WORKING PAPER

Citation-Grounded Case-Law Research for Jurists

Working paper on a technical tool for case-law research

Heuris, from the Greek *heuriskein*, 'to find, to discover', is the name of the case-law research tool studied in this paper.

Applied research · November 2026

Thesis. In law, the distrust of AI is structural: its outputs are opaque, can change from one run to the next, and can cite authorities that do not exist. A general-purpose integration inherits these properties; no jurist can rely on it before a court. Trust is therefore a matter of architecture, designed for the constraints of legal research:

1. Every displayed sentence rests on a source passage kept at exact character offsets; AI drafts, the source decides.
2. A verifier removes any sentence whose quotation cannot be found or does not support the claim, so the jurist does not have to second-guess whether a citation was hallucinated.
3. The system fails closed: when nothing survives verification, it says so rather than inventing an answer.
4. The client's file is treated as confidential by design: it is parsed in memory and retrieval runs on the local machine; the language-model call is the only point where text leaves it, and no real client file is processed before an EU-hosted or local model is in place.

Abstract

Generative assistants write fluent legal prose. In legal research, however, the difficulty lies not in fluency but in verifiability. A preregistered study of commercial AI legal research tools found that retrieval-augmented systems hallucinate less than a general-purpose chatbot, yet still produced incorrect or misgrounded answers in 17–33 % of queries.[1] Retrieval-augmented generation[2] changes where the model's words come from. It does not, by itself, guarantee that the passage cited is the right one, quoted exactly, and actually supportive of the sentence it is attached to.

The problem has a second aspect, which concerns the time and resources available for research. The official description of Juportal, the public case-law database of the FPS Justice, describes search by ECLI number, case number and dates, with filters; it mentions no semantic or natural-language search.[16] No Belgian or French measurement of the time jurists spend on case-law research was found. The resources of the Belgian justice system are, moreover, limited relative to the other EU Member States: according to the European Commission, "the budget spent on the justice system and the number of professional judges per 100 000 inhabitants remain below the EU average", and the disposition time for civil and commercial cases at first instance was 246 days in 2023.[18] For 2022, the CEPEJ puts this budget at €102.5 per inhabitant, above the Council of Europe median, but at 0.22 % of GDP, below it.[17] In its press release of 10 January 2025, the Court of Audit finds that the digital transformation of justice « n'est pas menée de manière satisfaisante » (is not being carried out satisfactorily).[19]

The problem has a third aspect: confidentiality. Preparing a file brings the client's facts, covered by professional secrecy, into the query and into the material the tool processes. A general-purpose assistant sends them to third-party services, under terms and in jurisdictions the jurist cannot inspect, and offers no guarantee that this secrecy is kept.

Heuris is intended to help jurists reduce the effort of case-law research while retaining traceability; that time saving for the jurist has not been measured (see requirement R8).

Heuris is a case-law research tool for jurists, designed for any legal field where lawful data is available. Its proof of concept runs on a first corpus of more than 25,000 Belgian labour-law decisions, in the form of French-language commentaries rather than the decisions themselves; other fields are planned, depending on data access. The problem is addressed at the level of the data and of verification: collection is limited to sources that permit it; every source passage is kept at exact character offsets; generation is subject to a verifier that removes any sentence whose quotation cannot be found in full at its cited location, or that a second automated check judges not to support the sentence. If no sentence survives verification, the system states this. For confidentiality, the case file is parsed in memory and retrieval runs on the local machine; only the language-model call leaves it, and the proof of concept is limited to public data and fictitious or anonymised files until an EU-hosted or local model is in place. In this paper, "verified" means that an excerpt matches the source text at its cited offsets; since the source is a commentary, a verified excerpt is not proof that the court said it.

Postulate. A legal research assistant is reliable only if every sentence it shows can be traced, character for character, to a lawfully obtained passage that supports it — and if it says "nothing found" when it cannot.

1. Statement of the problem

1.1 Hallucination in generative legal research

In a study of more than 800,000 queries on U.S. federal case law, general-purpose language models (GPT-4, GPT-3.5, PaLM 2 and Llama 2) hallucinated at least 58 % of the time, struggled to predict their own hallucinations, and often uncritically accepted users' incorrect legal assumptions.[3] Specialised tools built on retrieval do better without solving the problem: in the preregistered study cited above, commercial AI legal research tools still hallucinated 17–33 % of the time.[1] More recent work suggests that the problem is structural. Xu et al. formalise hallucination and show that "it is impossible to eliminate hallucination in LLMs";[4] Kalai et al. observe that hallucinations "persist even in state-of-the-art systems" and attribute this in part to evaluation practices that reward guessing under uncertainty.[5] On this literature, a plausible reading is that more recent models reduce the rate of hallucination without removing its mechanism. This reading is an inference from the cited work, not a measured result. Both legal studies concern U.S. law, and no equivalent measurement exists for Belgian or French law. The language of the question is unlikely to be the main factor. Training data are dominated by English: in the pretraining data of Llama 2, English accounts for 89.70 % and French for 0.16 %.[6] For a high-resource language such as French, the measured gap on a general benchmark is small: on a machine-translated MMLU, GPT-4 scored 83.6 % in French against 85.5 % in English (3-shot).[7] The larger documented weakness concerns regional and long-tail knowledge: in Global MMLU, 84.9 % of the questions requiring geographic knowledge focus on North America or Europe,[8] and BLEnD reports that models perform better for cultures highly represented online, with a difference of up to 57.34 % for GPT-4.[9] *Inference, not a measured result:* Belgian case law is barely represented in training data, so models are likely to recall specific Belgian decisions poorly; no benchmark measures this.

1.2 Judicial consequences

In *Mata v. Avianca* (S.D.N.Y., 22 June 2023), the court sanctioned lawyers who had "submitted non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT", imposing a penalty of US$5,000 jointly and severally.[10] Belgian courts have since met the same failure. RTBF reported in December 2025 that the Antwerp Court of Appeal had found submissions supported by "une jurisprudence inexistante et des sources juridiques inventées de toutes pièces", and ordered the lawyer concerned to pay €25,000 in total: a €2,500 fine for manifest abuse of procedure and €7,500 to each of the three parties. The same report states that on 15 December 2025 the Ghent business court (*tribunal de l'entreprise*) dealt with submissions citing decisions of the Court of Cassation and the Constitutional Court that never existed.[11] In October 2025 the Orde van Vlaamse Balies reported a judgment of the Ghent Court of Appeal concerning submissions that contained a long list of incorrect quotations from existing statutory provisions and references to non-existent case law and legal writing.[12] These Belgian decisions are cited as reported by the press and by the Orde van Vlaamse Balies; the judgments themselves, their dates and roll numbers were not consulted. None was given by a labour court; they are cited for the lawyer's duty of care in the use of AI.

1.3 Access to Belgian case law

About 1.4 million decisions are handed down each year in Belgium, according to the SPF Justice (2026).[13] Only a small share is published. An academic estimate puts it at about 1 %;[14] no official measurement of that share was found. Juportal, the official case-law database of the SPF Justice, is a selection: the individual courts decide which judgments they anonymise and publish there. At its launch in 2020, the labour courts of appeal contributed a selection of judgments (more than 21,000 publications).[15] The central register of judicial decisions (JustJudgment) exists, but as of mid-2026 only the parties concerned and their lawyers have access to their own judgments, and the judgments of the labour courts of appeal were still to be added.[13] As § 6 explains, Juportal's robots.txt closes it to automated agents other than the ministry's own crawler, and Heuris does not collect from it.

1.4 Research time and resources

Juportal replaced the Jure-Juridat search engine in December 2020, because the Flash technology on which Jure-Juridat had been built would no longer be supported. The official description presents it as « la seule banque de données publique comportant une sélection de jugements et d'arrêts anonymisés » and describes search by ECLI number, case number (*numéro de rôle*) and dates, with filters by court, legal field, type and year of publication, and RSS feeds.[16] That description mentions no semantic or natural-language search; this is a statement about the official description, not proof that no such function exists.

No Belgian or French measurement of the time jurists spend on case-law research was found. The figures available for other jurisdictions are vendor surveys or could not be verified in their primary source, and are not used here.

Comparative figures on the resources of the Belgian justice system depend on the comparison group. The CEPEJ (Council of Europe) reports for 2022 an implemented judicial-system budget of about €1.2 billion, or €102.5 per inhabitant, "exceeding the CoE median", whereas as a percentage of GDP (0.22 %) the budget is "below the CoE median"; it adds that the number of pending civil and criminal cases "is not available for 1st and 2nd instances".[17] The European Commission's 2025 Rule of Law Report, which compares Belgium with the other EU Member States, states that "the

budget spent on the justice system and the number of professional judges per 100 000 inhabitants remain below the EU average", with reference to the 2025 EU Justice Scoreboard. It reports a disposition time of 246 days in 2023 for civil and commercial cases at first instance, observes that "a comprehensive overview of the efficiency of justice remains unavailable due to a persistent lack of data on court proceedings", and names "online access to published judgments and machine-readable judicial decisions" among the matters raised by stakeholders.[18] The two findings are not contradictory: they rest on different reference groups (the member States of the Council of Europe; the EU Member States) and on different years. In its press release of 10 January 2025 on the federal State's steering of the digital transformation of justice, the Court of Audit (*Cour des comptes*) finds that this transformation « n'est pas menée de manière satisfaisante » (is not being carried out satisfactorily) and that the digitalisation of justice « reste en chantier depuis plusieurs décennies » (has remained a work in progress for several decades).[19]

1.5 Limits of current knowledge

The absence of a Belgian or French measurement of research time is stated in § 1.4. The hallucination rates above were measured on U.S. law only. The 1 % share is an estimate whose method is not stated.

