Factual State Discovery Benchmark: Evaluating Fact Elicitation in Polish Tax Law

Mateusz Bystroński[1] **Kamil Tagowski**[1] **Denis Janiak**[1] **Julia Farganus**[1]
Łukasz Augustyniak[1] **Monika Kajdanowicz**[2] **Tomasz Kajdanowicz**[1]
[1]Wroclaw University of Science and Technology, [2]KRUK S.A.

Abstract

Before a tax authority can issue a ruling, it must receive a complete description of the taxpayer's situation—yet no benchmark measures whether language models can systematically elicit all relevant facts through dialogue. We introduce **FSDBench** (**Factual State Discovery Benchmark**), in which a discovery agent questions a simulated taxpayer grounded in a real tax document. The dataset comprises 500 narratives from official Polish tax interpretations, decomposed into 32 874 atomic facts with validated supported precision (97.6%), atomicity (93.8%), and sentence coverage (96.0%). Experiments with four models show that even the best system recovers only 77% of facts on easy samples and under 49% on hard samples after 50 turns. These findings establish conversational fact elicitation as a challenging open problem requiring retrieval-augmented and adaptive questioning strategies.[1]

1 Introduction

Before a tax authority can rule on the applicability of a legal provision, it must have a complete picture of the taxpayer's factual situation. In the Polish tax interpretation process, a taxpayer—or, more often, a professional advisor acting on their behalf—submits a detailed narrative describing their circumstances alongside a legal question to the National Tax and Customs Information Office (*Krajowa Informacja Skarbowa*). The quality of the resulting ruling depends directly on the completeness of this narrative: missing facts may lead the authority to apply the wrong provision, or to issue a ruling that is inapplicable to the actual situation.

In practice, assembling a complete factual account is one of the most time-consuming and expertise-intensive stages of the interpretation process. A tax advisor must conduct a structured interview with the client, probing systematically across financial, contractual, regulatory, and domain-specific dimensions—often over multiple sessions—until all legally relevant details have been surfaced. This requires not only deep knowledge of tax law but also the ability to anticipate which facts will be decisive under competing legal theories, a judgement that is difficult to formalise and impossible to fully automate without first understanding what "complete discovery" looks like in practice. Despite its centrality to the quality of tax rulings, this discovery phase has received little attention in the NLP literature.

Our contributions are as follows:

1. **Factual state discovery task.** We formalise conversational factual-state discovery as a multi-turn elicitation task under controlled information asymmetry. The protocol couples a discovery agent, a source-grounded stateless QA agent, and a semantic scorer, enforcing information asymmetry and preventing leakage so agents must actively elicit facts rather than passively receive them.

2. **FSDBench dataset.** We release a reusable benchmark resource: 500 official Polish tax-interpretation narratives paired with 32,874 atomic factual targets. The fact sets are produced by an LLM-assisted atomization pipeline and checked with intrinsic diagnostics for support (97.6%), atomicity (93.8%), and sentence coverage (96.0%).

3. **Coverage-oriented evaluation.** We propose an atomic-fact scoring pipeline that combines embedding retrieval with LLM-based semantic judgement, together with two metrics—*Fact Coverage Ratio* and *Conversational Efficiency*—that capture the fundamental trade-off between completeness of elicited information and the number of interaction steps required.

[1]Source code is available at github.com

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1262–1279

4. **Empirical evaluation of LLMs as discovery agents.** We benchmark multiple frontier and open-weight models, showing that factual discovery remains challenging, especially for long and complex cases, with substantial variance across instances.

2 Related Works

Multi-Turn Conversations and Information Gathering. While Large Language Models (LLMs) demonstrate strong performance in exam-like, single-turn Question Answering (QA) (Singhal et al., 2023), their accuracy drops significantly in interactive, multi-turn environments that require active information gathering (Johri et al., 2025). Off-the-shelf models struggle to prioritize follow-up questions when faced with incomplete information, often resorting to repetitive inquiries (Li et al., 2024). To address this, research has heavily focused on the medical domain, deploying agentic "Doctor-Patient" simulations (Johri et al., 2024; Li et al., 2024; Johri et al., 2025; Tu et al., 2024) to break away from the conventional reliance on static medical exams. Nevertheless, benchmarks like MEDIQ (Li et al., 2024) reveal that multi-turn medical dialogues do not outperform single-turn summaries unless guided by highly structured clinical frameworks. While medical systems heavily weigh the accuracy of the final prediction (the diagnosis) and can rely on strict, formalized guidelines (McDuff et al., 2025), legal fact-gathering is largely unstructured. Consequently, multi-turn agentic approaches in the legal domain remain scarce, with frameworks like LeCoDe (Yuan et al., 2025) and LexRAG (Li et al., 2025) evaluating the ability of LLMs to actively extract information from users, though they are currently limited to the Chinese language.

Benchmarking LLMs in the Legal Domain. Historically, LLM performance in the legal domain has been evaluated using static, single-turn tasks with known ground truths. Popular benchmarks such as (Chalkidis et al., 2022; Fei et al., 2024; Guha et al., 2023) rely on traditional, n-gram-based evaluation metrics like ROUGE-L or basic accuracy, failing to capture the complexities of legal consultations that require iterative fact-finding. Recent work has shifted toward interactive legal benchmarks. Notably, LeCoDe (Yuan et al., 2025) introduces 3,696 legal consultation dialogues from Chinese social media to evaluate clarification and advice clarity. Similarly, LexRAG (Li et al., 2025) provides 1,013 multi-turn dialogues, establishing the first benchmark for conversational RAG in the legal domain.

LLMs evaluation. Evaluating open-ended, multi-turn dialogues is difficult and costly for human experts; consequently, the "LLM-as-a-judge" paradigm (Chiang and Lee, 2023) has emerged as a scalable, cost-effective, and sample-independent alternative. Studies have shown a strong correlation between LLM judges and human evaluators (Zheng et al., 2023); for example, evaluations of medical dialogues using ChatGPT (e.g., the MedEval (Shi et al., 2023) dataset) demonstrated that automated evaluation can effectively replace experts in assessing relative performance between models. However, LLM judges can struggle with hallucination and domain-specific factual knowledge if left ungrounded (Bai et al., 2024; Zheng et al., 2023) . To remedy this, recent methodologies have shifted toward evaluation based on Atomic Facts—small, irreducible units of information. The FactScore (Min et al., 2023) framework pioneered this by extracting atomic facts from generated text via a strong language model and verifying them against a knowledge source (such as Wikipedia). In interactive settings, grounding evaluation in atomic facts has been shown to significantly reduce hallucination and provide a highly granular metric for conversational success (Gong et al., 2026, 2025; Li et al., 2024) . The agent's performance is objectively measured by how effectively its retrieved facts match the underlying atomic facts of the Client Agent, ensuring that evaluation is driven by strict factual extraction rather than conversational fluency alone.

3 Benchmark Description

We introduce the **Factual State Discovery Benchmark**, a task-oriented evaluation framework for measuring the ability of conversational agents to elicit factual information from a simulated taxpayer through multi-turn dialogue. The benchmark targets a core challenge in AI-assisted tax advisory: systematically uncovering all relevant facts from a client before any legal analysis can begin.

3.1 Motivation

In the Polish tax interpretation process, a taxpayer submits a detailed description of their factual situation alongside a legal question to the tax authority. The completeness of this factual description is

