LLM Evaluations for Bharat Tax Laws (‘LE-BTL’)

A Framework to Evaluate and Benchmark the Accuracy of Large Language Models (‘LLMs’) in the Context of the Indian Tax Laws

Nitish Jain, Naveen Wadhwa[1], Pawan Goyal[2], Saptarshi Ghosh[2], Sankalp Pawar, Abhishek Shinde, Karthik Boinepally, Vrinda V Malpani, Raaga K

[1] Taxmann.AI, India; [2] IIT Kharagpur, West Bengal, India

{nitishj, pawang.iitk, saptarshi.ghosh, pawar.sankalp.sanjay, caabhishekshinde, karthikboinepally, vrinda.v.malpani, raaga.upr} @gmail.com
naveen@taxmann.ai

Abstract

LLMs are increasingly being used to support legal and tax workflows by enhancing research efficiency, facilitating compliance analysis, and automating document management. However, the intricate and evolving nature of Indian tax law, marked by frequent amendments, complex statutory interpretation, and divergent judicial precedents, requires domain-specific evaluation tools. Existing global LLM benchmarks fail to adequately assess jurisdictional legal reasoning in this context.

Benchmarking serves as a critical tool for evaluating the performance and effectiveness of LLMs, ensuring that they meet the specific requirements and challenges posed by the Indian tax system. By establishing clear benchmarks, we can systematically assess the capabilities of LLMs in handling complex tax-related queries, providing accurate interpretations, and delivering reliable outcomes.

Through the paper, we present the LE-BTL framework, the first benchmark designed to systematically evaluate LLM performance in Indian tax law. Built in collaboration with tax law professionals, the benchmark uses a structured IRAC+ framework (Issue, Rule, Application, Conclusion, Interpretation and Justification) to evaluate models on realistic legal tasks. Twelve leading LLMs are assessed across three prompting strategies: base, persona, and few-shot persona, highlighting the interplay between model architecture, instruction tuning, and the depth of legal reasoning.

Our results reveal a clear performance stratification: proprietary models such as GPT o3 Pro, GPT o3 and Gemini 2.5 Pro consistently outperform open-weight and lightweight models, especially on complex reasoning tasks. Although some recently introduced models demonstrate strong performance within agentic frameworks with tool assistance, they often fall short of delivering the level of detailed reasoning and depth of legal analysis that tax professionals expect. We also evaluated LLM-as-a-Judge by comparing GPT-4o’s scoring against human experts in tax laws and an alternative LLM judge (Gemini Flash 2.5) (code repo available at Github repository). While GPT 4o exhibits mild score inflation for top models, all three evaluators converge on relative model rankings with very high correlation, confirming the comparative reliability of LLM-based assessment frameworks.

This evaluation exercise thus offers a comprehensive, explainable, and reproducible

evaluation tool for legal AI in India. The evaluation prompts and scoring rubric used in LE-BTL framework closely mirror the approach a human expert would take in assessing performance. This alignment is further supported by the directional consistency observed between the model rankings produced by LLM judges GPT-4o and Gemini Flash 2.5 and those assigned by human evaluators. It provides a reliable foundation for researchers, policymakers, and legal-tech developers seeking to deploy LLMs responsibly, enhancing, rather than replacing, expert judgment in the Indian tax ecosystem.

Contents

1. Introduction

Tax services, as part of the legal industry, often characterised by their reliance on meticulous research and reasoning, are undergoing a transformation driven by the advent of LLMs [1] [2].

The ability of LLMs to process and analyze vast amounts of textual data has opened new avenues for their application in tax research, contract analysis, document summarization, and compliance monitoring. By automating time-intensive legal tasks, LLMs can enhance efficiency, allowing professionals to focus on high-value activities such as strategic decision-making, client advocacy, and nuanced legal reasoning. This transformation has the potential to democratize access to tax services, making them more affordable for individuals and small businesses that previously faced financial barriers to professional legal assistance [3].