2. The jurist's situation

The sources available on this point are limited. This section separates what is documented from what is not.

2.1 Where case-law research sits in the preparation of a file

The use case recorded for this project describes the following sequence. A jurist receives a case file. The facts, the legal question, the provisions relied on, the parties and the key dates are identified; the client's questions and the opposing party's arguments are listed; case law is sought for each of them; and the jurist drafts an answer that cites the decisions relied on. Heuris follows this sequence: no search is run before the jurist has validated the case sheet, questions inferred from the file are not searched until the jurist validates them, and the answer remains the jurist's work (D46–D53; see Appendix B). (Sources: *project documentation (not public)*; *project documentation (not public)*, steps 6–12g.)

This sequence is the designer's recorded use case of 2 November 2026. It is not the result of a study of how Belgian jurists work, and no such study was found. Nor was any Belgian figure found on the time jurists spend on case-law research, or on how often they check the citations they rely on (§ 1.5).

2.2 The professional significance of verifiable provenance

The jurist needs to know what a court decided, on which facts, and where it is written: court, date, roll number (*R.G.*) and a passage that can be opened and read in context (*project documentation (not public)*). The decisions reported in § 1.2 show the professional consequences of citations that cannot be traced. In *Mata v. Avianca*, the court found that the lawyers had "abandoned their responsibilities" by submitting non-existent opinions.[10] In Belgium, the Antwerp Court of Appeal ordered a lawyer to pay a fine and an indemnity to each of the three parties, as the submissions rested on non-existent case law,[11] and the Orde van Vlaamse Balies noted that such errors cause a loss of time for courts and for fellow lawyers.[12] Preparing a file also involves the client's facts,

which are covered by professional secrecy; in an assisted workflow these facts enter the query and the material submitted to the language model (*project documentation (not public)*; *project documentation (not public)*). The applicable rules are not analysed here (§ 13, limit 20). The rules of professional conduct that apply in Belgium to the use of AI are not analysed here. The Orde van Vlaamse Balies refers to guidelines drawn up with Avocats.be;[12] they were not consulted.

3. Reliability requirements for a case-law research tool

The problem set out in §§ 1 and 2 implies requirements for any system that assists case-law research. They are stated independently of any implementation, and each is retained only because § 1 or § 2 motivates it.

- **R1 — Verbatim traceability.** Every claim presented to the jurist must be traceable to a passage of an identified source that it quotes verbatim and that can be opened in context. *Motivated by* § 1.1 (a citation that does not support its proposition counts as a hallucination) and § 2.2 (court, date, roll number and an openable passage).
- **R2 — Verification independent of generation.** Whether a quoted passage supports a claim must be checked by a step distinct from the one that generated the claim; retrieval alone does not suffice. *Motivated by* § 1.1 (retrieval-based legal tools still hallucinate; hallucination is not eliminated in recent systems).
- **R3 — Refusal rather than an unsupported answer.** When no claim can be supported, the system must say so instead of answering. *Motivated by* § 1.2 (sanctions for submissions resting on non-existent case law).
- **R4 — Visible scope and gaps of the corpus.** The system must state what its corpus covers and what it omits, so that an absence of results is not read as an absence of case law. *Motivated by* § 1.3 (about 1 % of decisions published, by estimate; Juportal a selection) and § 1.5 (the gaps stated).
- **R5 — Clarity about the nature and authority of the source.** Each result must identify the court, the date and the roll number, and state whether the source is the decision itself or a secondary text about it. *Motivated by* § 2.2 (what a court decided, on which facts, and where it is written) and § 1.3 (official texts selected and access restricted).
- **R6 — The jurist keeps the decision.** The system may propose; the jurist validates, drafts and remains the author of what is filed. *Motivated by* § 2.1 (the jurist drafts the answer) and § 2.2 (the lawyer's responsibility for what is submitted).
- **R7 — Confidentiality of the client's facts.** The confidentiality of the client's facts must be protected throughout processing, including when they are submitted to a language model and when results are stored. *Motivated by* § 2.2 (the client's facts, covered by professional secrecy, enter the query).
- **R8 — Reduced search effort without loss of traceability.** The system should reduce the effort of finding relevant case law compared with search by number, date and filters, without weakening R1. *Motivated by* § 1.4 (search by ECLI, case number, date and filters, with no semantic search in the official description; no measurement of research time; constrained resources).

4. Related work

4.1 Retrieval-augmented legal research tools

Retrieval-augmented generation conditions a language model on passages retrieved from an external corpus.[2] Magesh et al. evaluated commercial legal research tools built on this approach, on 202 preregistered queries on U.S. law, and report hallucination rates 17–33 %. Their definition of hallucination includes a response whose citation does not support the proposition to which it is attached.[1] Dahl et al. had documented higher rates for general-purpose models used without retrieval.[3]

4.2 Attribution and citation evaluation

Several works treat the question of whether generated text is supported by a source as an evaluation problem. Rashkin et al. define "Attributable to Identified Sources" (AIS), a framework and a two-stage human annotation procedure for judging whether model-generated statements are supported by underlying sources.[20] Bohnet et al. formulate attributed question answering, in which a system returns an answer together with a supporting source, and propose a reproducible evaluation framework in which an automatic metric correlated with human judgements is used for development.[21] Gao et al. introduce ALCE, a benchmark for systems that retrieve evidence and generate answers with citations, with automatic metrics for fluency, correctness and citation quality.[22]

4.3 Claim decomposition and entailment checking

Min et al. (FActScore) break a generated text into atomic facts and compute the share of those facts supported by a knowledge source; they also propose an automated estimator based on retrieval and a language model.[23] Honovich et al. (TRUE) compare factual-consistency metrics on a standardised collection of annotated texts and report that large-scale natural-language-inference (entailment) and question-generation-and-answering approaches achieve strong and complementary results.[24]

4.4 Position of Heuris

Heuris combines elements found in this literature: retrieval of passages (§ 8), decomposition of answers into claims, and an entailment check by a language model (§ 9). Its particular choices are to require each quotation to be found in full at its cited character offsets, and to withhold, rather than score, any claim that fails a check (D9, D15). These are design choices, not results. Heuris has not been evaluated on any of the benchmarks cited above, and no comparison of effectiveness is made or implied.

5. Fluency and verifiable provenance

The requirements of § 3 are used below to situate Heuris. The status of each is assessed only against facts stated in this paper; it is not an evaluation and implies no performance claim.

Table 5.1 · Requirements of § 3 and how Heuris addresses them

R	How Heuris addresses it	Status
R1	Passages kept at exact offsets; each claim must quote its passage in full at the cited offsets; jurist-mode quotations are cut by the server (D2, D9, D24, D53)	Met, with respect to the commentary text (see R5)
R2	The verbatim check is deterministic code; the support check is a second call to the same language model that drafted the claim (D9; *project documentation (not public)*); jurist-mode sentences are not support-checked (D53)	Partly met
R3	Failed claims are removed; at least one claim must be verified against a decision passage, otherwise « Absence de réponse » (D9, D15, D16; § 9.4)	Partly met: implemented, but its reliability depends on the support check, which has not been measured
R4	Scope and gaps are stated in this paper (§ 1.3; § 13, limit 19); the interface shows no notice of the corpus scope (*project documentation (not public)*, `app.js`)	Partly met
R5	Court (with division where recorded), date and R.G., with a link to the commentary; but the corpus consists of commentaries and the interface labels items « décision » (§ 6.2; § 13, limit 12)	Partly met
R6	Case sheet validated before any search; inferred items require validation; drafting is optional; the jurist validates answers (D46–D53)	Met in design; exercised in simulated mode only (§ 13, limit 2)
R7	Retrieved passages, the case sheet and case-file excerpts are sent to the DeepSeek API, operated outside the EU (D11; note 28); drafts and validated answers are stored unencrypted and without a retention limit (`06_confidentiality.md`; § 13, limit 7); a local or EU-hosted model is planned, not implemented (§ 11)	Not met
R8	Hybrid retrieval combines lexical (BM25) and semantic (bge-m3) search, fused by RRF and reranked on the local machine; each sub-question of a file receives at least one result slot (D4–D7, D17, D18, D27); every claim remains subject to the verbatim check (R1)	Not demonstrated: search effort has not been measured and no time study exists; fifteen self-written questions do not discriminate between retrieval modes (§ 13, limit 6)

5.1 What the jurist requires of an answer

An unsourced plausible paragraph is of limited use to a jurist, in labour law as in any other field. The jurist needs to know what a court decided, on which facts, and where it is written: court, date, roll number (*R.G.*), and a passage they can open and read in context. Before a court, an unsupported citation is more damaging than the absence of an answer. The design rests on one constraint: every statement must point to an exact, openable passage, and the system must refuse when it cannot. (Source: *project documentation (not public).*)

5.2 Limits of retrieval alone

Retrieval restricts the material the model reads; it does not constrain what the model writes. Three failures remain possible after retrieval: the right document but the wrong passage; the right passage, paraphrased into something it does not say; and a correct-looking quotation that does not support the claim it is attached to. The study cited above classes the last kind, a citation that does not support its proposition, as a hallucination in its own right.[1] Heuris treats each of the three as a verification problem, not a prompting problem.

