Nuffield Foundation

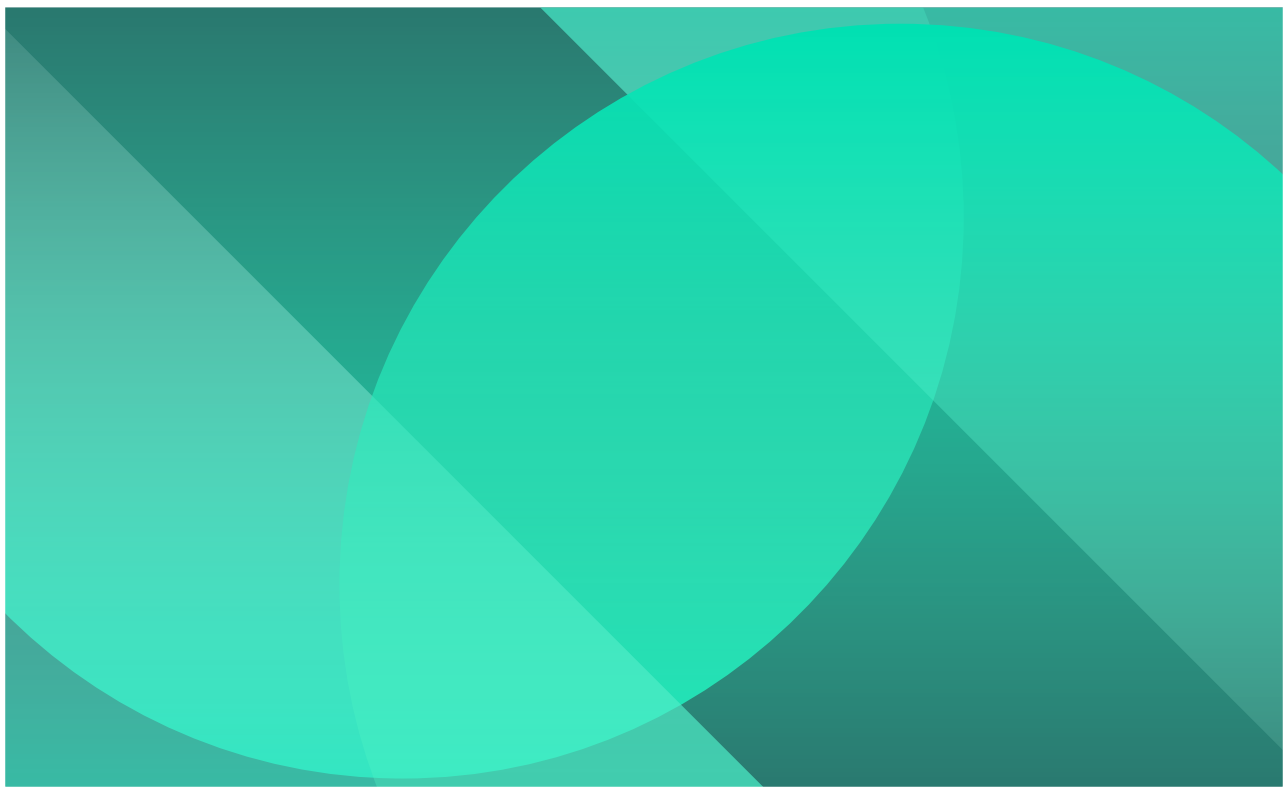
Artificial intelligence & justice

An evidence scoping review

Full report

Dr Holli Sargeant, University of Cambridge

July 2026



This report was commissioned by the Nuffield Foundation as part of its Public right to justice programme. The views expressed are those of the author and not necessarily the Foundation. The report forms part of a series of evidence reviews and briefing papers critically examining different dimensions of the justice system and access to justice in England and Wales. The programme aims to develop an interconnected body of research, providing policymakers, practitioners and the public with a better understanding of how the justice system operates, where it falls short, and how it could better meet the needs of those it serves.

As an independent charitable trust with a mission to advance social well-being, the Foundation funds and undertakes rigorous research, encourages innovation and supports the use of sound evidence to inform social and economic policy, and improve people's lives. It is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute, and the Nuffield Family Justice Observatory.

Find out more at: nuffieldfoundation.org
Bluesky: @nuffieldfoundation.org
LinkedIn: Nuffield Foundation

Executive summary

Artificial intelligence (AI) is being deployed across the justice system in England and Wales faster than its full effects on the public can be independently or transparently evaluated. Some deployments include user-centred design or pilot processes; other use, including the informal use of AI by judges, practitioners, and litigants in person, sits outside any such process. In either case, deployment is materially outpacing the production of independent evidence about its effects on the people the system serves.

In July 2025 the Ministry of Justice (MoJ) published its *AI Action Plan for Justice*, setting out a three-year strategy for embedding AI across the department and the wider justice system, and established a dedicated Justice AI Unit to oversee implementation. Judges across His Majesty's Courts and Tribunals Service (HMCTS) now have access to Microsoft Copilot on the eJudiciary platform, supported by recent judicial guidance. Proprietary large language models (LLMs) are in use across government agencies and legal advice organisations. Speaking at Microsoft's AI Tour in February 2026, the Lord Chancellor described the MoJ as one of the fastest-growing users of Copilot across government. Senior judiciary have publicly endorsed and modelled the use of generative AI in judicial work. Adoption is being framed, in policy and in practice, as a core priority to achieve "swifter, fairer, and more accessible justice".

There is well-motivated ambition and optimism for AI to improve access to justice. A strong case can be made especially where the system is currently under the most pressure – court and tribunal backlogs are creating significant delays, the legal aid and advice sector are capacity-constrained, and the level of unmet legal need is high. Impactful opportunities are visible across various stakeholders of the justice system, and a gain realised for one group frequently depends on or produces a gain for another. For people facing everyday legal problems, AI could drastically improve accessibility. Public-facing legal information, triage, navigation, and intake tools – available at low cost, at any time, and in plain and multiple languages – may reach those whom conventional provision does not, and help them recognise a problem as legal and actionable. For the judiciary, practitioners, and court staff, significant opportunities arise for efficiency and capacity. Automating transcription, document processing, legal research, and case management may reduce administrative load and redirect scarce professional time towards complex, people-centred work – provided AI augments rather than displaces professional judgment. Better access to empirical insight and evaluation may assist justice administrators and policymakers in understanding how the justice system functions. Data-driven analysis of otherwise poorly utilised justice data could show where delay accumulates and legal needs go unmet, and could support more accountable and responsive policy.

At the same time, the public have a right to a justice system that is accessible, fair, and effectively meets their legal needs. Whether the system is delivering on that right under conditions of AI deployment is not only a normative question – it is also an empirical one.

This evidence scoping review, commissioned by the Nuffield Foundation as part of its *Public right to justice* programme, examined the available evidence on AI and data-driven technologies in administrative, civil, and family justice settings in the UK, drawing on international evidence where transferability could be argued. The review mapped 45 tools in use or in development, concentrated in internal-facing and generative AI applications, with publicly available evaluation for a small minority. It characterised the academic, policy, and grey literature by temporal distribution, publication type, domain, and methodology, and synthesised the evidence thematically across individual and group experiences and outcomes, societal factors including public trust and legitimacy, and system efficiency and effectiveness.

The central finding of the review is that AI evaluations have not yet measured the outcomes that matter most from the standpoint of the people the justice system serves. Efficiency and task performance are well represented and show important opportunity: a faster, more accessible system is itself a contribution to access to justice. However, fairness, comprehension, lived experience, differential impact, and procedural legitimacy are under-represented. Tools may be assessed as effective against the metrics by which they are evaluated while the conditions on which the public right to justice depends remain under-measured. Four structural gaps run through the findings:

A measurement gap. No study identified in this review examined the effects on an individual whose matter was triaged, advised on, or decided through an AI-assisted process in a UK justice setting. Evaluations examine the experience of system operators rather than the people whose legal problems are being addressed. They rely on self-reported time savings and satisfaction with outputs rather than independent assessment of whether AI tools improved the efficiency of the resolution of legal issues or the broader justice system. They test tools on narrow tasks in controlled conditions rather than in the complex settings where real-world legal problems arise. Related evidence reveals that certain demographic groups experience worse outcomes in the justice system, and court digitisation projects revealed uneven uptake of digital processes across user groups. Evaluations should measure specifically whether similar patterns of disadvantage or exclusion will arise in the deployment of AI tools. No UK justice-specific study has measured how AI-assisted processes affect individuals, or how effects differ across groups.

This not only is a gap in what has been studied but also reflects a structural limit in the data itself. Many people's legal needs never reach formal advice or adjudication and so are not recorded in legal data – a lack which only increases in publicly accessible legal data. Legal data also rarely contains ground truth. It reflects only those issues that are recorded in formal stages, for example in court judgments. It can be difficult to distinguish systematic biases across groups from valid legal reasoning and differing institutional contexts. The people most likely to be poorly served, and to have limited access to justice, are likely the least visible to empirical analysis of the system.

A deployment gap. The tools studied in the research literature are increasingly not the tools used in practice. UK deployments are dominated by general-purpose proprietary models, which cannot be subjected to the structural inspection of weights, training data, and architecture in the way that open-source or domain-specific systems permit. Behavioural evaluation through benchmarking, red-teaming,[1] and human-interaction studies remains possible, and is being developed, but further

1 Benchmarking produces a score by comparing one model performance against another in constrained settings. Red-teaming is an adversarial testing method where evaluators attack a system to reveal vulnerabilities and flaws. For discussions of strengths and weaknesses of these methods, see Elliot Jones, Mahi Hardalupas and William Agnew, 'Under the Radar?' (Ada Lovelace Institute 2024).

work is required. Evaluative research indicates that legal-domain LLMs are more reliable than general-purpose LLMs for UK legal tasks, and that smaller non-generative models can outperform general-purpose LLMs for some legal applications. The current direction of deployment appears to be shaped by commercial availability rather than by evidence about which designs best serve users with unmet legal needs. These issues also extend beyond formal deployments: publicly accessible general-purpose LLMs such as ChatGPT and Claude are increasingly used informally, by litigants in person and practitioners, introducing challenges that currently sit beyond the purview of formal justice processes but already shape how people encounter them.

A legitimacy gap. While evaluations focus on performance metrics, users primarily experience AI in terms of procedural fairness, dignity, transparency, human engagement, and perceived independent legal judgment. These dimensions are under-represented in the current evidence. Even limited assistive uses of AI in tribunal processes may reduce perceived fairness and dignity among members of the public. Attitudinal and experimental research of this kind offers valuable early insight into how people may respond to AI in justice settings, though it cannot substitute for evidence of how individuals experience AI-mediated processes in practice. Human oversight is perceived as vital to the procedural fairness of AI use, and is treated in policy and judicial guidance as the principal safeguard, yet no UK deployment evaluation identified in this review examines whether that oversight is effective, calibrated, or sustainable in practice. Judicial and professional concerns about deskilling and over-reliance are articulated but untested.

A transparency gap. Where evaluation exists, it is often internal, partial, or methodologically thin, and the dominance of proprietary models makes external evaluation more difficult. Internal and vendor-led evaluations cannot easily escape the actual or perceived risks of optimism bias, and the apparent effectiveness of some tools may reflect selective reporting rather than demonstrated performance. Independent evaluation capacity needs to grow to enable verification of claims and assumptions about AI's opportunities and risks. The justice system has, historically, a less developed tradition of independent evaluation than other parts of the public sector, and building that capacity is itself a precondition for meaningful scrutiny.

The conditions the review identifies – rapid adoption of proprietary general-purpose models, reliance on self-reported metrics, untested human oversight, and internal evaluation – are present across other high-stakes public service contexts in which AI is being deployed, including criminal justice, social care, and benefits administration. The findings of this review are therefore likely to have application beyond justice settings.

AI is already shaping the conditions under which people encounter the justice system. There is reasonable ground to think that AI, designed and used well, can extend access and reduce friction for users who would otherwise be poorly served. Whether it will advance or undermine the public's right to justice is, at present, an open question, and one the public is entitled to have answered.

Contents

List of abbreviations

Acas	Advisory, Conciliation and Arbitration Service
AI	Artificial intelligence
ASR	Automatic speech recognition
BERT	Bidirectional Encoder Representations from Transformers, a family of encoder-based natural language processing models
Cafcass	Children and Family Court Advisory and Support Service
DBT	Department for Business and Trade
DSIT	Department for Science, Innovation and Technology
FOS	Financial Ombudsman Service
HMCTS	His Majesty's Courts and Tribunals Service
i.AI	Incubator for Artificial Intelligence
ICO	Information Commissioner's Office
LLM	Large language model
ML	Machine learning
MoJ	Ministry of Justice
NLP	Natural language processing
PRTJ	Public right to justice, programme of the Nuffield Foundation
RAG	Retrieval-augmented generation
SCTS	Scottish Courts and Tribunals Service
SEND	Special Educational Needs and Disability
TAR	Technology-assisted review

1 Introduction

Artificial intelligence (**AI**) and data-driven systems are moving rapidly from aspiration to implementation in justice systems. Across the courts and tribunals, public bodies, legal advice organisations, and private legal service providers across the United Kingdom (**UK**), AI tools are actively being explored or deployed at multiple stages of the justice process. This development is taking place against a backdrop of acute systemic pressure: delays, backlogs, and unmet legal need have combined with a long-running shift towards digitisation and self-service. It is unsurprising then that policy now places AI at the centre of "swifter, fairer, and more accessible justice".[2] In England and Wales, among other jurisdictions, this increasing use of AI is framed as an access-to-justice project. AI is expected to reduce backlogs, improve administrative capacity, inform better legal decision-making, support litigants in person, and broadly "improve outcomes for the public and contribute to wider economic growth".[3] The turn to AI in justice also forms part of a wider governmental ambition to extend AI across the public sector by "rewiring the state".[4]

The pace of this adoption has accelerated markedly in recent years. In July 2025, the Ministry of Justice (**MoJ**) published its *AI Action Plan for Justice*, setting out a three-year strategy for embedding AI across the department and the wider justice system.[5] This was accompanied by the creation of a dedicated Justice AI Unit to oversee implementation and a commitment to provide staff with secure, enterprise-grade AI assistants.[6] Speaking at Microsoft's AI Tour in February 2026, the Lord Chancellor and Secretary of State for Justice David Lammy stated that the MoJ is "one of the fastest growing users [of Copilot] across government", and announced the expansion of AI use in courts and tribunals, including to support speech transcription, summarising judgments, and case scheduling.[7]

2 Ministry of Justice, 'AI Action Plan for Justice' (2025) https://www.gov.uk/government/publications/ai-action-plan-for-justice accessed 4 February 2026.

3 ibid; OECD, 'AI in Justice Administration and Access to Justice: Governing with Artificial Intelligence' (OECD 2025) https://doi.org/https://doi.org/10.1787/795de142-en; Ben Hawes and others, 'Just Outcomes: How Can AI Make People's Lives Better?' (Bennett Institute for Public Policy and the Web Science Institute, funded by the Nuffield Foundation 2025) https://www.bennettschool.cam.ac.uk/publications/just-outcomes-how-can-ai-makepeoples-lives-better/ accessed 21 January 2026.

4 Department for Science, Innovation and Technology, 'AI Opportunities Action Plan: One Year On' (*UK Government*, 29 January 2026) https://www.gov.uk/government/publications/ai-opportunities-action-plan-one-year-on/ai-opportunities-action-plan-one-year-on accessed 19 May 2026.

5 Ministry of Justice, 'AI Action Plan for Justice' (n 2).

6 Ministry of Justice, 'Justice AI Unit' https://ai.justice.gov.uk accessed 25 March 2026.

7 The Rt Hon David Lammy MP, 'We Are Calling Time on the Justice System of the Past' (*Speech at Microsoft's AI tour*, 24 February 2026) https://www.gov.uk/government/speeches/we-are-calling-time-on-the-justice-system-of-the-past accessed 4 April 2026.

Judges in His Majesty's Courts and Tribunals Service (**HMCTS**) have had access to Microsoft's Copilot Chat through eJudiciary since April 2025, and refreshed judicial guidance on AI use was issued in October 2025.[8] Judicial use has begun to appear openly in reported practice, for example, in September 2025, a First-tier Tax Tribunal judge became the first in England and Wales to confirm using Copilot in a reported decision, employing it to summarise party documents while retaining all evaluative judgment personally.[9]

Senior judges have also described concrete uses of AI in their own work. Master of the Rolls Sir Geoffrey Vos has spoken extensively about the potential of AI to assist judges, support litigants in person, and improve case management, identifying judicial use of AI to refine judgments, low-cost case-transcript production, and AI support for litigants in person as applications of particular promise.[10] Recently, the Chancellor of the High Court and Lead Judge for AI, Lord Justice Birss, set out four concrete ways in which judges are now using AI: as an in-house transcription tool through ongoing collaborative work with HMCTS and the MoJ; to identify information combinations that risk "jigsaw identification" not obvious to the judge in the production of anonymised judgments; to identify internal inconsistencies in draft judgments (a practice drawing on his own use of the Copilot system); and for administrative tasks including searching emails and files.[11] In his account, these uses do not displace judicial responsibility. Rather, they are presented as tools that support the judge, with the final evaluative decision remaining human. The President of the King's Bench Division, Dame Victoria Sharp, develops this point by contrasting the potential benefits of AI – including greater efficiency and improved access to justice – with the risk that, if insufficiently regulated, AI may undermine the core legal values that sustain the legitimacy of the law and the inherently human dimensions of judicial reasoning.[12]

This broader institutional posture is one of active but bounded engagement. It is visible not only in ministerial policy and judicial speeches, but also in the surrounding procedural and professional infrastructure. In April 2025, the Scottish Courts and Tribunals Services (**SCTS**) published its approach to exploring and developing AI that "will benefit our service users, staff and the judiciary whilst supporting their rights and freedoms".[13] However, it clarifies limits of AI adoption – for example, it is not currently using nor considering the use of AI for any form of judicial decision-making in courts, devolved tribunals, or the Office of the Public Guardian.[14] The Bar Standards Board has published *Guidance on the Use of Artificial Intelligence and Other Technologies* to explain how existing duties

8 Courts and Tribunals Judiciary, 'Artificial Intelligence (AI) – Guidance for Judicial Office Holders' https://www.judiciary.uk/guidance-and-resources/artificial-intelligence-ai-judicial-guidance-october-2025/ accessed 25 March 2026.

9 *VP Evans v HMRC* [2025] UKFTT 1112 (TC), [42]-[49].

10 See recent speeches for example, 'Speech by The Master of the Rolls: Justice for All, Justice for the Accused' (*Courts and Tribunals Judiciary*, 5 February 2026) https://www.judiciary.uk/speech-by-the-master-of-the-rolls-justice-for-all-justice-for-the-accused/ accessed 24 May 2026; 'Speech by the Master of the Rolls: Artificial Intelligence and the Judiciary' (*Courts and Tribunals Judiciary*, 5 May 2026) https://www.judiciary.uk/speech-by-the-master-of-the-rolls-artificial-intelligence-and-the-judiciary/ accessed 24 May 2026.

11 'Speech by the Chancellor of the High Court: Legal Professional Privilege in the Age of AI' (*Courts and Tribunals Judiciary*, 24 April 2026) https://www.judiciary.uk/speech-by-the-chancellor-of-the-high-court-legal-professional-privilege-in-the-age-of-ai/ accessed 24 May 2026.

12 'Speech by the President of the King's Bench Division: The Mayflower Lecture 2025' (*Courts and Tribunals Judiciary*, 19 December 2025) https://www.judiciary.uk/speech-by-the-president-of-the-kings-bench-division-the-mayflower-lecture-2025/ accessed 24 May 2026.

13 Scottish Courts and Tribunals Service, 'Our Approach to the Development of Services Using Artificial Intelligence' (2025) https://www.scotcourts.gov.uk/media/xalno3ff/scts-ai-policy.pdf accessed 13 March 2026.

14 ibid.

and rules apply when using AI.[15] The Civil Justice Council, through a working group chaired by Lord Justice Birss, is consulting on the use of AI by legal representatives in preparing court documents, including whether parties should be required to declare such use.[16] The Online Procedure Rule Committee has developed an Inclusion Framework and Pre-Action Model for the Digital Justice System intended to improve accessibility for online users, by supporting an integrated digital system, alongside the appropriate adoption of AI.[17] Related work on the doctrinal legal issues raised by AI deployment is being explored in consultations run by the Law Commission,[18] and LawTech UK's Jurisdiction Taskforce.[19]

Alongside formal developments, AI is being used informally by judges, lawyers, and litigants in person across a rapidly evolving spectrum of publicly available AI.[20] At one end sit the freely accessible general-purpose large language model (**LLM**) chatbots – such as ChatGPT, Gemini, and Claude – which are widely available without subscription and require no specialist configuration to use. At the next level sit subscription-based products that offer more recent and capable models, longer context windows, and additional features (for example, ChatGPT, Gemini, and Claude all provide a 'pro' subscription). These systems can be customised with domain-specific data and tools: vendors are releasing connectors and plugins that allow the same underlying foundation models to be grounded in legal sources, drafting workflows, or sector-specific materials. In May 2026, Anthropic released over 20 new connectors and plugins for the legal industry, including linking Claude to the United States' Free Law Project, and tools designed for litigants in person and legal aid clinics.[21] In the UK, several government departments collaborated to develop and release a software interface for researchers as well as LLM agents (specifically Claude, Cursor, Copilot) to access and search UK legislation data.[22] These developments in publicly accessible AI, whether free or by subscription, amount to a substantial change in the conditions under which AI is accessible across stakeholders in the justice system. Informal AI use, encompassing the full spectrum, is also visible in the courts: as of May 2026, there had been 64 recorded UK cases involving the poor use of LLMs, including cases in which fabricated legal references were identified; 37 of these arose from litigants in person.[23]

15 Bar Standards Board, 'Guidance on the Use of Artificial Intelligence and Other Technologies' (18 May 2026) https://www.barstandardsboard.org.uk/static/5e1baa5c-614c-4105-afae9a9ccdf8d97b/c0d1ed62-fcad-4a68-8ab810de5fa9b116/Artificial-Intelligence-Guidance-May-2026.pdf accessed 21 May 2026.

16 Civil Justice Council, 'Use of AI in Preparing Court Documents' (*Courts and Tribunals Judiciary*, 17 February 2026) https://www.judiciary.uk/related-offices-and-bodies/advisory-bodies/cjc/current-work/use-of-ai-in-preparing-court-documents/ accessed 24 May 2026.

17 Online Procedure Rule Committee, 'Digital Justice System: Inclusion Framework and Pre-Action Model' (*GOV.UK*, 11 July 2025) https://www.gov.uk/government/consultations/digital-justice-system-inclusion-framework-and-pre-action-model accessed 24 May 2026.

18 Law Commission, 'AI and the Law: A Discussion Paper' (30 July 2025) https://cdn.websitebuilder.service.justice.gov.uk/uploads/sites/54/2025/07/AI-paper-PDF.pdf accessed 1 May 2026.

19 UK Jurisdiction Taskforce, 'Consultation on the Legal Statement on Liability for AI Harms under the Private Law of England and Wales' (2026) https://lawtechuk.io/ukjt/public-consultation-liability-for-ai-harms-under-the-private-law-of-england-and-wales/ accessed 8 May 2026.

20 Patricia Clarke, 'Meet the Vibe Lawyers: The AI Chatbots Wreaking Havoc in the Justice System' (*The Observer*, 24 February 2026) https://observer.co.uk/news/science-technology/article/meet-the-vibe-lawyers-the-custom-chatgpt-wreaking-havoc-in-the-justice-system accessed 25 March 2026; Giles Peaker, 'AI Issues in the First Tier Tribunal (Property Chamber) and Upper Tribunal (LC)' (*Nearly Legal*, 18 January 2026) https://nearlylegal.co.uk/2026/01/ai-issues-in-the-first-tier-tribunal-property-chamber-and-upper-tribunal-lc/ accessed 5 February 2026; Matthew Lee, 'UK AI Hallucination Cases Tracker (Verified List & Case Law)' (30 January 2026) https://naturalandartificiallaw.com/uk-ai-hallucination-cases-tracker/ accessed 25 March 2026.

21 Claude, 'Claude for the Legal Industry' (*Claude*, 12 May 2026) https://claude.com/blog/claude-for-the-legal-industry accessed 23 May 2026.

22 UK Government, 'Lex API' (*Incubator for AI*) https://lex.lab.i.ai.gov.uk/ accessed 24 May 2026.

23 Lee (n 20); for important guidance on the use of AI in UK courts and tribunals, see *R (on the application of Frederick Ayinde) v Haringey London Borough Council* [2025] EWHC (Admin) 1040.

Interest in AI-enabled tools in the UK justice system spans multiple institutional settings, from central government departments, courts and tribunal administration, and publicly funded and third-sector advice services to private legal service providers.[24] At the same time, many tools are procured or built through partnerships with private vendors, which creates variability in documentation and in the availability of public evaluation. The landscape is characterised not only by growing adoption but by considerable diversity in the types of tools deployed, the institutional contexts in which they operate, and the degree to which their use has been subject to evaluation.

The gap between adoption and evaluation is a defining feature of the current landscape. Recent work has provided important contextualisation of the risks and opportunities of implementing AI technologies in the UK justice system.[25] However, the balance of available policy and research remains heavily weighted towards the articulation of potential benefits and risks rather than their empirical evaluation. Where evidence does exist, it tends to concentrate on the outcomes that are most readily measurable, such as operator-reported efficiency and satisfaction, rather than the outcomes that matter more from the standpoint of the people the system serves. It has been argued that: "Empirical measures should be adopted to assess the impact of these tools on the right of access to justice in the real world, alongside other factors."[26]

1.1 PRTJ rationale for this review

This review examines the existing evidence on the use of AI and data-driven technologies in justice settings. It was commissioned by the Nuffield Foundation as part of its *Public right to justice* (**PRTJ**) programme. The PRTJ programme proceeds from the assumption that the public have a right to a system of justice that accessibly, fairly, and effectively meets their legal needs, having regard for their characteristics or circumstances, and that the system should be open, accountable, trustworthy, and supportive of wider societal values.[27]

The PRTJ programme seeks to explore the extent to which the public's right to a fair, accessible, and effective justice system is being realised, and to develop evidence-informed proposals for improvement. The role and impact of new digital technologies, specifically AI and data-driven systems, is a critical area of inquiry, given its rapid development and profound implications for the justice system and the people it serves. The need for strong evaluations of AI impact is being recognised more broadly,[28] and warrants specific attention in the legal system.

This review is designed to directly support the PRTJ programme's objectives. It is a systematic evidence scoping review. Its purpose is not only to synthesise what the evidence says about the impacts of AI in justice settings, but to characterise the evidence base itself, to identify what has been studied, what has not, and where the balance of evidence is sufficient to support provisional conclusions and where it is not.

24 Ministry of Justice, 'AI Action Plan for Justice' (n 2); UK Government, 'AI Opportunities Action Plan' (13 January 2025) https://www.gov.uk/government/publications/ai-opportunities-action-plan-government-response accessed 6 February 2026.
25 Justice UK, 'AI in Our Justice System' (Justice UK 2025) https://perma.cc/BW8N-RZ9N ; Hawes and others (n 3).
26 Natalie Byrom, 'Computational Law and Access to Justice' (2024) 2 Journal of Cross-disciplinary Research in Computational Law.
27 Nuffield Foundation, 'Public Right to Justice' https://www.nuffieldfoundation.org/evidence-and-impact/our-programmes/public-right-to-justice accessed 25 March 2026.
28 Oliver Hauser and others, 'Why Evaluating the Impact of AI Needs to Start Now' (2025) 643 Nature 910.

A theme that recurs throughout this review, which is helpful to identify at the outset, is that the evidence base on AI in justice settings is characterised as much by its absences as by its findings. In many areas, deployment materially outpaces evaluation. The general absence of high-quality independent evaluations and evidence of the impact of AI tools in UK justice settings, and the narrow focus of the evaluations that do exist, is a significant finding of this review with direct implications.

1.2 Research questions

The review is organised around three core research questions, aligned to the PRTJ programme's aims.

- **RQ1 – Individual and Group Experiences and Outcomes.** What evidence exists and what does it reveal about how AI technologies, positively or negatively, affect individuals and groups who are addressing or resolving legal problems, including differential impacts?
- **RQ2 – Societal Factors.** What evidence exists and what does it reveal about the impacts on wider societal issues such as public trust in justice, public understanding or scrutiny of justice, and the acceptability of AI use in justice settings?
- **RQ3 – System Efficiency and Effectiveness.** What evidence exists and what does it reveal about the justice system's capacity to use AI technologies to address legal needs promptly and fairly?

These three questions are distinct but not independent. A single study may speak to more than one question simultaneously: a tool designed to improve administrative efficiency, for example, may also affect the experience and outcomes of individuals interacting with the system, and may carry implications for public perceptions of legitimacy and fairness. Where the evidence provides insights across questions, the review makes these intersections explicit rather than confining findings artificially to a single category.

1.3 Scope and boundaries

1.3.1 Legal domains

The primary focus of this review is on administrative, civil, and family law. The concern is particularly with the everyday legal problems of individuals and families and how they resolve them: money and debt problems, housing disputes, benefits access and entitlements, education provision (including matters of Special Educational Needs and Disability (**SEND**)), employment problems, accident and health-related legal issues, neighbour disputes, and the full range of public and private family law matters. These are the areas where, for most people, the justice system is encountered not as an abstraction but as a practical mechanism for resolving problems that affect their daily lives, their financial security, and their well-being.

