

AI-Enabled Access to Justice: Adaptive Questioning for Personalized Online Dispute Resolution

MARTIN LOEBL, Charles University, Czech Republic
ZBYNĚK LOEBL, Algostratus, Czech Republic
QUANYAN ZHU, New York University, USA

Large language models (LLMs) could help people without lawyers state their legal claims, but a system that rephrases questions or drafts filings must never change their legal meaning. For vulnerable persons, access to justice through online dispute resolution (ODR) remains an unsolved problem, because no single online court can afford to accommodate every disability, language barrier, or cognitive limitation. We propose to invert the problem: the person brings tools personalized to her needs, and an *open ODR standard* of numbered claim elements lets any court that adopts it read the result as if it had been filed on its own platform. Accommodations provided by each court thus become *portable* accommodations that the person carries from court to court, at the stage where many people drop out: turning a grievance into a claim.

We design an AI-enabled system for question creation, adaptive elicitation, filing assembly, and adaptation. Its principle is that *the LLM proposes and the standard disposes*: every generated question and every generated filing must map back to the numbered elements of the standard. At its core is adaptive questioning: eliciting what a person intends through short questions answered with *yes*, *no*, or *I do not care*, with occasional response errors. We prove that, given the same confirmed answers, the questions asked and the filing assembled do not depend on how the questions are worded or rendered, define a *tension score* for inconsistent answers, bound the number of questions needed, give a deterministic, published elicitation algorithm that chooses certified questions by value per unit of attention, with explicit definitions of question complexity, risk, and value, a ceiling on the value of one answer, and a stopping rule, generate questions from the applicable law, and accompany each filing with raw data on errors and incomplete answers, with a separate probabilistic bound, for certification, that separates elicitation errors from LLM errors. In simulation, adaptive questioning matches the accuracy of a fixed form with about a third fewer questions; flagging uncertain elements for read-back cuts the share of filings with an unseen error from 16% to 5%; and learning each person's error rate during the session nearly halves the accuracy gap between simulated higher-error and typical claimants. We state design requirements grounded in procedural justice, foremost that a measured inconsistency must never become a statement about a person's credibility, and identify the legal questions the system raises.

CCS Concepts: • **Applied computing** → Law; • **Theory of computation** → Active learning; Algorithmic game theory and mechanism design; • **Human-centered computing** → Accessibility; • **Computing methodologies** → Natural language processing.

Additional Key Words and Phrases: artificial intelligence, large language models, access to justice, online dispute resolution, adaptive questioning, vulnerable persons, open standards, rational inattention, searching with lies

1 Introduction

Anna is 78 and has low vision. A washing machine she bought online broke within a month, and the seller refuses a refund. The few hundred euros at stake matter to her but are far too little to pay a lawyer. If she is to obtain a remedy, she must state her claim herself, and the forms through which she would do so are designed for people who read a screen easily, understand legal categories, and can concentrate through a long questionnaire.

Authors' Contact Information: Martin Loeb, Charles University, Prague, Czech Republic, loebl@kam.mff.cuni.cz; Zbyněk Loeb, Algostratus, Prague, Czech Republic, Zbynek.Loeb@algostratus.com; Quanyan Zhu, New York University, Brooklyn, NY, USA, quanyan.zhu@gmail.com.

Anna’s situation is common. An estimated 5.1 billion people have justice needs that are not met [57], and equal access to justice is a Sustainable Development Goal (Target 16.3) [52]. Low-income Americans receive no or insufficient legal help for 92% of their substantial civil legal problems [29], and in a large study of U.S. state courts at least one party was self-represented in roughly three quarters of civil cases [26]. People without lawyers must translate their problems into the categories of the law themselves, and vulnerable people are the least able to do so.

Online courts and online dispute resolution (ODR) promised to lower these barriers, and for many people they have [28, 46]. For vulnerable people they often add new ones: a web form that is cheap for a court to run can be impossible to complete for someone with impaired vision, limited literacy, or cognitive decline. Easy and fair access for vulnerable persons is therefore widely regarded as the hardest problem in the digitization of justice, and significant challenges remain across jurisdictions. According to the designers of several online courts, in personal communication with one of us, attempts to build fully accessible court platforms have not succeeded in a number of countries, including Canada, the United States, and the Scandinavian countries.

One important source of the difficulty is structural. With N institutions and M kinds of vulnerability, the current approach asks every institution to accommodate every vulnerability: $N \times M$ separate adaptations, each paid from a single court’s budget and reusable nowhere else. The two other responses do not close the gap. Human assistance, from legal aid to court navigators, works but is scarce [29]. General-purpose AI chatbots are abundant, but they produce unstructured text that a court cannot verify, they invent legal content [8, 33], and they drift into giving legal advice. What is missing is a way for *personalized* assistance to produce output that *any* institution can accept and trust.

We propose to turn the problem upside down. Suppose that a vulnerable person could prepare her filing with tools suited to her own needs, and that the ODR institution could nevertheless read the result as if it had been produced on its own platform. This new concept was proposed by one of us in 2023 [32]. Two things make this possible. The first is an *open ODR standard*: a shared, machine-readable vocabulary of numbered elements for the content of ODR forms, so that what is filed is independent of how it was produced [31]. The second is an *open-source ODR environment* implementing the standard, into which producers of assistive technology can plug their tools. The $N \times M$ problem becomes an $N + M$ problem: each institution accepts the standard once, and each tool is built once and works everywhere. (This counts interfaces, not total cost: schemas for each jurisdiction, certification, and maintenance still take work.) Because the environment reaches a mass market, tool producers have an incentive to build for it, as they did for the plugin ecosystems of open-source e-commerce.

The legal significance of this inversion is that it changes who carries the accommodation. The Convention on the Rights of Persons with Disabilities obliges states to ensure effective access to justice through “procedural and age-appropriate accommodations” (Art. 13) [51, 53], and in the United States the Supreme Court has held that Title II of the Americans with Disabilities Act requires access to the courts [47]. In the European Union, Article 5(2) of the consumer ADR Directive 2013/11/EU, as amended by Directive (EU) 2025/2647, requires the digital tools of ADR entities to be easily accessible and inclusive, including for vulnerable consumers; Member States must adopt implementing measures by 20 March 2028 and apply them from 20 September 2028 [20]. Today each court must provide such accommodations itself. Under the open standard, a person’s accommodations become *portable*: they travel with her tool to any court that accepts standard-conforming filings. Accepting such filings helps a court meet its duty; it does not excuse an inaccessible official channel.

At the heart of every such tool is a conversation, and in a justice setting each design choice in that conversation has legal stakes. *How many questions* can be asked? Attention is the scarcest

resource of many vulnerable users, and every extra question is a burden that can make a person give up. *How should conflicting answers be handled?* People make mistakes, and a machine that reports inconsistency can turn an honest error into doubt about the person’s credibility. *When should the questions stop?* Stopping too early misstates the claim, and a court cannot award a remedy that was never asked for. And the procedure must never feel like an examination, because questioning formats that resemble a test have been observed to diminish the very people they are meant to support [55]. Large language models (LLMs) can phrase questions for each person and turn answers into a written claim, but they may also change what is asked or claimed. This paper begins to supply precise answers to these questions, both mathematical and architectural.

Contributions and organization. Sections 2–3 describe the claiming stage and the burdens of access, and derive design requirements from procedural justice. Section 4 presents the open ODR framework and an LLM-assisted architecture whose verification gates keep legal content fixed by the standard, and proves that, given the same confirmed answers, its decisions do not depend on how questions are rendered. Sections 5–7 develop the elicitation core: detecting, locating, and resolving conflicting answers, including conflicts with the law, and a published elicitation algorithm that chooses certified questions by value per unit of attention, with practical rules for simplifying, re-asking, pausing, and switching modality. Sections 8–9 cover questions generated from the applicable law, filings accompanied by raw data on errors and incomplete answers, and adaptation within a session, across people, and as the law changes. Sections 10–11 report simulations, including adaptation to claimants with higher error rates, and the legal questions that deployment raises.

2 Background and Problem Setting

2.1 Online dispute resolution and online courts

ODR began in e-commerce, where marketplaces built systems that resolve tens of millions of low-value disputes a year without a hearing [28]. Public courts followed. British Columbia’s Civil Resolution Tribunal has resolved disputes online since 2016 [39], England and Wales offer online money claims, and China opened the first of its internet courts in Hangzhou in 2017 [45]; Susskind surveys these and similar initiatives [46], as does one of us [30]. Our framework sits at the entry to online evaluation, facilitation, and judging [46], and to the staged ODR process of the UNCITRAL Technical Notes [50], since each depends on a statement of the person’s claim; the same literature warns that online processes can exclude people who are not digitally capable [46]. The EU’s consumer ODR platform [15], repealed by Regulation (EU) 2024/3228 [19], is a cautionary lesson: a central platform that institutions and traders do not adopt does not improve access. The proposal here is a standard plus an ecosystem, not another platform.

2.2 Vulnerability and the claiming stage

Legal instruments usually define vulnerability through categories such as disability, age, or language. The CRPD follows a social model, in which disability arises from the interaction between impairments and barriers in the environment [51]. Vulnerability theory goes further and treats vulnerability as universal and constant in kind but highly variable in degree and form across persons and over time [22]. Both views imply that vulnerabilities are too heterogeneous and situational for a fixed menu of accessible forms, so accommodation must be personal (Section 4).

Access to justice is access to a just resolution of the problems people actually have, not only to lawyers or courts [43]. Socio-legal research describes how a grievance becomes a dispute in three steps [21]: *naming* an experience as injurious, *blaming* someone for it, and *claiming* a remedy from that person or an institution. Most problems never reach the third step, and the burdens described next weigh most heavily on vulnerable people.

Table 1. Burdens of access to justice [27] and the parts of the framework that address them.

Burden	Example for Anna	Framework component	Formal object
Learning cost	not knowing which remedies exist or which form to use	self-assessment, tool recommendation, open standard (Sec. 4)	intention space Θ ; shared prior (Sec. 9.2)
Compliance cost	reading a long web form with low vision; long sessions	assistive plugins; short adaptive questions	attention cost $c(q)$, stopping rule (Sec. 7)
Psychological cost	feeling tested or disbelieved	local explanation of conflicts; no credibility inference	tension witness (Sec. 6); Design Requirements 1–2
Exclusion by channel	a filing that reveals her disability	equivalence of channels under the standard	Design Requirement 4

The framework targets claiming. Anna knows that her washing machine broke and that the seller is responsible. What she cannot easily do is state a claim in the form an institution requires: which remedy she seeks, on which facts, and what she would accept in settlement. The elicitation problem of Sections 5–7 is the formal version of this step. Its output is a statement of claim, and its errors are errors in what the institution is asked to decide.

