śliwicka. The article constitutes an overview. It provides a synthetic

description of the current state of research and the directions for the development of the thought, which is focused on the issue of the language of the regulations and legal norms, as well as the language of the legal theory and practice, in the previous century starting from the thirties (Swlicka, 2018).

In terms of models of legal reasoning (law interpretation) in the context of LLM, those studies that present models of legal reasoning seem particularly useful. For Polish legal sciences, two concepts of legal interpretation are widely accepted: the clarification concept (Dascal and Wróblewski, 1988) and the derivation concept (Zieliński, 2017).

The advantage of the concepts presented is that their operational character allows them to be presented as principles and guidelines. The clarification concept is based on the distinction between a direct understanding of a legal provision (situation of isomorphy) and an indirect understanding of a legal text through its interpretation (situation of interpretation).

The concept of 'direct understanding' of linguistic expressions occurs when, in the context of the use of legal language, there is no doubt that the factual situation under consideration either falls within or outside the scope set by the provision. A consequence of the introduction of the concept of ‘direct understanding’ is the adoption as a principle of the paremma “clara non sunt interpretanda” in the version referred to here as ‘primary’, and therefore as a principle that does not allow interpretation when the subject understands the phrase in question ‘directly’.

According to this conception, the interpretation process is carried out according to eighteen general rules: seven linguistic rules, six systemic rules, two functional rules, and three second-degree rules (Wroblewski, 1959).

The derivational concept (Zieliński, 2917) distinguishes between the interpretation of a legal provision and the interpretation of a legal text. Interpretation of a legal provision means a mental activity that replaces a legal provision P with an expression N equivalent to a provision P (on the basis of R rules), which is a norm of conduct or an element thereof. Interpretation of a legal text, on the other hand, consists of replacing, by means of R rules, a legal text with a set of norms of conduct equivalent to R rules to the legal text.

According to this conception, the process of interpretation is carried out according to four directives (general, ordering phase of interpretation, reconstructive phase of interpretation, perceptual phase of interpretation), within which principles, rules and guidelines are defined.

Another essential element of legal reasoning is the rules of legal interpretation, which include logical, instrumental, and axiological rules (Atria, 2001) and the scope of their application in tax law (Slupczewski, 2023).

Based on the literature presented, the following critical elements of the legal interpretation process can be identified: the legal qualification of a particular state of affairs, i.e. the assignment of legal rules or norms to a particular state of affairs, the principles of legal interpretation applied, the rules of legal inference, the correct reference in the argumentation to court decisions and the substantive correctness of the conclusion (conclusion).

When it comes to assessing the quality of LLM understanding and reasoning in the case of law, methods based on the evaluation of domain experts play an important role.

Below is key research on the application of the LLM model to the study of the quality of reasoning in expert judgment law.

Shui R., Cao Y., Wang X. Chua T. 2023. A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. Findings of the Association for Computational Linguistics: EMNLP 2023.

This study analysed the ability of LLMs to predict court judgements. The authors designed and tested practical solutions based on LLMs in a judgment prediction task. The models were used both stand-alone and in collaboration with information retrieval systems. Experts evaluated the accuracy and relevance of the generated answers in a legal context.

Jayakumar T., Farooqui F., Farooqui L. (2023). Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. arXiv:2311.08890.

In this study, the authors compare the performance of generic LLMs with models adapted to the legal domain in classifying contractual clauses. The experts evaluated the correctness and precision of the answers generated by the models in a legal context, highlighting the need for powerful law-specific models.

Thalken R., Stiglitz E. Mimno D., Milkens M. (2023). Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement. arXiv:2310.18440.

This paper presents the results of a study of LLMs' ability to classify legal reasoning according to the philosophy of jurisprudence. A team of experts created a new dataset with annotations and evaluated the performance of different models in this task, analysing the correspondence between human annotations and model predictions.

Janatian S., Westermann H., Tan J., Savelka J., Benyekhlef K. (2023). From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems. rXiv:2311.04911.

The authors investigated the extent to which LLMs can automatically extract structural representations from legal texts. Experts assessed the quality and accuracy of the structures generated by the models in the context of creating expert systems in the legal domain.

The above studies highlight the importance of expert analysis in evaluating the correctness and adequacy of responses generated by LLMs in the legal context, which is crucial for their practical application in legal practice.

