TR-2026-01 · VIRTUALVAKIL RESEARCH LAB · APRIL 2026

VIM-1: A Privacy-Preserving, India-Hosted Legal AI for India

VIM-1 (VirtualVakil Intelligence Model 1) is Virtual Vakil's proprietary, domain-specialised legal AI, served from Indian infrastructure. It powers the next generation of our WhatsApp assistant with grounded offline answers, persistent case comprehension across uploaded documents, multi-day conversation continuity, and locally-solved court-record ingestion.

Indian Legal AI | India-Hosted | Privacy-First | Case Dossier | Session Continuity | Grounded Answers

ABSTRACT

Abstract

Legal practice in India is constrained by a shortage of specialists, fragmented statutory sources, and the absence of consumer channels that deliver legal guidance at mobile latencies. The August 2025 Virtual Vakil paper introduced a multi-agent legal intelligence system that addressed the first two constraints by decomposing the problem into eight role-specialised agents delivered through WhatsApp. This paper introduces its successor, **VIM-1** — the VirtualVakil Intelligence Model 1 — a privacy-preserving, India-hosted, domain-specialised legal AI that serves real users in production today.

VIM-1 advances the 2025 system along four axes. First, a proprietary retrieval layer lets the assistant answer bare-act and common-statute queries directly from our own corpus, keeping the data path entirely within Indian infrastructure. Second, a persistent case-comprehension service — the Case Dossier — accumulates understanding across every document a user uploads over the life of a matter, so that cross-document questions are answered without the user having to re-explain the matter. Third, a conversation-continuity service — the Case Session — carries a matter-focused consultation across days, through a clear lifecycle of active, resting, and closed, so that returning users pick up where they left off. Fourth, a locally-solved vision component reads court-record portal captchas inside our own infrastructure, keeping the court-data pipeline fully Indian from capture to result.

VIM-1 is positioned as a privacy-first alternative for Indian legal AI: served from Indian servers, grounded on Indian statutory sources, aligned with the Digital Personal Data Protection Act 2023, and built for the delivery surface — WhatsApp — that Indian users actually reach for.

AUTHORS & AFFILIATIONS

Who wrote this paper

Mahir Gupta

PRINCIPAL INVESTIGATOR

VirtualVakil Research Lab · Advocate, Delhi High Court

Piyush Gupta

CO-INVESTIGATOR

VirtualVakil Research Lab

VirtualVakil Research Lab

CONTRIBUTING RESEARCHERS AND REVIEWERS

New Delhi, India

Correspondence: research@virtualvakil.com

1. INTRODUCTION

From specialist agents to VIM-1

The August 2025 Virtual Vakil paper established three design commitments that this paper carries forward unchanged. Routing precedes generation: a lightweight classifier chooses the specialist persona before any generation is performed. Delivery is a first-class constraint: the WhatsApp customer-care window, the approved-template regime for outbound broadcasts, and the plain-text formatting that renders inside a chat bubble are treated as primary design inputs. Evaluation is by practising advocates on sampled production traffic, not by crowdsourced labels.

Between August 2025 and April 2026, Virtual Vakil's research programme has concentrated on four capabilities that a consumer legal assistant for India must possess and that the 2025 architecture was not yet organised around. First, the assistant must be able to answer common statutory and procedural questions without reaching outside Indian infrastructure — both for latency and for the privacy expectations that Indian users reasonably hold of their legal counsel. Second, the assistant must carry comprehension across multiple documents in a single matter, because a legal matter is not a single question but an evolving file. Third, the assistant must preserve a consultation across days, because legal matters do not resolve in one sitting. Fourth, the ingestion of court records must remain entirely within Indian infrastructure, including the vision component that reads portal captchas.

VIM-1 is the system that introduces these four capabilities. It sits underneath the specialist layer described in the 2025 paper rather than replacing it, so that the conversational front-end users are already familiar with is unchanged. The rest of this paper describes each advance, the privacy architecture that surrounds the whole system, and the direction in which VIM-1 continues to evolve.