2.3 Burdens of access

Public-administration research divides the burden of dealing with the state into *learning costs* (finding out what to do), *compliance costs* (doing it), and *psychological costs* (stress, stigma, and loss of autonomy) [27]. For a vulnerable person filing a claim all three are high, and each can deter a valid claim. Table 1 previews how the design requirements of Section 3, the components of the framework, and the formal results of this paper address each burden.

2.4 Existing building blocks

Standards for legal data. Machine-readable legal data is well established: the European Case Law Identifier numbers judgments [7], Akoma Ntoso structures legislative and judicial documents [35], the OASIS Electronic Court Filing specification carries filings between filers and courts [36], and the EU’s standard forms for small claims and payment orders use numbered fields that are identical across languages [13, 14]. None of them addresses the *content* of an ODR claim independently of its presentation. The open ODR standard proposed by one of us [31] does exactly that. It uses comprehensive numbering conventions for the elements of a claim and is designed to be flexible enough for continual adaptation, local procedural specifics, additions, and amendments.

Guided interviews and AI assistants. Courts and legal-aid organizations already use guided interviews to help self-represented litigants complete forms. A2J Author [6] and the open-source docassemble [40] are widely used examples, and JusticeBot guides laypeople through legal rules and similar cases [56]. Their core interview logic is generally authored by hand and does not explicitly model response errors. LLM assistants converse flexibly but hallucinate legal content [8], with no guarantee about what was asked or how reliable the answers are. Our elicitation layer lies between the two: it produces a structured, standard-conforming claim like a guided interview, adapts to the person like an assistant, and, unlike either, has explicit bounds on how many questions it asks, how it handles inconsistent answers, and when it stops. Technical related work, on query learning [1–3], searching with lies [37], active learning [9, 23, 24], rational inattention [34, 44], and feedback loops in learned systems [12, 38], is discussed where it is used.

3 Design Requirements

The burdens of Section 2.3 translate into requirements that the rest of the paper is designed to satisfy. The psychological cost is the one a court would scrutinize first, since question-and-answer formats are often experienced as testing and can diminish the person asked [55]. Procedural-justice research finds that people accept outcomes when they have a *voice*, the decision maker is *neutral*, they are treated with *respect*, and they can *trust* the authority [48]. The requirements below turn these elements into system properties, with their legal basis in brackets.

Design Requirement 1 (No credibility inference). Measures of inconsistency, including the tension score, are used only to decide which questions to ask again. They are never reported to the counterparty or to the decision maker as a property of the person. [*Respect: a machine-elicited answer with a known error rate must not become a prior inconsistent statement used against credibility.*]

Design Requirement 2 (Local explanation). When answers conflict, the system names the specific answers in tension, ideally a witness of size two (Proposition 6.3), and invites the person to revisit one of them. It does not issue a global verdict of inconsistency. [*Voice and trust: the person corrects her own statement and sees why she is asked.*]

Design Requirement 3 (Honest uncertainty). When the stopping rule ends the elicitation, the residual uncertainty is part of the filing. The system does not resolve it by default choices invisible to the person. A protected elicitation log, kept apart from the filing under access controls, records how each element was elicited (wording, modality, number of times asked) and is disclosed to the court only where needed to assess an element. [*The court's duty to clarify; fair-trial guarantees.*]

Design Requirement 4 (Equivalence of channels). A filing produced through any conforming tool is treated exactly as one produced through the institution's own platform. Equivalence is a property of the standardized filing: the choice of tool, and hence the person's vulnerability, cannot be inferred from it. [*Neutrality; equality of arms; non-discrimination, CRPD Arts. 5 and 13.*]

Design Requirement 5 (Voluntary sharing). Sharing of self-assessments is opt-in and never a condition for access. Withdrawal stops further use and deletes stored records; whether past contributions are also removed from trained models is disclosed in advance. [*Consent and its withdrawal, GDPR Art. 7, and Art. 9 where health and other sensitive data are involved.*]

4 The Open ODR Framework and System Architecture

4.1 The open ODR standard and ecosystem

The standard specifies (a) the hierarchy of ODR processes, from roles (for example, the complainant) to processes and sub-processes, for a catalogue of dispute types, together with the elements that a filing of each type must or may contain; (b) numbering conventions that give every element a stable identifier independent of language and presentation; and (c) a machine-readable format for a completed filing. A filing is then a partial assignment of values to numbered elements. Any tool that produces a valid assignment produces a valid filing, however the questions were presented to the person. In the language of the model below, the standard fixes the space Θ of possible intentions and the meaning of each question; the tool decides only how, and in what order, questions are presented. A working draft of the standard, the Open ODR Model, was published by the National Center for Technology and Dispute Resolution (NCTDR) on 13 October 2026; the complete model, including its initial numbering, is still being developed [31]. Our guarantees are therefore conditional on a schema of the kind described here, not on an adopted specification.

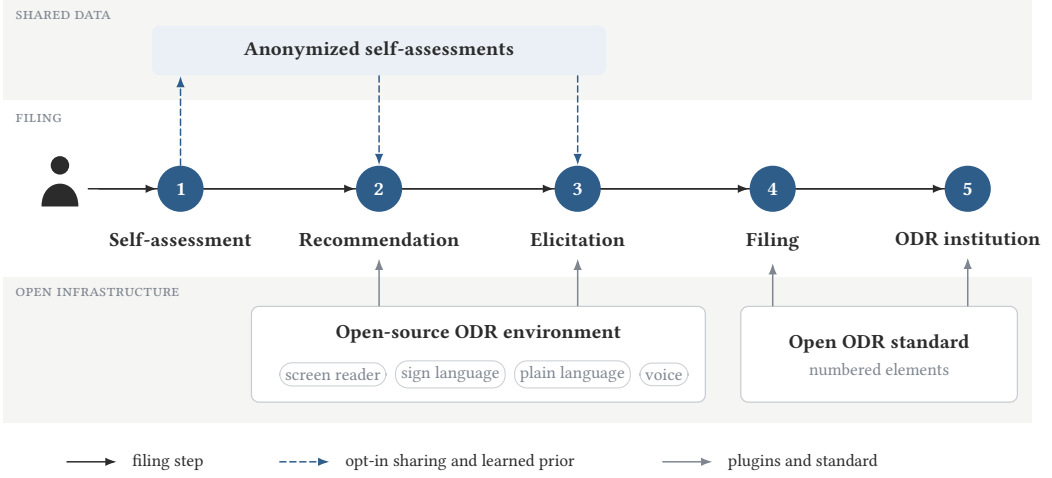


Fig. 1. The open ODR framework. The person passes through five steps, from self-assessment to receipt by the ODR institution, which reads the standard filing as if it came from its own platform. Shared self-assessments improve recommendations and the prior; the open-source environment provides assistive plugins; the open standard defines the numbered elements.

4.2 Personalized filing

The person first assesses her needs with AI assistance and may share the assessment in anonymized form (Section 9.2). She then receives a recommendation of personalization tools, such as a screen reader, a sign-language interface, or a plain-language mode. Using these tools, she prepares her claim by answering a sequence of questions (Sections 5–8) and files it with the ODR institution, whose platform receives it as if it came from its own interface (Figure 1). The open-source environment offers *modelling tools* that edit form content under the numbering conventions, *editing tools* that adapt the forms to the party’s needs, where assistive-technology producers will plug in, and *translating tools* that turn forms into executable process code such as BPMN. We expect the dialogue studied in this paper to take place in the person’s dashboard, as the first part of her digital solution explorer.

4.3 An LLM-assisted system architecture

The first four steps of personalized filing are carried out by a person-facing tool, which Figure 2 decomposes into three modules: *question creation*, *adaptive elicitation*, and *filing assembly*. An *adaptation process* improves all three over time. LLMs help in every module, and every module faces the same risk: an LLM may silently change what a question asks or what a filing claims [8]. The system therefore follows one principle: *the LLM proposes, the standard disposes*. Every LLM output passes a verification gate that maps it back to the numbered elements of the claim schema, and it is rejected if the mapping changes. Gates are themselves fallible, since a second model can misjudge equivalence, so we treat each as passing a wrong output with some probability (p in Proposition E.1), fall back to lawyer-reviewed templates where a gate is unsure, and leave the final check to the person’s read-back. Legal meaning lives in the standard, which lawyers review; language is produced by the LLM; and decisions about what to ask, when to ask again, and when to stop are made by the elicitation core, with the guarantees of Sections 5–7.

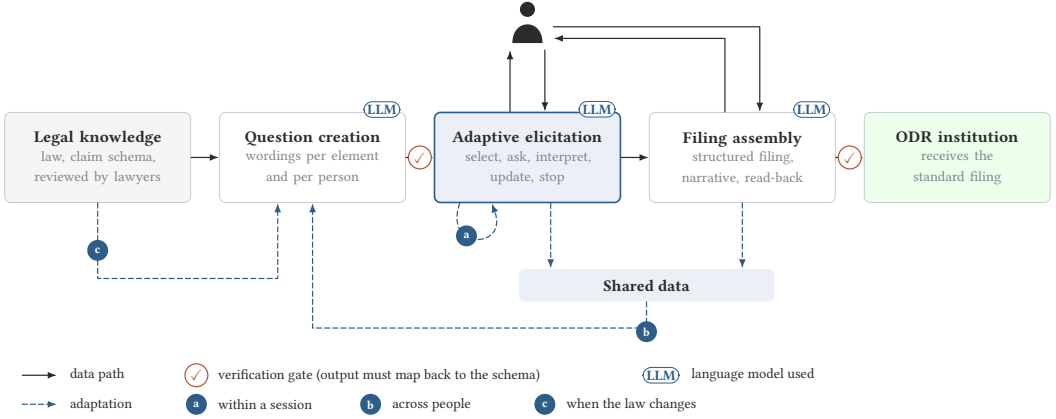


Fig. 2. System architecture. LLMs draft questions, interpret answers, and write narratives; verification gates ensure every output maps back to the claim schema. Adaptation acts (a) within a session, (b) across people through shared data, and (c) when the law changes.