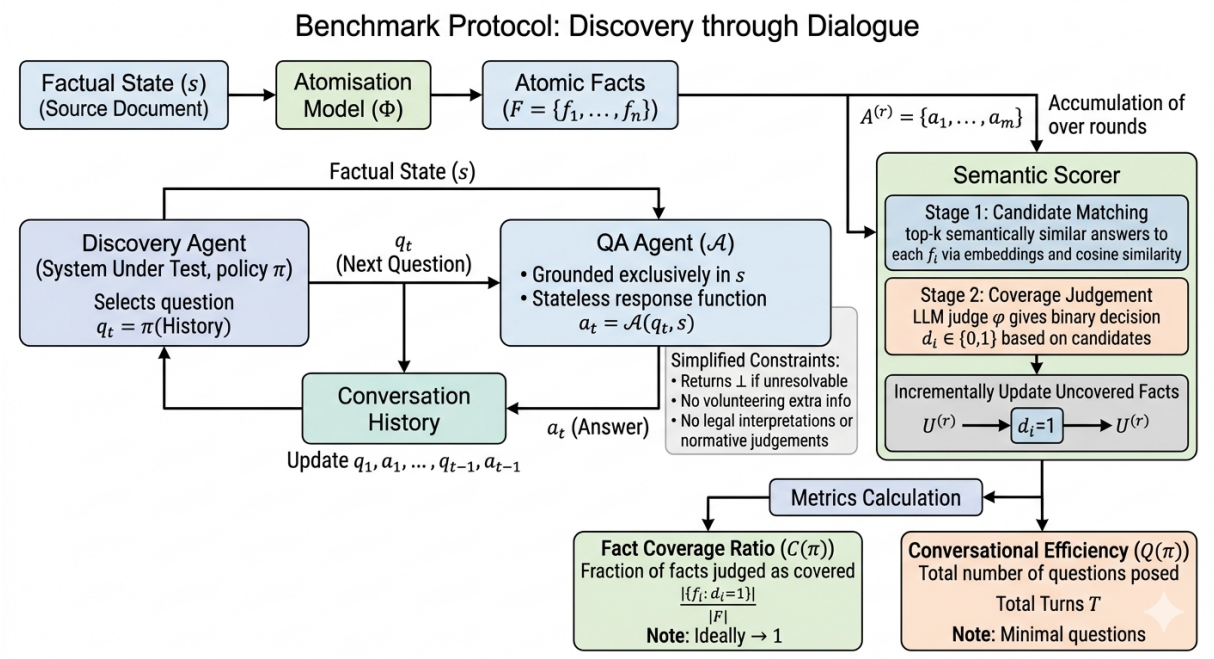

Figure 1: **Benchmark Protocol Architecture.** The workflow begins with the extraction of atomic facts (F) from a source document (s). The Discovery Agent (system under test) attempts to uncover these facts through iterative dialogue with a constrained QA Agent. A two-stage semantic scorer evaluates the accumulated answers ($A^{(r)}$) against the atomic facts to compute the final metrics: Fact Coverage Ratio ($C(\pi)$) and Conversational Efficiency ($Q(\pi)$).

critical—an incomplete or inaccurate account may lead to an incorrect or inapplicable ruling. In practice, eliciting this factual state requires a skilled advisor who asks targeted, comprehensive questions. Our benchmark operationalises this process as a measurable task for large language models.

3.2 Dataset

The dataset is derived from official Polish individual tax interpretations published by the National Tax Information Service (*Krajowa Informacja Skarbowa*). The source corpus comprises 488 917 interpretations published between July 2007 and June 2025, collected via the Eureka public API. Factual state narratives are extracted from raw HTML documents through an acquisition and parsing pipeline described in Appendix A. Each sample consists of two components:

1. **Factual state** (s): The verbatim factual narrative from a tax interpretation request, describing the taxpayer's situation in Polish.

2. **Atomic facts** ($F = \{f_1, f_2, \ldots, f_n\}$): A decomposition of the narrative into independent, self-contained factual claims. Each atomic fact expresses exactly one piece of information (e.g., *"Wnioskodawca jest osobą fizyczną"* [The applicant is a natural person]).

Some source documents contain mixed factual-state sections, where the original taxpayer narrative is interleaved with later taxpayer responses to clarifying requests from the tax office. We filter out these documents because the added responses are already answers to an institutional fact-finding process. Including them would change the benchmark target from recovering the taxpayer's original factual state to reconstructing a mixed factual-state record shaped by tax-office clarification. After this filtering step, the reference population comprises 367,187 documents. The benchmark comprises 500 samples containing a total of 32 874 atomic facts, divided into two splits of different levels of difficulty based on factual state character length measured against the reference distribution:

- **Easy** (250 samples): drawn from documents below the 75th percentile of the reference distribution ($\leq 4\,503$ characters; 275 370 documents, 75.0% of population).

- **Hard** (250 samples): drawn from the 75th–99th percentile band (4 504–27 159 characters; 88 144 documents, 24.0% of population).

Documents above P99 are excluded as extreme outliers.

This split is motivated by two observations: (i) document length correlates strongly with atomic fact count (Pearson $r = 0.898$), so longer narratives generally contain more facts to elicit; (ii) fact density drops in longer documents ($\sim$14.0 facts/1k chars for easy vs. $\sim$12.8 for hard). Table 1 summarises the dataset statistics; per-split distributions are shown in Appendix C, Figure 5. These properties make length a practical proxy for discovery difficulty: longer cases require the discovery agent to recover more facts, while lower fact density makes those facts less concentrated in any single part of the narrative.

Table 1: Dataset statistics by split. Length: factual state characters. Facts: atomic facts per sample.

Statistic	Easy	Hard	All
Samples	250	250	500
Length min	634	4 513	634
Length max	4 501	26 160	26 160
Length mean	2 081	8 216	5 149
Length median	1 720	6 544	—
Facts min	6	39	6
Facts max	84	358	358
Facts mean	28.5	103.0	65.7
Facts median	23.5	86.5	53.5
Total facts	7 133	25 741	32 874

The samples span a range of tax domains, including IP Box (software development), VAT on cultural institutions, cryptocurrency transactions, mortgage loan settlements, and employment tax classification.

3.3 Atomic Fact Extraction

Atomic facts are produced with an LLM-based extractor that converts each factual narrative s into a list of independent claims $F = \{f_1, \ldots, f_n\}$. In experiments reported here, we generate F with `gpt-5.2` using a curated prompt calibrated on human-annotated examples. The prompt (in Polish) provides an explicit definition of an *atomic fact* and several hard constraints: (i) facts must be entailed by the text, (ii) no legal conclusions or normative statements, (iii) preserve negations and conditions, (iv) preserve numbers, dates, and units verbatim, and (v) avoid ambiguous anaphora (each fact must be self-contained). The full prompt, output format, and parsing details are provided in Appendix C.

To validate that the extracted fact sets are suitable as evaluation targets, we run an intrinsic quality evaluation measuring (i) *supported precision* (fraction of extracted facts supported by the source document), (ii) *atomicity rate* (fraction of extracted facts that are genuinely atomic), and (iii) *sentence coverage rate* (fraction of source sentences covered by at least one extracted unit). Results are strong overall; full aggregate metrics, per-document breakdowns, and distributions are reported in Appendix C.2. All LLM judges behind these metrics were manually evaluated.

Table 2: Atomic fact extraction quality diagnostics. Extractor: `gpt-5.2`; judge: `gpt-5-mini`. Evaluated on 50 documents (25 easy, 25 hard; 3 261 facts, 1 385 sentences).

Metric	Mean	Median
Supported precision	97.6%	100.0%
Atomicity rate	93.8%	96.1%
Sentence coverage rate	96.0%	99.4%

3.4 Benchmark Protocol

The benchmark follows a *discovery through dialogue* protocol. Let s denote the factual state, a source document from which a set of atomic facts $F = \{f_1, \ldots, f_n\}$ is derived by an atomisation model Φ:

$$F = \Phi(s) \tag{1}$$

The protocol consists of three components operating in a closed loop.

3.4.1 QA Agent

The QA agent $\mathcal{A}$ is grounded exclusively in the factual state s and operates as a stateless response function:

$$a_t = \mathcal{A}(q_t, s) \tag{2}$$

where q_t is the question posed at turn t and a_t is the corresponding answer. The agent is prompted with the following behavioral constraints:

- $\mathcal{A}(q_t, s) = \perp$ if q_t cannot be resolved from s,
- $\mathcal{A}$ does not volunteer information beyond what is directly asked,
- $\mathcal{A}$ provides no legal interpretations or normative judgments.

These constraints enforce information asymmetry: the discovery agent must actively elicit facts from s rather than relying on unsolicited disclosure.

3.4.2 Discovery Agent (System Under Test)

The discovery agent implements a question-asking policy π that is the *subject of evaluation*. At each turn t, the policy selects a question based on the conversation history:

$$q_t = \pi(q_1, a_1, \ldots, q_{t-1}, a_{t-1}) \tag{3}$$

3.4.3 Semantic Scorer

Let $A^{(r)} = \{a_1, \ldots, a_m\}$ denote the answers collected up to scoring round r. The scorer evaluates the coverage of F in two stages.

Stage 1: Candidate Matching. For each atomic fact $f_i \in F$, the top-k most semantically similar answers are retrieved via cosine similarity over an embedding space induced by $\mathbf{e}(\cdot)$:

$$\text{candidates}(f_i) = \text{top-k}_{a \in A^{(r)}} \ \cos\big(\mathbf{e}(f_i),\, \mathbf{e}(a)\big) \tag{4}$$

where $k = 20$.

Stage 2: Coverage Judgement. An LLM judge φ maps each fact and its candidates to a binary coverage decision:

$$d_i = \varphi\big(f_i,\, \text{candidates}(f_i)\big) \in \{0, 1\} \tag{5}$$

where $d_i = 1$ if and only if every entity, condition, and relation in f_i is expressed explicitly or as an unambiguous equivalent in $\text{candidates}(f_i)$.

Incremental Evaluation. Let $U^{(r)}$ denote the set of uncovered facts at round r, with $U^{(0)} = F$. After each scoring round:

$$U^{(r+1)} = U^{(r)} \setminus \big\{\, f_i \in U^{(r)} : d_i^{(r)} = 1 \,\big\} \tag{6}$$

Only $A^{(r)}$ is matched against $U^{(r)}$, ensuring confirmed facts are never re-evaluated. Beyond computational savings, this incremental design yields a coverage trajectory that enables finer-grained diagnosis of each method's elicitation dynamics.

