Can AI Think Like a Lawyer? Evaluating Generative AI in South African Law

D Thaldar*

Online ISSN
1727-3781

P·E·R
Pioneer in peer-reviewed,
open access online law publications

Author

Donrich Thaldar

Affiliation

University of KwaZulu-Natal,
South Africa

Email

ThaldarD@ukzn.ac.za

Date Submitted

17 March 2025

Date Revised

15 August 2025

Date Accepted

15 August 2025

Date Published

16 October 2025

Editor

Prof Germarié Viljoen

Journal Editor

Prof Wian Erlank

How to cite this contribution

Thaldar D "Can AI Think Like a Lawyer? Evaluating Generative AI in South African Law" *PER / PELJ* 2025(28) - DOI http://dx.doi.org/10.17159/1727-3781/2025/v28i0a21584

Copyright

DOI
http://dx.doi.org/10.17159/1727-3781/2025/v28i0a21584

Abstract

Artificial intelligence (AI) is increasingly being explored as a tool for legal research and reasoning, yet its effectiveness in applying South African legal principles remains under-examined. This study evaluates the performance of five generative AI models - ChatGPT 4o, Claude 3.7 Sonnet, DeepSeek R1, Gemini 2.0 Flash, and Grok3 Beta - across three private law scenarios involving *actio de pauperie*, *negotiorum gestio* and *actio legis Aquiliae*. Each AI model's response was assessed against seven criteria: identification of the correct legal action, accuracy of legal requirements, application to facts, case law citation, relevance of case law, consideration of defences, and clarity of the final answer. The findings reveal that while AI models generally identify and apply South African legal principles correctly, their performance varies significantly. Claude performed the strongest overall, demonstrating structured legal reasoning and engagement with statutory provisions, while ChatGPT followed closely but was undermined by hallucinated case law. DeepSeek provided sound reasoning but occasionally misapplied legal principles. Gemini and Grok were the weakest, with incomplete legal analyses and limited case law engagement. A key limitation across all models was the unreliable retrieval and application of case law, with frequent misinterpretations and fabricated references. Additionally, most models failed to incorporate statutory law unless explicitly prompted. These results underscore the potential of AI as a supplementary legal tool while highlighting its current limitations. Future research should explore AI's competency in broader areas of South African law, including statutory interpretation and constitutional analysis, to better understand its role in legal practice and academia. While AI can assist legal professionals, human oversight remains essential to ensure doctrinal accuracy and case law reliability.

Keywords

AI in law; AI legal reasoning; case law citation accuracy; legal research automation.

.....

1 Introduction

Artificial intelligence (AI) is increasingly shaping legal practice worldwide, raising both optimism and scepticism about its capabilities in different legal systems. While AI-powered tools offer efficiency in legal research, document drafting, and analysis, concerns persist about their accuracy, particularly in jurisdictions with complex legal traditions such as South Africa. This study critically examines the extent to which general-purpose generative AI models can effectively interpret and apply South African law.

The question of AI's ability to engage with South African legal principles has become particularly relevant following recent court cases that highlight both its potential and its risks. In *Makunga v Barlequins Beleggings*,[1] the Western Cape High Court recognised the utility of AI tools when a self-represented litigant used AI-generated submissions that were commended for their quality. Conversely, in *Parker v Forsyth*[2] the Johannesburg Regional Court cautioned against the dangers of AI-generated misinformation, reprimanding legal representatives for submitting fabricated case law sourced from AI tools. The court emphasised that AI should not replace human diligence in legal research. Similarly, *Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs KZN*[3] reinforced this principle, stressing that legal practitioners bear ultimate responsibility for verifying AI-generated outputs.

While these cases focus on AI's role in legal practice, they raise broader questions about AI's ability to engage meaningfully with South African law. If AI is to be effectively used in legal research and practice, it must be able to correctly identify legal principles, apply them accurately and cite case law reliably. However, AI models differ significantly in their approach to retrieving legal information. Some, such as ChatGPT and Claude, rely solely on their inherent knowledge - drawing upon data available up to the point of their last training update - while others, such as Grok, have real-time internet access and can retrieve case law, legislation and legal commentary at the time of a query. This distinction raises an important methodological consideration: to what extent are AI models relying on their internalised knowledge, and to what extent can they supplement gaps through online research? The ability to retrieve external legal sources could provide an

* Donrich Thaldar. BLC LLB MPPS PhD. Professor, University of KwaZulu-Natal. E-mail: ThaldarD@ukzn.ac.za. ORCiD https://orcid.org/0000-0002-7346-3490.

1 *Makunga v Barlequins Beleggings (Pty) Ltd t/a Indigo Spur* (19733/2017) [2023] ZAWCHC 332 (1 December 2023).

2 *Parker v Forsyth* (1585/20) [2023] ZAGPRD 1 (29 June 2023).

3 *Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs KwaZulu-Natal* 2025 3 SA 534 (KZP).

advantage, but it also introduces risks, such as a reliance on unverified or non-authoritative materials.

Recent empirical studies have sought to assess how well AI models perform in legal reasoning and case law application. The study by Bhambhoria *et al*[4] provides one of the most comprehensive evaluations to date, testing multiple AI models on their ability to correctly answer legal questions and apply legal principles across different jurisdictions. Their findings highlight substantial variation in AI performance, with some models demonstrating strong logical reasoning but struggling with jurisdiction-specific nuances. Notably, AI models frequently misapplied legal precedents, omitted key legal elements, or defaulted to overly generic reasoning. These deficiencies were particularly pronounced in jurisdiction-specific applications, where AI models struggled to account for the contextual and layered nature of legal authority. The study underscores the necessity of jurisdictional adaptation in AI legal research and highlights the risks of relying on models that have not been sufficiently trained on local legal materials.

Complementing this, Magesh *et al*[5] found that even legal AI tools specifically designed for case law retrieval, such as those developed by LexisNexis and Thomson Reuters, continue to generate hallucinated case citations in approximately 17% of responses. While retrieval-augmented generation has reduced the frequency of completely fabricated cases, these tools still frequently misapply legal precedents, leading to misleading but plausible-sounding citations. This finding reinforces the need for legal practitioners and researchers to independently verify AI-generated references, as misplaced reliance on these tools could lead to serious legal consequences.

In the South African context, existing research has predominantly focussed on the legal and policy implications of AI rather than on assessing AI's ability to function as a legal technology. Studies by Sihlahla *et al*[6] and Bottomley and Thaldar[7] explore legal liability for harm caused by AI, while Naidoo *et al*[8] propose policy reforms to better regulate AI in healthcare. Mtuze and Morige[9] go further, advocating a *sui generis* AI regulatory framework for South Africa. However, there has been no systematic analysis of general-purpose generative AI models' ability to engage with South African law at a doctrinal level, leaving a gap in understanding how these models perform when applied to local legal principles.

4 Bhambhoria *et al* 2024 https://arxiv.org/abs/2404.12349.
5 Magesh *et al* 2024 https://arxiv.org/abs/2405.20362.
6 Sihlahla *et al* 2025 *Developing World Bioethics* 35-45.
7 Bottomley and Thaldar 2023 *Frontiers in Pharmacology*.
8 Naidoo *et al* 2022 *South African Journal of Bioethics and Law* 11-16.
9 Mtuze and Morige 2024 *Obiter* 161-179.

A notable recent initiative is the development of a legal chatbot[10] by researchers at the University of KwaZulu-Natal's School of Law.[11] Unlike general-purpose generative AI models, this chatbot has been specifically trained on the data-sharing laws of 12 African countries, including South Africa, to assist researchers in navigating complex regulatory frameworks. Early reports indicate that it has successfully helped facilitate the legal transfer of datasets between jurisdictions, demonstrating AI's potential in legal compliance tasks. However, the focus of this project differs fundamentally from that of the present study, as it involves a domain-specific AI trained by legal researchers for a targeted regulatory function. It does not assess how untrained, publicly available generative AI models engage with South African law in a doctrinal sense, which remains an open question.

Given these challenges, it is critical to assess how general-purpose generative AI models perform when applied to South African law. This article addresses this gap by conducting a systematic evaluation of five prominent general-purpose generative AI models on their ability to handle South African legal scenarios. By evaluating their responses against established legal criteria, this research provides insights into their strengths and limitations, contributing to the discourse on AI's role in South African law.

2 Methodology

To ensure a meaningful evaluation of AI's ability to engage with South African legal principles, three legal actions within private law were selected. Specifically, the study focussed on legal actions in the law of delict and unjustified enrichment. Private law disputes involving contract law typically follow straightforward principles: there must be an agreement, and legal consequences flow from that agreement. Similarly, in public law disputes - such as those involving administrative justice or access to information - the relevant legal action is usually grounded in statutory provisions, and AI's task would merely involve identifying and interpreting the applicable statute.