2. FOUR ADVANCES OVER THE 2025 SYSTEM

What VIM-1 adds

VIM-1 introduces four capabilities that together move the assistant from a turn-by-turn responder into an India-hosted legal intelligence that carries a matter across documents, across days, and entirely within Indian infrastructure.

Offline retrieval, grounded answers

VIM-1 answers common statutory questions and bare-act queries directly from our own legal corpus, without leaving Indian infrastructure. Retrieved passages ground the response, making answers auditable and citation-ready. For users this means faster replies; for the practice it means every answer can be traced back to a statutory source.

Persistent case comprehension

Users can upload FIRs, chargesheets, orders, notices, and agreements over the life of a matter. VIM-1 maintains a structured, continually-updated dossier for each case so that every subsequent question benefits from every prior document. The user no longer has to re-explain the matter on each return visit; the assistant already knows what is on file.

Conversation continuity across days

Every matter-focused conversation is a persistent session that progresses through a clear lifecycle of active, resting, and closed. When a user pauses a consultation and

returns hours or days later, VIM-1 resumes with the same understanding — no retelling, no lost context, no repeated disclosures.

Locally-solved court records

A proprietary vision component reads court-record portal captchas entirely within our own infrastructure, and a resilient caching layer keeps the court-record pipeline responsive even when upstream sources are slow. The court-data path stays fully inside Indian territory, from captcha to result.

3. OFFLINE RETRIEVAL, GROUNDED ANSWERS

Answering from our own corpus, without leaving India

A large share of incoming queries on an Indian legal assistant are structurally familiar: what does a given section of the criminal code say, what is the procedure for filing a particular kind of complaint, what are the rights of an arrestee, what is the limitation period for a civil claim of a given kind. These questions have settled, authoritative answers anchored in statutory text. VIM-1 answers them from a curated Indian-legal corpus — bare-act sections, canonical question-answer pairs, procedural checklists, and advocate-verified answers accumulated over months of production use — held entirely inside Indian infrastructure.

A learned representation over this corpus lets the assistant match each incoming query against the stored material before any generation is performed. When the match is strong, the corresponding advocate-verified answer is returned directly, with the exact statutory passages that ground it. When the match is partial, the retrieved passages are used as grounding for the generation that follows, so that the response is anchored to our own sources rather than to free-form recall.

For users this translates into two concrete benefits. Responses are faster, because the common path does not have to synthesise an answer from scratch. And responses are auditable, because every grounded answer carries the statutory text it relies on, which is the condition that lets a practising advocate verify the assistant's output in the way they would verify a junior's research note.

A self-learning loop closes over this layer: when the assistant produces a generated response that clears advocate review, the response is folded back into the corpus and becomes available to future queries as a direct answer. Over time the share of queries served from our own material grows, and the assistant becomes progressively more Indian-grounded in the register its users actually speak.

4. PERSISTENT CASE COMPREHENSION

The Case Dossier: cross-document understanding for a single matter

Legal matters arrive at the assistant through documents — an FIR photographed on a phone, a chargesheet forwarded from a lawyer, a bail order, a notice, an agreement. In the 2025 system each of these was treated as an isolated attachment whose content was summarised for the user and then left behind. VIM-1 promotes these uploads to members of a persistent Case Dossier that accumulates understanding across the life of a matter.

Each uploaded document is read, structured, and matched against the user's existing open cases. The key facts — parties and their roles, dates and their labels, courts and case numbers, the statutory sections invoked, the relief sought — are extracted and folded into a running composite understanding of the matter. The assistant does not store the document as a photograph to be re-read later; it stores a structured record that subsequent questions can benefit from.

The user-facing effect is immediate. A question asked on Thursday about an order received that morning is answered with knowledge of the first-information report that was uploaded on Monday. A question about the next procedural step is answered with reference to both the current order and the original cause of action. The user does not have to re-explain the matter; the assistant already knows what is on file.