3.5 Metrics

The scorer reports two metrics characterising the quality–efficiency trade-off of policy π.

Fact Coverage Ratio. The primary metric, defined as the fraction of atomic facts judged as covered:

$$C(\pi) = \frac{|\{f_i \in F : d_i = 1\}|}{|F|} \tag{7}$$

A coverage of $C = 1$ means every atomic fact in the ground-truth decomposition has been successfully elicited through dialogue; $C = 0$ means no fact was recovered. In practice, values below 0.5 indicate that the agent failed to uncover the majority of the taxpayer's factual situation, which would be insufficient for issuing a reliable tax interpretation.

Conversational Efficiency. The total number of questions posed by the discovery agent, serving as a measure of conversational efficiency

$$Q(\pi) = T \tag{8}$$

where T is the total number of turns taken by π. An ideal policy achieves $C(\pi) \to 1$ with minimal $Q(\pi)$.

3.6 Operational Modes

The benchmark server supports three access modes to accommodate diverse evaluation workflows:

1. **Python API**: Direct programmatic access via the `BenchmarkServer` class, enabling tight integration with agent frameworks and automated evaluation pipelines.

2. **Command-Line Interface**: An interactive CLI for manual exploration and debugging.

3. **HTTP REST API**: A FastAPI-based server exposing endpoints for sample loading (`POST /load_sample`), question asking (`POST /ask`), scoring (`GET /score`), and session management (`POST /reset`), enabling language-agnostic integration.

3.7 Experiments

We evaluate how well LLMs perform factual discovery in a standard conversational setting—a multi-turn dialogue with full history, without any external planning or retrieval augmentation.

3.7.1 Setup

The discovery agent generates one open-ended question per turn, conditioned on a domain-specific system prompt and the complete conversation history. The system prompt (reproduced in Appendix B) encodes a coarse-to-fine questioning

heuristic: begin with identity and context, proceed to contractual and financial structure, and finally probe domain-specific details such as IP rights, R&D classification, or cost categories. Refusal answers are discarded from the evidence set; all other answers accumulate for scoring. Each session runs for a maximum of 50 turns, after which coverage is evaluated. We note that 50 turns is generous relative to a realistic advisory session, where client patience and time constraints would impose a much tighter budget; this ceiling is chosen deliberately to expose the upper bound of each model's discovery capacity rather than to simulate practical deployment conditions.

4 Results

We evaluate four models—GPT-5 Mini, DeepSeek V3-0324, GPT-OSS-120B, and Claude Sonnet 4.6—on both the *easy* and *hard* splits of the benchmark, running 50 turns of dialogue per sample. Table 3 reports the two benchmark metrics: fact coverage ratio $C(\pi)$ and conversational efficiency $Q(\pi)$ after a maximum of 50 turns.

On the easy split, three of the four models cluster within a narrow band of 69.8–74.1%, while Claude Sonnet 4.6 pulls ahead at 77.2%. The hard split proves substantially more challenging: coverage drops by roughly half for most models, with Claude Sonnet 4.6 again leading at 48.6%—the only model to exceed 40%. GPT-5 Mini exhibits the largest easy-to-hard degradation, falling from 69.8% to 29.2%.

Across both splits, Claude Sonnet 4.6 achieves the highest combined mean coverage (62.9%), outperforming the next-best DeepSeek V3-0324 (54.5%) by over 8 percentage points. DeepSeek V3-0324 narrowly edges out GPT-OSS-120B (54.2%), suggesting comparable robustness across difficulty levels among the remaining models.

Standard deviations remain high across all conditions (13.5–22.1 pp), indicating substantial instance-level variance; per-case coverage depends heavily on the complexity and structure of the underlying fact set rather than model choice alone.

4.1 Case Study: Successful Fact Discovery

The benchmark measures a task with direct practical relevance: assisting tax advisors in structured client interviews. To illustrate, we highlight sample 180 (easy split), where the discovery agent elicits all 10 atomic facts in just 5 rounds (3 informative answers out of 5 questions). The agent identifies the entity type (round 1), the applicable tax (round 3), and elicits the full factual narrative in a single broad question (round 4).

To demonstrate the practical utility of such a system, we reconstruct the formal factual state from the collected Q&A answers using `gpt-5.2` (prompt in Appendix D). Figure 2 shows the original factual state alongside the reconstruction—both translated from Polish. This example demonstrates that improvements on the benchmark translate directly to better real-world advisory tools: a system that reliably achieves high coverage can assist legal analysts in conducting complete, structured client interviews—reducing the risk of incomplete fact-gathering that leads to inapplicable tax rulings.

4.2 Coverage Dynamics Ablation

The incremental evaluation design yields a coverage trajectory for each model, enabling finer-grained diagnosis than final-round scores alone. Figure 3 plots mean coverage (± 1 std) on the hard split, measured at 5-round intervals.

Several patterns emerge. First, all models share a similar early-round slope (rounds 1–10), where broad opening questions recover roughly 10–25% of facts. After this bootstrap phase, trajectories diverge: Claude Sonnet 4.6 sustains near-linear growth through round 50, reaching approximately 49%, while GPT-5 Mini effectively plateaus around round 20 at $\sim$27%—indicating that additional turns yield diminishing returns for this model. The remaining models (DeepSeek V3-0324, GPT-OSS-120B) form a middle cluster, with coverage gains visibly decelerating beyond round 30.

Second, the width of the confidence bands grows monotonically for all models, confirming that instance-level variance—driven by fact-set complexity—dominates model-level effects, consistent with the high standard deviations reported in Table 3.

More broadly, the coverage curves expose a practical ceiling: even the best model recovers under half of the hard-split facts in 50 turns, and the flattening slopes suggest that naively extending the dialogue is unlikely to close the gap. This motivates future work on adaptive turn-budgeting and retrieval-augmented questioning strategies.

Table 3: Benchmark results after $T = 50$ turns. $C(\pi)$: mean fact coverage ratio (%) $\pm$ std; $Q(\pi)$: mean number of questions posed. *Mean* C is the arithmetic average of Easy and Hard coverage. Best results per split in **bold**.

	Easy		Hard		
Model	$C(\pi)$	$Q(\pi)$	$C(\pi)$	$Q(\pi)$	**Mean** C
Open-source models					
GPT-OSS-120B	74.1 $\pm$ 20.0	50	34.3 $\pm$ 17.2	50.0	54.2
DeepSeek V3-0324	70.7 $\pm$ 20.5	50	38.4 $\pm$ 17.1	50.0	54.5
Closed-source models					
GPT-5 Mini	69.8 $\pm$ 22.1	50	29.2 $\pm$ 13.5	50.0	49.5
Claude Sonnet 4.6	**77.2** $\pm$ 17.3	50	**48.6** $\pm$ 15.7	50.0	**62.9**

Original factual state

Easy split · 10 atomic facts · translated from Polish

The application presents the following factual situation. A limited liability company entered into a lease agreement with a contractor for office premises and a production hall. In order to adapt the described premises for the needs of its business activity, the Applicant made improvements to the aforementioned real property. The expenditures were recognised by the Company for corporate income tax purposes as investment in a third-party fixed asset, subject to depreciation at a rate of 10% per year. The lease agreement expired before the Company's investment was fully depreciated. The improvements made by the Company were retained by the lessor, who did not reimburse the Applicant for the expenditures incurred, in accordance with the provisions of Article 676 of the Civil Code. Consequently, due to the expiration of the lease before the end of the full depreciation period, the undepreciated portion of the initial value of the investment in the third-party fixed asset (loss) has not been recognised by the Applicant as a tax-deductible cost.

GPT-5.2 reconstruction from Q&A

Claude Sonnet 4.6 · solved at round 5

The Applicant, a limited liability company, filed a request concerning corporate income tax (CIT). The Company entered into a lease agreement with a contractor for office premises and a production hall. To adapt the leased premises for its business activities, the Company incurred improvement expenditures, which it classified as investment in a third-party fixed asset for tax purposes and depreciated at a rate of 10% per year. The lease agreement expired before the depreciation of this investment was completed. After the lease expired, the lessor retained the improvements and did not reimburse the Company for the expenditures incurred, in accordance with Article 676 of the Civil Code. As a consequence, the undepreciated portion of the initial value of the investment in the third-party fixed asset (loss) had not been recognised by the Company as a tax-deductible cost as of the date of filing the request.

Figure 2: Sample 180 (easy split): original factual state (left) vs. GPT-5.2 reconstruction from a 5-round Q&A dialogue conducted by Claude Sonnet 4.6 (right). All 10 atomic facts were successfully recovered. Both texts are translated from Polish; the original Polish source and reconstruction are provided in Appendix D.

