Trustworthy by design: data contribution, anonymisation, and quality control in an AI system for Dutch pro bono lawyers

Chenxi Deng, Hao Lin, Tiina Otala, Maurits van Kuijk, Silver Zweers

Abstract: Pro bono lawyers in the Netherlands face a structural knowledge gap: most legal case data is inaccessible or locked behind expensive platforms. and pro bono lawyers often work in isolation. A potential solution could be a community-controlled AI system where lawyers collectively contribute anonymised case documents to a shared knowledge base powered by retrieval-augmented generation (RAG); however, it raises critical questions about whether such a pipeline can be made trustworthy enough to drive adoption. This paper investigates design characteristics for a community-controlled legal AI system through a qualitative-dominant mixed-methods study. A technical feasibility evaluation confirms that (local) LLM-based anonymisation is viable, informing the design of prototypes evaluated with four Dutch pro bono lawyers. Qualitative sub-studies examine anonymisation interfaces, confidence score communication, semantic chunking, upload mechanisms, and peer validation. Findings indicate that adoption is encouraged by human-in-the-loop control, factual accuracy as a veto criterion, and predictable schema-based output structures, yielding actionable design principles for community-controlled legal AI.

Keywords: Community-controlled AI, retrieval-augmented generation, legal aid, knowledge contribution, human-in-the-loop, trust

1 Introduction

In the Netherlands, the right to legal aid is constitutionally guaranteed and delivered through a system of government-subsidised legal assistance (*gesubsidieerde rechtsbijstand*). Citizens who cannot afford legal representation may receive a government-funded legal aid voucher (*toevoeging*) that covers their lawyer's fees. Compensation for pro bono lawyers is determined through a points-based system, with each point intended to correspond to approximately one hour of legal work (Raad voor Rechtsbijstand, n.d.). Unlike commercial lawyers who bill by the hour, pro bono lawyers (*sociaal advocaten*) receive a fixed, case-based payment regardless of hours spent, creating a structural incentive for efficiency: tools that reduce time on legal research directly increase their capacity to serve more clients.

This system is under significant pressure. The number of active pro bono lawyers has declined by over 22% in the past five years, while demand remains stable at approximately 380,000 legal aid cases per year (Derksen et al., 2026). Over half (56%) of all practices are now one-person firms, and practitioners increasingly work in isolation (Derksen et al., 2026). This structural isolation is compounded by a significant knowledge gap: of the roughly 1.44 million court cases completed annually, only 4.4% are made publicly accessible through rechtspraak.nl (De Rechtspraak, 2025). The remaining case data, along with expert commentary and analytical overviews, is only available via paid platforms such as Wolters Kluwer and SDU, which large firms can afford but solo practitioners typically cannot (Aanbod voor advocaten en juristen | Lefebvre Sdu, n.d.).

A community-controlled AI system offers a potential response: pro bono lawyers could collectively contribute and structure case documents into a shared knowledge base powered by retrieval-augmented generation (RAG). RAG grounds AI outputs in curated external documents, enhancing training transparency and reducing the risk of hallucinations (Lewis et al., 2020). However, for such a system to function, large volumes of knowledge must be contributed, which raises critical questions about how data can be contributed safely, anonymised properly, and validated for quality by the community itself. When raw, unstructured legal texts are ingested without preprocessing, outputs may exhibit factual errors and semantic drift, undermining the trust that RAG was designed to provide (Bender et al., 2021; Shi et al., 2024).

This paper investigates what makes such a data contribution pipeline trustworthy by design. Through prototype testing and in-depth semi-structured interviews with four Dutch pro bono lawyers, we examine three dimensions of the contribution pipeline: (1) data contribution mechanisms: how lawyers are prompted to upload and annotate case documents; (2) anonymisation processes: how automation in identifying sensitive information affects trust; and (3) quality control: how semantic structuring and peer validation shape the perceived reliability of AI-generated outputs. We frame our inquiry around the following research question: How do the design characteristics of a community-controlled RAG system's data contribution pipeline influence the trust and willingness of Dutch pro bono lawyers to contribute to a shared legal knowledge base?

The findings yield actionable design guidance for future community-controlled legal AI systems, contributing to the broader discourse on trustworthy AI in professional domains.

2 Related work/Literature Review

Community-controlled AI refers to systems in which the people affected by a technology collectively shape its data, governance, and outputs, rather than these being determined by a centralised provider (Costanza-Chock, 2020). Such arrangements are particularly relevant where

existing infrastructures distribute knowledge or power unequally, and where a community holds situated expertise that external actors cannot adequately supply (Hsu et al., 2022).

A key element to fostering adoption, its intended users must trust it. User trust in AI-enabled systems is best understood relative to context, not as one universal definition. The factors influencing trust are dominated by user characteristics , and that trust can develop over time through repeated user-system interaction, although system design and socio-ethical factors also remain important. (Bach et al., 2024). Acceptance is not a single appraisal but a structured one. Davis's Technology Acceptance Model holds that user acceptance is shaped chiefly by two beliefs: perceived usefulness and perceived ease of use, which influence users' attitude or motivation toward using a system and, ultimately, system use (Davis, 1989), while more recent work on generative AI maps acceptance and resistance toward generative AI as distinct but interrelated outcomes, shaped by functional, risk-related, and sociolegal factors (Shrivastava, 2025). Consistent with this, AI adoption depends on employees' perceptions of risks and benefits, and that concerns around data security, accountability, transparency, and reliability should be addressed directly to build trust and support use. (Leutheuser et al., 2026).

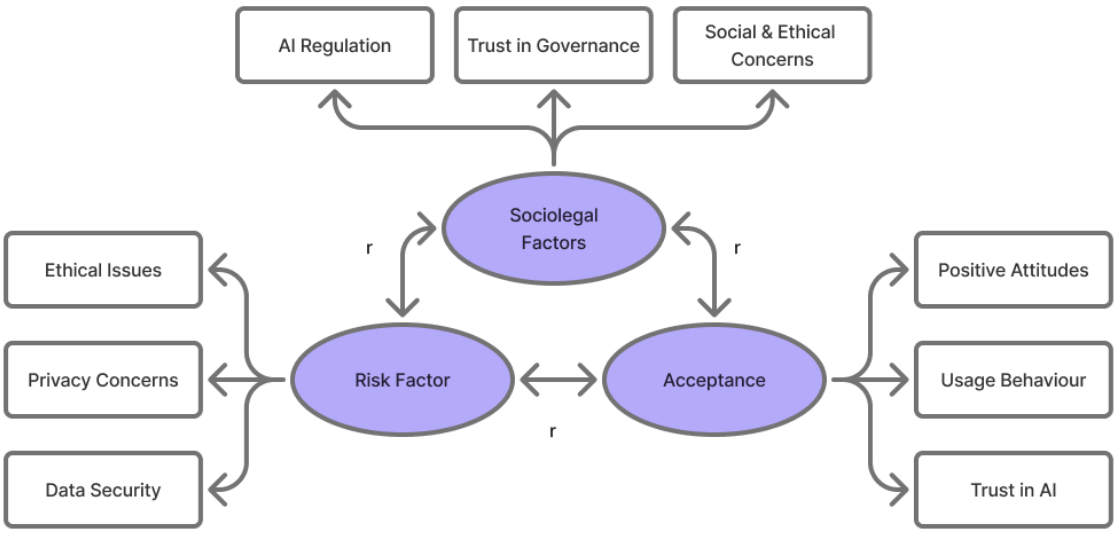


Figure 1: Our theoretical framework based on AI acceptance-resistance model (Shrivastava, 2025)

The technical foundation for a contributory legal system of this kind is retrieval-augmented generation (RAG), which grounds a model's responses in retrieved source documents rather than in its pre-trained parameters alone (Lewis et al., 2020). This properly suits legal work, where retrieval must identify precise and relevant legal text snippets, both to reduce hallucination risk and to support citation-based verification by the user (Pipitone & Alami, 2024). The benefit is conditional. Large language models can produce fluent and plausible but factually unsupported text. This hallucination problem is well documented and includes both outputs that contradict provided source material and outputs that cannot be verified from the source or from established knowledge. (Huang et al., 2023). Retrieval-based grounding can mitigate this problem, but only when the retrieved material is relevant and usable: irrelevant retrieval can introduce noise, and models use long contexts unevenly, often favouring information placed at the beginning or end of the input. This makes retrieval quality, ordering, and chunking of contributed text important design choices (Liu et al., 2024). The conditions under which the model can function therefore constrain the design of every stage that feeds it.

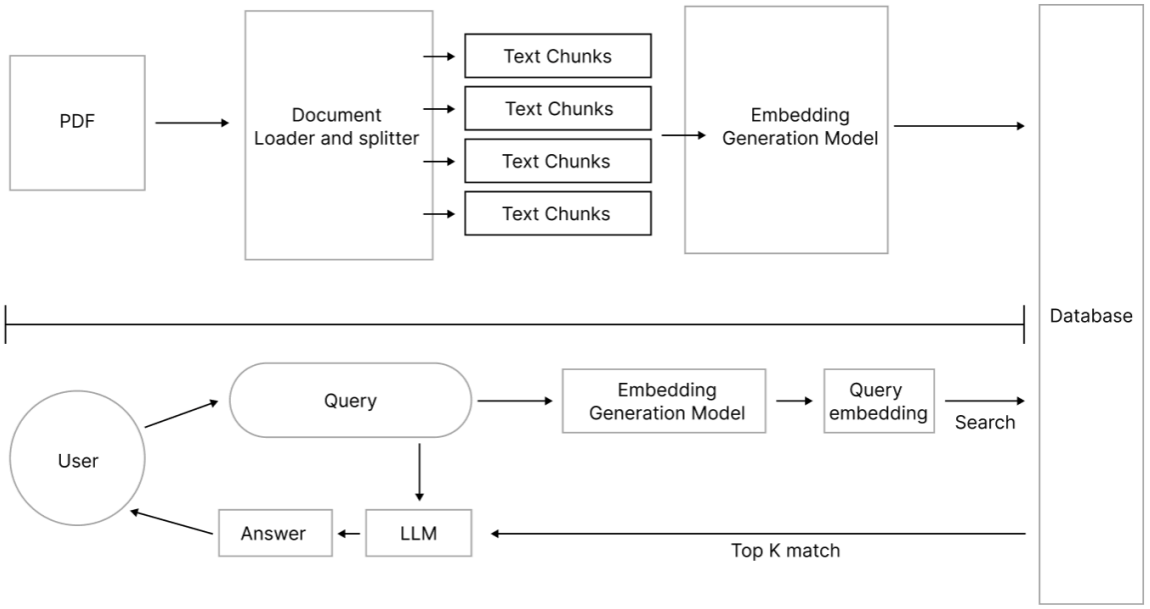


Figure 2: A standard RAG (Retrieval-Augmented Generation) pipeline. In the ingestion phase (top), a source PDF is split into text chunks by a document loader and splitter, which are then converted into numerical vector representations by an embedding generation model and stored in a database. In the retrieval phase (bottom), a user submits a query, which is similarly converted into a vector embedding and used to search the database for the most semantically similar chunks (Top K matches). These retrieved chunks are passed, alongside the original query, to a large language model (LLM), which generates a contextually grounded answer returned to the user.

The first of these stages is anonymisation, which should take place before any upload to the cloud where identifiable or confidential material is not necessary for the task. Under the GDPR, controllers and processors must implement technical and organisational measures appropriate to the risk, including pseudonymisation and encryption(European Parliament & Council of the European Union, 2016). Dutch case material is governed by the Rechtspraak anonymisation guidelines, which require the removal of names, addresses, and other directly or indirectly identifying details before a ruling may be published (Anonymization Guidelines | Dutch Judiciary, n.d.). The difficulty is that confidentiality is breached not only through explicit identifiers but through semantic inference, so adequate anonymisation must reason about context and be assessed with metrics able to capture that subtlety (Lison et al., 2021). Earlier rule-based or pattern-matching NER approaches can be effective in narrow legal contexts but lack scalability and adaptability. In contrast, domain-adapted transformer models such as LegNER use legal-domain pretraining, span-level supervision, and vocabulary adaptation to capture nuanced legal expressions, multi-word entities, and rare domain terms more reliably, thereby improving entity recognition and anonymization performance. (Karamitsos et al., 2025). Anonymisation is thus built directly upon what the model requires: the contextual capacity that makes a model useful is what makes it an effective anonymiser.