This section concludes with a discussion of the study of LLMs' understanding of tax law presented in 'Large language models as tax attorneys: a case study in legal capabilities emergence (Nay *et al.*, 2024.) This study's focus is similar to the research presented in this article. The correctness of the responses generated by the tax law GPT models was assessed in various experimental settings and on different LLM models, including the largest and most modern GPT-4 OpenAI available. For questions relating to the tax provisions of the United States Code of Federal Regulations (CFR), the number of answers found to be correct exceeded 90%, and for Title 26 of the United States Code, 70%.

By analysing the responses generated by successively larger models, the scholars felt that they had found evidence of the emergence of LLM's ability to understand the law, which improved with each newer version of the GPT model. The increasing performance of LLMs in these tasks could have profound implications for legal practice in tax law.

The article points out that even if LLMs do not replace professional tax advisers, they can help them, mainly by providing a first draft of a document. This could significantly increase the efficiency of lawyers and reduce the cost of legal services, improving access to legal aid for many people who cannot afford it.

3. Research Methodology

An experimental method was used to verify the research hypotheses, in which selected GPT models simulated the responses of a tax law expert. The research used two GPT models by OpenAI: GPT - 4 (available 14.03.2023) and GPT o1-preview (available 12.09.2024). The method used is an extension of the Turing Test concept (Turing, 1950), in which the AI model is intended to mimic human communication by assessing the ability to think logically, be creative and understand the context.

Two advanced GPT models from OpenAI were available during the experiments: the GPT - 4o and GPT o1-preview. The GPT o1-preview model is designed to analyse complex and complicated problems, the solution of which requires an extensive inference process and the analysis of extensive sources of information. The provision of legal advice in the field of tax law belongs to such a class of problems. The GPT o1 - preview model was therefore used for the experimental study.

Experiments were conducted based on the following sources of information:

- questions from substantive tax law (588 questions) included in the list of questions and examination tasks for the examination for tax advisor attached to Resolution No. 4/VIII/2023 of the State Examination Commission for Tax Advising of 28 April 2023;
- database of tax questions in the Lex Wolters Kluwer system (tens of thousands of questions);
- database of private tax ruling (approximately 400 000 documents).

Individual tax rulings have a special place in these sources of information. According to the IMF (Waerzeggers and Cory, 2016) a private tax ruling (PTR) consists of advice that a taxpayer may seek from the tax authority in relation to the application of the tax law to their particular arrangement. PTRs are a valuable instrument for limiting or even removing tax risks of significant economic transactions. They expose primary reasoning of tax authorities and often form a basis of later interpretations (Strąk and Tyszynski, 2020).

Taxpayers do not file applications for a ruling lightly. As a rule, they only do it in complex and unusual cases. The corpus of tax rulings shows clearly the real-life problems of taxpayers and acts as a window into otherwise unobservable behaviours of taxpayers (Strąk and Tyszyński, 2020).

In a PTRs application, the taxpayer describes a state of affairs (a factual or future event) and, based on this, asks a question concerning its tax consequences. PTRs are structured as follows: the taxpayer's presentation of the fatal state of affairs, questions addressed to the National Revenue Administration Information Centre (NRAIC), the taxpayer's position with legal justification, the NRAIC's position with legal justification.

Four research experiments were conducted:

Experiment I aimed to verify hypothesis I and proceeded as follows:

- A random selection of 100 questions from the tax advisor tests;
- Asking these questions to the GPT model with a request to indicate the correct answer;
- Evaluation of the correctness of the answer given by the model.
- This experiment was conducted twice: in May 2024 using the GPT - 4 model and in October 2024 using the o1 - preview model.

Experiment II aimed to verify hypothesis II and included the following steps:

- A selection from Wolters Kluwer's LEX service of 40 tax law questions asked by users of that service in 2024;
- asking these questions to the GPT o1 - preview model;

- Comparison of the answer generated by the GPT model with its justification with the answer prepared by tax law experts;
- evaluation of the results obtained.

Experiment III aimed to verify research hypothesis III and proceeded as follows:

- From the period 2019 - IX 2024, a random selection of 45 individual tax interpretations on VAT, PIT and CIT;
- selecting from these interpretations descriptions of the facts along with the questions asked;
- making stylistic modifications to the factual descriptions and questions so as not to alter their substantive sense;
- feeding the GPT o1-preview model with facts and questions (45 experiments);
- assessing the compatibility of the responses received from the GPT model and the NIS position.