Replayable decisions. The *canonical history* of a session is its sequence of pairs (numbered question, confirmed answer in $\{Y, N, I\}$). Its *declared inputs* are the schema version, the prior, the error and cost models, and the parameters λ and δ of Section 7.

PROPOSITION 4.1 (PRESENTATION INVARIANCE). *If two sessions have the same declared inputs and the person confirms the same answers to the same questions, then they ask the same questions in the same order, stop at the same point, and assemble the same structured filing, whatever wording, modality, or LLM each of them used.*

The proof is in Appendix A. This is a noninterference property: the elicitation algorithm cannot be steered through changes of presentation while the declared inputs and confirmed answers stay the same, and every session can be replayed from its log by an auditor or a court, which supports Design Requirement 4.

This division of labour is the point of the design: lawyers own the legal meaning, the LLM owns the language, the elicitation core owns the decisions, and the person has the final say.

5 Formal Model of Elicitation

5.1 Intentions, questions, and answers

Let Θ be a finite set of possible *intentions*; in the standard, an intention is a complete specification of what the person wants. A *question* is a map $q : \Theta \rightarrow \{Y, N, I\}$, where $q(\theta) = I$ means that a person with intention θ is genuinely indifferent to what q asks. Let $\mathcal{Q}$ be the family of admissible questions, those simple enough to be asked; Section 8 shows how they are generated from the open standard and the applicable law. A person with hidden intention θ^* is asked questions $q_1, q_2, \dots$, possibly chosen adaptively, and gives answers $r_1, r_2, \dots \in \{Y, N, I\}$. An answer is *correct* if $r_t = q_t(\theta^*)$ and *wrong* otherwise.

Two features distinguish this from textbook exact learning with membership queries [1]. First, the third answer I is not an omission of information but information: it tells us that the intention is indifferent on this point, and it makes the target a partially specified function. This differs from incomplete membership oracles, where “I don’t know” is a failure of the oracle [2, 3]. In our system “I do not know”, “please explain”, and a failed recognition are not answers but *deferrals*: they leave the element open and lead to an explanation, another modality, a pause, or human help. Second,

wrong answers are not independent coin flips. Their probability $\varepsilon(q)$ depends on the complexity $\kappa(q)$ of the question and on the person’s fatigue and attention, and in a dispute their content may depend on the person’s interests (Section 5.1).

An intention can be represented as a point, as a set of acceptable outcomes, or as a preference over remedies, each with its own natural questions (Appendix D). Each answer (q, r) defines an *acceptance set* $A(q, r) = \{\theta \in \Theta : q(\theta) = r\}$. A history $h = ((q_1, r_1), \dots, (q_n, r_n))$ is *consistent* if $\bigcap_t A(q_t, r_t) \neq \emptyset$. Figure 4 in Appendix E shows the elicitation loop that the rest of the paper analyzes.

LLMs inside the core. When the person answers in her own words (“I just want my money back”), an LLM maps the utterance to $\{Y, N, I\}$. Its errors compose with the person’s into one effective response channel; for independent symmetric errors $\varepsilon_{\text{eff}} = \varepsilon_h + \varepsilon_L - \frac{3}{2}\varepsilon_h\varepsilon_L$, and in general a confusion matrix is calibrated jointly. A low-confidence reading triggers a short confirmation.

Scope: noise, not strategy. Some misreports in a dispute will be strategic rather than noisy. This paper treats noise; strategic misreporting calls for incentive compatibility and is the natural sequel. Consistency is not truthfulness, so the tension score below measures the reliability of the elicitation, not the honesty of the person.

5.2 A legal reading of the model

Each object in the model has a legal counterpart; Table 7 in Appendix E lists them. Three correspondences explain why the mathematical questions matter for justice. *Elicitation errors change what can be awarded:* under *ne ultra petita* a court may not grant more than, or other than, what was requested, so a missed “yes” to a remedy removes it from the case unless the claim is amended, and identification accuracy bounds the remedies a claimant can receive. *Not every answer belongs in the filing:* the minimum amount Anna would accept is a reservation value for the negotiation or mediation stage and normally confidential, so the standard must mark which elements are disclosed to whom. *Residual uncertainty has an addressee:* courts in many civil-law systems must point out incomplete or unclear submissions (for example §139 of the German and §118a of the Czech Code of Civil Procedure), and the uncertainty left when the stopping rule ends questioning is handed to the judge in exactly the form this duty addresses. In the United States, pro se pleadings are held to less stringent standards than those drafted by lawyers [25].

6 Conflicting Answers: Detect, Locate, Resolve

Conflicting answers are routine in real filings. Anna says she would accept €300, and a few questions later that she would not accept €500; she says the machine broke within a month, and later that it worked for six weeks. Such conflicts come from a misread question, fatigue, anxiety, or an LLM that misheard. Like a good interviewer, the tool must find the clash and ask again without implying that the person is unreliable, in three steps: *detect* that the answers conflict, *locate* the few answers responsible, and *resolve* the conflict in a way the person can follow.

Detecting and locating. The answers given so far are consistent if some intention agrees with all of them. The useful measure is how far they are from consistency.

Definition 6.1. The *tension score* of a history h is $\tau(h) = \min_{\theta \in \Theta} |\{t : r_t \neq q_t(\theta)\}|$, the smallest number of answers that must change to make h consistent. A *tension witness* is an inclusion-minimal set W of answers with $\bigcap_{t \in W} A(q_t, r_t) = \emptyset$.

The witness is what the tool shows the person, so it matters how large it can be. In general, computing τ is NP-hard (Proposition E.2 in Appendix E). Real claim schemas are much simpler:

most questions ask about a single numbered element, such as the remedy sought, the date of purchase, or the acceptable amount.

PROPOSITION 6.2. *Let $\Theta = D_1 \times \dots \times D_m$ and let every question concern a single element. Then $\tau(h) = \sum_j \tau_j(h)$, where τ_j is the tension score of the answers about element j alone, so τ can be computed in $O(\sum_j n_j |D_j|)$ question evaluations, where n_j is the number of answers about element j . Every tension witness consists of answers about a single element j and contains at most $|D_j|$ answers.*

The proof is in Appendix A.

For amounts and dates, which are ordered, the bound improves to two.

PROPOSITION 6.3. *Let Θ be totally ordered and let every acceptance set be an interval of Θ , as for threshold questions and their negations. If h is inconsistent, then it has a tension witness of at most two answers, and of one answer if some acceptance set is empty.*

The proof is in Appendix A; Figure 6 in Appendix E shows Anna’s case. In practice, therefore, every conflict about an amount or a date can be explained by pointing to two of the person’s own answers.

Internal and legal tension. A second source of conflict is the law itself. Let $\Theta_R \subseteq \Theta$, assumed nonempty, be the intentions that the rules R of the claim schema allow (Section 8). Under the Sale of Goods Directive, for example, a price reduction requires one of several grounds, such as the seller’s refusal or failure to repair or replace, a defect that persists despite an attempted repair, or a defect serious enough to justify it immediately (Art. 13(4)) [17]. If Anna asks for a price reduction, says that the seller has agreed to a repair, and reports no other ground, her answers are consistent with each other but not with the law. The tool therefore computes two scores: the *internal* tension $\tau(h)$, minimized over Θ , and the *legal* tension $\tau_R(h) - \tau(h)$, where τ_R is the same minimum taken over Θ_R only. Internal tension is an error to be resolved by asking again. Legal tension is not an error of the person: the tool explains the rule in plain language and lets her choose among the remedies that are available.

Resolving. When the internal tension rises, the tool never reports inconsistency as such. It shows the witness in the person’s modality, quoting her own words where an LLM interpreted them: “Earlier you said you would accept €300. Just now you said you would not accept €500. Which is closer to what you want?” The reply is a *resolution question* that enters the history as a noisy answer like any other. If she confirms both answers, the element is flagged as uncertain in the filing (Design Requirement 3) rather than resolved by the tool. The conflict raises the estimate of her error rate, which changes how the tool asks from then on (Section 7.3), but is never reported as a property of the person (Design Requirement 1).

7 Choosing the Next Question: The Elicitation Algorithm

A fixed form asks every question on it, whatever Anna has already said. An experienced legal-aid interviewer skips what is clear, asks first about what decides the claim, keeps questions short when the client is tired, re-checks what sounded doubtful, and stops when she has enough to file. This section turns that judgement into an explicit elicitation algorithm.

Questions come from a certified bank. The algorithm chooses among questions; it does not invent them. Every question in the bank Q is part of the standard: it has a number, a legal meaning $q : \Theta \rightarrow \{Y, N, I\}$ reviewed by lawyers, and a recorded complexity, and new questions are added only with the approval of the body that maintains the standard (Section 8). An LLM may only re-word or re-render a certified question so that the person can hear it, in her language and modality;

the rendering passes the verification gates of Section 8 and keeps the number and meaning of the question. The elicitation algorithm itself is deterministic and published: from the declared inputs and the confirmed answers, anyone can recompute which question was asked and why (Proposition 4.1). This is a critical advantage over a free-form assistant, whose choice of questions can be neither predicted nor audited.

7.1 Complexity, risk, and value

Three quantities drive the choice; each is computable in one pass over the bank and measurable (Table 6, Appendix E).

Definition 7.1 (Complexity, cost, and risk). The *complexity* $\kappa(q) \in \{1, 2, \dots\}$ of a question is the number of elementary judgements it requires: the number of schema elements it combines plus the number of conditions and negations it contains. It is recorded with the question in the bank. The *attention cost* $c(q) = c_m(\kappa(q))$ and the *risk* $\varepsilon(q, t) = \min\{\frac{1}{2}, \phi \varepsilon_m(\kappa(q), t)\}$, the probability that the answer to q asked as the t -th question of a session is wrong, are read from tables calibrated for each modality m and scaled by the person’s error scale ϕ (Section 9.2). A wrong answer is either of the two other symbols, each with probability $\varepsilon/2$.