However, like the legal services domain, the adoption of LLMs in the tax services domain necessitates a cautious and measured approach due to the high-stakes nature of the decision-making involved in this industry [4]. Unlike general text-processing applications, tax analysis requires strict adherence to statutory frameworks, judicial precedents, and jurisdiction-specific interpretations. The opacity of LLM reasoning, the limited visibility of training data sources, and the inherent complexity of the language make it imperative to establish rigorous objective evaluation mechanisms for assessing the reliability and accuracy of LLM-generated legal outputs.

In response to the challenges faced in both legal and tax services industries, the legal experts worldwide have identified legal benchmarks as a crucial tool for systematically evaluating the capabilities and limitations of LLMs in legal applications. LLM benchmarks are standardized tools used to evaluate how well language models perform in various facets such as language comprehension, mathematics, coding, etc. Some of the prominent LLM benchmarks are HellaSwag[5], CodeContests[6], CodeEval[7], MMLU[8], etc. They include a range of tasks and questions, both quantitative and qualitative metrics, to test how effectively an LLM understands language, reasons through complex questions, and produces relevant and coherent responses. For legal services and tax services, LLM benchmarks would provide standardized methodologies for testing LLM performance on legal tasks, ensuring that their outputs meet the necessary standards of precision, consistency, and interpretability[9].

Indian tax laws' intricate regulatory framework, coupled with frequent amendments and contradictory court rulings, presents unique challenges that require localised AI solutions. In this paper, we present our work in developing a customized LLM benchmark for Indian tax laws. Our study explores the performance of LLMs across various tasks undertaken by tax practitioners. By establishing a structured benchmarking framework, this work aims to contribute to the responsible deployment of AI in the tax services industry in India, ensuring that LLMs serve as reliable tools that complement, rather than replace, human expertise.

1.1 LLMs in the legal sector

LLMs like GPT-4, GPT-4o, Gemini 2.5 Pro, Gemini Nano, Claude 3 Opus, Grok3, and DeepSeekV3 are transforming legal practice by enabling rapid analysis, summarization, and interpretation of complex legal data. They streamline routine tasks, enhance accuracy, and increase accessibility to legal services[10].

LLMs are being integrated across legal workflows, including research, drafting, compliance, advisory, and education. Key applications include:

(a) Streamlined Legal Research

- *Automated Research and Interpretation*: LLMs can quickly extract relevant precedents and statutory provisions from statutes, such as the Income Tax Act, 1961, and the Central Goods and Services Tax Act, 2017, and assist in interpretations, thereby reducing research time while ensuring completeness.
- *Summarization*: LLMs can summarize lengthy legislative texts, regulatory changes, and government circulars and notifications into concise, accessible and actionable summaries for practitioners and taxpayers.

(b) Efficient Contract Review, Drafting, and Analysis

- *Review and Risk Analysis*: LLMs analyze complex agreements (e.g., M&A, commercial contracts), identify key clauses, flag risks, and suggest improvements.
- *Automated Drafting*: Based on input parameters, they generate legally compliant drafts, saving time and improving accuracy.

(c) Legal Document Standardisation

- *Document Generation*: LLMs assist in drafting contracts, submissions, opinions, and checklists.
- *Standardisation*: Ensures consistency across documents, aligning with legal standards and minimising errors.

(d) Legal Advisory and Client Interaction

- *Preliminary Consultation*: AI chatbots provide general guidance on tax, contract law, *and* regulations, filtering routine queries before escalation to professionals.
- *Legal Help*: Simplifies legal content for individuals and SMEs, promoting legal and tax literacy.

(e) Due Diligence and Compliance

- *Automated Due Diligence*: LLMs analyze financial and legal documents in M&A or investment scenarios, flagging risks and liabilities.
- *Compliance Monitoring*: Track and summarize legal changes, helping organizations remain compliant with evolving laws.

(f) Predictive Legal Analysis

- *Outcome Forecasting*: By analysing past judgments, LLMs estimate litigation success *probabilities*, guiding decision-making.