Experiment IV was the most elaborate and aimed at verifying research hypothesis IV. The first four stages of this experiment are the same as experiment three. The fifth stage consisted of an analysis of the quality of the legal reasoning carried out by the GPT o1-preview model, i.e. an assessment of the quality of the legal reasoning formulated by the model.

Based on the given facts and the formulated questions, the GPT model generated an answer with elaborate legal reasoning, which included an analysis of the rules relating to the facts under consideration (legal qualification), legal inference (inference methods and rules of interpretation) and an analysis of case law (court decisions and PTRs). Given that the reasoning assessment is qualitative and requires a high level of qualification in tax law fields, five experts with experience in applying tax law were engaged to carry it out.

Based on two concepts of legal interpretation, namely the clarification concept (Dascal and Wróblewski, 1988) and the derivation concept (Zieliński, 2017), the Quality of Legal Reasoning Indicator (QLR) was developed. The measure of this indicator is a qualitative assessment of five key aspects of legal reasoning: legal qualification, method of legal interpretation, rules of legal inference, jurisprudence and factual correctness of the conclusion. The expert assessed each of these factors for the response generated by the GPT model using the following scale: 0 - negative rating; 0.5 - partially positive rating; 1 - positive rating.

The final value of the QLR was determined as the arithmetic mean of the assessments made by the individual experts. The individual evaluation criteria are characterised below.

Legal qualification- the assessment verified whether the GPT model provided adequate tax law provisions for the tax problem analysed. If the answer did not

contain the underlying provisions for the problem analysed, the expert awarded 0 points; if it contained some of the underlying provisions, it awarded 0.5; if it contained all the underlying provisions, it awarded 1 point.

Methods of interpretation of the law - in this criterion, the experts assessed whether correct methods of interpretation of the law were used in justifying the position of the answer generated by the GPT model and in a manner consistent with the principles of interpretation of tax law (for example primacy of linguistic interpretation, in dubio pro tributario, compliance with Supreme Administrative Court decisions, compliance with established interpretative practice).

If the expert assessed that the methods of interpretation of the law were applied correctly in the analysed answer, 1 point was awarded; if the correct methods were applied but with some shortcomings, 0.5 points were awarded; in the case of an incorrect choice of a method of interpretation of the law or its application in an inappropriate manner, 0 points were awarded. In the case of this criterion, the experts' assessments were most divergent.

Rules of legal inference –in this criterion, the experts assessed whether the regimes of legal inference (logical, instrumental, axiological rules) relevant to the tax law were applied correctly. This criterion has the same problems as the previous criterion. There are, in practice, universal principles of legal inference that apply to tax law (e.g., analogia legis and analogia iuris to the advantage of the taxpayer) as well as should not be used in tax law (e.g., analogia legis and analogia iuris to the disadvantage of the taxpayer). Scoring rules for this criterion: 1 - correct application of the rules of legal interpretation; 0.5 - rules of legal inference applied correctly in principle, but the experts point out some shortcomings; 0 - incorrect application of the rules of legal interpretation.

Court decisions - the assessment verified that the GPT model cited vital court rulings and PTRs for the tax problem under analysis. If the GPT model did not provide any rulings or PTRs to justify its position, then the expert awarded 0 points for this criterion. Some experts also awarded 0 points if there was a hallucinatory effect (citing a non-existent ruling). 0.5 points were awarded if at least two rulings were given, and 1 point was awarded if at least four rulings were given or two rulings were given but crucial to the facts.

Substantive correctness of the conclusion - assessing whether the GPT model formulated the correct legal conclusion for the facts. This criterion was assessed as follows: 0 - wrong conclusion; 0.5 - partially correct; 1 - correct.

The QLR value for each expert was determined using the following formula:

$$QLR = \frac{\sum_i^5 c_i}{5},$$

Where:
QLR - Quality of Legal Reasoning Indicator.
C_i - assessment criteria, where:
C_1 - legal qualification,
$C_{2.}$ - methods of interpretation of the law,
C_3 - rules of legal reasoning,
C_4 - court decisions,
C_5 - substantive correctness of conclusions.

