Received: 21 FEB 2025 | Accepted: 25 JUN 2025 | Published Online: 30 JUN 2025
Recoletos Multidisciplinary Research Journal Vol. 13 no. 1 (June 2025)
DOI: https://doi.org/10.32871/rmrj2513.01.15

Orginal Article

Using Large Language Models to Transform Estate Planning

Richard Yeaw Chong Seow

Pôle Paris Alternance Business School, Paris, France

Correspondence:
richard-seow@hotmail.com

Abstract

Background: *Advancements in generative artificial intelligence (AI) and large language models (LLMs) have introduced new possibilities in democratizing access to professional services. Malaysia's non-Muslims can rely on AI-driven tools to draft their wills without legal assistance, but empirical evaluations of LLM chatbots' reliability are absent.*

Methods: *Grounded in the Technology Acceptance Model, this study assesses the accuracy, legal validity, comprehensibility, and reliability of five prominent LLM chatbots, namely ChatGPT 3.5, ChatGPT 4.0, Claude Sonnet, Gemini Pro, and Microsoft Copilot.*

Results: *ChatGPT 4.0 consistently outperformed other models across all complexity levels in succession and drafting-related question tasks, showing the highest reliability and accuracy. Gemini Pro performed well for introductory and intermediate queries, particularly in drafting simple wills. In contrast, Copilot and Claude Sonnet exhibited high variability and struggled with complex queries. Across all chatbots, performance declined significantly with increased query complexity. Qualitative assessment reveals inconsistencies, misinterpretations, and occasional legal inaccuracies, particularly when prompts contain incomplete information.*

Conclusion: *While specific LLM chatbots, particularly ChatGPT 4.0, demonstrate potential as reliable tools for basic estate planning, their limitations in handling complex legal instructions underscore the need for caution. By shedding light on the role of AI in legal contexts, this research significantly enriched both scholarly and practical dialogues, enhancing our understanding of AI's potential to revolutionize the legal landscape, particularly in estate planning.*

Keywords

generative artificial intelligence, large language models, chatbots, estate planning, Malaysia, wills

INTRODUCTION

The rapid advancement in artificial intelligence (AI) technology, propelled by innovations in large language models (LLMs) and natural language processing (NLP), has significantly broadened its applicability across professional sectors (Deveci et al., 2023). These technologies underpin the development of sophisticated chatbots capable of interpreting and generating human-like language, enabling complex, interactive

communication (Rahmanti et al., 2022; Weeks et al., 2023). This progress marks a critical shift in AI's capacity to perform tasks that once required human expertise, enhancing precision, automation, and decision-making in data-intensive environments (Emenike & Emenike, 2023; Sætra, 2020). AI's accelerated development has liberated unprecedented opportunities in various sectors, including medicine, healthcare, education, and more, enhancing automation, efficiency, precision, and the efficacy of data-driven processes (Armour & Sako, 2020; Egli et al., 2020). The release of ChatGPT 3.5 in late 2022, followed by tools such as ChatGPT 4.0, Microsoft Copilot, Gemini Pro, Claude Sonnet, Grok, and DeepSeek, marked a pivotal moment in the expanding use and popularity of generative AI, enabling users to perform complex tasks with greater efficiency and precision (Brynjolfsson et al., 2025; Ooi et al., 2025).

The proliferation of generative AI tools in multiple domains has revealed their potential to undertake complex analytical tasks traditionally reserved for humans (Mojadeddi, 2023). Empirical studies further affirm LLMs' capability to manage high-level analytical functions. AI systems have demonstrated strong performance on professional examinations, including those in medicine, law, and science, often outperforming humans on essay-based assessments that require nuanced reasoning (Choi et al., 2022; Gilson et al., 2023; Kung et al., 2023). In healthcare, AI chatbots are increasingly used to support clinical decisions, improve patient interaction, and assist in disease management and diagnostics (Aslan et al., 2014; Lim et al., 2023; Piya et al., 2021). Their application extends to supporting medical professionals, including microbiologists and nursing staff, by streamlining diagnostic workflows and enhancing patient communication (Egli, 2023; Gazulla et al., 2023; Mendoza et al., 2022). Together, these developments highlight AI's growing potential to complement and amplify human expertise, paving the way for more efficient, informed, and accessible professional practice across disciplines.

In Malaysia, the Wills Act of 1959 permits any adult of sound mind to draft a will independently without requiring legal representation. The Act underscores that a will's validity depends on proper attestation rather than who drafts it. While some individuals still seek professional legal services, others increasingly turn to digital solutions, including online templates and AI tools like ChatGPT and Gemini. Given that LLM chatbots can now generate human-like text (Deveci et al., 2023), their application in will-writing appears plausible. However, despite growing interest in AI's broader capabilities, its integration into legal services, particularly estate planning, remains underexplored (Armour & Sako, 2020). Scholars such as Basir et al. (2023) and Zandi et al. (2017) have noted limited academic engagement with estate planning, especially in developing countries. Crucially, there is little empirical research evaluating how effectively LLM chatbots support non-Muslim Malaysians in this context, despite the potential risks of laypersons relying on these tools without legal training.

This study contributes significantly to the growing discourse on AI integration in the legal field by advancing theoretical understanding and practical insight. First, it assesses the accuracy of LLM chatbots in delivering succession-related guidance, providing a clearer view of their reliability as tools for legal support. It underscores AI's potential to broaden access to legal services, particularly in contexts where traditional legal assistance is financially or socially prohibitive. Second, the findings help legal practitioners evaluate the practical viability of generative AI in their workflows, laying the groundwork for its ethical and informed adoption. Finally, the results have policy implications, highlighting the urgent need for regulatory frameworks that ensure responsible use of AI in legal contexts, safeguard users from misinformation, and uphold professional standards.

Estate Planning

Estate planning encompasses financial planning, wealth distribution, and succession planning, and is a crucial strategy for managing asset transfers in the event of death or incapacitation (Basir et al., 2023). Beyond ensuring the orderly distribution of assets, it promotes financial security for dependents, prevents family disputes, preserves legacies, and facilitates intergenerational wealth transfer (Basir et al., 2023; Harrington, 2012; Horkey, 2009; Ismail et al., 2013; Madura & Gill, 2015). A well-designed estate plan can also offer tax advantages and ensure liquidity, allowing beneficiaries' needs to be met promptly. Sometimes, the testator

may benefit directly from the arrangement during their lifetime, highlighting estate planning's adaptability to diverse personal and financial goals (Horton, 2017).