Once confidentiality is secured, willingness to contribute becomes a determinant of trust in its own right. Contribution in professional communities is driven less by direct reciprocity than by the reputational returns of being seen to contribute, a pattern observed directly within a legal professional network (Wasko & Faraj, 2005), and is further shaped by social-psychological factors and the expectations of colleagues (Bock et al., 2005). It is also fragile: when Wikipedia tightened quality control to protect article quality, it deterred the very newcomers it depended upon (Halfaker et al., 2013). How contributions are validated is therefore consequential. Warranting theory suggests that observers place more trust in information that is difficult for the target to manipulate. In online settings, other-generated information can therefore be seen as more credible than self-generated claims because it is less directly controlled by the person or

organization being evaluated (Walther et al., 2009), a property that can be treated as a graded warranting value rather than a binary one (DeAndrea, 2014). Validation by peers can function as a reputation signal in peer-to-peer knowledge communities, because reputation is commonly used as a collective indicator of trustworthiness based on ratings or referrals from other members. However, such mechanisms are vulnerable to manipulation, especially in virtual communities where users may create dummy profiles or trade positive evaluations (Havakhor & Sabherwal, 2013).

Because these mechanisms operate through what the contributor sees, they are realised at the level of the interface, and the features intended to build trust must be legible to the lawyers using them. Sundar's MAIN model explains how users rely on technological cues and heuristics when judging credibility, including an authority heuristic in which information attributed to an official or expert source is perceived as more credible (Sundar, 2008). Transparency about why a system produces a given output can strengthen trust in retrieval-augmented systems. Ravi and Sindhgatta show that, for financial domain experts, confidence scores alone did not significantly increase trust, whereas source attribution and highlighted document sections improved both trust and understanding. This supports surfacing the confidence, source evidence, and validation behind outputs rather than concealing them (Ravi et al., 2025). A feature whose purpose is not understood cannot perform its reassuring function, however sound its underlying design.

These requirements converge on a single principle: trustworthiness depends on combining automation with human oversight rather than choosing between them. Shneiderman (2020) framework of human-centred AI holds that high automation and high human control are not competing quantities, but can be raised together to produce systems that are reliable, safe, and trustworthy. He distinguishes between recommender, consequential, and life-critical applications, and places legal systems within the consequential category: systems that may bring substantial benefits and harms, but where decision makers often have time to reflect, consult colleagues, and probe recommendations more deeply. Legal document handling is therefore well suited to designs that pair strong automation with meaningful professional control. This principle governs the pipeline as a whole: at each stage where a lawyer's trust is won or lost, from anonymisation through contribution to the interpretation of output, automation must remain visible, reviewable, and answerable to professional judgement.

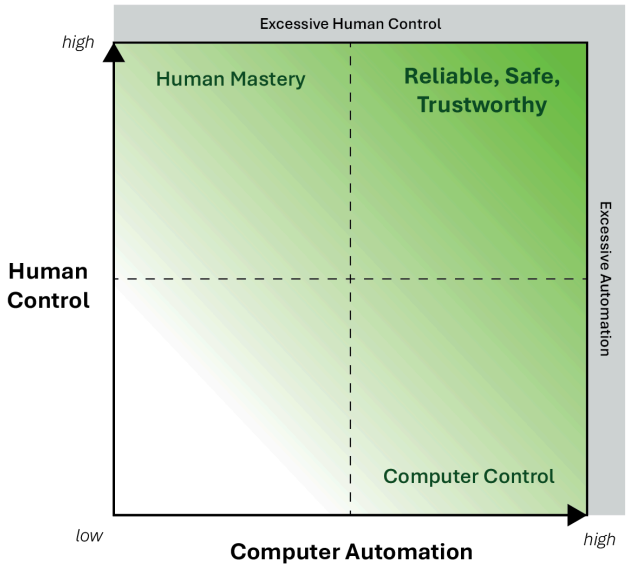


Figure 3 Original two-dimensional framework of human control and computer automation. Adapted from "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," by B. Shneiderman, 2020, International Journal of Human–Computer Interaction, 36(6), p. 498. (https://doi.org/10.1080/10447318.2020.1741118).

3 Methodology

3.1 Research Paradigm

This study adopts an interpretivist paradigm (Creswell & Poth, 2018), treating lawyers' trust in AI as a subjective, context-dependent perception rather than an objective, measurable quantity. Knowledge is co-constructed through direct interaction with participants and analysed through reflexive thematic analysis (Braun & Clarke, 2019). This orientation governs the qualitative sub-studies, which concern human perception and professional judgement. A complementary technical benchmarking experiment employs quantitative metrics to assess pipeline performance.

3.2 Research Approach

This study follows an inductive research approach. Rather than testing hypotheses derived from existing theory, we move from specific observations to broader generalisations: participant responses are coded, patterns are identified, and themes emerge that inform design principles.

The inductive process unfolds in four stages:

1. Technical benchmarking: A sub-study tests the assumption that LLM-based pipelines are viable for legal anonymisation, which is a prerequisite for the system's feasibility and informs possible interfaces for the next stages.
2. Data collection: Four sub-studies gather rich qualitative data through prototype interaction (think-aloud protocol) and post-interaction interviews.
3. Pattern identification: Transcripts are coded using reflexive thematic analysis (Braun & Clarke, 2006, 2019), an iterative method that moves from initial codes to candidate themes to refined themes.
4. Theory dialogue: Existing theoretical lenses are used to interpret emergent themes. These include the AI acceptance-resistance model(Shrivastava, 2025). The purpose is not to test these theories, but to enrich the analysis.

This research approach is appropriate to explore an innovative design space: community-controlled AI for legal professionals. While existing theories can provide a useful framework, they cannot predict the specific design characteristics that will engender trust.

3.3 Research Method

Prior to the qualitative sub-studies, a quantitative technical benchmarking experiment compared three anonymisation pipeline configurations: rule-based NER, a locally-deployable LLM, and a hybrid of both against an annotated ground-truth dataset. The dataset consisted of five PDF documents, which were real-life anonymized court documents that were subsequently reverse-anonymized for testing purposes. Performance was quantified using precision, recall, and F1 score, with recall prioritised as the critical metric given the asymmetric risk of missed identifiers. Results informed the design of the anonymisation prototypes evaluated in the qualitative sub-studies.

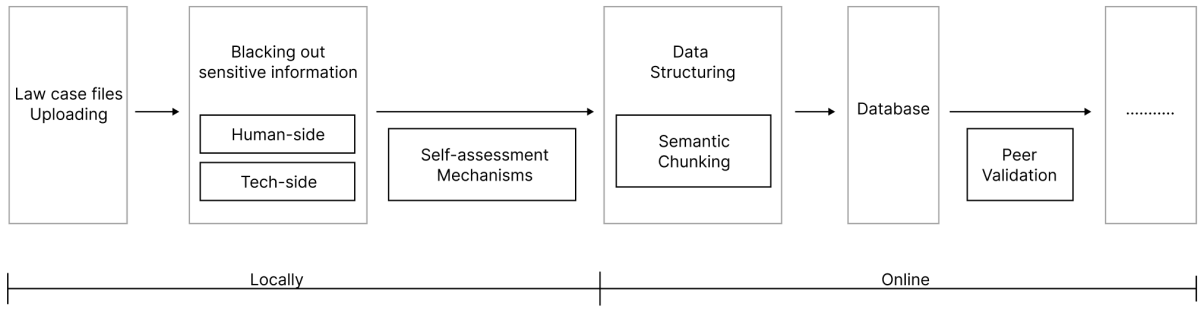


Figure 4: Document ingestion pipeline for the community-controlled RAG system. Law case files are uploaded and locally anonymized before being structured and processed online, then stored in a cloud database for retrieval.

The experiment was structured as a modular, workflow-based design: each researcher owned a thematically distinct sub-study investigating one stage of the data contribution pipeline (figure xx), and the sub-studies were sequenced to mirror the user journey through the RAG platform, from data contribution (anonymisation and upload) to output evaluation (summary quality and chunking), and finally to community trust (validation). Each module is independently testable yet connected to the others through shared participants, a common analytical method, and a unified research question. The structure allowed for investigation of the full contribution pipeline while maintaining methodological coherence.

This study employs a qualitative-dominant mixed-methods design. The primary data are qualitative (interview transcripts, think-aloud verbalisations, open-ended responses). Quantitative measures (Likert scales, categorical preferences) provide descriptive support and triangulation but are not used for statistical inference due to the small sample size.

Data collection combines two primary techniques grounded in established HCI methodology. Think-aloud protocol (Ericsson & Simon, 1993) is used during prototype interaction: participants verbalise their reasoning in real time as they evaluate anonymization tools, summaries, upload interfaces, and validation mechanisms. Semi-structured interviews (Flick, 2022) follow each interaction, allowing the researcher to probe emergent themes while maintaining comparability across participants. Low-fidelity prototypes function as design probes (Gaver, Dunne & Pacenti, 1999), which are artefacts designed to elicit thoughtful responses regarding values, preferences and professional norms.

Table 1. Four sub-studies investigate the following stages

Sub-Study	Focus	Data Collection	Analysis
Anonymisation: technical benchmark	Feasibility of LLM-based anonymisation vs. NER	3 pipeline configurations × annotated ground-truth dataset (5 documents)	Quantitative: precision, recall, F1 score
Anonymisation: confidence scoring	Confidence score communication informed by benchmark findings	3 confidence display variants × 3 lawyers; think-aloud +	Semi-structured interview

		semi-structured interview	
Anonymisation: Automation levels	Automation levels in AI-assisted anonymisation	3 interface conditions × 4 lawyers; think-aloud + semi-structured interview	Semi-structured interview
Semantic Structuring	Chunking methods and summary quality	3 summarisation conditions, 2 chunking methods × 4 lawyers; Likert ratings + failure annotation + interview	Mixed: descriptive statistics + thematic analysis
Contribution Interface	Upload mechanisms and willingness to contribute	3 interface conditions × 4 lawyers; think-aloud + 5-item, 7-point Likert scale on Knowledge Sharing Intention (KSI) subscale of Usman (2018) + interview	Mixed: descriptive statistics + thematic analysis
Validation Mechanism	Peer and institutional validation	3 validation conditions × 4 lawyers; think-aloud + interview	Reflexive thematic analysis

All sub-studies share a common analytical method; **reflexive thematic analysis** (Braun & Clarke, 2006, 2019) enabling cross-study synthesis. Two sub-studies utilized Likert scales for quantitative measures of agreement or disagreement with statements (Joshi et al., 2015). Participants were also asked to do failure annotation in AI-generated summaries, following a taxonomy grounded in factuality research (Pagnoni et al., 2021). It provided structured diagnostic data beyond overall quality ratings, and it prompted participants to articulate the specific criteria underlying their trust judgements.

Participants were recruited through **purposive sampling** (Patton, 2015), selected for variation in legal specialisation, firm size, and AI experience. Sample sizes (n=5) are consistent with qualitative research norms, where the goal is depth of understanding rather than statistical generalisability (Nielsen & Landauer, 1993). Participants 1-4 were interviewed for four sub-studies: anonymisation (automation levels), semantic structuring, contribution interface, and validation mechanism. Participants 4, 5 and 6, 5 and 6 were interviewed for the anonymisation (confidence scoring) sub-study.