“Would you accept €100?” has complexity 1. “If the seller does not repair the machine within a reasonable time, would you prefer part of the price back or to return the machine?” combines two remedies under a condition and has complexity 3; it is expensive and error-prone, especially for Anna. We assume throughout that c_m and ε_m are nondecreasing in complexity and that ε_m is nondecreasing in t : harder questions cost more attention and are answered less reliably, and every question asked makes the next answer slightly less reliable. The intention of the design is that the algorithm moves to simpler questions as Anna gets tired, and Corollary 7.5 below shows that it does.

Definition 7.2 (Value). After history h the tool’s belief is a posterior π_h over intentions, starting from a prior learned from shared self-assessments. For a quantity X that depends on θ , let $H_h(X) = -\sum_x \pi_h(X = x) \log_2 \pi_h(X = x)$ be its entropy, in bits, under π_h , and let $h \cdot (q, r)$ be the history h extended by the answer r to q . The *value* of q is the expected reduction of uncertainty about the filing $f(\theta)$ (Section 8):

$$\text{IG}_h(q) = H_h(f(\theta)) - \sum_{r \in \{Y, N, I\}} \text{Pr}_h(r \mid q) H_{h \cdot (q, r)}(f(\theta)),$$

where $\text{Pr}_h(r \mid q)$ is the probability of answer r predicted from π_h and the risk of q .

“Do you want a refund?” is worth much at the start and nothing once the remedy is settled. Value depends on risk, and the dependence is at most linear. Let $\rho(q) = 1 - \frac{3}{2}\varepsilon(q)$ be the *reliability* of the answer.

PROPOSITION 7.3 (LINEAR DISCOUNT FOR RISK). *A noisy answer is distributed as the true answer $q(\theta)$ with probability $\rho(q)$ and as a uniformly random symbol otherwise. Consequently $\text{IG}_h(q) \leq \rho(q) \text{IG}_h^0(q)$, where $\text{IG}_h^0(q)$ is the value the same question would have if it were answered without error.*

The proof is in Appendix A. An algorithm-friendly alternative is the value $\rho(q) U_h(q)$, with $U_h(q)$ the value of a reliable answer, in bits or, without entropy, the expected reduction in the probability that the most probable filing is wrong; our simulations use Definition 7.2.

The value of one answer is bounded. However complex a question is, its answer is one noisy symbol out of three. This bounds the *value* of a question, not its cost: a complex question may demand much attention, but it cannot return more information than its answer carries.

PROPOSITION 7.4. For every history h and question q , $IG_h(q) \leq \log_2 3 - h_2(\varepsilon(q)) - \varepsilon(q)$, where h_2 is the binary entropy function. Consequently, a question with $\lambda(\log_2 3 - h_2(\varepsilon(q)) - \varepsilon(q)) < c(q)$ is never asked by the algorithm of Section 7.2, at any point of the session.

COROLLARY 7.5. The bound is strictly decreasing in ε on $[0, \frac{2}{3})$. Hence, under the monotonicity assumption above, as a person tires or her error rates rise, the set of questions that can ever be worth asking shrinks toward simpler ones.

The proofs are in Appendix A. With $\lambda = 4$, a question about three issues ($c = 3$) is never worth asking once its measured error rate exceeds 17%, and a question about two issues ($c = 2$) once it exceeds 26%. The conditional question above is therefore split into “Did the seller refuse to repair it?” and, only if needed, “Would you accept part of the price back?”.

7.2 The elicitation algorithm

Elicitation algorithm. At each turn, ask the question with the most value per unit of attention, and stop when no question is worth its cost:

$$q^*(h) \in \arg \max_{q \in Q} \frac{IG_h(q)}{c(q)}, \quad \text{stop when} \quad \max_{q \in Q} [\lambda IG_h(q) - c(q)] \leq 0,$$

with ties broken by question number. The price λ says how much attention one bit of information about the claim is worth. An operator sets it from a session budget, and a person may choose a short or a thorough mode. When the algorithm stops, what remains uncertain is flagged for the read-back and the court, not hidden (Section 9.1).

A packing heuristic, and work in progress. Questions are items with a value and a size, their attention cost, and the person’s attention is the capacity. Choosing by value per unit of attention is the density heuristic for knapsack- and bin-packing-type problems, here in an adaptive and noisy form, since values change with every answer. When the tool offers breaks (Section 7.3), questions are in effect packed into several sittings of bounded attention, a bin-packing problem. The rule is inspired by rational inattention [34, 44] but does not solve a dynamic rational-inattention problem optimally, and it looks one question ahead: two questions can be jointly informative while each alone is worthless, for instance when the filing depends on whether exactly one of two answers is yes; schemas with such interactions should offer the pair as one question. Variants of the algorithm (ratio or net value, lookahead, packing into sittings) and their complexity and approximation guarantees are work in progress, and in our view they are the core mathematical question that the model raises (Appendix C).

7.3 Running the algorithm in a deployed tool

At each turn the tool scores every question in the certified bank, asks the best one, updates its belief, and checks the stopping rule. In our simulation (1944 intentions, 32 questions) a turn takes about a millisecond; larger schemas need a factored or sampled posterior. Table 6 lists how each parameter is obtained. The algorithm motivates five operational rules (Appendix E): simplify questions when answers become unreliable (Corollary 7.5), re-ask what surprised (Table 4, Appendix B), offer a pause rather than file while elements are flagged, switch to the modality in which the person answers most reliably, and explain why a question matters where the stakes are high.

8 Question Creation: Legally Grounded Questions

In ODR the question bank cannot be arbitrary: each question must concern something the law makes relevant, have a fixed legal meaning, and be worded so that the person can answer reliably without being steered. This section develops the question-creation module of Section 4.3.

8.1 From law to questions

Claim schema. For each dispute type, the standard defines a *claim schema*: a set of numbered elements $e_1, \dots, e_m$, each with a domain D_j that includes a value $\perp$ for “left open”, and a set R of legal rules that state which combinations of elements form a valid claim and which elements become relevant under which conditions. An intention is an assignment $\theta = (\theta_1, \dots, \theta_m)$, and the *filing map* f sends an intention to the complete filing package it produces under the rules R ; each audience view (Section 9.1) is a projection of $f(\theta)$, so the reservation value belongs to $f(\theta)$ but only to the mediator’s view.

Example. In the European Union, Anna’s claim is governed by the Sale of Goods Directive [17]. The seller is liable for a lack of conformity that becomes apparent within two years of delivery (Art. 10). The consumer may first require repair or replacement, and may ask for a price reduction or terminate the contract only if, among other cases, the seller has not completed repair or replacement or has refused to do so (Art. 13). Table 5 in Appendix E shows elements of the schema and the questions derived from them. Facts that condition a remedy (“Did the seller refuse to repair or replace it?”) are asked before the remedy that depends on them. A remedy whose conditions are not yet established is marked *eligibility-uncertain* and stays open, so that the person can clarify the facts; it is never suppressed on a provisional legal inference.

Question types. Three kinds of question arise from the schema. *Element questions* ask whether θ_j lies in a set of values $V \subseteq D_j$, and their answer is I when $\theta_j = \perp$. *Threshold questions* are element questions for ordered domains such as amounts or dates, with $V = \{v \in D_j : v \leq x\}$. *Condition questions* ask about the facts on which a rule in R makes another element relevant, and are asked only while that element is still open. *Compound questions* combine several elements and have higher complexity κ . For a remedy, Y means it is sought, N that it is not wanted, and I ($\theta_j = \perp$) that the person leaves it open. N and I give the same court view (“not claimed”), but only I lets the mediator offer the remedy in settlement; deferrals (Section 5) are neither.

Separating semantics from presentation. The meaning of a question, the map $q : \Theta \rightarrow \{Y, N, I\}$, is fixed by the standard and reviewed by lawyers once for each schema and jurisdiction. These certified questions form the bank from which the elicitation algorithm of Section 7 chooses; a new question enters the bank only with the approval of the body that maintains the standard. Only the wording and its rendering are generated per person, by an LLM, and admitted through the verification gates below (Figure 5 in Appendix E). A hallucination can therefore make a question harder to answer, which the complexity estimate penalizes, but, as long as the equivalence gate is sound, it cannot change what an answer means.

Verification gates. For each element of the claim schema, the legal rule that makes it relevant, and the person’s profile (language, modality, reading level), an LLM drafts candidate wordings. A candidate enters the question bank only if it passes three gates: an *equivalence* check, in which the wording is mapped back, by a second model or a reviewed template, to the element and value set it was drafted from; a *neutrality* check that rejects presupposed answers and evaluative language, since evidence law restricts leading questions for the same reason [54] and vulnerable persons may be more inclined to agree with the questioner; and a *complexity* check, in which the complexity $\hat{\kappa}$ of the wording (Definition 7.1), adjusted for length and legal concepts, must be low enough for the wording to pass the ceiling of Proposition 7.4 for the person’s profile.

Amounts without anchoring. “Would you accept €300?” is suggestive in a way no wording check removes: the amount named can anchor what the person says she would accept, and a person who is unsure may simply agree. Possible remedies are to let her state an amount before any is

proposed and use thresholds only to confirm it, and to ask about the reservation value only in the confidential mediation view. Which mechanism serves vulnerable persons best is open and must be settled in a pilot; the tension analysis of Section 6 applies to whatever threshold answers are collected.

What the legal structure buys. Intentions with the same filing are legally equivalent, so questions that cannot change the filing are never asked, and the number of questions needed depends on the number of distinct filings rather than on $|\Theta|$ (Appendix C).

9 Filing, Error Reporting, and Adaptation

The elicitation algorithm ends with one intention and a record of how it was obtained. The filing of that intention is done by the standard, through the filing map f , so assembly involves no modelling choice. What remains to design is the information on errors and incomplete answers that accompanies the filing, and how the tool improves over time.

9.1 The filing and its error record

One intention, one filing. When the algorithm stops, it returns an intention $\hat{\theta}$ whose filing is most probable, $f(\hat{\theta}) = \hat{F} \in \arg \max_F \pi_h(\{\theta : f(\theta) = F\})$, with ties broken by element number. The *structured filing* $\hat{F}$, an assignment of values to the numbered elements, is the legal instrument. The *narrative* written by an LLM is only a rendering of it, and the standard should state that where the two differ the numbered elements prevail. Each element carries an audience label (court, counterparty, or mediator); the reservation value, for example, appears only in the mediator's view.