Consumer problems, certain residential property issues, and areas of commercial or corporate law are less central to the review's interests. Commercial and corporate legal practice has historically led the adoption of AI tools, and the maturity of deployment in these domains exceeds that in the everyday-justice domains on which this review focuses. Therefore, available evidence from these domains is included where it generates relevant insights.

Criminal justice is excluded as a primary area of focus. The criminal justice system presents a distinct set of risks, institutional structures, and accountability mechanisms, and has been the subject of a more developed body of research on AI applications, particularly in policing and sentencing. Where evidence from the criminal justice sector offers direct overlap with the review's primary scope or where findings are transferable to civil, administrative, or family justice settings, it is drawn upon selectively.

The review focuses on formal legal action, including adjudication by courts and tribunals and the processes leading up to them, as well as any modes of formal action to resolve legal problems, including seeking legal advice. Some emerging uses of AI in downstream enforcement and regulatory work are mentioned but only as relevant to the scope. The review also treats as within its analytical concern the informal use of general-purpose AI tools by practitioners, judges, and litigants in person. Such use currently sits beyond the purview of formal justice processes but already shapes how people encounter them: a litigant in person who drafts legal documents using ChatGPT, or a judge who uses Copilot to summarise party submissions, is participating in a justice process whose conditions are partly being set by tools outside that process. The tool-mapping captures formal deployments and in-development tools; informal use is treated separately throughout the synthesis chapters.

The evidence base also includes studies from adjacent public service contexts – including social care, benefits administration, policing, and general civil service productivity – that generate insights relevant to justice settings. The review draws on this evidence where appropriate but does so with explicit attention to the question of whether findings transfer. Where evidence is absent on a specific issue, a comparative review of other regulated public sectors has been considered. It also provides a lens through which to examine the extent of 'legal exceptionalism' in relation to AI deployment and its effects.

1.3.2 Technology

The review focuses on applications of AI and related data-driven technologies that go beyond basic digitisation and rule-based automation of existing processes, and instead play a more assistive, adaptive, or autonomous role in justice processes. The review uses the term 'artificial intelligence' as a practical category rather than a definitional claim; the intention is to be inclusive in a manner appropriate for a scoping review. Several types of technology are in scope: machine learning; predictive analytics; natural language processing (**NLP**); LLMs; automatic speech recognition (**ASR**); and other data-driven systems that infer, classify, generate, or recommend outputs from patterns in data. The review examines settings where these models, or a combination of them, are used to build either justice-specific products or general-purpose tools applied to legal tasks. Tools such as e-filing portals, online forms, and digital case-management systems are treated as out of scope unless they also include AI-enabled functions, such as semantic search, AI transcription or summarisation, or data-driven decision-support.[29] Without providing a full set of definitions, it is worth noting a few important terms recurring in this review.[30]

29 For example, the recent evaluation of HMCTS Digital Services reform is considered out of scope as it does not include use of AI systems, Frontier Economics, IFF Research and Ministry of Justice, 'HMCTS Reform Digital Services Evaluation Overarching Report' (*GOV.UK*, 11 September 2025) https://www.gov.uk/government/publications/hm-courts-tribunals-service-reform-digital-services-evaluation accessed 21 January 2026.

30 For a more detailed introduction, see Elliot Jones, 'What Is a Foundation Model?' (Ada Lovelace Institute 2024) https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ accessed 9 April 2026; David Lehr and Paul Ohm, 'Playing with the

Rule-based AI systems operate on explicit 'if-then' logic specified in advance. They do not learn from data. Their behaviour is therefore interpretable, in that the path from input to output can in principle be clearly reviewed,[31] but they are less scalable and cannot perform the more sophisticated inference that data-driven methods allow. In justice settings, these may convert legal logic or categorisations into 'if-then' rules and may appear in tools in isolation or in combination with other technologies.

Machine learning is a broad category of AI in which the system learns patterns from data rather than following pre-defined rules. A model is trained on examples in the data and learns patterns to classify, score, predict, or otherwise produce predicted outcomes for new data. Statistical models, deep learning, and other predictive machine learning methods are central to legal risk assessment, resource prioritisation, mapping patterns in routing and triage, and other types of tasks that require data-driven outcome prediction.[32]

Legal data is predominantly textual; most AI systems considered in this review are therefore largely based on language-model architectures that fall within the field of NLP, focused on computational approaches to interpret, analyse, and produce human language. Transformer models – a neural-network architecture designed to track relationships across words, text, and language – underpin most current language models. There are two primary variants: encoder-based models, oriented towards interpreting and labelling existing text (for example, the Bidirectional Encoder Representations from Transformers (**BERT**) model family), and decoder-based models, oriented towards generating new text (for example, LLMs).

LLMs are a prominent and specific example of general-purpose AI, also called foundation models and often referring to generative AI. LLMs are trained on vast and heterogeneous text and are capable of a broad range of tasks where the outputs resemble human writing – including drafting, summarisation, question-answering, and open-ended dialogue. These models perform well across a wide range of tasks without being built for a specific domain or task.

Also relevant in justice domains are AI tools that produce written text from spoken speech. ASR tools have been in use for many years, but recent versions are built on foundation-model architectures and offer improved accuracy and summarisation. ASR shares two properties with LLMs: it can produce fabricated content; and it may handle some accents, dialects, and domain-specific vocabulary less reliably than mainstream speech. ASR tools are not neutral or transparent recording devices.

Given the prevalence of LLM-based tools in this review, it is worth briefly elaborating on a few core features:

First, LLMs can vary considerably in purpose and design. General-purpose LLMs are designed to perform a wide range of tasks without being trained on any single domain. These include models in the GPT (OpenAI, underpinning ChatGPT and Microsoft Copilot), Claude (Anthropic), Gemini (Google), Llama (Meta), and Qwen (Alibaba) families, among others. A growing set of legal-domain tools is built on top of these models by proprietary vendors, integrating foundation-model capabilities with curated legal databases or practice-specific workflows to support research, drafting, and case analysis in professional settings. Examples include Lexis+AI (LexisNexis), Westlaw AI-Assisted

Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 UC Davis Law Review 654.

31 Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1 Nature Machine Intelligence 206.

32 Ministry of Justice, 'AI Action Plan for Justice' (n 2).

Research and CoCounsel (Thomson Reuters), Harvey, and Legora. Major providers also offer mechanisms that let a user configure and publish their own version of a general-purpose chatbot without any coding or model training (informally called 'wrappers'). Using OpenAI's GPTs feature, for example, a person can take ChatGPT, give it a fixed set of instructions, upload a body of reference documents, and publish it as a named assistant that runs inside GPT (comparable services exist on other platforms). This requires very little technical expertise and no institutional or professional setting; a law firm, individual, or start-up can each produce a branded legal assistant in the same way and on the same underlying model. A number of consumer-facing UK legal tools aimed at members of the public appear to be built along these lines,[33] offering legal information with disclaimers that it is not legal advice.

Tools built on the same underlying model are not necessarily alike in their suitability for legal work. A general-purpose LLM is designed to serve the largest possible range of users and tasks, which may mean it handles domain-specific legal language and tasks less well than a non-generative system adapted to legal materials. The methodological point follows directly: claims about 'what AI can do' in justice settings depend on which tool, applied to which task, under which conditions, and within which institutional infrastructure.

Second, several types of systems can be adapted and customised for legal settings. AI models are pre-trained on large, general datasets. A pre-trained model can then be fine-tuned for a different but related task using a smaller, specialised dataset, so that it performs better on a specific task or type of material. For example, a pre-trained model can be fine-tuned on legal data to enhance its understanding of domain-specific language, legal terminology, and the structure of legal documents, in order to deliver more tailored and consistent outputs. A prominent NLP model used in academic research is the encoder-based model BERT, which has been fine-tuned on legal domain-specific data to create Legal-BERT for defined tasks such as legal citation detection, case classification, and information extraction.[34] Retrieval-augmented generation (**RAG**) is a different form of customisation that connects the model to an external, curated dataset of source material at the point of use, so that responses are grounded in specified documents rather than in the model's general training alone. RAG is the most prominent of these methods in the evidence reviewed here, and several deployed tools rely on it to constrain outputs to authoritative legal sources.

A recurring issue across generative systems is hallucination – also described as confabulation or fabrication – the production of output text that is fluent and plausible but factually incorrect, unsupported, or invented. Hallucination is a structural feature of how LLM architectures work and are not a removable defect; mitigation methods such as fine-tuning and RAG can reduce its incidence but cannot drive error rates to zero.[35] It follows that the appropriate evaluative question for LLM-based legal tools is not whether errors can be eliminated, but what level and distribution of error is acceptable in a given legal context, and how the level of error interacts with the risks of user over-reliance.

33 See Appendix A, OLS Solicitors, 'Nessa - AI Family Law Chatbot' https://www.ols-solicitors.co.uk/family-law-ai-assistant/ accessed 7 April 2026; AskEllie, 'About Ellie' https://www.askellie.co.uk/about-ellie/ accessed 30 January 2026.

34 Ilias Chalkidis and others, 'LEGAL-BERT: The Muppets Straight out of Law School' *Findings of the Conference on Empirical Methods in Natural Language Processing* (ACL 2020) https://doi.org/10.18653/v1/2020.findings-emnlp.261.

35 Emily Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM 2021) https://doi.org/10.1145/3442188.3445922 ; Joseph Dumit and Andreas Roepstorff, 'AI Hallucinations Are a Feature of LLM Design, Not a Bug' (2025) 639 Nature 38.

Third, an important distinction is how openness varies across systems and its impacts on how they can be independently inspected.[36] Proprietary or enterprise models – such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini – are closed: their model weights and full training data are not publicly released, and only limited information about their architecture and training process is typically disclosed. Open-weight models, by contrast, make their weights publicly available, which permits a greater degree of inspection, adaptation, and fine-tuning, even though the training data and full development pipeline may remain undisclosed. Examples include Google's Gemma, OpenAI's gpt-oss, Meta's Llama, Alibaba's Qwen, and DeepSeek's and Mistral's open-weight models. More fully open-source models, such as OLMo (Allen AI) and Pythia (EleutherAI), go further by releasing weights together with training code and either the training data itself or tools to reconstruct exact training dataloaders, enabling a substantially higher degree of scientific scrutiny. Structural inspection of training data, model weights, and architecture is therefore possible for fully open systems, partly possible for open-weight systems, and much more limited for proprietary systems. Behavioural evaluation, through benchmarking, red-teaming (adversarial testing), and human-interaction studies, remains possible across all three and is the main route by which closed systems can be assessed. Operational deployments in UK justice settings rely overwhelmingly on proprietary systems; many academic studies in this review evaluate a broader range of models across both BERT-style encoder models and different types of LLMs. This asymmetry between deployment and study has implications for transparency, accountability, and the reproducibility of evaluations – themes that recur throughout the report.

The greater the level of impact associated with the technology's use, the greater the salience for this review. Where AI is deployed in ways that affect people's obligations, rights, health, well-being, or financial position, and particularly where the effects are difficult to reverse or are ongoing, the stakes are higher and the need for evidence more pressing. Where the evidence allows, the review also attends to the cumulative effect of multiple technologies deployed at different points in the justice process, potentially by different actors and decision-makers. This cumulative dimension is poorly captured by studies that evaluate single tools in isolation, but it matters for understanding how AI reshapes the experience of justice as a whole.

36 The landscape of general-purpose AI, particularly LLMs, is highly dynamic, with constant iteration and the release of new models. The examples mentioned here are representative of current leading model families, see 'List of Large Language Models', *Wikipedia* (2026) https://en.wikipedia.org/wiki/List_of_large_language_models accessed 29 May 2026.

1.3.3 Jurisdiction

The review prioritises the jurisdictions of England and Wales, followed by the UK as a whole where the evidence base and institutional context warrant. International evidence is drawn upon where there is high-quality, directly relevant evidence from other jurisdictions, and where transferability to the UK context can be argued. The review does not assume transferability; it draws on evidence from similar or related jurisdictions where that evidence illuminates shared challenges or offers instructive contrasts.

1.4 Structure

The report is organised as follows.

Section 2 sets out a brief methodology covering the review design, approach, and limitations, with full details provided in Appendices B to D.

Section 3 maps the current deployment landscape. It describes the 45 AI tools and systems identified as operating in, or directly adjacent to, UK justice settings, and examines their distribution across legal domains, institutional contexts, technology types, and user orientations. It documents the gap between deployment and publicly available evaluation, a structural feature of the field that shapes the analysis throughout the report.

Section 4 provides a descriptive account of the evidence base itself, mapping the included studies by temporal distribution, publication type, domain, methodology, and alignment with the three research questions. The section treats the character of the evidence base as an analytical finding in its own right, distinct from the substantive conclusions that the evidence supports.

Sections 5, 6, and 7 present the thematic synthesis of evidence, organised by research question. Section 5 addresses individual and group experiences and outcomes (RQ1), including access to legal services and information, procedural experience, and differential impact. Section 6 addresses societal factors (RQ2), examining public trust and acceptability, professional and judicial attitudes, and open justice. Section 7 addresses system efficiency and effectiveness (RQ3), covering time savings, accuracy and reliability, and the implementation burden associated with AI tools in justice workflows. Each synthesis section opens with an evidential orientation that characterises the type and quality of evidence available for that question and closes with a summary of findings.

Section 8 draws together the cross-cutting insights that emerge from the evidence as a whole and organises them around four structural gaps: measurement gap, deployment gap, legitimacy gap, and transparency gap. Section 9 sets out priorities arising from the analysis, including research priorities, evaluation paradigm questions, and policy and procurement considerations.

2 Methodology

2.1 Review design and approach

The review was designed to give a clear picture of what empirical and evaluative evidence currently exists on how AI and data-driven technologies are impacting the UK justice system. It adopts a systematic scoping method to map the range and character of available evidence, identify gaps, and synthesise findings across diverse literature, while allowing a flexible and practical approach.[37]

The review focuses on empirical and evaluative evidence, prioritising real-world evidence from evaluations, pilots, technical assessments, and applied studies. The review describes these records as 'real-world evaluations' which is simply an umbrella term for applied studies that assess an AI or data-driven tool as it operates in an actual or piloted deployment setting. This category may include process evaluations (how an intervention was implemented and what lessons can be learned), impact evaluations (examining the difference an intervention made, to the extent observed changes are attributable to it), and the experimental and quasi-experimental studies used to support these evaluations (such as randomised controlled trials).[38] A narrower category of evidence identified is 'technical benchmarking' research. These studies use a pre-defined scoring metric to compare AI models against each other, or against prior results, on a narrowly defined task using a fixed dataset. While benchmarks provide insight into model performance for specific legal tasks or processes, these scores do not measure real-world outcomes or impacts.[39]

Conceptual or theoretical literature was used only to help frame issues and identify gaps – it does not form part of the main evidence base.

2.2 Evidence gathering, screening, and mapping

In order to optimise inclusion, the search strategy was designed to gain broad coverage across a rapidly evolving evidence base spanning academic literature, policy, and grey literature. In practice, the review conducted more manual searching than is traditional for an academic scoping review, which was necessary to capture a wider range of publication types and AI applications. Contacts across public, academic, and civil society organisations suggested further relevant sources.

37 Andrea Tricco and others, 'PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation' (2018) 169 Annals of Internal Medicine 467.

38 Evaluation Taskforce, 'Magenta Book: Central Government Guidance on Evaluation (HTML)' (*UK Government*, 15 May 2026) https://www.gov.uk/government/publications/the-magenta-book/magenta-book-central-government-guidance-on-evaluation-html accessed 27 May 2026.

39 Inioluwa Deborah Raji and others, 'AI and the Everything in the Whole Wide World Benchmark' (2021) 34 Advances in Neural Information Processing Systems (Datasets and Benchmarks); see discussion in Elliot Jones, Mahi Hardalupas and William Agnew, 'Under the Radar?' (Ada Lovelace Institute 2024) https://www.adalovelaceinstitute.org/report/under-the-radar/ accessed 27 May 2026.

Five stages shaped the process: preliminary landscape mapping, targeted source searching, iterative refinement of search terms, expansion to recent applications and pilots, and final compilation. Full details of search strings, dates, and yields are reported in **Appendix B**.

All records were screened for both topic relevance and an evidence threshold to exclude studies that did not present relevant evidence, following the workflow outlined in Appendices C and D. Included materials were then mapped against the three research questions: individual and group experiences and outcomes, societal factors, and system efficiency and effectiveness. A structured data extraction framework supported coding and synthesis.

2.3 Limitations

The review captures a field that is developing rapidly and not consistently documented. Evidence is dispersed across academic, practitioner, and grey literature sources, and many operational assessments or vendor-led evaluations remain unpublished. The review's capacity to capture such literature is inherently partial. For example, the Algorithmic Transparency Records provide a partial window onto government deployment, but coverage may be incomplete and disclosures vary in depth. The Public Law Project, for instance, has reported that there are far fewer tools listed in the Algorithmic Transparency Records than the number of public sector algorithms it has identified in practice.[40]

Legal and technical evidence is spread across multiple formats and sources, requiring pragmatic inclusion choices. Some potentially relevant materials may not have been located or may have been excluded due to insufficient information. Overall, the dataset provides a representative but not exhaustive picture of UK practice in AI and justice.

40 Mia Leslie, 'Around the World in AI Regulation - How the UK Can Become a Leader in Transparency' (*Public Law Project*, 11 October 2024) https://publiclawproject.org.uk/resources/around-the-world-in-ai-regulation-how-the-uk-can-become-a-leader-in-transparency/ accessed 21 May 2026.

3 Deployment landscape

This section maps the current landscape of AI tools operating in or relevant to the UK justice system. The mapping draws on publicly available information including government publications, academic sources, and vendor documentation. Two boundaries should be noted at the outset. First, the mapping captures formal deployments and in-development tools; it does not record informal use of general-purpose AI tools by practitioners, judges, or litigants in person, which is discussed separately. Second, where no formal evaluation was identified, the mapping records only the tool's existence, stated functionality, and deployment status. The absence of a recorded evaluation does not necessarily mean no internal assessment has taken place; it means no evaluation was publicly available at the time of the review.

3.1 Deployment scale and legal domain

The mapping identifies 45 AI tools and systems operating in, or directly adjacent to, justice settings in the UK. Of the 45 tools, 27 are live, 12 are at pilot stage, and 6 remain exploratory, which indicates that the landscape is already operational rather than merely prospective. That said, 'live' does not imply uniform scale. The Home Office's asylum casework tools operate across a national caseload. HMCTS tools such as ListAssist and the Secure AI Assistants serve courts across multiple jurisdictions. Caddy, the RAG-based copilot for Citizens Advice, has been deployed across six offices with national rollout planned. Thus, while some tools operate across national caseloads or across multiple jurisdictions, others remain confined to a single service, local jurisdiction, or pilot site. Similarly, tools in the exploratory stage are those that have been publicly announced, but there are likely to be many legal AI tools currently being explored across the justice system that are not yet publicly documented. The full tool-mapping is provided in Appendix A, which details the primary developer, deployment status, jurisdiction, whether public or internal facing, legal domain, AI type, primary function, and evaluation status.

The mapped landscape is also geographically broad. The review identified 26 tools operating UK-wide, 10 in England and Wales, 4 in Scotland, 2 in England, 2 in Northern Ireland, and 1 in Wales, although these jurisdiction labels capture primary deployment rather than full territorial reach in every case.

As can be seen in Figure 1, the largest domain cluster is courts and tribunals, with 12 tools, covering transcription, listing and scheduling, form processing, anonymisation, e-discovery, case-management support, and knowledge retrieval, all of which are internal-facing. Civil law across legal aid settings, negligence, and small claims is the next largest cluster, with 8 tools, followed by family law with 6. Criminal, employment and immigration each contain 3 tools, and smaller clusters appear in welfare, education (related to SEND), child rights, data protection, housing, financial service ombudsman, and related procedural settings.

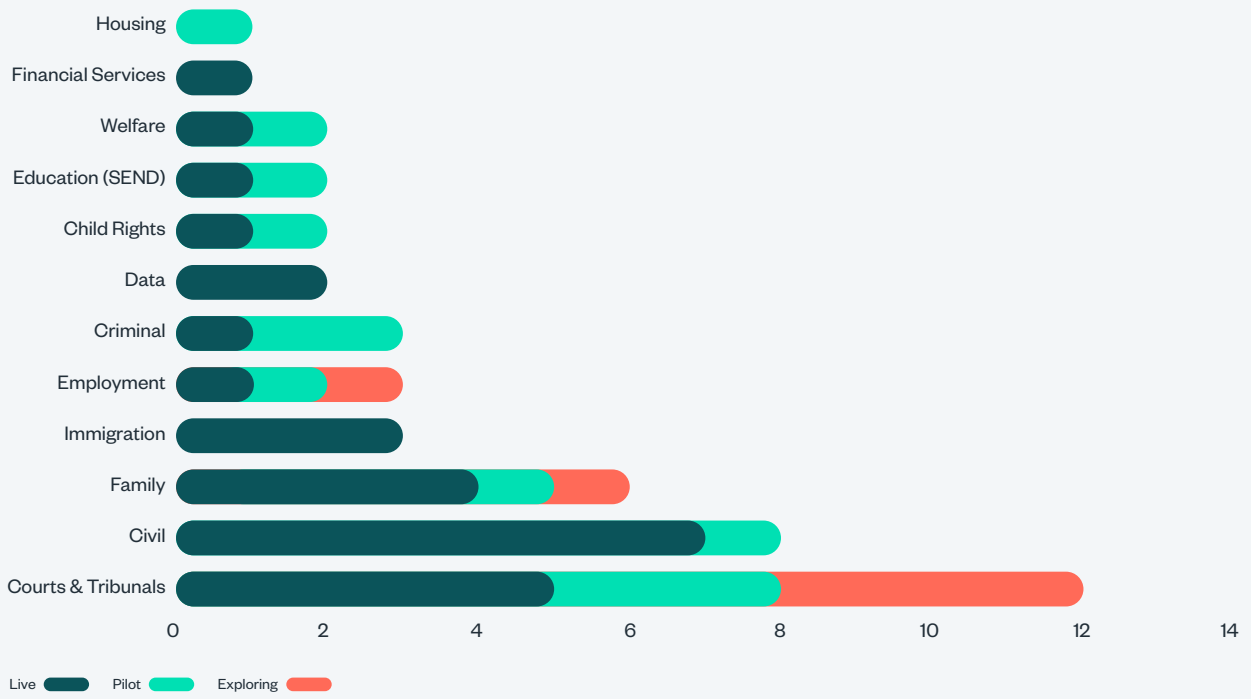


Figure 1: Distribution of mapped tools by deployment status and across legal domain.

This distribution is important for understanding the present direction of AI use. The most institutionally mature cluster sits in courts and tribunals, but it is concentrated in administrative and professional support functions rather than direct public interaction. By contrast, public-facing tools are more visible in civil advice, family, education, housing, immigration, and child-rights settings.

3.2 Institutional character and user orientation

Given this domain distribution, the landscape is institutionally led by government. Government bodies account for 22 of the 45 mapped tools, including HMCTS, MoJ, Home Office, Children and Family Court Advisory and Support Service (**Cafcass**), SCTS, Information Commissioner's Office (**ICO**), Advisory, Conciliation and Arbitration Service (**Acas**), and Financial Ombudsman Service (**FOS**). Private developers account for 11 tools, while access-to-justice organisations and mixed-sector collaborations account for 5 each, with a smaller academic contribution also present in the mapping.

The landscape is predominantly internal facing, as shown in Figure 2. Of the tools, 29 are designed or judges, civil servants, advisers, lawyers, court staff, or other professionals, while 16 are public facing, designed for direct interaction with the public. Internal tools are designed for tasks such as transcription, case summarisation, legal research, search, drafting, listing, routing, and case-management support. Public-facing tools, by contrast, tend to provide legal information, triage, guided intake, signposting, or document support, rather than making or automating final legal decisions.

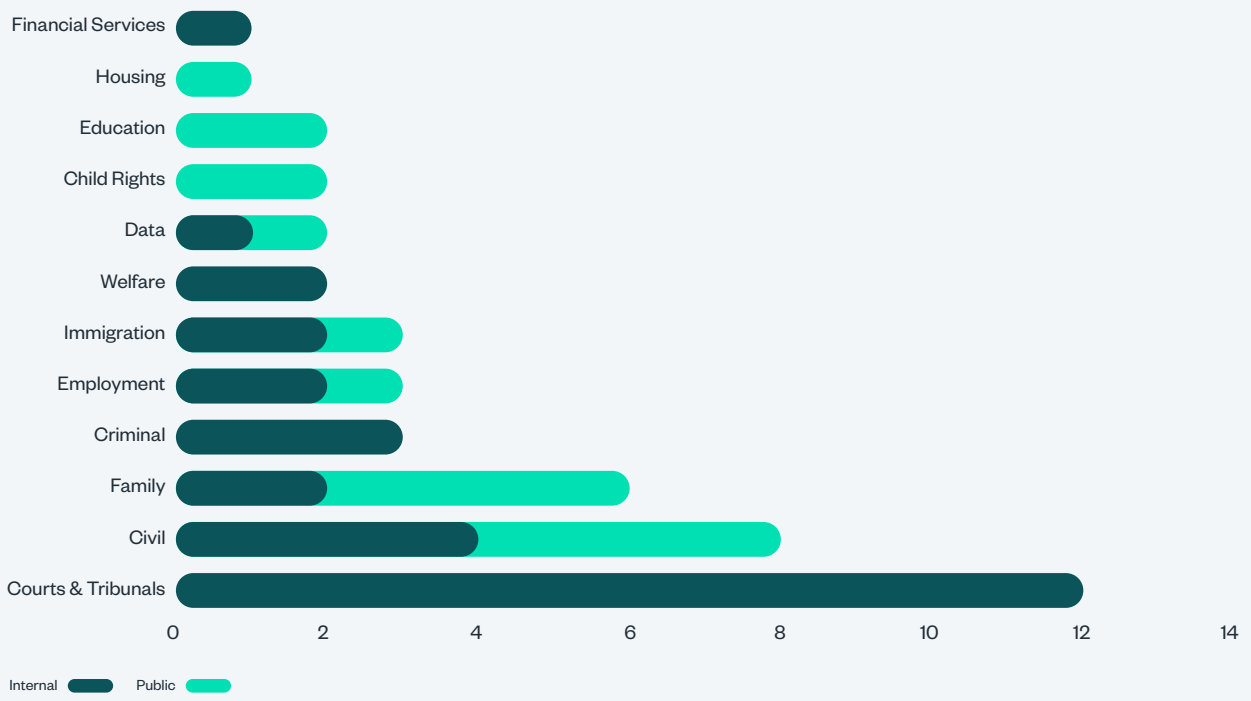


Figure 2: Distribution of mapped tools by legal domain, across internal versus public orientation.

3.3 Technology profile

The technology profile is led by LLMs. Of the 45 mapped tools, 28 are built on LLMs in some form, 8 incorporate RAG, 4 use ASR, 4 use machine learning, 2 rely on NLP approaches, and 1 is a rules-based algorithm. Nine tools do not report enough information for their underlying AI technology to be identified clearly. The mapped field is best understood as LLM-centred, although not exclusively so.

Thirteen tools disclose the type of AI models they are built from, 11 of which are built on proprietary general-purpose LLMs, such as Open AI's GPT family or Microsoft's Copilot; only 2 are built on open-source LLM architectures. A further 26 tools do not disclose the specific model used, and 6 are explicitly built with non-disclosed proprietary AI. As discussed in Section 1.3.2, proprietary models cannot be subjected to the structural inspection of weights, training data, and architecture that open-source or open-weight systems permit, although the depth of openness varies considerably. Evaluation through benchmarking, red-teaming, and human-interaction studies remains possible for both proprietary and open-weight models and is the principal route by which independent assessment of closed systems is conducted.

The mapping also records what tools disclose about specific design features, though the picture here is less complete: 17 tools are recorded as involving human review of outputs; 10 are recorded as not involving human review; and 18 do not report the position clearly. These figures capture vendor or developer disclosures about intended system design, not observed practice. Even where the intended human-in-the-loop arrangement is clearly stated, this does not establish whether or how

such review is exercised at the point of use; that question requires empirical evaluation and is largely absent from the evidence base reviewed here. The figures are therefore not a confident measure of the prevalence of human oversight in deployed tools, but they are an indication of how developers describe their systems.

3.4 Deployment–evaluation gap

The clearest overall feature of the mapping is the gap between deployment and publicly available evaluation. Only 7 of the 45 mapped tools have any publicly available evaluation, and 37 have no publicly available evaluation identified in this review. One further tool has an evaluation in progress. The evaluation deficit is not confined to early-stage pilots; among the 27 live tools, 22 have no publicly available evaluation.

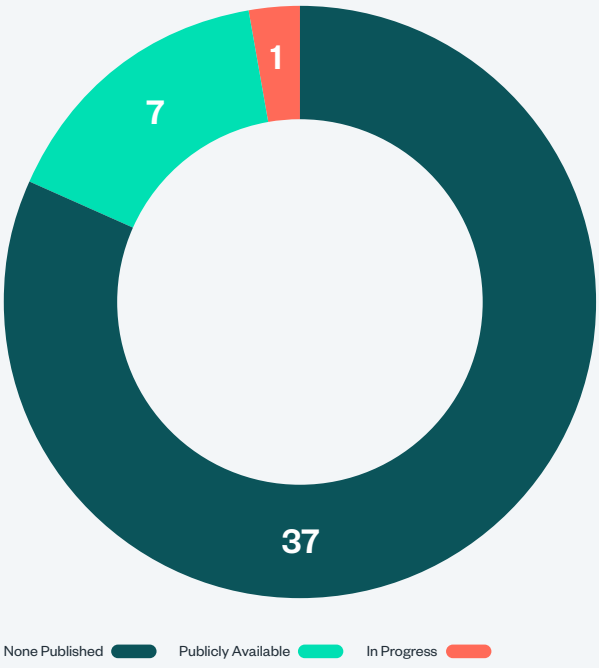


Figure 3: Evaluation status of mapped tools.