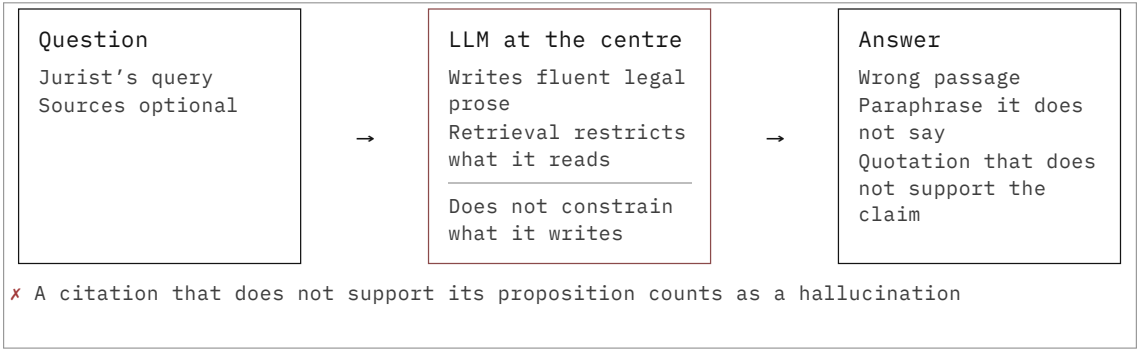


Figure 1 · Default pattern: LLM at the centre, sources optional

Retrieval restricts the material the model reads; it does not constrain what the model writes.

Three failures remain possible after retrieval: the right document but the wrong passage; the right passage, paraphrased into something it does not say; and a correct-looking quotation that does not support the claim it is attached to.

6. Lawfulness of the data

The decision records cited in this section are explained for jurists in Appendix B.

6.1 Sources excluded

Juportal is the official case-law database of the SPF Justice. Its robots.txt disallows all automated agents except the ministry's own crawler.[25] Heuris does not crawl Juportal: no crawl, and no browser-based workaround (decision record D1). Official full-text decisions and their ECLI identifiers are therefore not part of the corpus.

6.2 Sources retained (first corpus: labour law)

The first corpus consists of French-language commentaries on decisions of the Belgian labour tribunals (*tribunaux du travail*) and labour courts of appeal (*cours du travail*). The publisher's robots.txt allows article pages. Its terms authorise reproduction of textual information provided the source is clearly cited, and forbid any commercial exploitation.[26] The collector respects a 1.5-second interval between requests, above the declared crawl delay, and stores the terms verbatim in each record. Every passage returned by the system carries its source URL and decision reference. (Sources: *project documentation (not public)* § 3; *project documentation (not public)*; *project documentation (not public)* § 5.)

The documents are commentaries, not decisions: the collector types every record as a commentary (*project documentation (not public)*). They concern judgments (*arrêts*) of labour courts of appeal and judgments (*jugements*) of labour tribunals. Each comprises a title, a reference line, an introductory summary, the legal provisions and an analysis in which the commentator reports

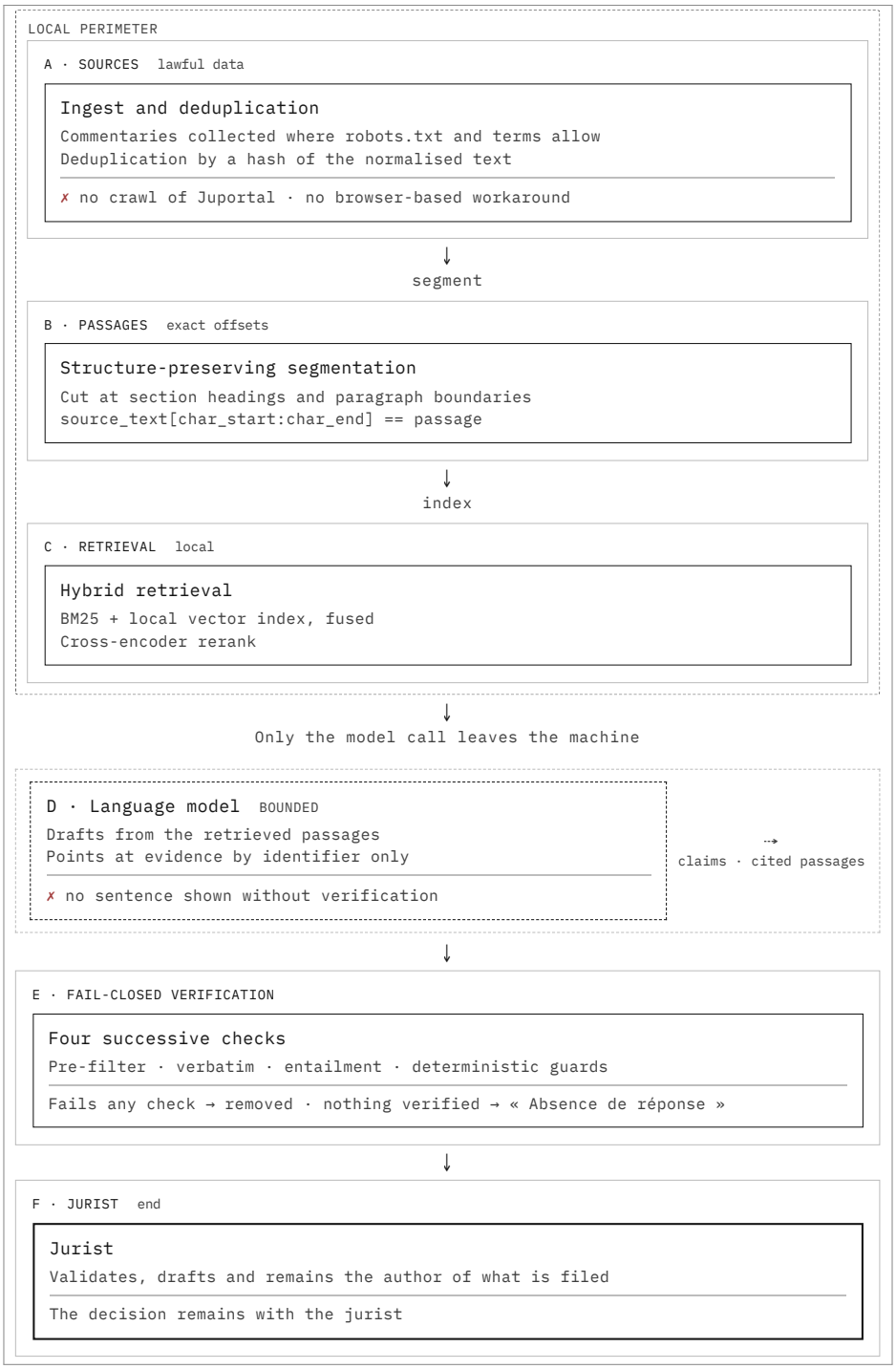


Figure 2 · Heuris spine: sources, passages, hybrid retrieval (BM25 + local vector index), fail-closed verification, jurist

Every statement must point to an exact, openable passage, and the system must refuse when it cannot.

The model may only point at evidence by identifier. Every generated sentence is decomposed into claims; a claim that fails any check is removed. The decision remains with the jurist.

the court's reasoning, partly in paraphrase and partly in quoted extracts, whether of the court or of other texts. The operative part (*dispositif*) is mostly absent. The commentaries carry no ECLI; a decision is identified by court (with division where recorded), date and roll number, and each result links to the commentary on the publisher's site. (Sources: *project documentation (not public)*, *project documentation (not public)*; `01_data-sources.md`.)

Definition. Throughout this paper, "verified" means that the excerpt matches the source text — here, the commentary — at its cited offsets. It does not mean that the court said it: a verified passage may be the commentator's paraphrase or analysis.

Table 6.1 · Source comparison

Source	Content	Status in Heuris	Basis
Juportal (SPF Justice)	Official full-text decisions, ECLI-identified	Not collected	robots.txt disallows all robots but the ministry crawler[25]
European e-Justice ECLI search	ECLI index including Belgian decisions	Not used	Reported closed to robots (01 § 1; not re-checked)
Publisher of case-law commentaries	Commentaries on labour tribunal and labour court of appeal decisions	In use	robots.txt allows article pages; terms permit attributed reproduction, forbid commercial use[26]

6.3 Corpus

Heuris works with a corpus of more than 25,000 decisions. The publisher sometimes publishes one commentary under several article ids; deduplication by a hash of the normalised text keeps the smallest id. Records that share a decision reference but differ in body or title are kept. Hashing the content rather than the reference was deliberate: a shared reference does not prove identical text (D3). Near-duplicates that differ only in punctuation are a known residual gap. (Sources: `01_data-sources.md` §§ 4–5; *project documentation (not public)*; *project documentation (not public)*.)