By contrast, delictual and unjustified enrichment claims originate in South Africa's common law, requiring a more nuanced understanding of legal principles beyond mere statutory interpretation. These areas of law involve various possible actions, each with distinct requirements that must be met, based on the cause of action. Additionally, unjustified enrichment claims are frequently confused with delictual claims, making them particularly intricate.

10 DS-I Africa Law Research Group 2024 https://www.datalaw.bot.
11 Nordling 2024 https://doi.org/10.1038/d44148-024-00249-w.

Given that this body of law is not as easily searchable online as statutory provisions in public law, it provides a more challenging test for AI models.

Based on these considerations, three legal scenarios were constructed, each focussing on the following causes of action:

- *Actio de pauperie* (liability for harm caused by domesticated animals);
- *Actio negotiorum gestorum* (management of another's affairs without mandate); and
- *Actio legis Aquiliae* (delictual liability for financial loss caused by wrongful and negligent conduct).

For ease of reference, each of the three scenarios is presented in full within a textbox in the results analysis section. Each scenario was carefully designed to specify a South African location, ensuring that the AI models recognised the applicability of South African law. Furthermore, each scenario concluded with a legal question - whether the relevant party had a claim against the other and on what grounds. AI models were instructed to limit their responses to 300 words and to cite relevant case law.

The scenarios were presented one at a time to five prominent generative AI models - ChatGPT 4o,[12] Claude 3.7 Sonnet,[13] DeepSeek R1,[14] Gemini 2.0 Flash[15] and Grok3 Beta.[16] After a response to each scenario had been received, a new session was initiated before the next one was presented. This ensured that each scenario was treated as a separate legal problem, preventing the AI from drawing on prior interactions. The responses of all five AI models were then transcribed into separate Word documents, which are included as supplementary material. All of this interaction took place on 1 March 2025.

Each AI-generated response was subsequently evaluated against the following seven criteria:

1. Correct Identification of the cause of action: Did the AI correctly identify the relevant cause of action or legal doctrine applicable to the scenario?
2. Accurate description of legal requirements: Did the AI provide a correct and comprehensive explanation of the elements required to establish the claim?

12 OpenAI 2025 https://chat.chatbotapp.ai/?model=gpt-4o.
13 Anthropic 2025 https://www.anthropic.com/claude/sonnet.
14 DeepSeek-AI 2025 https://www.deepseek.com.
15 Google DeepMind 2025 https://gemini.google.com/app.
16 xAI 2025 https://grok.com/?referrer=website.

3. Application to the facts: Did the AI correctly apply the legal principles to the specific facts of the case?
4. Case law references: Did the AI cite case law, and were the citations correct and complete?
5. Relevance of case law: Was the case law cited relevant to the issue at hand?
6. Consideration of defences or counter-arguments: Did the AI anticipate potential defences or counter-arguments that could be raised?
7. Clarity and accuracy of the final answer: Was the response logically structured, clear, and legally sound?

3 Results and analysis

3.1 Scenario 1

Scenario 1: The Dachshund's Unexpected Bite

Sipho Mthembu, a schoolteacher from Pietermaritzburg, was walking through his neighbour's property to visit his friend, Lerato Dlamini, for afternoon tea. As he strolled along the paved walkway, he noticed his shoelace had come undone. Not wanting to trip, he stopped, bent down, and calmly tied his shoe.

Unbeknownst to him, Lerato's dachshund, Boetie, had been watching him closely. For reasons known only to the little sausage dog, Boetie suddenly lunged forward and bit Sipho's hand - completely unprovoked. Shocked and bleeding slightly, Sipho pulled back and shouted in pain.

Sipho now wants to claim compensation from Lerato for his medical costs and pain. Does Sipho have a claim against Lerato? If so, on what grounds?

Write an answer of 300 words, and make sure to refer to relevant case law.

3.1.1 Identification of the cause of action

All five AI models correctly identified *actio de pauperie* as the applicable cause of action for this scenario. This is a positive outcome, as it demonstrates that the models were able to recognise the appropriate cause of action under South African law without being explicitly prompted with its name.

3.1.2 Accuracy of the legal requirements

All the AI models correctly outlined the legal elements of *actio de pauperie*, namely that liability arises on the part of the owner when a domesticated animal causes harm in a manner *contra naturam sui generis*, meaning contrary to the nature of its species.[17] All but DeepSeek noted that the owner's negligence is not a requirement for liability, as the action is based on ownership rather than fault. However, DeepSeek introduced an error by suggesting that negligence could be a factor in determining liability, which is incorrect under *actio de pauperie*. This inaccuracy resulted in a mischaracterisation of the legal basis for liability and led to difficulties in correctly applying the principles later in its response.

3.1.3 Application to the facts

Most models correctly applied the legal requirements to the facts of the case. ChatGPT, Claude, Gemini and Grok all explicitly confirmed that ownership was established, that the dog's attack was unprovoked and therefore *contra naturam sui generis*, and that there was a causal link between the attack and the harm suffered. DeepSeek, however, applied its incorrect interpretation of negligence to the scenario, which led to a flawed reasoning structure. Because it misrepresented the fundamental elements of *actio de pauperie*, its subsequent analysis did not align with the correct legal framework.

3.1.4 Case law references

All five AI models cited *O'Callaghan v Chaplin*,[18] which is the *locus classicus* for *actio de pauperie*. ChatGPT, Claude, Gemini and Grok all referenced the case correctly, while Gemini cited only the case name. ChatGPT further cited the recent case *Van Meyeren v Cloete*,[19] providing additional context on potential defences. Claude and Grok also mentioned *Lever v Purdy*,[20] though Claude misrepresented its significance, framing it as affirming liability under *actio de pauperie* regardless of negligence, when in reality the case focussed on whether third-party negligence could serve as a defence. Grok correctly summarised the content of the case but provided an incorrect year for the case.

3.1.5 Relevance of case law

The case law cited by ChatGPT, Claude, Gemini and Grok was generally relevant to the issue at hand, particularly *O'Callaghan v Chaplin*, which is a

---

17 *O'Callaghan v Chaplin* 1927 AD 310; *Da Silva v Otto* 1986 3 SA 538 (T); *Portwood v Svamvur* 1970 4 SA 8 (RA).

18 *O'Callaghan v Chaplin* 1927 AD 310.

19 *Van Meyeren v Cloete* (636/2019) [2020] ZASCA 100 (11 September 2020).

20 *Lever v Purdy* 1993 3 SA 17 (A).

foundational case for *actio de pauperie*. ChatGPT's reference to *Van Meyeren v Cloete* added further depth by discussing defences. However, DeepSeek's citation was entirely irrelevant, citing an entirely unrelated 19th-century case[21] that demonstrates a clear difficulty in retrieving legally appropriate precedents.

3.1.6 Consideration of defences or counterarguments

Three models - ChatGPT, Claude, and Gemini - considered potential defences to the claim. ChatGPT was the most comprehensive in this regard, discussing provocation (*Van Meyeren v Cloete*), trespassing[22] and third-party fault.[23] Claude and Gemini also acknowledged provocation and trespassing but did not explicitly mention third-party fault. DeepSeek correctly suggested that an assumption of risk defence could apply.[24] Grok discussed possible defences but framed them in a broader legal context rather than specifically within *actio de pauperie*.

3.1.7 Clarity and accuracy of the final answer

Four models - ChatGPT, Claude, Gemini and Grok - provided clear, well-reasoned conclusions affirming that Sipho does have a valid claim under *actio de pauperie*. Their answers were structured logically, leading from the legal principles to the application of facts before arriving at the final determination. DeepSeek's answer, however, was weakened by its erroneous reference to negligence, which introduced an unnecessary complexity that detracted from the clarity of its conclusion.

3.1.8 Conclusion on Scenario 1

Overall, ChatGPT provided the strongest response, correctly identifying the legal action, accurately describing and applying its requirements, citing and applying relevant case law correctly, providing a clear conclusion, and comprehensively considering potential defences. Claude and Gemini also performed well but were slightly weaker in their consideration of defences. Grok was mostly accurate but provided an incorrect citation for *Lever v Purdy*, which slightly weakened the precision of its response. DeepSeek faced the greatest challenges, as its initial misidentification of negligence as a relevant element resulted in incorrect reasoning throughout its response. Additionally, its selection of an unrelated case further illustrates its difficulty in retrieving and applying appropriate South African legal precedents. This analysis suggests that while generative AI models can produce legally sound responses in a South African context, variations in the accuracy of

21 *Mackintosh v Mackintosh* 1864 2 M 1357.
22 Regarding trespassing as a defense, see *O'Callaghan v Chaplin* 1927 AD 310.
23 Regarding third-party fault as a defence, see *Lever v Purdy* 1993 3 SA 17 (A).
24 Regarding *volenti non fit iniuria* as a defence, see *Joubert v Combrinck* 1980 3 SA 680 (T).

their foundational legal knowledge and case law retrieval capabilities affect their overall reliability.