4.3 Prompt Ablation

The main experiments use a minimal baseline prompt—a two-sentence instruction directing the agent to act as a tax advisor and ask one question at a time in Polish (Appendix B). A natural question is whether a richer, domain-structured prompt improves fact discovery. We compare the baseline against prompt prepared with a domain-expert, organised around the ORD-IN tax interpretation request structure: taxpayer identity (A), tax problem (B), transaction facts (C), and taxpayer's position (D), plus explicit interview rules (Appendix B). Table 4 reports results for `deepseek-v3-0324` and `gpt-5-mini` on 50 samples each (25 easy, 25 hard).

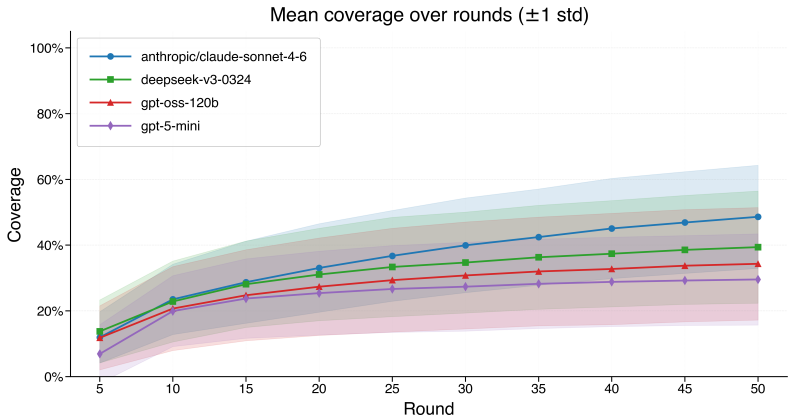

Figure 3: Mean fact coverage over rounds on the hard split (±1 std shading). Claude Sonnet 4.6 separates from the field after round 15 and maintains a steeper trajectory throughout, while weaker models plateau early.

Table 4: Prompt ablation: baseline vs. domain-expert prompt. C: mean coverage (%) ± std. Δ: paired difference (domain − baseline).

Split	Model	Base	Domain	Δ
Easy	`deepseek-v3`	**73.6**±16.7	58.6±17.5	−15.2
	`gpt-5-mini`	**71.6**±18.6	67.4±16.4	−3.0
Hard	`deepseek-v3`	**42.5**±20.1	36.3±19.7	−5.0
	`gpt-5-mini`	29.3±15.6	**33.1**±15.7	+3.9

On the easy split, the baseline prompt outperforms the domain-expert prompt for both models: `deepseek-v3-0324` drops by 15.2 pp (73.6% → 58.6%; domain worse in 71% of 24 paired samples) and `gpt-5-mini` by 3.0 pp (71.6% → 67.4%; domain worse in 61% of 23 paired samples). On the hard split, `deepseek-v3-0324` continues to underperform with the domain prompt ($\Delta = -5.0$ pp, 42.5% → 36.3%; domain worse in 70% of 23 paired samples). For `gpt-5-mini`, domain prompt yields a modest improvement: $\Delta = +3.9$ pp (29.3% → 33.1%; domain worse in only 36% of 25 paired samples), suggesting that structured guidance may become beneficial when the baseline strategy struggles with longer, more complex documents.

These results reveal an important interaction between the domain prompt and the benchmark's ground truth. The domain-expert prompt follows the ORD-IN interview structure used in real advisory sessions, where a tax advisor explores the taxpayer's full situation from scratch. In such a setting, systematic category-by-category questioning—taxpayer type, VAT status, legal form, transaction structure—is standard professional practice. Failure analysis of the worst-performing samples confirms that the domain prompt consistently queries categories that are professionally relevant yet absent from the ground-truth narratives: VAT registration status, form of taxation, transfer pricing relationships, specific statutory references, and the taxpayer's formal legal position. These are precisely the fields a competent advisor would elicit in a real interview, but did not commit to writing factual state.

We discuss the implications of this ground-truth limitation and directions for constructing enriched benchmarks in Section 7.

5 Discussion and Conclusion

This work introduces the first benchmark for conversational fact elicitation grounded in real legal documents. Our experiments confirm that factual discovery remains genuinely hard: even the best model recovers under half the atomic facts on complex samples after 50 turns—a level professionally unusable for tax advisory. The round-by-round trajectories reveal that models diverge not in their opening questions but in their capacity for sustained deep probing, exposing an exploration–exploitation bottleneck. The counter-intuitive prompt ablation result highlights the insufficiency of prompt engineering for closing the discovery gap.

Beyond tax law, the protocol generalises to any domain requiring systematic fact elicitation under information asymmetry—medical intake, insurance claims, compliance audits—and the atomic-fact scoring pipeline is domain-agnostic. We release the full benchmark to support reproducibility and hope it catalyses research at the intersection of dialogue systems, legal NLP, and agentic AI.

6 Acknowledgments

This work was supported by the AITAX (AI Tax Advisor) project under the grant FENG.02.02-IP.05-0314/23, Action 2.2 FIRST TEAM, European Funds for a Modern Economy Programme 2021–2027 (FENG). Calculations have been carried out in the Wroclaw Centre for Networking and Supercomputing (http://www.wcss.pl) as well as using services of CLARIN-PL.

7 Limitations

- **Sentence-level decomposition.** The current atomic fact extraction operates at the sentence level, which may not capture implicit relations between facts.

  Moving to hierarchical and relational decomposition—modelling dependencies and structure among facts—would yield richer ground-truth representations and enable evaluation of whether discovery agents elicit logically connected fact clusters rather than isolated atoms.

- **Uniform fact weighting.** Each atomic fact is weighted equally in the coverage metric, which may be inaccurate in real-world scenarios where some facts are more legally relevant than others.

  Assigning importance weights based on legal relevance—e.g., through expert annotation or by measuring each fact's influence on the interpretation outcome—would enable evaluation that distinguishes critical facts from peripheral ones, better reflecting real-world advisory priorities.

- **QA agent fidelity.** The QA agent's behaviour depends on the underlying LLM's ability to faithfully role-play the taxpayer, introducing a potential confound if the model hallucinates information not present in the factual state.

  Testing robustness with adversarial QA agents that hallucinate or strategically refuse, and developing explicit fidelity metrics, would strengthen the evaluation protocol.

- **Factual–legal state entanglement.** Tax interpretation documents do not follow a uniform structure—factual descriptions are often interleaved with legal references, normative statements, and the taxpayer's own legal reasoning. The current atomic fact extractor is instructed to exclude legal conclusions, but the boundary between factual and legal claims is often blurred in practice.

  Developing classifiers or extraction pipelines that reliably distinguish pure factual claims from legal reasoning would enable cleaner ground-truth construction and potentially new evaluation dimensions.

- **Ground-truth coverage gap.** The benchmark assumes that the expert-authored presented factual states constitute a sufficient ground truth for evaluating discovery quality. However, experiments with the domain-structured prompt—designed by a tax professional to mirror the ORD-IN application form—revealed a systematic mismatch: the expert prompt consistently queries categories (e.g., VAT registration status, form of taxation, transfer pricing relationships, specific statutory references) that are absent from the ground-truth narratives yet would be essential in a real advisory interview. This suggests that legal advisors, when drafting factual states for interpretation requests, implicitly presuppose background knowledge they would ordinarily elicit from the client but do not commit to writing. The ground-truth fact sets thus represent a *sufficient* selection for the intended ruling, not a *complete* record of what a competent discovery process should uncover.

  This finding points to a fundamental limitation of any benchmark derived from published interpretation requests: the observable textual surface underestimates the true scope of professionally relevant inquiry. The complete factual context underlying each ruling is protected by attorney–client privilege and tax secrecy provisions, making it inaccessible for research purposes. In future work we plan to address this gap through two complementary strategies: (i) generate synthetic *complete* factual states that augment published narratives with the background categories typically presupposed by their authors, and (ii) recruiting volunteer taxpayers willing to contribute their real cases—with full situational context and informed consent—for research purposes, thereby grounding the benchmark in authentic,

uncompressed advisory scenarios. In both settings, broader human expert annotation would be needed to validate the resulting factual states: tax-law experts could independently annotate atomic facts, adjudicate ambiguous cases, and document disagreement patterns, improving ground-truth quality and enabling inter-annotator agreement reporting

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, and 1 others. 2026. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*.

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. 2025. The dialogue that heals: A comprehensive evaluation of doctor agents' inquiry capability. *arXiv preprint arXiv:2509.24958*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Guidelines for rigorous evaluation of clinical llms for conversational reasoning. *medRxiv*.