- *Judicial Trend Analysis*: Identifies patterns in judge/court decisions to shape more strategic legal arguments.

(g) Legal Education and Training

- *Custom Learning Tools*: LLMs support legal education with tailored modules, interactive case studies, and auto-generated quizzes.
- *Simulation Exercises*: Real-time drafting and legal scenario simulations help students gain *practical* skills.

The integration of LLMs in legal workflows has the potential to transform legal practice across the value chain. By enhancing efficiency, reducing costs, improving accessibility, and enabling innovation, LLMs are transforming the way legal professionals interact with legal information and perform their daily tasks [10].By enhancing efficiency, reducing costs, improving accessibility, and enabling innovation, LLMs are transforming the way legal professionals interact with legal information and perform their daily tasks[10].

Below are key areas where LLMs drive efficiency in legal work:

(a) *Cost Efficiency:* LLMs automate routine tasks like legal research, drafting, review, summarization, and compliance checks, reducing billable hours, lowering operational costs, and increasing the affordability of quality legal services.

(b) *Improved Accessibility*: LLMs democratise legal knowledge through:

- *Self-Service Tools*
- *Simplified Legal Explanations*
- *Multilingual Support*

(c) *Enhanced Accuracy and Consistency*: LLMs (with human-in-a-loop) help minimise human error, ensure compliance with legal standards, and standardise documentation workflows. Though expert review remains essential, overall precision improves significantly.

(d) *Innovation in Legal Services*: LLMs enable new models like:

- *AI-driven legal platforms*
- *Predictive analytics*
- *Smart contract* lifecycle tools

These innovations boost scalability and service quality.

The application of LLMs to tax and legal workflows is progressing rapidly, but the current state of the art shows clear limitations that prevent fully autonomous use. Across multiple independent evaluations in legal reasoning and tax compliance, consistent patterns emerge:

- LLMs can meaningfully improve productivity but are prone to domain-critical mistakes if left unsupervised.
- Many models still fail in logical reasoning, multi-step application, and numerical computation.

- Generic benchmarks are insufficient as evaluation must match the specific tasks, risks, and structures of the intended use case.
- Modular, tool-augmented architectures outperform “vanilla” LLMs for high-stakes workflows.

These points are well-supported by recent research in tax reasoning, legal benchmarking, and rule-guided task evaluation (e.g., Nay et al., 2024; LegalBench, 2023; RuleArena, 2025; TaxCalcBench, 2025).

Thus, it can be said that LLMs are not replacing legal professionals but are enhancing efficiency in legal workflows by automating routine tasks, increasing accessibility, improving accuracy, and enabling innovation. However, their responsible adoption requires careful oversight, validation, and compliance with ethical and legal standards. As AI-driven legal tools continue to evolve, structured benchmarking and evaluation frameworks will be crucial in determining their reliability and effectiveness. The limitation of existing benchmarks in the context of evaluation of various LLMs on tax reasoning, specifically with respect to the Indian Tax Laws, has been discussed in detail in Table 3 in point no 2.3.

Deploying LLMs requires addressing the ethical and regulatory safeguards for the following issues:

- Inaccurate information and outdated law: LLMs may generate responses based on superseded or incorrect legal provisions, risking non-compliant or misleading outputs.
- Hallucinations and misinformation: LLMs can fabricate facts or cite non-existent cases, requiring validation to ensure factual and legal accuracy.
- Bias and Fairness: LLMs may reflect or amplify societal and dataset biases, leading to discriminatory or unbalanced legal interpretations.
- Data Privacy: Handling sensitive legal and personal data with LLMs necessitates strict compliance with confidentiality and data protection laws.
- Emerging regulations on AI-generated legal advice: As jurisdictions develop rules around AI in legal services, compliance with regulatory standards and professional liability norms becomes essential.

A balanced approach with human oversight and ethical compliance is critical to responsible adoption.