In summary, the proposed method belongs to the quality evaluation class and employs an approach specific to task-based evaluation and expert review. These are classic methods for evaluating the quality of understanding in GPT models. The experts also used the formal-dogmatic method to assess the quality of legal reasoning.

4. Research Results and Discussion

Experiment I:
Experiment I aimed to assess whether the most developed universal GPT models understand Polish tax law. In order to achieve such a goal, an objective criterion was used as a test from the state examination for tax advisors. The written part of the examination for tax advisors in Poland consists of solving a test of 100 questions and solving a task of drafting an address on behalf of a client to a tax authority or court. In order to pass the test, you must obtain 80% of the maximum number of points to solve the test.

The experiment used 100 randomly drawn test questions from the list of examination questions and tasks for the tax adviser examination attached to Resolution No. 4/VIII/2023 of the State Examination Commission for Tax Advising of 28 April 2023. Table 1 presents the structure of the tax advisor exam test questions drawn for Experiment I.

Table 1. *Structure of the test questions used in Experiment I*

Tax	Number of observations
PIT	20
CIT	20
VAT	20
Excise tax	10
Property tax	10
Inheritance and gift tax	10
Tax on means of transport	10
Total	100

Source: *Own study.*

Experiment I was attempted three times. The first time the Experiment I attempted was in November 2023. However, the GPT-4 model answered only some of the questions, giving clauses that it did not have up-to-date data and needed to be qualified to give legal advice.

Experiment I was conducted again (this time successfully) in May 2024 using the GPT—4 model to provide the answers. It was repeated in October 2024 using the GPT o1—preview model (the latest and most extensive GPT model) to provide the answers. Table 2 presents the results obtained using the GPT - 4 model (May 2024) and the GPT o1 - preview model (October 2024).

Table 2. *Results of experiment I*

Tax	GPT --4	GPT o1-preview
PIT	55%	80%
CIT	50%	85%
VAT	55%	80%
Excise tax	50%	80%
Property tax	50%	100%
Inheritance and gift tax	40%	70%
Tax on means of transport	30%	70%
Total	49%	81%

Source: *Own study.*

The results of the experiments show substantial progress in understanding Polish tax law for the generic GPT models, which was achieved in a short time. The GPT - 4 model in May 2024 had a score close to purely random and could, therefore, be considered not to understand Polish tax law. For model o1 - preview in October 2024, a score at the level required to pass this exam was obtained, and therefore, we are entitled to conclude that the GPT model understands Polish tax law. Therefore, the results obtained in Experiment I allowed for positive verification of hypothesis 1.

Experiment II:

Experiment II aimed to test whether the latest and most extensive generic GPT model available in October 2024 can solve practical tax problems, i.e. whether it can apply its knowledge of tax law to solve practical tax problems. The experiment used a tax advice database available to users of Poland's most famous legal information system (LEX), owned by Wolters Kluwer. This database contains questions from LEX users and answers prepared by Wolters Kluwer experts. Individual questions are presented as follows: question, answer, and justification. The GPT model generated an answer with the same structure, i.e., answer and justification.

During the experiment, for each answer generated by the GPT model, the consistency of the obtained answer and justification with the answer and justification

prepared by Wolters Kluwer experts was assessed. Table 3 presents the results obtained.

The result obtained demonstrates the GPT model's high skill in solving tax problems. The level of correctness of the answers indicates the GPT model's usefulness in supporting the expert in preparing answers to questions on tax problems. However, the accuracy of the legal advice provided needs to be higher (31 cases of incorrect advice out of 100 advice) for the model to be used to provide legal advice on its own.

Table 3. *Results of experiment II*

Tax	Correct answers	
	Number of observations	% observations
CIT	4	66,67%
PIT	5,5	55,00%
VAT	18	75,00%
Total	27,5	68,75%

Source: *Own study.*

There is a much higher correctness of VAT answers compared to income taxes. This is a consequence of the harmonisation of VAT at the European Union level, according to which national VAT rules must be consistent with those of the European Union. The GPT models answer the question of applying VAT based on VAT knowledge for all EU countries. Therefore, the knowledge base is much larger than other taxes in this case.