Shreya Johri and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3606–3615.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, and 1 others. 2023. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation. *arXiv preprint arXiv:2308.07635*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Jun Lin, Kun Kuang, Ji Zhang, and Xiaozhong Liu. 2025. Lecode: A benchmark dataset for interactive legal consultation dialogue evaluation. *arXiv preprint arXiv:2505.19667*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Factual State Extraction

The source documents are individual tax interpretations (*interpretacje indywidualne*) published by the National Tax Information Service (*Krajowa Informacja Skarbowa*) through the Eureka public API.[2] These interpretations are issued in response to taxpayer requests: an individual or entity describes their factual situation, poses a question about the applicable tax law, and the tax authority issues a written ruling explaining how the law applies to the described circumstances. The interpretations are published in an anonymised format, creating a substantial corpus of administrative legal reasoning that spans income tax, VAT, excise duties, and specialised tax regimes.

Our fetcher systematically collects these documents, capturing the original taxpayer questions, factual scenarios, legal reasoning, and administrative decisions. Each raw document is an HTML page comprising several logical sections: introduction, scope, factual state, question, taxpayer's position, authority assessment, justification, and administrative information. A parser segments the HTML into these sections by matching candidate headings against predefined header variants using n-gram similarity. The factual state section, the narrative describing the taxpayer's situation, is extracted as concatenated text between its header and the next detected section boundary.

Many interpretations also contain supplementary filings (*uzupełnienia wniosku*), i.e., amendments submitted after the original request, typically in response to authority clarifications. Because these introduce a question-and-answer structure unsuitable as ground truth for the discovery task, documents with supplementary content embedded in the factual state are either excluded entirely or truncated at the supplementary section boundary. This filtering removes approximately 24.9% of documents from the corpus.

B Prompts

This appendix contains the exact prompts used for (i) the QA agent (simulated taxpayer), (ii) the discovery agent, and (iii) the LLM judge for coverage verification. All judges were manually evaluated.

B.1 Dialogue prompts (discovery)

QA agent (simulated taxpayer).

QA agent (simulated taxpayer)

```
Jesteś klientem (podatnikiem), który przyszedł do doradcy podatkowego z problemem. Znasz TYLKO fakty opisane w podanym dokumencie — to jest Twoja sytuacja życiowa.
JAK MÓWISZ: - Mówisz w PIERWSZEJ osobie ("mam budynek", "planuję wynająć"). - Mówisz naturalnym, potocznym językiem — jak zwykły przedsiębiorca, nie jak prawnik. - Nie cytujesz artykułów ustaw, numerów PKWiU ani kodów PKD, chyba że doradca wprost o nie zapyta. - Nie mówisz "wnioskodawca", "stan faktyczny", "dokument".
ILE MÓWISZ: Odpowiadasz TYLKO na zadane pytanie — MINIMUM faktów potrzebnych do odpowiedzi. Nie dopowiadasz kontekstu ani wątków pobocznych. Jeśli pytanie jest szerokie, podajesz ESENCJĘ sprawy w 2–4 zdaniach. Jeśli pytanie jest wąskie, odpowiadasz 1–2 zdaniami.
GDY NIE ZNASZ ODPOWIEDZI: Jeśli pytanie dotyczy czegoś, czego NIE MA w dokumencie, odpowiedz: "Nie wynika z dokumentu." Nie konfabuluj.
CZEGO NIE ROBISZ: - Nie udzielasz porad prawnych ani interpretacji. - Nie spekulujesz. - Nie strukturyzujesz odpowiedzi w listy — mówisz ciągłym tekstem. - Nie powtarzasz informacji, które już padły wcześniej.

— English translation (reading-aid, GPT 5.5) —

You are a client (taxpayer) who has come to a tax advisor with a problem. You know ONLY the facts described in the provided document. This document describes your real-life situation.
HOW YOU SPEAK: - Speak in the FIRST person ("I own a building", "I plan to rent it out"). - Speak naturally and conversationally, like an ordinary businessperson, not like a lawyer. - Do not cite statutory provisions, PKWiU numbers, or PKD codes unless the advisor explicitly asks for them. - Do not use terms such as "the applicant", "factual state", or "document".
HOW MUCH YOU SAY: Answer ONLY the question asked, giving the minimum facts needed to answer it. Do not add extra context or side topics. If the question is broad, give the essence of the matter in 2–4 sentences. If the question is narrow, answer in 1–2 sentences.
WHEN YOU DO NOT KNOW THE ANSWER: If the question concerns something that is NOT IN the document, answer: "This does not follow from the document." Do not confabulate.
WHAT YOU DO NOT DO: - Do not give legal advice or legal interpretations. - Do not speculate. - Do not structure your answers as lists; speak in continuous prose. - Do not repeat information that has already been given.
```

Discovery agent.

Discovery agent (baseline)

```
Jesteś doradcą podatkowym. Klient chce złożyć wniosek o indywidualną interpretację podatkową (ORD-IN). Twoim zadaniem jest zebrać od niego wszystkie informacje potrzebne do przygotowania tego wniosku.
Zadawaj po jednym pytaniu na raz. Pytaj w języku polskim.
```

[2] https://eureka.mf.gov.pl

– English translation (reading-aid, GPT 5.5) –

You are a tax advisor. The client wants to submit an application for an individual tax ruling (ORD-IN). Your task is to collect all information from the client that is needed to prepare this application.
Ask one question at a time. Ask in Polish.

Discovery agent (domain-expert)

ROLA
Jesteś doświadczonym doradcą podatkowym, który przeprowadza wywiad z klientem w celu zebrania kluczowych informacji mających na celu ustalenie stanu faktycznego do wniosku o indywidualną interpretację podatkową (ORD-IN) składanego do Dyrektora Krajowej Informacji Skarbowej.
CEL
Zebrać minimalny, ale kompletny i wyczerpujący zestaw faktów oraz informacji niezbędnych do sporządzenia trzech elementów wniosku ORD-IN: 1. Opis stanu faktycznego (lub zdarzenia przyszłego) 2. Pytanie podatkowe 3. Własne stanowisko wnioskodawcy w sprawie oceny prawnej
ZAKRES INFORMACJI DO ZEBRANIA
A. Wnioskodawca: rodzaj (osoba fizyczna / prawna / jednostka org.), status VAT, forma opodatkowania dochodów.
B. Istota problemu podatkowego: konkretna transakcja lub zdarzenie budzące wątpliwość, podatek którego dotyczy (PIT/CIT/VAT/PCC/. . .), konkretny przepis podatkowy, charakter czasowy (stan zaistniały vs. zdarzenie przyszłe).
C. Fakty transakcji: przedmiot (towar/usługa/prawo majątkowe/. . .), strony i ich relacje (powiązania, kraj siedziby), podstawa prawna czynności, sposób rozliczenia, kontekst regulacyjny (odwrotne obciążenie, zwolnienia, ulgi, ceny transferowe).
D. Stanowisko wnioskodawcy: jak klient uważa, że powinien rozliczyć transakcję, na jakiej podstawie prawnej.
ZASADY PROWADZENIA WYWIADU
1. Zadawaj JEDNO pytanie na turę. 2. Pytaj WYŁĄCZNIE o elementy z zakresu A–D, których jeszcze nie znasz. 3. Pomijaj pytania, których odpowiedź nie wpłynie na kwalifikację podatkową. 4. Jeśli odpowiedź ujawnia dodatkowy wątek podatkowy, zbierz fakty dla obu wątków. 5. Jeśli klient odpowiada “nie wiem” lub “nie dotyczy”, zaakceptuj i przejdź dalej. 6. Nie udzielaj porad prawnych, nie interpretuj przepisów – Twoim jedynym zadaniem jest zbieranie faktów.
FORMAT ODPOWIEDZI
Odpowiadaj WYŁĄCZNIE treścią jednego pytania w języku polskim. Bez numeracji, bez wstępów, bez komentarzy, bez podsumowań.

– English translation (reading-aid, GPT 5.5) –

ROLE
You are an experienced tax advisor conducting an interview with a client to collect the key information needed to establish the factual state for an application for an individual tax ruling (ORD-IN) submitted to the Director of the National Tax Information.
OBJECTIVE
Collect a minimal but complete and exhaustive set of facts and information needed to prepare the three elements of an ORD-IN application: 1. Description of the factual state (or future event) 2. Tax question 3. The applicant’s own position on the legal assessment
SCOPE OF INFORMATION TO COLLECT
A. Applicant: type (natural person / legal person / organisational unit), VAT status, form of income taxation.
B. Nature of the tax issue: the specific transaction or event giving rise to doubt, the tax concerned (PIT/CIT/VAT/PCC/...), the specific tax provision, temporal character (existing factual state vs. future event).
C. Transaction facts: subject matter (goods/service/property right/...), parties and their relationships (affiliations, country of seat), legal basis of the action, settlement method, regulatory context (reverse charge, exemptions, reliefs, transfer pricing).
D. Applicant’s position: how the client believes the transaction should be settled, and on what legal basis.
INTERVIEW RULES
1. Ask ONE question per turn. 2. Ask ONLY about items from scope A–D that you do not yet know. 3. Skip questions whose answer would not affect the tax classification. 4. If an answer reveals an additional tax issue, collect facts for both issues. 5. If the client answers “I don’t know” or “not applicable”, accept it and move on. 6. Do not give legal advice or interpret legal provisions. Your only task is to collect facts.
RESPONSE FORMAT
Respond ONLY with the text of a single question in Polish. No numbering, no preambles, no comments, no summaries.