In Malaysia, a wide range of legal instruments, including wills, trusts, nominations, powers of attorney, and buy-sell agreements, form the foundation of estate planning. Among these, wills play a particularly central role for non-Muslim Malaysians, providing a legally recognized mechanism for posthumous asset allocation under the Wills Act 1959 (Horton, 2017; Kim & Stebbins, 2021). The Act grants significant autonomy to individuals, allowing them to create, modify, or revoke wills with minimal formal constraints. Drafting a will can help avoid intestate succession under the Distribution Act 1958, mitigate legal delays, and prevent familial conflict (Basir et al., 2023; Choi et al., 2019; Cox & Stark, 2005). Additionally, wills can empower marginalized groups, such as women and LGBTQ+ individuals, to exercise control over their assets and ensure the well-being of their partners and heirs (Mendenhall et al., 2007; Westwood, 2015). They also enable testators to appoint executors, designate beneficiaries, make charitable contributions, and plan memorial arrangements (Street, 2006). Multiple factors influence individuals' estate planning decisions, including personal attitudes, family dynamics, financial literacy, and cultural or religious beliefs (Aubry et al., 2023; Ismail et al., 2013). Financially literate individuals and property owners are more likely to draft wills, and lifestyle practices such as meditation have been linked to proactive estate planning behaviors (Gill et al., 2017). With the emergence of generative AI tools, non-Muslim Malaysians now have access to an additional support mechanism that could simplify will drafting and expand participation in estate planning processes.

Use of Artificial Intelligence (AI) in the Legal Process

The evolution of AI, initially driven by efforts to replicate human intelligence, has been shaped by key milestones, including neural network development and machine learning (ML) (Izenman, 2008). ML marked a turning point by enabling systems to learn autonomously from large datasets, replacing rule-based programming with adaptive, data-driven decision-making that emulates aspects of human cognition such as reasoning and optimization (Bathaee, 2018; Izenman, 2008; Mittelstadt et al., 2016). Drawing from diverse fields, statistics, linguistics, and neuroscience, AI now underpins applications ranging from the automation of routine tasks to the execution of complex, knowledge-intensive functions (Surden, 2019; Trajtenberg, 2019). Despite these advancements, current AI systems remain fundamentally distinct from human intelligence; while they excel at pattern recognition and generating contextually appropriate outputs, they lack abstract reasoning and emotional understanding (Surden, 2014).

In the legal domain, an increasing number of people employ AI to assist with tasks such as document review, planning, and client guidance, particularly in areas like estate planning, where LLM-powered chatbots can address straightforward queries (Bench-Capon et al., 2012; Surden, 2019). LLM chatbots respond to straightforward legal queries, helping individuals navigate the complexities of legal documentation independently (Surden, 2019). However, legal reasoning often requires contextual sensitivity and interpretive nuance that AI cannot replicate (Rissland et al., 2003). As such, while AI enhances efficiency in routine legal processes, it cannot substitute for human judgment in tasks involving ambiguity, ethical discernment, or emotional complexity (Armour & Sako, 2020; Susskind, 2019). Moreover, the growing autonomy of AI also raises critical concerns around legal accountability. As AI systems increasingly influence or make decisions independently, traditional human oversight and liability frameworks are challenged, prompting urgent ethical and regulatory questions (Taeihagh, 2021). These issues are particularly pressing in law, where the stakes of decision-making are high and the consequences far-reaching.

Underpinning Theory

Reliability is critical to user trust in LLM chatbots, particularly when applied to high-stakes tasks such as will drafting. Although this study does not employ a single guiding theory, it draws conceptually from the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT/UTAUT2) to interpret user adoption behavior (Davis et al., 1989; Venkatesh et al., 2003, 2012). Rooted in the

theory of reasoned action, TAM posits that perceived usefulness, strongly linked to reliability, drives users' intentions to adopt technology (Cheng, 2019; Granić & Marangunić, 2019; Solomovich & Abraham, 2024). This model underscores that users' intention to engage with technology is molded mainly by their beliefs about its utility and ease of use (Davis et al., 1989; Lee & Lehto, 2013).

Reliability, defined as consistent and accurate performance (Balagurunathan et al., 2021), enhances users' confidence in a technology's utility. Furthermore, UTAUT and its extension, UTAUT2, introduce additional dimensions, such as performance expectancy and perceived risk, further elaborating on the adoption process (Venkatesh et al., 2003, 2012). A reliable system raises performance expectancy, the belief that the technology will deliver tangible benefits while reducing perceived risks, such as concerns about errors or adverse outcomes (Hess et al., 2014). This balance is essential in legal contexts where mistakes can have serious consequences. When LLM chatbots demonstrate high reliability by delivering accurate and pertinent responses, they boost performance expectancy and diminish perceived risks. Therefore, designing LLM chatbots that consistently deliver accurate and contextually appropriate outputs is essential for fostering trust and promoting adoption in professional settings. This study addresses this research gap by evaluating the reliability of LLM chatbots in estate planning, offering insights into their role in expanding access to legal services, and raising critical questions about their practical utility and implications for legal practice in Malaysia. This research aims to assess the practical utility of AI tools in personal legal processes and their potential implications for legal practice in Malaysia by understanding whether LLM chatbots can reliably assist non-Muslim Malaysians in making well-informed decisions related to their wills and determine the effectiveness and reliability of LLM chatbots in aiding non-Muslim Malaysians in the preparation of their wills.

METHODS

Research Design and Protocols

This study evaluates the accuracy and reliability of LLM chatbots in two key aspects of will preparation: guiding testamentary decision-making and assisting with will drafting. The methodological framework outlined in the research protocols (see Figure 1) was purposefully designed to address these objectives.

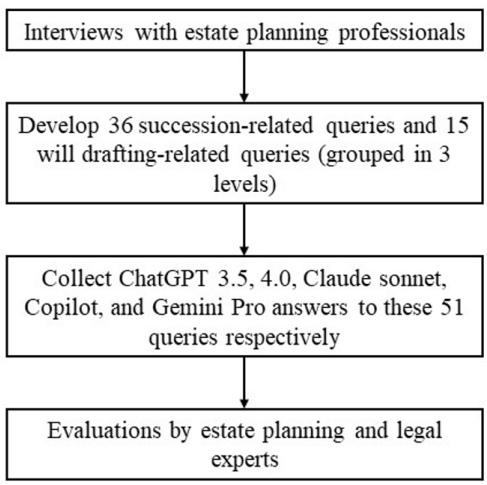

Figure 1. *Research protocols*

Estate planning typically involves two core stages: decision-making and will drafting. To reflect this process, the study categorized queries into two phases, succession planning, and drafting, and developed corresponding sets of questions. Ten estate planning professionals, each with over a decade of experience advising non-Muslim Malaysians, interviewed the respondents. The interviews identified common concerns and instructions raised by testators, forming the basis for 36 succession-related and 15 will-drafting questions

(see Table 1). These inquiries were further classified into basic, intermediate, and advanced levels to assess the performance of LLM chatbots across varying complexities (Lim et al., 2023).

The questions were administered on March 15, 2024, to five LLM chatbots: ChatGPT 3.5, ChatGPT 4.0, Claude Sonnet, Copilot, and Gemini 1.0 Pro. Each question was submitted once, with no further prompting or clarification. This single-attempt design reflects realistic user behavior, as most individuals lack the legal expertise to evaluate or refine chatbot-generated responses. Repeated submissions could distort the natural interaction flow and reduce the ecological validity of the evaluation.