7. Structure-preserving segmentation

The decision records cited in this section are explained for jurists in Appendix B.

A verifier can only check a quotation against a passage if the passage is exactly what the source says. Heuris therefore cuts each commentary only where its author did: at section headings and paragraph boundaries. Oversize paragraphs (more than 1,500 characters) are split only at a sentence end that lies outside a « » quotation. No passage ends mid-sentence and no passage crosses a heading. For every passage, `source_text[char_start:char_end] == passage` is asserted at build time and in tests. A document with fewer than two recognised section types is flagged and produces no passages rather than badly cut ones (D2).

Section types follow the commentary layout: summary, legal provisions, facts, procedure, parties' positions, the lower court's decision, the court's decision, the prosecutor's opinion, and notes. No passage comes from a flagged document. (Sources: *project documentation (not public)* §§ 1–5; *project documentation (not public)*; *project documentation (not public)*; *project documentation (not public)*.)

Table 7.1 · Fixed-size chunks vs structure-aware passages

	Fixed-size chunks	**Structure-aware passages (Heuris)**
Cut points	Every *n* tokens, anywhere	Headings, paragraphs, sentence ends outside quotations
Section identity	Lost	Kept (summary, facts, court's decision …)
Exact offsets	Usually approximate	Asserted for every passage
Unstructured input	Silently chunked	Flagged, no passages
Suitable for verbatim verification	Weakly	By construction

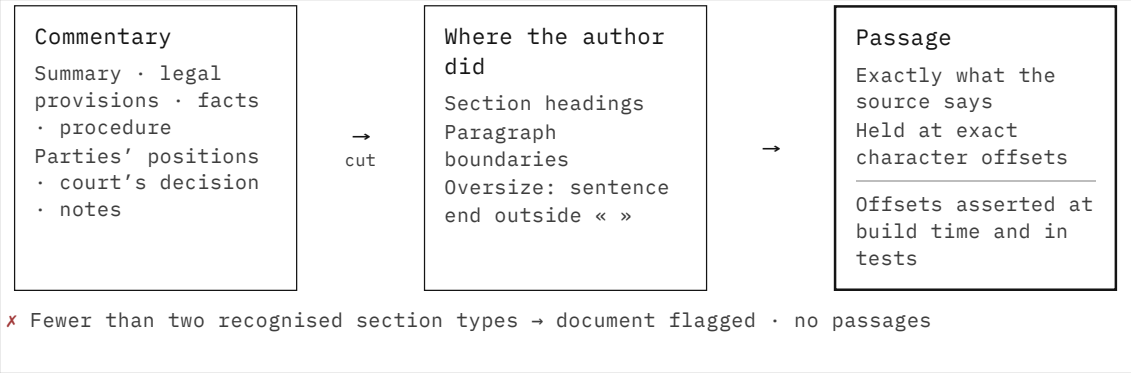


Figure 3 · Structure-preserving segmentation: passages cut where the author did

A verifier can only check a quotation against a passage if the passage is exactly what the source says. No passage ends mid-sentence and no passage crosses a heading. A document with fewer than two recognised section types is flagged and produces no passages rather than badly cut ones.

8. Retrieval

The decision records cited in this section are explained for jurists in Appendix B.

In labour law, exact terms matter — "motif grave", an article number, a roll number — and so does meaning. Heuris combines both. A sparse index (BM25 with a French stemmer, keeping legal references such as "art. 35" as tokens) and a dense index (the multilingual bge-m3 model, run locally, stored in a local LanceDB table) are fused by reciprocal rank fusion with k = 60.[27] Both indexes, lexical and vector, run on the local machine. The top 50 fused passages are reranked by the bge-reranker-v2-m3 cross-encoder. When a case file is analysed, each sub-question produces its own query, and the per-decision ranks are fused again by reciprocal rank fusion (D4–D7, D17). Embedding and reranking run on the local machine; no corpus text is sent to a hosted embedding service. Generation is different: the retrieved passages are sent, with the question, to the external LLM API of the configured provider — DeepSeek by default — which drafts the answer (D11; *project documentation (not public)*).

The retrieval evaluation is small and is reported as such. Fifteen self-written, provisional questions give recall@5 of 1.00 in every mode, and mean reciprocal rank of 1.00 (BM25), 0.92 (dense), 0.97 (hybrid) and 0.96 (hybrid with rerank). The set does not discriminate between modes, so it supports no claim that one mode is better. The hybrid design rests on the engineering argument above, not on these numbers. (Sources: *project documentation (not public)*; *project documentation*

(not public); *project documentation (not public)*; *project documentation (not public)*, *project documentation (not public).*)

Table 8.1 · Retrieval modes

Mode	Strength	Typical failure	Role in Heuris
BM25 only	Exact legal terms, references	Misses paraphrase	Component
Dense only	Meaning, paraphrase	Blurs exact terms and numbers	Component
Hybrid (RRF)	Both signals, rank-based fusion	Inherits both indices' blind spots	Default candidate list
Hybrid + rerank	Best passages first	Slower on CPU	Default ranking

9. Fail-closed verification

The decision records cited in this section are explained for jurists in Appendix B.

The main difference from a conventional retrieval-augmented assistant lies in verification. The model may only point at evidence by identifier: decision passages, and case-file paragraphs whose quotation was located verbatim in the file. Verification covers the sentences generated by the model: every such sentence is decomposed into claims, and each claim is subjected to four successive checks. If it fails any of them, it is removed. Sentences written by the jurist in jurist mode are not verified this way: their quotations are cut verbatim by the server, hence exact, but no entailment check is run on them (D53; § 10). (Source: *project documentation (not public)*; *project documentation (not public).*)

9.1 Verification checks

1. **Pre-filter.** Rule-based removal of claims that cannot be verified as such, for example negative claims or meta-claims about a decision (D16).
2. **Verbatim check.** The claim's quotation must be found in full at its cited offsets, tolerant only of whitespace and apostrophe variants, and is displayed with the source's original characters (D9, D24).
3. **Entailment check.** A second LLM call must judge that the quoted passage supports the claim (D9).
4. **Deterministic guards.** Every number, month, date or amount in a case-fact claim, and every date or amount in a conclusion, must come from the quotation or from a verified premise (D23, D29).

9.2 Answer-level rules

- A result requires at least one claim verified against a **decision** passage; case-file claims alone do not make a case-law answer (D15).
- For each decision, the system attempts one **fact → rule link**: a fact from the case file and a passage of the commentary reporting the ground on which the court relied, both quoted, both verified. The passage may be in the commentator's words; the label « fait → règle » does not imply that the decision creates a rule. If either half fails, the interface says « Lien avec le dossier non établi » (D30, D31).

- The **application step**, which relates a rule to the case, uses verified premises only and is worded as an indication. When premises are missing it does not conclude: the interface shows « Application au dossier — absence de conclusion » ("application to the case file — no conclusion") followed by « Éléments manquants : … » ("missing elements"), listing what is missing (internal status value `insufficient`; D19, D25).