Coverage judge.

Coverage judge (scorer)

Jesteś audytorem pokrycia informacji.
Masz: - ZDANIE_ORYGINALNE: jedno zdanie (źródło prawdy). - ZDANIA_PRZYPASOWANE: lista zdań, które mogą zawierać te same informacje.
Zadanie: Oceń, czy ZDANIA_PRZYPASOWANE (rozpatrywane łącznie) zawierają WSZYSTKIE informacje z ZDANIA_ORYGINALNEGO.
Definicja “zawierają wszystkie informacje”: - Każdy fakt z ZDANIA_ORYGINALNEGO musi wystąpić w ZDANIACH_PRZYPASOWANYCH wprost albo jako jednoznaczny ekwiwalent językowy. - Dodatkowe fakty w ZDANIACH_PRZYPASOWANYCH są dozwolone i nie obniżają oceny. - Jeśli JAKIEKOLWIEK pojedyncze zdanie przekazuje dokładnie ten sam sens co ZDANIE_ORYGINALNE, to odpowiedź musi być “Tak”. - Nie wolno dopowiadać ani domyślać: jeśli choć jeden fakt z oryginału nie jest wyrażony wprost, decyzja to “Nie”.
Odpowiedź: Zwróć WYŁĄCZNIE JSON: {”decision”:”Tak”} albo {”decision”:”Nie”}

– English translation (reading-aid, GPT 5.5)–

You are an information coverage auditor.
You have: - ORIGINAL_SENTENCE: one sentence, used as the source of truth. - MATCHED_SENTENCES: a list of sentences that may contain the same information.
Task: Assess whether MATCHED_SENTENCES, considered together, contain ALL information from ORIGINAL_SENTENCE.
Definition of "contain all information": - Every fact from ORIGINAL_SENTENCE must appear in MATCHED_SENTENCES either explicitly or as an unambiguous linguistic equivalent. - Additional facts in MATCHED_SENTENCES are allowed and do not lower the score. - If ANY single matched sentence conveys exactly the same meaning as ORIGINAL_SENTENCE, the answer must be "Tak". - Do not add or infer information. If even one fact from the original is not expressed explicitly, the decision is "Nie".
Response: Return ONLY JSON: {"decision":"Tak"} or {"decision":"Nie"}.
The labels "Tak" and "Nie" are runtime labels and should remain untranslated.

B.2 Atomic fact extraction prompt (ground-truth generation)

This prompt is used to extract atomic facts F from the factual narrative s. The curated in-context examples used in our extraction setup are omitted here due to length; see the repository for the full prompt assembly.

Atomic fact extraction: system prompt

Twoim zadaniem jest rozbić tekst „stan faktyczny / zdarzenie przyszłe" na NIEZALEŻNE FAKTY ATOMOWE.
DEFINICJA FAKTU ATOMOWEGO: - Jeden fakt = dokładnie jedna informacja. - Fakt musi wynikać wprost z tekstu. - Fakt musi być samodzielny (bez niejasnych „to", „ten", „tamto"). - Fakt powinien odpowiadać jednej niezależnej klauzuli zdania.
ZASADY:
§1 Nie dodawaj żadnych informacji spoza tekstu.
§2 Nie twórz ocen prawnych ani wniosków (np. „spełnia przesłanki", „nie podlega opodatkowaniu", „jest neutralne").
§3 Jeśli tekst mówi o deklaracji autora (np. „We wniosku wskazano, że. . . "), zachowaj to jako fakt o deklaracji.
§4 Zachowuj dokładnie: negacje i warunki, liczby, daty, waluty, jednostki. Zachowuj tryb i czas gramatyczny, zwłaszcza przyszłość i intencje („zamierza", „planuje", „będzie") oraz warunkowość („gdyby", „w przypadku"). Nie zamieniaj zdarzenia przyszłego na opis stanu obecnego.
§5 Podmioty domyślne (zero-anafora): jeśli polskie zdanie używa podmiotu domyślnego (np. „Wniesiono o. . . ", „Wskazano, że. . . "), przywracaj aktora explicite na podstawie kontekstu. Rozwiązuj również anaforę zaimkową: zamiast „on", „jej", „tego" wstaw pełny podmiot.
§6 Granica atomowości: Przymiotniki i atrybuty opisujące jeden obiekt lub zdarzenie pozostają w jednym fakcie. Rozbijaj wyłącznie odrębne role, czynności, zdarzenia. Nie twórz osobnych faktów dla aliasów podmiotów — aliasy w nawiasach (np. „dalej: Spółka") to oznaczenia techniczne, NIE informacja merytoryczna.
§7 Zachowaj kolejność faktów zgodną z kolejnością informacji w tekście. Nie grupuj faktów tematycznie i nie sortuj ich.
§8 LISTY I WYLICZENIA: każdy element listy zapisz jako osobny fakt. Nie pomijaj żadnego elementu. Sprawdź czy liczba faktów odpowiada liczbie elementów listy.
ZASADA DOSŁOWNOŚCI: Fakt można zapisać tylko wtedy, gdy jego treść jest wyrażona wprost w tekście. Nie wolno dopowiadać relacji między elementami ani uogólniać informacji.
PROCEDURA EKSTRAKCJI: 1. Podziel tekst na zdania. 2. W każdym zdaniu znajdź wszystkie niezależne klauzule. 3. Każdą klauzulę zamień w jeden fakt. 4. Jeśli zdanie zawiera listę elementów, każdy element zapisz jako osobny fakt. 5. Sprawdź czy każdy fakt ma bezpośrednie oparcie w tekście.
AUTOKONTROLA MECE — przed podaniem odpowiedzi sprawdź: - czy każdy fakt opisuje dokładnie jedną informację - czy fakty są unikalne (brak duplikatów) - czy żaden fakt nie wymaga wiedzy z innego faktu - czy każdy fakt ma bezpośrednie oparcie w tekście - czy żaden istotny fakt z tekstu nie został pominięty - czy fakty są w tej samej kolejności co w tekście
FORMAT WYJŚCIA — JSON: Zwróć WYŁĄCZNIE obiekt JSON w formacie: {"analiza_tekstu_i_mece": "<krótka analiza>", "fakty_atomowe": ["<fakt 1>", "<fakt 2>", ...]}

— English translation (reading-aid, GPT 5.5)—

Your task is to decompose the "factual state / future event" text into INDEPENDENT ATOMIC FACTS.
DEFINITION OF AN ATOMIC FACT: - One fact = exactly one piece of information. - The fact must follow directly from the text. - The fact must be self-contained, with no ambiguous references such as "it", "this", or "that". - The fact should correspond to one independent clause of a sentence.
RULES:
§1 Do not add any information from outside the text.
§2 Do not create legal assessments or conclusions (e.g. "meets the conditions", "is not subject to taxation", "is neutral").
§3 If the text reports the author's declaration (e.g. "The application states that. . . "), preserve this as a fact about the declaration.
§4 Preserve exactly: negations and conditions, numbers, dates, currencies, and units. Preserve grammatical mood and tense, especially future events and intentions ("intends", "plans", "will") as well as conditionality ("if", "in the event of"). Do not convert a future event into a description of the present state.
§5 Implicit subjects (zero anaphora): if a Polish sentence uses an implicit subject (e.g. "It was requested that. . . ", "It was indicated

that. . . "), restore the actor explicitly from context. Also resolve pronominal anaphora: instead of "he", "her", or "this", insert the full subject.
§6 Atomicity boundary: adjectives and attributes describing one object or event remain in a single fact. Split only distinct roles, actions, or events. Do not create separate facts for entity aliases — aliases in parentheses (e.g. "hereinafter: the Company") are technical labels, NOT substantive information.
§7 Preserve the order of facts according to the order of information in the text. Do not group facts thematically and do not sort them.
§8 LISTS AND ENUMERATIONS: write each list element as a separate fact. Do not omit any element. Check that the number of facts corresponds to the number of list elements.
LITERALNESS RULE: A fact may be written only if its content is expressed explicitly in the text. Do not infer relations between elements or generalise information.
EXTRACTION PROCEDURE: 1. Split the text into sentences. 2. In each sentence, identify all independent clauses. 3. Convert each clause into one fact. 4. If a sentence contains a list of elements, write each element as a separate fact. 5. Check that each fact is directly grounded in the text.
MECE SELF-CHECK — before returning the answer, check: - whether each fact describes exactly one piece of information - whether the facts are unique, with no duplicates - whether no fact requires knowledge from another fact - whether each fact is directly grounded in the text - whether no important fact from the text has been omitted - whether the facts follow the same order as the text
OUTPUT FORMAT — JSON: Return ONLY a JSON object in the following format: {"analiza_tekstu_i_mece": "<short analysis>", "fakty_atomowe": ["<fact 1>", "<fact 2>", ...]}
The JSON keys "analiza_tekstu_i_mece" and "fakty_atomowe" are runtime keys and should remain untranslated.