Table 1. *Estate planning queries*

	Succession-related queries (SRQ)	**Will drafting-related queries (WDRQ)**
Basic	12 queries – general inquiries	5 queries – simple and specific testamentary instructions
Intermediate	12 queries – specific non-complex inquiries	5 queries – Specific, but complicated testamentary instructions with relevant information
Advanced	12 queries – specific and complex inquiries	5 queries – Specific, but complicated testamentary instructions with missing some relevant information

Evaluation

Responses generated by ChatGPT 3.5 and 4.0, Claude Sonnet, Copilot, and Gemini 1.0 Pro were independently evaluated by a panel of three estate planning and legal experts, each with over ten years of experience. To ensure impartiality, all responses were converted to plain text to mask model-specific identifiers (Lim et al., 2023). Before the evaluation, the panel developed standardized matrices reflecting key performance criteria: legal accuracy and clarity for SRQ and legal soundness and clause clarity for WDRQ. Each matrix featured five performance tiers aligned with professional estate planning standards.

Evaluators assessed the chatbot outputs over three weeks using a five-point scale across two dimensions per query, producing scores between 2 and 10. These were aggregated across all three experts to generate an overall rating for each chatbot. The panel also provided qualitative feedback, offering profound insights into the chatbot responses' practical reliability and legal adequacy in real-world estate planning scenarios.

Statistical Analysis

Researchers performed statistical analyses using the Statistical Package for the Social Sciences (SPSS) version 28. To assess the reliability and consistency of expert ratings, intraclass correlation coefficients (ICC) were calculated using a two-way mixed-effects model focused on consistency, an appropriate choice given the use of predetermined raters (Bujanga & Baharum, 2017; Koo & Li, 2016). The ICC is particularly suitable for assessing the consistency or agreement of quantitative assessments among two or more raters, making it highly applicable in this context (Müller & Büttner, 1994). Descriptive statistics provided an overview of the score distributions, and the Shapiro-Wilk test ($p < 0.001$) confirmed non-normality in the data, justifying the use of non-parametric methods (González-Estrada et al., 2022). Accordingly, the Kruskal-Wallis H test, followed by Dunn's post hoc test, was employed to compare chatbot performance across different complexity levels (Tomita et al., 2013). This approach ensured a robust analysis suited to the skewed distribution of the dataset.

RESULTS

Interclass Correlation Coefficients (ICC)

The ICC values were robust, registering at 0.803 for both SRQ and WDRQ and 0.801 for the combined SRQ and WDRQ, indicating good reliability among the raters (see Table 2) (Koo & Li, 2016). Additionally, the study reports a Cronbach's Alpha of 0.803, well above the 0.7 threshold, thereby confirming a high level of reliability in the measurements (Bujang et al., 2018).

Table 2. *Interclass correlation coefficients analysis*

	Intraclass Correlation	95% Confidence Interval		
		Lower Bound	Upper Bound	p-value
SRQ	0.803	0.747	0.848	<0.001
WDRQ	0.803	0.710	0.869	<0.001
SRQ+ WDRQ	0.801	0.754	0.840	<0.001

The researchers used the Kruskal-Wallis H test to assess five LLMs' performance differences across complexity levels. Results for ChatGPT 3.5 (p = 0.434), ChatGPT 4.0 (p = 0.177), Claude Sonnet (p = 0.885), and Gemini Pro (p = 0.145) indicated no statistically significant differences, suggesting that complexity did not meaningfully impact their performance, possibly due to sample size limitations (Mishra et al., 2019; Yan et al., 2024). In contrast, Copilot showed a significant effect (p = 0.014), prompting further analysis via Dunn's test. Post hoc results revealed significant differences between intermediate- and advanced-level queries (p = 0.021) and basic- and advanced-level queries (p = 0.007). In contrast, basic vs. intermediate showed no significant difference (p = 0.689) (Mishra et al., 2019).

Pairwise model comparisons further highlighted Copilot's performance disparity. Researchers found significant differences between Copilot and ChatGPT 3.5 (p = 0.010), ChatGPT 4.0 (p < 0.001), Claude Sonnet (p = 0.003), and Gemini Pro (p < 0.001). Similarly, ChatGPT 4.0 significantly outperformed ChatGPT 3.5 (p < 0.001), Claude Sonnet (p < 0.001), and Gemini Pro (p = 0.004). In contrast, no significant differences were observed between ChatGPT 3.5 and Claude Sonnet (p = 0.732), ChatGPT 3.5 and Gemini Pro (p = 0.204), or Claude Sonnet and Gemini Pro (p = 0.353), suggesting broadly comparable performance among these three models.

Frequency

Figure 2 presents the score distributions for five LLM chatbots, highlighting marked differences in performance. Low-reliability responses (scores ≤5) were most frequent in Copilot (29.4%), Claude Sonnet (26.1%), and Gemini Pro (19.1%), while ChatGPT 3.5 and 4.0 recorded lower rates at 16% and 8%, respectively, indicating superior reliability in the ChatGPT models, particularly version 4.0. Notably, 52% of ChatGPT 4.0's outputs scored in the top bracket (9–10), outperforming Claude Sonnet and Gemini Pro (42% each), ChatGPT 3.5 (27%), and Copilot (18%). In the mid-range (scores 6–8), ChatGPT 3.5 led with 56%, followed by ChatGPT 4.0 (40%), Gemini Pro (39%), Copilot (53%), and Claude Sonnet (31%). These results affirm ChatGPT 4.0's overall dominance in reliability and accuracy, demonstrating its consistent ability to generate high-quality responses across varying complexity levels.

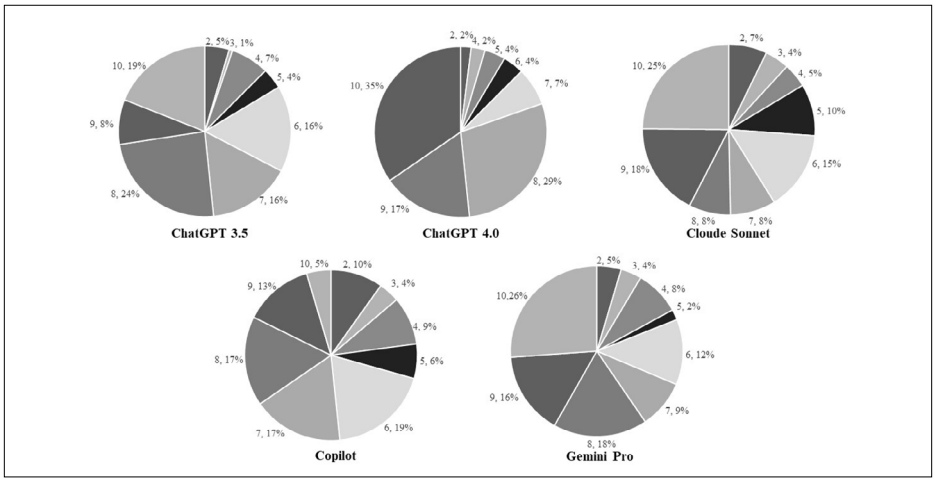

Figure 2. *Score distributions of LLM chatbots*

Descriptive Statistics

Descriptive statistics offer a comparative overview of LLM chatbot performance across SRQ and WDRQ tasks at varying complexity levels (see Table 3). ChatGPT 4.0 consistently outperformed other models, particularly excelling in SRQ-intermediate queries with a mean score of 26.83, reflecting its strength in handling moderately complex legal tasks. Gemini Pro also performed well, especially in WDRQ-basic, where it achieved a high mean score of 26.20. In contrast, ChatGPT 3.5 showed balanced performance across both SRQ and WDRQ but struggled with advanced drafting tasks, as evidenced by a drop in mean score from 22 to 16.40. However, it demonstrated the capacity for high-quality output, reaching maximum scores of 28 (SRQ) and 23 (WDRQ). Claude Sonnet and Copilot exhibited greater performance variability, particularly in advanced queries, with lower mean scores (19.40 and 17, respectively) and higher standard deviations (6.88 and 4.53), suggesting limitations in managing complex legal reasoning and drafting.