3.2 Scenario 2

Scenario 2: The Unwanted Storage Fees

Thabo Nkosi, a businessman in Pretoria, had his Toyota Hilux serviced at Mpho's Auto Repairs. After a full service, including an oil change and brake check, the garage called him to collect the car.

Since Thabo was busy in meetings Mpho, the owner, kindly arranged for one of his mechanics to drop the car off at Thabo's house. But when the car arrived, Thabo took one look at it and refused to accept it, claiming the car still made a strange noise. He told the driver to take it back, despite Mpho's assuring him that everything was in order.

Not wanting to leave the vehicle out in the open, Mpho decided to store it safely in a secure lot at his own cost. After a several months Thabo changed his mind and called to say he wanted his car back. However, Mpho insisted that before he returned the Hilux Thabo needed to reimburse him for the storage costs, arguing that Thabo had benefitted from his efforts in safeguarding the vehicle.

Thabo refused to pay, claiming he never agreed to the storage. Does Mpho have a claim against Thabo for the storage costs? If so, on what grounds?

Write an answer of 300 words, and make sure to refer to relevant case law.

3.2.1 Identification of the cause of action

The relevant legal doctrine in this scenario is *negotiorum gestio*. Among the five AI models, only Claude correctly identified *negotiorum gestio* without introducing alternative causes of action. ChatGPT and DeepSeek identified *negotiorum gestio* but also referenced unjustified enrichment, creating doctrinal uncertainty as to which elements should be applied.

Gemini and Grok not only failed to identify *negotiorum gestio* but also suggested two alternative causes of action: unjustified enrichment and implied contract. The reference to unjustified enrichment is incorrect, as there is no general enrichment claim in South African law. The reference to implied contract, however, is a more complex issue. It is legally plausible to consider implied contract as an alternative claim in this context, but only

under certain circumstances. Implied contract could be relevant if it could be established that Thabo was aware that Mpho had placed the vehicle in storage and, by failing to object or intervene, tacitly accepted the benefit. If Thabo was not ignorant of the management of his affairs, *negotiorum gestio* would not apply, but liability could still arise under implied contract.

However, neither Gemini nor Grok engaged in this line of reasoning. Instead, they seemed to assume that Thabo's refusal to accept the vehicle was sufficient to create an implied contractual obligation. This is incorrect. A refusal to accept the return of an object does not in itself establish a contract, nor does it constitute tacit acceptance of services provided as a consequence of that refusal. The models' misapplication of implied contract principles further contributed to their doctrinal confusion.

3.2.2 Accuracy of the legal requirements

The first requirement of *negotiorum gestio* is that the *gestor* must have managed the affairs of another.[25] Claude, ChatGPT and DeepSeek explicitly identified this requirement, correctly recognising that *negotiorum gestio* applies to situations where one party has taken steps to manage another's affairs. Gemini and Grok, however, mischaracterised the scenario in terms of unjustified enrichment and failed to identify the element at all. Instead of recognising that the doctrine applies to situations of voluntary management, they focussed on the benefit derived by the *dominus*, which is not a requirement of *negotiorum gestio*.

The second requirement is that the *dominus* must have been unaware that its affairs were being managed.[26] None of the AI models explicitly stated this requirement. Claude, ChatGPT and DeepSeek only implied this element but failed to articulate it as a distinct requirement. Gemini and Grok did not address this element in any form, as they were operating under the incorrect assumption that unjustified enrichment was the basis for the claim.

The third requirement is that the *gestor* must have had the intention to manage another's affairs, including the intention to claim reimbursement for necessarily or usefully incurred expenses.[27] Claude explicitly recognised the requirement of intention but framed it incorrectly by referring to a general intent to act in the interest of the *dominus* rather than an intention to claim reimbursement. ChatGPT and DeepSeek implied this requirement but failed to articulate it properly. Gemini and Grok did not engage with this element at all, as their responses were based on unjustified enrichment rather than *negotiorum gestio*. Instead of addressing the *gestor's* intent, they replaced

25 *Immaculate Truck Repairs CC v Capital Acceptances Ltd* (1153/2014) [2016] ZAFSHC 3 (14 January 2016) (hereafter *Immaculate Truck Repairs*) para 22.
26 *Immaculate Truck Repairs* para 22.
27 *Immaculate Truck Repairs* para 22.

it with a focus on whether Mpho had suffered impoverishment, which is not a relevant factor under *negotiorum gestio*.

The fourth requirement is that the management of the affairs of the *dominus* should have been conducted in a reasonable way.[28] All five AI models explicitly identified this element, although Gemini and Grok did so in a distorted form by linking reasonableness to unjustified enrichment rather than to *negotiorum gestio*. While their references to reasonableness were technically present, they were not framed within the correct legal doctrine.

Claude provided its own list of four requirements, but only one fully aligned with the correct elements, while another partially aligned, and the remaining two did not align at all. The element that fully aligned was reasonableness, which it correctly identified as a necessary requirement. The element that partially aligned was the requirement that the *gestor* must have managed the affairs of another, which Claude framed in terms of acting without a mandate. However, its framing suggested that the absence of a mandate was sufficient on its own, without addressing that the *gestor* must also have actively taken over the management of the affairs of the *dominus*. Claude's other two elements - acting in Thabo's interest and the benefit derived by Thabo - were incorrect and introduced legal confusion.

ChatGPT and DeepSeek's lists of requirements were incomplete. ChatGPT referred to *negotiorum gestio* but immediately followed this with unjustified enrichment, making it unclear which elements were actually relevant. Its formulation of the requirements was vague, listing voluntariness, necessity, and acting in the interest of the *dominus*, but it did not explicitly identify the requirement that the *dominus* must have been unaware or that the *gestor* must have intended to claim reimbursement. DeepSeek made similar errors, focussing on reasonableness and usefulness but failing to properly outline the requirement of the *gestor's* intention.

Gemini and Grok were entirely incorrect in their formulations, as their analyses were grounded in unjustified enrichment rather than *negotiorum gestio*. They set out elements such as impoverishment, enrichment at another's expense, and an unjustified transfer of benefit, none of which apply to *negotiorum gestio*. As a result, their responses failed to engage with any of the four requirements.

3.2.3 Application to the facts

This section assesses whether the AI models applied the facts logically to the legal elements they identified, even where those elements were

---

28 *Immaculate Truck Repairs* para 22.

incorrect. The focus is on whether they demonstrated an understanding of the factual scenario and applied it coherently to their chosen framework.

Claude applied the facts in a structured manner, linking Mpho's actions to the elements it proposed. While it incorrectly framed intent as acting in Thabo's interest rather than intending to claim reimbursement, at least it assessed whether Mpho's intervention was beneficial. It also connected the absence of a mandate to its misinterpretation of *negotiorum gestio*, making its application internally coherent, though legally flawed.

ChatGPT and DeepSeek introduced uncertainty by referencing unjustified enrichment alongside *negotiorum gestio*, but they nevertheless applied the facts logically to their own confused framework. ChatGPT assessed voluntariness and necessity, considering whether Mpho acted without obligation and whether storage was required. DeepSeek focussed on usefulness, evaluating whether Mpho's actions ultimately benefitted Thabo. While both misunderstood the legal doctrine, they followed a clear reasoning process in applying the facts.

Gemini and Grok failed entirely to apply the facts coherently, even within their own misidentified framework. By relying on unjustified enrichment they assessed Mpho's financial loss without properly linking it to an enrichment on Thabo's part. They assumed that because Mpho incurred costs, Thabo was enriched, which does not follow logically. Their factual analysis was not only legally incorrect but also internally inconsistent.

Claude, ChatGPT and DeepSeek demonstrated an ability to process facts and apply them within their flawed formulations, while Gemini and Grok lacked even internal coherence in their application.

3.2.4 Case law references

Each AI model cited at least two cases, but there was very little overlap between the cases referenced. This suggests that the models retrieved case law based on different underlying algorithms or datasets. Of particular concern is that ChatGPT cited a non-existent case, "*Pascoal v Wurdemann* 1908 TS 201", which raises reliability issues regarding its legal research capabilities. While the other models cited real cases, the accuracy of their application of these cases varied considerably.