Table 2. Summary of research participants

Participant identifier	Legal experience	Legal specialties	Firm size	Prior AI tool usage
1	25 years	· Administrative *(bestuursrecht)* · Environmental *(omgevingsrecht)* · Migration *(migratierecht)*	~60 (only pro bono lawyer in firm)	· Microsoft Copilot · Zeno · GenIA-L
2	12 years	· Administrative *(bestuursrecht)* · Criminal *(strafrecht)*	1 (works alone, but shares office space with other pro bono lawyers)	· Microsoft Copilot · LegalMike
3	8 years	· Criminal *(strafrecht)* · Juvenile criminal *(jeugdstrafrecht)*	1	—
4	27 years	· Criminal *(strafrecht)* · Tenancy *(huurrecht)*	1	· Andri AI · LegalMike
5	4 years	· Privacy *(Privacyrecht/ gegevensbescherming)*	~50 (mix of pro bono and commercial law)	· Copilot · Harvey · LegalMike
6	2 years	Legal consult *for a legal AI startup*	~25 (6 of them with a law degree)	· Copilot · Legora · LegalMike · Claude
6	2 years	Legal consult for a legal AI startup	~25 (6 of them with a law degree)	· Copilot · Legora · LegalMike · Claude

3.4 Research Design

Each participant session lasted approximately 60–90 minutes and followed a standardised sequence mirroring the user journey through the RAG platform (Table 3). All sessions were audio recorded and transcribed. Data were analysed using reflexive thematic analysis (Braun & Clarke, 2006, 2019).

Table 3. Interview procedure

Phase	Duration	Activity	Sub-Study
1. Introduction & Consent	5 min	Researcher explained the study purpose, obtained written informed consent, and requested permission for audio recording	—
2. Background Interview	10 min	Questions on legal specialisation, years of practice, current AI usage, and attitudes toward technology	Cross-cutting
3. Anonymisation Prototype Testing	15–20 min	Participants interacted with three anonymisation variants (manual, semi-automatic, fully automatic) while thinking aloud	Anonymisation
4. AI Summary Evaluation	15 min	Participants read three summaries of the same case, rated each on Likert items, annotated failures, and stated preferences	Semantic Structuring
5. Upload Interface Testing	10 min	Participants interacted with three upload interface variants while thinking aloud	Contribution Interface
6. Validation Mechanism Evaluation	10 min	Participants viewed three validation variants and stated their preference and reasoning	Validation Mechanism
7. Post-Evaluation Questions	5 min	Contribution willingness, trust factors, driving factors, and open feedback	Cross-cutting

3.5 Design probes

Prior to the primary qualitative evaluation, a comparative study was conducted to assess the efficacy of Large Language Model (LLM) architectures against Named Entity Recognition (NER) algorithms. This sub-study addressed the risky assumption that LLMs represent a superior foundation for anonymisation pipelines due to their capacity to identify indirect personal identifiers. The assessment involved the systematic comparison of three distinct pipeline configurations: a traditional NER-based algorithm, a standalone LLM (Gemma4; 12B parameters), and a hybrid model integrating both technologies. The performance of these pipeline configurations was quantified through standard evaluation metrics, specifically assessing their efficacy in terms of F1 scores, precision, and recall.

Table 4. The experiment used 6 categories of low-fidelity prototypes, which can be seen in the Appendix. Low-fidelity prototype variants by sub-study. Each of the six prototype categories was presented to participants in randomised or counterbalanced order, with variants differing along a single design dimension, ranging from manual control to full automation in anonymisation, absent to categorical confidence display, granular to preset

processing control, unstructured to schema-based summarisation, baseline to annotated upload, and no validation to institutional validation.

Prototype	Duration	Description
Anonymisation prototypes	Variant 1a	Manual input (full user control)
	Variant 1b	Auto pre-select with editable suggestions (balanced)
	Variant 1c	Fully automatic (maximum automation)
Confidence scoring prototypes	Variant 2a	Absent (no indication of reliability)
	Variant 2b	Numerical (percentage-based probability)
	Variant 2c	Categorical (qualitative levels: possible, likely, and very likely)
Control prototypes	Variant 3a	Granular token processing control (enabling manual definition of scan intervals)
	Variant 3b	Preset-based interface (choice between a comprehensive document scan or incremental semantic chunking, with intermediate findings displayed during processing)
AI-generated case summaries and chunking method comparison	Variant 4a	Unstructured
	Variant 4b	Schema-Based semantic chunking with predefined legal fields
	Variant 4c	Structure-Aware chunking following the document's own section order
Upload interface prototypes	Variant 5a	Plain drag-and-drop baseline with no self-assessment
	Variant 5b	Checkbox condition offering structured prompts to guide reflection
	Variant 5c	Free-text condition inviting contributors to annotate on their files
Validation mechanism prototypes	Variant 6a	No validation (baseline, no review)
	Variant 6b	Peer validation (review by identified colleague lawyers)

Variant 6c	Institutional validation (review by an editorial authority)

The cases used as examples in the experiment were either provided by the lawyers themselves or carefully selected from publicly available court rulings. In cases where lawyers provided their own documents, they were already highly familiar with the content, which not only lended high credibility to their quality judgements of the AI-generated summaries, but also meant that their responses across all experiment stages (anonymisation, upload, and validation) more closely simulated the real-world usage scenario of a lawyer contributing their own case material to the platform.

Table 5. Ethical Considerations. The study followed the ethical guidelines of TU Delft's Human Research Ethics Committee (HREC). Given the sensitive nature of legal practice, several additional measures were taken

Phase	Activity
Informed consent	All participants signed a consent form before the experiment, with the right to withdraw at any time without giving a reason.
Anonymisation of participants	All participant data was anonymised; no personally identifiable information is published. Participants are referred to by pseudonyms.
Anonymisation of case data	All test cases were sourced from publicly available court rulings (Rechtspraak.nl) to avoid handling confidential client information during the experiment.
Data security	Audio recordings were stored securely and deleted after transcription.
No deception	Participants were informed that they would evaluate AI-generated summaries. The specific processing methods were revealed only during debrief to avoid biasing their evaluations.
Professional sensitivity	The researcher was mindful that legal professionals may have heightened concerns about confidentiality; participants were explicitly assured that their responses would not be shared with employers, bar associations, or any third party.

4 Findings

4.1 Results

Across all four participants and all prototype dimensions, a consistent pattern emerged: trust was not expressed as a single judgement but as a layered evaluation. Participants varied in their current AI usage. Three participants used a combination of legal and general AI tools daily, while one participant had tried ChatGPT and abandoned it due to perceived unreliability.

Anonymisation

Technical Performance. The LLM-based pipeline demonstrated superior recall (87%), significantly outperforming the traditional NER approach (72%) on the metric most critical to legal confidentiality: the prevention of false negatives that could expose sensitive identities. Furthermore, the LLM achieved the highest precision (91%), surpassing both NER (48%) and the hybrid configuration (54%), resulting in a leading F1 score of 89%. While NER exhibited high velocity (processing in under 10 seconds), it failed to capture quasi-identifiers requiring semantic context. The hybrid model yielded the peak recall (89%) and intermediate precision (54%), with an F1 score of 67%; critically, it enabled a proxy for confidence scoring by identifying tokens flagged by both systems as high-confidence. The successful execution of the full pipeline on 4GB VRAM confirms technical feasibility for local, privacy-preserving deployment on standard professional hardware.

***Confidence Scoring and Agency*.** Evaluation of three confidence display variants absent, numerical, and a 3-level categorical scale revealed a universal preference for categorical labels (possible, likely, very likely). Participants dismissed numerical percentages as lacking semantic relevance for professional workflows. Similarly, when presented with granular controls over LLM technical parameters such as chunking size, lawyers expressed that such distinctions lacked utility, preferring an intuitive, streamlined interface that balances automation with meaningful oversight.

Level of automation. An initial observation was that participants varied in their current methods of anonymisation: Participant 1 used a PDF editor to completely erase information, Participants 2 and 4 drew black boxes over information, and Participant 3 had minimal experience with anonymising. All four participants had anonymised documents by hand. Reasons for anonymising included education (e.g. guest lectures) and critical analysis with peers. All four participants chose variant 1b (balanced automation and control) as their preferred interface.

The thematic analysis yielded three themes, eight codes, and 52 quotes (Figure 4). AI-related concerns included the risk of errors, fear of skill erosion, and uncertainty about products' data protection terms (e.g. Microsoft Copilot, LegalMike). Design considerations for an anonymisation tool included always having some degree of control to make edits, interface visual responsiveness to track changes between original documents and AI-assisted edits, and awareness that participants categorize even high automation solutions as simple "tools" if AI makes errors. Finally, all four participants emphasized the necessity of double checking all of the AI's work regardless of AI ability or automation level. The participants described their trust as provisional, with Participant 2 stating, "it has to pass 100% for me to be comfortable, over and over…If I find one mistake after four times of 100% passes, and then the fifth time is 99%, [my trust] will drop."

THEME	CODES	FREQUENCY (# of quotes)	DISTRIBUTION (participants quoted)			
			P1	P2	P3	P4
Concerns	Fear of skill erosion	5				
	Potential errors	7				
	Uncertainty about AI data protection terms	9				
Design considerations	Ability to edit	4				
	Visual responsiveness	5				
	Categorization as intelligence or tool	6				
Trust through verification	Necessity of checking	12				
	Provisional trust	4				

Figure 5: Themes and codes with code frequency and distribution (anonymisation sub-study)

Theme	Subtheme	Code	#	Distribution among participant			
				P1	P2	P3	P4
Belonging *Understanding the community and identity in legal work*	Professional identity	Awareness of self-check	5				
		Confidencial concerns	4				
		Client's rights/benefits	4				
	Community Awareness	Community Awareness	5				
		Data sharing in exsiting workflow	5				
Source *Data quality matters and needed to be transparent*	Trust	Trust on data transparency	6				
		Value/quality of data source	6				
	Sharing barrier	Exsiting tool/data base sufficient	3				
		Self data quality	5				
Intergration *The interface should fit in daily legal work*	Efficiency	Version / identification clarity	3				
		Annotation guidelines	3				
		Uploading speed	4				
	Clearity	Hybrid interface preference	2				
		Context relevance	3				
		Document-type categorisation	3				

Figure 6: Themes and codes with code frequency and distribution (data contribution sub-study)

Quality Control: Summary Evaluation and Chunking

When evaluating three AI-generated summaries of the same legal case, all participants identified factual errors, including misidentified genders, incorrect conviction counts, and omitted damages claims. One participant described a summary as "just Wikipedia level summary for non-legal experts... too many fatal errors." Another noted that a summary missed the core legal dispute entirely, rendering it unusable despite being factually partially correct.

When comparing two chunking methods, three of four participants preferred Schema-Based (fixed legal fields) over Structure-Aware (AI-determined keywords). The primary reason was predictability: "When it's a fixed structure, I always know where to go to find relevant information." One participant suggested that fixed structures should be adapted by document type, court rulings, legal advice, and appeals each warrant their own template. Additionally, all participants expressed the desire to manually edit or correct AI-generated summaries before finalising them , the ability to amend factual errors or add missing information was considered essential for the system to be usable.

Data Contribution Mechanisms

Of the three upload conditions, participants preferred the free-text field as it captured richer, more context-specific information about their files. The checkbox condition was the least popular, with participants noting that its fixed prompts were not well suited to different document types.

Beyond the individual conditions, all participants agreed that the upload process should include a categorisation mechanism to create a clearer shared understanding between contributing lawyers. They also stressed the importance of efficiency, observing that the process must not become so complex as to disrupt a legal aid lawyer's existing workflow.

Peer Validation

All three participants located trust in the source itself rather than in the validation layer. Court rulings were treated as self-validating and the unvalidated baseline was judged sufficient to use; one participant reported that validation gave "maybe not the trust, but the usability," and another that it "doesn't change anything for me." All three distinguished substantive validation from mere approval. Expert annotation was recognised through the existing note on a ruling ("it's quite familiar... to have some notes from a specialist") and valued as reusable ("I can use it in my own statements, like this professor said this"), with weight tracking author seniority. Bare peer signals were dismissed by two participants: "thumbs up, thumbs down, I don't think that adds much." One participant noted peer approval could be an adverse signal in adversarial cases: "what if the other one likes it because it's bad for your client?"

4.2 Findings

The results directly address the research question of which design characteristics make the contribution pipeline trustworthy, examining this across all four dimensions.