Table 3. *Descriptive statistics*

LLM Chatbot	Mean	Standard Deviation	Minimum	Maximum
ChatGPT 3.5				
SRQ Basic	23.17	3.881	15	28
SRQ Intermediate	21.58	4.078	14	28
SRQ Advanced	23.50	5.196	16	28
WDRQ Basic	22.00	1.000	21	23
WDRQ Intermediate	20.20	1.924	18	23
WDRQ Advanced	16.40	4.930	10	23
ChatGPT 4.0				
SRQ Basic	24.67	1.723	22	28
SRQ Intermediate	26.83	1.992	22	29
SRQ Advanced	26.00	2.629	19	28
WDRQ Basic	25.40	1.817	23	27
WDRQ Intermediate	24.40	2.074	21	26
WDRQ Advanced	20.40	6.107	10	26
Claude Sonnet				
SRQ Basic	21.00	5.721	6	26
SRQ Intermediate	20.42	6.921	8	27
SRQ Advanced	22.58	5.125	11	26
WDRQ Basic	22.20	5.450	13	27
WDRQ Intermediate	23.60	1.342	22	25
WDRQ Advanced	19.40	6.877	11	27
Copilot				
SRQ Basic	19.00	5.427	7	25
SRQ Intermediate	19.08	6.921	6	27
SRQ Advanced	16.50	4.964	8	22
WDRQ Basic	24.00	0.707	23	25
WDRQ Intermediate	21.60	1.673	19	23
WDRQ Advanced	17.00	4.528	10	21

Table 3. *continued*

LLM Chatbot	Mean	Standard Deviation	Minimum	Maximum
Gemini Pro				
SRQ Basic	22.67	3.676	16	28
SRQ Intermediate	24.00	4.023	17	28
SRQ Advanced	19.00	8.169	8	28
WDRQ Basic	26.20	1.304	24	27
WDRQ Intermediate	23.60	2.510	20	26
WDRQ Advanced	21.20	3.834	16	26

Abbreviations: SRQ, succession-related queries; WDRQ, will drafting-related queries.

The analysis reveals a consistent pattern: all LLM chatbots performed better on basic and intermediate queries than advanced ones, underscoring their limitations in addressing complex legal tasks (Alawida et al., 2023). This decline in performance suggests that while these models are increasingly adept at handling straightforward inquiries, their capacity to manage nuanced legal reasoning remains underdeveloped. High score variability in advanced queries further indicates reliability concerns, particularly in maintaining consistency under greater complexity. ChatGPT 3.5 and 4.0 notably showed marked performance drops in advanced WDRQ tasks, pointing to a broader challenge that warrants targeted improvements in model training. Among the evaluated models, ChatGPT 4.0 demonstrated the most consistent accuracy across complexity levels, with Gemini Pro showing strong results in less complex queries (Lim et al., 2023).

LLM Chatbots Performance

Figure 3 delineates the performance of five LLM chatbots in handling SRQ and WDRQ. Across all levels of SRQ complexity, ChatGPT 4.0 emerged as the most consistently reliable model. The panel distinguished its responses with a high degree of legal precision, accessible language tailored for non-experts, and a nuanced awareness of Malaysia's jurisdictional distinctions. The expert panel reached a unanimous consensus on its superior performance, particularly noting the model's clarity, legal soundness, and practical applicability in the context of will-writing. These qualities underscore its potential as a dependable tool for facilitating informed and accurate succession planning.

In contrast, Copilot underperformed across the board, revealing significant limitations in legal reasoning and drafting accuracy. For SRQ specifics, ChatGPT 4.0 and Gemini Pro show pronounced proficiency at the intermediate complexity and advanced and basic levels, respectively. ChatGPT 3.5 and Claude Sonnet were most effective at the advanced level, with performance tapering at the basic and intermediate levels. Interestingly, Copilot's best results were at the intermediate level, descending to the basic and least at the advanced level, challenging the expected norm that performance should improve from advanced to basic levels.

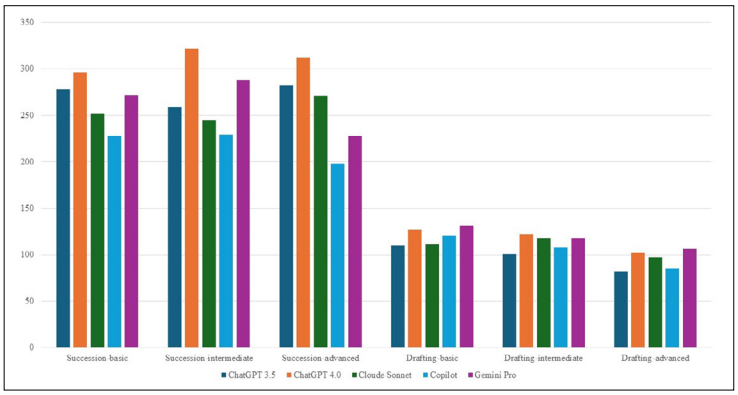

Figure 3. *LLM chatbots performance*

Performance variations were more pronounced across complexity levels. For SRQ-basic queries, ChatGPT 4.0 scored the highest average (24.67), followed by ChatGPT 3.5 and Gemini Pro. Although their mean scores were acceptable, their lower minimum scores (15 and 16, respectively) indicate potential reliability risks, especially for users lacking legal expertise who may not detect subtle inaccuracies, highlighted by Expert 1. In the SRQ-intermediate tier, ChatGPT 4.0 again led with the highest aggregated and average scores, with Gemini Pro performing relatively well but showing inconsistency. Other models, including Claude Sonnet and Copilot, fell short of the expert-defined thresholds for dependable responses.

For SRQ-intermediate, ChatGPT 4.0 again leads with reliability, scoring an aggregated 322 and achieving a maximum score of 29 and an average of 26.83. Gemini Pro also demonstrates commendable reliability with an aggregated score of 288, though its minimum score of 17 suggests variability in performance. The experts agreed that a minimum individual score of 7 or an aggregated score of 21 for a response is necessary for laypersons to perceive the responses as dependable. Given the lower scores, ChatGPT 3.5, Claude Sonnet, and Copilot are deemed less reliable for these queries.

At the SRQ-advanced level, both versions of ChatGPT showed strong performance. ChatGPT 4.0 outshines its counterparts with an aggregated score of 312, an average score of 26, and a minimum of 19. The expert panel submitted that ChatGPT 4.0 generates more improved responses, demonstrates an enhanced understanding of queries, uses layperson's language over legal jargon, and can distinguish laws applicable to different regions in Malaysia. Claude Sonnet also showed promise in handling complex legal reasoning with an aggregated score of 271 and a mean score of 22.58. However, experts noted its inconsistency across queries—experts 1 and 3 highlight Claude Sonnet's varied performance as a significant observation. However, Copilot and Gemini Pro struggled significantly when interpreting complex or jurisdiction-specific queries.