3.2.5 Relevance of case law

Across the AI responses, a common issue was the misinterpretation of case law principles. Some models, such as Claude and DeepSeek, framed legal principles too broadly, attributing generalised, unjustified enrichment doctrines to cases that were far more specific in their rulings. For example,

Claude's citation of *McCarthy Retail Ltd v Shortdistance Carriers CC*[29] did not accurately reflect the judgment's actual emphasis on necessary and useful expenses rather than the proactive protection of property. Similarly, *Gouws v Jester Pools*[30] was misapplied by both Claude and Grok, with each AI incorrectly suggesting that the case supported a broad principle of compensating service providers who confer a benefit. DeepSeek's use of *Nortje v Pool*[31] was another example of misinterpretation, as the case explicitly rejected a general enrichment action in South African law rather than supporting a broad right to claim reimbursement for beneficial expenditures.

Gemini and Grok similarly cited cases whose holdings did not fully support the principles they were claimed to establish. Gemini's reference to *Kommissaris van Binnelandse Inkomste v Willers*[32] did involve unjustified enrichment, but it did not emphasise the principle that "enrichment must be without legal cause" in the broad sense that the AI suggested. Similarly, Grok's reference to *Conradie v Rossouw*[33] overstated the case's stance on implied agreement, applying it in a context where it was not directly relevant.

Although each AI model cited at least two cases and some cases were correctly interpreted, there was a high frequency of misinterpretation. The primary issue was not necessarily the choice of cases but rather the way in which the AI models extracted and applied legal principles. This suggests a broader limitation in their ability to engage with nuanced judicial reasoning, often attributing broader doctrinal significance to cases than they actually support.

3.2.6 Consideration of defences or counterarguments

Claude considered Thabo's explicit rejection of the vehicle as a potential defence, suggesting that this could create a dispute over whether Mpho had the right to store it at his own expense and then claim reimbursement. It also speculated that the extended period of storage could weaken Mpho's claim if he did not take reasonable steps to resolve the situation earlier. While there is no requirement that a *gestor* must repeatedly attempt to return the property, Claude at least engaged with the issue of whether prolonged storage could become unreasonable. This is a thought-provoking argument, though ultimately unlikely to succeed as a defence.

ChatGPT and DeepSeek considered Thabo's lack of explicit consent but framed it in terms of unjustified enrichment rather than *negotiorum gestio*.

---

29 *McCarthy Retail Ltd v Shortdistance Carriers CC* 2001 3 SA 482 (SCA).
30 *Gouws v Jester Pools (Pty) Ltd* 1968 3 SA 563 (T).
31 *Nortje v Pool* 1966 3 SA 96 (A).
32 *Kommissaris van Binnelandse Inkomste v Willers 1994 3 SA 283 (A).*
33 *Conradie v Rossouw* 1919 AD 276.

They assumed that because Thabo eventually reclaimed the vehicle, he accepted the benefit and should compensate Mpho. Their approach was legally flawed because later enrichment does not retroactively satisfy the requirements of *negotiorum gestio*.

Gemini and Grok introduced implied contract but did not frame it as a defence. Instead, they treated it as an alternative cause of action. Their reasoning, however, was flawed. They implied that Thabo's refusal to accept the vehicle itself gave rise to an implied contractual obligation. This is incorrect. A refusal does not amount to tacit acceptance of liability for subsequent expenses. A more legally valid counterargument - one that neither Gemini nor Grok considered - would have been to argue that Thabo was aware that Mpho had stored the vehicle, meaning that he was not ignorant of the management of his affairs. If this had been the case, *negotiorum gestio* would not apply, but Thabo could still be liable under an implied contract theory. However, this would require factual speculation, as there is no indication in the scenario that Thabo had such knowledge.

Claude was the only model to provide a counterargument that, while unlikely to succeed, at least engaged with a potential weakness in Mpho's claim within the *negotiorum gestio* framework. ChatGPT and DeepSeek misunderstood the issue by assuming later enrichment was sufficient to justify reimbursement. Gemini and Grok fundamentally misapplied the principles of contract law, treating implied contract as arising from a refusal rather than from tacit acceptance. None of the models correctly analysed the strongest potential counterargument: that if Thabo was aware of the storage, *negotiorum gestio* would not apply, but liability could still arise under implied contract.

3.2.7 Clarity and accuracy of the final answer

Claude was the only AI model that provided an accurate final answer, correctly concluding that Mpho had a claim against Thabo under *negotiorum gestio*. However, its reasoning contained some ambiguities, particularly in its discussion of intent and the reasonableness of prolonged storage, which, while thought-provoking, introduced unnecessary complexity. Nevertheless, its final answer remained clear and legally structured.

ChatGPT, DeepSeek, Gemini and Grok did not provide accurate final answers. They either misidentified the legal doctrine or incorrectly framed the claim under unjustified enrichment or implied contract. Their conclusions were therefore legally incorrect, and their responses lacked the necessary precision to be considered valid assessments of the scenario.

3.2.8 Conclusion on Scenario 2

Claude provided the most legally accurate and structured response, correctly identifying *negotiorum gestio* and concluding that Mpho had a claim against Thabo. However, its reasoning introduced some unnecessary complexity, particularly in its discussion of intent and the reasonableness of prolonged storage. ChatGPT and DeepSeek created doctrinal confusion by referencing unjustified enrichment alongside *negotiorum gestio*, leading to an unclear assessment of the claim. Gemini and Grok failed entirely to identify the correct doctrine, instead misapplying unjustified enrichment and implied contract, which resulted in legally incorrect conclusions. Across all responses there were significant issues in the identification of the legal doctrine, the articulation of the correct legal requirements, and the logical application of the facts. While some models demonstrated internal coherence, none fully engaged with the strongest potential counterargument - that Thabo's awareness of the storage could have displaced *negotiorum gestio* in favour of implied contract. The engagement of the AI models with case law was also inconsistent, with frequent misinterpretations of judicial reasoning.

3.3 Scenario 3

Scenario 3: Fire in the Midlands

On a small farm in the rolling hills of the KwaZulu-Natal Midlands, Jacob Zungu, a cattle farmer, was preparing his fields for winter grazing. As was common practice in the area, he decided to burn some of the old dry grass to clear space for new growth.

Meanwhile on the neighbouring farm Maria van der Merwe had a herd of prized Nguni cattle, known for their rich, patterned coats. She had warned Jacob many times to be careful with his fire-making, as the wind often carried embers across their shared fence line.

Ignoring Maria's warnings, Jacob lit his fire on a windy afternoon. Within minutes a strong gust carried burning embers into Maria's field, setting the dry grass alight. The flames spread rapidly, destroying part of her grazing land and injuring two of her cattle. The damage was severe, and Maria demanded compensation for her loss.

Jacob insisted that the fire was an accident and refused to compensate her. Does Maria have a claim against Jacob? If so, on what grounds?

Write an answer of 300 words, and make sure to refer to relevant case law.

3.3.1 Identification of the cause of action

All five AI models recognised that delictual liability applies to the scenario. However, only ChatGPT, Claude and DeepSeek explicitly identified *actio legis Aquiliae* as the correct cause of action, demonstrating a stronger grasp of South African legal principles. Gemini and Grok framed the claim more generally as negligence under delict, failing to specify *actio legis Aquiliae*. This omission is significant because it overlooks the structured requirements of South African delictual liability for patrimonial loss.

Given that this case involves a fire, it is crucial not only to recognise *actio legis Aquiliae* but also to acknowledge the applicability of the *National Veld and Forest Fire Act* (NVFFA).[34] This Act establishes statutory duties of care on landowners regarding fire prevention and introduces a statutory presumption of negligence, both of which significantly influence how *actio legis Aquiliae* should be applied in fire-related cases.

The NVFFA creates a statutory duty of care, reinforcing the wrongfulness element in a delictual claim. Section 12(1) requires landowners to maintain firebreaks to prevent the spread of veld fires. Failure to comply with this requirement strengthens the wrongfulness argument in Maria's claim. Section 25(5) further establishes wrongfulness by making it an offence for a landowner not to take reasonable steps to contain a fire on its property.

The NVFFA also introduces a statutory presumption of negligence, which directly affects the fault element of a delictual claim. Section 34(1) states that if a veld fire spreads from a landowner's property and causes damage, that landowner is presumed to have been negligent unless able to prove otherwise. This provision shifts the burden of proof onto the defendant, making it more difficult for Jacob to argue that he was not at fault. However, section 34(1) provides an exception: if the landowner is a member of a registered fire protection association, the landowner is not automatically presumed negligent. Jacob's liability could therefore depend on whether he belongs to a registered fire protection association and whether he can show that he took reasonable precautions.

Claude was the only AI model that identified the relevance of the NVFFA, correctly recognising that it imposes a statutory duty of care and creates a presumption of negligence. ChatGPT, DeepSeek, Gemini and Grok failed to mention the NVFFA, meaning they did not fully engage with the statutory framework governing fire-related liability. This is a critical oversight because the NVFFA directly strengthens Maria's claim under *actio legis Aquiliae*. Additionally, Grok incorrectly referred to a "delict of negligence", displaying

34 *National Veld and Forest Fire Act 101 of 1998.*

a fundamental misunderstanding of South African law by conflating delict with common law tort principles.