The independence of the available evaluations is also limited. Of the 7 evaluations, 3 were government-led (assessing government-developed tools), 3 were academic, and 1 was vendor-commissioned. Many tools have reported performance claims in public documentation, government case studies, or vendor materials. These range from processing time reductions and user satisfaction scores to broader assertions about improved access or quality. They are self-reported and unverified claims, presented without published methodologies. The content and quality of the published evaluations are discussed in Sections 5 to 7.

There is one notable in-progress evaluation. Change Please's 'Housing Helper' is an AI chatbot designed to help users assess their housing situation, offer targeted housing advice, draft custom letters and emails related to housing relief, and connect users to relevant resources and services. It is built on an LLM and RAG system that constrains advice to publicly accessible, authoritative, and ethically approved sources. A randomised controlled trial is being conducted by a collaborative team comprising Change Please, the Centre for Homelessness Impact, Southwark Council, and King's College London, with results expected in spring 2027.[41]

The deployment of AI in UK justice settings is, in this sense, proceeding well ahead of the publicly available evidence. Tools are live, in daily use, processing cases and shaping advice, without a publicly available account of whether they work as intended, whether they produce errors that matter, or whether they serve the interests of the individuals seeking resolution of their legal problems.

41 Centre for Homelessness Impact, 'AI Chatbot to Offer Housing Advice' (2 February 2026) https://www.homelessnessimpact.org/projects/ai-chatbot-to-offer-housing-advice accessed 3 February 2026; King's College London, 'Trial Protocol: Generative AI Chatbot "Housing Helper": A Randomised Controlled Trial' (15 December 2025) https://perma.cc/LFD8-FMZT

4 Evidence landscape

4.1 UK evidence landscape

This section provides a purely descriptive account of the studies included in the UK strand of the evidence review. It maps the corpus across temporal distribution, jurisdictional coverage, publication type, legal domain, technology type, research design, evidence quality, and alignment with the three core research questions. Substantive thematic synthesis follows in Sections 5 to 7.

The systematic search and screening process identified 48 studies meeting the inclusion criteria for the UK-specific evidence base. This is a modest corpus given the breadth of AI applications now being deployed or piloted in UK justice settings, and it should be read in the context of a broader finding discussed in subsequent sections: the deployment of AI tools in justice settings is materially outpacing the production of evaluative evidence.

The 48 studies vary considerably in their methodological rigour, thematic focus, and the strength of evidence produced.

4.1.1 Jurisdiction and temporal distribution

Of the 48 studies identified within the UK, 30 adopt a UK-wide frame, 11 cover England and Wales together, and 5 use a UK frame with an international comparison. One study each was limited to Wales and to England respectively; no studies specifically focused on Northern Ireland or Scotland.

Figure 4 maps the jurisdiction to publication year, showing the evidence base exhibits a pronounced recency skew: 22 (45%) were published in 2025; 7 (14.5%) in 2024; and 6 (12.5%) in 2026 (year-to-date at the time of the review). A further 5 studies (11%) were published in 2023. The remaining 8 studies (16.5%) span the period 2013–2022. During the screening and selection of the evidence, it was clear that earlier literature on AI in the UK justice system consisted of predominantly theoretical and normative contributions. While these are valuable contributions to the field, they fall outside the scope of this review, which focuses on empirical and evaluative work.

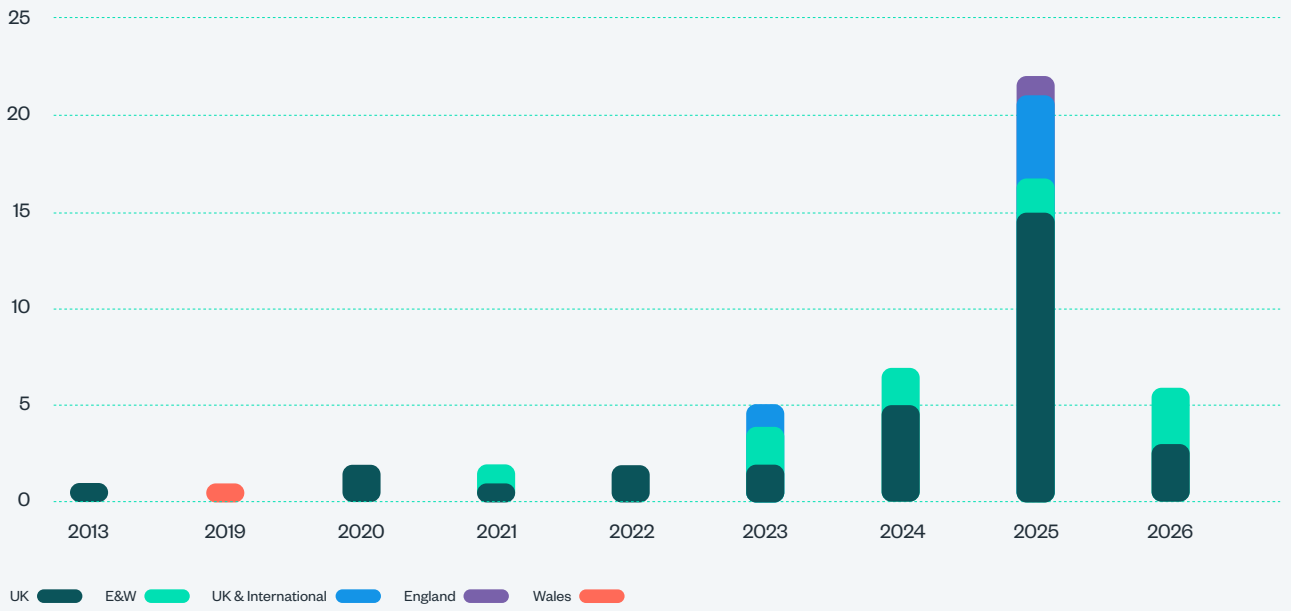


Figure 4: Distribution of studies by publication year and primary jurisdiction.

4.1.2 Publication type

The literature spans a range of publication types. Peer-reviewed and academic articles constitute the largest single group (n=24). Independent or commissioned reports account for 8 studies, while government reports account for 7. The remainder comprises preprints not yet subject to formal peer review (n=4), workshop papers (n=2), vendor reports (n=2), and a government blog post (n=1).

Methodological and contextual differences between government, vendor, and academic outputs are flagged at appropriate points in the synthesis. Studies published without peer review or methodological transparency are reported and weighted accordingly. Government reports and vendor materials, in particular, are read with explicit attention to the fact that internal and vendor-led evaluations cannot easily escape the actual or perceived risk of optimism bias.

4.1.3 Legal domain and setting coverage

General cross-domain work is the most represented category with 13 studies, followed closely by general civil law with 12 studies (see Figure 5). Of those civil law studies, 4 specifically focus on legal aid settings. Commercial law accounts for 5 studies, welfare settings account for 4, while 3 studies evaluate non-legal, public service settings. The remaining domains appear in more isolated records, including employment, human rights, immigration, and administrative law.

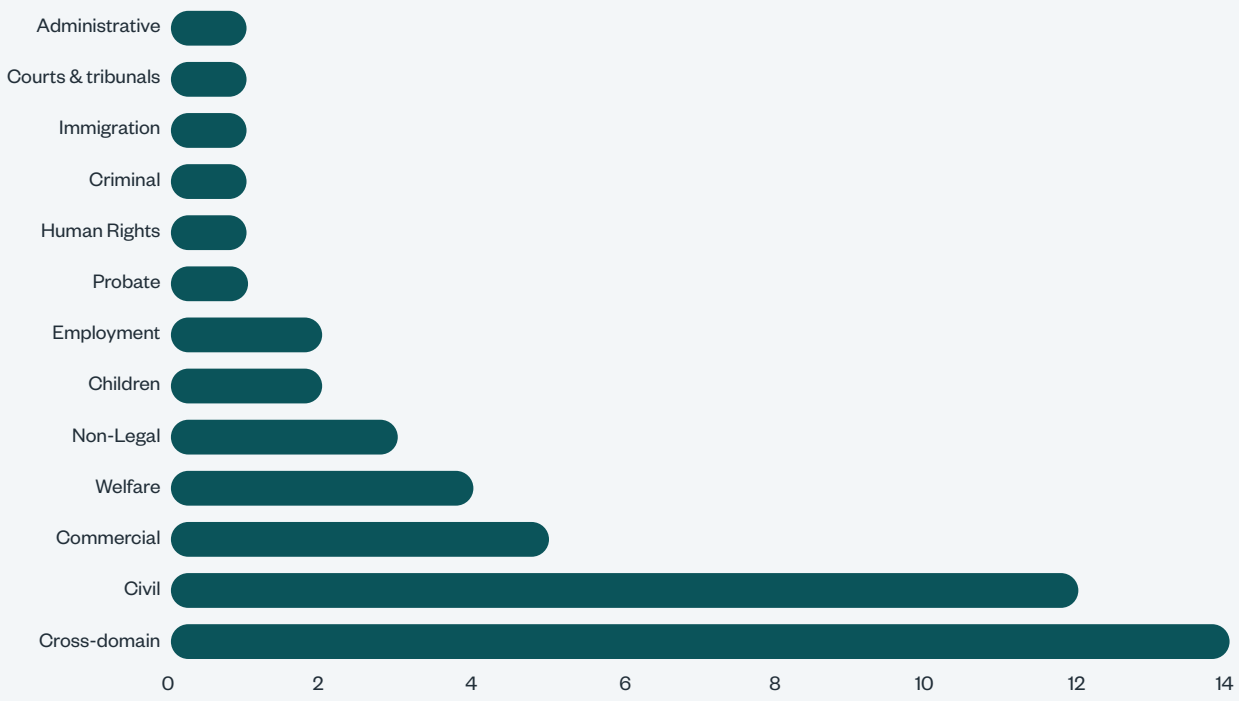


Figure 5: Distribution of studies by legal domain.

While criminal justice (n=1) and non-legal settings (n=3) fall strictly outside the primary scope of this review, a small number of these records have been included where the evidence demonstrates direct transferability to the review's core focus. Examples include studies that examine a procedural or administrative process (rather than substantive criminal law) closely mirroring in-scope justice settings, and records providing a broader, overarching government evaluation applicable across multiple public sectors.

4.1.4 Technology coverage

LLM-based tools dominate the evidence base. Table 1 summarises the distribution across technology types.

Primary technology	Studies
LLM (standalone)	**15**
LLM combined with ASR	**5**
LLM combined with RAG	**3**
LLM combined with BERT-style models	**3**
ASR combined with NLP	**3**
NLP	**4**
Predictive ML	**4**
Digitalisation	**2**
Technology-assisted review (TAR)	**1**
Not specified	**8**

Table 1: Distribution of studies by primary technology type

LLM-based work, taken together, accounts for 26 studies (54% of the corpus). This concentration reflects both the visibility of generative AI in current research and the actual prevalence of LLM-based deployments mapped in Section 3. Non-generative machine learning, rule-based, and NLP-only work is comparatively under-represented, even where such systems remain in active deployment.

4.1.5 Research design and evidence types

The studies employ a range of research design approaches, as shown in Figure 6. Given the nature of the review, real-world evaluations – which assess the design, implementation, and outcomes of an AI tool in deployed settings – are the most common evidence type, accounting for 16 studies. Technical benchmark papers, which assess AI performance on standardised tasks using fixed legal datasets, account for 15. Proof-of-concept (PoC) papers with accompanying evaluative data account for 8, survey-based perception studies for 7, while 2 studies use other research designs.

In terms of method type, quantitative approaches account for 17 studies, mixed methods for 14, and survey-based approaches (both mixed and qualitative) for 11. Experimental approaches are taken in 5 studies, and 1 is theoretical.

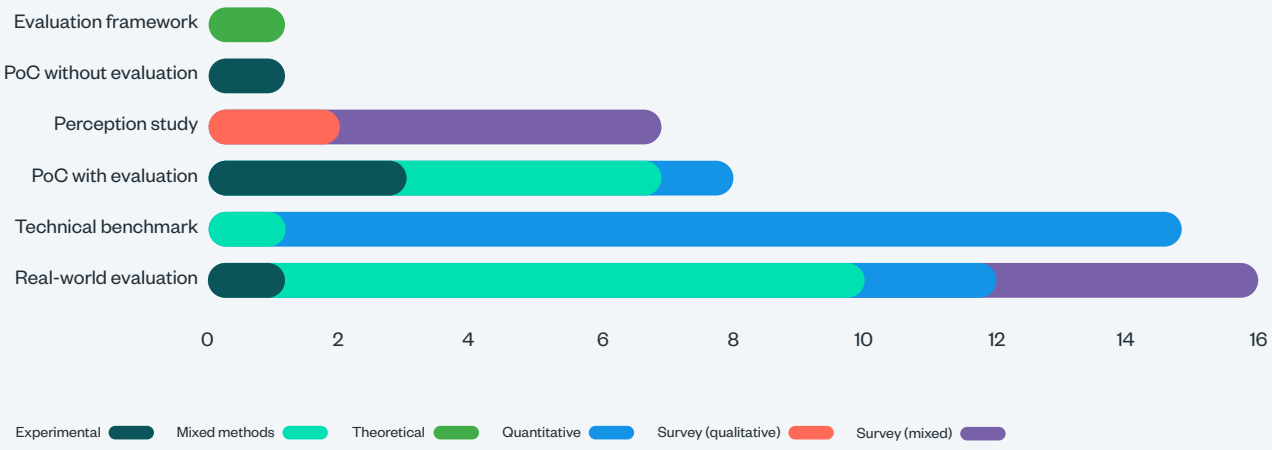


Figure 6: Distribution of studies by research design and method type.

The methodological spread reflects the different types of questions that different designs are equipped to answer. Technical benchmarks can assess task accuracy and error rates but cannot speak to broader system effects, user experience, or distributional impacts. Real-world evaluations vary considerably in rigour: several in this corpus are single-organisation pilots or evaluations conducted by the tool developer or commissioner, which limits independence and generalisability.

The distribution of evidence strength ratings confirms that the UK evidence base is mostly moderate. Only nine studies (19%) were assessed as producing strong evidence; the majority were rated moderate (n=27, 55%) or limited (n=12, 26%) in quality.

4.1.6 Research question coverage

The three research questions are not addressed with equal frequency or depth across the evidence.

Research question 1 – on individual and group experiences and outcomes – is addressed as a primary focus in 6 studies (13%), as a secondary focus in 12 (26%), and incidentally in 4 (9%). Over half the studies (n=26, 53%) do not address user or group outcomes at all. The under-representation of individual outcomes and differential impacts is a substantive concern that is returned to in the synthesis.

Research question 2 – exploring societal factors, including public trust, perceived legitimacy, and the acceptability of AI use in justice settings – has the fewest studies overall: 10 studies (21%) treat it as a primary focus and 5 address it secondarily, leaving 33 studies (68%) that do not address it at all. The near absence of empirical work on how AI deployment affects public understanding of, and confidence in, the justice system as a systemic institution represents a significant gap.

Research question 3 – relating to system efficiency and effectiveness – is the dominant concern of the evidence base: 31 studies (64%) treat it as their primary focus, with a further 2 addressing it secondarily and 3 incidentally. It is not addressed at all by 12 studies (26%).

4.2 International comparator landscape

The review's primary evidential commitment is to UK-based literature, and the synthesis in subsequent sections prioritises that corpus accordingly. However, the UK evidence base contains a number of domain-specific and thematic gaps that are significant enough to warrant drawing on international evidence where it can be applied with reasonable justification. In those areas, the review draws on a purposively selected international comparator corpus.

This is not a parallel systematic review of international literature. Instead, it is a deliberately selective comparator corpus of **109 studies**. Each study was included only where it addresses a research question or domain materially under-represented in the UK evidence base, and where a credible argument can be made that its findings are at least partially transferable to the UK legal context. The international studies are not representative and should not be read as a comprehensive account of the global literature on AI in justice settings; the corpus is a targeted resource drawn on to address specific UK gaps.

Much like the UK landscape, the international evidence base is heavily skewed towards recent developments, driven by the rapid emergence of generative AI. Approximately 77% of the included studies were published between 2023 and early 2026, with 15% between 2020 and 2023, and the remaining 7% spanning 2003 to 2020. Geographically, this rapidly expanding corpus is heavily anchored in North America – the US and Canada together account for nearly half of the included studies – alongside key comparator evidence from the European Union, China, and a diverse long tail of other global jurisdictions.

Similar to the UK evidence, LLMs are the most heavily studied technology, featuring as the primary AI type in over 40% of the international records, alongside strong representation of broader machine learning and NLP architectures. Thematically, the research spans a wide array of settings. While general cross-domain legal work (30 studies) and broad civil law applications (28 studies) are the most common categories, the international corpus also provides important comparative clusters in specific, high-impact domains such as housing, family, employment, and human rights law, which offer useful analogical insights for UK deployments.

The international corpus leans heavily towards rigorous, structured testing. Technical benchmarks constitute more than half of the evidence base (61 studies), reflecting a global academic focus on evaluating AI task accuracy and model performance against established legal datasets.

In terms of alignment with the review's core objectives:

- Exploratory evidence on differential impacts and on how lay users interact with AI-driven legal tools in practice (RQ1) provides valuable insights into the UK evidence gap, with RQ1 the primary focus of 15 studies.
- Perception studies and real-world evaluations within the international evidence base supply methodologically robust insights into judicial and lay attitudes towards AI (RQ2), explored primarily in 17 studies.
- Consistent with the high volume of technical benchmarking literature, system effectiveness and efficiency, model accuracy, and administrative burden dominate the international corpus (RQ3), serving as the primary focus for 77 of the 109 studies.

Where the synthesis chapters draw on international evidence, the transferability argument is set out explicitly. Where international findings appear to diverge from UK ones, the divergence and its possible reasons are reported.

5 Synthesis of evidence – individual and group experiences and outcomes (RQ1)

5.1 Evidential orientation

This section synthesises the available evidence bearing on Research question 1: what does the evidence reveal about how AI technologies affect people who are addressing or resolving legal problems, including differential impacts? The evidence available to address this question is limited and uneven. Despite the deployment of several tools in justice settings, the review did not identify any study that examined whether an AI tool had produced a measurable change in legal outcomes for individuals or groups within the justice system.

5.2 Access to legal services

Perhaps the most well-known UK deployment of an AI tool intended to expand access to legal services is Caddy, an LLM-based assistant tool developed by Citizens Advice Stockport, Oldham, Rochdale, and Trafford in collaboration with the Department for Science, Innovation and Technology (**DSIT**) and the Incubator for Artificial Intelligence (**i.AI**).[42] Caddy is designed to support frontline advisers handling enquiries by providing AI-generated draft responses. It is built on Anthropic's Claude LLM with RAG over Citizens Advice's own guidance and training materials. The development team built in a randomised controlled trial in which advisers were randomly assigned on each request, with queries directed either to Caddy or to a supervisor as per the pre-existing process. However, the results of the trial have not yet been published in full. The available evidence comes from an i.AI blog post which reports selected headline findings, but does not report the trial design, sample size, methods of analysis, or limitations. In the published results, there was no discussion of the experiences of the individuals seeking legal help, with reports only of the experiences of the advisers. The blog post itself notes that the trial "wasn't without issues" and promises a follow-up discussion, which has not appeared at the time of writing. The absence of a full published evaluation

42 AI Knowledge Hub, 'Caddy' (*AI.GOV.UK*, 23 May 2025) https://ai.gov.uk/knowledge-hub/tools/caddy/ accessed 21 January 2026; Andreas Varotsis, 'Transforming Civic Engagement with Caddy' (*Incubator for Artificial Intelligence*, 13 March 2025) https://ai.gov.uk/blogs/transforming-civic-engagement-with-caddy/ accessed 20 January 2026.

is a notable gap, given that Caddy has been rolled out to six local Citizens Advice offices, national expansion is planned for 2026, and it is being piloted in government departments.[43]

Several other academic studies explore the use of AI tools and methods to improve legal triage and intake in legal aid settings. Most of these studies report only the accuracy of the tools against benchmarks rather than effects on access in practice; they are accordingly discussed in Section 7. One international study warrants discussion here because it directly addresses differential impact across demographic groups, the kind of evidence the UK base does not contain.

Mistica and others (2021) conducted a study of Smart Assist, a tool deployed by Justice Connect Australia to route free legal assistance requests to appropriate legal teams.[44] Conducted in collaboration between the University of Melbourne, University of Queensland, and Justice Connect, they study a fine-tuned BERT model on real-world legal-help requests across multiple areas of law. Performance varied substantially by legal category. Nine categories achieved scores too low for operational use, including intellectual property, privacy, public and administrative law, and native title. The weakest-performing categories were typically those with the smallest training samples, suggesting that data sparsity rather than inherent difficulty was the primary constraint. The legal problems rarest in training data, and potentially those experienced by the most marginalised individuals, are the ones the system handles least well.

The most significant finding concerned differential performance by demographic subgroup. Fairness analysis across six subgroups revealed systematic disparities. Seniors experienced the worst overall performance (F1[45] of 0.540 compared with the general F1 of 0.671), and this underperformance was not explained by area of law: it was observed even in the tool's strongest categories (an F1 of 0.571 for seniors in the "Employees and Volunteers" category compared with 0.913 overall; and 0.545 compared with 0.793 overall in "Family"). Aboriginal and Torres Strait Islander users also experienced below-average performance overall (F1 of 0.638) and in criminal law specifically (F1 of 0.571 compared with 0.650 overall). The consistent gap in F1 score between demographic groups across the same task is analytically relevant and revealing of demographic bias.

The deployment context shares structural similarities with UK legal aid and advice triage, and the common-law framework offers a degree of comparability; the specific demographic subgroups, legal categories, and institutional context, however, differ materially. The findings transfer at the level of the general lesson (differential impacts on demographic groups can be substantial and may not be visible without disaggregated analysis) rather than at the level of specific quantification. No UK study identified in this review reports disaggregated tool performance or conducts an equivalent fairness

43 Incubator for AI, 'Caddy' (*GOV.UK*, 20 January 2026) https://ai.gov.uk/our-work/frontline-services/#caddy accessed 21 January 2026.

44 Meladel Mistica and others, 'Semi-Automatic Triage of Requests for Free Legal Assistance' *Proceedings of the Natural Legal Language Processing Workshop 2021* (ACL 2021) https://doi.org/10.18653/v1/2021.nllp-1.23.

45 The F1 score is a standard measure of classification accuracy, ranging from 0 to 1, where 1 represents perfect performance. It is calculated as a combination of precision (the proportion of predicted positives that are correct) and recall (the proportion of true positives correctly identified). There is no universal threshold for what counts as an acceptable F1 score; it depends on the task and context. As a rule of thumb in applied machine-learning practice, scores above 0.7 are often described as acceptable and scores above 0.8 as indicating strong performance, though these are conventions rather than formal rules. See 'F-Score', *Wikipedia* (2026) https://en.wikipedia.org/w/index.php?title=F-score&oldid=1344848713 accessed 27 May 2026; 'Machine Learning Glossary: Metrics' (*Google for Developers*) https://developers.google.com/machine-learning/glossary/metrics accessed 27 May 2026.

evaluation in a justice setting, and none evaluates whether correct classification translates into improved legal outcomes for the individuals classified.

A UK study from a parallel sector offers an instructive methodological comparator. Neto and others (2024) evaluated whether a machine learning model could accurately predict the need for social care interventions among young people referred to Early Help Services in local authorities in England and Wales.[46] The best model achieved a modest ability to differentiate between young people requiring some action and those who do not, with relatively high recall but a considerable number of false positives. No significant bias was detected for gender or for deprivation (as measured by the Income Deprivation Affecting Children Index), although statistically significant bias was identified for age and for school attendance rates. The authors advocate for further work on technical bias mitigation paired with responsible processes and practices, and on developing methods to deal with missing, uncertain, and sparse data on sensitive features. A limitation of the study is that the dataset excluded 23,420 young people who were never referred but potentially should have been. The measured bias may therefore understate the true differential impact, because it cannot capture the population for whom the system never produced an output at all. A similar issue is likely to arise in curating legal data: it will be challenging to identify data for legal issues that never reach formal advice or adjudication settings, even where the underlying need exists. The absence of ground truth in much legal data – taken together with the systematic under-representation of unmet legal need in the records the system has – places a meaningful methodological limit on how much fairness the analysis of justice-AI tools can establish without complementary qualitative work.

5.3 Access to legal information

5.3.1 Justice-specific legal information chatbots

There is limited evidence on user expectations of, and the impact of, AI legal information tools. Two early academic proof-of-concept evaluations examine users' comprehension and experiences of vulnerable users.

Morgan and others (2018) developed a proof-of-concept chatbot for the Children's Legal Centre Wales.[47] The system classifies children's messages by speech act and legal issue type, and extracts entities to generate a case summary for a human adviser. Technical performance, measured as the highest F1 score achieved across the system's classification tasks, was high (0.98). The user experience study asked 14 participants to converse with the chatbot and complete a qualitative questionnaire. User experience scores were positive on ease of use (average 7.42 out of 10), politeness (7.57), and whether they would use the chatbot again in future (8.29), but notably lower on whether participants believed they had been understood by the chatbot (5.00). This study relied on a small sample (n=14), and its experiment was limited in scenario and scope; nonetheless, the gap between technical performance and perceived comprehension is notable. Children and young people represent a population that may feel particularly unheard by formal institutions, and for whom access to justice takes a distinct form.

46 Eufrásio de A Lima Neto and others, 'Identifying Early Help Referrals for Local Authorities with Machine Learning and Bias Analysis' (2024) 7 Journal of Computational Social Science 385.

47 Jay Morgan and others, 'A Chatbot Framework for the Children's Legal Centre' *Frontiers in Artificial Intelligence and Applications*, vol 131 (IOS Press 2018).

Sasidharan and others (2013) conducted a proof-of-concept evaluation of EQUALS, a rule-based legal decision aid designed to help people with mental health challenges understand their rights under the Equality Act 2010.[48] The study introduces a different technological approach and a different set of user experience findings. EQUALS uses a rule-based architecture: its outputs are determined by a structured decision tree rather than a probabilistic language model. The design offers transparency and predictability (the user can see why the tool reached its conclusion) but constrains its scope and flexibility. The evaluation found that 7 of 22 participants were unaware, before using the tool, that mental illness could constitute a disability under the Act, suggesting that the tool served an awareness-raising function that participants would not otherwise have accessed. Usability improved between iterations when legal jargon was reduced. However, 3 of 22 participants experienced emotional distress during use, and 2 were identified as unsuitable to use the tool alone. A legal information tool designed for vulnerable users can, by surfacing the user's legal situation, cause distress as well as empowerment. The design implication then is that public-facing legal tools for sensitive domains should be embedded within support structures, not offered as standalone self-service products.

In 2020, the MoJ tested and conducted a limited evaluation of a Child Arrangements Information Tool, a public-facing closed chatbot on family law.[49] During the 12-day pilot, 1,121 users visited, although only 26% initiated a conversation and just 15% of those continued after the initial interaction. Users quickly perceived the tool's limited scope and described it as a closed loop. While referrals to external support reportedly tripled, the low continuation rate and user feedback suggest that the tool, in its proof-of-concept form, did not meet users' expectations. The experience suggests that unmet expectations can deter sustained engagement, a finding that is likely to apply to any public-facing legal tool whose scope is narrower than users anticipate.

That tool was discontinued, though with the recent development of LLMs, the MoJ is now testing related Digital Justice AI Assistants.[50] It is then helpful to consider AI justice chatbots in comparator jurisdictions.

JusticeBot, developed by academics and deployed in Québec for landlord-tenant disputes, provides an instructive comparator.[51] It uses a hybrid case-based and rule-based reasoning approach guiding users through structured legal reasoning pathways and providing references to previous similar cases. Several studies provide the most sustained examination of a domain-specific tool for self-represented litigants. JusticeBot reports more than 20,000 uses and over 140,000 page views. User surveys indicate that 63% of users felt they understood their situation better and 66% understood their next steps, with 89% willing to recommend the tool. The system correctly identified relevant legal issues in 93.5% of cases. The architecture, being non-generative, avoids hallucination

48 Padmaja Sasidharan and others, 'User Centered Evaluation of EQUALS, a Rule-Based Legal Decision-Aid' *Front. Artif. Intell. Appl.* (IOS Press BV 2013) https://doi.org/10.3233/978-1-61499-359-9-131.

49 Ministry of Justice, 'Child Arrangements Informational Tool' https://github.com/ministryofjustice/help-with-child-arrangements accessed 3 February 2026.

50 Justice AI Unit, 'Justice Chatbots' (*AI.GOV.UK*, 2025) https://ai.justice.gov.uk/ai-in-action/justice-chatbots accessed 21 January 2026.

51 Hannes Westermann and others, 'Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice' *Proc. Int. Conf. Artif. Intell. Law, ICAIL* (ACM 2019) https://doi.org/10.1145/3322640.3326732; Hannes Westermann and others, 'Bridging the Gap: Mapping Layperson Narratives to Legal Issues with Language Models' *CEUR Workshop Proc.* (CEUR-WS 2023) https://www.scopus.com/inward/record.uri?eid=2-s2.0-85167821739&partnerID=40&md5=0b1a34f155669aeb6c1dd5f212353c2c; Hannes Westermann and Karim Benyekhlef, 'JusticeBot: A Methodology for Building Augmented Intelligence Tools for Laypeople to Increase Access to Justice' *Proc. Int. Conf. Artif. Intell. Law, ICAIL* (Association for Computing Machinery 2023) https://doi.org/10.1145/3594536.3595166.

risks; an experiment comparing JusticeBot with ChatGPT illustrated this directly, with ChatGPT fabricating case references the rule-based tool did not.[52] The trade-off is that JusticeBot is limited to a narrow domain and lacks the conversational fluency of general-purpose LLMs.