1.2 Application of LLMs for the Indian Tax Laws

India, with its burgeoning caseload, as detailed in Table 1, and complex legal landscape, presents a unique opportunity for LLMs to transform legal and tax workflows, from judicial assistance to compliance monitoring [11]

Table 1: (Details of the caseload in the Indian judiciary)

Particulars	Details
No. of appeals pending at CIT(A) level as on 01 April 2025	5,38,000[12]

Particulars	Details
No. of appeals pending at the Income Tax Appellate Tribunal (ITAT), High Court and SC at the end of FY 2023-24	64,311[13]
Typical time period required to reach a final resolution on tax matters through the appellate system	15 to 20 years[13]

Recent advances show growing momentum in AI adoption within India's legal system, such as AI-assisted research platforms, live court transcriptions, and the use of ChatGPT in judicial reasoning by courts. These developments highlight the utility of LLMs in making legal processes more accessible and efficient.

However, Indian tax law poses distinct challenges that complicate the direct application of LLMs:

- Statutory and procedural complexity (frequent amendments, overlapping regulations).
- Judicial inconsistency and layered legal interpretation (statutes, rules, circulars, precedents).
- Co-existence of multiple tax regimes and jurisdictional variances.
- Risk of outdated or biased LLM outputs without domain awareness.

While promising, these factors necessitate a human-in-the-loop deployment model and a domain-specific evaluation framework to assess whether LLMs can reason accurately within India's unique legal and tax context. Section 2.4 elaborates on the structural limitations of current benchmarks and the need for the LE-BTL framework.

2. Literature Review

LLM benchmarks are standardized activities to evaluate how well language models perform. They include a range of tasks and questions, both quantitative and qualitative metrics, to test how effectively an LLM understands language, reasons through complex questions, and produces relevant and coherent responses. Benchmarking helps researchers, developers, and users assess model strengths and weaknesses, compare models, and guide future improvements [14].

2.1 Need for LLM benchmarks

As LLMs rapidly expand across domains like law, finance, and healthcare, structured benchmarks are crucial to evaluate their performance, guide development, and ensure responsible use. Benchmarks offer standardized, objective frameworks to test model accuracy, fairness, and real-world reliability[14].

Key reasons LLM benchmarks are critical:

(a) Evaluation and Comparison:

- *Objective Metrics*: Benchmarks quantify LLM performance in tasks like legal reasoning, summarization, and contract analysis.
- *Model Comparison*: Consistent criteria allow for comparing models like GPT, Gemini, and Claude.
- *Domain Relevance*: General LLMs often fall short in fields like tax law. Specialised benchmarks (e.g., LE-BTL) ensure domain-specific reliability.

(b) Progress Tracking:

- *Monitoring Improvements*: Benchmarks track performance across model generations.
- *Identifying Gaps*: They highlight issues like hallucinations, bias, and weak reasoning.

(c) Guide Development:

- *Focused Research*: Results identify areas needing better legal reasoning or data quality.
- *Training Optimisation*: Insights help refine data, model architecture, and prompts.
- *Innovation Driver*: Promotes new tools like AI legal summarizers or compliance *checkers*.

(d) Collaboration and Transparency:

- *Standardized Ground*: Enables researchers and regulators to compare and audit results.
- *Open Research*: Facilitates shared learning and safe AI development.

(e) Responsible AI:

- *Bias Detection*: Benchmarks uncover ethical and fairness issues in outputs.
- *Regulatory Alignment*: Helps meet AI compliance norms (e.g., EU AI Act, Indian guidelines).

Benchmarks are not just scoring tools. They are foundational to the development and deployment of trustworthy LLMs. In high-stakes domains like legal and tax advisory, where inaccuracies can lead to regulatory violations, financial loss, or miscarriage of justice, benchmarks play a vital role. They provide a structured and objective mechanism to assess whether an LLM can reason with legal precision, interpret complex statutes, and apply case law appropriately. Moreover, by identifying gaps such as hallucinations, bias, or ethical oversights, benchmarks help developers refine model behavior and align outputs with professional standards. In doing so, benchmarks ensure that LLMs adhere to the principles of fairness, explainability, and reliability, which are crucial for the adoption of responsible AI in legal and tax advisory contexts.