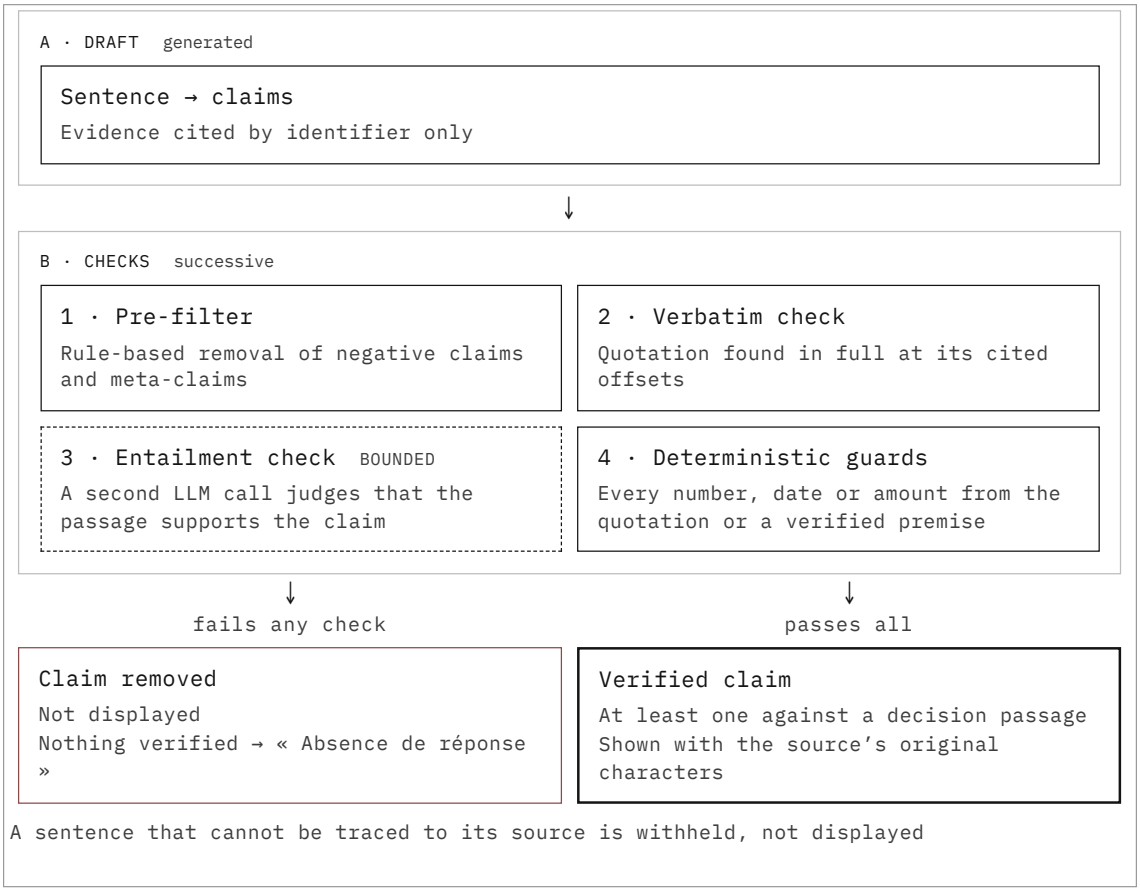


Figure 4 · Fail-closed verification: four checks per claim, refusal by default

Every sentence generated by the model is decomposed into claims, and each claim is subjected to four successive checks. If it fails any of them, it is removed.

A result requires at least one claim verified against a decision passage. When no claim can be supported, the system says so instead of answering.

9.3 A heuristic confidence score

Each result carries a score combining reranker relevance (weight 0.40), the share of drafted claims that passed verification (0.30), factual similarity between the case and the decision (0.20) and recency (0.10, halving every five years). It is forced to 0 when nothing is verified. Plain-language bands translate the score: « Lien fort » requires a score of at least 0.78 **and** a verified fact → rule link; « Lien partiel » starts at 0.65; « Lien faible » below. The weights and thresholds are hand-set and uncalibrated, and the interface labels them so (D10, D18, D34; *project documentation (not public)*; *project documentation (not public)*).

Table 9.1 · What each approach guarantees about a cited passage

	Free generation	**Retrieval-augmented generation**	**Fail-closed grounding (Heuris)**
Passage exists	No guarantee	Retrieved, usually	Identified by server-owned id
Quotation exact	No	Not checked	Full span checked at offsets
Passage supports claim	No	Not checked	Entailment check, else removed
Numbers and dates	Unchecked	Unchecked	Deterministic guard
When evidence is missing	Answers anyway	Often answers anyway	Removes the claim, or refuses

Refusal by default entails a cost: some correct statements are removed because they could not be verified. This cost is accepted deliberately, since a jurist can read a decision the system did not summarise, whereas a sentence the system should not have written is considerably harder to detect.

9.4 Worked example (mock mode, not an evaluation)

To illustrate the stages, one of the fifteen self-written retrieval questions (q02: « Les chauffeurs Uber sont-ils liés par un contrat de travail ? ») was run through retrieval and verification on 4 November 2026 in mock mode. The language model was replaced by the deterministic stand-in used in the automated tests; no external API was called and no case file was supplied. In this mode the entailment check always passes, and the stand-in, by design, emits an unverifiable quotation when it finds no usable passage. The run therefore illustrates the mechanism only; it is not an evaluation and supports no performance claim. (Command, script and log: *project documentation (not public).*) Grounding prompt `legal-ground-v4`; output schema `legal-grounding-v1`. The example run uses the same pipeline with a mock model (entailment always true), so it shows the flow, not verification quality (*project documentation (not public)*, `schema.py:7`). The example says nothing about the legal framework governing the qualification of a work relationship, which this paper does not describe.

Retrieval returned eight passages. Six, including the first five, came from the commentary listed as the provisional expected document for this question (`4200`); the remaining two came from two other commentaries. The stand-in drafted a single claim, citing the top-ranked passage, whose quotation does not occur in that passage. Verification rejected it (`verify: fail` in the audit trace), no claim remained, and the system returned « Absence de réponse : aucune affirmation vérifiable dans les sources. » The example thus shows the fail-closed rule at work: a sentence that cannot be traced to its source is withheld, not displayed.

10. The jurist's authority and the workbench

The decision records cited in this section are explained for jurists in Appendix B.

System output is subject to the jurist's validation; the decision remains with the jurist. The interface is a three-column workbench: the case sheet and case file on the left, results in the middle, the full text of the decision on the right, with numbered passage markers (D13, D43, D45).

- **Case sheet.** A typed, sourced summary of the case file — facts, provisions, dated events — in which every field quotes the file. The jurist validates it before searching (D28, D35).
- **Questions and arguments.** Client questions and opposing arguments become target items. Items the system infers must be validated by the jurist before they are searched (D46, D47).
- **Indicative stances.** For each item, a decision may be labelled as tending to support or contradict it, but only when the pairing is verified. The label describes the direction of one decision's reasoning as reported in a commentary and carries no precedential weight. Group summaries count these labels (e.g. « 2 tendent à soutenir »; *project documentation (not public)*); such a count is not a measure of the state of the case law. Otherwise the label is « non établi ». A stance is an indication for the jurist, never a conclusion.
- **Drafting assistant (optional).** The draft is displayed below the retrieved decisions and does not replace them. A draft answer has at most six sentences, each verified like any other claim. When the linked decisions contain no sufficiently verified passage, the panel shows « Les décisions rattachées à cette question ne comportent pas de passage vérifié suffisant pour proposer un projet de réponse. » ("the decisions linked to this question contain no sufficiently verified passage to propose a draft answer") and « Cela ne préjuge pas de leur pertinence. » ("this does not prejudge their relevance"), with « Consulter les décisions » and « Rédiger soi-même » (internal status value `insufficient`; D48, D51, D52).
- **Jurist mode.** The jurist writes the answer and attaches citations by selecting text in the decision or the case file; the server cuts the exact quotation from the offsets. These quotations are exact, but no entailment check is run on them (D53).
- **Validated answers.** Validation creates an immutable version; withdrawal is recorded, not erased. Answers export to Word with their citations; per-sentence verification status appears only in the export's annex of initial drafts, and the export does not indicate that jurist-mode quotations were not entailment-checked (*project documentation (not public)*); and an audit trace records each step (D49, D50).

The drafting assistant and the workbench features above have been exercised in simulated-LLM mode only (§ 13). (Sources: *project documentation (not public)* D12–D13, D35–D54; *project documentation (not public)*; *project documentation (not public)*, *project documentation (not public)*, *project documentation (not public)*.)

11. Confidentiality and data protection

Preparing a file brings the client's facts into the query and into the material the system processes (§ 2.2). Professional secrecy and the GDPR[29] therefore make the path taken by the text of the case file a design constraint, not an operational detail. The generic retrieval-augmented architecture works against this constraint: passages are embedded through a hosted API, stored in a managed vector-database service and sent to a remote language-model API. Each of these steps transmits passages, embeddings or queries to a third party, under that party's terms and in its jurisdiction.