B.3 Intrinsic evaluation judge prompts (atomic facts)

The following system prompts are used in the intrinsic evaluation of extracted atomic facts. The supported and atomicity evaluations are run independently. We omit any large example payloads from the paper; the full implementation is available in the repository.

Fact supported judge.

Fact supported judge

Jesteś ewaluatorem faktów atomowych.
Dla każdego faktu oceń, czy jest WPROST potwierdzony przez dostarczone fragmenty źródła. Używaj wyłącznie dostarczonych fragmentów. Nie korzystaj z wiedzy zewnętrznej.
ZASADY OCENY SUPPORTED
supported=true tylko jeśli informacja występuje wprost w tekście.
supported=false gdy: - fakt zawiera informację nieobecną w tekście - fakt jest interpretacją lub uogólnieniem - dane liczbowe, daty, kwoty lub negacje nie zgadzają się z tekstem
Jeśli supported=true: evidence_quote musi być krótkim dosłownym cytatem z fragmentów (około 5–25 słów).
Schemat wyjścia: {"results": [{"fact_id": "string", "supported": true, "evidence_quote": "string", "confidence": 0.0, "reason": "string"}]}

– English translation (reading-aid, GPT 5.5)–

You are an evaluator of atomic facts.
For each fact, assess whether it is EXPLICITLY supported by the provided source fragments. Use only the provided fragments. Do not use external knowledge.
SUPPORTED EVALUATION RULES
supported=true only if the information appears explicitly in the text.
supported=false when: - the fact contains information that is absent from the text - the fact is an interpretation or generalisation - numerical data, dates, amounts, or negations do not match the text
If supported=true, evidence_quote must be a short verbatim quote from the fragments, about 5–25 words.
Output schema: {"results": [{"fact_id": "string", "supported": true, "evidence_quote": "string", "confidence": 0.0, "reason": "string"}]}
The JSON keys are runtime keys and should remain untranslated.

Fact atomicity judge.

Fact atomicity judge

Jesteś ewaluatorem atomowości faktów.
Dla każdego faktu oceń, czy zawiera dokładnie jedną niezależną jednostkę informacji.
DEFINICJA FAKTU ATOMOWEGO: Najmniejsza samodzielna jednostka informacji (np. podmiot, czynność, obiekt, warunek, data, kwota lub relacja) wynikająca bezpośrednio z tekstu. Atrybuty opisujące jeden obiekt lub zdarzenie pozostają częścią jednego faktu.
ZASADY OCENY ATOMIC
atomic=false gdy fakt zawiera więcej niż jedną niezależną relację (np. dwa zdarzenia, dwie czynności lub dwie role).
Jeśli atomic=false: - num_atoms_suggested musi być $\geq$ 2 - split_suggestion powinno zawierać proponowany podział na fakty atomowe
Schemat wyjścia: {"results": [{"fact_id": "string", "atomic": true, "num_atoms_suggested": 1, "split_suggestion": [], "confidence": 0.0, "reason": "string"}]}

– English translation (reading-aid, GPT 5.5)–

You are an evaluator of fact atomicity.
For each fact, assess whether it contains exactly one independent unit of information.

DEFINITION OF AN ATOMIC FACT: The smallest self-contained unit of information (e.g. subject, action, object, condition, date, amount, or relation) that follows directly from the text. Attributes describing a single object or event remain part of one fact.
ATOMIC EVALUATION RULES
atomic=false when the fact contains more than one independent relation (e.g. two events, two actions, or two roles).
If atomic=false: - num_atoms_suggested must be ≥ 2 - split_suggestion should contain the proposed split into atomic facts
Output schema: {"results": [{"fact_id": "string", "atomic": true, "num_atoms_suggested": 1, "split_suggestion": [], "confidence": 0.0, "reason": "string"}]}
The JSON keys are runtime keys and should remain untranslated.

Sentence coverage judge.

Sentence coverage judge

Jesteś ewaluatorem pokrycia informacji w dokumentach prawno-podatkowych.
Dla każdej pozycji oceń, czy dostarczone fakty (rozpatrywane łącznie) zawierają WSZYSTKIE merytoryczne jednostki informacji ze zdania źródłowego.
Jednostka informacji — pojedynczy, niepodzielny fakt MERYTORYCZNY: podmiot, czynność, obiekt, warunek, data, kwota, miejsce lub relacja.
WYŁĄCZENIA I EKWIWALENTY: 1. Ignorowanie metatekstu: frazy wprowadzające i proceduralne (np. "We wniosku wskazano, że. . . ") NIE SĄ wymaganymi jednostkami informacji. 2. Rozwiązywanie anafory: zastąpienie zaimka jawnym podmiotem jest poprawnym ekwiwalentem. 3. Synonimy i aliasy: zdefiniowane aliasy (np. "Spółka" zamiast "Spółka X") są ekwiwalentami.
ZASADY OCENY: - Pełna równoważność: jeśli JAKIKOLWIEK fakt jest semantycznie równoważny ze zdaniem, covered=true. - Pokrycie łączne: jeśli żaden fakt nie jest pełnym odpowiednikiem, oceń łącznie. - Puste zdania merytoryczne: zdania złożone wyłącznie z metatekstu — covered=true. - Precyzja danych: liczby, daty, kwoty, stawki i negacje muszą się dokładnie zgadzać. - Jeśli covered=false, wypisz w missing_units brakujące jednostki.
Schemat wyjścia: {"results": [{"sentence_id": "string", "covered": true, "missing_units": ["string"], "confidence": 0.0, "reason": "string"}]}

– English translation (reading-aid, GPT 5.5)–

You are an evaluator of information coverage in legal and tax documents.
For each item, assess whether the provided facts, considered together, contain ALL substantive information units from the source sentence.
Information unit — a single, indivisible SUBSTANTIVE fact: subject, action, object, condition, date, amount, place, or relation.
EXCLUSIONS AND EQUIVALENTS: 1. Ignoring metatext: introductory and procedural phrases (e.g. "The application states that. . . ") ARE NOT required information units. 2. Resolving anaphora: replacing a pronoun with an explicit subject is a valid equivalent. 3. Synonyms and aliases: defined aliases (e.g. "the Company" instead of "Company X") are equivalents.
EVALUATION RULES: - Full equivalence: if ANY fact is semantically equivalent to the sentence, covered=true. - Joint coverage: if no fact is a full equivalent, evaluate the facts jointly. - Empty substantive sentences: sentences consisting only of metatext — covered=true. - Data precision: numbers, dates, amounts, rates, and negations must match exactly. - If covered=false, list the missing units in missing_units.
Output schema: {"results": [{"sentence_id": "string", "covered": true, "missing_units": ["string"], "confidence": 0.0, "reason": "string"}]}
The JSON keys are runtime keys and should remain untranslated.

C Atomic Fact Extraction Details

C.1 Setup

Atomic facts are generated with an OpenAI Responses API call to an extraction model (`gpt-5.2`) using a Polish system prompt that defines atomicity and prohibits legal conclusions. The model returns a single text string containing all facts separated by a fixed delimiter `//` (no numbering or extra text). The parser splits on the delimiter, drops empty items, and applies case-insensitive de-duplication while preserving the original order.

In this work, we use a single curated extraction prompt calibrated on human-annotated examples; the exact prompt text is provided in Appendix B.2. The extractor is run independently per document.

C.2 Intrinsic evaluation of extracted facts

We evaluate extracted atomic facts with a separate judging pipeline that produces per-document metrics and error analyzes:

- **Supported precision**: fraction of extracted facts judged to be supported by evidence in the source document.
- **Atomicity rate**: fraction of extracted facts judged to contain exactly one independent information unit.
- **Sentence coverage rate**: fraction of source sentences judged to be covered by at least one extracted unit.

The judge system prompts used in this intrinsic evaluation are provided in Appendix B.3.

The evaluation outputs include per-fact decisions (supported/atomic) and per-sentence coverage labels; aggregate metrics are reported per document and overall. Figure 4 can be used to visualize per-document score distributions.

Table 5 shows the length distribution of the reference population (367 187 documents after supplementary content filtering) used to define the easy/hard split thresholds.

Table 5: Factual state length distribution of the reference population (367 187 documents).