Raw data on errors and incompleteness, not probabilities. The court receives facts about the elicitation, not a model's probabilities. For each element the record states whether it was answered directly or inferred, how often it was asked and what was answered, any conflict (with its witness) or deferral, and whether the person confirmed it at the read-back. An element is *flagged* when it was left open, only inferred, or part of an unresolved conflict, or when the tool's internal confidence $\gamma_j = \pi_h(\{\theta : f_j(\theta) = \hat{F}_j\})$ is below $1 - \delta$; the flag states its reason in factual terms, for low confidence that the answers about the element do not settle it (Design Requirement 3). Probabilities stay inside the tool, where they decide what to ask, what to read back, and when to stop. The flags and their reasons are part of the court's view and address its duty to clarify (Section 5.2); the full record is the protected elicitation log, kept with the schema version and outside every audience view.

Read-back. An LLM renders each audience view as a narrative, and a *read-back* for the person in her modality and language. Every generated text passes a faithfulness gate: a parser maps it back to schema elements and must reproduce the view exactly, with no added facts and no omitted remedies. Reading back every element would cost too much attention, so the read-back, a short sequence of confirmations, covers the remedies, whose omission cannot be cured under *ne ultra petita*, the flagged elements, and every element the person never answered directly. Even when a single consistent intention remains, nothing is filed that the person has not confirmed. A rejected element is re-opened as a question, exactly like an answer in a tension witness.

A probabilistic analysis for certification. Probabilities have one legitimate use outside the tool: in deciding whether a tool is good enough to be certified. Proposition E.1 in Appendix E bounds the probability that what is filed misstates the claim by $(1 - d)P_e + gp$, where P_e is the probability that the most probable filing is wrong, d the probability that the read-back detects such an error, g the probability of an unfaithful narrative, and p the probability that the faithfulness gate accepts one.

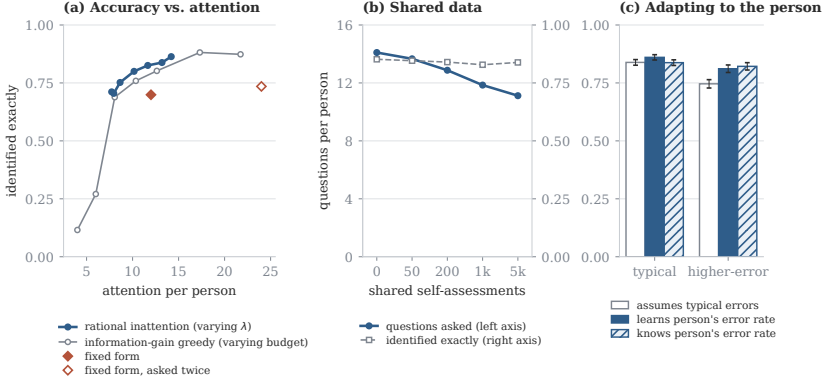


Fig. 3. Simulation results. (a) Exact identification against attention spent ($M = 1000$); RI for $\lambda \in \{2, \dots, 16\}$, IG for budgets $B \in \{4, \dots, 20\}$. (b) Questions asked and accuracy as shared self-assessments grow (RI, $\lambda = 8$). (c) Accuracy for typical and simulated higher-error claimants when the tool assumes typical errors, learns the person’s error rate, or knows it; bars show one standard error.

The stopping rule and the read-back control the elicitation term, better gates shrink the LLM term, and under the precedence rule the LLM term cannot change the legal content of the claim.

9.2 Adaptation

(a) *Within a session.* The tool models the person’s error rate on the t -th question as $\varepsilon_i(q, t) = \min\{\frac{1}{2}, \phi_i \varepsilon(q, t)\}$, where $\varepsilon(q, t)$ is the typical error rate and ϕ_i a person-specific scale whose prior comes from her self-assessment, and it updates its belief about ϕ_i jointly with its belief about θ . The rules of Section 7.3 then follow automatically: a rising $\hat{\phi}_i$ simplifies the questions, makes the tool trust single answers less, and leads it to re-ask surprising ones.

(b) *Across people.* Pooled session logs calibrate three things: the prior over intentions; the measured error rate of each *wording* (Table 6), so that wordings that confuse people are retired; and the model that interprets free answers. The prior over intentions must not be conditioned on a person’s vulnerability profile, which would let a diagnosis shape her claim. Because the learned prior decides which questions are asked, it can lock in the biases of early users [12, 38]; exploration and audits of question policies across groups are minimal safeguards.

(c) *When the law or the model changes.* The claim schema is versioned: when the law changes, only questions touching changed elements are regenerated and re-verified, and when the LLM is replaced, the whole bank is re-run through the gates before deployment.

10 Simulation

We simulate the elicitation loop on a stylized version of Anna’s complaint. Code and data will be released with the final version.

Setup. Intentions specify five issues, each Y, N, or I, and a minimum amount on eight levels ($|\Theta| = 1944$). The tool can ask 32 questions of complexity $\kappa \in \{1, 2, 3\}$ and cost $c(q) = \kappa(q)$, with initial error rates from 3% to 16% that grow with fatigue. The prior is learned from M shared intentions of a three-type population. We compare a fixed form of twelve simple questions (Fixed; Fixed $\times 2$ asks it twice), greedy information gain with a budget of B questions (IG(B)), and the elicitation algorithm of Section 7.2 (RI(λ)), on 2000 simulated people each. A further experiment

gives 40% of people a higher error scale ($\phi_i = 2.5$, otherwise 1), a synthetic stand-in, not a model of disability, and compares RI(8) tools that assume typical errors (*naive*), learn ϕ_i in session (*adaptive*), or know it (*oracle*). Appendix B gives details.

We report five findings; Table 2 in Appendix B lists selected values.

Adaptive questioning saves attention. RI(2) reaches the accuracy of the fixed form (0.71 versus 0.70) with 7.7 instead of 12 questions, about 36% fewer. With 11.7 questions, RI(8) identifies 83% of intentions exactly, against 70% for the fixed form with 12. Asking the fixed form twice raises accuracy only to 0.74, because fatigue raises the error rate of the repeated questions.

Pricing attention keeps questions simple; shared data shortens sessions. At equal attention, RI is on or above the IG frontier (0.80 versus 0.76 at about ten units) and never asks compound questions; with 5000 rather than 0 shared self-assessments, it asks 11.1 instead of 14.1 questions at the same accuracy.

Adapting to the person closes much of the gap. A tool that assumes typical errors identifies 84% of typical claimants' intentions but only 75% of higher-error claimants'. Learning the person's error scale during the session raises the higher-error group to 81%, close to the 82% of a tool that knows it, and nearly halves the gap between the groups, from 9 to 5 points. The price is about two more questions for higher-error claimants (14.1 instead of 12.2); the gain comes from re-asking, not from simpler questions. In the session of Table 4 (Appendix B), a mistaken "I do not care" to the refund question leads the naive tool to file without the refund, which could prevent the court from awarding it absent clarification or amendment; the adaptive tool re-asks and files it.

Flags make the remaining errors visible. Beyond the remedies, which are always confirmed, flags with $\delta = 0.1$ add on average 0.23 elements per filing for RI(8) and catch two thirds of all wrong elements. Without flags, 15.9% of RI(8) filings contain a wrong element; with them, 5.3% contain one that is not flagged (fixed form: 29.2% and 16.5%). The posterior is slightly overconfident: it predicts a wrong filing with probability 14.0%, against $15.9 \pm 0.8\%$ observed. All of this evidence comes from a stylized error model on a synthetic population without strategic misreports; LLM wording, gate failures, and real users require an end-to-end study.

11 Legal and Regulatory Questions

Status of filings: legislation, as for the European standard forms [13, 14], or procedural rules must oblige or permit institutions to accept filings from conforming tools, which raises who certifies tools and who is liable for a misstated claim. *Advice versus assistance:* the tool structures a claim without advising on its merits, which matters where legal advice is reserved to lawyers. *AI regulation:* Annex III, point 8(a) of the EU AI Act covers AI intended to assist a judicial authority in researching and interpreting facts and law and in applying the law, or used similarly in alternative dispute resolution [18]. A tool that only structures a party's filing arguably falls outside; one that assesses the merits for the decision maker may fall inside. *Data protection:* a self-assessment may reveal health data, a special category under Article 9 GDPR, for which explicit consent is the natural ground, though not every accessibility preference is health data [16]; small populations are hard to anonymize, so sharing needs formal guarantees such as differential privacy [11].

12 Research Agenda and Conclusion

We have proposed that vulnerable persons bring their own tools while an open ODR standard makes the result legible to any court that adopts it. The next step is a pilot that calibrates Table 6 with real users, vulnerable ones included, measuring faithful filings, abandonment, and undetected omissions.