Anonymisation

A comparative evaluation of Named Entity Recognition (NER) and Large Language Model (LLM) architectures demonstrates that LLM-based anonymisation is not only technically feasible but preferred for its superior recall and capacity to identify indirect personal identifiers. This finding aligns with recent research on domain-specific legal models (Karamitsos et al., 2025). However, think-aloud interactions and scoring revealed that pro-bono lawyers preferred rough confidence scoring (3 levels of confidence), a feature currently limited by the next-token prediction nature of LLMs. To reconcile these requirements, a hybrid pipeline integrating NER and LLM outputs appears optimal, leveraging algorithmic agreement as a proxy for confidence to ensure the human-in-the-loop oversight that professionals preferred.

Analysis of lawyer preferences indicated a desire for granular agency over technical parameters, such as scanning chunk size; however, the distinction between local and cloud-based deployment models appeared irrelevant to their perception.

The universal preference for human-in-the-loop anonymisation resonates with Shneiderman's (2020) human-controlled AI framework and Lee and See (2004) automation-control tradeoff. Desire for human-in-the-loop was echoed in the results of the automation level sub-study: across all four participants, there were consistent requests for the user to be able to edit (take control) and a repeated emphasis on the need to double check and pay attention to AI outputs. Interface design could reinforce the feeling of control by making AI edits live and transparent, as well as leaning into the perception of automation as a simple tool rather than as an intelligence equivalent to a human.

Quality control

Factual accuracy acted as a veto criterion: any factual error, such as misidentifying gender or providing an incorrect conviction count, was enough to cause participants to reject the entire summary, regardless of its other qualities. This establishes a clear hierarchy: the system must prioritise accuracy before optimising for any other dimension.

To ensure this level of quality, specialised chunking processing tailored to legal documents is necessary; to build trust, demonstrating to lawyers that the AI tool has correctly understood their legal documents is highly effective. The preference for schema-based chunking is explained by Rumelhart's (1980) schema theory, which states that predictable structures reduce cognitive load and enable fast navigation. However, the suggestion to adapt schemas by document type indicates that a one-size-fits-all structure is insufficient.

A further finding concerns lawyers' tolerance for length: unlike consumer-level tools where longer text reduces user patience, lawyers prioritise substance over brevity. So, if a summary contains important, accurate, and clearly structured information, length is not a deterrent. Moreover, the comparison of different presentation formats revealed that presentations with a clear structure, divided into sections with labelled, schema-based headings, were favoured over a single, continuous block of text.

Data Contribution mechanisms

A lawyer's sense of belonging to a legal practice influences their trust in governance and ethical concerns, which affects their willingness to contribute. Those working in law firms are more open to adopting community-based AI systems, whereas independent lawyers prefer to rely on existing AI tools. While sharing case files among colleagues is common practice, some lawyers are reluctant to share with unfamiliar attorneys who may represent opposing parties in future cases.

Another factor is the source of the contributed data, which can lower the barrier to sharing by strengthening the perception of data security. Transparency about who authored a case file is essential, as lawyers use this information to evaluate the quality and context of the material they work with. Lawyers were also more willing to share complex cases, which they considered more valuable to the shared knowledge base than routine ones, as the perceived value of their own files influenced their WTC.

A further barrier to sharing concerned whether a shared database would offer enough beyond the resources lawyers already control, together with their confidence in the quality of their own data. Participants reported working primarily from their own past case files, turning to the court library or to colleagues only when they needed more. Because their own material was already substantial and trusted, some found an office-level tool more appealing than a shared community database. As one participant explained: "I would rather have a localized AI which goes through all my own cases or my office's case, because we already have a lot of data here.”

Peer validation

Source validation operated as a usability layer rather than a foundation for trust. Across participants, trust was anchored in a source's standing within the legal hierarchy rather than in any platform endorsement: court rulings were treated as self-validating and the unvalidated baseline was judged sufficient to use, with one participant noting validation added "maybe not the trust, but the usability." This reflects an authority-based credibility heuristic: trust is shaped by the perceived authority of the institutional source, rather than by the interface itself (Sundar, 2008). In this case, the community RAG system appears to inherit credibility from the legal materials it retrieves, rather than creating that credibility on its own.

A consistent distinction emerged between substantive validation and mere approval. Expert validation was readily accepted because it reproduces an established professional convention, the annotated note on a published ruling, and was valued as directly reusable: "I can use it in my own statements, like this professor said this." Its weight tracked the author's seniority, recognised specialists carrying epistemic authority that unfamiliar names did not. Peer validation through likes or dislikes was dismissed as adding little, since aggregated approval has no analogue in how legal knowledge is conventionally authorised.

Peer validation was further weakened by the adversarial and case-specific nature of legal work: because relevance is bound to the individual matter and parties are opposed, a colleague's approval can act as an adverse rather than a quality signal ("what if the other one likes it because it's bad for your client?"). Consequently, willingness to contribute was high but largely independent of the validation model. What conditioned it was not validation itself but the prospect of being evaluated; consistent with reputation and evaluation concerns in knowledge-sharing networks (Wasko & Faraj, 2005), participants separated contributing knowledge from submitting their work to peer judgement, a distinction developed in the data contribution findings. This preference for substantive, human-authored input over automated or crowd signals aligns with human-in-the-loop principles of professional AI (Shneiderman, 2020)

Collectively, these findings suggest that trustworthiness in a community-controlled legal AI system is not achieved through any single feature but through the alignment of contribution friction, privacy assurance, and output quality with lawyers' professional norms and time constraints.

5 Discussion

5.1 Interpretation

The findings converge on a single principle: across all aspect, trust depends on preserving the lawyer's professional judgement rather than displacing it.

The preference for hybrid, human-in-the-loop anonymisation is best read through Shneiderman's (2020) two-dimensional model of human-centred AI, which argues that high automation and high human control are not opposites but can be maximised together. Legal anonymisation is a consequential task: errors are serious but reversible, so it sits in the quadrant where high automation paired with high control produces reliable, safe, and trustworthy systems. Participants' worry that AI misses context-dependent identifiers, a place name harmless in a city but identifying in a village, explains why they accepted AI suggestions but refused to surrender the final decision. Confidentiality is a professional obligation that cannot be delegated, so automation is welcome only as an assistant, not an authority.

The veto power of factual accuracy is better understood through Shrivastava's (2025) AI acceptance-resistance model, which treats acceptance and resistance as distinct forces shaped by functional, risk, and sociolegal factors. A factual error not only weakens the functional benefit, but also activates the risk dimension, causing lawyers to actively resist. Consequently, trust does not decline gradually, but collapses. This veto reflects the primacy of professional judgement seen elsewhere: lawyers treat the system as something they must verify rather than an authority to defer to. Therefore, a single error that survives their scrutiny marks the tool as unreliable, resulting in the withdrawal of trust.

The preference for schema-based chunking over document-structure-aware chunking points to the same underlying value. Fixed legal fields (parties, facts, issue, reasoning, ruling) map onto the mental model lawyers already use to read a case, which lowers the cognitive cost of verification and makes outputs predictable. Predictability matters precisely because lawyers intend to check the output: a structure they can anticipate is a structure they can audit. The simultaneous demand for customisable schemas shows this is not a request for rigidity but for a predictable frame that still adapts to document type and specialisation.

Source validation findings highlight professional autonomy: participants trusted the inherent authority of primary legal documents over platform-specific endorsements. Specialist commentary was valued only when aligned with academic norms, reinforcing Sundar's (2008) authority-based credibility heuristic where trust stems from institutional origins. Conversely, peer ratings were dismissed as lacking traditional legal standing and potentially acting as negative indicators in adversarial contexts. Per Shrivastava's (2025) model, these social signals can trigger resistance by exposing lawyers to unwanted scrutiny without improving utility, explaining why practitioners decouple knowledge contribution from formal peer review.

Together these patterns extend the notion of trust by design. Trust here is earned less through technical robustness than through a pipeline that keeps the professional in control, holds accuracy as a non-negotiable threshold, and presents information in forms practitioners can predict, verify, and correct.

5.2 Proposed Applications

The findings support the feasibility of an open-source legal research tool for Dutch pro bono practitioners, and participants were encouraged about full implementation. Beyond the immediate context, the contribution pipeline designed in this study has potential relevance in any domain where professionals handle sensitive documents under time pressure and privacy obligations; most notably healthcare, where GDPR-compliant anonymisation of patient records faces similar challenges.

Within the legal domain, participants identified adjacent applications during think-aloud sessions: legal template generation and automated population of standard forms were mentioned as natural extensions of the same infrastructure. These use cases share the same underlying requirement; structured, anonymised, professionally validated documents, suggesting that the pipeline architecture developed here could serve as a foundation for a broader legal practice toolkit rather than a single-purpose research tool.

The current prototypes relied on general-purpose LLMs. Now that the pipeline architecture provides a reasonable proof of concept, a compelling next step is domain-specific model specialisation through reinforcement learning from human feedback. Domain-adapted models such as LegNER (Karamitsos et al., 2025), a transformer built specifically for legal named entity recognition and text anonymisation, indicate that such specialisation is both feasible and valuable in the legal domain. Lawyers are uniquely positioned to provide high-signal corrective feedback, they can identify not just that an output is wrong, but why it is wrong in legal terms. Designing

interfaces that capture this feedback efficiently, without adding burden to an already time-pressured workflow, represents a significant and largely unexplored design research opportunity. Concretely, this could take the form of in-line annotation affordances on AI-generated summaries, structured error categorisation, or lightweight approval flows that simultaneously serve as training signals. The critical design challenge is that such interfaces must feel like quality control, something the lawyers already do, rather than unpaid AI training labour.

5.3 Limitations / future research

Several limitations should be acknowledged. The sample size across sub-studies (n=3–4 per sub-study) limits statistical generalisability, though the depth of 60–90 minute sessions provides rich contextual data consistent with qualitative research norms (Nielsen & Landauer, 1993). Session length however can also carry risk: back-to-back prototype evaluations may introduce cognitive fatigue, particularly in later conditions of within-subjects designs, which could bias preference toward simpler interfaces regardless of actual suitability.

The study was conducted exclusively with Dutch legal professionals operating within the *gesubsidieerde rechtsbijstand* system. Findings, particularly around institutional trust hierarchies and the adversarial dynamics of peer validation, may not transfer directly to other jurisdictions or legal aid structures. All evaluations were conducted with low-fidelity prototypes, which limits ecological validity; lawyers may respond differently when using a functioning system embedded in their actual caseload workflow, where real confidentiality obligations and time pressure are present. In addition, pro bono practice in the Netherlands spans a wide variety of legal fields. Our experiments and interviews showed that workflows differ widely across these fields, so the findings of this research are difficult to generalise across legal specialisations.

Future work should validate these findings with larger and more diverse participant samples, including practitioners across different legal specialisations, career stages, and firm sizes. Testing contribution mechanisms in live workflow settings would provide stronger evidence for actual adoption intent beyond stated willingness. Most importantly, longitudinal research is needed to understand how trust evolves as the knowledge base grows and output quality improves, and whether the validation and feedback mechanisms identified here can generate sufficient labelled data to enable meaningful model specialisation over time.

6 Conclusion

This paper set out to investigate what design characteristics make the data contribution pipeline of a community-controlled RAG system trustworthy for Dutch pro bono lawyers. Through prototype testing and semi-structured interviews with four practitioners spanning criminal, immigration, and rental law, we examined four dimensions of the contribution pipeline: anonymisation processes, data contribution mechanisms and quality control, source validation.

Regarding anonymisation, participants consistently preferred a hybrid approach in which AI suggests sensitive terms for redaction but the lawyer retains final control. All participants expressed concern that current AI anonymisation tools miss contextual clues, such as a location name that may be innocuous in a city but identifying in a small village. They also emphasised that confidentiality is a non-negotiable professional obligation.

The preference for human-in-the-loop anonymisation resonates with Shneiderman's (2020) framework of human-controlled AI and underscores that automation in high-stakes domains must be calibrated to preserve professional agency.

Regarding quality control, two sub-findings emerged. First, factual accuracy functions as a veto criterion: any factual error in an AI-generated summary was sufficient to destroy participant trust in the entire system, regardless of other qualities. Second, when comparing text-processing

approaches, participants preferred schema-based semantic chunking with predefined legal fields (parties, facts, issue, reasoning, ruling) over document-structure-aware chunking, citing predictability and cognitive efficiency. However, they also advocated for user-customisable schemas that adapt to document type and legal specialisation.