In addition, Polish users and users from other EU countries ask VAT questions about the GPT model. GPT models can learn from the activity of their users. In the case of VAT, the GPT model is constantly reviewed by a much larger number of users than in the case of income taxes. In their case, national regulations are not universal to the legal system. For LLMs, which are stochastic models, these factors significantly affect the quality of the answers obtained.

The result obtained allowed a negative verification of hypothesis 2. Despite this, the progress in this area is impressive. In May 2023, the GPT model needed help to provide helpful advice on Polish tax law. Currently (October 2024), they already have this ability.

Experiment III:

Experiment III aimed to assess whether the GPT model could formulate an appropriate conclusion based on a factual description made in formal language for the issue of a PTR. The factual description of the question that the GPT model analysed was 1 and 3 A4 pages long. The test examples included:

- was not legally complex and required the application of a linguistic interpretation of one or more provisions;
- complex factual situations that required the application of not only linguistic interpretation but also a purposive or systemic interpretation and referred to provisions that were not only covered by a single law;
- factual situations in which the taxpayer knowingly, in order to gain tax advantages, used legal institutions other than those relevant to the facts described to present the facts.

It should be noted that the test set was dominated by examples relating to points 2 and 3 (approximately 82% of cases). Therefore, the cases used for the experiment were mainly those with complicated tax implications. The results of Experiment III are presented in Table 4.

Table 4. *Results of experiment III*

Tax	Correct answers	
	Number of observations	% observations
CIT	10	66,67%
PIT	11	73,33%
VAT	12	80,00%
Total	33	73,33%

Source: *Own study.*

The results obtained allowed a positive verification of hypothesis 3. The model with a relevance of 73.33% predicts the NRAIC position. The level of relevance obtained is helpful for professionals preparing a PTR disbursement application.

As in Experiment II, better results were obtained for VAT than for income taxes. This further confirms the thesis that the size of the domain training set and the number of questions asked by users influence the quality of the answers obtained.

Experiment IV:

Experiment IV aimed to assess the quality of the legal argumentation generated by the GPT model and the correctness of the conclusions. It was, therefore, the most demanding test for the GPT model of all those conducted so far. This is because the model's response was compared with models of legal reasoning specific to the Polish legal system. However, the models are universally applicable to ius civile and to the legal system of the European Union.

Experiment IV assessed whether the GPT model understands tax laws and whether it can supportively apply the legal reasoning (model of interpretation) inherent in ius civile.

Five experts (three tax advisors, a lawyer, and a tax law lawyer) participated in the experiment. Each expert independently assessed the quality of the responses they

received by individually assessing the five criteria outlined in section 3. Table 5 presents the results of the experts' assessment of the Quality of Legal Reasoning Indicator.

Table 5. *Results of the expert assessment of the Quality of Legal Reasoning Indicator*

Expert	Quality of legal reasoning indicator	Legal qualification	Methods of interpretation of the law	Rules of legal reasoning	Court decisions	Substantive correctness of conclusions
Expert 1	65,78%	83,33%	71,11%	71,11%	30,00%	73,33%
Expert 2	68,00%	81,11%	76,67%	77,78%	31,11%	73,33%
Expert 3	59,33%	84,44%	70,00%	67,78%	0,00%	73,33%
Expert 4	67,33%	78,89%	70,00%	67,78%	46,67%	73,33%
Expert 5	62,44%	83,33%	77,78%	77,78%	0,00%	73,33%
Average value	64,58%	82,22%	73,11%	72,44%	21,56%	73,33%

Source: *Own study.*

The spread in expert judgement for individual parameters was as follows:

- Quality of legal reasoning indicator - 8.67%;
- Legal qualification - 5.56%;
- Methods of interpretation of the law - 7.78%
- Rules of legal reasoning - 10.00%;
- Court decisions - 46.67%
- Substantive correctness of conclusions - 0%.

The large discrepancies in the assessment of the court decisions criterion were due to the experts' different opinions on the identification of key judgements and their different approaches in the case of a hallucination effect. Expert 3 and expert 5 automatically awarded 0 points in the case of a hallucination effect, i.e., the GPT model giving a non-existent court decisions reference or PTR. Table 6 shows the averaged scores of the individual experts for the QLR indicator and its components.