References

- [1] Dana Angluin. 1988. Queries and concept learning. *Machine Learning* 2, 4 (1988), 319–342.
- [2] Dana Angluin, Mārtiņš Kriķis, Robert H. Sloan, and György Turán. 1997. Malicious omissions and errors in answers to membership queries. *Machine Learning* 28, 2–3 (1997), 211–255.
- [3] Dana Angluin and Donna K. Slonim. 1994. Randomly fallible teachers: Learning monotone DNF with an incomplete membership oracle. *Machine Learning* 14, 1 (1994), 7–26.
- [4] Maria-Florina Balcan and Nicholas J. A. Harvey. 2018. Submodular functions: Learnability, structure, and optimization. *SIAM J. Comput.* 47, 3 (2018), 703–754.
- [5] Elwyn R. Berlekamp. 1968. Block coding for the binary symmetric channel with noiseless, delayless feedback. In *Error-Correcting Codes*, Henry B. Mann (Ed.). Wiley, New York, 61–85.
- [6] Center for Computer-Assisted Legal Instruction and Chicago-Kent College of Law. 2026. A2J Author: Guided Interviews for Self-Represented Litigants. <https://www.a2jauthor.org>.
- [7] Council of the European Union. 2011. Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law. OJ C 127, 29.4.2011.
- [8] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–93.
- [9] Sanjoy Dasgupta. 2005. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 18. Curran Associates, Red Hook, NY, 337–344.
- [10] Brian W. Dolhansky and Jeff A. Bilmes. 2016. Deep submodular functions: Definitions and learning. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 29. Curran Associates, Red Hook, NY, 3404–3412.
- [11] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [12] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*. ACM, New York, NY, 160–171.
- [13] European Parliament and Council. 2006. Regulation (EC) No 1896/2006 creating a European order for payment procedure. OJ L 399, 30.12.2006.
- [14] European Parliament and Council. 2007. Regulation (EC) No 861/2007 establishing a European Small Claims Procedure. OJ L 199, 31.7.2007.
- [15] European Parliament and Council. 2013. Regulation (EU) No 524/2013 on online dispute resolution for consumer disputes. OJ L 165, 18.6.2013.
- [16] European Parliament and Council. 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). OJ L 119, 4.5.2016.
- [17] European Parliament and Council. 2019. Directive (EU) 2019/771 on certain aspects concerning contracts for the sale of goods. OJ L 136, 22.5.2019.
- [18] European Parliament and Council. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L, 2024/1689, 12.7.2024.
- [19] European Parliament and Council. 2024. Regulation (EU) 2024/3228 of 19 December 2024 repealing Regulation (EU) No 524/2013 with regard to the discontinuation of the European Online Dispute Resolution Platform. OJ L, 2024/3228, 30.12.2024.
- [20] European Parliament and Council. 2025. Directive (EU) 2025/2647 of 16 December 2025 amending Directive 2013/11/EU on alternative dispute resolution for consumer disputes. OJ L, 2025/2647, 30.12.2025.
- [21] William L. F. Felstiner, Richard L. Abel, and Austin Sarat. 1980. The Emergence and Transformation of Disputes: Naming, Blaming, Claiming *Law & Society Review* 15, 3/4 (1980), 631–654.
- [22] Martha Albertson Fineman. 2008. The Vulnerable Subject: Anchoring Equality in the Human Condition. *Yale Journal of Law & Feminism* 20, 1 (2008), 1–23.
- [23] Daniel Golovin and Andreas Krause. 2011. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research* 42 (2011), 427–486.
- [24] Daniel Golovin, Andreas Krause, and Debajyoti Ray. 2010. Near-optimal Bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 23. Curran Associates, Red Hook, NY, 766–774.
- [25] Haines v. Kerner. 1972. 404 U.S. 519 (U.S. Supreme Court).
- [26] Paula Hannaford-Agor, Scott Graves, and Shelley Spacek Miller. 2015. *The Landscape of Civil Litigation in State Courts*. Technical Report. National Center for State Courts.
- [27] Pamela Herd and Donald P. Moynihan. 2018. *Administrative Burden: Policymaking by Other Means*. Russell Sage Foundation, New York.

- [28] Ethan Katsh and Orna Rabinovich-Einy. 2017. *Digital Justice: Technology and the Internet of Disputes*. Oxford University Press, Oxford.
- [29] Legal Services Corporation. 2022. *The Justice Gap: The Unmet Civil Legal Needs of Low-income Americans*. Technical Report. Legal Services Corporation, Washington, DC.
- [30] Zbyněk Loeb. 2019. *Designing Online Courts: The Future of Justice Is Open to All*. Kluwer Law International, Alphen aan den Rijn, The Netherlands.
- [31] Zbyněk Loeb. 2026. Open ODR Model: Discussion Draft. National Center for Technology and Dispute Resolution (NCTDR), <https://odr.info>. Published 13 October 2026.
- [32] Zbyněk Loeb, Davide Carneiro, Ciprian Ciubotariu, and Tereza Rezabkova. 2023. Personalized e-Justice ODR Forms. *International Journal of Online Dispute Resolution* 10, 1 (Nov. 2023), 97–122.
- [33] Mata v. Avianca, Inc. 2023. 678 F. Supp. 3d 443 (S.D.N.Y.).
- [34] Filip Matějka and Alisdair McKay. 2015. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105, 1 (2015), 272–298.
- [35] OASIS. 2018. Akoma Ntoso Version 1.0. OASIS Standard, LegalDocumentML (LegalDocML) Technical Committee.
- [36] OASIS. 2023. Electronic Court Filing Version 5.01. OASIS Committee Specification 01, LegalXML Electronic Court Filing Technical Committee.
- [37] Andrzej Pelc. 2002. Searching games with errors—fifty years of coping with liars. *Theoretical Computer Science* 270, 1–2 (2002), 71–109.
- [38] Juan C. Perdomo, Tijana Zrnic, Celestine Mender-Dünner, and Moritz Hardt. 2020. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, Virtual, 7599–7609.
- [39] Province of British Columbia. 2012. Civil Resolution Tribunal Act. SBC 2012, c. 25.
- [40] Jonathan Pyle. 2026. docassemble: A Free, Open-Source Expert System for Guided Interviews and Document Assembly. <https://docassemble.org>.
- [41] Sofya Raskhodnikova and Grigory Yaroslavtsev. 2013. Learning pseudo-Boolean k -DNF and submodular functions. In *Proceedings of the 24th ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, Philadelphia, PA, 1356–1368.
- [42] Alfréd Rényi. 1961. On a problem of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 6 (1961), 505–516.
- [43] Rebecca L. Sandefur. 2019. Access to What? *Daedalus* 148, 1 (2019), 49–55.
- [44] Christopher A. Sims. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50, 3 (2003), 665–690.
- [45] Supreme People’s Court of the People’s Republic of China. 2019. *Chinese Courts and Internet Judiciary*. Technical Report. Supreme People’s Court, Beijing.
- [46] Richard Susskind. 2019. *Online Courts and the Future of Justice*. Oxford University Press, Oxford.
- [47] Tennessee v. Lane. 2004. 541 U.S. 509 (U.S. Supreme Court).
- [48] Tom R. Tyler. 1990. *Why People Obey the Law*. Yale University Press, New Haven.
- [49] Stanislaw M. Ulam. 1976. *Adventures of a Mathematician*. Scribner, New York.
- [50] UNCITRAL. 2017. *UNCITRAL Technical Notes on Online Dispute Resolution*. United Nations, Vienna.
- [51] United Nations. 2006. Convention on the Rights of Persons with Disabilities. G.A. Res. 61/106, U.N. Doc. A/RES/61/106.
- [52] United Nations. 2015. Transforming our World: The 2030 Agenda for Sustainable Development. G.A. Res. 70/1, U.N. Doc. A/RES/70/1.
- [53] United Nations High Commissioner for Human Rights. 2017. Right to Access to Justice under Article 13 of the Convention on the Rights of Persons with Disabilities. U.N. Doc. A/HRC/37/25.
- [54] United States. 2011. Federal Rules of Evidence, Rule 611: Mode and Order of Examining Witnesses and Presenting Evidence.
- [55] Joseph Webb, Val Williams, Sandy Read, Harry Davis, and Roy James. 2024. A social practice theory approach to exploring the ubiquity of quizzes in dementia care settings. *Ageing & Society* 44, 8 (2024), 1809–1830. doi:10.1017/S0144686X22001052
- [56] Hannes Westermann and Karim Benyekhlef. 2023. JusticeBot: A Methodology for Building Augmented Intelligence Tools for Laypeople to Increase Access to Justice. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAAIL)*. ACM, New York, NY, 351–360.
- [57] World Justice Project. 2019. *Measuring the Justice Gap: A People-Centered Assessment of Unmet Justice Needs Around the World*. Technical Report. World Justice Project, Washington, DC.

A Proofs

PROOF OF PROPOSITION 4.1. The elicitation core and the legal layer of assembly are deterministic functions of the declared inputs and the canonical history, with ties broken by element number. By

induction on the turn, both sessions hold the same belief and choose the same question, and by hypothesis they then record the same confirmed answer. $\square$

PROOF OF PROPOSITION E.2. The first claim follows by evaluating each candidate θ against all n answers. For the second, membership in NP is clear (guess θ). For hardness, reduce from MAX-SAT: given clauses $C_1, \dots, C_n$ over m variables, let $q_t(\theta) = Y$ if θ satisfies C_t and N otherwise, and let every answer be $r_t = Y$. Then $\tau(h)$ is the minimum number of clauses left unsatisfied by an assignment. $\square$

PROOF OF PROPOSITION 6.2. The number of answers violated by θ is a sum of terms that each depend on one coordinate θ_j , so its minimum over the product is the sum of the per-element minima. A product of sets is empty exactly when one factor is empty, so a minimal infeasible family lies within one element. In a minimal infeasible family F of subsets of D_j , each member A misses some point $x_A \in \bigcap (F \setminus \{A\})$; these points are distinct, so $|F| \leq |D_j|$. $\square$

PROOF OF PROPOSITION 6.3. If some acceptance set is empty, that answer alone is a witness. Otherwise let $I_t = [a_t, b_t] \cap \Theta$ be the nonempty acceptance sets and suppose they pairwise intersect. Let s maximize a_t and s' minimize b_t . Since $I_s \cap I_{s'} \neq \emptyset$, it contains some $\theta \in \Theta$ with $a_s \leq \theta \leq b_{s'}$, and hence $a_t \leq \theta \leq b_t$ for all t . So all sets intersect, and an inconsistent history must contain an inconsistent pair. $\square$

PROOF OF PROPOSITION 7.3. Write $\varepsilon = \varepsilon(q)$ and $\rho = 1 - \frac{3}{2}\varepsilon$. The answer channel has matrix C with $1 - \varepsilon$ on the diagonal and $\varepsilon/2$ elsewhere, so $C = \rho \text{Id} + (1 - \rho)U$, where every entry of U is $\frac{1}{3}$; indeed $\rho + \frac{1-\rho}{3} = 1 - \varepsilon$ and $\frac{1-\rho}{3} = \frac{\varepsilon}{2}$. Hence the answer R can be generated by passing $q(\theta)$ through an erasure channel E that erases with probability $1 - \rho$, independently of θ and h , and replacing an erasure by a uniformly random symbol. By the data-processing inequality, $I(f(\theta); R) \leq I(f(\theta); E) = \rho I(f(\theta); q(\theta))$, and the right-hand side is $\rho \text{IG}_h^0(q)$. $\square$

PROOF OF PROPOSITION 7.4. $\text{IG}_h(q)$ is the mutual information between the filing $f(\theta)$ and the answer R under π_h . Since R depends on the filing only through θ , the data-processing inequality gives $I(f(\theta); R) \leq I(\theta; R) = H(R) - H(R \mid \theta) \leq \log_2 3 - H(R \mid \theta)$. Given θ , the answer has distribution $(1 - \varepsilon, \varepsilon/2, \varepsilon/2)$, whose entropy is $h_2(\varepsilon) + \varepsilon$. $\square$