Regarding source validation, trust derived from the authority of the underlying legal materials rather than from validation features layered onto them. Participants accepted expert annotation that mirrored established professional conventions but dismissed peer signals such as likes or dislikes, and their willingness to contribute proved largely independent of the validation model, conditioned instead by the prospect of being evaluated by colleagues.

These findings contribute to the literature on trustworthy AI in professional domains in two ways. Theoretically, they extend the concept of *trust by design* beyond technical robustness to encompass the full contribution pipeline: from the moment a lawyer decides to share a document to the moment another lawyer reads the AI-generated output. Practically, they yield actionable design guidance: anonymisation must blend automation with human oversight; and quality control must prioritise factual accuracy above all else, using structured formats that lawyers can predict, verify, and correct;

Acknowledgment of AI use: Artificial intelligence has been used to compact text, or give recommendations for phrasing and grammar.

7 References

Aanbod voor advocaten en juristen | Lefebvre Sdu. (n.d.). *Sdu.* https://www.sdu.nl/advocaten-juristen

Anonymization guidelines | Dutch judiciary. (n.d.). De Rechtspraak. https://www.rechtspraak.nl/english/anonymization-guidelines

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction, 40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bock, G. W., Zmud, R. W., Kim, Y. G., & Lee, J. N. (2005). Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological factors, and organizational climate. *MIS Quarterly, 29*(1), 87–111. https://doi.org/10.2307/25148669

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need.* The MIT Press. https://doi.org/10.7551/mitpress/12236.001.0001

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches* (4th ed.). SAGE Publications.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

DeAndrea, D. C. (2014). Advancing warranting theory. *Communication Theory, 24*(2), 186–204. https://doi.org/10.1111/comt.12033

De Rechtspraak. (2025). *Jaarverslag Rechtspraak 2024.* https://www.rechtspraak.nl/binaries/_rts_1768832993433/content/assets/rvdr/jd/rvdr-jd-jaarverslag-rechtspraak-2024.pdf

Derksen, T., van Gammeren-Zoeteweij, M., van den Brink, J., & El Alili, H. (2026). *Aanbod in beeld: Inzicht in ontwikkelingen rondom het aanbod in de sociale advocatuur.* Kenniscentrum Stelsel Gesubsidieerde Rechtsbijstand.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). MIT Press.

European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Article 32.*

European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), Recital 26.*

Flick, U. (2022). *Doing interview research: The essential how to guide.* SAGE Publications.

Gaver, B., Dunne, T., & Pacenti, E. (1999). Cultural probes. *Interactions, 6*(1), 21–29.

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist, 57*(5), 664–688. https://doi.org/10.1177/0002764212469365

Havakhor, T., & Sabherwal, R. (2013). Knowledge sharing in peer-to-peer online communities: The effects of recommendation agents and community characteristics. In *2013 46th Hawaii International Conference on System Sciences* (pp. 3551–3560). https://doi.org/10.1109/HICSS.2013.376

Hsu, Y., Huang, T., Verma, H., Mauri, A., Nourbakhsh, I., & Bozzon, A. (2022). Empowering local communities using artificial intelligence. *Patterns, 3*(3), Article 100449. https://doi.org/10.1016/j.patter.2022.100449

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A survey on hallucination in large language models.* arXiv. https://doi.org/10.48550/arXiv.2311.05232

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Karamitsos, I., Roufas, N., Al-Hussaeni, K., & Kanavos, A. (2025). LegNER: A domain-adapted transformer for legal named entity recognition and text anonymization. *Frontiers in Artificial Intelligence, 8.* https://doi.org/10.3389/frai.2025.1638971

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leutheuser, V., Schäfer, F., Müller, J. M., & Voigt, K.-I. (2026). Building trust in AI: A qualitative study of human-AI interaction. *Procedia Computer Science, 277,* 71–80. https://doi.org/10.1016/j.procs.2026.02.048

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33,* 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

Lison, P., Pilán, I., Sánchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 4188–4203). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.323

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12,* 157–173. https://doi.org/10.1162/tacl_a_00638

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (pp. 206–213). Association for Computing Machinery. https://doi.org/10.1145/169059.169166

Pagnoni, A., Balachandran, V., & Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 496–506). https://arxiv.org/abs/2104.13346

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). SAGE Publications.

Pipitone, N., & Alami, G. H. (2024, August 19). *LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain.* arXiv. https://arxiv.org/abs/2408.10343

Raad voor Rechtsbijstand. (n.d.). *Art. 02 Vergoeding.* Retrieved February 1, 2026, from https://www.rvr.org/kenniswijzer/zoeken-kenniswijzer/vaststellen/alle-rechtsterreinen/art-02-vergoeding/#highlight=basisbedrag

Ravi, D., & Sindhgatta, R. (2025). Exploring trust and transparency in retrieval-augmented generation for domain experts. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Article 254, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3706599.3719985

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 33–58). Erlbaum.

Shi, Y., Liu, H., Hu, Y., Song, G., Xu, X., & Wei, H. (2024). *PLawBench: A rubric-based benchmark for evaluating LLMs in real-world legal practice* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.16669

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shrivastava, P. (2025). Understanding acceptance and resistance toward generative AI technologies: A multi-theoretical framework integrating functional, risk, and sociolegal factors. *Frontiers in Artificial Intelligence, 8,* Article 1565927. https://doi.org/10.3389/frai.2025.1565927

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press. https://doi.org/10.1162/dmal.9780262562324.073

Usman, B., & Yennita. (2018). A glimpse into the virtual community of practice (CoP): Knowledge sharing in the Wikipedia community. *Asian Journal of Business and Accounting, 11*(2), 249–276. https://doi.org/10.22452/ajba.vol11no2.8

Walther, J. B., Van Der Heide, B., Hamel, L. M., & Shulman, H. C. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication. *Communication Research, 36*(2), 229–253. https://doi.org/10.1177/0093650208330251

Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly, 29*(1), 35–57. https://doi.org/10.2307/25148667

8 Appendix

8.1 Prototype: Levels of automation in anonymisation

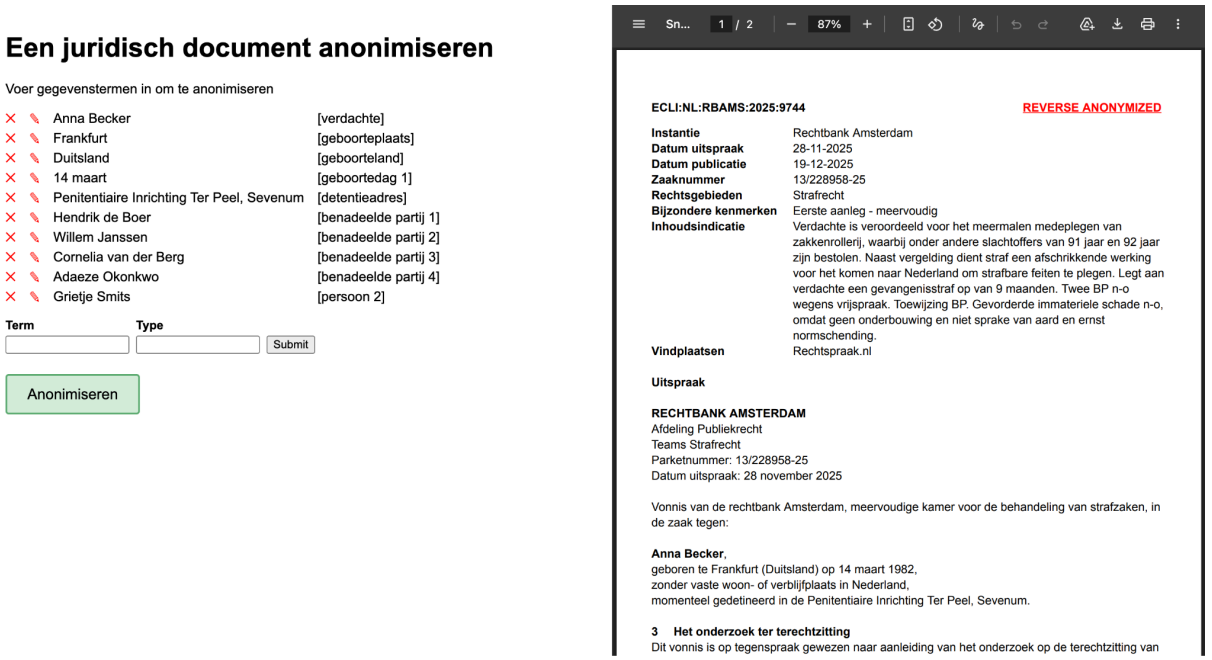


Figure 7: Variant 1a, Manual input (full user control)

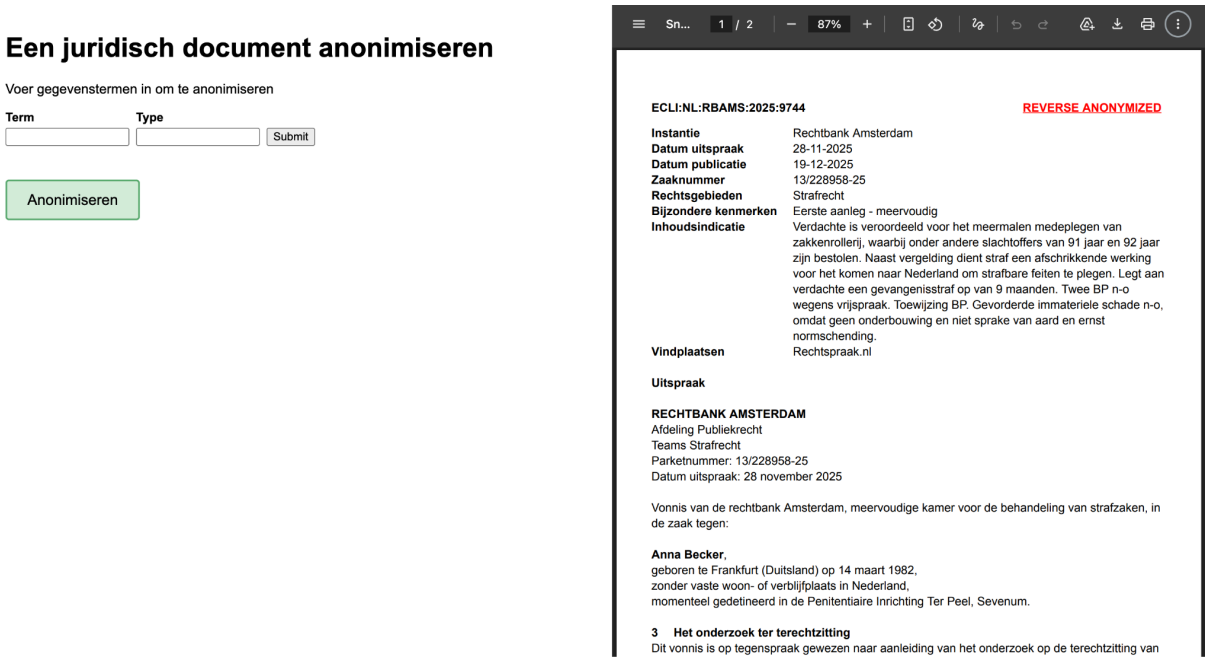


Figure 8: Variant 1b, Auto pre-select with editable suggestions (balanced)