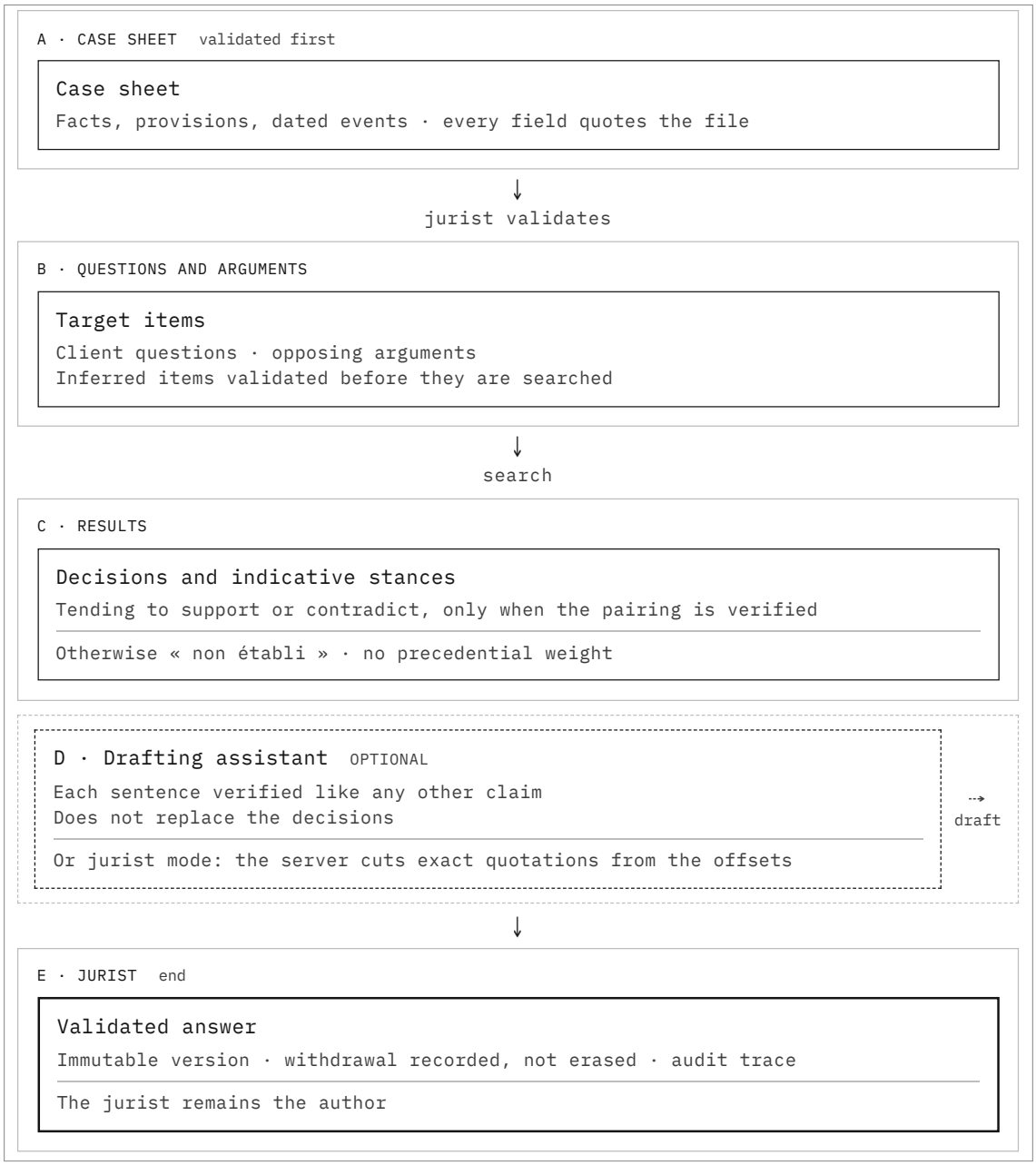


Figure 5 · The workbench: the system proposes, the jurist validates

System output is subject to the jurist's validation; the decision remains with the jurist.

No search is run before the jurist has validated the case sheet, and questions inferred from the file are not searched until the jurist validates them. A stance is an indication for the jurist, never a conclusion.

Heuris is built the other way for every step except generation. The measures are listed below, with two exceptions stated at the outset: drafts and validated answers, which may quote the case file, are stored on disk unencrypted and without a retention limit; and the language model used by the proof of concept is operated outside the European Union.

- The file is parsed in memory with local libraries (`pypdf`, `python-docx`); no hosted parsing service is used (D8).
- Case text is never written to disk for indexing and is never indexed into the corpus. Its paragraphs are citable evidence only within the session.
- Sessions live in memory for at most four hours and are purged on request, on expiry and at shutdown. Logs record session ids and sizes, never prompts, responses or case content.
- Retrieval runs on the local machine: the BM25 index, the bge-m3 embeddings stored in a local LanceDB file, and the bge-reranker-v2-m3 cross-encoder (§ 8). No corpus text, embedding or query leaves the machine for retrieval. A vector-database server was not adopted: at the present corpus size it adds nothing operationally (D5). Embeddings are computed locally, so no corpus text is sent to a hosted embedding service (D4).
- The server binds to the local machine only. There is no authentication.

One documented exception. Drafts, jurist edits and validated answers are written to disk as JSON, one file per session. They can contain quotations from the case file, are not encrypted, have no retention limit, and outlive the four-hour session; they are deleted only when the session is explicitly closed (06 § 1b).

The language model. Generation and the entailment check (§ 9) form a separate boundary. Both are calls to an external language model, and both send it the retrieved passages, the question, the case sheet and case-file excerpts (D11). The provider is set by a configuration parameter. The proof of concept uses DeepSeek, for which API credit was already available (D11). Its privacy policy states that it may collect "your text input, voice input, prompt, uploaded files, photos, feedback, chat history, or other content that you provide to our model and Services"; that it retains personal data "for as long as necessary to provide our Services and for the other purposes set out in this Privacy Policy", purposes that include "to train and improve our technology"; and that "we directly collect, process and store your Personal Data in People's Republic of China".[28] DeepSeek's separate API platform terms were not reviewed, and no zero-retention commitment was found. The project is therefore restricted to public data and fictitious or anonymised case files. When the provider is DeepSeek, the interface asks for explicit confirmation before the case file is transmitted, and repeats that only fictitious or anonymised files may be submitted (06 § 1).

Live use is intended on an EU-hosted provider (for example Mistral) or on a local model run on rented infrastructure (for example an H200 GPU server), with no retention and with processing terms consistent with the GDPR and professional secrecy. This is the design target, not the default of the proof of concept (D11; `06_confidentiality.md` § 4; planned, not implemented). The terms of the providers named here have not been reviewed.

The EU AI Act classifies as high-risk AI systems "intended to be used by a judicial authority or on its behalf to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts" (recital 61; Annex III, point 8(a)).[30] Heuris is designed for jurists in general. Whether a given deployment falls within that category depends on who uses it and on whose behalf, and is not assessed here. The wider regulatory analysis remains outside this paper.

12. Comparison by category

Heuris is not a substitute for the tools jurists already use, in any field, and this paper does not claim otherwise.

General-purpose assistants offer breadth and speed of drafting. Commercial legal databases offer what Heuris lacks entirely: official texts, editorial curation, wide coverage, and in some cases their own AI layers. The study cited in §§ 1 and 5 found that legal-specific tools hallucinated less than a general chatbot, and that the best of them answered a majority of queries accurately.[1]

The proof of concept applies a narrower method end to end to a small, lawful corpus — today a single labour-law corpus. The method is not specific to labour law; extending it to other fields is planned, depending on data access, and is not implemented. A sentence is displayed only if it is traced to its source. The comparison below is by category; no product is named or assessed.

The retrieval, verification and traceability architecture is not specific to social law. Moving to another field would mean replacing the corpus and its ingestion and chunking adapter, and adapting a domain layer: the case-sheet schema (dispute types, party and date roles), the LLM prompts, the query-expansion terms and the interface labels. The current build is also tied to French (stemming, date parsing, prompts) and to Belgian court conventions. No cross-field measurement exists: the only retrieval evaluation, on 15 self-written questions, concerns the labour-law corpus alone and does not distinguish the retrieval modes (§ 8, § 13). Extension to other fields remains planned, depending on data access. (Source: *project documentation (not public)*, *project documentation (not public)*, *project documentation (not public)*, *project documentation (not public)*.)

Table 12.1 · Category comparison

	General assistant	**Commercial legal database**	**Heuris (PoC)**
Source coverage	Open web / training data	Broad, curated, official texts, many fields	One field (labour law): more than 25,000 decisions, no official texts
Provenance per sentence	Not by design	Varies by product	Required, verified at offsets
When evidence is missing	Tends to answer	Varies by product	Removes claims or refuses
Where data is processed	Provider's infrastructure	Provider's infrastructure	Local, except the LLM call (§ 11)
Commercial availability	Yes	Yes	No

A proof of concept does not replace established tools; the two approaches may be complementary.

13. Limits of the study

The limits of the study are set out below.

1. **No accuracy figure.** No accuracy or hallucination rate is claimed. No lawyer has evaluated any answer.
2. **Mock-mode testing.** The 105 automated tests pass using fakes, not a live model. The interface and the drafting assistant (D46–D54) were exercised in simulated-LLM mode only.

3. **Single live runs.** The live model has been run only in single runs on fictitious test files. Run-to-run variance is unknown. These runs are not a benchmark.
4. **Application step.** In the live test runs of the latest grounding version (v4), the application step concluded on 0 of 3 legal questions, against 3 of 3 for similar questions with the previous version (v3). The cause has not been established.
5. **Uncalibrated scores.** Confidence scores and the « Lien fort / partiel / faible » bands are heuristic, with hand-set thresholds.
6. **Retrieval evaluation.** Fifteen self-written questions do not discriminate between retrieval modes and are not evidence of retrieval quality.
7. **Stored answers.** Drafts and validated answers on disk are not encrypted, have no retention limit, outlive the session and may contain case-file quotations.
8. **No access control.** The API has no authentication; it binds to the local machine only.
9. **Non-EU model.** DeepSeek operates outside the EU, and retrieved passages are sent to its API for generation (D11).[28] Only public data and fictitious or anonymised files may be used.
10. **Licence.** The publisher's terms state: « La reproduction des informations textuelles du site est autorisée pour autant que la source soit clairement citée. » and « Toute exploitation commerciale des informations est interdite. »[26] They say nothing on non-commercial professional use; whether use inside a firm that bills its clients counts as commercial exploitation is not settled by the terms. Any paid offering would require an agreement with the publisher or another source.
11. **No official texts.** Heuris does not crawl Juportal, the official SPF Justice database (robots.txt). There are no official full-text decisions and no ECLI.
12. **Commentaries, not decisions.** The corpus consists of commentaries that quote extracts of the reasoning; an operative part (*dispositif*) is mostly absent. The interface nevertheless labels items « décision » and the source panel with the publisher's name (*project documentation (not public)*), never « commentaire »; relabelling is a proposal not yet decided, and no decision record exists.
13. **Near-duplicates.** Records that differ only in punctuation survive deduplication.
14. **Jurist-mode quotations.** Quotations inserted in jurist mode are exact but not entailment-checked.
15. **Name and status.** Trademark and domain checks for "Heuris" have not been done. Heuris is not commercially available. Nothing here is legal advice.
16. **One field only.** Only the Belgian labour-law corpus is implemented; all figures concern it. No other legal field is implemented.
17. **No appeal or cassation follow-up.** The corpus records nothing on appeal, cassation or later developments, nor on changes in the legislation applied; a result may concern a decision later reversed.
18. **Recency in the score.** The confidence score includes recency (weight 0.10, halving every five years; *project documentation (not public)*, *project documentation (not public)*). Recency is not authority; the weighting is a design choice that no measurement supports.
19. **Sample bias.** The corpus of more than 25,000 decisions comes from commentaries selected by one publisher, in French only; Dutch-language case law is absent, and published case law is itself a selection (§ 1.3). An answer, or the absence of one, says nothing about the state of the case law.