For practising advocates the dossier serves a parallel role. When a senior reviews an assistant-drafted response, the dossier surface lets them see at a glance what the assistant is working from — which documents, which dates, which courts. Cross-document legal work, which is where junior mistakes most often occur, is supported rather than obscured.

5. CONVERSATION CONTINUITY

The Case Session: matter-focused consultations across days

Legal consultations are not single exchanges. A user sends a long message about a property dispute on Tuesday, returns a day later to ask a clarifying question, steps away for a week, and comes back with a fresh document. Treating every turn as a standalone event misses the matter that holds the turns together. VIM-1 therefore promotes matter-focused conversations to persistent Case Sessions that progress through a clear lifecycle of active, resting, and closed.

A session becomes active when the conversation turns substantive — when the user begins describing a matter in enough detail that subsequent turns clearly belong to the same discussion. A session enters a resting state when the user pauses without closing, so that the matter is remembered but the assistant is not holding the full conversational context in the foreground. A session closes when the user indicates resolution, or when the matter has not been touched for long enough that it is safe to archive.

When a user returns to a resting session — a day later, a week later — VIM-1 resumes with the same understanding. The assistant does not ask the user to re-summarise what was discussed. It already has a condensed understanding of the prior turns: the summary, the advice given, the open questions that were never resolved. The user simply continues the matter from where it paused.

Conversation continuity is the feature that allows VIM-1 to behave like a legal assistant that one has a relationship with, rather than like a stranger one explains a matter to each time. For consumer users, particularly those without access to a dedicated advocate, this is the difference between a one-shot query tool and a continuing legal companion.

6. LOCALLY-SOLVED COURT RECORDS

Keeping the court-data pipeline inside Indian infrastructure

The case-tracking and traffic-challan features depend on upstream court-record sources maintained by Indian government portals. These portals impose captchas as a standard anti-bot measure. In the 2025 system, captchas were solved by an external vision component, which introduced both a cost burden and a jurisdictional concern: the image of a court-record query briefly left Indian infrastructure on its way to the solver.

VIM-1 replaces this with a proprietary vision component trained specifically for the captcha styles used by Indian court-record portals, served entirely within our own infrastructure. A post-processing step handles systematic character confusions that are typical of low-resolution captchas. The assistant therefore reads court-portal captchas, fetches the underlying record, and returns results to the user without any part of that path leaving Indian territory.

A resilient caching layer sits in front of upstream sources so that repeated queries — a common pattern for hearing-date lookups and pending-challan checks — do not re-fetch identical records from already-strained government portals. Cached records are refreshed in the background as they age, which keeps WhatsApp round-trips responsive even when an upstream source is slow.

Geographic coverage spans twenty-two states of court-record retrieval through the virtual-courts ecosystem, with cross-state search configured so that a vehicle or party registered in one state is also looked up in the neighbouring states where related enforcement action is most commonly recorded. The user sees a single consolidated result; the fan-out work runs inside our infrastructure.

7. PRIVACY ARCHITECTURE

Privacy by design, not afterthought

VIM-1 is designed so that client-privileged information is protected by the shape of the system, not only by the policies that govern it. The data path is Indian from end to end.

India-Hosted Inference

VIM-1 is served from Indian infrastructure. The gateway, the retrieval layer, the case store, and the conversation store all run on Indian-territory servers.

No Cross-Border Data Path

Queries, uploaded documents, and conversation state do not traverse foreign jurisdictions on the hot path. Client-privileged information is protected by architecture, not by policy alone.

DPDP Act Alignment

Personal-data handling — identifiers, uploaded case documents, conversation history — is architected against the Digital Personal Data Protection Act 2023, with minimisation and purpose-limitation as defaults.

Audit-Friendly by Design

Because responses are grounded on retrievable statutory text, every answer carries a traceable provenance path. Advocate reviewers can verify what the assistant said and why in a way that closed black-box systems do not permit.