PROOF OF COROLLARY 7.5. The derivative of $h_2(\varepsilon) + \varepsilon$ is $\log_2 \frac{1-\varepsilon}{\varepsilon} + 1$, which is positive exactly when $\varepsilon < \frac{2}{3}$; error rates in the model are at most $\frac{1}{2}$. $\square$

PROOF OF PROPOSITION E.1. A misstatement requires either a wrong structured filing that is not detected at the read-back (a detected one is corrected or withheld), with probability at most $P_e(1 - d)$, or an unfaithful narrative that passes the gate, with probability at most gp . The union bound gives the claim. $\square$

PROOF OF PROPOSITION C.1. An adaptive strategy is a ternary decision tree of depth n . Fix θ and an error pattern, consisting of a set of at most k positions together with a wrong symbol at each. Together they determine a unique root-to-leaf path. Two different patterns for the same θ first differ at some position t ; up to that point the histories coincide, so the same question is asked at t and the two answers differ. Hence the two paths reach different leaves. There are $\sum_{j \leq k} \binom{n}{j} 2^j$ patterns per θ . Correct identification requires that leaves reached from different intentions be distinct, and there are at most 3^n leaves. $\square$

B Simulation Setup

Intentions. An intention specifies five issues (refund, replacement, repair, reimbursement of costs, apology), each Y, N, or I, together with the minimum acceptable amount a on a grid of eight levels. This gives $|\Theta| = 3^5 \cdot 8 = 1944$. In the simulation every intention is its own filing package (the amount belongs to the mediator’s view), so information about $f(\theta)$ equals information about θ .

Questions. There are 32 admissible questions: five *issue questions* (“Do you want a refund?”, $\kappa = 1$), seven *threshold questions* (“Would you accept level x ?”, $\kappa = 1$), ten *pairwise compound questions* (“Do you want both a repair and an apology?”, $\kappa = 2$), and ten *triple compound questions* ($\kappa = 3$). A compound question (“Do you want both...?”) has truth Y if all of its issues are Y, I if all are I, and N otherwise: a person who wants only some of the remedies does not want the bundle, a modelling choice that matters little because RI never asks compound questions. The attention cost of a question is $c(q) = \kappa(q)$.

Errors. The t -th question is answered wrongly with probability $\varepsilon(q, t) = \min\{0.5, 0.03 \kappa(q)^{1.5}(1 + 0.04 t)\}$, so a fresh simple question has a 3% error rate and a fresh triple question 16%; each question asked raises error rates by 4% of their initial value, so that at the 40th question they reach 8% and 40%. A wrong answer is either incorrect symbol with equal probability.

Population and prior. True intentions are drawn from a mixture of three claimant types (“wants the money back”, “wants the machine working”, “wants recognition”) with weights 0.5, 0.3, and 0.2; within a type, issues and amount are independent, with the distributions of Table 3. The tool’s prior is the empirical distribution of M shared intentions from the same population with add-one smoothing; $M = 0$ gives the uniform prior. The tool knows the error model and reports the most likely intention when it stops.

Policies. Fixed asks the twelve simple questions once in a fixed order, as a paper form would, and Fixed×2 asks them twice. IG(B) greedily maximizes information gain over all 32 questions for a budget of B questions. RI(λ) is the elicitation algorithm of Section 7.2: it maximizes $\text{IG}_h(q)/c(q)$ and stops when $\max_q [\lambda \text{IG}_h(q) - c(q)] \leq 0$, with at most 40 questions. Each configuration is run for 2000 simulated people; the standard error of every accuracy reported is at most 0.011.

Adapting to the person. In the third experiment each person’s error probability is $\min\{\frac{1}{2}, \phi_i \varepsilon(q, t)\}$ with $\phi_i = 1$ (probability 0.6) or $\phi_i = 2.5$ (probability 0.4). All three tools use RI(8) with the prior learned from 1000 shared intentions. The adaptive tool holds a joint posterior over (θ, ϕ) with $\phi \in \{1, 1.75, 2.5\}$ and a uniform prior on ϕ ; information gain is computed on θ alone, $\text{IG}_h(q) = I(\theta; R)$. Each tool is run on the same 1500 people (917 typical, 583 higher-error), so the standard error of each group accuracy is at most 0.018. The session in Table 4 is the first higher-error person, in simulation order, whom the naive tool misidentifies and the adaptive tool identifies, with both tools facing the same random answer errors. Code: `sim/simulate_adapt.py`, seed 20261013.

Flagging and calibration. With the prior learned from 1000 shared intentions, 2000 people are run through Fixed and RI(8). The tool assembles the most probable filing, computes the confidence γ_j of each of the six elements, and flags elements with $\gamma_j < 1 - \delta$. For $\delta = 0.1$, Fixed flags 0.37 elements per filing and 46% of wrong elements; RI(8) flags 0.23 and 67%. In this setting both policies ask about every element at least once, so no element is read back merely for being inferred. Results for $\delta = 0.05$ and 0.2 are similar for RI(8), whose sessions end with elements that are either nearly certain or clearly uncertain. Code: `sim/simulate_flags.py`, seed 20261017. Experiments 1 and 2 use `sim/simulate.py`, seed 20261011.

Table 2. Selected results ($M = 1000$ shared self-assessments, 2000 people per row). “Exact” is the fraction of people whose full intention is identified; “elements” is the fraction of the six elements of the filing that are correct.

Policy	questions	attention	compound questions	exact	elements
Fixed	12.0	12.0	0%	0.70	0.943
Fixed×2	24.0	24.0	0%	0.74	0.952
IG(12)	12.0	12.6	5.3%	0.80	0.965
IG(16)	16.0	17.3	8.2%	0.88	0.980
RI(2)	7.7	7.7	0%	0.71	0.947
RI(8)	11.7	11.7	0%	0.83	0.970
RI(16)	14.2	14.2	0%	0.86	0.977

Table 3. Population model: probability of Y/N/I for each issue, and distribution of the amount level $a \in \{0, \dots, 7\}$, for each claimant type.

Type (weight)	refund	replacement	repair	costs	apology	amount $a = 0, \dots, 7$						
money back (0.5)	.8/.1/.1	.1/.7/.2	.1/.6/.3	.6/.2/.2	.1/.2/.7	.02	.03	.05	.10	.15	.25	.15
machine working (0.3)	.1/.7/.2	.8/.1/.1	.6/.2/.2	.2/.3/.5	.1/.3/.6	.30	.25	.20	.10	.07	.04	.02
recognition (0.2)	.4/.3/.3	.3/.3/.4	.2/.3/.5	.7/.1/.2	.8/.1/.1	.10	.15	.20	.20	.15	.10	.05

Table 4. A simulated session with a higher-error claimant. Both tools face the same answer errors; * marks a wrong answer. Confidence is the posterior probability of the most likely intention.

#	Question	Answer	naive tool	adaptive tool	$\hat{\phi}$
			confidence	confidence	
1	Do you want it repaired?	I	0.01	0.01	1.75
2	Do you want a replacement?	Y	0.02	0.02	1.75
3	Do you want your costs reimbursed?	Y	0.06	0.05	1.75
4	Do you want a refund?	I*	0.11	0.10	1.75
5	Do you want an apology?	Y	0.25	0.22	1.74
6	Would you accept €100?	Y	0.49	0.44	1.73
7	Would you accept €50?	N	0.78	0.65	1.75
8	Would you accept €100?	Y	0.83	0.73	1.71
9	Would you accept €50?	N	0.87	0.79	1.68
10	Do you want a replacement?	Y	0.90	0.84	1.65
11	Do you want an apology?	Y	0.93	0.88	1.63
12	Do you want a refund?	Y	stops	0.51	1.84
13	Do you want a refund?	Y		0.87	1.83
14	Do you want a refund?	Y		0.92	1.80

C Query Complexity with Errors and Abstention

What the legal structure buys. Intentions with the same filing are legally equivalent, so the value of a question is measured on $f(\theta)$: questions that cannot change the filing are never asked, an instance of equivalence-class determination [24]. The information-theoretic lower bound on the number of questions (Proposition C.1) depends on the number of distinct filings rather than on $|\Theta|$ (Corollary E.3); matching upper bounds for legally admissible question families remain open.

When all answers are correct and only Y/N answers are possible, identifying θ^* requires $\lceil \log_2 |\Theta| \rceil$ questions in the worst case. The Rényi–Ulam searching game asks the same question when up to k answers may be lies [42, 49]; see [37] for a survey. The classical volume bound [5] says that n adaptive yes/no questions can tolerate k lies only if $|\Theta| \sum_{j=0}^k \binom{n}{j} \leq 2^n$. The bound extends to three answers.

PROPOSITION C.1. *Suppose that up to k of the n answers may be wrong, and that each wrong answer may be either of the two symbols other than the correct one. If an adaptive strategy with n questions always identifies θ^* , then*

$$|\Theta| \sum_{j=0}^k \binom{n}{j} 2^j \leq 3^n.$$

The proof is in Appendix A.

For $k = 0$ the bound gives $n \geq \log_3 |\Theta|$, so abstention can save up to a factor $\log_2 3 \approx 1.58$ in the number of questions, but only when indifference genuinely discriminates between intentions. When I is rare, the gain disappears. A precise theory for structured question families (thresholds, membership queries for R2, and comparisons for R3; see Appendix D) and for errors that depend on question complexity rather than being adversarial is open. Pelc’s survey records sharp results for several of these variants in the binary case [37].

Approximation guarantees and the respondent’s side. For noiseless answers, greedy cost-sensitive selection has logarithmic approximation guarantees [9, 23]. With noise, information gain is not adaptive submodular, so these guarantees do not transfer directly. The equivalence-class edge-cutting objective of [24] is an adaptive-submodular surrogate with guarantees, and adapting it to three answers and complexity-dependent noise is a concrete next step.

The person is also rationally inattentive. With Shannon attention costs, her answers follow a logit rule [34]: she errs more often where the stakes appear low. This turns $\varepsilon(q)$ into a design lever. Explaining why a question matters lowers its error rate, at the price of a longer question.

D Representations of Intentions

The model does not fix what an element of Θ is. Three representations are natural, and each determines a natural class of questions.

R1. A point. θ is an element of a structured space, such as an assignment to the numbered elements of the standard. Questions ask whether θ lies in a subset: “Do you want your money back?”