20. **Professional secrecy and data protection.** Professional secrecy and the GDPR[29] are named but not analysed. This paper does not assess when a file counts as anonymised in the GDPR sense. No real client file may be processed until a local or EU-hosted model with no retention is in place (§ 11).
21. **Meaning of "verified".** "Verified" means that the excerpt matches the source text, here a commentary; the entailment check is a model judgement and has not been audited by a human.
22. **Export.** The Word export does not carry the verification status of the validated answer's citations.

(Sources: `05_evaluation-protocol.md`; `04_grounding-verification.md` § 13; `06_confidentiality.md`; `01_data-sources.md`; decision records D10, D11, D34, D46–D54; pytest run at `9ceeefe`.)

14. Outlook and planned work (all items planned)

- **Other legal fields.** New corpora in further fields where lawful data is available — planned, depending on data access.
- **Evaluation.** A question set written by labour lawyers for the first corpus; grounding and refusal metrics; calibration of the confidence score against lawyer judgements.
- **Model.** An EU-hosted provider or a model run on rented infrastructure, with no retention, before any real file.
- **Hardening.** API authentication; local OCR for scanned files; near-duplicate detection; a refusal threshold on reranker scores; encryption and retention limits for stored answers.
- **This paper.** French version — done (`heuris_whitepaper_fr.md`).
- **Licensing.** Any commercial step only with the publisher's agreement or another lawful source.

(Sources: *project documentation (not public)*, planned rows; `05_evaluation-protocol.md`; `06_confidentiality.md` § 4; `01_data-sources.md` § 7.)

ID	Content	Source

Figures and tables

ID	Content	Source
F1	Default pattern: LLM at the centre, sources optional	*project documentation (not public)* (figure plan)
F2	Heuris spine: Sources → Ingest and deduplication → Passages → Retrieval → Fail-closed verification → Jurist	`00_architecture.md` pipeline
T1	Corpus: more than 25,000 decisions · French · deduplicated by a hash of the normalised text	`01` §§ 4–5; `02` § 5; `dedupe_report.md`
T2	Retrieval evaluation (15 self-written questions): recall@5 1.00 all modes; MRR BM25 1.00 · dense 0.92 · hybrid 0.97 · hybrid + rerank 0.96 — non-discriminating	`retrieval_report.md`; `questions.jsonl`
P1	The problem (§ 1): general LLMs ≥ 58 % hallucination (U.S. law); legal RAG tools 17–33 % (U.S. law); about 1.4 million Belgian decisions a year; about 1 % published (estimate); Antwerp sanction €25,000; Belgian justice budget 102.5 € per inhabitant (2022, above CoE median), below EU average (Commission 2025); 246 days first instance (2023); Court of Audit press release (10 January 2025)	Notes 1, 3, 11, 13, 14, 16–19
T3	Automated tests: 105 passed at `9ceeefe`, fakes only	pytest, re-run 2026-11-04
Table 5.1	Requirements R1–R6 and their status for Heuris (assessed against the facts in this paper)	§ 3, § 5
Tables 6.1, 7.1, 8.1, 9.1, 12.1	Qualitative comparisons	§§ 6–12

Interface illustrations, if added, are labelled “Interface specimen — simulated LLM, fictitious data”.

Glossary

Term	Meaning in this paper
Passage	A structure-aware segment of a source, with exact character offsets
Commentary	A text by the publisher about a decision; not the decision itself
Decision reference	Court (with division where recorded), date and *R.G.* roll number, as recorded, e.g. « C. trav. Liège (div. Liège), 14 janvier 2026, R.G. 2025/AL/438 »; the chamber is not recorded
Claim	One atomic statement in a generated answer
Verbatim check	Full-span match of a claim’s quotation at its cited offsets (rule: the excerpt must appear verbatim in the source)

Term	Meaning in this paper
Entailment check	Model check that the quoted passage supports the claim
Fail-closed	On failed verification, the claim is removed or the answer refused; never shown
Case sheet	Typed, sourced summary of a case file, validated by the jurist
Fact → rule link	One verified pair of a case-file fact and a commentary passage reporting the ground on which the court relied
Verified	The excerpt matches the source text (here, a commentary) at its cited offsets; not proof that the court said it
Labour tribunal (*tribunal du travail*)	First-instance labour court; gives *jugements*
Labour court of appeal (*cour du travail*)	Appeal court in labour matters; gives *arrêts*
Court of Cassation (*Cour de cassation*)	Belgium's supreme court in the judicial order
Application step	Relating a rule to the case from verified premises only; otherwise no conclusion (« Application au dossier — absence de conclusion », with the missing elements)
Indicative stance	Verified "tends to support / contradict" label; never a conclusion
Simulated LLM (mock mode)	Deterministic stand-in model for tests and demonstrations
Interface specimen	An illustration produced in mock mode with fictitious data
Language model (LLM)	A statistical model that generates text; in Heuris it drafts and checks, but never decides what is shown
Retrieval-augmented generation (RAG)	Generation conditioned on passages retrieved from a corpus[2]
Hallucination	In Magesh et al., a response that is incorrect, or whose citation does not support the proposition[1]
Attribution	Whether a generated statement is supported by an identified source (cf. AIS)[20]
BM25	A keyword-based ranking function used for the sparse index
Dense embedding	A numerical representation of a passage's meaning (here, bge-m3), used for semantic search
Vector database	A store of dense embeddings searched by similarity; in Heuris, a local LanceDB file
Reciprocal rank fusion (RRF)	A method that merges several rankings using the rank of each item, with constant k = 60
Reranker (cross-encoder)	A model that rescores a question–passage pair jointly (here, bge-reranker-v2-m3)
Offset	The position of a character in a source text; quotations are located by start and end offsets
Entailment (natural-language inference)	The relation in which one text supports the truth of another[24]

Term	Meaning in this paper
Prompt	The instructions and sources sent to the language model for one call
API	Programming interface through which the software calls the language-model provider
robots.txt	A file by which a website states which automated agents may collect its pages
ECLI	European Case Law Identifier
Central register (JustJudgment)	The Belgian central register of judicial decisions; not open to the public as of mid-2026[13]
Decision record (D-record)	A numbered architecture decision of the project, D1–D54; see Appendix B

Notes

Internal sources are shown as project documentation (not public).
Each note states what the source supports.

1. V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools," arXiv:2405.20362, https://arxiv.org/abs/2405.20362 (accessed 2026-11-04). *Supports:* a preregistered evaluation of RAG-based commercial legal research tools; they still hallucinate 17–33 % of the time, and the authors find them "less prone to hallucination than GPT-4" (§ 5); a citation that does not support its proposition counts as a hallucination; the best-performing tool tested, Lexis+ AI, answered 65 % of the queries accurately (§ 1, § 5). (Study of U.S. law; cited for the category, not for Belgian law.)
2. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, https://arxiv.org/abs/2005.11401 (accessed 2026-11-04). *Supports:* the definition of retrieval-augmented generation, and that provenance for model outputs was an open problem the technique addressed only in part.
3. M. Dahl, V. Magesh, M. Suzgun, D. E. Ho, "Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models," *Journal of Legal Analysis* (2024), arXiv:2401.01301, https://arxiv.org/abs/2401.01301 (accessed 2026-11-04). *Supports:* LLMs "hallucinate at least 58% of the time, struggle to predict their own hallucinations, and often uncritically accept users' incorrect legal assumptions"; more than 800,000 queries on GPT-4, GPT-3.5, PaLM 2 and Llama 2, on U.S. federal case law.
4. Z. Xu, S. Jain, M. Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models," arXiv:2401.11817, https://arxiv.org/abs/2401.11817 (accessed 2026-11-04). *Supports:* "we formalize the problem and show that it is impossible to eliminate hallucination in LLMs".
5. A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, "Why Language Models Hallucinate," arXiv:2509.04664, https://arxiv.org/abs/2509.04664 (accessed 2026-11-04). *Supports:* hallucinations "persist even in state-of-the-art systems"; they persist "due to the way most evaluations are graded", guessing when uncertain improving test performance.
6. H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, https://arxiv.org/abs/2307.09288 (accessed 2026-11-04). *Supports:* Table 10, pretraining-data language shares "en 89.70%" and "fr 0.16%"; "Most data is in English".
7. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, https://arxiv.org/abs/2303.08774 (accessed 2026-11-04). *Supports:* figure "GPT-4 3-shot accuracy on MMLU across languages" (machine-translated MMLU): English 85.5 %, French 83.6 %.