In British Columbia, Beagle+ is a legal information AI chatbot developed by People's Law School in partnership with TangoWork. It is designed to provide plain-language civil-legal information on issues such as housing, employment, family, and small-claims matters. Beagle+ initially launched in 2020 based on Rasa technology but is now deployed on the ChatGPT platform, using RAG to ground responses in a curated, updated knowledge base, with continuous human-review and quality-control loops. The development team reports that pre-launch testing on 42 challenging questions showed accuracy of around 70%, but after iterative refinement of prompts and content this improved to over 99% accuracy on a dataset of 5,300 lawyer-reviewed conversations.[53]

By contrast, the Alaska Virtual Assistant (AVA) is a court-delivered, LLM-based chatbot focused narrowly on probate-related self-help, built in collaboration with the vendor LawDroid.[54] The project is most notable for its struggles with hallucination and reliability. During development, the team tested AVA on 91 probate-related questions but found that maintaining "100% accuracy" required more labour-intensive human review than they could sustain; they narrowed the evaluation set to 16 core questions and put the deployment on a more cautious footing.[55] AVA is now live, but no final evaluation nor accuracy results have been published.

The design of legal information tools engages specific design tensions. Chatbots that rely on rules-based or non-generative architectures (such as JusticeBot or EQUALS) avoid hallucination and reduce over-reliance, but they constrain scope and conversational fluency. LLM-based tools may offer more natural interaction but introduce reliability risks that, on the current evidence, cannot be eliminated even with extensive RAG, prompt engineering, and human review. The Beagle+ experience indicates that higher reliability is achievable with sustained institutional capacity for review and correction; the AVA experience gives some indication that such capacity is difficult to sustain at scale. Given that hallucination is a structural feature of LLM architectures, there is a challenging question of how much error is acceptable in legal information and advice settings.

5.3.2 Accessing legal help through general-purpose LLMs

The evidence in this subsection concerns the informal use of general-purpose LLMs by members of the public for legal information and assistance. It addresses three questions in turn: how widespread the informal use is; how reliable the outputs are for UK legal questions; and how lay users interpret, trust, and act on those outputs.

52 Jinzhe Tan, Hannes Westermann and Karim Benyekhlef, 'ChatGPT as an Artificial Lawyer?' *CEUR Workshop Proc.* (CEUR-WS 2023) https://www.scopus.com/pages/publications/85167430056.

53 People's Law School, 'Beagle+' (2024) https://www.peopleslawschool.ca/about/beagle/ accessed 3 April 2026; People's Law School, 'How to Test a GenAI Chatbot' (17 March 2025) https://www.peopleslawschool.ca/news/testing-genai-chatbot/ accessed 3 April 2026.

54 Alaska Court System, 'Alaska Virtual Assistant (AVA)' (1 April 2026) https://courts.alaska.gov/shc/AVA/faq.htm accessed 9 April 2026.

55 Jared Perlo, 'Alaska's Court System Built an AI Chatbot. It Didn't Go Smoothly.' (*NBC News*, 3 January 2026) https://www.nbcnews.com/tech/tech-news/alaskas-court-system-built-ai-chatbot-didnt-go-smoothly-rcna235985 accessed 7 April 2026; Frank Landymore, 'Court System Says Hallucinating AI System Is Ready to Be Deployed After Dramatically Lowering Expectations' (*Futurism*, 6 January 2026) https://futurism.com/artificial-intelligence/court-hallucinating-ai-ready-deployed accessed 7 April 2026.

The available survey evidence indicates that use of general-purpose LLMs is already widespread, including for matters with legal relevance. In late 2024, the Ada Lovelace Institute and The Alan Turing Institute conducted a nationally representative survey asking people about their experience with general-purpose LLMs. The survey showed 40% of respondents had used a general-purpose LLM such as ChatGPT, and around 8% had used a general-purpose LLM for guidance on issues such as legal disputes or taxation.[56] In Seabrooke and others (2024), survey evidence indicates that while relatively few UK participants have used LLMs for legal advice in the past (17%), nearly half reported they would be somewhat or extremely likely to do so in the future (further survey findings are discussed in Section 6.2).[57]

This survey evidence is also consistent with a growing number of reported UK court cases in which litigants in person, or those acting on their behalf, appear to have used publicly accessible LLMs in preparing documents. As of April 2026, at least 37 cases have been identified in which litigants in person prepared documents with LLMs, revealed through hallucinated citations.[58] Therefore, it is no longer hypothetical that the public are accessing legal help through general-purpose LLMs. Public-facing systems through which individuals may seek legal guidance are also proliferating, including CustomGPT tools such as AskEllie,[59] and Nessa.[60]

While hallucination rates and related performance issues are examined in more detail in Section 7.3, a separate discussion is needed here because the risks to lay users depend in part on how those users interpret, trust, and act on the outputs they receive.

Several informative studies examine the risks of the public using general-purpose LLMs for UK legal settings.

Ryan and Hardie (2024) tested general-purpose LLM performance on queries about English law developed from the authors' pro bono legal clinic:[61] 42% of responses were too generic to be useful, 25% were wrong in law, and 21% defaulted to the wrong jurisdiction (predominately US law). The evaluation of ChatGPT-3.5 (free), ChatGPT-4 (paid subscription), Bing Chat, and Google Bard found that overall only 13% of initial queries produced correct advice.[62] The authors evaluated the LLM-generated legal advice for accuracy by assessing whether it reflected current law, was comprehensive and correct, appropriately applied to the scenario, provided without additional prompting, and was clear to a non-lawyer audience. The authors distinguish three principal error types. First, jurisdictional errors: many first responses defaulted to US laws and did not make the jurisdictional difference sufficiently clear, with 21% of initial replies referring to the wrong jurisdiction. Second, generic or shallow answers: 42% of initial answers were so general that they would not enable a lay enquirer to identify their rights or the source of law. Third, substantive legal errors: 25% of responses were wrong in law, not through invented cases, but through reliance on outdated statutes,

56 Roshni Modhvadia and others, 'How Do People Feel About AI? Wave Two of a Nationally Representative Survey of UK Attitudes to AI Designed through a Lens of Equity and Inclusion' (Ada Lovelace Institute and The Alan Turing Institute 2025) https://attitudestoai.uk/.

57 Tina Seabrooke and others, 'A Survey of Lay People's Willingness to Generate Legal Advice Using Large Language Models (LLMs)' *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems* (ACM 2024) https://doi.org/10.1145/3686038.3686043.

58 Lee (n 20).

59 AskEllie (n 33).

60 OLS Solicitors (n 33).

61 Francine Ryan and Liz Hardie, 'ChatGPT, I Have a Legal Question? The Impact of Generative AI Tools on Law Clinics and Access to Justice' (2024) 31 International Journal of Clinical Legal Education 166.

62 ibid.

omission of key legal requirements or timescales, and incorrect summaries of the applicable legal regime. These errors were material rather than merely superficial because they could cause litigants in person to miss limitation periods, pursue ineffective or impossible remedies, or be misdirected about enforcement. Performance varied by legal domain: consumer law and housing queries scored highest; family law, employment law, and child maintenance queries produced potentially misleading advice, including reliance on superseded legislation, omission of critical time limits, and incorrect statements about court procedures. ChatGPT-4 (paid subscription) substantially outperformed free tools, creating a quality differential that tracks existing socio-economic inequalities in access to legal information. Ryan and Hardie identify the significant risks of litigants in person relying on LLMs for legal information. They explain that while these tools might be "useful for someone with some understanding of the law, who was able to review and ask appropriate follow-up questions (such as jurisdiction), many litigants in person have limited legal knowledge and may not appreciate the need to check the jurisdiction or source of law".[63]

Curran and others (2025) also identify geographic variation in LLM hallucination rates across three common-law jurisdictions.[64] For legal questions situated in Los Angeles, the average hallucination rate was 0.45; for London, 0.55; for Sydney, 0.61. Across all locations, housing had the highest hallucination rate (0.649), followed by crime and traffic (0.547), family (0.511), and employment (0.440). The authors attribute the geographic variation in part to differences in training data density: where online legal texts and related publications are more prevalent, hallucination rates on related queries tend to decrease. The implication is that UK users seeking help with everyday legal problems from general-purpose LLMs may receive systematically less reliable information than users in comparable jurisdictions. Such risk is concentrated in precisely the areas of law where unmet legal need is highest.

Deroy and Maity (2023) identify gender and jurisdiction biases in AI-generated case judgment summaries produced by both legal-domain and general-purpose LLMs.[65] General-purpose LLMs demonstrated gendered terminology skew in UK case summaries, while legal-specific LLMs showed a strong bias towards US content. This suggests that the composition of training data may produce systematic distortions in how cases are summarised, with implications for both lay users and practitioners who rely on such tools. More broadly, the findings raise the possibility that AI-generated legal summaries may frame cases in ways that reflect, and potentially reinforce, gendered assumptions, with implications for both lay users and practitioners who rely on such tools. It is also consistent with earlier studies suggesting that general-purpose LLMs display a broader US bias in legal contexts.

Two more studies reinforce these findings of high citation hallucination for UK law, although in more limited settings. The Linklaters AI English Law Benchmark (2023) tested 50 questions from 10 different practice areas on GPT2, GPT3, GPT4, and Bard. Across all models, none of the answers had adequate citations; in a third of all responses the cited law was totally fabricated. Bard achieved the highest scores overall.[66] McLaren and Rowe (2025) tested ChatGPT-4, Gemini 1.5, Claude

63 ibid 185.

64 Damian Curran and others, 'Place Matters: Comparing LLM Hallucination Rates for Place-Based Legal Queries' (arXiv, 10 November 2025) https://doi.org/10.48550/arXiv.2511.06700.

65 Aniket Deroy and Subhankar Maity, 'Questioning Biases in Case Judgment Summaries: Legal Datasets or Large Language Models?' (arXiv, 1 December 2023) https://doi.org/10.48550/arXiv.2312.00554.

66 Peter Church, 'The LinksAI English Law Benchmark – Can You Rely on AI for Legal Advice?' (*Passle*, 30 October 2023) https://techinsights.linklaters.com//post/102irc3/the-linksaienglish-law-benchmark-can-you-rely-on-ai-for-legal-advice accessed 24

Sonnet, Llama, Copilot, and Perplexity by requesting information about a fabricated UK case citation.[67] Most platforms generated plausible but entirely fabricated case summaries, appearing legitimate with professional legal structure, formatting and referencing; only ChatGPT-4 and Claude Sonnet immediately questioned the authenticity of the citation. The study is limited to a single fabricated citation with no systematic variation but evidences the ongoing challenges of LLM citation hallucination on UK law.

Similarly, two specific international studies reveal certain biases in general-purpose LLMs used in legal settings.[68] Harašta and others (2025) present preliminary evidence of gender bias in LLM outputs in a Czech family law context, with descriptive patterns suggesting that some models may favour the female parent in shared parenting allocations when certain risk factors apply.[69] The pattern is strongest for some models (notably Claude Haiku 4.5), and weaker but still visible for others (Gemini 2.5 Flash). The authors flag high uncertainty due to the absence of statistical testing and treat this as exploratory work for future development. Pulvino and others (2026) study prosecutorial recommendations from ChatGPT in US criminal law and find that the model systematically recommends prosecution regardless of prompt framing, evidence quality, or constitutional violations.[70] Prosecutor-framed prompts increased prosecution recommendations by 6.6 percentage points; defence-framed prompts reduced but did not reverse this tendency. No statistically significant racial bias was detected in outputs, though the authors note that a consistent pro-prosecution tendency is likely to produce racially disparate real-world impacts given existing disparities in who enters the criminal justice system. The finding is included here for completeness; criminal justice falls outside the primary scope of this review, but the underlying mechanism (an LLM showing strong directional default behaviour regardless of fact-pattern framing) has potential transferability to civil and administrative legal settings where similar default behaviours may operate undetected.

There are no observational studies on how people use LLMs for legal help in the UK. The closest international evidence is provided by two studies.

Kuk and Harašta (2025) provide the most detailed empirical account of lay-user behaviour when seeking AI-generated legal assistance.[71] The study conducted an observational analysis of 3,847 queries from 1,252 users submitted to a GPT4-powered tool for addressing Czech legal queries. The findings reveal a substantial gap between how users communicate with LLMs and how a legal expert would frame a legal information query. Of the queries analysed, 70.1% contained no factual context, meaning users asked legal questions without providing the circumstances that would

January 2026.

67 Sally McLaren and Lily Rowe, '"You're Right to Be Skeptical!": The Role of Legal Information Professionals in Assessing Generative AI Outputs' (2025) 25 Legal Information Management 19.

68 Sukanth Korkanti, 'Detecting Bias in Legal Language Models through Data Analytics' *Int. Conf. Self Sustain. Artif. Intell. Syst., ICSSAS* (IEEE 2024) https://doi.org/10.1109/ICSSAS64001.2024.10760917; Zongyue Xue and others, 'JustEva: A Toolkit to Evaluate LLM Fairness in Legal Knowledge Inference' *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (ACM 2025) https://doi.org/10.1145/3746252.376149.

69 Jakub Harašta and others, 'Gender Bias in LLMs: Preliminary Evidence from Shared Parenting Scenario in Czech Family Law' *Workshop on AI for Access to Justice, Dispute Resolution, and Data Access (AIDA2J) at Jurix 2025* (2025) https://doi.org/10.48550/arXiv.2601.05879.

70 Rory Pulvino, Dan Sutton and JJ Naddeo, 'Hiding in Plain Sight: An Empirical Study of Prosecutorial Bias in AI Legal Analysis' (2026) 27 Science and Technology Law Review 1.

71 Michal Kuk and Jakub Harašta, 'LLMs & Legal Aid: Understanding Legal Needs Exhibited Through User Queries' *AI for Access to Justice Workshop, JURIX 2024* (2025) https://doi.org/10.48550/arXiv.2501.01711.

otherwise enable a lawyer, let alone a model, to give a meaningful answer. Most queries (64.9%) sought legal information, while others sought advice on a course of action (35.1%), a distinction that matters both for professional responsibility and for the kind of reasoning the model must perform. Open-ended queries, which granted the model discretion over the form and content of its response, constituted 71.4% of queries, compared to 28.6% that attempted to impose requirements on the model answer. Only 3.4% of queries combined all three characteristics of factual grounding, specificity, and structural clarity, that would approximate how a legal professional would frame a legal query or factual information.

Two further findings carry particular weight. First, users frequently overshared personal and sensitive information, including details about health conditions, family situations, and financial circumstances, without apparent awareness that this information might be stored, processed, or used in ways they did not intend. Second, the conversational mode of interaction led users to treat the chatbot as a trusted interlocutor rather than as a tool with known limitations. These patterns suggest that the risk associated with AI legal information tools is informed by user engagement. The gap between casual lay communication and the precision required for reliable legal reasoning is itself a source of risk for lay users. It is also a risk that this could be a source of differential impact. Users with greater legal literacy, education, and digital confidence will receive better outputs than those without, reproducing existing inequalities through a new mechanism.

Hagan's (2024) qualitative study of community members' interactions with Google Bard in a legal help scenario provides a complementary, and in some respects more fine-grained, account.[72] The study drew on structured interviews and design sessions with 15 US adults, who were asked to use Bard to respond to a fictional eviction notice and then to evaluate the tool's usefulness, suggest improvements, and indicate their preferences regarding interface design and complexity. The sample was small, non-representative, and skewed towards higher-income participants who were relatively confident with technology. These limitations restrict the study's generalisability, but they do not negate the value of its observational findings.

There are three particularly relevant findings. First, participants' trust in the AI tool increased after using it: the average trust rating rose from 2.7 out of 6 before the exercise to 4.2 out of 6 afterwards. The participants found the tool's output clear, specific, and apparently authoritative. Second, the study reveals that participants' increased trust was not matched by an increased capacity to identify quality problems in the AI's output. Hagan identifies three categories of problematic output which she collectively terms "ersatz legal help": outright hallucinations (fabricated legal cases), correct information presented without sufficient context, and correct but inapplicable information (such as referrals to organisations in the wrong jurisdiction or organisations that offer no direct services). Participants did not recognise these errors. Third, Hagan identifies three distinct user archetypes in how participants engaged with and intended to use the AI's output. A minority (2 of 15) treated the AI's response as evidence to be deployed directly, for example, screenshotting or copying the output to send to a landlord as a statement of their legal rights. A second group (4 of 15) sought definitive legal answers, cherry-picking specific details, and sometimes using these without verification. A third group, the majority (7 of 15), treated the AI's output as a framework for further research that helped them identify relevant questions and organisations, which they would then verify through other sources. Only this third group displayed the kind of critical engagement that would mitigate the risks

72 Margaret Hagan, 'Towards Human-Centred Standards for Legal Help AI' (2024) 382 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20230157.

of reliance on unreliable output. The study also found that participants largely disregarded or actively resisted warnings about the limitations of the tool for legal advice.

These findings, taken together, carry implications that extend beyond the individual studies. The studies address different aspects of lay engagement: Kuk and Harašta capture the input side (how people formulate queries), while Hagan captures the output side (how people assess, trust, and act on responses), but they converge on a common conclusion. The design of AI legal tools should account for how the public interact with them. If lay users often omit key facts in queries or over-trust flawed outputs, it raises risks for self-help scenarios. Studies predating LLMs have also identified the challenges of deploying AI methods on lay-user text as it does not match the language in training data that is generated by the legal profession.[73] The challenge is exacerbated by the increasing accessibility of these publicly available general-purpose LLMs.

5.4 Procedural experience

The evidence on the procedural experience of AI tools in legal settings is sparse. In that gap, it is worth noting the structural conditions into which such tools are deployed.

The recent HMCTS Reform Evaluation, while generally outside the scope of this review, provides relevant evidence of demographic disparity in existing digital justice processes.[74] The evaluation reports that digital uptake varies by ethnicity, disability status, age, and main language. Black, African, Caribbean, or Black British probate applicants had the lowest digital uptake. Applicants with disabilities that significantly limited them were more likely to use the paper channel. In the probate evaluation, quantitative evidence identified that applicants from ethnic minority backgrounds experience consistently longer case durations, higher rates of application 'stops', and different stop reasons compared with White applicants.[75] Where the underlying digital system already disadvantages some groups, AI tools could amplify those inequalities unless their design explicitly accounts for pre-existing divides.

Given the large number of pilots and trials of AI tools being undertaken and explored by the MoJ and HMCTS, the absence of any reported findings or evaluations that reflect on user impacts or procedural experience is a notable gap.

Even tools that have a published evaluation do not measure details relevant to the individual or group experience. The Asylum Case Summarisation (ACS) and Asylum Policy Search (APS) tools were designed to assist asylum decision-makers: ACS uses OpenAI's GPT-4 to summarise asylum interview transcripts, which can extend to 50 pages or more, while APS is a chat-based search assistant that retrieves and summarises Country Policy Information Notes and Country of Origin

73 Elisabeth Uijttenbroek and others, 'Retrieval of Case Law to Provide Layman with Information about Liability: Preliminary Results of the BEST-Project' *Lect. Notes Comput. Sci.* (2008) https://doi.org/10.1007/978-3-540-85569-9-19; Karl Branting and others, 'Judges Are from Mars, pro Se Litigants Are from Venus: Predicting Decisions from Lay Text' *Front. Artif. Intell. Appl.* (IOS Press BV 2020) https://doi.org/10.3233/FAIA200867; Westermann and others, 'Bridging the Gap: Mapping Layperson Narratives to Legal Issues with Language Models' (n 51); M Büttner and Ivan Habernal,'Answering Legal Questions from Laymen in German Civil Law System' *Conf. European Chapter Assoc. Comput. Linguist., ECAL* (ACL 2024) https://doi.org/10.18653/v1/2024.eacl-long.122.

74 Gemma Pilling and others, 'HMCTS Reform Evaluation Thematic Report: Digitalisation' (Ministry of Justice Analytical Series 2026) https://www.gov.uk/government/publications/hmcts-reform-evaluation-thematic-report-digitalisation accessed 26 March 2026.

75 Dr Ryan Wilkinson, Tyler Opoku and Katherine Jones, 'Reform Evaluation Data Summary: Probate' (2026) HMCTS Reform Evaluation.

Information Requests.[76] Both tools operate under a human-in-the-loop design and neither can be used to make a decision directly. A full rollout of both tools was planned for January 2026. The evaluations were conducted by Home Office research teams, using comparison groups drawn from decision-making units. However, the Home Office evaluations did not report on individual's procedural experience of the tools and the research teams were unable to conduct demographic bias analysis due to the limited volume of pilot cases. Equality Impact Assessments were reportedly completed in compliance with the Public Sector Equality Duty, but have not been published.[77] The tools reportedly produced inaccurate outputs in approximately 9% of cases. Without disaggregated analysis it is not possible to determine the source or distribution of those errors, whether error rates varied across individuals with different national origins, languages, or case profiles, or whether any demographic group was systematically disadvantaged.

76 Home Office, 'Evaluation of AI Trials in the Asylum Decision Making Process' (*GOV.UK*, 29 April 2025) https://www.gov.uk/government/publications/evaluation-of-ai-trials-in-the-asylum-decision-making-process/evaluation-of-ai-trials-in-the-asylum-decision-making-process accessed 20 January 2026; AI Knowledge Hub, 'Asylum Case Summarisation (ACS)' (12 September 2025) https://ai.gov.uk/knowledge-hub/ accessed 3 April 2026.

77 Home Office (n 76).

5.5 Computational analysis of differential impact

The integration of data-driven technologies into the justice system presents a dualistic paradigm. While AI systems risk codifying new inequalities or exacerbating entrenched disparities,[78] they concurrently offer unprecedented opportunities for the empirical evaluation of systemic unfairness. By leveraging computational scale, researchers can scrutinise judicial and procedural patterns that were previously obscured by the sheer volume of analogue records. This subsection briefly reviews two studies that illustrate the opportunities and the limits of computational approaches to differential-impact analysis.

Although not strictly an AI-based study, Blackham (2021) provides a foundational example of how increased digitalisation and computational methods can be used for evaluations of differential impact in the justice system.[79] Following the publication of UK Employment Tribunal decisions online, Blackham curated a dataset of 1,208 age discrimination cases and undertook an empirical analysis. The study found that claimant demographics in decisions were skewed towards older claimants, particularly men, revealing substantial gaps in the enforcement of age discrimination law, particularly for young workers and older women. Furthermore, the data quantified a representation gap, finding that successful claimants with legal representation (46.5%) significantly outperformed the baseline success rate across all cases (34.7%). Such data-driven findings demonstrate how computational methods can better inform attention and effort to overcome differential impacts within the justice system.

As evaluations shift towards machine learning, new epistemological challenges emerge. Barale and others (2025) conduct a machine learning-driven fairness assessment on the AsyLex dataset of 59,112 Canadian refugee decisions.[80] The study detects demographic disparities across gender, sexual orientation, age, as well as judge, year, and city. However, the different methods yield divergent and sometimes contradictory signals. The central conclusion is that standard machine learning methods for fairness evaluations are insufficient for domains in which legitimate discretion is exercised because they cannot distinguish legally justified variation from unjustified bias. The authors emphasise that legal reasoning and institutional context, which is not always identifiable solely in case judgment data, is indispensable to meaningful fairness evaluations.[81] This conclusion has broad implications for how technical bias evidence in justice-related AI should be interpreted and for the design of future evaluations.

While the presence of biases in legal data that may be used for training AI has been theoretically described for UK legal data,[82] it has not been empirically investigated as in other jurisdictions.[83] Several studies explore the detection of bias in legal corpora, consistently finding that the specific

78 Justice UK (n 25).

79 Alysia Blackham, 'Enforcing Rights in Employment Tribunals: Insights from Age Discrimination Claims in a New "Dataset"' (2021) 41 Legal Studies 390.

80 Claire Barale, Michael Rovatsos and Nehal Bhuta, 'When Fairness Isn't Statistical: The Limits of Machine Learning in Evaluating Legal Reasoning' (arXiv, 4 June 2025) https://doi.org/10.48550/arXiv.2506.03913.

81 See also Václav Janeček, 'Judgments as Bulk Data' (2023) 10 Big Data & Society.

82 Holli Sargeant and Måns Magnusson, 'Bias in Legal Data for Generative AI' *Generative AI & Law Workshop at ICML 2024* (2024) https://blog.genlaw.org/pdfs/genlaw_icml2024/9.pdf accessed 3 April 2026.

83 Noa Baker Gillis, 'Sexism in the Judiciary: The Importance of Bias Definition in NLP and In Our Courts' *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing* (ACL 2021) https://doi.org/10.18653/v1/2021.gebnlp-1.6 Jackson Sargent and Melanie Weber, 'Identifying Biases in Legal Data: An Algorithmic Fairness Perspective' (arXiv, 21 September 2021) https://doi.org/10.48550/arXiv.2109.09946; Ece Gumusel and others, 'An Annotation Schema for the Detection of Social Bias in Legal Text Corpora' *Lect. Notes Comput. Sci.* (Springer 2022) https://doi.org/10.1007/978-3-030-96957-8_17; Nurullah Sevim, Furkan Şahinuç and Aykut Koç, 'Gender Bias in Legal Corpora and Debiasing It' (2023) 29 Natural Language Engineering 449; Mustafa Bozdag, Nurullah Sevim and Aykut Koç, 'Measuring and Mitigating Gender Bias in Legal Contextualized Language Models' (2024) 18 ACM Transactions on Knowledge Discovery from Data 79:1.

character of bias in legal language requires domain-specific methods rather than general NLP approaches. The distinction between statistical bias in outputs and systemic impact on individuals will be a significant challenge for evaluations of AI tools in algorithmic settings.

5.6 Summary of findings

The evidence base is inadequate for answering whether AI tools improve or worsen the experiences and outcomes of people addressing or resolving legal problems. The deployments most directly oriented to improving access (such as Caddy) are operating without publicly available evaluation of their effects on the people seeking help. Where evaluations exist, they typically omit analysis of actual impacts on users, material impacts on legal outcomes, or differential impacts across individuals or groups.

Similarly, some studies have identified risks of demographic bias in AI tools used for legal triage and information, but they do not translate measured statistical disparities into evidence of effects in real-world deployment settings. They do highlight the risk of using AI tools across domains with more limited training data, or where data are likely to be systematically missing for groups whose legal needs do not reach formal action. Patterns of gender, jurisdiction, and prosecutorial bias have also been identified in LLM tools. Examining patterns of disparity is made more accessible by computational methods, although, given the absence of ground truth in most legal data, those methods cannot distinguish issues of legal reasoning and justified differentiation from bias.

Increasingly, the UK public are using public-facing, proprietary, general-purpose LLMs for legal help. Despite the potential for such everyday intuitive tools, there are several significant risks of such use. First, LLM outputs are confident but often inaccurate. These errors and hallucinations are exacerbated for UK law and in different legal domains. Second, user trust and reliance on LLM outputs are often misaligned with users' own capacity to evaluate outputs. Finally, the public often describe legal issues without the precise language used by legal experts that would help improve the quality of output. Model accuracy, quality, and reliability are discussed in more detail in Section 7, although this analysis makes clear that several core risks arise from individual experiences, particularly for less legally informed users.

A balanced reading of this evidence does not, however, support a wholly negative conclusion. Hagan's study indicates that a majority of the small sample (7 out of 15 participants) treated AI outputs as a starting point for further enquiry rather than as definitive answers. There are real risks in current public use of general-purpose LLMs for legal help, and there are also meaningful groups of users who engage critically with AI outputs and would benefit from better-designed tools. AI tools, if designed for legal tasks and supported by appropriate scaffolding, may provide entry points and orientation for users who would otherwise be poorly served.

The design of legal information tools requires balancing competing interests. Chatbots that rely on rules-based or non-generative architectures remove the risk of hallucination and over-reliance but may be limited in scope and reduce user comprehension. While LLM-based tools may offer more natural interaction, their reliability raises serious risks for general public use. Considering differential uptake across certain groups in HMCTS digital processes, it is possible that uptake of public-facing legal AI tools will also differ. Whether and how AI advances access to justice in this domain will depend not only on the tools but on the institutional conditions of their use.

6 Synthesis of evidence – societal factors (RQ2)

6.1 Evidential orientation

This section synthesises the available evidence bearing on Research question 2: what does the evidence reveal about the impacts of AI on wider societal issues such as public trust in justice, public understanding or scrutiny of justice, and the acceptability of AI use in justice settings? It also addresses the closely related questions of professional attitudes towards these technologies and the conditions for open justice.

The character of this evidence base is dominated by surveys, vignette experiments, and field studies on perceptions of AI in justice. Most studies ask participants how they would respond to hypothetical scenarios involving AI in justice, rather than how they responded to actual encounters. The distinction between hypothetical and experiential evidence is important for interpretation. Attitudinal research tells us something about predispositions and preferences, but it cannot tell us with confidence how people will behave, or how they will feel, when they encounter AI in a real justice process with real consequences. There is a near absence of longitudinal or outcome data capturing how trust, legitimacy, or public understanding shift over time following the deployment of AI tools in justice settings.

6.2 Public trust, acceptability, and perceived fairness

6.2.1 UK public perceptions

The strongest UK evidence on lay perceptions of AI in justice settings comes from a small number of experimental surveys, each addressing different aspects of perceived acceptability and fairness. Taken together, these studies suggest a public that is neither uniformly hostile to AI in justice nor uncritically accepting of it. Attitudes are shaped by context, domain, and the availability of information about how AI tools are used.