Reliability declined across all models in the WDRQ category, which involved actual drafting tasks as query complexity increased. Gemini Pro outperformed other models at the basic level, followed closely by ChatGPT 4.0 and Copilot (refer to Figure 3). ChatGPT 4.0 shows robust performance at the basic and advanced levels by attaining the second spot. Conversely, ChatGPT 3.5 lags at all levels of WDRQ complexity, illustrating a significant variance in performance between SRQ and WDRQ. This disparity underscores that proficiency in SRQ does not necessarily translate into equivalent reliability in drafting responses, marking a critical consideration for users relying on these models for complex drafting tasks.

For basic WDRQs involving straightforward and specific testamentary instructions, Gemini Pro leads with the highest aggregated score, demonstrating strong reliability, closely followed by ChatGPT 4.0 and Copilot. Claude Sonnet and ChatGPT 3.5 also show commendable reliability, though slightly less consistent. Notably, Claude Sonnet exhibits the broadest range between minimum and maximum scores, with Expert 2 indicating a misunderstanding in the bequest instruction, resulting in a low satisfactory response. The other two experts collaborated with Expert 2 by giving low scores. In contrast, Expert 1 highlights that Gemini Pro was lauded for its precise drafting, which was deemed accurate and sufficient for laypersons drafting simple wills.

ChatGPT 4.0 maintained high reliability at the intermediate drafting level, while Claude Sonnet and Gemini Pro showed moderate success. However, Copilot and ChatGPT 3.5 only achieved moderate reliability, with ChatGPT 3.5 particularly noted for its lesser adequacy in handling more complex instructions, as highlighted by Expert 2. Notably, Claude Sonnet was the only model that did not experience a noticeable performance drop between basic and intermediate queries.

Advanced WDRQs posed the most significant challenge. Only ChatGPT 4.0 and Gemini Pro produced moderately reliable drafts under incomplete information. Other chatbots, Claude Sonnet, Copilot, and ChatGPT 3.5, frequently introduced ambiguities or omitted essential legal components. Expert 1 points out that despite the clarity of the language used, some of the drafted clauses were legally incorrect. Expert 2 observes that specific clauses contained ambiguities, which could lead to difficulties in determining the testators' intentions in court. Expert 3 noted that the inability of the LLM chatbots to recognize the implications of omitted information resulted in several clauses being unreliable and potentially misleading to users. These issues carry profound implications for users relying solely on AI-generated wills without legal oversight.

Qualitative Assessment of LLM Chatbots Performance

Beyond quantitative scores, the expert panel observed that none of the LLM chatbots consistently performed across similar query types or complexity levels. All models exhibited varying degrees of inconsistency, particularly in citing relevant legislation, interpreting legal provisions, and grasping the core intent of queries. While some responses were concise and legally sound, earning perfect scores, such quality was not reliably reproduced. This internal variability suggests that chatbots often deliver a mix of accurate and flawed outputs even at the same difficulty level, undermining their reliability in assisting Malaysian non-Muslims with will preparation.

Despite these shortcomings, the panel acknowledged some commendable patterns. All chatbots consistently advised users to consult legal professionals in response to will-drafting queries. However, they struggled to distinguish between jurisdictional differences across Peninsular Malaysia, Sabah, and Sarawak, failed to recognize that vehicles cannot have multiple registered owners, and, except for Gemini Pro, overlooked the importance of including national identification numbers. Further concerns include frequent misinterpretation of queries. While factually correct, some responses were contextually irrelevant, while others reflected legal and interpretive errors. These issues are particularly problematic for users without legal training, who may be unable to judge response reliability. The panel also flagged imprecise language, for example, using "cannot" instead of "should not" or "may not" instead of "will not," as potentially misleading.

Chatbots often include excessive and sometimes inaccurate information without any discernible pattern. At the same time, some elaboration may be helpful, but excessive detail, particularly when legally incorrect, risks confusing users. The panel emphasized that practical legal guidance should be clear, concise, and appropriately scoped, especially for laypersons.

DISCUSSION

More people use generative AI tools to support user decision-making across various domains. However, Miller (2019) highlights a critical gap in developers' appreciation of the role of explanation in AI systems. Similarly, Lim and Taeihagh (2019) underscore the limited adaptability of most AI applications to diverse user contexts. In will-writing scenarios, ChatGPT 4.0 consistently outperformed other LLM chatbots across different levels of query complexity, aligning with earlier findings (Lim et al., 2023) and affirming its reliability for assisting non-Muslims during decision-making stages. With strong aggregated scores, it inspires greater confidence in output quality. ChatGPT 3.5 and Gemini Pro also performed well for general succession inquiries, though Gemini Pro is better suited to simpler tasks. While ChatGPT 3.5 struggled with detailed queries, it and Claude Sonnet showed promise in handling more complex legal issues. However, given the general public's limited legal knowledge, users may misjudge query complexity, making it advisable to use ChatGPT 3.5, Claude Sonnet, and Gemini Pro cautiously. Errors in their responses risk leading to faulty testamentary instructions. Gemini Pro, ChatGPT 4.0, and Copilot could generate competent legal drafts for basic will-drafting tasks. In contrast, more detailed testamentary instructions are best handled by ChatGPT 4.0, which remains the most dependable among current tools. Nonetheless, all tested chatbots, including ChatGPT 4.0, struggled when key information was missing, highlighting a persistent limitation in managing complex, incomplete instructions.

Contrary to expectations, performance did not consistently decline with increasing query complexity. While some chatbots exhibited this pattern in WDRQ, it was absent in SRQ, notably for Claude Sonnet in WDRQ. Surprisingly, ChatGPT 3.5 and Claude Sonnet performed better at the most advanced SRQ levels, raising concerns about algorithmic consistency and reliability (Glikson & Woolley, 2020). Expert panel feedback reinforced these concerns, citing issues such as incomplete understanding, omission of relevant laws, misuse of legal terminology, and errors that can significantly distort meaning for legal professionals and lay users. The interchangeable use of terms like "may not" and "will be" was particularly problematic, affecting scoring and interpretation. Other chatbot evaluations have noted similar issues (Alawida et al., 2023).

AI users criticize LLM chatbots for delivering overly detailed and occasionally irrelevant explanations, which can confuse users, particularly in will-drafting contexts (Hassani & Silva, 2023). Rather than focusing on core legal guidance, these tools often digress into tangential topics, overlooking critical issues such as the distinction between specific and general legacies or the legal impact of divorce on a will's validity. Some responses included inappropriate discussions on residuary or small estates while failing to mention that a newer will typically revoke prior ones. Such lapses reflect a broader issue of inconsistency and lack of relevance in chatbot-generated legal content, a concern also raised by Hassani and Silva (2023) and Wach et al. (2023). This problem may stem from an inability to manage linguistic complexity and ensure coherent flow, as Deveci et al. (2023) noted, undermining readability and user comprehension.