While all models recognised delictual liability, only ChatGPT, Claude and DeepSeek correctly identified *actio legis Aquiliae* as the applicable cause of action. Claude provided the most complete answer by recognising the NVFFA's role, which significantly strengthens Maria's claim. The other models weakened their responses by failing to incorporate the statutory framework governing fire liability. Gemini and Grok's omission of *actio legis Aquiliae*, and Grok's incorrect reference to a "delict of negligence", further undermined their responses.

3.3.2 Accuracy of the legal requirements

The *actio legis Aquiliae* requires proof of five elements: conduct, wrongfulness, fault (in this case, negligence), causation (both factual and legal), and patrimonial loss.[35] The AI models varied in their ability to correctly identify and articulate these elements in the abstract. ChatGPT and DeepSeek provided the most structured and comprehensive identification of all five elements, clearly distinguishing between them. Claude also identified all five elements but presented them in a more case law-driven manner rather than as a doctrinally structured analysis. Gemini and Grok identified the core elements of delict but in a less systematic way. Grok's response was the weakest, as it lacked precision in distinguishing between the different elements.

Beyond merely listing the elements, a doctrinally sound response should also recognise statutory modifications to common law requirements where applicable. As this claim arises from a veld fire, the NVFFA imposes statutory duties and presumptions that influence the legal assessment. The most significant statutory modification is the presumption of negligence in section 34(1), which shifts the burden of proof onto the defendant. Additionally, section 12(1) establishes a statutory duty of care regarding fire prevention, which strengthens the wrongfulness element in a delictual claim.

Among the AI models only Claude incorporated these statutory considerations into its abstract analysis of delictual liability. It correctly acknowledged the presumption of negligence and the statutory duty of care, thereby demonstrating a more integrated understanding of how South African law modifies the common law delictual framework in fire-related cases. ChatGPT, DeepSeek, Gemini and Grok, while correctly identifying the five elements of *actio legis Aquiliae*, failed to consider how the NVFFA

35 *Evrigard (Pty) Ltd v Select PPE (Pty) Ltd* (2022-22743) [2024] ZAGPJHC 183 (26 February 2024) para 45.

alters the standard application of these elements. As a result, their analyses remained incomplete, lacking the necessary statutory dimension.

Claude thus provided the most legally robust response, not only identifying all five elements but also contextualising them within the relevant statutory framework. ChatGPT and DeepSeek closely followed in terms of structure and doctrinal accuracy but failed to account for the statutory modifications. Gemini and Grok demonstrated a weaker grasp of the required elements, with Grok's response being the least structured. The omission of the statutory considerations in all models except Claude highlights a recurring limitation in AI-generated legal reasoning - namely, an inability to integrate statutory law with common law principles unless explicitly prompted.

3.3.3 Application to the facts

The application of the five elements of *actio legis Aquiliae* to the facts of the case varied across the AI models. The strongest responses correctly identified how each element should be applied to the given scenario, while weaker responses either omitted key elements or applied them in a manner inconsistent with South African law.

In terms of wrongfulness, the correct legal approach is to recognise that where a *positive act* causes physical damage, wrongfulness is *prima facie* established.[36] This is the case in the present scenario, where Jacob's affirmative conduct - lighting a fire - led directly to harm. There is no immediate need to inquire into a duty of care unless wrongfulness is challenged. However, the statutory duty created by section 12(1) of the NVFFA further reinforces wrongfulness, making it even clearer that Jacob's actions were unlawful. Claude, which referenced the NVFFA, had the strongest foundation for establishing wrongfulness. The other models, which failed to engage with the statute, needed to rely on the *prima facie* wrongfulness of the act but did not always do so correctly. Some models instead introduced the concept of a duty of care prematurely, treating it as a prerequisite rather than a response to a potential challenge.

With regard to fault, specifically negligence, the facts of the case already indicate that Jacob acted negligently by lighting a fire despite Maria's prior warnings, and by doing so on a windy day. Even without the NVFFA, this would be sufficient to establish fault under the test in *Kruger v Coetzee*,[37] as a reasonable person in Jacob's position would have foreseen the risk of harm and taken steps to prevent it. However, section 34(1) of the NVFFA introduces a presumption of negligence in cases where a fire spreads from a landowner's property, shifting the burden onto Jacob to prove that he was

36 *Trustees for the Time Being of Two Oceans Aquarium Trust v Kantey & Templer (Pty) Ltd* 2006 3 SA 138 (SCA) para 10.

37 *Kruger v Coetzee* 1966 2 SA 428 (A).

not negligent. Once again, Claude had the advantage of incorporating the statutory presumption, while the other models relied only on common law reasoning. ChatGPT and DeepSeek nevertheless provided strong arguments for negligence, as they correctly applied *Kruger v Coetzee* to the facts. Gemini and Grok offered weaker analyses, with Grok in particular displaying confusion in its reasoning.

Causation consists of two elements: factual and legal causation. Factual causation is established through the *but-for* test - *but for* Jacob's act of setting the fire, Maria's grazing land and cattle would not have been harmed. All AI models correctly identified factual causation. Legal causation, which considers whether the harm is sufficiently proximate or too remote, is also satisfied in this case, as the damage was a direct and foreseeable consequence of Jacob's conduct. The fire spread *immediately* due to the wind, leaving no intervening causes that might break the chain of causation. All models correctly acknowledged causation in some form, but ChatGPT and DeepSeek were the most precise in distinguishing between factual and legal causation.

Finally, patrimonial loss is evident from the facts, as Maria suffered economic harm in the form of damage to her grazing land and injured cattle. This element was correctly recognised by all models.

Claude provided the most legally robust response by applying the statutory framework to its analysis of wrongfulness and fault, thereby reinforcing its reasoning. ChatGPT and DeepSeek followed closely, offering well-structured applications of delictual principles but failing to incorporate the NVFFA. Gemini and Grok demonstrated a weaker grasp of the application of legal principles to the facts, with Grok's reasoning being the least coherent. The primary weakness across the models was the failure to recognise that wrongfulness was *prima facie* established due to Jacob's *positive act*, with some models instead resorting to an unnecessary discussion of the duty of care. Similarly, the omission of the statutory presumption of negligence in all but Claude's response meant that most models did not fully appreciate the evidentiary burden placed on Jacob to exonerate himself.

3.3.4 Case law references

Each AI model cited case law in its response, though the number of cases referenced varied. ChatGPT, Claude and DeepSeek cited multiple cases, while Gemini referenced only one case. Grok cited four cases but provided incomplete citations, listing only the parties and year. While this made verification more difficult, it did not significantly impact its reliability, as the cases were still identifiable.

ChatGPT made a significant error by hallucinating a non-existent case, "*Mokgokong v Minister of Police* 2000 2 SA 92 (T)". Additionally, it relied on a criminal law case[38] to explain causation 9nstead of referencing a delictual precedent. Claude cited two cases,[39] one of which was correctly applied, while the other - *Roshcon (Pty) Ltd v Anchor Auto Body Builders CC*[40] - was entirely unrelated to fire-related liability, instead dealing with contractual ownership disputes. DeepSeek referenced three cases,[41] all of which were real, but two were misapplied - *Van der Merwe v Road Accident Fund*[42] did not establish a general duty to prevent harm as the AI claimed, and *Herschel v Mrupe*[43] was incorrectly presented as authority on fire-related negligence. Gemini referenced only one case, *Minister van Polisie v Ewels*,[44] which was correctly cited and properly applied.

3.3.5 Relevance of case law

A recurring issue across the AI responses was the frequent citation of case law that was not the best or even an applicable precedent for the principles discussed. While all of the models correctly stated the general legal principles, they often did not identify the most relevant case law to support their conclusions. This is indicative of a failure to find appropriate precedents for the legal principles. Four of the AI models referenced cases that were either from an unrelated legal domain (criminal law in the case of ChatGPT and contract law in the case of Claude) or did not support the principle they were cited for (as seen with DeepSeek's reliance on *Van der Merwe* and *Herschel and* Grok's use of *Cape Town Municipality v Paine).*[45]

3.3.6 Consideration of defences or counterarguments

The AI models varied significantly in their engagement with potential defences or counter-arguments. Claude was the most comprehensive, recognising that Jacob could attempt to rebut the presumption of negligence under section 34(1) of the NVFFA by proving that he took reasonable precautions or was a member of a registered fire protection association. However, Claude did not explore whether Jacob might argue that the spread of the fire was due to an unforeseeable intervening force, such as an exceptionally strong gust of wind beyond his control.

---

38 *S v Mokgethi* 1990 1 SA 32 (A).