2.2 Prominent LLM benchmarks

Several leading benchmarks evaluate LLMs across tasks like language understanding, reasoning, and contextual comprehension. These evaluations offer structured insights into model strengths and limitations.

Around 2018, with the release of GLUE (General Language Understanding Evaluation, a pivotal moment that established standardized, multi-task testing for NLP models. As models evolved in scale and complexity, more challenging benchmarks like SuperGLUE and SQuAD were introduced, pushing models beyond syntactic understanding toward semantic reasoning and contextual comprehension.

By end of 2025, the field will see an explosion in the number of benchmarks, with over 150 active public and private benchmarks assessing everything from general-purpose reasoning to highly specialised domains like mathematics, programming, law, medicine, and taxation.

Benchmarks today are often categorised into:

- General-purpose benchmarks – evaluating core NLP abilities (e.g., GLUE, SuperGLUE, MMLU).
- Domain-specific benchmarks – targeting specialised reasoning tasks in a domain (e.g., LegalBench, MedQA, MathBench).

This section highlights some of the most influential general-purpose benchmarks, which provide a foundation for evaluating LLM performance but often fall short in niche legal or tax applications.

Table 2 compares leading LLM evaluation benchmarks (e.g., GLUE, SuperGLUE, SQuAD) based on their purpose, task types, examples, and relevance. It helps illustrate the scope and limitations of general-purpose benchmarks for domain-specific applications like legal or tax reasoning.

Table 2: Summary of Prominent General LLM Benchmarks

Benchmark	Purpose	Tasks Included	Example	Relevance
GLUE (General Language Understanding Evaluation) [Multi-task]	Tests general natural language understandin g (NLU) across multiple linguistic tasks.	Sentiment analysis; Question answering; Textual entailment (determining if a statement logically follows from a previous sentence).	Evaluate whether "This movie was fantastic!" is positive sentiment.	Provides a comprehensive multi-task evaluation, useful for assessing how well LLMs understand and process natural language.
SuperGLUE [Multi-task]	Evaluates advanced reasoning and comprehensi on in LLMs (Challenging version of GLUE).	Complex question answering; Commonsense reasoning; Reading comprehension challenges.	Answer: "If a person buys a life insurance policy but forgets to disclose a pre-existing condition, and then files a claim after diagnosis, should the insurance company honour the policy?"	Pushes LLMs beyond basic text processing, requiring stronger logical and inferential capabilities.
MMLU (Measuring Massive Multitask Language Understanding) [Multi-task]	Measures LLM subject knowledge.	Answer multiple-choice questions on 57 subjects including law, math, and science.	"What is the minimum number of members required to form a cooperative society under Indian law?"	Covers broad subjects including legal/tax but lacks deep contextual nuance.
Big-Bench-Hard [Multi-task]	Evaluates beyond-surface abilities.	Answer complex riddles or uncommon reasoning tasks like counterfactuals, symbolic	"If Alice had gone to court instead of settling, how might the outcome differ?"	Probes limits of LLM generalization and reasoning.

Benchmark	Purpose	Tasks Included	Example	Relevance
		reasoning, and logic puzzles.		
GSM8K [Single-task]	Test math word problem solving.	Grade-school math.	“Priya earns Rs. 500 per week. If she saves Rs. 150 each week, how many weeks will it take to save Rs. 3000?”	Reflects reasoning under constraints, helpful in tax calculations.
SQuAD (Stanford Question Answering Dataset) [Single-task]	Measures reading comprehension.	Locate the correct answer within a given passage.	From a legal passage - “What is the statute of limitations for contract claims in India?”	Essential for evaluating fact retrieval and contextual understanding, crucial for legal and research-based applications.
LAMBADA [Single-task]	Tests long-context understanding.	Assessing the ability to track long-range dependencies in text.	Complete: “The appellant argued extensively throughout the…”	Measures deep contextual comprehension, important for legal document analysis and contract interpretation.
Winograd Schema Challenge [Single-task]	A test for common-sense reasoning and pronoun resolution.	Disambiguating pronouns in complex sentence structures.	“The contract clause was vague because it was overly broad. What does ‘it’ refer to?”	Evaluates contextual and logical reasoning, critical for contract review and case law interpretation.
HellaSwag [Single-task]	Tests common-sense reasoning.	Predicting plausible story conclusions using logical coherence.	Given a courtroom narrative, select the plausible next action.	Assesses narrative comprehension and inference-making, crucial for applications like legal argument construction and judgment summarization.