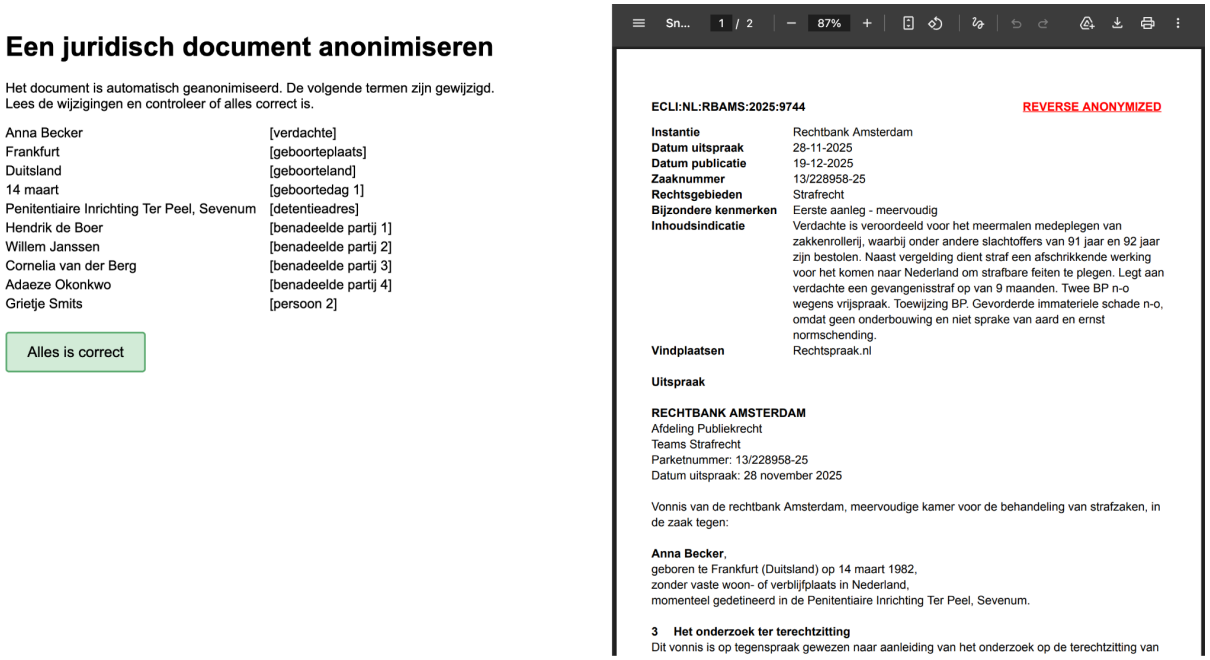
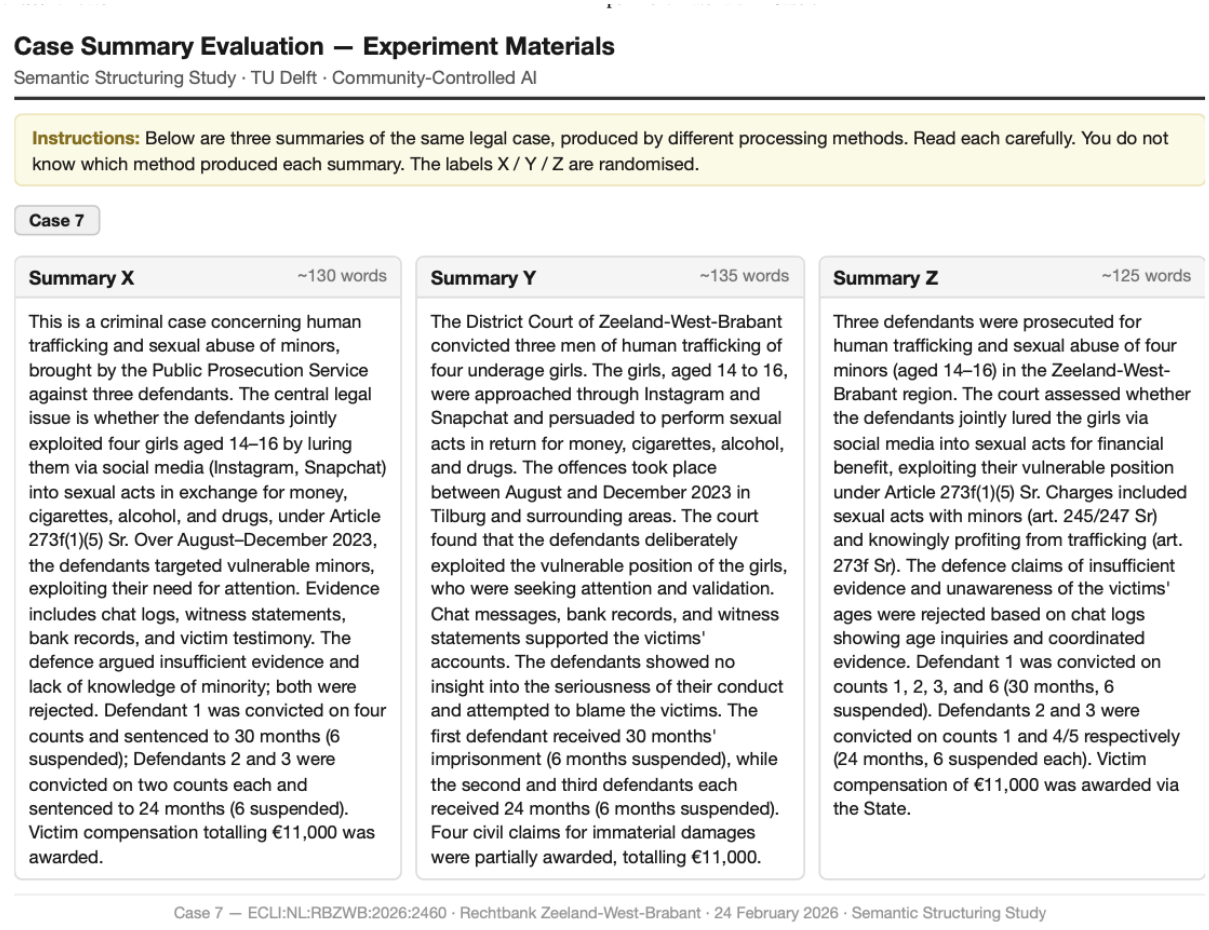


Figure 9: Variant 1c, Fully automatic (maximum automation)

8.2 Semantic Structuring Experiment Materials



Case Summary Evaluation — Experiment Materials

Semantic Structuring Study · TU Delft · Community-Controlled AI

Instructions: Below are three summaries of the same legal case, produced by different processing methods. Read each carefully. You do not know which method produced each summary. The labels X / Y / Z are randomised.

Case 7

Summary X ~130 words

This is a criminal case concerning human trafficking and sexual abuse of minors, brought by the Public Prosecution Service against three defendants. The central legal issue is whether the defendants jointly exploited four girls aged 14–16 by luring them via social media (Instagram, Snapchat) into sexual acts in exchange for money, cigarettes, alcohol, and drugs, under Article 273f(1)(5) Sr. Over August–December 2023, the defendants targeted vulnerable minors, exploiting their need for attention. Evidence includes chat logs, witness statements, bank records, and victim testimony. The defence argued insufficient evidence and lack of knowledge of minority; both were rejected. Defendant 1 was convicted on four counts and sentenced to 30 months (6 suspended); Defendants 2 and 3 were convicted on two counts each and sentenced to 24 months (6 suspended). Victim compensation totalling €11,000 was awarded.

Summary Y ~135 words

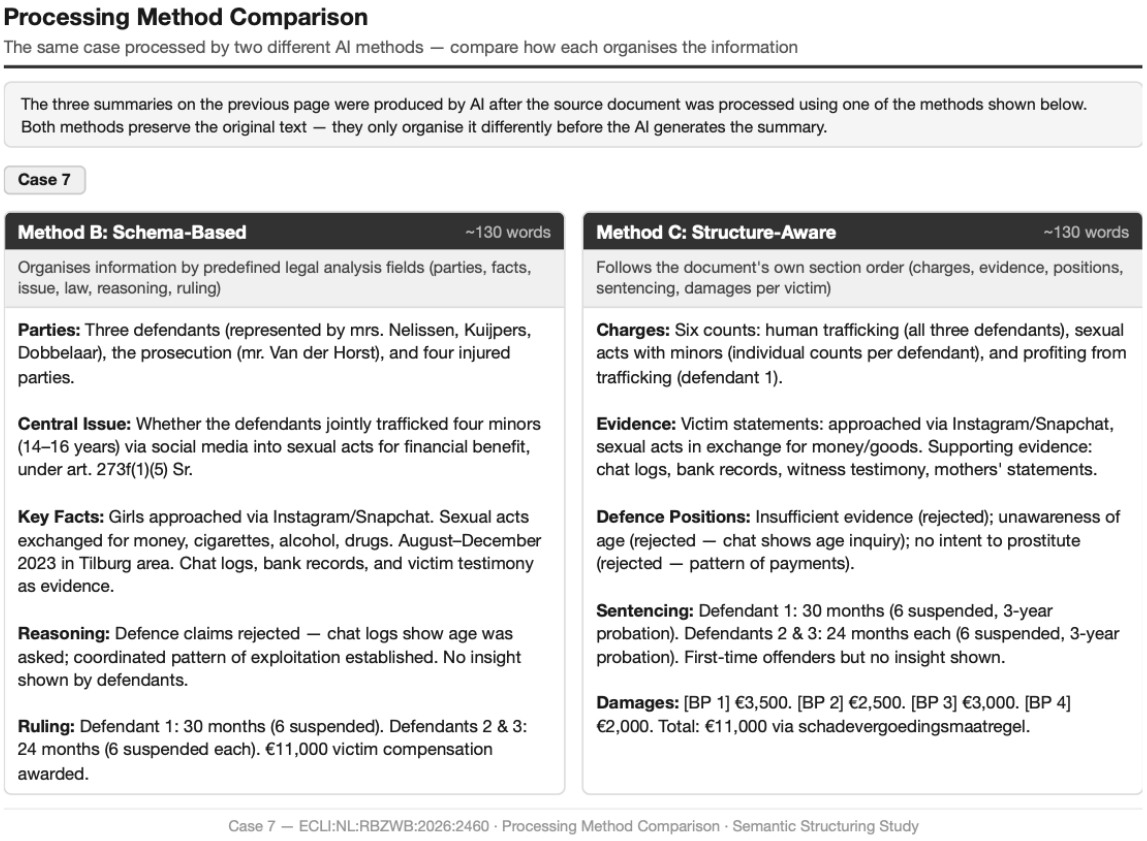
The District Court of Zeeland-West-Brabant convicted three men of human trafficking of four underage girls. The girls, aged 14 to 16, were approached through Instagram and Snapchat and persuaded to perform sexual acts in return for money, cigarettes, alcohol, and drugs. The offences took place between August and December 2023 in Tilburg and surrounding areas. The court found that the defendants deliberately exploited the vulnerable position of the girls, who were seeking attention and validation. Chat messages, bank records, and witness statements supported the victims' accounts. The defendants showed no insight into the seriousness of their conduct and attempted to blame the victims. The first defendant received 30 months' imprisonment (6 months suspended), while the second and third defendants each received 24 months (6 months suspended). Four civil claims for immaterial damages were partially awarded, totalling €11,000.

Summary Z ~125 words

Three defendants were prosecuted for human trafficking and sexual abuse of four minors (aged 14–16) in the Zeeland-West-Brabant region. The court assessed whether the defendants jointly lured the girls via social media into sexual acts for financial benefit, exploiting their vulnerable position under Article 273f(1)(5) Sr. Charges included sexual acts with minors (art. 245/247 Sr) and knowingly profiting from trafficking (art. 273f Sr). The defence claims of insufficient evidence and unawareness of the victims' ages were rejected based on chat logs showing age inquiries and coordinated evidence. Defendant 1 was convicted on counts 1, 2, 3, and 6 (30 months, 6 suspended). Defendants 2 and 3 were convicted on counts 1 and 4/5 respectively (24 months, 6 suspended each). Victim compensation of €11,000 was awarded via the State.

Case 7 — ECLI:NL:RBZWB:2026:2460 · Rechtbank Zeeland-West-Brabant · 24 February 2026 · Semantic Structuring Study

Figure 10: Example case of 3 summarisation conditions



Processing Method Comparison

The same case processed by two different AI methods — compare how each organises the information

The three summaries on the previous page were produced by AI after the source document was processed using one of the methods shown below. Both methods preserve the original text — they only organise it differently before the AI generates the summary.