39 *Roshcon (Pty) Limited v Anchor Auto Body Builders CC* (49/13) [2014] ZASCA 40 (31 March 2014); *Steenberg v De Kaap Timber (Pty) Ltd* 1992 2 SA 169 (A).

40 *Roshcon (Pty) Limited v Anchor Auto Body Builders CC* (49/13) [2014] ZASCA 40 (31 March 2014).

41 *Van der Merwe v Road Accident Fund* 2006 4 SA 230 (CC); *Herschel v Mrupe* 1954 3 SA 464 (A); *International Shipping Co (Pty) Ltd v Bentley* 1990 1 SA 680 (A).

42 *Van der Merwe v Road Accident Fund* 2006 4 SA 230 (CC).

43 *Herschel v Mrupe* 1954 3 SA 464 (A).

44 *Minister van Polisie v Ewels* 1975 3 SA 590 (A).

45 *Cape Town Municipality v Paine* 1923 AD 207.

ChatGPT and DeepSeek considered possible counter-arguments but did so in a more general manner. Both models acknowledged that Jacob might argue that he acted reasonably under the circumstances or that the spread of the fire was beyond his control. However, neither incorporated the statutory framework, meaning that they failed to recognise that the burden of proof would shift to Jacob under section 34(1) of the NVFFA.

Gemini and Grok displayed the weakest engagement with defences. While both suggested that Jacob might argue that the fire was accidental, they did not meaningfully consider how South African law would assess such a claim, nor did they address the role of the statutory presumption of negligence. Grok in particular failed to offer any structured counter-argument analysis.

Overall, Claude provided the most structured approach to defences by engaging with the statutory burden-shifting mechanism, while ChatGPT and DeepSeek considered counter-arguments but lacked statutory grounding. Gemini and Grok offered superficial or incomplete defences, weakening their analyses.

3.3.7 Clarity and accuracy of the final answer

The clarity and accuracy of the final answers varied among the AI models. Claude provided the most precise and well-reasoned conclusion, explicitly affirming that Maria has a valid claim under *actio legis Aquiliae*, reinforced by the statutory presumption of negligence under the NVFFA. Its final answer was both legally sound and clearly articulated. ChatGPT and DeepSeek also concluded that Maria had a valid claim, correctly applying *actio legis Aquiliae*. However, their final statements lacked statutory reinforcement, making their conclusions legally accurate but somewhat incomplete. Gemini reached the correct conclusion but presented it in a less structured manner, making its reasoning less persuasive. Grok was the weakest, as its final answer was vague and contained legal inaccuracies, particularly in its mischaracterisation of negligence within South African delict.

Overall, Claude provided the strongest final answer in terms of both clarity and accuracy, while ChatGPT and DeepSeek followed closely behind. Gemini was less structured but still broadly correct, whereas Grok's final answer was the weakest due to a lack of precision and doctrinal clarity.

3.3.8 Conclusion on Scenario 3

The evaluation of Scenario 3 highlights both the strengths and limitations of AI models in engaging with South African delictual law, particularly in cases involving statutory modifications to common law principles. While all models correctly recognised that the claim falls within the realm of delict, only

ChatGPT, Claude, and DeepSeek explicitly identified *actio legis Aquiliae* as the applicable cause of action. However, Claude was the only model to incorporate the NVFFA, recognising its role in establishing a statutory duty of care and a presumption of negligence. The other models, despite correctly identifying the common law elements of *actio legis Aquiliae*, failed to integrate this crucial statutory framework, limiting the accuracy of their analyses.

A key weakness across the models was the failure to appreciate that wrongfulness was *prima facie* established due to Jacob's positive act of lighting the fire. Several models unnecessarily engaged with duty of care considerations, which would become relevant only if wrongfulness were challenged. The engagement with case law was also inconsistent, with multiple models citing cases that were either misapplied or not the most relevant. Additionally, while some models considered potential defences, only Claude correctly acknowledged the statutory burden-shifting mechanism under the NVFFA.

Overall, Claude provided the most legally precise and structured response, followed by ChatGPT and DeepSeek, which offered strong but incomplete analyses. Gemini and Grok were the weakest, with Grok's response being particularly unclear. The findings underscore the importance of human oversight when using AI for legal research, as even the strongest models struggle to consistently integrate the principles of statutory and common law.

3.4 Overall assessment

The overall performance of the AI models across the seven evaluation criteria is summarised in Table A below, providing a comparative overview of their relative strengths and weaknesses.

Table A: Overall performance of the AI models across the seven evaluation criteria

Criteria	Claude	ChatGPT	DeepSeek	Grok	Gemini
Identification of Cause of Action	Consistently accurate, correctly identified cause of action in all scenarios	Correctly identified cause of action in most cases, but introduced uncertainty in Scenario 2	Correct in most cases but misinterpreted key elements in Scenario 1 and 2	Missed the correct cause of action in Scenario 2 and 3, incorrectly framed negligence	Struggled with identifying the correct cause of action, weakest overall
Accuracy of Legal Requirements	Structured analysis, correctly incorporated	Well-structured but failed to incorporate	Identified key elements but	Identified basic elements but lacked precision	Inconsistent, missed key legal elements in

Criteria	Claude	ChatGPT	DeepSeek	Grok	Gemini
	statutory modifications	statutory modifications	frequently misapplied them		multiple scenarios
Application to Facts	Strongest factual application, integrated statutory framework	Strong factual application but lacked statutory engagement in Scenario 3	Applied legal principles well but lacked statutory considerations	Weaker application, often confused legal tests	Weak factual application, often misclassified legal principles
Case Law References	Cited 3-4 cases per scenario, all real and correctly formatted	Cited 2-4 cases per scenario, but hallucinated cases in two scenarios	Cited 2-3 real cases per scenario but misapplied some	Cited 2-4 cases per scenario, but with incomplete citations	Cited 1-2 cases per scenario, often incomplete
Relevance of Case Law	Mostly relevant case law, though some misinterpretations	Some relevant case law but also incorrect or fabricated references	Frequent misinterpretation of case law principles	Limited relevance, often cited cases without full reasoning	Least engagement with case law, weakest citations
Consideration of Defences	Best engagement with defences, recognised statutory burden shifting	Engaged with defences but lacked statutory integration	Considered counter arguments but without depth	Weak engagement with defences, lacked legal depth	Superficial consideration of defences, lacked structured reasoning
Clarity and Accuracy of Final Answer	Most precise and well-reasoned conclusions, legally structured	Legally accurate but weakened by missing statutory elements	Accurate but lacked depth in statutory considerations	Final answers vague, lacked doctrinal clarity	Final answers broadly correct but poorly structured

3.4.1 Identification of the cause of action

The AI models varied in their ability to correctly identify the applicable cause of action across the three scenarios. All five models correctly identified *actio de pauperie* in Scenario 1, demonstrating a strong foundational understanding of liability for harm caused by domesticated animals. However, in Scenario 2 only Claude correctly identified *negotiorum gestio*, while ChatGPT and DeepSeek introduced doctrinal uncertainty by referencing unjustified enrichment alongside *negotiorum gestio*. Gemini and Grok entirely failed to identify *negotiorum gestio*, instead misapplying unjustified enrichment and implied contract principles. In Scenario 3 ChatGPT, Claude and DeepSeek correctly identified *actio legis Aquiliae*, while Gemini and Grok framed the claim more generally as negligence, failing to specify the correct delictual action. Across the three scenarios, Claude demonstrated the most consistent accuracy in identifying the appropriate cause of action, while Gemini and Grok displayed the greatest weaknesses.

3.4.2 Accuracy of the legal requirements

Most models demonstrated a solid understanding of the fundamental elements of *actio de pauperie* in Scenario 1, though DeepSeek introduced an error by suggesting that negligence was relevant to liability. In Scenario 2 only Claude provided a structured breakdown of the requirements for *negotiorum gestio*, while the other models either confused it with unjustified enrichment or failed to correctly outline its elements. In Scenario 3 ChatGPT and DeepSeek provided structured lists of the five elements of *actio legis Aquiliae*, while Claude framed its analysis through case law rather than as a doctrinal structure. However, only Claude incorporated statutory modifications, specifically the NVFFA, demonstrating a more integrated understanding of South African law. Overall, Claude performed best in correctly identifying legal requirements, while Gemini and Grok consistently failed to structure their responses around the correct legal principles.