The most striking finding is found in Tomlinson and others (2025), which provides the largest-scale experimental investigation of public perceptions of AI in UK administrative tribunals.[84] The study comprises two randomised between-subject survey experiments in which UK respondents (quota-sampled as representative of the UK) were assigned either to a control vignette or to a vignette describing AI assistance in the Social Entitlement Chamber of the First-tier Tribunal. One experiment focused on bundle summarisation (n=1,186) and the other on AI-generated hearing

84 Joe Tomlinson and others, 'The Hidden Cost of Quick Wins: How Artificial Intelligence in Administrative Appeals Risks Trust in Justice' *Cambridge Handbook of Artificial Intelligence and Public Law* (Cambridge University Press 2025).

summaries (n=1,242). Perceptions were then measured across topics including fairness, dignity, privacy, transparency, and informed decision-making.

The study finds that even limited, assistive uses of AI in the tribunal processes can reduce perceived procedural justice and legitimacy, with statistically significant differences across all measured dimensions in both experiments. In the first vignette, in the status quo condition, 33.3% of participants strongly agreed that the process allowed the judge to make an informed decision; under the AI-summarised bundle condition, this fell to 7.5%. Only 21.7% of the AI group agreed that the process ensured fairness, compared with 57.9% in the control group. The second vignette produced smaller but still negative disparities. In the status quo condition, 30.4% agreed that the process allowed for informed decision-making, which fell to 19.0% where AI was used to record and summarise a hearing. On fairness, 14.5% of the AI group agreed the process ensured fairness compared to 23.3% in the status quo group. Qualitatively, respondents often viewed AI tools as shortcuts inconsistent with the judicial role, as a threat to voice, dignity, and careful human consideration, and in some cases as a potential instrument of governmental manipulation. The findings provide strong evidence of negative attitudinal responses among the UK public to AI in tribunal vignettes; they cannot, on their own, establish how individuals would respond to AI in tribunal processes that affected their own cases.

A different dimension of the public trust question is explored in Schneiders and others (2025), which examines attitudes towards AI-generated legal advice across three experiments (total n=288).[85] When the source of legal advice was unknown, UK participants report significantly higher willingness to act on the LLM-generated advice than on the lawyer-generated advice. However, when the source of the advice was known, no significant differences were observed. Additionally, participants were able to distinguish LLM from lawyer-generated texts with 59% accuracy. While the study cannot answer why participants had increased trust in LLM-generated text, the authors note that the LLM advice appeared more complex and often more confident than the lawyer-generated text, a function of the LLM training. The authors suggest future work to examine the reasons for the potential over-reliance on AI.

Seabrooke and others (2024) offer a broader mapping of UK participants' willingness to use LLMs for legal advice across subdomains.[86] Only 17% of participants (n=105) reported having used an LLM for legal advice, but 45% indicated they would be somewhat or extremely likely to do so in future. Such willingness varied by legal subdomain: it was highest for tenancy issues (58%), planning (55%), tax law (53%), and traffic issues (52%), but considerably lower for divorce (39%) and civil disputes (25%). The authors observe that the differential may reflect either the perceived likelihood of encountering each type of problem, or the perceived stakes, emotional weight, and complexity of the legal problem at issue. They highlight the likelihood that non-experts will use LLMs for accessing legal information and the importance of supporting users to recognise the risks involved.

In relation to AI notetaking tools in social care, Meers and others (2026) contribute a typological approach, identifying four attitudinal clusters among the public: "enthusiasts", "cautious adopters", "pragmatists", and "resistant".[87] The study compares attitudes towards manual human notetaking,

85 Eike Schneiders and others, 'Objection Overruled! Lay People Can Distinguish Large Language Models from Lawyers, but Still Favour Advice from an LLM' *Conference on Human Factors in Computing Systems* (ACM 2025) https://doi.org/10.1145/3706598.3713470.

86 Seabrooke and others (n 57).

87 Jed Meers and others, 'What Do the Public Think about Artificial Intelligence Note-Taking Tools in Social Care?' (2026) 1 European Social Work Research 1.

fully automated notetaking, and human-in-the-loop systems. AI-supported notetaking without human review was consistently ranked lower across all procedural fairness dimensions, while maintaining a human-in-the-loop preserved similar levels of trust to traditional manual notetaking methods. The typological variation suggests that binary framings of public acceptance or rejection are misleading, and the consistent preference for human-in-the-loop indicates that the presence (rather than mere availability) of human oversight is central to perceptions of procedural fairness.

6.2.2 International lay perceptions

Several international studies present a more varied picture than the UK evidence alone would suggest.

Flanagan and others (2025) provide a notable counterweight to the UK findings.[88] The nationally representative survey of Kenyan participants (n=2,246) examined perceptions of judicial decisions informed by legal analysis from either a human or an AI law clerk in a randomised between-subject vignette experiment. The study found no significant difference in perceived legitimacy between the two conditions, and AI-clerk decisions were rated above the scale mid-point, indicating overall acceptance. This is a notable divergence from the UK evidence and raises the question whether the trust deficit observed in UK studies is a feature of specific institutional contexts, levels of pre-existing institutional trust, or framings used in the UK experiments, rather than a universal response to AI in judicial settings.

Fine and others (2025) conducted a US-based survey of public perceptions of judicial use of AI tools.[89] In a survey of 1,800 participants, the study reports perceptions across three conditions: judging with no AI tools, judging with AI as a supplement, and judging relying solely on AI. Similar to the UK findings, the respondents viewed judging without AI tools most favourably, with the sharpest drop where the judge relied entirely on AI, particularly on perceived legitimacy and procedural justice. The study also considered demographic variation. Black participants rated legitimacy and procedural justice somewhat higher overall than White and Hispanic participants, and perceived judicial trust in AI was positively associated with participants' own trust in AI, with a stronger effect among Black participants. The authors interpret this as suggesting that AI may be perceived by some groups as a potential corrective to human bias, even as it is perceived by the general population as reducing legitimacy. This identifies an important complexity in that the legitimacy effects of AI in justice may vary across communities – and for groups that perceive the existing system as biased, AI may be seen not as a threat to legitimacy but as an opportunity to improve it.

As mentioned in Section 5.3, Hagan (2024) is briefly reiterated here as the study provides a distinctive finding on public trust. In the small US study, participants' trust in an AI legal help tool increased substantially after use (from 2.7/6 to 4.2/6), despite the tool producing hallucinated case citations, decontextualised information, and incorrect jurisdictional referrals.[90] Three user archetypes were identified: those who would use AI output as direct evidence (2/15), those who cherry-picked legal details without verification (4/15), and those who used the output as a starting

88 Brian Flanagan and others, 'The Rule of Law or the Rule of Robots? Nationally Representative Survey Evidence from Kenya' [2025] Information and Communications Technology Law.
89 Anna Fine, Emily Berthelot and Shawn Marsh, 'Public Perceptions of Judges' Use of AI Tools in Courtroom Decision-Making: An Examination of Legitimacy, Fairness, Trust, and Procedural Justice' (2025) 15 Behavioral Sciences.
90 Hagan (n 72).

point for further research. The post-use trust increase, in the face of objectively poor outputs, raises questions about the calibration of lay trust and the adequacy of current mechanisms for enabling users to evaluate the reliability of AI-generated legal information.

6.2.3 Bridging public and professional attitudes

Ludwig and others (2024) provide a pre-registered survey experiment directly comparing the German general population (n=298) and judges at Bavarian regional courts (n=267).[91] The study focuses on whether perceived inequality increases acceptability of AI use in sentencing decisions and reveals two findings of particular relevance. First, judges were significantly more sceptical than laypeople about AI-informed decision-making, with a large effect across all attitudinal dimensions. Second, when participants were exposed to information about sentencing inequality, lay acceptance increased, but this effect was absent among judges, who were already aware of such disparities. The implication is that public attitudes towards AI in justice may be conditioned by perceptions of existing human failings in the system, a dynamic with potential relevance to UK debates about AI tools in contexts where human decision-making is already subject to criticism.

6.3 Practitioner and professional attitudes

This section examines evidence on how judges, lawyers, and other legal professionals perceive AI tools. Adoption rates and barriers are discussed within the evidence on legal professions as it is generally embedded within broader attitudinal findings. The treatment is strictly evidential: it reports what surveyed and interviewed practitioners say, and what observational and field-study evidence indicates about adoption patterns. Commentary and strategic direction set by the UK judiciary, which the available evidence base does not fully capture, are discussed in Section 8.

6.3.1 Judicial attitudes and concerns

The UK evidence on judicial attitudes is anchored by Solovey and others (2025), a qualitative study of 12 judges ranging from the Supreme Court to the County Court.[92] Judges expressed openness to AI for legal research, summarisation, and some administrative tasks, but insisted on personal responsibility for all outputs. Concerns centred on three themes: the risk of deskilling, particularly in lower courts where judges feared that reliance on AI-generated summaries or draft reasoning could erode the analytical competencies required for the judicial role; the normative importance of human judgment, including the value that parties derive from knowing that a human being has engaged with their case; and the reliability of AI outputs, particularly the risk of hallucinations in legal research contexts. The study suggests that any judicial resistance to AI is not necessarily technophobia but may reflect a substantive view about the conditions under which justice is perceived as legitimate. The small sample size and self-selected participants limit the generalisability of these findings, but the study is the only directly relevant UK evidence on judicial attitudes currently available.

91 Jonas Ludwig and others, 'Inequality Threat Increases Laypeople's, but Not Judges', Acceptance of Algorithmic Decision Making in Court' (2024) 48 Law and Human Behavior 441.

92 Erin Solovey, Brian Flanagan and Daniel Chen, 'Interacting with AI at Work: Perceptions and Opportunities from the UK Judiciary' *Proceedings of the 4th Annual Symposium Human-Computer Interaction for Work* (ACM 2025) https://doi.org/10.1145/3729176.3729192.

International evidence presents similar findings. Dhungel and Heine (2024), interviewing 20 German judges, found cautious openness to AI as a support tool rather than an autonomous decision-maker.[93] The judges' main concerns were that AI might erode the distinctly human features of adjudication, especially face-to-face interaction, contextual understanding, and the judge's personal responsibility for deciding the case. In particular, a primary concern about the 'black box' nature of AI tools was viewed not merely as a usability problem but as an issue with the core duty of transparency. Judges argued that the obligation to provide reasons for decisions, and the right of parties to understand and challenge those reasons, placed constitutional constraints on the use of opaque AI tools in justice.

Fine and others (2023) provide the largest judicial sample (n=131 US judges), finding consistent emphasis on the advisory, rather than determinative, role of AI tools; frustration with proprietary opacity; and insistence on the retention of human discretion.[94] Similar to existing findings, this study shows predominately negative judicial sentiment towards AI as a bias-reduction tool: 64% of judges did not believe AI held promise for removing bias. Judges expressed concern about courtroom AI due to biased data, loss of judicial discretion, the need for human element in decision-making, threats to voice and individualised justice, and fears that AI could weaken legitimacy.

Liu and Li (2024) present a field study on Shenzhen courts, which is the first globally to routinely use LLMs for judicial reasoning generation across almost all civil case types.[95] The authors describe a stable interaction pattern: the judge makes an initial decision, the LLM generates draft reasoning, and the judge revises. The authors argue that this pattern is likely to remain stable regardless of AI advances because judicial accountability rests with the human judge. Judges report the LLM tool is valuable for moderately complex cases but less useful for either very simple or very complex cases. One judge explained, "For simple cases, there is no need to do anything on the system, as judges' documents (templates of judgments) are already well written. For overly complex cases, AI is not yet smart enough at this point, and judges also find it not very useful".[96] While the AI system used in Shenzhen uses fine-turning and RAG to attempt to limit the LLM to known legal references, the authors acknowledge that hallucination issues persist and it poses a significant risk to judicial decision-making. They also identify a possible "echoes of bias" dynamic in which AI reasoning selectively supports judges' initial decisions, potentially reinforcing errors while reducing the corrective function of judges writing their own reasons. The transferability of the Shenzhen study to UK judicial settings is limited; the institutional and procedural context differs substantially. It is unlikely that a similar approach will be adopted in the UK, although it provides an insight into judicial adoption of AI for legal reasoning.

6.3.2 Legal professional attitudes and adoption

Two UK surveys provide methodologically strong evidence on solicitor attitudes towards technology adoption; however, both studies predate the rapid expansion of generative AI since 2023.

93 Anna-Katharina Dhungel and Moreen Heine, 'Cui Bono? Judicial Decision-Making in the Era of AI: A Qualitative Study on the Expectations of Judges in Germany' (2024) 33 Zeitschrift fur Technikfolgenabschatzung in Theorie und Praxis / Journal for Technology Assessment in Theory and Practice 14.
94 Anna Fine, Stephanie Le and Monica Miller, 'Content Analysis of Judges' Sentiments Toward Artificial Intelligence Risk Assessment Tools' (2023) 24 Criminology, Criminal Justice, Law and Society 31.
95 John Zhuang Liu and Xueyao Li, 'How Do Judges Use Large Language Models? Evidence from Shenzhen' (2024) 16 Journal of Legal Analysis 235.
96 ibid 240.

Hodgkinson and others (2023) conducted a random-sample survey of 656 solicitors in England and Wales.[97] The study found that regular usage of 'lawtech' tools was modest, with only legal databases showing substantial adoption (28.2% using "to a large extent"). Many respondents expressed mistrust or indifference towards technology and low confidence to experiment; although over 50% intended to use lawtech more over the following five years. Sako and others (2021) provide complementary survey evidence and document a significant gap in AI adoption between large commercial firms and legal services directed at individuals with everyday legal problems.[98] Only 3.2% of UK lawtech funding reached what the authors term "PeopleLaw" services, as distinct from "BigLaw". Planned adoption of lawtech was shifting towards consumer-facing technologies: portals (21.2% planning vs 15.4% using), interactive document-generation websites (19.5% vs 9.9%), and chatbots or virtual assistants (14.0% vs 6.2%). Rogers and others (2023) also contribute a mixed-methods analysis of how technology is (or is not) changing law firms, finding that the evidence favours augmentation over substitution: lawyers are learning to work alongside AI rather than being displaced by it.[99]

These adoption numbers and attitude perceptions are low by comparison to more recent findings. Recently the LexisNexis *AI Culture Clash* report (August 2025), which surveyed over 700 UK legal professionals, found that 61% of lawyers were using AI for work purposes, a sharp increase even from 46% in January 2025.[100] Among those using AI, 51% had adopted tools designed specifically for legal work, with the figure rising to 58% in private practice and 70% at medium-sized firms. A majority of lawyers using AI reported spending the time saved on increasing billable work (56%) or enjoying better work–life balance (53%). However, 77% remained concerned about relying on inaccurate or fabricated information, followed by concerns about confidential data leakage (47%) and becoming too dependent on AI (47%). The LexisNexis report is a vendor survey with unpublished methodology; the figures should be read with that caveat in mind, although the trajectory of rising adoption is consistent across the available sources.

The combination of these findings suggests a profession that has moved rapidly from cautious experimentation to widespread use. The concentration of purpose-built legal AI tools in medium-sized and large firms, combined with the Sako and others' finding that only 3.2% of UK lawtech funding flows to "PeopleLaw", raises a continuing concern that the benefits of legal AI may accrue disproportionately to well-resourced segments of the profession and their clients. The LexisNexis findings on accuracy concerns are consistent with hallucination error evidence reviewed in Section 7.3 and suggest widespread adoption has not been accompanied by widespread confidence.

In a Czech experimental survey, Harašta and others (2024) reveal a clear preference for legal documents perceived as human-drafted over those believed to be generated by AI.[101] Czech lawyers and law students (n=75) evaluated two human-drafted versions of a legal document, one brief and one more detailed or verbose, each presented under either an "AI-generated" or "human-drafted"

97 Gerard Hodgkinson and others, 'Attitudes towards Lawtech Adoption: Findings from a Survey of Solicitors in England and Wales' (The Law Society, June 2023).

98 Mari Sako and others, 'Technology and Innovation in Legal Services: Final Report for the Solicitors Regulation Authority' (2021) https://www.sra.org.uk/globalassets/documents/sra/research/full-report-technology-and-innovation-in-legal-services.pdf accessed 6 April 2026.

99 Ian Rodgers, John Armour and Mari Sako, 'How Technology Is (or Is Not) Transforming Law Firms' (2023) 19 Annual Review of Law and Social Science 299.

100 LexisNexis, 'The AI Culture Clash: AI Adoption Has Now Outpaced the Slow-Moving Corporate Culture of UK Law' (September 2025) https://www.lexisnexis.co.uk/insights/the-ai-culture-clash/index.html accessed 6 April 2026.

101 Jakub Harašta, Tereza Novotná and Jaromir Šavelka, 'It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human' (2024) 34 Artificial Intelligence and Law 153.

label. Using a between-subjects design, one group saw the brief document labelled "AI-generated" and the detailed document labelled "human-drafted"; the other group saw the labels reversed. Participants were unaware that both documents were entirely human-drafted and were asked to score each document on correctness and language quality (1 to 5). Documents labelled as human-drafted received significantly higher mean scores on both dimensions: correctness (4.69 versus 4.21) and language quality (4.55 versus 3.97). In direct pairwise comparisons, participants preferred the human-labelled document on correctness 35 times against 7 for the AI-labelled equivalent; on language quality, 43 times against 6. These differences are statistically significant. The same respondents believed (93%) that full automation of legal document production would become possible. The data indicate a tension between a present-tense reservation about AI outputs and a future-tense expectation of AI capability that may shape how professionals integrate or resist AI tools during a transitional period.

Several international studies illuminate adoption dynamics with potential UK relevance.

Two studies by Cheong and others address different facets of the adoption question. In a 2024 study, Cheong and others convened seven workshops with 20 US legal experts, where the experts evaluated simulated LLM response strategies to 33 realistic lay legal queries spanning family, criminal, housing, and employment law.[102] The response strategies ranged from outright refusals to legal action recommendations and the study identified 25 evaluative dimensions across four categories: user attributes and behaviours, query nature, AI capabilities, and social impacts. The legal experts favoured providing information-focused responses rather than offering definitive judgements on particular situations. Expert concerns centred on accuracy, hallucinations, confidentiality gaps, accountability voids, bias, and the risk that LLM outputs cross into unauthorised practice of law. The small sample and US focus limits generalisability, but the structured identification of 25 evaluative dimensions is itself a useful contribution, providing a framework against which future tools could be assessed.

In 2025, Cheong and others then conducted a qualitative interview study of 14 US public defence practitioners, addressing the specific constraints on AI adoption in under-resourced criminal-defence settings.[103] The study identifies four adoption barriers: cost (particularly acute for self-employed attorneys and under-resourced offices), restrictive office norms and court orders, confidentiality concerns (including risks associated with third-party vendors and protective orders), and unsatisfactory output quality (hallucinations, superficial reasoning, and subtle errors). Public defenders repeatedly stressed the necessity of mandatory human verification, the dangers of over-reliance, and the preservation of relational lawyering. The implication is that professional confidence in AI is contingent on systems being subordinate to professional judgement rather than substituting for it.

Chien and Kim (2025) explore the relationship between professional adoption and structural incentives, in a study with particularly clear implications for UK legal aid settings.[104] The study

102 Inyoung Cheong and others, '(A)I Am Not a Lawyer, But?: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice' *ACM Conference on Fairness, Accountability, and Transparency, FAccT* (ACM 2024) https://doi.org/10.1145/3630106.3659048.

103 Inyoung Cheong and others, 'How Can AI Augment Access to Justice? Public Defenders' Perspectives on AI Adoption' *ACM Conference on Fairness, Accountability, and Transparency, FAccT* (ACM 2026) https://doi.org/10.48550/arXiv.2510.22933.

104 Colleen Chien and Miriam Kim, 'Generative AI and Legal Aid: Results from a Field Study and 100 Use Cases to Bridge the Access to Justice Gap' (2025) 57 Loyola of Los Angeles Law Review 903.

combines a baseline survey of 202 US legal aid professionals with a 91-person field pilot examining how generative AI tools are actually used in legal aid work, and whether structured support changes uptake and outcomes. At baseline, only 21% of respondents were already using generative AI, with a marked gender gap: 47% of men reported current use compared with 17% of women. In the pilot, exposure to paid tools produced broadly positive but cautious results: 86% reported increased use during the pilot, 90% reported some productivity gain, 25% reported a medium or significant productivity increase, 63% reported greater satisfaction, and 75% intended to continue using generative AI tools. A significant finding, however, is that over 70% of participants were as concerned or more concerned about the technology after the pilot than before. The most analytically important result concerns the effect of structured support. Participants who received "concierge" services (office hours, tailored training, weekly use-case emails, and curated materials) had better outcomes on every measure except comparative usage, with statistically significant improvements on five of nine outcomes. Three of those differences remained significant after adjusting for prior use, suggesting that structured support materially improves adoption and perceived value rather than mere access to tools alone. The pre-pilot gender gap in adoption was closed by the end of the pilot, with no significant gender differences in outcomes, satisfaction, productivity, or intentions.

The findings from both studies have direct implications for the design of AI implementation programmes in UK legal aid settings, where similar resource constraints, demographic variation, and training gaps are likely to apply. It also supports the broader proposition that the conditions for effective use of AI in justice settings include training, oversight design, and institutional support.

6.4 Open justice and scrutiny

The review identified a small subset of UK contributions that connect principles of open justice to AI deployment. In theoretical contributions, Byrom (2024) and Aidinlis and others (2024) argue, from different angles, that asymmetries in access-to-justice data shape who can develop AI tools and for whose benefit, with Byrom contending that unless disparities in data access are addressed, computational legal tools risk deepening "existing inequalities of arms": those with access to comprehensive legal data, typically commercial actors and large law firms, will produce more capable tools than those serving access-to-justice purposes.[105] The argument has structural implications for the development of legal AI in the UK, where commercial database access is concentrated.

As discussed in Section 5.5, Blackham (2021) explains the "online publication of ET [Employment Tribunal] decisions potentially heralds a new era of accessibility and openness in tribunal decision-making. However, this new system is only as good as the data that is inputted into it. Many cases in this particular sample did not have full written reasons attached to the online record; ETs often deliver their decisions verbally, and written reasons are not always requested by the parties."[106]

The most direct empirical evidence on the extent of the access gap comes from Somers-Joce and others (2022), who compared coverage of Administrative Court judgments between the free-access platform BAILII and the commercial database vLex Justis across 2015 to 2020. Only 55% of the

105 Byrom (n 26); Stergios Aidinlis and others, 'Lawful Grounds to Share Justice Data for Lawtech Innovation in the UK' (2024) 140 Law Quarterly Review 544.
106 Blackham (n 79) 408.

5,408 judgments in their dataset were available on BAILII; the remaining 45% were accessible only through commercial subscription.[107]

Conroy and others (2022) describe the development of Find Case Law, a platform built by The National Archives for the publication of UK judgments. The authors argue that integrating existing machine learning research on summarisation, rhetorical role labelling, and entity recognition into the platform could materially improve accessibility for non-expert users. Since these papers, more UK legal data is available for computational research through the Cambridge Law Corpus,[108] and considerable expansion of access to The National Archives Find Case Law.[109]

6.5 Summary of findings

The evidence reviewed suggests that AI use in justice processes may undermine procedural legitimacy, particularly where users perceive a loss of human judgment or where transparency about AI use is limited. Even limited assistive applications are perceived in vignette experiments by the UK public as reducing fairness, dignity, transparency, and the perceived capacity for informed decision-making. Public perceptions are variable across settings, including type of task, area of law, and degree of automation; perceptions may also be conditioned by views on the status quo system. The presence (not merely the availability) of human oversight is central to perceptions of procedural fairness. Public trust in AI is sensitive to process, not only to outcomes.

In the surveyed practitioner populations, judges and lawyers express greater scepticism than lay users. The evidence captures principled professional concerns about accountability, transparency, deskilling, and the calibration of human oversight that are widely articulated in the practitioner literature. Both public and practitioner perceptions are more open to AI in advisory or supporting deployments than to AI in determinative or autonomous roles. The evidential picture should be read alongside the strategic direction now publicly set by the senior judiciary in England and Wales, which is more openly supportive of AI adoption in the judicial role. The disjoint between the evidence-based attitudinal picture and the strategic posture of the senior judiciary is itself a feature of the present moment, and is discussed in Section 8.

Accessible legal data can improve public understanding of UK law and create conditions for downstream research and AI development. Given the dominance of commercial legal technology products and the significant asymmetry in coverage between free-access and commercial legal databases, the further development of open data and accessible tools is one means by which existing inequalities in access to legal information may be reduced.

Attitudinal research provides valuable early insights into the societal factors informing conditions of public and professional acceptance or resistance. However, hypothetical studies reveal contradictory findings. It is possible that real-world experience will moderate attitudes in either direction. For example, early research shows that increased use and familiarity might increase trust in AI tools, or it might deepen concerns.

107 Cassie Somers-Joce, Daniel Hoadley and Editha Nemsic, 'How Public Is Public Law? The Current State of Open Access to Administrative Court Judgments' (2022) 27 Judicial Review 95.
108 Andreas Östling and others, 'The Cambridge Law Corpus: A Dataset for Legal AI Research' (2023) 36 A dvances in Neural Information Processing Systems.
109 The National Archives, 'Find Case Law - When You Need Permission' https://caselaw.nationalarchives.gov.uk/when-you-need-permission accessed 8 April 2026.

7 Synthesis of evidence – system efficiency and effectiveness (RQ3)

7.1 Evidential Orientation

This section synthesises the available evidence bearing on Research question 3: what does the evidence reveal about the justice system's capacity to use AI technologies to address legal needs promptly and fairly? It examines effects on timeliness, accuracy, quality, and the reliability of AI for justice.

Evidence relating to this research question constitutes the largest portion of the overall evidence base. Measures of system efficiency, typically assessed through timeliness and narrow task-specific evaluations, are the most amenable to quantification and therefore reported across a wide range of studies and applications. By contrast, evidence on system effectiveness, understood as the capacity to meet legal needs promptly, fairly, and reliably, remains considerably weaker.

7.2 Time efficiency

The question addressed in this section is whether AI tools reduce the time taken for identifiable legal or justice-related tasks. It is a narrow question confined to examination of direct time-saving evidence. A finding that a particular tool saves time on a particular task does not, without further exploration, establish that the tool improves system efficiency or benefits the people whom the system serves. These second-order effects are discussed in Section 7.4.

7.2.1 Courts and tribunals

The HMCTS tool landscape includes seven tools covering transcription, listing, form processing, e-discovery, anonymisation, knowledge assistance, and case management. Two are live, three are at pilot stage, and two at exploratory stage. The SCTS has deployed tools for initial writ data extraction and transcription. Only one of these deployments has any form of evaluation: the TAR in litigation pilot evaluation. No published evaluation assesses how these tools have changed court operations, case processing times, or the experiences of parties appearing before courts and tribunals.

In England and Wales, the closest in-scope assessment is Mulheron's 2020 Third Interim Report on the TAR in litigation Pilot in the Business and Property Courts, which found generally negative practitioner views: 85% reporting it did not save costs, approximately 42% said it made disclosure

less accurate, and approximately 71% felt it increased burdens on the courts.[110] These findings should, however, be treated cautiously, because the report relied on self-selected practitioner responses and did not directly evaluate a single AI tool in isolation. Even so, the report suggests that potential benefits depend not only on the tool itself, but also on judicial familiarity, practitioner uptake, and broader culture change in court practice.

7.2.2 UK government AI pilots

Some UK government pilot evaluations report task-level time savings from AI tools deployed in justice workflows.

The Caddy deployment, also discussed in Section 5.2, reportedly delivered significant time savings, i.AI reports that messages were generated, validated, and returned in approximately 4 minutes, around half the time an adviser would take to type a response of comparable length (~400 words), before accounting for research time.[111] Advisers with access to Caddy were more than twice as likely to report being confident in giving advice compared with those in the control group, and more than one-and-a-half times as likely to report that the client's issue had been resolved. These are self-reported survey outcomes, not independently verified resolution rates, and the underlying trial methodology has not been published in full.

The Home Office's Asylum Case Summarisation (ACS) and Asylum Policy Search (APS) evaluation (2025) reports considerable time savings.[112] The ACS pilot found that the test group reviewed interview transcripts an average of 23 minutes faster than the comparison group, a 32% time saving. Users reported that the summary output served as a refresher and oriented them to the basis of claim, although some noted limited savings where the summary lacked source references within the transcript. The APS pilot found larger savings: an average saving of 37 minutes per case, broken down into approximate savings of 12 minutes at the pre-interview stage and 25 minutes at the decision-writing stage. All time-efficiency data is self-reported data without further evaluations as to overall improved system efficiency.

While no legal evaluation of M365 Copilot's use across HMCTS or MoJ has been published, the Department for Business and Trade (**DBT**) published its evaluation in 2025.[113]

The evaluation combined diary studies (32% response rate, n=300), qualitative interviews (n=19), observed task sessions (n=11), and Microsoft usage data to study UK-based staff's use of M365 Copilot between October and December 2024. Small time savings were observed across most use cases, with written tasks presenting the largest gains. After adjustments for unused outputs and novel tasks (those only performed because M365 Copilot was available), drafting written documents showed a mean saving of 1.3 hours per task, summarising research 0.8 hours, and transcribing or summarising meetings 0.7 hours. Some tasks, however, took longer with M365 Copilot: scheduling added 0.6 hours and generating images added 0.5 hours. Roles with heavy administrative burdens reported the greatest benefits. Individuals in roles requiring high levels of contextual awareness,

110 Rachael Mulheron, 'Disclosure Pilot Monitoring in the Business and Property Courts: Third Interim Report' (2020) https://perma.cc/A9FJ-5VUK.