Case 7

Method B: Schema-Based ~130 words

Organises information by predefined legal analysis fields (parties, facts, issue, law, reasoning, ruling)

Parties: Three defendants (represented by mrs. Nelissen, Kuijpers, Dobbelaar), the prosecution (mr. Van der Horst), and four injured parties.

Central Issue: Whether the defendants jointly trafficked four minors (14–16 years) via social media into sexual acts for financial benefit, under art. 273f(1)(5) Sr.

Key Facts: Girls approached via Instagram/Snapchat. Sexual acts exchanged for money, cigarettes, alcohol, drugs. August–December 2023 in Tilburg area. Chat logs, bank records, and victim testimony as evidence.

Reasoning: Defence claims rejected — chat logs show age was asked; coordinated pattern of exploitation established. No insight shown by defendants.

Ruling: Defendant 1: 30 months (6 suspended). Defendants 2 & 3: 24 months (6 suspended each). €11,000 victim compensation awarded.

Method C: Structure-Aware ~130 words

Follows the document's own section order (charges, evidence, positions, sentencing, damages per victim)

Charges: Six counts: human trafficking (all three defendants), sexual acts with minors (individual counts per defendant), and profiting from trafficking (defendant 1).

Evidence: Victim statements: approached via Instagram/Snapchat, sexual acts in exchange for money/goods. Supporting evidence: chat logs, bank records, witness testimony, mothers' statements.

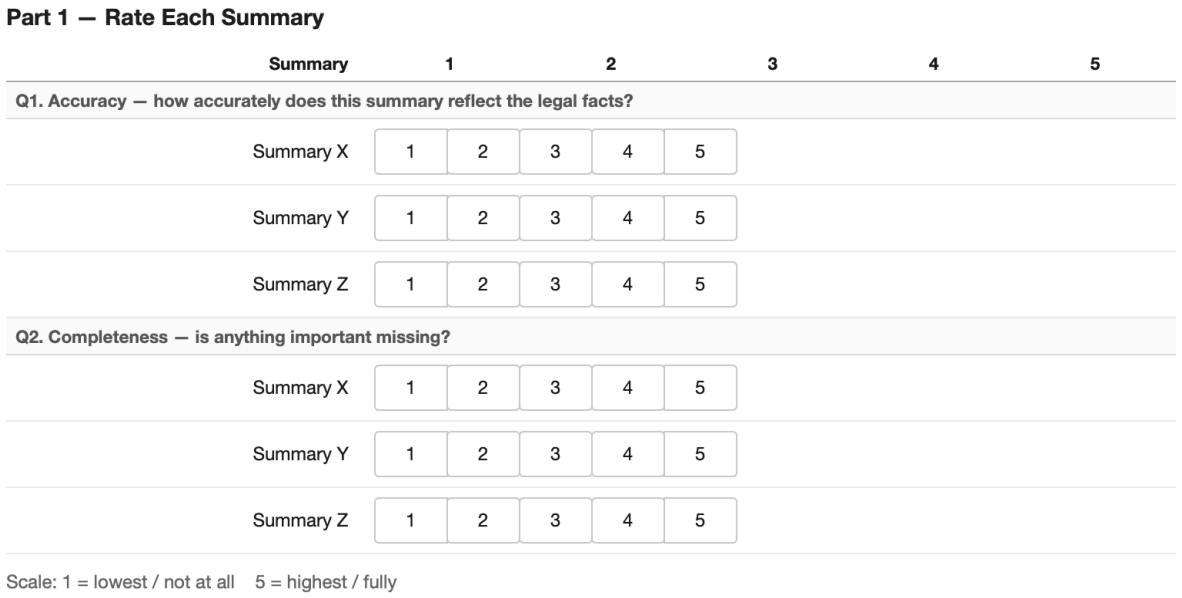
Defence Positions: Insufficient evidence (rejected); unawareness of age (rejected — chat shows age inquiry); no intent to prostitute (rejected — pattern of payments).

Sentencing: Defendant 1: 30 months (6 suspended, 3-year probation). Defendants 2 & 3: 24 months each (6 suspended, 3-year probation). First-time offenders but no insight shown.

Damages: [BP 1] €3,500. [BP 2] €2,500. [BP 3] €3,000. [BP 4] €2,000. Total: €11,000 via schadevergoedingsmaatregel.

Case 7 — ECLI:NL:RBZWB:2026:2460 · Processing Method Comparison · Semantic Structuring Study

Figure 11: Example case of 2 chunking methods



Part 1 — Rate Each Summary

Summary	1	2	3	4	5
Q1. Accuracy — how accurately does this summary reflect the legal facts?					
Summary X	1	2	3	4	5
Summary Y	1	2	3	4	5
Summary Z	1	2	3	4	5
Q2. Completeness — is anything important missing?					
Summary X	1	2	3	4	5
Summary Y	1	2	3	4	5
Summary Z	1	2	3	4	5

Scale: 1 = lowest / not at all 5 = highest / fully

Figure 12: Example of Likert ratings in questions

Term	Full definition
Semantic drift	Content is topically related but does not address the core issue
Missing info	A key legal fact or element is absent from the summary
Wrong position	Information exists but is incorrectly located or attributed
Redundancy	The same information is repeated unnecessarily
Factual error	Information in the summary contradicts the original document

Figure 13: Failure annotation table

8.3 Confidence level prototypes

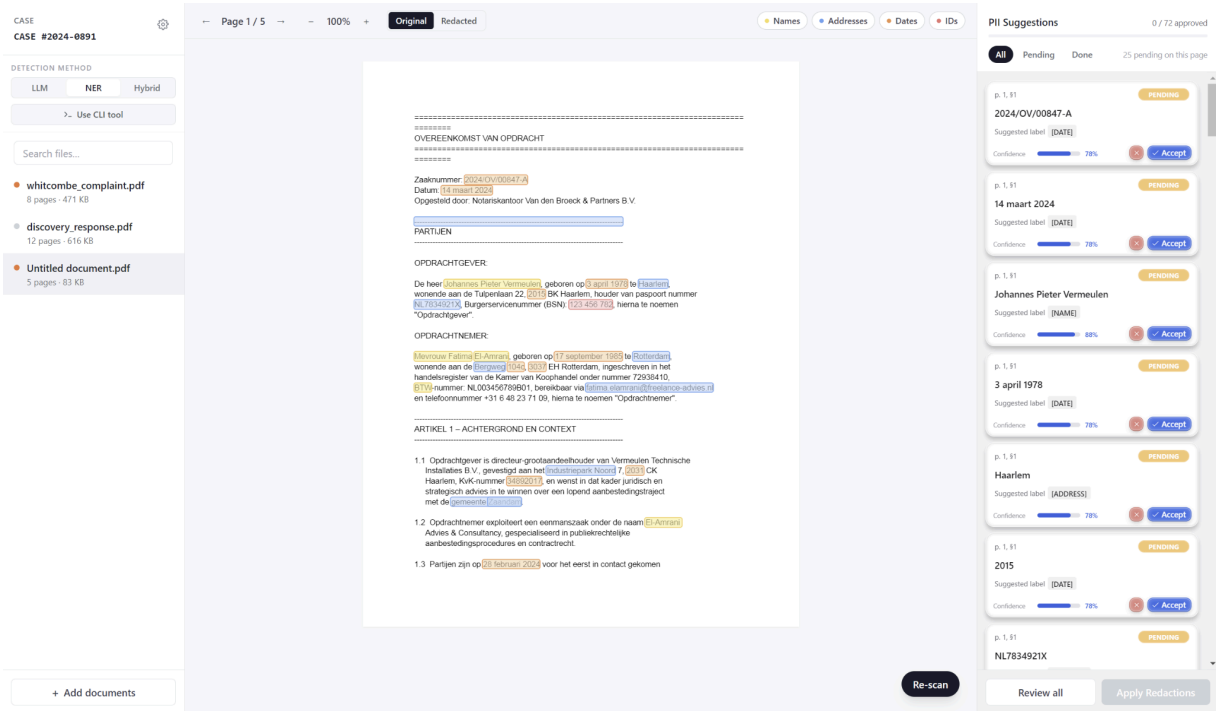


Figure 14: High-fidelity interface designs illustrating confidence scoring. demonstrating a functional document scan utilizing the Named Entity Recognition (NER) operational mode.

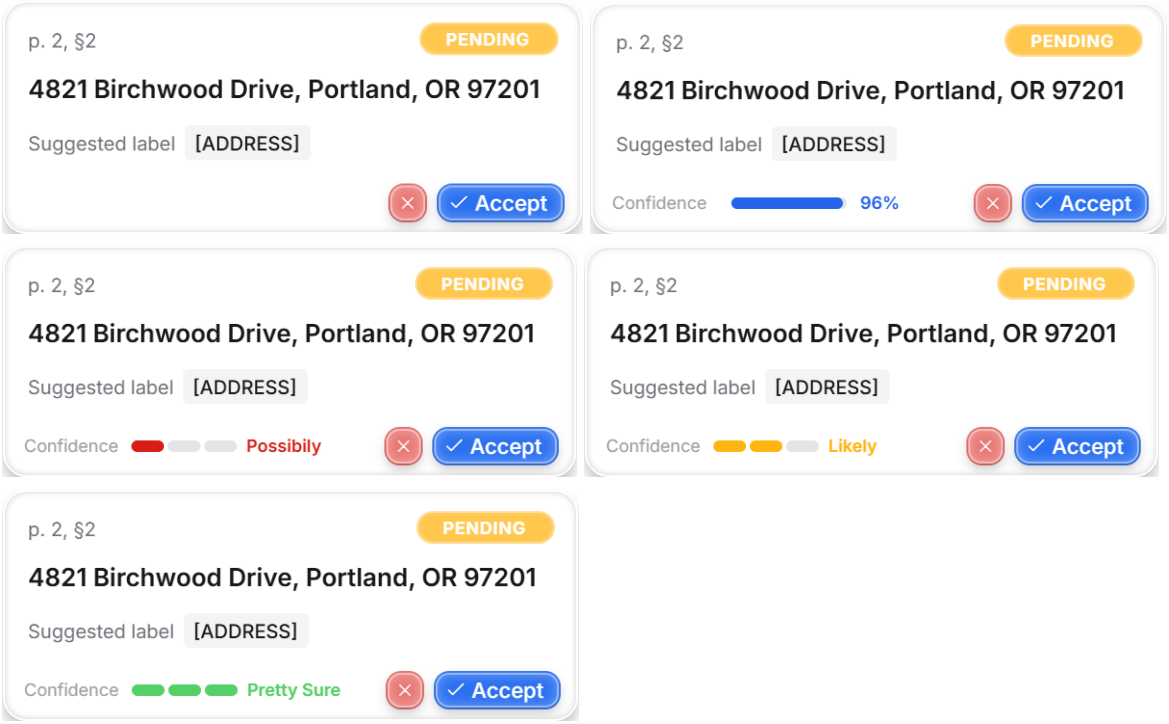


Figure 15: Low-fidelity prototypes for confidence scoring. Sequence (left to right, top to bottom) illustrates: Variant A (baseline/absent), Variant B (numerical/percentage-based), and Variant C (qualitative/3-level categorical).