Table 6. *Averaged scores of individual experts' assessments of the QLR dlave and its components*

Tax	Quality of legal reasoning indicator	Legal qualification -	Methods of interpretation of the law	Rules of legal reasoning	Court decisions	Substantive correctness of conclusions
CIT	60,13%	75,33%	67,33%	68,67%	22,67%	66,67%
PIT	65,73%	84,00%	74,00%	73,33%	23,33%	73,33%
VAT	67,87%	87,33%	78,00%	75,33%	18,67%	80,00%
Average value	64,58%	82,22%	73,11%	72,44%	21,56%	73,33%

Source: *Own study.*

The results obtained regarding the quality of the model's legal reasoning depend on two main factors: the correctness of the identification of the relevant court decisions and PTRs for a given factual situation and the accuracy of the conclusions.

To analyse in more depth the impact of the accuracy of the GPT model's conclusion on the value of the QLR indicator, its value was determined for those cases in which the GPT model formulated the correct conclusion (Table 7).

Good results were obtained for this group: the average QLR = 80.67% and very high values for the individual sub-indices, with the exception of Court decisions. OpenAI's GPT models generally do not perform well in correctly retrieving court decisions and PTRs for a particular state of affairs. However, it should be pointed out that solutions already exist today that use much less extensive LLMs than OpenAI models and allow for precise semantic searches of court judgements or tax interpretations (Strąk and Tuszyński, 2022).

In order to solve the identified problem, the Retrieval Augmented Generation) (RAG) architecture or retrieval augmentation of the language model is used in LLM (Chen, 2023; Gao, Yunfan *et al.*, 2023; Liu *et al.*, 2024).

Table 7. *Averaged scores of individual experts for QLR and its components for correct answers*

Tax	Quality of legal reasoning indicator	Legal qualification -	Methods of interpretation of the law	Rules of legal reasoning	Court decisions	Substantive correctness of conclusions
CIT	78,80%	90,00%	87,00%	88,00%	29,00%	100,00%
PIT	82,36%	96,36%	94,55%	92,73%	28,18%	100,00%
VAT	80,67%	99,17%	91,67%	89,17%	23,33%	100,00%
Average value	80,67%	95,45%	91,21%	90,00%	26,67%	100,00%

Source: *Own study.*

The RAG architecture comprises, broadly speaking, a language model combined with a search engine for the texts used to teach the model. The found text is then fed into the model's input, thus grounding the model in the virtual reality of the Internet. An example of this architecture is the Microsoft Bing search engine, which returns, among other things, information about the source of the information.

The RAG architecture is also the only primary method for reducing the hallucination of LLM models adapted to the legal domain (Padiu *et al.*, 2024).

Therefore, it can be assumed with a high degree of probability that, in the case of the GPT o1—preview model, supplementing it with a semantic search engine, i.e.,

building LLM models in the RAG architecture, will lead to very good results in terms of identifying judicial decisions and PTRs relating to a specific fatal condition. If, for the correctly identified examples, the Court decisions value was equal to 100%, then the QLR value would be 95.33%. LLM models, therefore, can learn legal reasoning at the level required for a professional lawyer, meaning that hypothesis 4 has been positively verified.

In addition, it should be pointed out that there are significant differences in the values of the sub-indices for Methods of interpretation of the law and Rules of legal reasoning for cases for which the model formed its conclusions correctly and for cases for which its conclusions were incorrect. As an example, the following results were obtained for VAT:

- Case with correct conclusion: methods of interpretation of the law = 91.67%; rules of legal reasoning = 89.17%;
- Case with wrong conclusion: methods of interpretation of the law = 50%; rules of legal reasoning = 50%.

Therefore, the quality of the legal inference impacts the correctness of the conclusion formed by the GPT model.

5. Conclusion

The obtained results of the conducted research about LLM models allow the formulation of the following main conclusions for Polish tax law:

- In 2024, there has been excellent progress in understanding and legal reasoning about OpenAI GPT models.
- The GPT models' quality of understanding and legal reasoning makes them useful tools for professional tax advisers. Their use can improve the quality of prepared positions and significantly reduce their costs.
- The accuracy of the legal advice provided by the GPT o1 preview model is too low for it to be used alone to provide legal advice.
- The GPT o1 - preview model can predict the NRAIC position for a given factual situation with high probability.
- GPT model o1 - preview has legal reasoning skills at the level of a professional lawyer except for the ability to analyse court decisions.
- There was a solid hallucinatory effect in the court decisions analysis in the studies conducted, which affected 50% of the cases analysed.
- The main limitation of the practical application of GPT models by lawyers is the strong hallucinatory effect when identifying rulings relating to the facts under analysis.
- However, the limitation mentioned in point 5 can easily be eliminated using the RAG architecture.