111 Andreas Varotsis (n 42); AI Knowledge Hub (n 42).

112 Home Office (n 76).

113 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade, 'The Evaluation of the M365 Copilot Pilot in the Department for Business and Trade' (28 August 2025) https://www.gov.uk/government/publications/microsoft-365-copilot-pilot-dbt-evaluation-report.

nuance, and attention to detail – including legal and policy roles – found M365 Copilot less suitable. They experienced minor benefits for simple tasks but reported that it was not particularly useful for their core work. One user noted: "In a work context, especially dealing with legal text, it's important that every word is right".[114]

The evaluation's most significant finding is that the diary study did not find evidence that self-reported time savings translated into demonstrated productivity improvements at the aggregate level. The diary study had a response rate of only 32%, limiting the evidential weight of its findings. Control group participants had not observed productivity improvements from pilot colleagues. Observed task sessions partially corroborated the self-reported data on some tasks (summarising, email drafting) but found inconsistencies on others (data analysis, presentations). The observation that M365 Copilot was least useful for roles requiring contextual awareness and attention to detail should moderate expectations about the capacity of general-purpose AI assistants to improve justice-system performance, where precision in language, legal reasoning, and case-specific judgment are routine requirements.

7.2.3 AI transcription and case recording

AI transcription tools represent one of the most rapidly adopted categories of AI in public services. The HMCTS and MoJ are currently testing the use of AI transcription tools across the justice sector.[115] The MoJ reported a positive pilot of Justice Transcribe in the Probation Service with significant reduction in notetaking time by probation officers in their meetings with offenders.[116] The government has just announced it is "testing transcription in the courts and tribunals based on the same technology".[117] No specific details of the pilot nor evaluations have been reported for such tools.

In the absence of published UK justice-sector evidence, evaluations of AI transcription tools in adjacent public-sector workflows are informative. Such tools are used in social work assessments, care records, and witness statements; the transcribed documents may subsequently be used in legal proceedings, meaning their accuracy and reliability bear directly on the quality of evidence entering the justice system.

In an evaluation of the Anathem Digital Assistant (ADA), which was piloted in the Hertfordshire Constabulary for AI-generated witness statements, Walley and Glasspoole-Bird (2025) found that AI-generated witness statements did not reduce call or report times.[118] Quality-control burdens, necessitated by the tool's weak narrative coherence, hallucinations, and overly formal style, offset any potential time savings. The ADA evaluation demonstrates that AI tools do not invariably produce time savings, and that quality-control requirements can consume or exceed any time nominally saved.

By contrast, a vendor-commissioned evaluation of Beam Magic Notes, which is a tool used across UK local authorities for social care, reports perceived weekly time savings of 6.8–7.2 hours on written

114 ibid 27.
115 Ministry of Justice, 'AI Action Plan for Justice' (n 2); Incubator for AI, 'Justice Transcribe' (*AI.GOV.UK*, 20 January 2026) https://ai.justice.gov.uk/ai-in-action/speech-ai accessed 21 January 2026.
116 ibid.
117 The Rt Hon David Lammy MP (n 7).
118 Paul Walley and Helen Glasspoole-Bird, 'An Evaluation of the Pilot Application of Artificial Intelligence to Witness Statement and Report Generation at Hertfordshire Constabulary' (Centre for Policing Research and Learning 2025).

administration and documentation, a 35–41% reduction.[119] Per-assessment time savings were approximately 2.7–2.8 hours (a 48% reduction for care needs assessments and occupational therapy assessments). Report submission times were reduced by approximately 2.0–3.5 days. These figures are substantial, but the evaluation relies on self-reported data with no comparator group.

The Ada Lovelace Institute conducted a detailed study on AI transcription in social care, particularly instructive on the specific dynamics of AI-based efficiency.[120] The study conducted 39 semi-structured interviews across 17 local authorities in England and Scotland with social workers and senior staff involved in procuring and evaluating AI transcription tools, principally Beam's Magic Notes and Microsoft Copilot. Almost all social workers reported that AI transcription tools brought meaningful benefits, with the most commonly cited being reduced documentation time. Some reported workload reductions of "a few hours"; others described their workloads being cut by "half". However, time savings were not experienced uniformly. Some social workers found that complex documentation formats within their local authority prevented effective integration. Others reported that management expectations increased following the introduction of AI, with assessments becoming longer and more detailed, absorbing the time saved. The study identified that the primary focus of these evaluations is efficiency savings, but it highlights limited evidence connecting self-reported time savings with broader productivity benefits. Additionally, the rapid adoption of AI transcription tools has meant evaluations focus on efficiency rather than on potential systemic impacts of AI on the profession of social care or on the experiences of people who draw on care.

The study's central framing, that AI-created records are not neutral administration but documents that can affect thresholds, court decisions, and individuals' ability to understand and challenge decisions made about them, is a finding with direct implications for how time-saving claims should be interpreted in justice contexts.

7.2.4 Commercial legal AI tools

While generally out of scope for this review, the commercial legal AI market is the most mature segment of the UK legal technology landscape. There are two relevant studies that explore time efficiency for lawyers.

Ashurst's *VoxPopulAI* report (2024) documents three global generative AI trials involving 411 partners, lawyers, and staff across 23 offices in 14 countries between November 2023 and March 2024. The trials combined blind studies, experiments, and feedback surveys.[121] It reports controlled experiment time savings of approximately 80% on UK corporate filings, 59% on sector research from public filings, and 45% on first-draft legal briefings. The tools tested are not disclosed and a detailed methodology was not published.

The RSGI study (2025), commissioned by legal AI company Harvey, surveyed 40 organisations (29 law firms and 11 in-house legal teams, 20% based in the UK) through standardised questionnaires

119 Unity Insights, 'Understanding the Impact of Magic Notes through Validating Evaluation Findings' (2025) https://unityinsights.co.uk/our-insights/understanding-the-impact-of-magic-notes-through-validating-evaluation-findings/ accessed 13 March 2026.
120 Oliver Bruff and Lara Groves, 'Scribe and Prejudice?' (Ada Lovelace Institute 2026) https://www.adalovelaceinstitute.org/report/scribe-and-prejudice/ accessed 24 February 2026.
121 Ashurst, 'Vox PopulAI: Lessons from a Global Law Firm's Exploration of Generative AI' (2024) https://www.ashurst.com/en/insights/vox-populai-lessons-from-a-global-law-firms-exploration-of-generative-ai/ accessed 24 January 2026.

and structured interviews.[122] It reports estimated time savings of 36.9 hours per month for "power users" (approximately 20–30% of the user base) and 15.7 hours per month for standard users in law firms. In-house power users reported 28.3 hours per month. However, the time-savings figures are self-reported estimates rather than independently measured, the study does not include a control group or pre-post comparison, and value measurement across participants relied predominantly on adoption rates (83% of participants) and usage intensity (75%) rather than on outcome measures. The study itself notes that "formal ROI frameworks for AI are still developing across the legal sector" and that cost savings are "still yet to be measured".

Self-reported time-saving metrics provides an indication as to the potential efficiencies of introducing AI to legal processes; however, it does not address whether it translates into actual productivity gains.

7.3 Accuracy, quality, and reliability

This section summarises evidence on the overlapping issues of accuracy, quality, and reliability of AI tools in justice settings.

7.3.1 UK public-sector evaluations

Despite detailed self-report data on time efficiency, reported UK government pilots do not publish accuracy or quality evaluations in equivalent detail.

For the Home Office ACS and APS pilots,[123] a dip-sample of decisions was reviewed using Calibre, the Home Office's quality assurance mechanism. The proportion of sustained decisions was similar across test and comparison groups, suggesting neither tool positively nor negatively affected decision quality. The specific accuracy data is more limited. Of the summaries used, 23% of users reported they were not fully confident in the summary information. For APS, 82% of queries provided sufficient information, and 79% of survey respondents were confident in the accuracy. For the ACS pilot, technical specialists reviewed all summaries prior to use; 9% were deemed inaccurate or had missing information and were removed. Although the reports do not explain the source or extent of these errors. The i.AI reports that 80% of Caddy's responses were judged to be of sufficient quality to pass directly to advisers without revision.[124] Similarly, no details of how these judgments were made or how "sufficient quality" was defined have been published.

The Bruff and Groves (2026) study on AI transcription in social care evidences concrete accuracy risks: hallucinated content (including, in one instance, invented references to suicidal ideation), mis-transcription of speakers with accents or dialects, irrelevant material entering children's files, and AI-generated language that was more formal and clinical than the person-centred language expected in social care documentation.[125]

122 RSGI Limited, 'Defining the Impact of Legal AI: How Harvey Customers Realise Value' (RSGI, 19 November 2025) https://www.harvey.ai/resources/reports/harvey-rsgi-report.
123 Home Office (n 76).
124 AI Knowledge Hub (n 42); Andreas Varotsis (n 42).
125 Bruff and Groves (n 120).

In the DBT's M365 Copilot evaluation, hallucinations were noted but only measured through self-reported identification.[126] When asked to recall hallucinations, 43% of respondents reported they did not identify a hallucination at any point throughout the pilot; while 22% did identify hallucinations, 11% were unsure if they encountered hallucinations, and 22% of respondents chose not to answer this question. Self-reported hallucination identification is a weak measure of actual hallucination rates: it captures hallucinations that users noticed, not hallucinations that occurred. The risk of hallucination is a repeated accuracy concern across LLM deployments and merits dedicated evaluation rather than reliance on user self-report.

7.3.2 LLM hallucination and accuracy benchmarks

In Section 5, several UK studies addressed hallucination risks in LLMs and the implications for lay users. This section turns to non-UK studies on the narrower technical question of how often such systems produce inaccurate, fabricated, or otherwise misleading legal outputs.

Dahl and others (2024) provide the first systematic empirical evidence of legal hallucination in public-facing LLMs, testing ChatGPT-4, ChatGPT-3.5, Google PaLM 2, and Meta Llama 2 across more than 200,000 queries on US law.[127] The study finds that LLMs hallucinate between 58% (GPT-4) and 88% (Llama 2) of the time when asked direct, verifiable questions about randomly selected US federal cases. Three additional findings are particularly significant. First, the models exhibit a tendency to over-represent prominent judges and cases, producing what the authors describe as a "flattening legal nuance and producing a falsely homogenous sense of the legal landscape". Second, LLMs are susceptible to contra-factual bias, in which the model accepts legally false premises and answers as though such premises were true. This is framed as a form of sycophancy: the model defers to the user's implicit assumptions even when those assumptions are wrong. The risk is directly relevant to lay users who may frame legal questions based on misconceptions and false legal premises. Third, the LLMs systematically overstate their confidence relative to their actual accuracy, meaning users are liable to receive hallucinated answers that carry the same apparent conviction as correct ones. Hallucination rates increase with task complexity, are higher for lower courts, vary by jurisdiction, and correlate with case prominence, suggesting that LLMs perform better on cases that appear frequently in training data.

In Magesh and others (2025), the same primary authors conducted an empirical evaluation of commercial AI-driven legal research tools, testing Lexis+ AI (LexisNexis), Westlaw AI-Assisted Research (Thomson Reuters), and Ask Practical Law AI (Thomson Reuters) against 202 hand-crafted legal queries spanning general research questions, jurisdiction-specific questions, false-premise questions, and factual recall.[128] The study directly challenges vendor marketing claims that through sophisticated techniques such as RAG they have mitigated if not entirely solved hallucination risk. All three tools produced hallucinated responses at substantial rates. Lexis+ AI hallucinated on 17% of queries, Ask Practical Law AI on 17%, and Westlaw AI-Assisted Research on 33%.[129] By comparison, GPT-4 in a closed-book setting hallucinated on 69% of queries.

126 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade (n 113).

127 Matthew Dahl and others, 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024) 16 Journal of Legal Analysis 64.

128 Varun Magesh and others, 'Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools' (2025) 22 Journal of Empirical Legal Studies 216.

129 Thomson Reuters disputed the findings of this study and asserted Westlaw's AI-Assisted Research product achieved roughly 90% accuracy, see Artificial Lawyer, 'Thomson Reuters Contradicts Stanford GenAI Study – "We Are 90% Accurate"' (10 June 2024) https://www.artificiallawyer.com/2024/06/10/thomson-reuters-contradicts-stanford-genai-study-we-are-90-accurate/ accessed 2 April 2026.

RAG-based tools thus reduced hallucination relative to general-purpose LLMs but did not eliminate it. Hallucinations persisted across all query types, including general legal research and bar-exam-style questions. Many errors were insidious: the systems struggled with elementary legal comprehension, including describing holdings, distinguishing between litigant arguments and court holdings, and respecting the hierarchy of legal authority. In several instances, tools cited over-ruled cases as good law.

Evidence from several studies indicates that there are methods to mitigate the hallucination rate in LLMs, such as RAG.[130] However, hallucination is an inherent risk of current LLM architecture, even if mitigation methods can reduce its incidence.[131]

Given the inherent hallucination risk of LLMs, whether general-purpose or specific to the legal domain, legal outputs cannot be relied upon to produce consistently accurate legal information without human verification. This constraint has direct implications for system efficiency. As explained by Magesh and others, "lawyers are faced with a difficult choice: either verify by hand each and every proposition and citation produced by these tools (thereby undercutting the efficiency gains that AI is promised to provide), or risk using these tools without full information about their specific risks and benefits (thereby neglecting their core duties of competency and supervision)."[132]

7.3.3 Access-to-justice tools and proof-of-concept evaluations

A distinct cluster of evidence comes from collaborative teams involving academic researchers, civil society organisations and technical teams that have developed and evaluated proof-of-concept legal AI tools. These studies differ from the deployed-tool evaluations above in that they are typically earlier-stage, more technically detailed, and evaluated under controlled conditions rather than in operational settings.

Ribary and others (2025) develop and test the Insolvency Bot, a RAG-based tool for insolvency law in England and Wales, which significantly outperformed each unmodified LLM.[133] The average score for GPT-4 was 21% and for the Insolvency Bot modification was 47%. When augmented with RAG from HMRC-approved sources, the best configuration (GPT-4o with retrieval) achieved 55%. This is a substantial improvement but still represents a system that produces incorrect answers nearly half the time. Case law retrieval remained a weak point: on test questions, the system identified correct cases with only 24% precision and 33% recall, reflecting the difficulty of matching lay queries to legal texts. These results demonstrate that domain-specific RAG with curated legal knowledge bases materially improves performance over general-purpose models in specialist legal domains, although even the best-performing configuration falls short of reliable legal guidance.

130 Marton Ribary and others, 'A Generative AI-Based Legal Advice Tool for Small Businesses in Distress' (2025) 12 Journal of International and Comparative Law 199; Jonathan Li and others, 'Experimenting with Legal AI Solutions: The Case of Question-Answering for Access to Justice' (arXiv, 27 July 2024) https://doi.org/10.48550/arXiv.2409.07713 Bakht Munir and others, 'Evaluating AI in Legal Operations: A Comparative Analysis of Accuracy, Completeness, and Hallucinations in ChatGPT-4, Copilot, DeepSeek, Lexis+ AI, and Llama 3' (2025) 53 International Journal of Legal Information 103; Abhishek Purushothama and others, 'Not Ready for the Bench: LLM Legal Interpretation Is Unstable and out of Step with Human Judgments' (arXiv, 29 October 2025) https://doi.org/10.48550/arXiv.2510.25356.

131 Bender and others (n 35); Dumit and Roepstorff (n 35).

132 Magesh and others (n 128) 232.

133 Ribary and others (n 130).

Designed by academics in Northern Ireland, Aura is an AI-enabled legal assistant trained on UK legal corpora including statutes, case law, and SRA practice rules.[134] A working paper reports over 99% efficiency in document classification and a 93% reduction in time for initial case assessments compared with traditional methods. However, these figures are reported by the tool's developers, and the evaluation methodology, sample sizes, and conditions under which these metrics were achieved are not described in sufficient detail for independent assessment. Qualitative user feedback from legal and community partners reported improvements in accessibility, consistency, and reduced administrative burden.

Saadany and others (2025), in collaboration with Just Access, have developed a tool for linking judgement text to court hearing video recordings and improving translation quality for UK Supreme Court hearing records.[135] The custom ASR system using a custom language model trained on 139 hours of edited Supreme Court transcripts and legal texts, plus NLP-extracted domain-specific vocabulary, achieved 9- and 8-percentage-point word-error-rate improvements over AWS and Whisper baselines respectively, with better handling of legal entities. Using the tool's custom interface significantly boosted legal expert productivity from 15 hours to validate 10 links to 3 hours for 220 links. One important observation of this work is the need for domain-specific work for UK legal settings.

The authors identified consistent errors from general-domain ASR with the unique pronunciations and linguistic etiquette of UK court hearings, and the specific legal terminology in the hearings (that is, instead of "my lady" the AWS ASR identified "melody", or instead of "all rise" the it identified "all right").[136] This tool also addresses a specific dimension of open justice, making Supreme Court proceedings more accessible by connecting written judgments to the spoken proceedings from which they derive, and suggests that AI can reduce barriers to public engagement with the courts.

134 Chaudhary Hamza Riaz and Muhammad Usman Hadi, 'Aura: An AI-Powered Legal Assistant for Enhancing Access to Justice in the UK Legal System' (Preprints, 27 August 2025) https://doi.org/10.36227/techrxiv.175632364.48017327/v1.

135 Hadeel Saadany and others, 'Employing AI for Better Access to Justice: An Automatic Text-to-Video Linking Tool for UK Supreme Court Hearings' (2025) 15 Applied Sciences; see earlier versions Hadeel Saadany and others, 'Linking Judgement Text to Court Hearing Videos: UK Supreme Court as a Case Study' *Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (ELRA 2024) https://aclanthology.org/2024.lrec-main.927/; Hadeel Saadany and others, 'Better Transcription of UK Supreme Court Hearings' *Workshop on Artificial Intelligence for Access to Justice (AI4AJ 2023)* (CEUR 2023) https://doi.org/10.48550/arXiv.2211.17094.

136 Saadany and others, 'Employing AI for Better Access to Justice: An Automatic Text-to-Video Linking Tool for UK Supreme Court Hearings' (n 135) 3.

Several international proof-of-concept tools and experiments have shown that legal triage and intake in legal aid and legal clinic settings can be improved with AI applications.[137] These studies show strong performance on smaller, transformer-based models, with one showing smaller models can outperform enterprise LLMs.[138]

7.3.4 Task-specific legal NLP research

A large body of academic literature in the field of legal NLP was identified in this review. Much of that literature was, however, excluded from detailed analysis because its findings have limited transferability to the questions examined here. In most cases, these studies assess narrow technical tasks on curated datasets, rather than the system-level question of whether AI tools improve the efficiency, quality, fairness, or accessibility of justice processes in operational settings.

The UK-specific technical evidence base is itself relatively modest and is concentrated mainly in benchmark studies of bounded legal-text tasks rather than operational evaluations. For example, fine-tuned or domain-adapted models, and often open-source or open-weight models, perform well on legal judgment prediction, citation detection in UK judgments, topic classification of summary judgment cases, Employment Tribunal outcome-prediction benchmarks, information extraction from Planning Court decisions, named entity recognition across Criminal Appeal judgments,

137 Daniel Rincón-Riveros and others, 'Automation System Based on NLP for Legal Clinic Assistance' *IFAC-PapersOnLine* (Elsevier BV 2021) https://doi.org/10.1016/j.ifacol.2021.10.460; Claire Barale, Michael Rovatsos and Nehal Bhuta, 'Automated Refugee Case Analysis: An NLP Pipeline for Supporting Legal Practitioners' *Proc. Annu. Meet. Assoc. Comput Linguist.* (ACL 2023) https://doi.org/10.18653/v1/2023.findings-acl.187; Quinten Steenhuis and Hannes Westermann, 'Getting in the Door: Streamlining Intake in Civil Legal Services with Large Language Models' *Front. Artif. Intell. Appl.* (IOS Press BV 2024) https://doi.org/10.3233/FAIA241242; Quinten Steenhuis, 'That's So FETCH: Fashioning Ensemble Techniques for LLM Classification in Civil Legal Intake and Referral' in Réka Markovich and others (eds), *Frontiers in Artificial Intelligence and Applications* (IOS Press 2025).

138 Steenhuis (n 137).

and studies on jurisdictional variations across UK statutes.[139] Such work suggests that AI systems can achieve technically competent performance on some narrow tasks under controlled conditions, particularly where models are fine-tuned or adapted to legal materials. This line of research indicates what current systems may be able to do on structured sub-tasks, while offering much less purchase on whether those capabilities translate into reliable gains in real justice settings.

Without attempting to review the wider legal NLP literature in full, a small number of themes appear consistently across international studies that merit noting. First, legal domain-specific and fine-tuned models generally outperform general-purpose models on legal tasks, although the degree of improvement varies by task and dataset.[140] Across recent legal benchmarking efforts, proprietary or enterprise LLMs often outperform open-weight models, although the size of the gap varies across tasks, datasets, and specific model releases.[141]

139 Benjamin Strickson and Beatriz De La Iglesia, 'Legal Judgement Prediction for UK Courts' *Proceedings of the 3rd International Conference on Information Science and Systems* (ACM 2020) https://doi.org/10.1145/3388176.3388183; Drish Mali, Rubash Mali and Claire Barale, 'Information Extraction for Planning Court Cases' *Proceedings of the Natural Legal Language Processing Workshop 2024* (ACL 2024) https://doi.org/10.18653/v1/2024.nllp-1.8; Huiyuan Xie and others, 'The CLC-UKET Dataset: Benchmarking Case Outcome Prediction for the UK Employment Tribunal' *Proceedings of the Natural Legal Language Processing Workshop 2024* (ACL 2024) https://doi.org/10.18653/v1/2024.nllp-1.7; Ahmed Izzidien, Holli Sargeant and Felix Steffek, 'LLM vs. Lawyers: Identifying a Subset of Summary Judgments in a Large UK Case Law Dataset' (arXiv, 4 March 2024) https://doi.org/10.48550/arXiv.2403.04791; Holli Sargeant, Ahmed Izzidien and Felix Steffek, 'Topic Classification of Case Law Using a Large Language Model and a New Taxonomy for UK Law: AI Insights into Summary Judgment' [2025] Artificial Intelligence and Law; Holli Sargeant, Andreas Östling and Måns Magnusson, 'Detecting Legal Citations in United Kingdom Court Judgments' *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (ACL 2025) https://doi.org/10.18653/v1/2025.emnlp-main.1361; Safia Kanwal and others, 'Can Large Language Models Reliably Extract Jurisdictional Variations? An Empirical Study on UK Statutory Texts' *2025 International Conference on Frontiers of Information Technology (FIT)* (2025) https://doi.org/10.1109/FIT67061.2025.11333744; Safia Kanwal and others, 'Leveraging LLMs and LegalDocML to Extract Legal Interpretations: A Case Study on UK Legislation and Case Law' (2025) https://cronfa.swansea.ac.uk/Record/cronfa69629; Waleed Abbas and others, 'The Application of Pre-Trained Transformer Models to UK Court of Appeal Legal Judgments' in Thanh Tho Quan and others (eds), *Multi-disciplinary Trends in Artificial Intelligence* (Springer Nature 2026) https://doi.org/10.1007/978-981-95-4963-4_21.

140 Mali, Mali and Barale (n 139); Li and others (n 130); Sargeant, Östling and Magnusson (n 139); Munir and others (n 130); Abbas and others (n 139); Oren Gazal Ayal, Zohar Elyoseph and Adir Solomon, 'Evaluating Large Language Models as Judicial Decision-Makers' (2026) 0 Justice Quarterly 1.

141 See summary comparisons compiled by Vals AI, 'LegalBench: Evaluating Language Models on a Wide Range of Open Source Legal Reasoning Tasks' (7 April 2026) https://www.vals.ai/benchmarks/legal_bench accessed 8 April 2026; Neel Guha, 'Legal LLM Leaderboard' (6 April 2026) https://neelguha.github.io/legal-benchmark-dashboard/ accessed 8 April 2026.

Second, performance is usually highest on well-bounded and structurally regular tasks, and tends to degrade as problems become more legally complex, factually messy, or procedurally unusual.[142] Third, where LLMs are used, retrieval and other design interventions may reduce error rates, but they do not remove hallucination or interpretive instability altogether.[143] Fourth, the relative explainability of certain model types – especially symbolic, factor-based, or otherwise interpretable approaches – may carry particular weight in legal contexts where the ability to inspect, challenge, and justify a result is itself normatively significant.[144]

Narrow NLP research, while not necessarily informative of real-world applications, provides important 'building block' and infrastructural research that still significantly informs the developments and evaluations of legal AI.

7.4 Implementation burden

This section expands the above findings to explore three questions relating to the real-world deployment of AI tools in justice settings. First, whether task-level time savings translate into organisational productivity gains. Second, whether the human oversight required to verify AI outputs creates hidden costs that offset nominal efficiency improvements. Third, whether the benefits of AI tools are evenly distributed across the skill and resource distribution of users, or whether they accrue disproportionately to some groups at the expense of others.

142 Nikolaos Aletras and others, 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective' (2016) 2 PeerJ Computer Science e93; Jason Lam and others, 'The Gap between Deep Learning and Law: Predicting Employment Notice' *CEUR Workshop Proc.* (CEUR 2020) https://www.scopus.com/pages/publications/85090899252; Rohit Pande and Shafiq Alam, 'Predicting the Outcome of Judicial Cases Using Semantic Analysis' *IEEE Symp. Ser. Comput. Intell., SSCI* (IEEE 2020) https://doi.org/10.1109/SSCI47803.2020.9308506; I Almuslim and D Inkpen, 'Legal Judgment Prediction for Canadian Appeal Cases' *Proc. - Int. Conf. Data Sci. Mach. Learn. Appl., CDMA* (IEEE Inc 2022) https://doi.org/10.1109/CDMA54072.2022.00032; DM Katz and others, 'GPT-4 Passes the Bar Exam' (2024) 382 Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences; Sophia Simeng Han and others, 'CourtReasoner: Can LLM Agents Reason Like Judges?' *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (ACL 2025) https://doi.org/10.18653/v1/2025.emnlp-main.1787; Lucia Zheng and others, 'A Reasoning-Focused Legal Retrieval Benchmark' *Proceedings of the 2025 Symposium on Computer Science and Law* (ACM 2025) https://doi.org/10.1145/3709025.3712219; Yu Fan and others, 'LEXam: Benchmarking Legal Reasoning on 340 Law Exams' *International Conference on Learning Representations* (2026) https://lexam-benchmark.github.io/.

143 Li and others (n 130); Ribary and others (n 130); Munir and others (n 130); Dahl and others (n 127); Magesh and others (n 128).

144 Karl Branting and others, 'Semi-Supervised Methods for Explainable Legal Prediction' *Proc. Int. Conf. Artif. Intell. Law, ICAIL* (ACM 2019) https://doi.org/10.1145/3322640.3326723; Karl Branting and others, 'Scalable and Explainable Legal Prediction' (2021) 29 Artificial Intelligence and Law 213; Westermann and others, 'Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice' (n 51); H Rodger, A Lensen and M Betkier, 'Explainable Artificial Intelligence for Assault Sentence Prediction in New Zealand' (2023) 53 Journal of the Royal Society of New Zealand 133; Joe Collenette, Katie Atkinson and Trevor Bench-Capon, 'Explainable AI Tools for Legal Reasoning about Cases: A Study on the European Court of Human Rights' (2023) 317 Artificial Intelligence 103861; Mohammad Mohammadi, Martijn Wieling and Michel Vols, 'An Explainable Approach to Detect Case Law on Housing and Eviction Issues within the HUDOC Database' (arXiv, 3 October 2024) https://doi.org/10.48550/arXiv.2410.02978.

7.4.1 Productivity costs and challenges

One specific implementation cost is intrinsic to the use of AI tools. Where AI outputs require correction, the time spent correcting them may directly offset the time that was nominally saved. For example, in Walley and Glasspoole-Bird (2025), the need to verify AI-generated witness statements for accuracy, completeness, and the absence of hallucinated content offset the nominal time savings from AI-assisted statement drafting.[145] Additionally, the Magesh and others (2025) finding that lawyers must either verify by hand each proposition and citation produced by AI tools, or risk using the tools without full information about their specific risks, points to the same dynamic in legal research.[146] Given the evidence profile on accuracy and hallucination risk, rework costs of integrating AI into existing workflows may be considerable.

Absorption of efficiency gains

Efficiency gains from AI may be absorbed by institutional responses, including raised throughput expectations and increased caseloads, rather than redirected towards improving the quality of outcomes or the experiences of those interacting with the system. Bruff and Groves (2026) document this dynamic in social care transcription: even where time savings are real, when management raises caseload targets in response to observed efficiency gains, the result from the practitioner's perspective is not greater efficiency but greater throughput at the same or higher level of intensity.[147] This is a dynamic about institutional behaviour rather than about AI, but it operates wherever efficiency-generating tools are introduced into stretched workforces, and it affects whether efficiency gains translate into the benefits that procurement decisions are typically intended to produce.