8.4 Upload interface prototypes

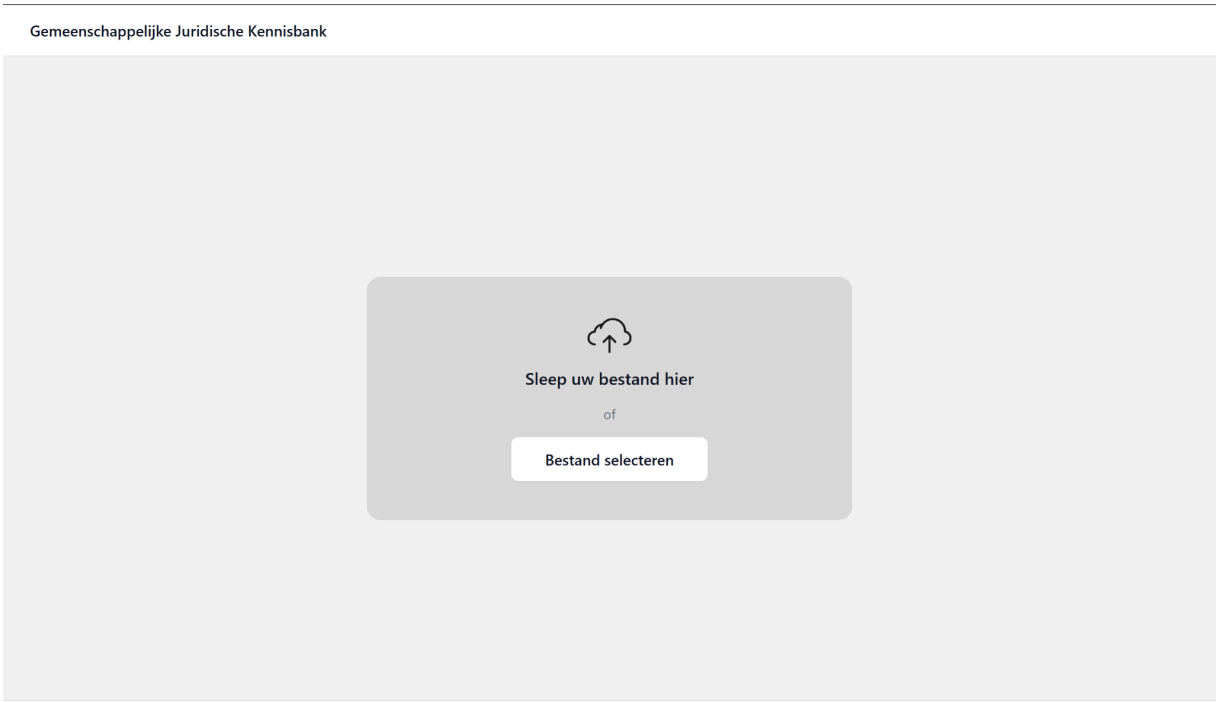


Figure 16: Variant 5a, plain drag-and-drop baseline with no self-assessment

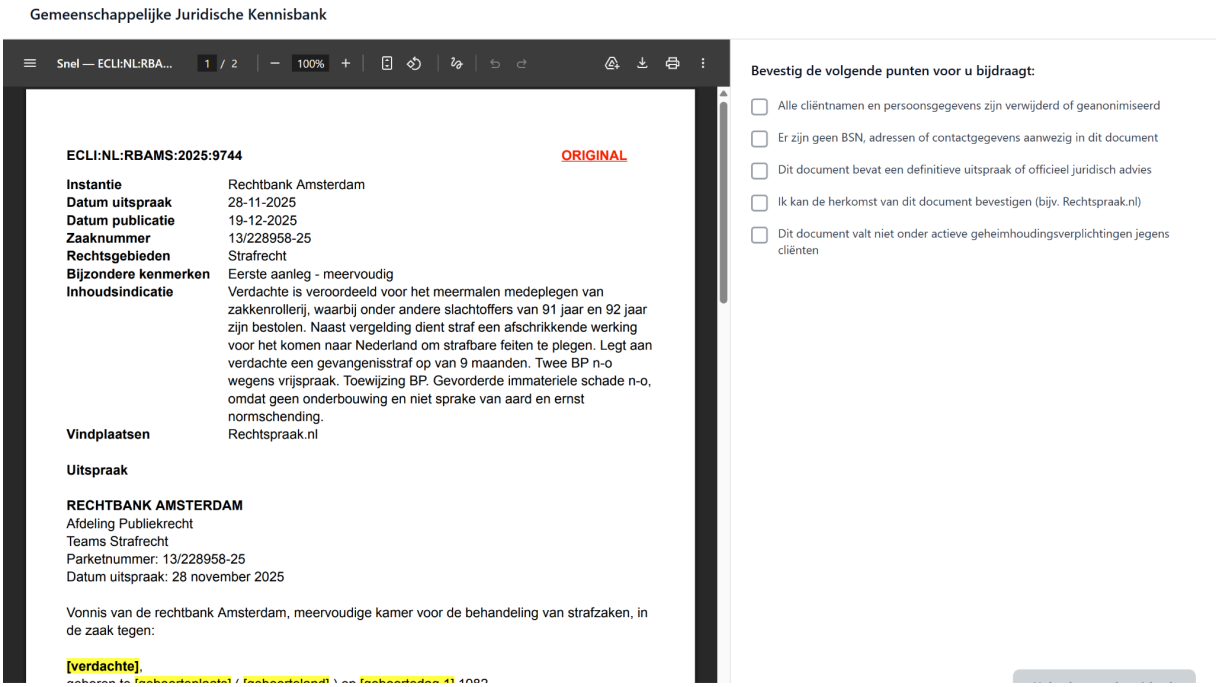


Figure 17:Variant 5b, a checkbox condition offering structured prompts to guide reflection

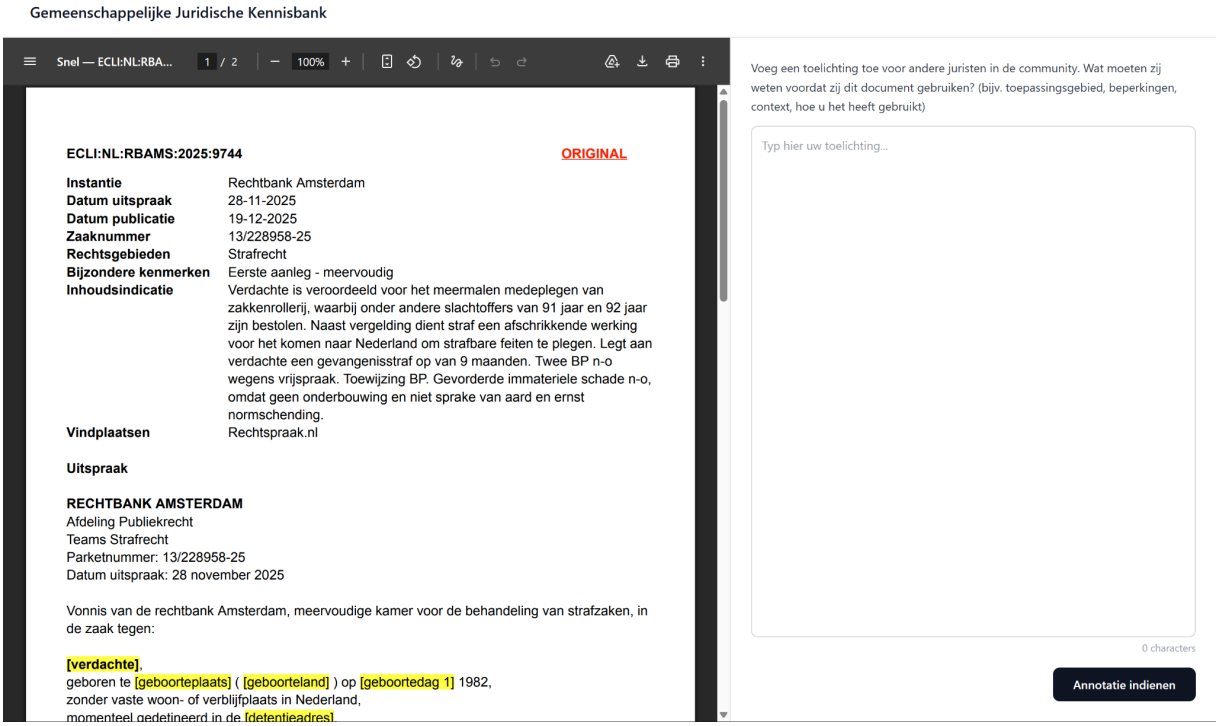


Figure 18: Variant 5c, a free-text condition inviting contributors to annotate on their files

8.5 Validation mechanism prototypes

Figure 19: Variant 6a, No validation (baseline, no review)

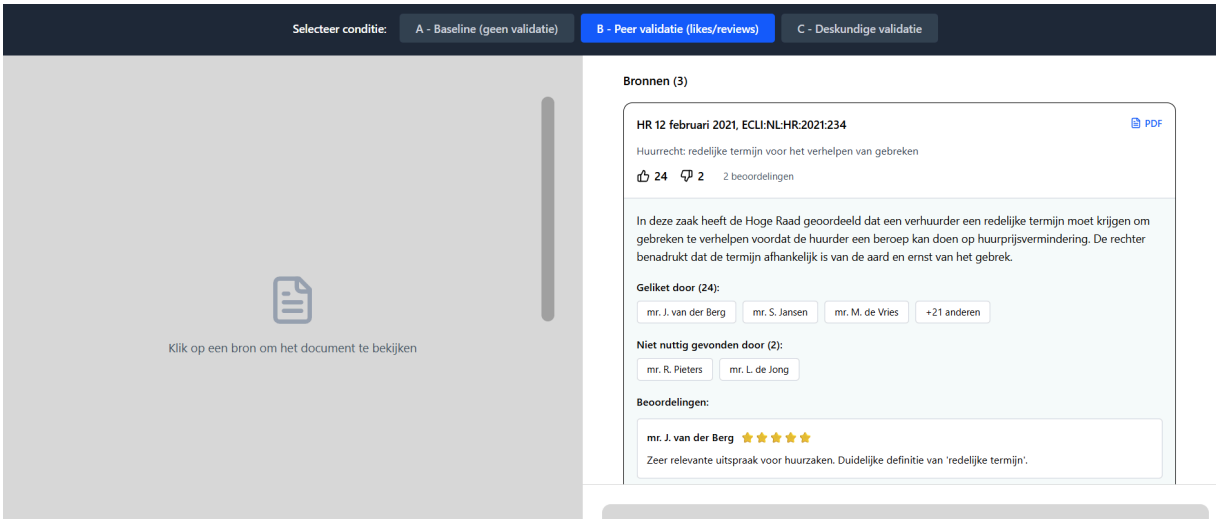


Figure 20: Variant 6b, Peer validation (review by identified colleague lawyers)

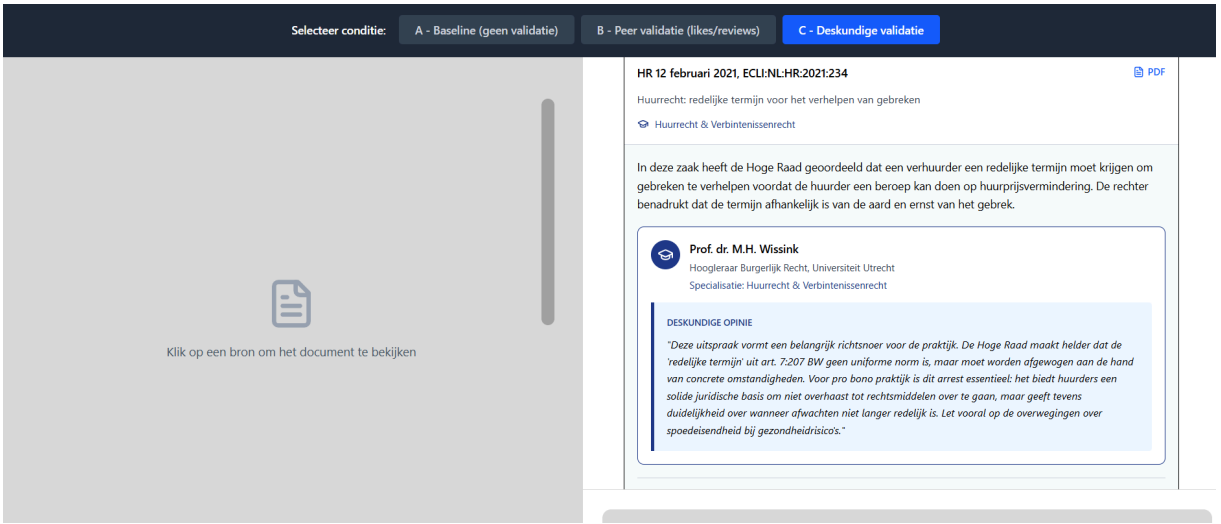


Figure 21: Variant 6c, Institutional validation (review by an editorial authority)