8. S. Singh et al., "Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation," arXiv:2412.03304, https://arxiv.org/abs/2412.03304 (accessed 2026-11-04). *Supports:* "for questions requiring geographic knowledge, an astounding 84.9% focus on either North American or European regions".
9. J. Myung et al., "BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages," arXiv:2406.09948, https://arxiv.org/abs/2406.09948 (accessed 2026-11-04). *Supports:* "LLMs perform better for cultures that are highly represented online, with a maximum 57.34% difference in GPT-4".
10. *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC) (S.D.N.Y. 22 June 2023), Opinion and Order on Sanctions, https://www.nhd.uscourts.gov/sites/default/files/pdf/Mata-v-Avianca-sanctions-order.PDF (accessed 2026-11-04). *Supports:* submission of "non-existent judicial opinions with fake quotes and citations created by the artificial intelligence tool ChatGPT"; "A penalty of $5,000 is jointly and severally imposed on Respondents".
11. M. Joris, « Un avocat condamné pour avoir remis des conclusions fantaisistes rédigées par l'intelligence artificielle », RTBF, 25 December 2025, https://www.rtbf.be/article/un-avocat-condamne-pour-avoir-remis-des-conclusions-fantaisistes-redigees-par-l-intelligence-artificielle-11653006 (accessed 2026-11-04). *Supports:* the Antwerp Court of Appeal decision (€2,500 fine, €7,500 per party to three parties, €25,000 in total; "jurisprudence inexistante"); the Ghent business court case of 15 December 2025. Press report; the judgments were not consulted.
12. A. Ockerman, "Arrest Gentse hof van beroep benadrukt noodzaak van zorgvuldigheid bij gebruik van artificiële intelligentie", Orde van Vlaamse Balies, 17 October 2025, https://www.ordevanvlaamsebalies.be/nl/nieuws-en-events/arrest-hof-van-beroep-onderstreept-de-noodzaak-voor-advocaten-om-zorgvuldig-om-te-springen-met-artifici%C3%ABle-intelligentie (accessed 2026-11-04). *Supports:* submissions with "een waslijst aan verkeerde citaten uit bestaande wetsartikelen en verwijzingen naar onbestaande rechtspraak en rechtsleer". Professional-body report; the judgment was not consulted.
13. SPF Justice, « Déjà 800 000 jugements disponibles sur Just-on-web » (2026), relayed at https://blog.forumforthefuture.be/fr/article/deja-800-000-jugements-disponibles-sur-just-on-web/31929 (accessed 2026-11-04). *Supports:* « près de 1,4 million de décisions sont prononcées chaque année en Belgique »; only the parties concerned and their lawyers have access to their own judgments; the judgments (*arrêts*) of the labour courts of appeal (*cours du travail*) are to be added to JustJudgment "in the coming months".
14. V. Hendrickx, "The Belgian database for case law – path towards legal analytics?", KU Leuven CiTiP blog, 3 February 2023, https://www.law.kuleuven.be/citip/blog/the-belgian-database-for-case-law-path-towards-legal-analytics/ (accessed 2026-11-04). *Supports:* "only a limited number of all judicial decisions are made public (e.g. 1% in Belgium or 4% in the Netherlands)". An estimate; method not stated.
15. SPF Justice, "Nieuwe JUPORTAL zoekmachine vervangt Jure-Juridat" (December 2020), https://justitie.belgium.be/nl/nieuws/andere_berichten/nieuwe_juportal_zoekmachine_vervangt_jure_juridat (accessed 2026-11-04). *Supports:* "de individuele rechtscolleges bepalen welke vonnissen en arresten zij anonimiseren en publiceren in JUPORTAL"; labour courts of appeal (*arbeidshoven*): "Een selectie van arresten (meer dan 21.000 publicaties)".
16. SPF Justice, « Le nouveau moteur de recherche JUPORTAL remplace Jure-Juridat » (December 2020), https://justice.belgium.be/fr/nouvelles/autres_communiques/le_nouveau_moteur_de_recherche_juportal_remplace_jure_juridat (accessed 2026-11-04). *Supports:* replacement because « la technologie flash [...] ne sera effectivement plus supportée »; « la seule banque de données publique comportant une sélection de jugements et d'arrêts anonymisés »; search « via n° ECLI, n° de rôle et dates » and « Options - Sélections Juridiction - Domaine juridique - Type - Année de publication »; RSS feed. No mention of semantic search.

17. CEPEJ (Council of Europe), *European judicial systems — CEPEJ Evaluation Report, 2024 Evaluation cycle (2022 data): Belgium*, https://rm.coe.int/belgium/1680b1df5c (accessed 2026-11-04). *Supports:* "In 2022, the implemented budget of the judicial system is €1,199,563,192, representing 102,5 € per inhabitant and exceeding the CoE median. Conversely, as a percentage of GDP (0,22%), this budget is below the CoE median"; "the number of pending civil and criminal cases is not available for 1st and 2nd instances".
18. European Commission, *2025 Rule of Law Report — Country Chapter on the rule of law situation in Belgium*, SWD(2025) 901 final, 8 July 2025, https://commission.europa.eu/document/download/5fe5bb71-2898-4079-ad5a-19a3ff595a62_en?filename=5_1_63936_coun_chap_belgium_en.pdf (accessed 2026-11-04). *Supports:* "remain below the EU average" (budget, judges per 100 000 inhabitants); "246 days in 2023"; "persistent lack of data on court proceedings"; "online access to published judgments and machine-readable judicial decisions".
19. Court of Audit (Cour des comptes), « Pilotage de la transformation numérique de la justice par l'État fédéral », press release, 10 January 2025, https://www.ccrek.be/sites/default/files/Docs/2025_01_PilotageTransformationNumeriqueJustice_Communique.pdf (accessed 2026-11-04). *Supports:* « la transformation numérique de la justice n'est pas menée de manière satisfaisante »; « la numérisation de la justice reste en chantier depuis plusieurs décennies ».
20. H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, D. Reitter, "Measuring Attribution in Natural Language Generation Models," arXiv:2112.12870, https://arxiv.org/abs/2112.12870 (accessed 2026-11-04). *Supports:* the AIS ("Attributable to Identified Sources") framework and its two-stage annotation pipeline for judging whether model-generated statements are supported by underlying sources.
21. B. Bohnet et al., "Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models," arXiv:2212.08037, https://arxiv.org/abs/2212.08037 (accessed 2026-11-04). *Supports:* the formulation of attributed question answering; a reproducible evaluation framework; human annotations as gold standard and a correlated automatic metric.
22. T. Gao, H. Yen, J. Yu, D. Chen, "Enabling Large Language Models to Generate Text with Citations," EMNLP 2023, arXiv:2305.14627, https://arxiv.org/abs/2305.14627 (accessed 2026-11-04). *Supports:* the ALCE benchmark and its automatic metrics for fluency, correctness and citation quality.
23. S. Min et al., "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation," EMNLP 2023, arXiv:2305.14251, https://arxiv.org/abs/2305.14251 (accessed 2026-11-04). *Supports:* decomposition of a generation into atomic facts and the share supported by a knowledge source; an automated estimator using retrieval and a language model.
24. O. Honovich et al., "TRUE: Re-evaluating Factual Consistency Evaluation," NAACL 2022, arXiv:2204.04991, https://arxiv.org/abs/2204.04991 (accessed 2026-11-04). *Supports:* across 11 datasets, "large-scale NLI and question generation-and-answering-based approaches achieve strong and complementary results".
25. Juportal, robots.txt, https://juportal.be/robots.txt (accessed 2026-11-04). *Supports:* `User-agent: * / Disallow: /`, with `Allow: /` only for `DG_JUSTICE_CRAWLER`.
26. The publisher's website terms, « Copyright et conditions d'utilisation du site » (accessed 2026-11-04). *Supports:* reproduction of textual information is authorised provided the source is clearly cited; any commercial exploitation is forbidden.
27. G. V. Cormack, C. L. A. Clarke, S. Büttcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," SIGIR 2009 — **citation pending** (publisher page not reachable for verification). *Would support:* the RRF method and the k = 60 constant.
28. DeepSeek, "DeepSeek Privacy Policy", https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html (accessed 2026-11-04). *Supports:* collection of "text input, voice input, prompt, uploaded files"; retention "for as long as necessary to provide our Services and for the other purposes set out in this Privacy Policy",

including “to train and improve our technology”; storage “in People’s Republic of China”. The separate API platform terms were not reviewed.

29. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1, https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed 2026-11-17). *Supports:* the EU legal framework for the processing of personal data named in §§ 11 and 13; not analysed in this paper.
30. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act), OJ L, 12.7.2024, recital 61 and Annex III, point 8(a), https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed 2026-11-17). *Supports:* AI systems “intended to be used by a judicial authority or on its behalf to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts” are to be classified as high-risk; the classification does not extend to “purely ancillary administrative activities”.

Figures and examples are illustrative. “Heuris” is a provisional working name. Not legal advice.