R2. A set. $\theta = S \subseteq U$ is the set of outcomes in a universe U that the person finds acceptable. Questions are membership queries: “Would you accept a replacement?” Learning S is the classical setting of [1].

R3. A relation or utility. θ is a preference order on outcomes, or a utility function $u : 2^U \rightarrow \mathbb{R}$ over bundles of remedies. Questions ask whether two items are in relation: “Would you rather have a repair, or half of the price back?” When u is submodular, which captures diminishing returns across remedies, a substantial theory of learning applies [4, 10, 41].

Formally R2 and R3 are instances of R1, with Θ a family of sets or orders, but the admissible questions differ and with them the query complexity. Representations R2 and R3 also raise a further question: *what does a given representation mean*, and which legal claim does it stand for? We suggest that people can learn the meaning with the help of AI, by seeing what a similar representation has typically meant for other participants. Doing this requires a protocol and data shared by others (Section 9.2).

E Additional Results, Figures, and Tables

Operational rules. The elicitation algorithm of Section 7.2 motivates five operational rules for adapting the conversation to the person.

- (1) *Simplify when answers become unreliable.* When the estimated error scale $\hat{\phi}_i$ pushes a compound question past the thresholds above, the tool splits it (Corollary 7.5).
- (2) *Re-ask what surprised.* After an answer that conflicts with the prior or with earlier answers, re-asking about the same element is often the cheapest way to remove uncertainty, so the ratio rule chooses it (Table 4, Appendix B).
- (3) *Pause before stopping.* Fatigue lowers the value per unit of attention of every question. If the stopping rule fires while elements are still flagged, the tool offers a break and resumes later instead of filing an uncertain claim.
- (4) *Switch modality.* Applied to pairs of question and modality (text, voice, sign language), the ratio rule moves to the modality in which the person answers most reliably.
- (5) *Explain where the stakes are high.* People err more where the stakes seem low [34], so for remedies the tool says why the question matters, trading a slightly higher $c(q)$ for a lower $\varepsilon(q)$.

Error bound for certification. The following bound supports the certification analysis of Section 9.1.

PROPOSITION E.1. *Suppose that at stopping the posterior is calibrated, so that $\hat{F}$ is wrong with probability $P_e = 1 - \pi_h(f^{-1}(\hat{F}))$; that, conditional on $\hat{F}$ being wrong, the read-back detects the error with probability at least d , and a detected error is corrected or the filing withheld; that the LLM produces an unfaithful narrative with probability at most g ; and that, conditional on an unfaithful narrative, the faithfulness gate accepts it with probability at most p . Then the probability that what is filed misstates the claim, in the structured filing or in the narrative, is at most*

$$(1 - d) P_e + g p.$$

The proof, a union bound, is in Appendix A.

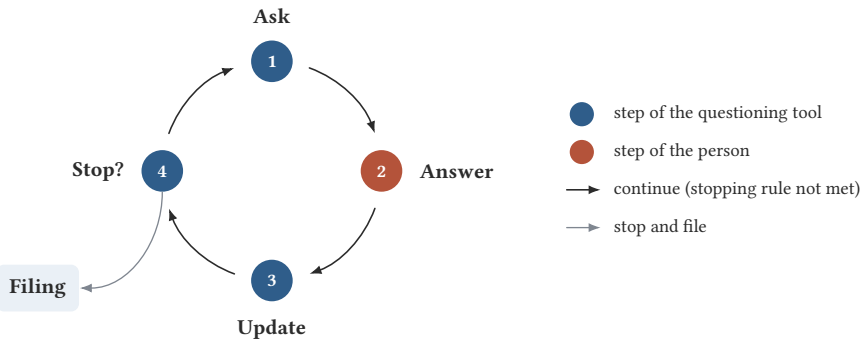


Fig. 4. The elicitation loop. (1) The tool asks the question q that maximizes information per unit of attention, $IG_h(q)/c(q)$ (Section 7). (2) The person answers $r \in \{Y, N, I\}$, wrongly with probability $\varepsilon(q)$, which grows with the complexity $\kappa(q)$ (Section 7). (3) The tool updates its belief π_h and the tension score $\tau(h)$, and re-asks an answer in a tension witness if needed (Section 6). (4) It stops when $\lambda IG_h(q) \leq c(q)$ for every q , that is, when the next question would cost more attention than the uncertainty it removes, and assembles the most probable filing (Section 9.1) together with the residual uncertainty.

Table 5. Part of the claim schema for Anna’s dispute under the Sale of Goods Directive [17], with derived questions.

Element	Legal basis	Generated question	Type
defect within 2 years	Art. 10	“Did the machine stop working within two years of delivery?”	fact
repair or replacement	Art. 13(2)	“Would you like the seller to repair or replace it?”	remedy
seller refused or failed (one ground)	Art. 13(4)	“Did the seller refuse, or fail, to repair or replace it?”	condition
price reduction	Art. 13(4)	“Would you accept part of the price back and keep the machine?”	remedy
termination	Art. 13(4)	“Do you want to return the machine and get all your money back?”	remedy
minimum amount	settlement	“Would you accept x euros?”	threshold

Table 6. Parameters of the elicitation algorithm and how each is set in a deployed tool.

	Meaning	How it is set in practice
$c(q)$	attention cost of a wording	estimated from length, negations, and legal concepts; calibrated by response times in pilots
$\varepsilon(q)$	error rate of a wording	test–retest: a random 5% of answers are asked again later; if the disagreement rate is D , then $\hat{\varepsilon} = (2 - \sqrt{4 - 6D})/3$, since two answers disagree with probability $2\varepsilon - \frac{3}{2}\varepsilon^2$; assumes independent repeats with equal error rates, which fatigue can violate
ϕ_i	the person’s error scale	prior from the self-assessment; updated in session from flipped answers and tension witnesses
λ	price of attention	set from a session budget ($\lambda = 8$ gives about 12 questions in our simulation); the person may choose a short or thorough mode
w_j	legal stakes of element j	set by lawyers, higher for remedies (<i>ne ultra petita</i>); used in an extension that ranks questions by the expected reduction in the weighted Bayes risk $\min_F \sum_j w_j \Pr(f_j(\theta) \neq F_j \mid h)$ per unit cost; Proposition 7.4 does not cover it
δ	flagging threshold	elements with confidence below $1 - \delta$ are read back and flagged for the court (Section 9.1)

PROPOSITION E.2. *If Θ is given as an explicit list, $\tau(h)$ can be computed in $O(n|\Theta|)$ question evaluations. If $\Theta = \{0, 1\}^m$ and each question is given as a Boolean formula, deciding whether $\tau(h) \leq k$ is NP-complete.*

The proof is in Appendix A.

COROLLARY E.3. *Under the assumptions of Proposition C.1, an adaptive strategy with n questions that always identifies the filing $f(\theta^*)$ requires $|f(\Theta)| \sum_{j=0}^k \binom{n}{j} 2^j \leq 3^n$.*

PROOF. Apply the argument of Proposition C.1 to one representative intention from each equivalence class. Leaves reached from different classes must be distinct. $\square$

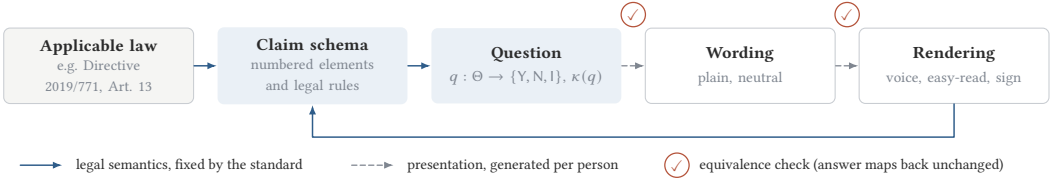


Fig. 5. Generating legally grounded questions. Legal meaning is fixed by the claim schema; only wording and rendering are generated per person, and each version must map back to the same element.

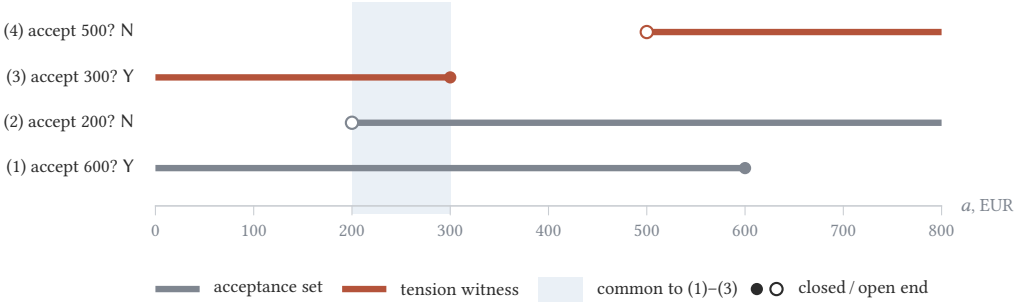


Fig. 6. Tension for threshold questions about a , the minimum amount Anna would accept. Each bar is the acceptance set of one answer to “Would you accept x euros?”. Answers (1)–(3) are consistent, with common region $200 < a \leq 300$, but no point lies in all four sets, so $\tau(h) = 1$. By Proposition 6.3 the conflict is explained by a pair, here answers (3) and (4). The system asks about one of them again instead of declaring Anna inconsistent.

Table 7. Legal counterparts of the formal objects.

Formal object	Legal counterpart
question q	a numbered element of the claim schema and the legal rule that makes it relevant (Sec. 8)
intention θ^*	the party’s claim: remedies sought (petitum) and supporting facts, as numbered elements of the statement of claim
answer l on a remedy	remedy not claimed before the court; open to settlement in mediation
threshold answers on a	settlement reservation value; confidential to the negotiation or mediation stage
filing $f(\theta^*)$	the statement of claim; legally equivalent intentions give the same filing
wrong answer	misstatement of the claim; may exclude a remedy (<i>ne ultra petita</i>)
tension score $\tau(h)$	reliability of the elicitation, <i>not</i> credibility of the party (Design Requirement 1)
tension witness	the specific statements to clarify with the party
stopping rule, residual uncertainty	the point at which clarification passes to the court’s duty to clarify
shared self-assessment	personal data, possibly health data under Art. 9 GDPR (Sec. 11)
strategic misreport	procedural bad faith; outside the scope of this paper (Sec. 5.1)