3.4.3 Application to the facts

In Scenario 1 most models correctly applied *actio de pauperie* to the facts, though DeepSeek's erroneous emphasis on negligence led to flawed reasoning. Scenario 2 revealed substantial weaknesses in application, as ChatGPT, DeepSeek, Gemini and Grok failed to apply the facts within the correct framework. Claude's response, while legally structured, contained some unnecessary complexity. Scenario 3 showed the strongest factual application, with ChatGPT, Claude and DeepSeek correctly applying *actio legis Aquiliae*, while Gemini and Grok displayed less coherence. A key weakness across multiple responses was the failure to recognise *prima facie* wrongfulness in Scenario 3, with some models resorting to an unnecessary discussion of the duty of care. Overall, Claude again demonstrated the most consistent factual application, followed closely by ChatGPT and DeepSeek, while Gemini and Grok exhibited weaker factual reasoning.

3.4.4 Case law references

Across all three scenarios the AI models demonstrated varying abilities in retrieving and correctly applying case law. Claude consistently cited 3-4 cases per scenario, all of which were real and correctly formatted. While it occasionally misinterpreted their significance, it did not fabricate case law, making it the most reliable model in this regard. ChatGPT cited 2-4 cases per scenario, but its reliability was significantly undermined by hallucinated cases in two of the three scenarios. While it often provided useful case references, the fact that it generated fictitious authorities presents a serious credibility issue. DeepSeek cited 2-3 real cases per scenario with correct formatting, but it frequently misapplied them, attributing broader doctrinal principles than the cases actually supported. However, unlike ChatGPT it

did not fabricate cases, making it slightly more reliable in citation accuracy. Gemini was the weakest in terms of case law engagement, citing only 1-2 cases per scenario, often in an incomplete format. This limited its ability to substantiate legal arguments with relevant precedents, weakening the depth and authority of its responses. Grok cited 2-4 cases per scenario, but its citations were frequently incomplete, listing only the parties and year without full references. While it did not hallucinate cases, this incomplete referencing made verification difficult. Moreover, its reasoning was often legally imprecise, leading to misapplications of case law.

3.4.5 Relevance of case law

While the models generally referenced case law that aligned with the legal principles discussed, there were frequent instances of misinterpretation. Scenario 1 saw mostly correct case citations, with ChatGPT's reference to *Van Meyeren v Cloete*[46] adding depth. In Scenario 2 multiple models misapplied case law, often attributing broader principles to judgments than they actually supported. In Scenario 3 ChatGPT, Claude, and DeepSeek cited relevant cases, but some references were drawn from unrelated legal domains. The inconsistent relevance of case law underscores a broader limitation in AI-generated legal reasoning - namely, the tendency to extract and apply legal principles in an overly broad manner without fully engaging with judicial reasoning.

3.4.6 Consideration of defences or counterarguments

ChatGPT, Claude and Gemini demonstrated the strongest engagement with defences in Scenario 1, correctly discussing provocation, trespassing and third-party fault. In Scenario 2 Claude was the only model to consider whether Thabo's explicit rejection of the vehicle could serve as a counterargument, though none of the models identified the strongest defence - that Thabo's awareness of the storage could displace *negotiorum gestio* in favour of implied contract. In Scenario 3 Claude again provided the most structured counterargument, recognising Jacob's potential rebuttal under section 34(1) of the NVFFA. ChatGPT and DeepSeek considered counterarguments in general terms but failed to incorporate the statutory burden-shifting mechanism. Gemini and Grok displayed the weakest engagement with defences across all three scenarios.

3.4.7 Clarity and accuracy of the final answer

Claude consistently provided the most precise and well-reasoned final answers, clearly affirming the correct legal conclusions in all three scenarios. ChatGPT and DeepSeek followed closely, producing mostly accurate conclusions, though their failure to incorporate statutory

46 *Van Meyeren v Cloete* (636/2019) [2020] ZASCA 100 (11 September 2020).

modifications in Scenario 3 weakened their responses. Gemini's final answers were generally correct but lacked structure, making their reasoning less persuasive. Grok was the weakest, frequently displaying legal inaccuracies and unclear conclusions. Across all scenarios the strongest responses were those that logically structured their reasoning from legal principles to factual application before arriving at a well-articulated conclusion.

3.5 Comparative evaluation and ranking

Considering performance across all three scenarios, Claude emerges as the strongest model, consistently identifying the correct legal framework, integrating statutory modifications where applicable, and engaging with counterarguments more effectively than the others. ChatGPT and DeepSeek follow closely, both providing structured reasoning and accurate conclusions but struggling with doctrinal confusion in Scenario 2 and failing to incorporate statutory law in Scenario 3. The evaluation of case law references reveals that Gemini consistently cited the fewest cases, which, combined with its weaker legal reasoning, warrants its ranking as the weakest performer. Despite its frequent legal inaccuracies, Grok at least attempted to engage with case law more frequently than Gemini.

Thus, the AI models rank as follows:

1. Claude – Strongest overall performance, with correct legal identification, statutory integration, and reliable case law references.
2. ChatGPT – Well-structured reasoning and strong engagement with case law, but hallucinated authorities undermine its credibility.
3. DeepSeek – Similar to ChatGPT but did not hallucinate cases. However, frequent misapplication of case law weakened its responses.
4. Grok – Provided more case references than Gemini but often lacked doctrinal clarity and used incomplete citations.
5. Gemini – The weakest performer due to consistently citing the fewest cases, providing incomplete references, and failing to substantiate legal arguments effectively.

This comparative assessment underscores that while AI models can produce legally sound analyses, their performance varies significantly depending on the complexity of the legal issue, the presence of statutory modifications, and their ability to engage with case law. Even the strongest models require human oversight to verify accuracy, particularly in the jurisdiction-specific applications of legal doctrine.

4 Discussion

The results of this study demonstrate that while AI models exhibit notable strengths in identifying and applying South African legal principles, they also reveal significant weaknesses in case law selection and statutory engagement. The models' ability to correctly determine the relevant cause of action and apply the legal elements to factual scenarios suggests that generative AI can serve as a useful tool for legal analysis. However, their inconsistent engagement with statutory provisions and frequent misapplication or fabrication of case law underscore the necessity for human oversight. These findings have important implications for the use of AI in legal practice and academia, particularly in assessing its reliability as a supplementary research tool.

4.1 Strengths in legal identification and application

The AI models performed well in recognising the relevant cause of action in most scenarios. In Scenario 1 all five models correctly identified *actio de pauperie*, suggesting a robust ability to classify strict liability claims for harm caused by domesticated animals. Similarly, in Scenario 3 ChatGPT, Claude and DeepSeek correctly identified *actio legis Aquiliae* as the basis for a delictual claim, demonstrating an awareness of South Africa's common law principles on patrimonial loss. The identification of *negotiorum gestio* in Scenario 2, however, was more inconsistent, with only Claude recognising it as the applicable doctrine. This indicates that AI models are more proficient in well-established areas of South African delict but struggle with more specialised doctrines.

Beyond legal identification, several models applied the relevant legal principles to the facts in a structured and logical manner. In Scenario 3 ChatGPT, Claude and DeepSeek correctly applied the five elements of *actio legis Aquiliae* to the facts, assessing wrongfulness, fault, causation and patrimonial loss in a legally coherent manner. Similarly, in Scenario 1 most models correctly established ownership, the *contra naturam* requirement, and causation. These results suggest that AI models are capable of performing structured legal reasoning, provided they correctly identify the governing principles. However, the errors observed in Scenario 2 indicate that their ability to apply legal doctrine weakens significantly when they fail to classify the claim correctly.

4.2 Deficiencies in case law selection and statutory engagement

One of the most striking limitations of the AI models was their inconsistent and often flawed engagement with case law. While some models correctly cited key authorities such as *O'Callaghan v Chaplin*[47] in Scenario 1, many

47 *O'Callaghan v Chaplin* 1927 AD 310.

either misinterpreted cases or cited irrelevant judgments. ChatGPT, despite demonstrating strong reasoning skills, hallucinated non-existent cases in two scenarios, raising serious concerns about its reliability for legal research. Gemini performed the weakest in case law engagement, consistently citing the fewest cases, often in an incomplete format, which weakened the authority of its responses. These deficiencies highlight the inability of AI models to systematically retrieve and apply authoritative case law, a fundamental requirement in legal practice.

In contrast to the explicit instruction to cite case law, no such instruction was given regarding statutes. Nonetheless, the failure of most AI models to engage with the NVFFA in Scenario 3 reflects a significant gap in their reasoning. A legal practitioner would instinctively consider both common law and statutory law when assessing liability, yet only Claude recognised the statutory duty of care and presumption of negligence imposed by the Act. The absence of statutory engagement in other models suggests that AI tools may not default to incorporating statutory provisions unless specifically prompted. This is a critical shortcoming, as legal reasoning requires a holistic approach that integrates both statutory and common law sources.