The expert panel further identified a recurring weakness in LLMs' ability to process complex legal queries. According to Expert 1, many responses lacked depth or included factual errors, indicating a limited understanding of nuanced issues. This finding aligns with observations by Kooli (2023) and Meyer et al. (2023), who attributed such deficiencies to insufficient contextual awareness. These limitations suggest that current LLM chatbots are not yet equipped to handle intricate legal scenarios, particularly those involving multifaceted testamentary wishes or complex family dynamics, echoing Lim and Taeihagh's (2019) broader critique of AI's contextual inflexibility. Omitting practical considerations when dealing with complex testamentary desires or intricate family dynamics can harm users.

Furthermore, the expert panel observed that LLM chatbots performed more reliably when queries were specific and complete. Their effectiveness declined at higher levels of the WDRQ, where missing information became more frequent, posing serious risks when legally untrained users attempted to draft complex wills. This concern is heightened by the irrevocable nature of wills once executed after death, making precision at the drafting stage critical (Choi & Carr, 2023). While all chatbots included disclaimers encouraging users to consult legal professionals, a practice endorsed by Bazzari and Bazzari (2024), this safeguard does not compensate for structural shortcomings.

A recurring issue was the omission of key details. Most LLMs, except Gemini Pro, failed to include basic identifiers such as the National Registration Identity Card (NRIC) number. Although not legally required, such details enhance clarity. ChatGPT 3.5's exclusion of the witness attestation clause was more concerning, a legal necessity under Section 5(2) of the Wills Act 1959. Its absence jeopardizes the will's validity and may trigger costly delays in probate (Nasrul & Mohd Salim, 2018).

Ambiguities in chatbot-generated clauses, particularly revocation provisions, were another critical flaw. The expert panel found ChatGPT 3.5's clauses unclear and overly complex, potentially obscuring the testator's intent. These issues highlight the limitations of LLMs in managing legal nuance and reinforce the need for human oversight in sensitive legal drafting. While AI tools may produce surface-level compliant documents, the subtleties of legal language remain a domain where human expertise is indispensable (Atkinson et al., 2020; Surden, 2019). As Meyer et al. (2023) and Taeihagh (2021) noted, such ambiguities may lead to unintended legal consequences, reinforcing the broader risks associated with AI-generated legal outputs.

Implications

This study highlights the growing potential of LLM chatbots to perform complex legal tasks, such as will drafting, which may reshape the delivery of legal services. Demonstrating the competence of ChatGPT 4.0 and similar models in handling succession-related queries, the findings suggest that these tools could become valuable assets in estate planning, particularly by improving access to legal assistance. In contexts where users perceive legal services as costly or intimidating, such accessibility may encourage more individuals, especially non-Muslims, to engage in proactive estate planning and reduce the legal complications of intestacy.

While focused on Malaysia, the implications are globally relevant, pointing to the broader applicability of LLMs across legal systems. Their integration could enable legal practitioners to delegate routine tasks to AI, allowing greater focus on complex legal matters. This evolution signals a shift toward a more efficient, hybrid legal practice model.

The findings also offer guidance for policymakers developing regulatory frameworks to ensure the ethical deployment of AI in legal contexts. Robust standards are needed to protect users from misinformation and mitigate risks arising from system inaccuracies. Moreover, the research underscores the importance of aligning legal education with technological advancements. Incorporating AI into legal training will better prepare future professionals to work alongside intelligent systems, ultimately enhancing the quality and efficiency of legal services.

CONCLUSION

Advancements in AI have dramatically empowered users to handle multidisciplinary tasks independently, revolutionizing how individuals approach tasks that traditionally require professional expertise. Particularly in Malaysia, where the law allows individuals to draft their wills, LLM chatbots hold significant promise for assisting non-Muslim Malaysians without legal training in the will-preparation process. The TAM suggests that generative AI tools are more likely to be utilized for will preparation if perceived as valuable and user-friendly. This study critically examines the performance of several LLM chatbots in assisting Malaysian non-Muslims with will preparation, focusing on key criteria such as accuracy, legal validity, comprehensibility, and overall reliability. The findings reveal a mixed landscape: while specific platforms deliver reasonably dependable support for basic testamentary tasks, others fall short of the consistency and precision required for practical legal guidance. To advance this line of research, future studies should adopt longitudinal approaches to better capture the evolving capabilities of LLM chatbots in legal contexts. Expanding the scope to include other generative AI platforms and extending the analysis to Muslim users or cross-jurisdictional applications would yield broader insights into their utility and reliability. In addition, exploring the use of AI for drafting a wider range of legal documents, such as employment contracts or tenancy agreements, could significantly deepen understanding of AI's role in democratizing legal services. Given the legal documentation's complexity and high stakes, future work should also consider frameworks for integrating expert oversight into AI-assisted processes to ensure legal soundness and reduce risk.

Funding
This research received no external funding.

Ethical Approval
The author declares no involvement of humans or animals, and no informed consent is needed.

Competing interest
The author declares no conflicts of interest.

Data Availability
All the data generated or analyzed during this study are included in this published article.

Declaration of Artificial Intelligence Use
During the preparation of this work the author used ChatGPT to improve the language and readability of the paper. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Acknowledgment
The author wishes to express his profound gratitude to the ten estate planning professionals and three estate planning and legal experts who generously contributed their time, expertise, and invaluable insights to this empirical study. Their unwavering support and cooperation were instrumental in the successful completion of this research.

REFERENCES

Alawida, M., Mejri, S., Mehmood, A., Chikhaoui, B., & Isaac Abiodun, O. (2023). A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. Information, 14(8), 462. https://doi.org/10.3390/info14080462

Armour, J., & Sako, M. (2020). AI-enabled Business Models in Legal Services: From Traditional Law Firms to Next-generation Law Companies? Journal of Professions and Organization, 7(1), 27–46. https://doi.org/10.1093/jpo/joaa001

Aslan, I., Çınar, O., & Özen, Ü. (2014). Developing Strategies for the Future of Healthcare in Turkey by Benchmarking and SWOT Analysis. Procedia - Social and Behavioral Sciences, 150, 230–240. https://doi.org/10.1016/j.sbspro.2014.09.043

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and Law: Past, Present and Future. Artificial Intelligence, 289, 103387. https://doi.org/10.1016/j.artint.2020.103387

Aubry, J.-P., Munnell, A. H., & Wettstein, G. (2023). Can Incentives Increase the Writing of Wills? An Experiment (Center for Retirement Research).