The *HMCTS Reform Evaluation Thematic Report: Digitalisation* (2026) provides contextual evidence on the institutional conditions for innovation in the justice system.[148] The report identified an unintended consequence: the move to self-service increased the administrative burden on the judiciary, contrary to programme expectations. Where services were not fully digitalised or were ineffectively implemented, there was an increase in workarounds and administrative tasks outside the digitalised systems, which increased the administrative burden on the judiciary. Limited support was found for the proposition that self-service functionality optimises judicial time. The gap between anticipated and realised efficiency gains in the HMCTS digitalisation reform programme is a cautionary precedent for expectations about AI-driven efficiency

Measurement challenges in moving from task-level efficiency to organisational productivity

Identifying this efficiency and productivity gap is a challenge in itself. The strongest UK evidence for the gap between task-level efficiency and organisational productivity comes from the DBT M365 Copilot evaluation, where self-reported task-level time savings did not translate into demonstrated

145 Walley and Glasspoole-Bird (n 118).
146 Magesh and others (n 128).
147 Oliver Bruff and Lara Groves, 'What Is an AI Transcription Tool?' (*Ada Lovelace Institute*, 5 February 2026) https://www.adalovelaceinstitute.org/resource/what-is-an-ai-transcription-tool/ accessed 24 February 2026.
148 Pilling and others (n 74).

department-level productivity gains.[149] Several possible explanations operate jointly: as mentioned, the absorption of saved time into additional tasks and the rework costs of correcting substandard AI outputs; and separately is the challenge of aggregating task-level and individually reported time savings into measurable organisational improvements. Therefore, gains may diminish, be redistributed, or be offset by costs not captured in the initial efficiency measurement. That is not to dismiss the opportunity for time savings but confidence in the scale of efficiency and productivity gains should be calibrated.

7.4.2 Human-AI collaboration and interaction

Despite the acknowledgement that human review is necessary for AI tools, there has not been a commensurate evaluation of over-reliance. For example, in the Home Office ACS and APS pilot, the risk of decision-makers becoming over-reliant on the tools was explored through interviews.[150] No ACS interviewees felt the summary had influenced their decision-making or had the potential to do so, as it did not replace any stage of the process. Similarly, interviews in the APS pilot did not suggest over-reliance. However, the evaluation notes that these findings reflect self-report, and that ongoing monitoring would be required to detect over-reliance in a full rollout.

Recent studies explore the dynamics of human-AI collaboration in more detail.

Alcaras and Ricci (2025) introduce "configuration work" as the category of hidden labour required to adapt LLMs to specific operational contexts: designing and iterating prompts, training users, auditing outputs, managing exceptions, and resolving failures that systems are not designed to handle.[151] This labour is real, time-consuming, and largely invisible in the task-completion metrics used to evaluate AI tools. It accrues at the organisational level rather than the individual task level, and it typically falls on a small number of engaged users rather than being evenly distributed across a workforce. In particular, participants identified that legal tasks "didn't translate well to the kind of short-term, local, iterative problems LLMs seemed able to tackle". For example, one participant reflected: "I think it shows that law, even when it looks like a simple legal task, always contains hidden complexity".

Nielsen and others (2024) show that different types of AI assistance can change efficiency gains.[152] In an experiment comparing AI-assisted highlighting and AI-assisted summarising with US law students (n=206) on completing private law tasks using a Thomson Reuters commercial model, AI-generated highlighting reduced task completion time by 30% without any reduction in measured quality indicators compared to no AI assistance. AI-generated summaries produced no change in performance metrics. Therefore, different human-AI workflows can produce different effects, and design choices about which form of AI assistance to deploy matter at least as much as the underlying model.

149 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade (n 113).
150 Home Office (n 76).
151 Gabriel Alcaras and Donato Ricci, 'Configuration Work: Four Consequences of LLMs-in-Use' (2025) https://hal.science/hal-05421828.
152 Aileen Nielsen and others, 'Building a Better Lawyer: Experimental Evidence That Artificial Intelligence Can Increase Legal Work Efficiency' (2024) 21 Journal of Empirical Legal Studies 979.

Zeng and others' (2025) study examined how Chinese attorneys (n=24) interact with commercial AI contract review tools, finding significant mismatches between how tools present recommendations and how attorneys conduct reviews.[153] When trust-calibration design features were introduced – showing AI confidence levels, comparing against templates, and receiving interactive outputs with recommendations rather than receiving static outputs – attorneys made more accurate review decisions. The study highlights that the usability and design of AI interfaces, not only their underlying accuracy, but shapes also whether AI assistance improves or degrades professional work.

7.4.3 Distributional effects on lower-performing users

Schwarcz and others (2026) conducted the first randomised controlled trial assessing next-generation AI tools in legal tasks,[154] assigning 137 upper-level law students to complete six realistic legal tasks using an LLM-and-RAG-powered tool (Vincent AI), an LLM reasoning model (OpenAI's o1-preview), or no AI. Both tools produced statistically significant productivity gains of 50% to 130% on five of six tasks, with the largest effects on tasks requiring structured reasoning and written advocacy. Both tools also improved work quality on at least four of six tasks, a marked contrast with prior research examining models such as GPT-4, which found speed gains but null effects on quality. Qualitatively, AI-assisted work was more polished and structurally coherent, with fewer surface errors. AI access did, however, lead participants to oversimplify legal questions in some tasks, and o1-preview users occasionally cited fabricated cases. The study found that AI disproportionately benefited lower-performing participants, raising the question of whether the human-AI interaction itself may introduce performance costs for more skilled users.

Choi and Schwarcz's (2025) study of AI-assisted legal analysis for US law students found a similar distributional pattern.[155] For multiple-choice questions, AI assistance improved average performance by 29 percentile points. For essays, no significant average improvement was found. The lowest-performing students gained approximately 45 percentile points; top students declined by approximately 20 percentile points. GPT-4 alone, with grounded prompting, outperformed both humans and AI-assisted humans.

These findings demonstrate an "equalising" or "raising the floor" effect observed in earlier non-legal studies, where less experienced and lower-skilled workers improve in both speed and quality of output with AI while the highest-skilled workers see small gains in speed and small declines in quality.[156] The performance-levelling finding across these studies raises a normative tension that is directly relevant to the justice system. On one reading, AI as a professional equaliser could democratise competence, extending the effective quality of legal work performed by less

153 Jian Zeng and others, 'ContractMind: Trust-Calibration Interaction Design for AI Contract Review Tools' (2025) 196 International Journal of Human-Computer Studies 103411.

154 Daniel Schwarcz and others, 'AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice' [2026] Journal of Law and Empirical Analysis.

155 Johnathan Choi and Daniel Schwarcz, 'AI Assistance in Legal Analysis: An Empirical Study' (2025) 73 Journal of Legal Education.

156 Erik Brynjolfsson, Danielle Li and Lindsey Raymond, 'Generative AI at Work' (2025) 140 The Quarterly Journal of Economics 889; see also Fabrizio Dell'Acqua and others, 'Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality' (2026) 37 Organization Science 403; Shakked Noy and Whitney Zhang, 'Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence' (2023) 381 Science 187; Schwarcz and others (n 153).

experienced or less well-resourced practitioners. On another reading, performance degradation among top performers on complex tasks poses risks given the high-stakes nature of legal contexts where the quality of legal work is not only critical but regulated.

7.5 Summary of findings

The evidence reviewed indicates that AI can improve efficiency and sometimes quality on bounded legal tasks, but the evidence that these gains translate into reliable, system-level effectiveness remains weak. Across the current literature, unresolved problems of verification, implementation burden, uneven benefit, and persistent error are more firmly evidenced than broad claims of justice-system improvement.

Efficiency gains are methodologically overstated. They rely heavily on self-reported measures and narrow task environments. Evaluations do not typically measure whether perceived time savings persist under real-world conditions or translate into demonstrated productivity improvements. The DBT M365 Copilot evaluation, the strongest UK evidence on this question, did not find evidence that self-reported time savings translated into demonstrated productivity improvements at the aggregate level.

Measurement of accuracy, quality, and reliability is approached differently across evidence types. Academic research verifies specific model performance on narrow task-specific deployments, while government evaluations do not report methods for independent evaluation and instead rely on self-reported data to assess AI outputs. Legal domain-specific tools generally outperform general-purpose models, although comparing accuracy across narrowly defined benchmarks does not, on its own, establish broader effectiveness.

Hallucination is a systemic and unresolved constraint on legal LLM deployment, as even domain-specific and RAG systems continue to produce fabricated and erroneous outputs. Because these errors are often difficult to detect, human verification is not optional but structurally required, which in turn limits achievable efficiency gains.

AI tools may produce real efficiency gains on bounded tasks, particularly where they are appropriately designed for the legal task at hand and where the institutional conditions for effective use are in place. The evidence on design choice indicates that legal-domain models and smaller non-generative systems outperform general-purpose models on some legal tasks. Any productivity benefits of AI may be unevenly distributed and potentially destabilising, with larger gains observed for lower-performing users, and mixed or negative effects for higher-performing users. It creates a trade-off where average performance may improve while expert reliability degrades, raising significant risks in the high-stakes nature of justice.

8 Insights and implications

The preceding sections synthesised evidence against each research question in turn. This section summarises the review's findings when the evidence bases are read together. It is organised around four structural gaps in how AI in justice is currently evaluated and deployed: a measurement gap, a deployment gap, a legitimacy gap, and a transparency gap.

8.1 Measurement gap: AI in justice is evaluated against narrow outcomes

There is a structural mismatch between how AI tools in justice settings are evaluated and the outcomes needed to assess their effects on justice.

Current evaluations of AI tools deployed in UK justice settings examine: first, the experience of system operators rather than the experience of the individuals whose cases are processed; second, self-reported metrics on time savings, satisfaction with outputs, and other process indicators rather than independent assessment of actual efficiency and effectiveness; and third, narrow, well-defined tasks in controlled environments rather than the complex settings in which legal problems arise.[157] Some evaluations indicate that self-reported time savings do not necessarily translate into realised productivity,[158] and that introducing new technology may decrease efficiency and productivity.[159] Evidence also suggests that efficiency gains from AI may be absorbed by institutional responses, including raised throughput expectations and increased caseloads, rather than redirected towards improving the quality of outcomes or the experiences of those interacting with the system.[160] As a result, the evidence base cannot support claims about system-level efficiency. What is measured is performance in isolation, not impact in practice. Establishing clear baselines and sound methodologies for evaluating system efficiency will be important in evaluating these effects.

The review did not identify any study examining the effects on an individual whose matter was triaged, advised on, or decided through an AI-assisted process in a UK justice setting. Experimental and proof-of-concept research reveals the likelihood of demographic bias and negative user experiences of such tools.[161] Evaluating fairness in this setting is made harder by the absence of a stable ground truth in most legal data, a point that limits what even well-designed bias testing can establish.[162]

157 See e.g., Incubator for AI (n 43); Home Office (n 76).
158 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade (n 113); see also Bruff and Groves (n 120).
159 Walley and Glasspoole-Bird (n 118); Pilling and others (n 74).
160 Bruff and Groves (n 120); Pilling and others (n 74).
161 Mistica and others (n 44); Blackham (n 79); Neto and others (n 46).
162 See Barale, Rovatsos and Bhuta (n 80); Neto and others (n 46); Blackham (n 79).

Existing evidence does not capture model–user interaction or cumulative system effects. Tools are typically evaluated in isolation, where they are evaluated at all. A transcription tool that produces errors, whether due to accent, audio quality, or model limitation, generates a record that may be treated as authoritative in subsequent proceedings. A triage system that fails to identify a legal issue may result in an individual not being directed to legal assistance they need, with consequences that extend well beyond the initial interaction. A summarisation tool that foregrounds certain facts and omits others shapes the information environment for every subsequent decision-maker who relies on that summary. In each case, the error or bias originates in an upstream tool but manifests in downstream decisions, making attribution and correction difficult.

Another important gap in the current evidence base is the informal use of publicly accessible AI tools – such as LLMs – by the public, litigants, legal practitioners, and the judiciary. The review did not identify any UK study that examines how members of the public or litigants in person use these tools to understand, prepare, or pursue a legal matter. Some evidence exists in other jurisdictions, drawn largely from interview-based and proof-of-concept work, which indicates that such use is already widespread and that outcomes are uneven, but its transferability to the procedural and substantive particularities of the jurisdictions of the UK cannot be assumed. Similarly, no research reveals how legal practitioners or members of the judiciary might be using these tools outside of official or institutional legal AI workflows.

A claim has recently begun to circulate that publicly available AI tools are producing, or are likely to produce, a surge in litigants in person bringing legal claims.[163] The "tsunami of small to moderate value civil claims"[164] is at present hypothetical and anecdotal, although possibly visible in some preliminary figures in other countries.[165] No systematic UK evidence currently establishes that AI is increasing the volume of claims or of litigants in person; although a rise in litigants in person bringing legal claims would not be unexpected on the evidence in this review. Current burdens on the justice system – combined with survey evidence that some of the UK public have already used, and many are inclined to use in future, a general-purpose LLM for guidance on legal matters – mean that an increase in AI-informed litigant in person claims may occur. It is likely then that any increase in such cases is due to AI being used by people who would otherwise have misidentified a problem, abandoned a legal issue with assistance, or did not previously have enough understanding or confidence to pursue formal legal action. Although, the evidence in this review gives reason to expect that AI-generated claims may be poorly framed or grounded in legal misunderstanding, because publicly accessible general-purpose LLMs produce fluent but frequently incorrect legal output, hallucinate and fabricate authority at higher rates for UK jurisdictions, and often default to the wrong jurisdiction.[166] The resulting issues are that the justice system may have limited capacity to receive the need that

163 Lord Briggs, 'AI and Civil Justice: Preparing for the Tsunami' (Oxford Civil Justice Systems in the 21st Century Conference, Oxford, 20 May 2026) https://supremecourt.uk/speeches/speech-lord-briggs-200526 accessed 26 May 2026; The Australian Fair Work Commission has received an unprecedented number of dismissal claims it says is being driven by applicants using generative AI assistance tools, see Fair Work Commission, 'General Manager Statement about Changes at the Commission' (29 May 2026) https://www.fwc.gov.au/about-us/news-and-media/news/general-manager-statement-about-changes-commission accessed 31 May 2026; While not included in the evidence review, a preprint paper posted to SSRN on 21 May 2026 revealed potential evidence of increased litigant in person claims arising in federal civil court cases, see Anand Shah and Joshua Levy, 'Access to Justice in the Age of AI: Evidence from U.S. Federal Courts' (SSRN, 21 May 2026) https://papers.ssrn.com/abstract=6766859 accessed 31 May 2026.

164 Lord Briggs (n 163).

165 Fair Work Commission (n 163); Shah and Levy (n 163).

166 Ryan and Hardie (n 62); 'ChatGPT, I Have a Legal Question? The Impact of Generative AI Tools on Law Clinics and Access to Justice' (2024) 31 International Journal of Clinical Legal Education 166; Curran and others (n 64); 'Place Matters: Comparing LLM Hallucination Rates for Place-Based Legal Queries' (arXiv, 10 November 2025) https://doi.org/10.48550/arXiv.2511.06700; Dahl and others (n 127); 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models' (2024) 16 Journal of Legal Analysis 64.

is surfaced, that the tools being used are not those best suited to legal tasks, and that the conditions of effective use have not been adequately studied. The priority that follows is to adequately measure how people are using AI to engage with the justice system, and what the impact of those tools are on the individuals seeking to resolve real legal problems and on the courts and tribunals responding to them.

Informal use of AI tools is consequential and likely to grow, yet it currently sits almost entirely outside the formal evidence base. Establishing how prevalent it is, who relies on it, and with what effect is therefore a priority for future research.

Taken together, these limitations mean that the evidence base cannot answer the central question of whether AI improves access to justice in practice. It remains focused on what AI tools can do in isolation, rather than on how they perform within the justice system as a whole.

Future work should explore:

- **Outcome-oriented evaluation frameworks for legal settings**, which should measure outcomes for users of the justice system, including fairness, differential impact, procedural legitimacy, comprehension, and legal effectiveness within real-world operational settings. Designing evaluations that are adapted to the normative and procedural demands of legal settings should be a methodological priority. General evaluative frameworks may not be most appropriate in justice settings that turn on rights, obligations, and questions of dignity, procedural fairness, and institutional legitimacy.
- **Methods for evaluating fairness in AI-informed justice outcomes**, given the risks of demographic bias, jurisdictional bias, and other differential effects. Evaluation in this context must be responsive to the distinctive characteristics of model and data biases in justice settings, in which legal data is often incomplete, does not provide a stable ground truth, and omits information where individuals should have, but did not, pursue formal legal action.
- **System-level research, rather than solely task-specific evaluations**, to inform both the functional impacts of AI tools on justice outcomes and also to reveal how the cumulative introduction of multiple AI tools across the whole justice process may impact how the public resolve legal problems. Identifying cascading effects also matters for tracing error or bias that originates upstream but manifests in downstream decisions.
- **Informal use of AI by litigants in person and practitioners**, in order to fill a significant gap in the current evidence of how AI is used in real-world legal settings. While informal use currently sits beyond the purview of formal evaluation, it already shapes how people encounter justice processes. Research designs that can study informal use without surveilling individual users will be a methodological challenge worth investing in.

8.2 Deployment gap: Real-world use diverges from evaluated systems

There is a growing disconnect between the systems evaluated in the literature and those deployed in practice.

Research tends to focus on domain-specific or strategically constrained tools, whereas real-world deployment is increasingly dominated by general-purpose, proprietary LLMs, in both public-facing contexts and internally deployed tools. The MoJ and HMCTS deploy Microsoft's Copilot which runs on OpenAI's GPT models, the Home Office's asylum casework tools and the ICO's automated case creation tool are built on OpenAI's GPT-4, and Caddy is built on Anthropic's Claude.[167] This creates a twofold problem. First, the systems studied may not be the systems deployed, which limits the transferability of the available evidence. Second, the systems deployed cannot be subjected to the structural inspection of weights, training data, and architecture that open-source or open-weight systems generally permit.

This divergence introduces unmeasured risk. General-purpose LLMs are used for legal assistance despite well-documented limitations, including hallucination, jurisdictional inconsistency, and sensitivity to input quality. Evaluative research indicates that legal-domain LLMs are more reliable than general-domain LLMs, particularly for the UK,[168] and that non-generative models can produce better outcomes for some legal applications.[169] The choice of model to evaluate and deploy should also be informed by the experience of users, whether legal advisers or members of the public. Given the spread of public-facing LLMs that present advice in a coherent and authoritative form, the effects of such tools outside traditional justice settings warrant examination.[170] Trust and willingness to act on AI outputs do not track a user's capacity to evaluate them: a minority engage critically with outputs while others treat them as authoritative.[171] The current direction of deployment, converging on general-purpose enterprise LLMs, appears to be driven by commercial availability rather than by evidence about which design approach best serves users with unmet legal need.

A further complication is that the foundation models underlying many deployed applications are not static. Performance observed at a single point in time may not persist, and it may improve or worsen for several distinct reasons. Commercial vendors release updated model versions frequently, sometimes with little notice, and older versions are deprecated or withdrawn. A newer model is not uniformly better, it may improve accuracy on some tasks while regressing on others, or behave differently in ways that are not immediately visible. Vendor changes may also adjust refusal behaviour and content filters in ways that bear on legal use, for instance by altering the handling of disclaimers or declining to engage with certain legal topics. The specific tools built on these models are themselves mutable, since their configuration may be updated with new retrieval corpora, guardrails, inference settings, or prompt-based instructions independently of the base model. The surrounding data environment shifts too. The data available to a system may become richer or poorer over time,

167 Courts and Tribunals Judiciary (n 8); Home Office (n 76); Andreas Varotsis (n 42).
168 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade (n 113); Saadany and others, (n 135); Choi and Schwarcz (n 155).
169 Rincón-Riveros and others (n 137); Barale, Rovatsos and Bhuta (n 137); Steenhuis and Westermann (n 137); Xie and others (n 139).
170 Seabrooke and others (n 57); Schneiders and others (n 85).
171 Ryan and Hardie (n 62); Hagan (n 72); Kuk and Harašta (n 71).

and a model trained or retrieved against an earlier corpus may fail to reflect subsequent changes in an evolving legal landscape, so that effective accuracy drifts even where the model itself is unchanged. There may also be shifts in the interaction between users and these systems, as trust, deference, and patterns of reliance evolve. Reliance may increase, through automation bias, complacency, or deskilling, or it may decrease as users encounter errors and recalibrate. For all these reasons, point-in-time evaluation, however rigorous, cannot by itself establish persistent performance. Longitudinal designs would help, but few exist, and the pace of model change may exceed the cadence of conventional research cycles.

This problem is compounded by the way AI is positioned in policy debate. It is frequently presented as a silver bullet for entrenched problems in policy and practice, including in justice, where unmet legal need and pressure on the system are acute. That framing, together with strong commercial and political momentum, makes it more likely that the pace of adoption will only continue to increase. Given the observed and likely continued acceleration of adoption while evaluative evidence remains slow to mature, the gap between deployment and evidence will widen.

One response is to complement rigorous independent evaluations with methods that are faster, more affordable, and better able to keep pace with change. The government's "test and learn" approach offers a model.[172] Rather than committing to system-wide fixed evaluation design at the outset, it adopts an iterative approach that tests critical assumptions early and through rapid learning cycles. The Government Digital Service has outlined a framework for testing AI for public services that adopts modular and iterative testing to ensure continuous adaptation, monitoring, and governance in the context of the evolving nature of AI.[173] While the recent MoJ policy focus on being data-driven and prioritising user-centred design is well placed, referring in general terms to measuring success and using data to inform decisions, it has not articulated a research or evaluation function directed at whether AI tools actually deployed deliver fair and reliable outcomes.[174]

Naturally, such an approach is no substitution for rigorous evaluations and trades robustness and validity for speed and relevance. It is, nonetheless, preferable to situations where deployment proceeds substantially ahead of any evaluation at all. It would be beneficial for the justice sector to adopt explicit evaluation regimes, both more rigorous and more nimble, to support the ongoing deployments of AI.

172 HM Treasury and Evaluation Task Force, 'Test and Learn' (*GOV.UK*, 15 May 2026) https://www.gov.uk/government/publications/the-magenta-book/test-and-learn-html accessed 31 May 2026.

173 Mibin Boban, 'How Do We Test and Assure AI in Government?' (*Government Digital and Data*, 22 September 2025) https://oddo.blog.gov.uk/2025/09/22/how-do-we-test-and-assure-ai-in-government/ accessed 31 May 2026.

174 Ministry of Justice, 'AI Action Plan for Justice' (n 2); Ministry of Justice, 'Ministry of Justice Digital Strategy 2025' (*UK Government*, 8 April 2022) https://www.gov.uk/government/publications/ministry-of-justice-digital-strategy-2025/ministry-of-justice-digital-strategy-2025 accessed 31 May 2026.

Therefore, future work should explore:

- **Comparative evaluation across different models and model types**, which would generate evidence to establish whether deployment decisions match performance on real-world legal use rather than on benchmarks or general non-legal tasks. Independent, like-for-like evaluation on real legal tasks should compare performance across different general-purpose models (for example, comparing GPT, Claude, or Gemini), and across model types (general-purpose models, legal-domain fine-tuned models, and non-generative models).
- **User-centric evaluation across design paradigms**, to identify whether model choice is aligned with not only performance but user comprehension, error detection, intended engagement, and differential outcomes. Evaluation should be centred on users, including legal practitioners, litigants in person, and the public. The evidence may inform both the model design and development (that is, for improved comprehension, fairness, and transparency), and deployment decisions about which model or model type to deploy in various settings.
- **The operational reality of human-AI collaboration or interaction in justice settings**, to examine empirical or observed disparities in distributional effects (positive and negative) across different users, the risk of deskilling or loss of independent discretion, and miscalibrated trust in AI outputs. It would be valuable to explore the specific conditions in which AI can be an effective assistive tool while maintaining human discretion and independence.

8.3 Legitimacy gap: Current approaches do not translate into perceived justice

A fundamental gap arises between how AI systems are evaluated and how they are experienced.

Evaluations focus on the system performance metrics, but users also experience AI in terms of how those outputs are produced. The evidence reviewed suggests that AI use in justice processes may undermine procedural legitimacy, particularly where users perceive a loss of human judgment or where transparency about AI use is limited. Even limited assistive applications are perceived in vignette experiments by the UK public as reducing fairness, dignity, transparency, and the perceived capacity for informed decision-making.

Current policy approaches and judicial guidance treat 'assistive' AI as categorically lower-risk than autonomous AI tools and treat responsibility for outputs as remaining with the human user exercising oversight.[175] While public and practitioner perspectives demonstrate more openness to administrative support from AI, even limited assistive uses of AI in UK tribunal processes may create significant reductions in perceived fairness, dignity, transparency, and informed decision-making among members of the public.[176] Litigants do not, however, experience AI only through the lens of procedural fairness. They also experience it through efficiency, speed, and the resolution of problems that would otherwise be inaccessible. For a litigant facing a long backlog, or an advice-seeker who would otherwise receive no help, the timeliness and availability of AI-assisted support is itself part of their experience of justice. The legitimacy question is therefore not whether AI use damages

175 Ministry of Justice, 'AI Action Plan for Justice' (n 2); Scottish Courts and Tribunals Service (n 13); Courts and Tribunals Judiciary (n 8); *R (on the application of Frederick Ayinde) v Haringey London Borough Council* (n 23).
176 Tomlinson and others (n 84).

procedural legitimacy in the abstract, but whether the balance between procedural protections and accessible resolution is being struck in the right place, with the right safeguards.

There is a gap not addressed in the main evidence base, but identifiable from public commentary, that reveals an active, public, and overall positive engagement and position on AI by the senior judiciary and ministerial leadership. The Master of the Rolls has spoken repeatedly about the potential of AI to assist judges, support litigants in person, and improve case management. The Chancellor of the High Court, as Lead Judge for AI, has described concrete applications – including AI-assisted transcription, support for anonymisation, the identification of internal inconsistencies in draft judgments, and routine administrative tasks. The MoJ has positioned itself as one of the fastest-growing users of Copilot across government, with the Lord Chancellor publicly endorsing the trajectory, and the first reported instance of a judge confirming use of Copilot in a court decision. The institutional posture, at the level at which strategic direction is set, is one of confidence in AI's contribution to the future of the justice system.[177]

The empirical evidence on how judges actually regard and use these tools is thin, and what little exists points towards principled hesitation and concerns, particularly in the lower courts. The UK evidence on judicial attitudes towards AI rests on Solovey and others (2025), a study of 12 judges ranging from the Supreme Court to the County Court.[178] Judges expressed three primary concerns: the risk of deskilling, felt most acutely in the lower courts, where reliance on AI-generated summaries or draft reasoning was thought capable of eroding core analytical competencies; the normative importance of human judgment, including the value to parties of knowing that a person has engaged with their case; and the reliability of AI outputs, particularly the risk of fabricated material in legal research. While this study is limited by small sample size and self-selected participants, it is also corroborated by three international studies that identified a similar cautious and sometimes sceptical approach of German and US judges that expressed concerns about judicial duties in conflict with AI tools.[179]

This tension is a feature of the current landscape that is worth identifying. Whether the senior leadership framing reflects, or will come to shape, actual judicial behaviour is an open question that the current evidence cannot answer. It is a priority area for future research to identify how widely, how competently, and to what effect AI is used across the judiciary as a whole, particularly in the high-volume first-instance courts and tribunals.

Human oversight is treated, in policy and guidance, as a principal safeguard, and it is crucial to perceptions of procedural fairness.[180] A reduction of independent discretion and judgment would likely have significant effects on the perceived role of AI in justice. Similarly, systems that shape how decisions are reasoned, what information is available, and what the record contains may have equally significant, but less visible, effects on justice outcomes. However, no UK deployment evaluation identified in this review examines whether that oversight is effective, calibrated, or sustainable.[181] Given the prominence of LLM deployments, the verification burden is high,[182] yet hallucination

177 See summary in Section 1.
178 Solovey, Flanagan and Chen (n 92).
179 Dhungel and Heine (n 93); J Ludwig, J Kleinberg and S Mullainathan, 'Cohort Bias in Predictive Risk Assessments of Future Criminal Justice System Involvement' (2023) 120 Proceedings of the National Academy of Sciences of the United States of America; Fine, Le and Miller (n 94).
180 Meers and others (n 87); Fine, Berthelot and Marsh (n 89).
181 Although the Home Office identified the risk and acknowledged it would be the basis for future evaluations, Home Office (n 76).
182 Dahl and others (n 127); ibid; Ribary and others (n 130); Munir and others (n 130).

detection has been recorded only through self-reported information.[183] The real costs of human oversight are largely invisible in current evaluations. Survey and experimental research provide valuable insights; however, it leaves a gap between expected perceptions under study conditions and real-world consequences in deployment.

Standard evaluation frameworks are not sufficient for justice settings because legal decision-making involves procedural obligations, contested reasoning, and high-stakes consequences that are not captured by measures of accuracy or efficiency alone. Evaluations designed for administrative or commercial contexts risk systematically mismeasuring impact, by overlooking legitimacy, reason-giving, and the quality of human engagement.[184] Tools may therefore appear effective against generic metrics while undermining the conditions required for lawful and legitimate adjudication.

Therefore, future work should explore:

- **Experiential research in deployment settings**, to progress current hypothetical and experimental research towards understanding how individuals and groups experience AI-mediated justice processes, and how those experiences compare with their expectations and with non-AI comparisons. While attitudinal research is valuable, it is possible that direct experience may moderate attitudes in different directions: familiarity may increase trust, as some findings suggest, or it may confirm and deepen concerns about the quality of human engagement.[185]
- **The perceived and real-world value of governance approaches**, such as the perceived safeguard of human oversight as a mitigation for automation bias and over-reliance.[186] Research is needed that can examine the real-world interactions between AI tools and various stakeholders to test whether such assumptions hold in practice or identify the conditions under which it may fail.

8.4 Transparency gap: AI is deployed faster than it is independently scrutinised

AI is entering justice workflows more quickly than it is being independently evaluated against user outcomes, equality impacts, or downstream legal effects.