- The quality of LLM understanding and reasoning is significantly influenced by the size of the training set and the number of domain-specific questions asked. For this reason, the study obtained better results for VAT (harmonised at the European Union level) than for other taxes (for which country-specific rules predominate).
- GPT models, due to their up-to-date skills and significant advances in legal understanding and reasoning, will significantly impact Poland's labour market and tax advisory services.

The above conclusion is consistent with Nay J. et al. (2024) argument about US tax law.

To significantly reduce the universal model's hallucinatory effects and improve the accuracy of the answers, it is reasonable to conduct research on the construction of domain-specific models based on universal models using the RAG architecture and external models to validate the quality of legal reasoning. The quality of universal models demonstrates that using them to build specialised models is rational.

References:

Atria, F. 2021. On Law and Legal Reasoning. Oxford- Portland Oregon: Hart Publishing.

Chen, J. et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v2 [cs.CL] 29 Dec 2023.

Dascal, N., Wroblewski, J. 1988, Transparency and Doubt: Understanding and Interpretation in Pragmatics and in Law, 7 (2), 427-450.

Frey, C.B., Osborne, M.A. 2013. The future of employment: how susceptible are jobs to computerisation? Oxford: Oxford Martin Programme on Technology and Employment.

Gao, Y. et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv abs/2312.10997.

Janatian, S., Westermann, H., Tan, J., Savelka, J., Benyekhlef, K. 2023 From Text to Structure: Using Large Language Models to Support the Development of Legal Expert Systems. rXiv:2311.04911.

Jayakumar, T., Farooqui, F., Farooqui, L. 2023. large Language Models are legal but they are not: Making the case for a powerful LegalLLM. arXiv:2311.08890.

Korgul, K., Witczak, J., Święcicki, I. 2024. AI on the Polish labour market. Warsaw: Polish Economic Institute.

Liu, D., et al. 2024. RetrievalAttention: Accelerating Long-Context LLM Inference via Vector Retrieval. ArXiv abs/2409.10516.

McKinsey Explainers, What is generative AI? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai#/%20McKinsey&Company.

Nay, J., et al. 2024. 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, volume 382, issue 2270.

Pieńkos, J. 1999 Fundamentals of Jurisprudence. Language in law - law in language. Warsaw: Oficyna Prawnicza Muza SA.

Shui, R., Cao, Y., Wang, X., Chua, T. 2023. A Comprehensive Evaluation of Large Language Models on Legal Judgment Prediction. Findings of the Association for Computational Linguistics: EMNLP.

Słupczewski, M. 2023 Reasoning per analogiam in tax law. Warsaw: Wolters Kluwer Polska.

Strąk, T., Tuszyński, M. 2020 Quantitative analysis of a private tax rulings corpus. Procedia Computer Science, 176, 2445-2455.

Strąk, T., Tuszyński, M. 2022. NLP Based Retrieval of Semantically Similar Private Tax Rulings. Procedia Computer Science, 207, 2853-2864.

Thalken, R., Stiglitz, E., Mimno, D., Milkens, M. 2023. Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement. arXiv:2310.18440.

Turing, A. 1950. Computing Machinery and Intelligence, Mind, vol. LIX, no. 236, October, 433-460.

Waerzeggers, C., Hillier, C. 2016. Introducing an Advance Tax Ruling (ATR) Regim. Tax Law IMF Technical Note. Vol. 1. Washington: International Monetary Fund.

Warner, R. 2005. Adjudication and Legal Reasoning. In: Golding, M.P., Edmundson, W.A. (ed.), The Blackwell Guide to the Philosophy of Law and Legal Theory, Malden - Oxford - Carlton: Blackwell Publishing, 259-270.

Wroblewski, B. 1948. Legal and juridical language. Kraków: Polish Academy of Arts and Sciences.

Zieliński, M. 2017. Interpreting the law. Principles - rules - guidelines. Warsaw: Wolters Kluwer Polska.