Balagurunathan, Y., Mitchell, R., & El Naqa, I. (2021). Requirements and Reliability of AI in the Medical Context. Physica Medica, 83, 72–78. https://doi.org/10.1016/j.ejmp.2021.02.024

Basir, F., Ahmad, W., & Rahman, M. (2023). Estate Planning Behaviour: A Systematic Literature Review. Journal of Risk and Financial Management, 16(2), 84. https://doi.org/10.3390/jrfm16020084

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law and Technology, 31(2), 889–934. https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

Bazzari, F. H., & Bazzari, A. H. (2024). Utilizing ChatGPT in Telepharmacy. Cureus, 16(1), e52365. https://doi.org/10.7759/cureus.52365

Bench-Capon, T., Araszkiewicz, M., Ashley, K., Atkinson, K., Bex, F., Borges, F., Bourcier, D., Bourgine, P., Conrad, J. G., Francesconi, E., Gordon, T. F., Governatori, G., Leidner, J. L., Lewis, D. D., Loui, R. P., McCarty, L. T., Prakken, H., Schilder, F., Schweighofer, E., … Wyner, A. Z. (2012). A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law. Artificial Intelligence and Law, 20(3), 215–319. https://doi.org/10.1007/s10506-012-9131-x

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. Malaysian Journal of Medical Sciences, 25(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Bujanga, M. A., & Baharum, N. (2017). A Simplified Guide to Determination of Sample Size Requirements for Estimating the Value of Intraclass Correlation Coefficient: A Review. Archives of Orofacial Sciences, 12(1), 1–12. https://aos.usm.my/docs/Vol_12/aos-article-0246.pdf

Cheng, E. W. L. (2019). Choosing Between the Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM). Educational Technology Research and Development, 67(1), 21–37. https://doi.org/10.1007/s11423-018-9598-6

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT Goes to Law School. Journal of Legal Education, 71(3), 387-400. https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=2055&context=faculty_articles

Choi, S. L., & Carr, D. (2023). Older Adults' Relationship Trajectories and Estate Planning. Journal of Family and Economic Issues, 44(2), 356–372. https://doi.org/10.1007/s10834-022-09839-y

Choi, S. L., McDonough, I. M., Kim, M., & Kim, G. (2019). Estate Planning Among Older Americans: The Moderating Role of Race and Ethnicity. Financial Planning Review, 2(3–4), e1058. https://doi.org/10.1002/cfp2.1058

Cox, D., & Stark, O. (2005). Bequests, Inheritances and Family Traditions. SSRN Electronic Journal, Art. WP#2005-09. https://doi.org/10.2139/ssrn.1148982

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science, 35(8), 982–1003. https://doi.org/10.1287/mnsc.35.8.982

Deveci, C. D., Baker, J. J., Sikander, B., & Rosenberg, J. (2023). A Comparison of Cover Letters Written by ChatGPT-4 or Humans. Danish Medical Journal, 70(12), A06230412.

Egli, A. (2023). ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology? Clinical Infectious Diseases, 77(9), 1322–1328. https://doi.org/10.1093/cid/ciad407

Egli, A., Schrenzel, J., & Greub, G. (2020). Digital Microbiology. Clinical Microbiology and Infection, 26(10), 1324–1331. https://doi.org/10.1016/j.cmi.2020.06.023

Emenike, M. E., & Emenike, B. U. (2023). Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-generation Software Programs for Chemists and Chemistry Educators. Journal of Chemical Education, 100(4), 1413–1418. https://doi.org/10.1021/acs.jchemed.3c00063

Gazulla, E. D., Martins, L., & Fernández-Ferrer, M. (2023). Designing Learning Technology Collaboratively: Analysis of A Chatbot Co-design. Education and Information Technologies, 28(1), 109–134. https://doi.org/10.1007/s10639-022-11162-w

Gill, A., Mand, H. S., Obradovich, J. D., & Mathur, N. (2017). Influence of Meditation on Estate Planning Decisions: Evidence from Indian Survey Data. Financial Innovation, 3(1), 27. https://doi.org/10.1186/s40854-017-0078-5

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Medical Education, 9, e45312. https://doi.org/10.2196/45312

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

González-Estrada, E., Villaseñor, J. A., & Acosta-Pech, R. (2022). Shapiro-Wilk Test for Multivariate Skew-normality. Computational Statistics, 37(4), 1985–2001. https://doi.org/10.1007/s00180-021-01188-y

Granić, A., & Marangunić, N. (2019). Technology Acceptance Model in Educational Context: A Systematic Literature Review. British Journal of Educational Technology, 50(5), 2572–2593. https://doi.org/10.1111/bjet.12864

Harrington, B. (2012). Trust and Estate Planning: The Emergence of a Profession and Its Contribution to Socioeconomic Inequality. Sociological Forum, 27(4), 825–846. https://doi.org/10.1111/j.1573-7861.2012.01358.x
Hassani, H., & Silva, E. S. (2023). The Role of ChatGPT in Data Science: How AI-assisted Conversational Interfaces are Revolutionizing the Field. Big Data and Cognitive Computing, 7(2), 62. https://doi.org/10.3390/bdcc7020062
Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability Generalization of Perceived Ease of Use, Perceived Usefulness, and Behavioral Intentions. MIS Quarterly, 38(1), 1–28. https://doi.org/10.25300/MISQ/2014/38.1.01
Horkey, C. (2009). Estate Planning Documents in Virginia Among Adults 50 and Over with at Least One Adult Child. Virginia Polytechnic Institute and State University. https://vtechworks.lib.vt.edu/items/6c86840e-2a52-44e9-ab5f-8b9c97ac168c/full
Horton, D. (2017). Tomorrow's Inheritance: The Frontiers of Estate Planning Formalism. Boston College Law Review, 58(2), 539–598. https://bclawreview.bc.edu/articles/468
Ismail, S., Hashima, N., Kamisa, R., Harunb, H., & Samad, N. N. A. (2013). Determinants of Attitude towards Estate Planning in Malaysia: An Empirical Investigation. Conference: International Conference on Economics & Business Research, 1–9. https://www.researchgate.net/publication/260986225_Determinants_of_Attitude_towards_Estate_Planning_in_Malaysia_An_Empirical_Investigation
Izenman, A. J. (2008). Modern Multivariate Statistical Techniques. Springer New York. https://doi.org/10.1007/978-0-387-78189-1
Kim, K. T., & Stebbins, R. (2021). Everybody Dies: Financial Education and Basic Estate Planning. Journal of Financial Counseling and Planning, 32(3), 402–416. https://doi.org/10.1891/JFCP-19-00076
Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
Kooli, C. (2023). Artificial Intelligence Dissociative Identity Disorder (AIDIS): The Dark Side of ChatGPT. QScience Connect, 2023(2). https://doi.org/10.5339/connect.2023.2
Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education Using Large Language Models. PLOS Digital Health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198
Lee, D. Y., & Lehto, M. R. (2013). User Acceptance of YouTube for Procedural Learning: An Extension of the Technology Acceptance Model. Computers & Education, 61, 193–208. https://doi.org/10.1016/j.compedu.2012.10.001
Lim, H. S. M., & Taeihagh, A. (2019). Algorithmic Decision-making in AVs: Understanding Ethical and Technical Concerns for Smart Cities. Sustainability, 11(20), 5791. https://doi.org/10.3390/su11205791
Lim, Z. W., Pushpanathan, K., Yew, S. M. E., Lai, Y., Sun, C.-H., Lam, J. S. H., Chen, D. Z., Goh, J. H. L., Tan, M. C. J., Sheng, B., Cheng, C.-Y., Koh, V. T. C., & Tham, Y.-C. (2023). Benchmarking Large Language Models' Performances for Myopia Care: A Comparative Analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine, 95, 104770. https://doi.org/10.1016/j.ebiom.2023.104770
Madura, J., & Gill, H. S. (2015). Personal Finance, Third Canadian Edition. Pearson Canada.
Mendenhall, E., Muzizi, L., Stephenson, R., Chomba, E., Ahmed, Y., Haworth, A., & Allen, S. (2007). Property Grabbing and Will Writing in Lusaka, Zambia: An Examination of Wills of HIV-infected Cohabiting Couples. AIDS Care, 19(3), 369–374. https://doi.org/10.1080/09540120600774362
Mendoza, S., Sánchez-Adame, L. M., Urquiza-Yllescas, J. F., González-Beltrán, B. A., & Decouchant, D. (2022). A Model to Develop Chatbots for Assisting the Teaching and Learning Process. Sensors, 22(15), 5532. https://doi.org/10.3390/s22155532
Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and Large Language Models in Academia: Opportunities and Challenges. BioData Mining, 16(1), 20. https://doi.org/10.1186/s13040-023-00339-9
Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
Mishra, P., Pandey, C., Singh, U., Keshri, A., & Sabaretnam, M. (2019). Selection of Appropriate Statistical Methods for Data Analysis. Annals of Cardiac Anaesthesia, 22(3), 297. https://doi.org/10.4103/aca.ACA_248_18
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679
Mojadeddi, Z. M. R. J. (2023). The Impact of AI and ChatGPT on Research Reporting. The New Zealand Medical Journal (Online), 136(1575), 60–64. https://doi.org/10.26635/6965.6122
Müller, R., & Büttner, P. (1994). A Critical Discussion of Intraclass Correlation Coefficients. Statistics in Medicine, 13(23–24), 2465–2476. https://doi.org/10.1002/sim.4780132310
Nasrul, M. A., & Mohd Salim, W. N. (2018). Administration of Estates in Malaysia: Determinant of Factors Behind the Delay in the Distribution of the Deceased's Asset. Journal of Nusantara Studies, 3(1), 75–86. https://doi.org/10.24200/jonus.vol3iss1pp75-86
Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., … Wong, L.-W. (2025). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. Journal of Computer Information Systems, 65(1), 76–107. https://doi.org/10.1080/08874417.2023.2261010
Piya, S., Shamsuzzoha, A., Khadem, M., & Al Kindi, M. (2021). Integrated Analytical Hierarchy Process and Grey Relational Analysis Approach to Measure Supply Chain Complexity. Benchmarking: An International Journal, 28(4), 1273–1295. https://doi.org/10.1108/BIJ-03-2020-0108
Rahmanti, A. R., Yang, H.-C., Bintoro, B. S., Nursetyo, A. A., Muhtar, M. S., Syed-Abdul, S., & Li, Y.-C. J. (2022). SlimMe, A Chatbot with Artificial Empathy for Personal Weight Management: System Design and Finding. Frontiers in Nutrition, 9, 870775. https://doi.org/10.3389/fnut.2022.870775