4.3 Implications for the use of AI in legal practice and academia

The findings of this study suggest that AI models can serve as useful aids in legal research and analysis, but their limitations must be carefully managed. In legal practice AI tools can assist in structuring legal arguments, identifying relevant legal principles and applying doctrinal tests to factual scenarios. However, their inconsistent engagement with case law and statutes makes them unreliable as stand-alone research tools. The hallucination of cases in particular presents a serious risk, as legal practitioners relying on AI-generated references without verification may inadvertently introduce fabricated authorities into their work. This underscores the need for AI outputs to be carefully reviewed and supplemented with independent legal research.

In academia AI models may serve as valuable tools for legal education, helping students grasp the fundamental principles of South African law. However, the findings indicate that AI-generated legal reasoning lacks depth in areas requiring statutory interpretation and doctrinal nuance. This suggests that while AI can be leveraged for preliminary research, it cannot replace traditional methods of legal analysis that require engagement with authoritative sources. Future research should explore ways to enhance AI's ability to integrate statutory law without explicit prompting and to improve the reliability of its case law retrieval. Until then, AI's role in legal education and practice should remain complementary rather than determinative.

5 Conclusion and future directions

This study has demonstrated that AI models exhibit notable strengths in identifying and applying South African private law principles, particularly in well-established areas of delict and unjustified enrichment. The ability of most models to correctly classify legal actions such as *actio de pauperie* and *actio legis Aquiliae* suggests that AI can competently navigate common-law-based claims. However, their struggle with more specialised doctrines, such as *negotiorum gestio*, highlights the unevenness of their legal reasoning. Furthermore, while some models applied legal principles to factual scenarios with logical consistency, their failure to engage comprehensively with statutory provisions - particularly in Scenario 3 - reveals a critical limitation in their approach to legal analysis.

One of the most concerning findings is the AI models' inconsistent and sometimes misleading engagement with case law. While some models correctly cited foundational precedents, others hallucinated non-existent cases or misapplied judgments, raising concerns about the reliability of AI-generated legal research. Moreover, the limited engagement with statutory law suggests that AI tools may not instinctively consider legislation unless explicitly prompted, a significant shortcoming given that statutory frameworks often modify or supplement common law principles. These limitations underscore the need for human oversight, as legal reasoning requires the ability to integrate both case law and statutory provisions in a holistic manner.

Future research should explore AI's competency across a broader spectrum of South African legal issues, particularly in areas beyond private law, such as statutory interpretation, constitutional law and procedural rules. Additionally, comparative studies assessing AI performance across different legal systems could provide valuable insights into how training data and retrieval mechanisms influence AI's legal reasoning. Understanding whether AI performs better in legal systems with codified law compared to those with a strong common law tradition could further clarify its potential and limitations.

Ultimately, while AI presents promising applications in legal research and analysis, its current limitations reinforce the irreplaceable role of human legal expertise. AI tools can support legal practitioners and scholars in structuring arguments and retrieving information, but their inconsistent engagement with statutes and case law means they cannot yet function as stand-alone research tools. As AI continues to develop, addressing these deficiencies - particularly in statutory engagement and case law accuracy - will be critical in determining its reliability and practical value in the legal field.

Acknowledgement

During the preparation of this work, ChatGPT 4o was used to enhance language and readability. The author subsequently reviewed and edited the content as necessary and assumes full responsibility for the content of the publication.

Bibliography

Literature

Bottomley and Thaldar 2023 *Frontiers in Pharmacology*
Bottomley D and Thaldar D "Liability for Harm Caused by AI in Healthcare: An Overview of the Core Legal Concepts" 2023 *Frontiers in Pharmacology* https://doi.org/10.3389/fphar.2023.1297353

Mtuze and Morige 2024 *Obiter*
Mtuze SSK and Morige M "Towards Drafting Artificial Intelligence (AI) Legislation in South Africa" 2024 *Obiter* 161-179 https://doi.org/10.17159/obiter.v45i1.18399

Naidoo *et al* 2022 *South African Journal of Bioethics and Law*
Naidoo S *et al* "Artificial Intelligence in Healthcare: Proposals for Policy Development in South Africa" 2022 *South African Journal of Bioethics and Law* 11-16 https://doi.org/10.7196/SAJBL.2022.v15i1.797

Sihlahla *et al* 2025 *Developing World Bioethics*
Sihlahla I *et al* "Legal and Ethical Principles Governing the Use of Artificial Intelligence in Radiology Services in South Africa" 2025 *Developing World Bioethics* 35-45 https://doi.org/10.1111/dewb.12436

Case law

Cape Town Municipality v Paine 1923 AD 207

Conradie v Rossouw 1919 AD 276

Da Silva v Otto 1986 3 SA 538 (T)

Evrigard (Pty) Ltd v Select PPE (Pty) Ltd (2022-22743) [2024] ZAGPJHC 183 (26 February 2024)

Gouws v Jester Pools (Pty) Ltd 1968 3 SA 563 (T)

Herschel v Mrupe 1954 3 SA 464 (A)

Immaculate Truck Repairs CC v Capital Acceptances Ltd (1153/2014) [2016] ZAFSHC 3 (14 January 2016)

International Shipping Co (Pty) Ltd v Bentley 1990 1 SA 680 (A)

Joubert v Combrinck 1980 3 SA 680 (T)

Kommissaris van Binnelandse Inkomste v Willers 1994 3 SA 283 (A)

Kruger v Coetzee 1966 2 SA 428 (A)

Lever v Purdy 1993 3 SA 17 (A)

Mackintosh v Mackintosh 1864 2 M 1357

Makunga v Barlequins Beleggings (Pty) Ltd t/a Indigo Spur (19733/2017) [2023] ZAWCHC 332 (1 December 2023)

Mavundla v MEC: Department of Co-Operative Government and Traditional Affairs KwaZulu-Natal 2025 3 SA 534 (KZP)

McCarthy Retail Ltd v Shortdistance Carriers CC 2001 3 SA 482 (SCA)

Minister van Polisie v Ewels 1975 3 SA 590 (A)

Nortje v Pool 1966 3 SA 96 (A)

O'Callaghan v Chaplin 1927 AD 310

Parker v Forsyth (1585/20) [2023] ZAGPRD 1 (29 June 2023)

Portwood v Svamvur 1970 4 SA 8 (RA)

Roshcon (Pty) Limited v Anchor Auto Body Builders CC (49/13) [2014] ZASCA 40 (31 March 2014)

S v Mokgethi 1990 1 SA 32 (A)

Steenberg v De Kaap Timber (Pty) Ltd 1992 2 SA 169 (A)

Trustees for the Time Being of Two Oceans Aquarium Trust v Kantey & Templer (Pty) Ltd 2006 3 SA 138 (SCA)

Van der Merwe v Road Accident Fund 2006 4 SA 230 (CC)

Van Meyeren v Cloete (636/2019) [2020] ZASCA 100 (11 September 2020)

Legislation

National Veld and Forest Fire Act 101 of 1998

Internet sources

Anthropic 2025 https://www.anthropic.com/claude/sonnet
Anthropic 2025 *Claude 3.7 Sonnet* https://www.anthropic.com/claude/sonnet accessed 1 March 2025

Bhambhoria *et al* 2024 https://arxiv.org/abs/2404.12349
Bhambhoria R *et al* 2024 *Evaluating AI for Law: Bridging the Gap with Open-Source Solutions* https://arxiv.org/abs/2404.12349 accessed 13 August 2025

DeepSeek-AI 2025 https://www.deepseek.com
DeepSeek-AI 2025 *DeepSeek-R1* https://www.deepseek.com accessed 1 March 2025

DS-I Africa Law Research Group 2024 https://www.datalaw.bot
DS-I Africa Law Research Group, University of KwaZulu-Natal 2024 *Welcome to datalaw.bot* https://www.datalaw.bot accessed 1 March 2025

Google DeepMind 2025 https://gemini.google.com/app
Google DeepMind 2025 *Gemini 2.0 Flash* https://gemini.google.com/app accessed 1 March 2025

Magesh *et al* 2024 https://arxiv.org/abs/2405.20362
Magesh V *et al* 2024 *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* https://arxiv.org/abs/2405.20362 accessed 13 August 2025

Nordling 2024 https://doi.org/10.1038/d44148-024-00249-w
Nordling L 2024 *New AI Chatbot Will Make African Health Data More Usable* https://doi.org/10.1038/d44148-024-00249-w accessed 13 August 2025

OpenAI 2025 https://chat.chatbotapp.ai/?model=gpt-4o
OpenAI 2025 *GPT-4o* https://chat.chatbotapp.ai/?model=gpt-4o accessed 1 March 2025

xAI 2025 https://grok.com/?referrer=website
xAI 2025 *Grok 3 Beta* https://grok.com/?referrer=website accessed 1 March 2025

List of Abbreviations

AI	artificial intelligence
NVFFA	*National Veld and Forest Fire Act 101 of 1998*