In several UK settings, tools appear to have been integrated into operational practice without public evidence of how they affect the people whose cases they shape. Where evaluation exists, it is often internal, partial, or methodologically thin, which makes claims of effectiveness difficult to test and weakens accountability.

183 Digital Data and Technology Monitoring and Evaluation Team and Department for Business and Trade (n 113).
184 See Solovey, Flanagan and Chen (n 92); Dhungel and Heine (n 93); Ludwig and others (n 91); Cheong and others (n 102); Fine, Berthelot and Marsh (n 89).
185 Hagan (n 72); Tomlinson and others (n 84).
186 Meers and others (n 87); Fine, Berthelot and Marsh (n 89).

Internal or vendor-led evaluations are subject to an actual or perceived risk of optimism bias. The incentives within the institution producing or procuring the tool are not the same as the incentives of an independent evaluator, and the resulting evidence base, however carefully assembled, requires external verification before claims can be confidently relied upon. As a result, the apparent effectiveness of some AI tools may reflect selective reporting rather than demonstrated performance. Independent evaluation capacity needs to grow to enable verification of claims and assumptions about AI's opportunities and risks, regardless of the rigour of internal processes. The dominance of proprietary models in UK deployments compounds this, because without disclosed weights, training data, or interpretable architecture, independent systematic evaluation is not possible. Independent behavioural evaluation through benchmarking, red-teaming, and human-interaction studies must therefore carry more of the verification burden than would be the case for open-source or even open-weight models.

At the same time, the structure of the legal data environment shapes both accountability and innovation. Efforts to improve access to judgments and legal data create important conditions for research, scrutiny, and tool development. However, uneven access and the dominance of proprietary systems risk reinforcing existing asymmetries in access to justice and limiting who can evaluate or build legal AI systems.[187]

A further factor reflects the institutional setting of justice. Historically, the justice system has a less developed tradition of independent and empirical evaluation than other parts of the public sector. The MoJ has itself identified barriers to evaluation, including limited central oversight to monitor evaluation activity and prioritise analytical resource, variable experience producing inconsistent coverage, limited access to high-quality data, small sample sizes that constrain impact evaluation, and operational and ethical contexts that make randomised controlled trials difficult.[188] On the specific evaluation of AI deployments, justice institutions would benefit from stronger interdisciplinary and methodological support for the technical and sociotechnical work that meaningful scrutiny requires. Evidence is, by definition, lagged and backward facing, while deployment is fast and likely to accelerate. An evaluation regime built solely on traditional independent studies will therefore always trail practice. Therefore, an appropriate response is to also approach more iterative and flexible evaluation methods, such as a structured test-and-learn, as explored above.

187 Byrom (n 26); Aidinlis and others (n 105); Somers-Joce, Hoadley and Nemsic (n 107).

188 Ministry of Justice, 'MOJ Evaluation and Prototyping Strategy' (UK 2023) Policy Paper https://www.gov.uk/government/publications/moj-evaluation-and-prototyping-strategy/moj-evaluation-and-prototyping-strategy accessed 27 May 2026.

Therefore, future work should explore:

- **Independent, collaborative, and interdisciplinary evaluation infrastructure**, to improve the strength and credibility of the current evidence base. Greater cross-sector collaborations,[189] and dedicated capacity for this type of research, such as a 'What Works Centre for AI in Public Services' body as proposed by the Ada Lovelace Institute,[190] are worth considering.
- **Employing and publishing responsive evaluation models alongside conventional studies**, such as structured test-and-learn approaches or action research, designed to generate evidence on a timescale closer to that of deployment. The aim is not to replace independent evaluation but to reduce the interval during which tools operate without publicly available information.
- **Longitudinal institutional research**, in order to map the institutional trajectory of AI adoption over time and to test the causal assumption that task-level efficiency translates into material improvements in justice outcomes.

189 Centre for Homelessness Impact (n 41); King's College London (n 41).
190 Bruff and Groves (n 120).

9 Reflections on our understanding of AI's role in justice

The Nuffield Foundation's *Public right to justice* (PRTJ) programme proceeds from the premise that the public have a right to a justice system that accessibly, fairly, and effectively meets their legal needs, having regard for their characteristics and circumstances. The system should be open, accountable, trustworthy, and supportive of wider societal values. The integration of AI into the justice system can and should support these goals.

There is meaningful potential to improve the justice system from the perspective of three constituencies: the people with everyday legal problems, the practitioners who help resolve them, and the administrators and policymakers who oversee the system. A gain realised for one group frequently depends on, or produces, a gain for another, and several of the most significant opportunities sit precisely at those intersections.

For people facing everyday legal problems, the central opportunity is accessibility. Many legal issues never reach formal legal processes, and many individuals are deterred by cost, distance, language, unfamiliarity with legal procedure, or distrust of the system more broadly. Public-facing triage,

navigation, and intake tools offer a potential route to people whom traditional legal information and advice do not reach. Tools available at lower cost, at any time, in plain language, and in more than one language carry particular potential for currently underserved groups. Better access to information and support could improve a person's ability to identify whether a problem is legal and actionable, to translate legal terminology, to clarify rights and responsibilities, and to navigate formal processes.

For legal practitioners and advisers, the judiciary, and court staff, the central opportunity is efficiency and capacity. AI-assisted transcription, document processing, case management, legal research, and client intake may redirect scarce professional time towards more complex and people-centred legal work. By absorbing routine tasks, such tools may extend the capacity of the legal aid and advice sector to serve more people with more challenging legal needs.

For justice administrators and policymakers, the central opportunity is greater data-driven insight. AI and data-driven methods can methodically analyse the large volume of structured and unstructured justice data that is presently poorly utilised. Better empirical insights can improve understanding of system functioning: where delay accumulates, when case management is successful and when it is not, which areas of legal need go unmet, and how resources might be allocated to greatest effect. This offers an operational benefit in identifying bottlenecks and

improving the capacity of courts and formal legal processes. It also produces better evidence to support more accountable and responsive policy. The same data capability that enables insight is the capability required to evaluate whether AI interventions deliver the claimed benefits in practice.

As demonstrated throughout this review, many such interventions are currently being trialled, deployed, or explored. The core finding of the review, however, is that AI tools are being deployed faster than their effects on the public's right to justice can be independently or transparently evaluated. Whether the justice system is accessible, fair, effective, open, and trustworthy under conditions of AI deployment is an empirical question, and one the current evidence is not well placed to answer.

The problem is not principally one of volume. The deeper issue is that the evidence which exists to date has not sufficiently measured people-centred outcomes from the standpoint of those the justice system serves. Efficiency and task performance are well represented, and they show meaningful possibilities for improving the performance of certain legal tasks within the system. Fairness, comprehension, lived experience, and procedural legitimacy remain comparatively under-studied. The result is that AI tools may be assessed as effective solely against the metrics by which they are evaluated, while the conditions on which the public right to justice depends are more challenging to measure. Deployment is proceeding under those conditions.

This has two consequences for the PRTJ programme. The first is that addressing the evidence gap is itself a precondition for the programme's substantive aims. Until the evidence base measures what matters for the public, claims about whether AI is advancing or undermining the right to justice cannot be tested or answered. The second is that the programme is distinctively positioned to advance the future work identified in Section 8.

9.1 Observations on the AI and justice agenda

Where the review has presented a synthesis of the evidence, this section draws together some insights from that evidence, and from the broader context in which these tools are being developed and deployed, to set out what I believe the review can suggest about priorities for AI and justice. The four observations are interpretive. They are anchored in the evidence reviewed but also consider issues more broadly, and are intended to support the PRTJ programme's next phase. Where an observation rests more on judgment than on the evidence base, I have tried to say so.

The distribution of benefit and risk. The clearest pattern, to my mind, is not about what AI can do but about who is presently positioned to benefit from it. Deployment in UK justice settings places these tools in the hands of actors who already navigate the system competently. Judges refine and check their judgments more efficiently; lawyers research and draft faster; civil servants summarise documents and triage casework. While some evidence reports time savings, measurable productivity and efficiency gains across the justice system have not been demonstrated. If designed and deployed well, future applications will hopefully reveal more material efficiency and capacity gains from these AI interventions.

These uses do not, however, in themselves reach the population for whom the formal justice system is, in practice, inaccessible. It is not difficult to see why deployment has taken this shape. Internal-facing tools are easier to scope, to govern, and to evaluate against existing professional workflows. Their risks are contained, at least nominally, by the expertise of the user. Building tools for the public

is a more difficult undertaking, and the caution shown by public bodies is in some respects warranted. The problem is what the resulting gap is being filled with. The people with the least access are not waiting for purpose-built support. Where they engage with these technologies at all, the available evidence suggests they are turning to general-purpose chatbots, informally and unsupervised, in the very legal domains where those tools are least reliable. If the productivity gains accrue mainly to the already served, and the reliability risks fall mainly on the underserved, AI could improve the efficiency of the overall justice system while leaving its central access problem untouched, or even widening it. There is a further asymmetry within the underserved group itself. The people least reached by the formal system are often those least equipped to use digital tools, whether by reason of digital access, literacy, language, or disability, so the very tools most often proposed to widen access may be the least usable by those who most need them.

The choice of technology. The pace at which general-purpose AI and generative AI have developed has, I think, distorted the conversation about what tools the justice sector should be deploying. The dominant framing is generative AI against nothing, when the relevant comparison is often generative AI against simpler, narrower, and better-understood alternatives. For many tasks, predictive machine learning models or rule-based systems may perform comparably or better, and they are typically easier to validate, more interpretable, and not subject to the same hallucination risk. The fluency of LLM outputs is genuinely valued by users, and the evidence on user preference reflects this. However, fluency is not the same as reliability, and the irreducible risk of hallucination has implications that apply to both internal and public-facing tools, regardless of how those tools are framed.

A second issue follows. The current deployment pattern shows a pronounced deference to proprietary models. There may be reasons for this, including support, performance at the frontier, and procurement convenience. There are also costs, including limited interpretability, change-of-behaviour risk when providers update models, sharing of data access with commercial vendors, and the absence of the audit access on which independent evaluation depends. Open-weight models such as Mistral, and open-source models such as OLMo (and, for classification and extraction tasks rather than generation, encoder models such as BERT), particularly when fine-tuned on domain-specific data or deployed on narrowly scoped tasks, may perform comparably while offering substantially greater transparency.[189] At a minimum, the choice between proprietary and open-weight infrastructure should be treated as a substantive design decision with consequences for transparency, accountability, and lock-in, rather than as a procurement default.

The true costs of adoption. The justice sector's current orientation is heavily weighted towards buying rather than building. There are sound reasons for this in the short term, including the absence of in-house capacity, the cost of building, and the pace at which the underlying technology is developing. The long-term implications, however, deserve more scrutiny than they have received. A 'buy' strategy concentrates dependency on a small number of external providers and leaves the public sector with limited ability to specify, evaluate, or alter the systems it uses. Where the state's relationship with its own citizens is mediated by tools it does not control, that dependency carries weight beyond procurement.

Many AI tools currently look inexpensive at the point of use. That appearance should be treated with caution. It does not tell us what these systems will cost once the market matures, once providers seek sustainable returns, once public bodies have become dependent on particular systems, and

once the full costs of oversight, audit, explainability, appeals, security, procurement, maintenance, and error correction are counted. The evidence already gestures at these hidden costs. The labour of configuring and supervising the tools is substantial and tends to fall on a small number of people, and the verification burden created by hallucination is recurring rather than one-off. Reliance on proprietary models, whose behaviour can change without notice and which cannot be independently inspected, adds a dependency that is easy to enter and hard to leave. A tool that is cheap to adopt may be costly to govern, and more costly still to leave.

The contested nature of legitimacy. The strongest UK studies in the review, strong in their internal design even if not yet tested in real-world settings, find that even modest assistive uses of AI can reduce the public's sense that a process is fair and dignified.[190] That is a significant finding, and not one to dismiss. It is not, however, the whole picture. International evidence, which must be read with care given the institutional differences involved, points in a different direction for some groups, particularly those who already perceive the existing system as biased against them. For those groups, AI may be received not as a threat to legitimacy but as an opportunity to improve it.[191] The relevant comparison for them is not AI against an ideal of critical human attention, but AI against a human system they already distrust. Public responses to AI in justice are conditional and contested. The same technology can both threaten and serve legitimacy, depending on who encounters it and what they are comparing it against. Policy that treats legitimacy as a single quantity to be optimised will, I think, miss this. The question worth asking is whose legitimacy, measured against what alternative, under which conditions.

9.2 Priorities and responsibilities

Framing this research through the public right to justice matters because it keeps access at the centre of the analysis. The evidence gaps documented in this review are significant. They are not, however, indications that AI cannot assist access to justice. The right response to uncertainty is to build the evidence and to design well.

What remains unresolved is much of what matters most for society: whether AI improves or worsens outcomes for individuals, and for which groups; whether human oversight, the safeguard on which policy most relies, actually works in practice; whether the negative legitimacy effects seen in experiments survive contact with real cases; and what these systems will truly cost. Three dimensions of future priorities follow.

Research. The priority is to measure what has not been measured: outcomes for the people whose problems are processed, disaggregated by group; the experience of AI-mediated justice in real settings rather than in vignettes; the effectiveness, not merely the presence, of human oversight; and the informal use of public-facing tools, studied with responsible evaluative methods. Alongside conventional independent evaluation, the sector should adopt faster, structured test-and-learn methods so that evidence is generated closer to the pace of deployment. This applies with most force to lower-stakes, internal-facing tools. Where AI bears directly on a person's rights or the outcome of their case, the pace of deployment does not lessen the case for independent evaluation in advance.

Policy and procurement. Independent evaluation capacity should be built, and evaluation against people-centred outcomes should, where possible, be a condition of deployment rather than an afterthought. Procurement should treat cost transparency, auditability, and the ability to change

provider as requirements, precisely because dependency on proprietary systems is among the most consequential and least reversible choices currently being made. Policy should also resist the framing of AI as a remedy for entrenched problems it cannot solve alone. Presenting it as an answer to unmet legal need risks substituting a tool for the investment the system requires.

Practice. Deployment should be oriented towards people facing everyday legal problems that are underserved by the current system. Public-facing tools in sensitive domains should generally be embedded within human support rather than offered as standalone self-service products, a lesson the evidence already teaches. Human oversight should be designed and tested as a real safeguard, with attention to the conditions under which it fails. The use of AI should be disclosed where it bears on a person's case, so that transparency is built into the process rather than left for a litigant to discover.

Some of these priorities will require the development of independent evaluation infrastructure and longitudinal institutional research, necessitating joint action by government, regulators, or sector bodies.

The implication is therefore not only a research agenda but a division of responsibility. Though the research and analytical community can develop the evidence base aligned with the public right to justice, how that evidence base translates into practice at the scale required will depend on decisions taken elsewhere: by funders, by government, and by those responsible for deploying AI in justice settings.

The findings of this review suggest these decisions cannot reasonably be deferred. AI is already shaping the conditions under which people encounter the justice system. There is reasonable ground to think that AI, designed and used well, can extend access and reduce friction for users who would otherwise be poorly served. Whether it will, in fact, advance or undermine the public's right to justice is, at present, an open question, and one the public is entitled to have answered.

Appendix A. Tool-mapping

Primary Developer	Tool	Status	JDX	Public / Internal	Domain	AI Type	Primary Function	Evaluation
Acas	Dispute Support	Expl.	UK	Int	Employment	Not reported	Decision support	None published
Access Social Care	AccessAva	Live	E	Int	Welfare	LLM	Legal information	None published
Anathem	Witness Statement Assistant	Pilot	E&W	Int	Criminal (Procedural)	ASR, LLM	Document drafting	Mixed methods (academic)
AskEllie Ltd	AskEllie	Live	UK	Pub	Education (SEND)	LLM	Legal information	None published
Beam	Magic Notes	Pilot	E	Int	Welfare	ASR, LLM	Transcription	User survey (vendor-commissioned)
Bequeathed	MyIntent	Live	UK	Pub	Family (Wills)	LLM	Document drafting	None published
Cafcass	Scribe	Live	E&W	Int	Family (Children)	LLM, RAG	Document drafting	None published
Cafcass	Chatbot (Genesys)	Live	UK	Pub	Family	LLM	Legal information	None published
CaseCraft.AI	CaseCraft.AI	Live	UK	Pub	Civil (Small Claims)	Not reported	Case preparation	None published
Children's Law Centre NI	Rights Responder	Live	NI	Pub	Child Rights	Not reported	Legal information	None published
Citizens Advice	Caddy	Live	E&W	Int	Civil (Legal Aid Setting)	LLM, RAG	Document drafting	A/B trial (government)
Citizens Advice Scotland	Casey	Live	S	Int	Civil (Legal Aid Setting)	LLM, RAG	Legal information	None published
Citizens Advice Scotland	Cassie	Live	S	Pub	Civil (Legal Aid Setting)	LLM, RAG	Legal information	None published
DFJ Trailblazer Project	Trailblazer AI Assistant	Pilot	UK	Pub	Family	LLM	Legal information	None published
Fletchers Solicitors	SIDSS	Live	UK	Int	Civil (Negligence)	ML	Triage and referral	None published
FOS	AI Casework Tools	Live	UK	Int	Financial Services (Ombudsman)	Not reported	Case summarisation	None published
Garfield.Law	Garfield.Law	Live	E&W	Pub	Civil (Small Claims)	LLM	Case preparation	None published
HMCTS	Case Management AI	Expl.	UK	Int	Courts & Tribunals	LLM, RAG	Legal research	None published
HMCTS	Document Anonymisation	Expl.	UK	Int	Courts & Tribunals	Not reported	Anonymisation	None published
HMCTS	ListAssist	Live	UK	Int	Courts & Tribunals	ML	Scheduling	None published
HMCTS	eDiscovery	Live	E&W	Int	Courts & Tribunals	TAR	Document review	User survey (academic)

HMCTS	AI Transcription	Pilot	UK	Int	Courts & Tribunals	ML	Transcription	None published
HMCTS	Court Forms Processing	Pilot	UK	Int	Courts & Tribunals	ML, CV	Case preparation	None published
HMCTS	Knowledge Assistant	Pilot	UK	Int	Courts & Tribunals	LLM, RAG	Legal research	None published
Home Office	Asylum Case Summarisation (ACS)	Live	UK	Int	Immigration	LLM	Case summarisation	A/B trial (government)
Home Office	Asylum Policy Search (APS)	Live	UK	Int	Immigration	LLM	Legal research	A/B trial (government)
ICO	ICE 360	Live	UK	Int	Data	LLM	Case summarisation	None published
ICO	Digital Assistant	Live	UK	Pub	Data	LLM, ML	Legal information	None published
KCL	Housing Helper	Pilot	E&W	Pub	Housing	LLM, RAG	Legal information	Trial (RCT) (in progress; academic)
LawConnect	LawConnect	Live	UK	Pub	Civil (Legal Aid Setting)	LLM	Legal information	None published
LawFairy	LawFairy	Live	E&W	Pub	Immigration	Rules-based	Case preparation	None published
LRA	Holiday Pay AI	Pilot	NI	Int	Employment	Not reported	Triage and referral	None published
MoJ	AI Scheduling	Expl.	UK	Int	Courts & Tribunals	Not reported	Scheduling	None published
MoJ	Justice Transcribe (Courts)	Expl.	UK	Int	Courts & Tribunals	LLM, ASR	Transcription	None published
MoJ	Digital Justice Assistant	Expl.	UK	Int	Family (Children)	Not reported	Legal information	None published
MoJ	Secure AI Assistants	Live	UK	Int	Courts & Tribunals	LLM	Legal research	None published
MoJ	Probation Semantic Search	Live	UK	Int	Criminal (Probation)	LLM, RAG	Case summarisation	None published
MoJ	Justice Transcribe	Pilot	E&W	Int	Criminal (Probation)	LLM, ASR	Transcription	None published
OLS Solicitors	Nessa	Live	E&W	Pub	Family	LLM	Legal information	None published
SCTS	Transcription Service	Live	S	Int	Courts & Tribunals	ASR	Transcription	None published
SCTS	Writ Data Extraction	Live	S	Int	Courts & Tribunals	NLP	Case preparation	None published
Spectrum Dynamics	SpektraBot	Pilot	UK	Pub	Education (SEND)	Not reported	Legal information	None published
Swansea University	Legal Centre Chatbot	Pilot	W	Pub	Child Rights	NLP	Legal information	Mixed methods (academic)
Tabled Technologies	Project Odyssey	Pilot	E&W	Int	Civil (Legal Aid Setting)	LLM, RAG	Legal information	None published
Valla	Valla	Live	UK	Pub	Employment	LLM	Case preparation	None published

Appendix B. Search strategies and sources

Databases, platforms, and grey literature sources

Academic and legal databases

- Scopus
- HeinOnline
- Lexis
- Westlaw

Grey literature and policy sources

- UK Government Algorithmic Transparency Records
- HMCTS and MoJ websites
- Tracking Automated Government Register
- Overton

Supplementary identification

- Backward and forward citation review, manual snowballing from key papers and reports
- Opportunistic and manual identification of salient developments not indexed in academic databases

Source	Search string	Date	Results	Screening approach
Scopus	TITLE-ABS-KEY ((“artificial intelligence” OR AI OR “machine learning” OR “deep learning” OR algorithm* OR “natural language processing” OR NLP OR “large language model” OR “large language models” OR LLM OR LLMs OR “generative AI” OR “automated decision” OR “automated decisions” OR “automated decision making” OR “automated decision-making” OR “algorithmic decision” OR “algorithmic decisions” OR “algorithmic decision making” OR “algorithmic decision-making” OR “predictive analytic” OR “predictive analytics” OR “data driven system” OR “data driven systems” OR “data-driven system” OR “data-driven systems”) AND (judiciar* OR “justice system” OR “justice systems” OR “legal advice” OR “legal service” OR “legal services” OR “legal dispute” OR “legal aid” OR “legal need” OR “legal needs” OR “legal argumentation” OR “access to justice” OR “dispute resolution” OR “litigant in person” OR “unrepresented litigant” OR “unrepresented litigants” OR “case law”) AND (empiric* OR evidence OR evaluat* OR assess* OR qualitative OR quantitative OR “mixed method” OR “mixed methods” OR “mixed-method” OR “mixed-methods” OR interview* OR survey* OR ethnograph* OR experiment* OR “field study” OR “field studies” OR observational OR “case study” OR “case studies” OR “impact assessment” OR pilot OR pilots OR trial OR trials OR implementation OR “implementation study” OR “implementation studies” OR “F1”))	14 Jan 2026	1,392	2 retracted, 2 duplicates removed Retained 1,388 records Exported all items
	AND (UK OR “United Kingdom” OR UK’s OR “United Kingdom’s” OR England OR Wales OR Scotland OR “Northern Ireland” OR HMCTS OR “Ministry of Justice” OR MoJ)	14 Jan 2026	50	Targeted jurisdiction query for first pass review
Scopus Preprints	As above	14 Jan 2026	298	6 duplicates removed Manual review of title / abstract retained 24 preprints for review
HeinOnline	(“artificial intelligence” OR AI OR “machine learning” OR “deep learning” OR algorithm* OR “natural language processing” OR NLP OR “large language model” OR “large language models” OR LLM OR LLMs OR “generative AI” OR “automated decision” OR “automated decisions” OR “automated decision making” OR “automated decision-making” OR “algorithmic decision” OR “algorithmic decisions” OR “algorithmic decision making” OR “algorithmic decision-making” OR “predictive analytic” OR “predictive analytics” OR “data driven system” OR “data driven systems” OR “data-driven system” OR “data-driven systems”) AND (judiciar* OR “justice system” OR “justice systems” OR “legal advice” OR “legal service” OR “legal services” OR “legal dispute” OR “legal aid” OR “legal need” OR “legal needs” OR “legal argumentation” OR “access to justice” OR “dispute resolution” OR “litigant in person” OR “unrepresented litigant” OR “unrepresented litigants” OR “case law”) AND (empiric* OR evidence OR evaluat* OR assess* OR qualitative OR quantitative OR “mixed method” OR “mixed methods” OR “mixed-method” OR “mixed-methods” OR interview* OR survey* OR ethnograph* OR experiment* OR “field study” OR “field studies” OR observational OR “case study” OR “case studies” OR “impact assessment” OR pilot OR pilots OR trial OR trials OR implementation OR “implementation study” OR “implementation studies” OR “F1”) AND (UK OR “United Kingdom” OR UK’s OR “United Kingdom’s” OR England OR Wales OR Scotland OR “Northern Ireland” OR HMCTS OR “Ministry of Justice” OR MoJ)	14 Jan 2026	125,264	Only allows full text search, due to inflation results were sorted by relevance and first 100 abstracts were screened.
Lexis	(“artificial intelligence” or “AI” or “automated decision!” or algorithm!) /p (pilot! or trial! or evaluat! or empiric! or survey! or “impact assessment” or qualitative or quantitative) and (“access to justice” or “litigant! in person” or “legal aid”)			Manual review of title / abstract retained 14 records for review
Westlaw	(“artificial intelligence” or “AI” or “automated decision!” or algorithm!) /p (pilot! or trial! or evaluat! or empiric! or survey! or “impact assessment” or qualitative or quantitative) and (“access to justice” or “litigant! in person” or “legal aid”)	14 Jan 2026	193	Manual review of title / abstract retained 13 records for review
Overton	(“artificial intelligence” OR “machine learning” OR algorithm*) AND (“evaluation” OR “evidence” OR pilot* OR case stud*) AND (“justice system” OR “legal system” OR “access to justice”) NOT “criminal justice” NOT policing NOT “Online Dispute Resolution”			Manual review of title / abstract retained 5 records for review
Algorithmic Transparency Records	Reviewed all 125 records.	14 Jan 2026	125	Manual review of all records and retained 8 records for review

Appendix C. Screening protocol and eligibility criteria

Inclusion criteria

Include records that meet all of the following:

- Technology: AI/data-driven systems beyond basic automation (e.g., ML, predictive analytics, NLP / LLMs / generative AI, algorithmic decision support).
- Domain: Administrative, civil, or family justice contexts (including advice/triage linked to these problems); or closely relevant comparative domains.
- Process: Formal legal action, adjudication, or upstream processes leading to them (including advice/triage where it affects entry into formal processes).
- Evidence: Contains empirical/evaluative content (quantitative, qualitative, mixed-methods, pilots, trials, implementation evaluation, or robust observational/case study evidence).
- Jurisdiction: Priority England & Wales; then UK; then comparable jurisdictions where relevant.

Exclusion criteria

Exclude where any of the following apply:

- Purely conceptual/normative commentary without empirical/evaluative evidence (coded: "Excluded: Not Evidence").
- Not meaningfully connected to justice/legal problem resolution or adjudication pipeline (coded: "Excluded: Scope").
- Criminal justice only, unless directly overlapping or uniquely informative for the primary scope (coded: "Excluded: Criminal Justice").
- Digitalisation only (no AI or data-driven inference) (coded: "Excluded: Scope").
- Inaccessible, either because paper is published in non-English language or only accessibly behind paywall (coded: "Excluded: Language" or "Excluded: No Full Text")

Screening workflow

1 Import records from searches into Zotero.

2 Deduplicate with Zotero merge and manual checks.

3 First pass on title and abstract:
- Exclude "Excluded: Scope"

4 Second pass on topically relevant items:
- Exclude "Excluded: Not Evidence"

5 For included items:
- Tag jurisdiction (E&W / UK / comparator)
- Tag status in synthesis workflow (Reviewed / Reported / To Report)
- Begin extraction into charting table

Appendix D. PRISMA-ScR

Records were identified through systematic searches of academic and legal databases (Scopus, Scopus Preprints, HeinOnline, Lexis, Westlaw) conducted on **14 January 2026**, supplemented by targeted grey literature searching and manual identification of high-salience deployments (including UK Algorithmic Transparency Records). Records were imported into Zotero and deduplicated. Titles and abstracts were screened for relevance to AI/data-driven technologies in administrative, civil, and family justice contexts. A second-pass screen excluded records that were conceptually relevant but contained no empirical or evaluative evidence. Remaining records were progressed to full-text review and data charting.

Stage	Category	Count
Identification	Records identified from databases	1,515
	Records identified from other sources (manual searches, snowballing, grey literature)	207
	Total records identified	1,722
Deduplication	Duplicates removed	76
	Records imported to Zotero	**1,656**
Screening – Title/Abstract	Records excluded (scope not relevant)	926
	of which: criminal justice (subset)	(257)
	of which: non-English language	(4)
	Records excluded (not empirical/evaluative)	289
	Records progressing to full-text review	**441**
Eligibility – Full Text	Full-text articles assessed	441
	Not reviewed in full (non-UK, technical only contributions)	158
	Records excluded: not empirical/evaluative	66
	Records excluded: insufficient relevance to review	55
	Records excluded: no full text available	5
	Studies included in charting	**157**
Included	UK evidence	48
	International evidence	109
	Tool entries mapped	45

This report was commissioned by the Nuffield Foundation as part of its Public right to justice programme. The views expressed are those of the author and not necessarily the Foundation. The report forms part of a series of evidence reviews and briefing papers critically examining different dimensions of the justice system and access to justice in England and Wales. The programme aims to develop an interconnected body of research, providing policymakers, practitioners and the public with a better understanding of how the justice system operates, where it falls short, and how it could better meet the needs of those it serves.

As an independent charitable trust with a mission to advance social well-being, the Foundation funds and undertakes rigorous research, encourages innovation and supports the use of sound evidence to inform social and economic policy, and improve people's lives. It is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute, and the Nuffield Family Justice Observatory.

Find out more at: nuffieldfoundation.org
Bluesky: @nuffieldfoundation.org
LinkedIn: Nuffield Foundation