Rissland, E. L., Ashley, K. D., & Loui, R. P. (2003). AI and Law: A Fruitful Synergy. Artificial Intelligence, 150(1–2), 1–15. https://doi.org/10.1016/S0004-3702(03)00122-X
Sætra, H. S. (2020). A Shallow Defence of A Technocracy of Artificial Intelligence: Examining the Political Harms of Algorithmic Governance in the Domain of Government. Technology in Society, 62, 101283. https://doi.org/10.1016/j.techsoc.2020.101283
Solomovich, L., & Abraham, V. (2024). Exploring the Influence of ChatGPT on Tourism Behavior Using the Technology Acceptance Model. Tourism Review, ahead-of-print(ahead-of-print). https://doi.org/10.1108/TR-10-2023-0697
Street, M. (2006). A Holistic Approach to Estate Planning: Paramount in Protecting Your Family, Your Wealth, and Your Legacy. Pepperdine Dispute Resolution Law Journal, 7(1), 141–163. https://digitalcommons.pepperdine.edu/drlj/vol7/iss1/5/
Surden, H. (2014). Machine Learning and Law. Washington Law Review, 89(1), 87–106. https://digitalcommons.law.uw.edu/wlr/vol89/iss1/5
Surden, H. (2019). Artificial Intelligence and Law: An Overview. Georgia State University Law Review, 35(4), 1305–1337. https://readingroom.law.gsu.edu/gsulr/vol35/iss4/8
Susskind, R. (2019). Online Courts and the Future of Justice. Oxford University Press. https://doi.org/10.1093/oso/9780198838364.001.0001
Taeihagh, A. (2021). Governance of Artificial Intelligence. Policy and Society, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377
Tomita, S., Komiya-Ito, A., Imamura, K., Kita, D., Ota, K., Takayama, S., Makino-Oi, A., Kinumatsu, T., Ota, M., & Saito, A. (2013). Prevalence of Aggregatibacter Actinomycetemcomitans, Porphyromonas Gingivalis and Tannerella Forsythia in Japanese Patients with Generalized Chronic and Aggressive Periodontitis. Microbial Pathogenesis, 61–62, 11–15. https://doi.org/10.1016/j.micpath.2013.04.006
Trajtenberg, M. (2019). Artificial Intelligence as the Next GPT: A Political-economy Perspective. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), The Economics of Artificial Intelligence: An Agenda (pp. 175–186). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0006
Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540
Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly, 36(1), 157–178. https://doi.org/10.2307/41410412
Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The Dark Side of Generative Artificial Intelligence: A Critical Analysis of Controversies and Risks of ChatGPT. Entrepreneurial Business and Economics Review, 11(2), 7–30. https://doi.org/10.15678/EBER.2023.110201
Weeks, R., Sangha, P., Cooper, L., Sedoc, J., White, S., Gretz, S., Toledo, A., Lahav, D., Hartner, A.-M., Martin, N. M., Lee, J. H., Slonim, N., & Bar-Zeev, N. (2023). Usability and Credibility of a COVID-19 Vaccine Chatbot for Young Adults and Health Workers in the United States: Formative Mixed Methods Study. JMIR Human Factors, 10, e40533. https://doi.org/10.2196/40533
Westwood, S. (2015). Complicating Kinship and Inheritance: Older Lesbians' and Gay Men's Will-Writing in England. Feminist Legal Studies, 23(2), 181–197. https://doi.org/10.1007/s10691-015-9287-3
Yan, Y., Oswald, E., & Roy, A. (2024). Not Optimal but Efficient: A Distinguisher Based on the Kruskal-Wallis Test. In H. Seo & S. Kim (Eds.), Information Security and Cryptology – ICISC 2023 (pp. 240–258). Springer Singapore. https://doi.org/10.1007/978-981-97-1235-9_13
Zandi, G. R., Abidin, S. Z., & Swee, K. N. (2017). The Preparations of Employees Towards Retirement and Estate Planning: The Case of Malaysia. International Journal of Applied Business and Economic Research, 15(22), 673–684. https://serialsjournals.com/abstract/32784_ch_51_f_-_gholamreza_zandi.pdf

How to cite this article:
Seow, R. Y. C. (2025). Using Large Language Models to Transform Estate Planning. Recoletos Multidisciplinary Research Journal 13(1), 185-199. https://doi.org/10.32871/rmrj2513.01.15