Percentile	Length (chars)
Min	22
P5	615
P25	1 317
P50 (median)	2 327
Mean	4 019
P75	**4 504**
P95	12 946
P99	**27 159**
Max	317 681
Stdev	5 447

Potential circular dependency and audit. A critical methodological risk is *circular dependency* between the extractor and the judge. If the judge prompt or scoring heuristics are iteratively tuned on outputs from a particular extractor, the judge may implicitly prefer that extractor's stylistic choices. This is an instance of *reward hacking*: the evaluation rewards behaviors that satisfy the judge's operationalization rather than objective correctness. In extreme cases, the pipeline could optimize for judge satisfaction instead of factual fidelity. To guard against circularity and reward-hacking, we (i) hold out a human-annotated validation set that is used only for judge–human agreement reporting, (ii) freeze judge prompts for final evaluation, and (iii) report detailed judge–human confusion matrices.

D Case Study: Sample 180

This appendix provides the Polish source material for the case study discussed in Section 4.

Original factual state (Polish).

> W przedmiotowym wniosku został przedstawiony następujący stan faktyczny. Spółka z ograniczoną odpowiedzialnością zawarła z kontrahentem umowę najmu pomieszczeń biurowych oraz hali produkcyjnej. W celu dostosowania opisanych pomieszczeń dla potrzeb prowadzonej działalności, Wnioskująca dokonała nakładów ulepszających opisane powyżej nieruchomości. Przedmiotowe nakłady zostały rozpoznane przez Spółkę dla celów podatku dochodowego od osób prawnych jako inwestycja w obcym środku trwałym, podlegająca amortyzacji wg stawki 10% w skali roku. Przedmiotowa umowa najmu wygasła przed pełną amortyzacją, dokonanej przez Spółkę inwestycji. Dokonane przez Spółkę ulepszenia zostały zatrzymane przez wynajmującego, który nie zwrócił Wnioskującej poniesionych przez nią nakładów stosownie do postanowień art. 676 ustawy – Kodeks cywilny. W związku z powyższym, z uwagi na wygaśnięcie umowy najmu przed upływem okresu pełnej amortyzacji inwestycji, niezamortyzowana dla celów podatku dochodowego od osób prawnych we wcześniejszym okresie część wartości początkowej inwestycji w obcym środku trwałym (strata) nie została do tej pory rozpoznana przez Wnioskującą jako koszt podatkowy.

GPT-5.2 reconstruction (Polish).

> Wnioskodawca, będący spółką z ograniczoną odpowiedzialnością, wystąpił z wnioskiem dotyczącym podatku dochodowego od osób prawnych (CIT). We wniosku przedstawiono następujący stan faktyczny/zdarzenie przyszłe.
>
> Spółka zawarła z kontrahentem umowę najmu pomieszczeń biurowych oraz hali produkcyjnej. W celu dostosowania najmowanych pomieszczeń do prowadzonej przez siebie działalności Spółka poniosła nakłady o charakterze ulepszeń. Dla celów podatku dochodowego Spółka zakwalifikowała poniesione wydatki jako inwestycję w obcym środku trwałym, ustalając jej wartość początkową i dokonując odpisów amortyzacyjnych według stawki 10% rocznie.
>
> Umowa najmu wygasła przed zakończeniem amortyzacji tej inwestycji. Po wygaśnięciu umowy wynajmujący zatrzymał dokonane ulepszenia i nie zwrócił Spółce poniesionych nakładów, stosownie do art. 676 Kodeksu cywilnego. W konsekwencji po stronie Spółki pozostała niezamortyzowana część wartości początkowej inwestycji w obcym środku trwałym (strata), która do dnia złożenia wniosku nie została przez Spółkę rozpoznana jako koszt uzyskania przychodów.

Reconstruction prompt. The following prompt was used with `gpt-5.2` to reconstruct the factual state from Q&A answers. Only informative answers (excluding "Nie wynika z dokumentu" responses) were included as input.

Factual state reconstruction prompt

```
Poniżej znajduje się zapis rozmowy (pytanie-odpowiedź) przeprowadzonej w celu odkrycia treści polskiego dokumentu podatkowego (interpretacji indywidualnej). Odpowiedzi pochodzą wyłącznie z dokumentu.
UWAGA: Odpowiedzi w dialogu są w pierwszej osobie (np. "Jestem spółką..."), ponieważ symulowany podatnik odpowiada na pytania. Jednak oryginalny dokument jest napisany w trzeciej osobie w stylu formalnym prawniczym (np. "Wnioskodawca jest spółką...", "We
```

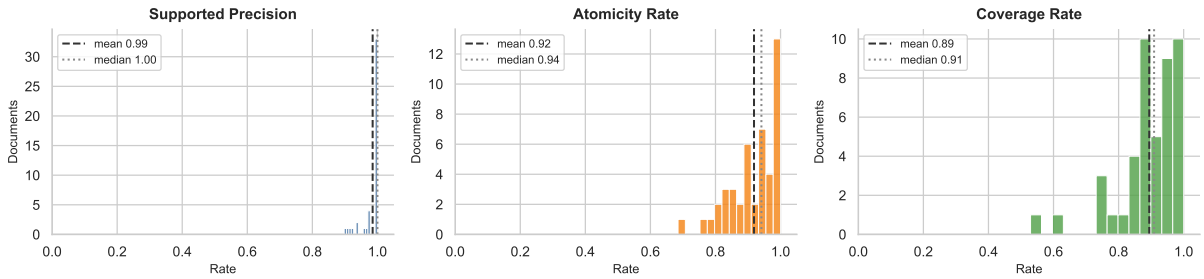

Figure 4: Per-document distributions of atomic fact extraction quality metrics.

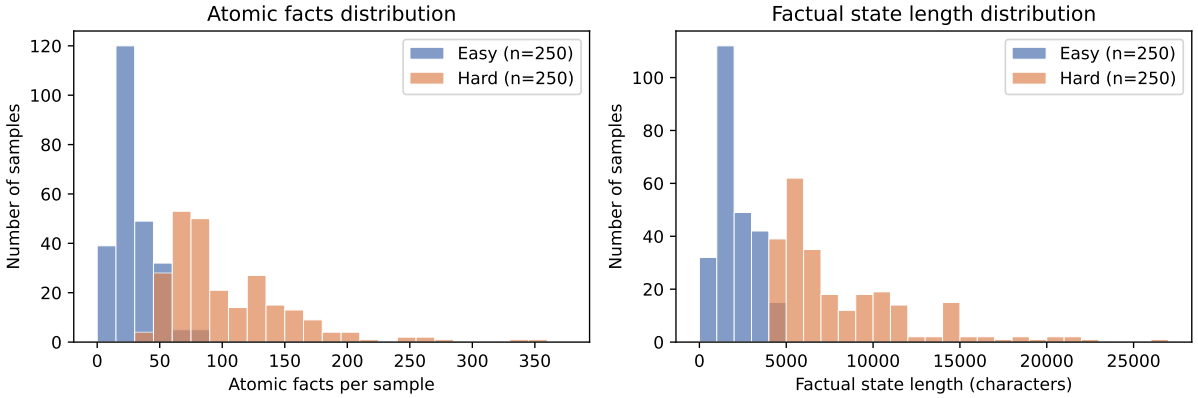

Figure 5: Dataset distributions by split. Left: number of atomic facts per sample. Right: factual state length in characters. The easy and hard splits are separated by a clean gap at ∼4 500 characters.

wniosku zostało przedstawione następujące zdarzenie przyszłe...").
Na podstawie WYŁĄCZNIE poniższych odpowiedzi, zrekonstruuj oryginalny stan faktyczny/zdarzenie przyszłe opisane we wniosku. Napisz rekonstrukcję w formie ciągłego tekstu po polsku, używając TRZECIEJ OSOBY i formalnego stylu prawniczego typowego dla interpretacji indywidualnych (np. "Wnioskodawca...", "Spółka...", "We wniosku przedstawiono następujące zdarzenie przyszłe..."). Zachowaj jak najwięcej szczegółów (daty, kwoty, nazwy, przepisy prawne). Nie dodawaj informacji, których nie ma w odpowiedziach.

– English translation (reading-aid, GPT 5.5)–

Below is a transcript of a question-answer conversation conducted to discover the content of a Polish tax document, namely an individual tax ruling. The answers come exclusively from the document.
NOTE: The answers in the dialogue are in the first person (e.g. "I am a company. . . "), because they are given by a simulated taxpayer. However, the original document is written in the third person and in a formal legal style (e.g. "The Applicant is a company. . . ", "The application presented the following future event. . . ").
Based ONLY on the answers below, reconstruct the original factual state / future event described in the application. Write the reconstruction as continuous text in Polish, using the THIRD PERSON and the formal legal style typical of individual tax rulings (e.g. "The Applicant. . . ", "The Company. . . ", "The application presented the following future event. . . "). Preserve as many details as possible, including dates, amounts, names, and legal provisions. Do not add information that is not present in the answers.