The retrieval layer, the case store, the conversation store, the vision component that reads court-portal captchas, and the caching layer in front of court-record sources all run on Indian-territory servers. Personal-data handling is architected against the Digital Personal Data Protection Act 2023, with minimisation, purpose-limitation, and user-directed deletion as defaults rather than as exceptions.

8. EVALUATION

What improved, qualitatively

As in the 2025 paper, evaluation is performed by practising advocates on sampled production traffic, rated on statutory correctness, procedural correctness, linguistic fidelity, and user-readiness. We prefer this protocol over crowdsourced labels because non-expert annotators mistake fluency for correctness on statutory material, and over frozen held-out benchmarks because the Indian statutory corpus itself is in transition — the Bharatiya Nyaya Sanhita and Bharatiya Nagarik Suraksha Sanhita are progressively replacing the Indian Penal Code and the Code of Criminal Procedure.

Under this protocol, VIM-1 is observed to improve on four dimensions. First, the style and statutory accuracy of responses to common legal questions became more consistent once grounded retrieval came online, because grounded answers are anchored in advocate-verified source text. Second, cross-document questions — "given the FIR and the chargesheet you have for my matter, what is my next step" — became directly answerable where they previously required the user to re-summarise both documents. Third, multi-day consultations no longer lose state between sessions: users return to a matter without having to restore context themselves. Fourth, court-record latency is lower and more predictable, because the captcha-to-record path stays within our infrastructure throughout.

Formal quantitative evaluation is ongoing. A held-out advocate-rated evaluation suite, with per-capability disaggregation, is under active construction and will be published as it matures. The current paper documents the deployed system and the qualitative changes that users and reviewing advocates have observed since VIM-1 came online.

9. WHAT COMES NEXT

The direction of travel

The VIM-1 programme continues along three directions. Language coverage is being broadened beyond the core English–Hindi–Hinglish register so that the assistant can serve users in the regional languages where legal help is most scarce. The on-device path is being deepened, so that the share of queries answered entirely inside Indian infrastructure continues to grow. Case-type coverage is being extended into commercial, family-law, and constitutional matters, with the same advocate-in-the-loop protocol that has shaped the system so far. None of these directions changes the basic architecture presented in this paper; each one reinforces it.

10. LIMITATIONS

Scope

⚠ **The system supports but does not replace a practising advocate.**

VIM-1 is a drafting, triage, and comprehension aid for litigants, junior advocates, and law students. On contested matters every user-facing response is framed as a starting point for human legal review. Formal quantitative evaluation is ongoing under the advocate-in-the-loop protocol described above.

HOW TO CITE

Citation

```
@techreport{gupta2026vim1,
  title        = {VIM-1: A Privacy-Preserving, India-Hosted, Domain-Specialised
                  Legal AI for WhatsApp-Native Delivery},
  author       = {Gupta, Mahir and Gupta, Piyush and {VirtualVakil Research Lab}},
  institution  = {VirtualVakil Research Lab},
  year         = {2026},
  month        = {April},
  number       = {TR-2026-01},
  url          = {https://virtualvakil.com/research},
}
```

Tap the **Cite** button near the top of the page to copy this BibTeX to your clipboard.

REFERENCES

References

[1] Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.

[2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP 2019.

[3] Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. NeurIPS 2022.

[4] Shi, B., Bai, X., & Yao, C. (2016). An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. IEEE TPAMI.

[5] Ministry of Electronics and Information Technology, Government of India. (2023). The Digital Personal Data Protection Act, 2023.

[6] Ministry of Law and Justice, Government of India. (2023). The Bharatiya Nyaya Sanhita (BNS) and Bharatiya Nagarik Suraksha Sanhita (BNSS).

[7] Supreme Court of India. (2024). National Judicial Data Grid: Pendency statistics, accessed April 2026.

Previous Research

The August 2025 foundational paper describes the eight role-specialised agents that VIM-1 continues to build on.

Read 2025 Paper

View full archive