As shown in Table 2, general LLM benchmarks address multiple natural language tasks such as sentiment analysis and question answering, but they fall short in testing legal reasoning or domain-specific comprehension.

Each of these benchmarks provides valuable insights into different dimensions of LLM capabilities. However, they primarily focus on general language understanding rather than domain-specific reasoning. In highly specialised fields like law and taxation, existing benchmarks may not sufficiently evaluate statutory interpretation, case law reasoning, or compliance-based decision-making. This underscores the need for domain-specific benchmarks.

2.3 Equivalent Global Tax & Legal AI benchmarks

The benchmarking of LLMs in tax and legal domains is still in its early stages compared to general AI benchmarks such as GLUE, SuperGLUE, and SQuAD. However, several efforts have been made globally to evaluate AI-assisted tax and legal reasoning. A literature review of existing benchmarks provides insights into their strengths, limitations, and applicability to tax law tasks.

Table 3 summarizes benchmarks that evaluate legal and tax reasoning in AI models. It outlines each benchmark's focus area, strengths, and limitations, highlighting the lack of support for Indian tax law evaluation and the need for a specialised benchmark tailored to Indian tax law.

Table 3: Comparative Overview of Global Tax and Legal AI Benchmarks

Benchmark	Description
Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence [26-Feb-2024] [https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0159#d1e491]	Evaluates LLMs on real-world U.S. federal tax questions, including deductions, credits, and filing logic. It uses IRS-style scenarios and case facts to assess reasoning accuracy and rule application. Measures interpretive capability, evolution across model generations. **Strengths:** Large-scale dataset; strong legal interpretive evaluation; tracks model progression (e.g., GPT-3.5 vs GPT-4). **Limitations:** U.S.-centric; dataset not publicly reusable or open-source.
LegalBench [22-Aug-2023] [https://arxiv.org/abs/2308.11462]	Legal reasoning benchmark across multiple jurisdictions and domains (contracts, torts, tax). Provides 160+ tasks using real case facts and legal principles.

Benchmark	Description
	Measures LLM accuracy, consistency, and explainability in jurisdictional legal reasoning. **Strengths:** Task diversity; real court language; multi-jurisdictional scope. **Limitations:** Limited coverage of Indian tax law; few statute-based reasoning tasks.
PLAT (Predicting the Legitimacy of Punitive Additional Tax) [May 2025] [https://arxiv.org/pdf/2503.03444]	Benchmark focused on tax legitimacy and punitive assessments. Covers multilingual Korean tax disputes; tests compliance logic, fairness, and judgment prediction. Evaluates classification of tax liability legitimacy and legal judgment soundness. **Strengths:** Realistic tax data; multilingual; includes ethical and fairness dimensions. **Limitations:** Korea-specific legal logic; not generalizable to common law (e.g., India).
TaxEval (v2) Benchmark [May 2025] [https://www.vals.ai/benchmarks/tax_eval_v2-05-05-2025]	Structured benchmark for tax compliance and statutory understanding. Contains updated tax problems for corporate tax, income tax, VAT. Measures both numerical accuracy and interpretive compliance with tax statutes. **Strengths:** Updated regularly; practical compliance scenarios; includes multiple tax types. **Limitations:** Private dataset; restricted access; not tailored to Indian laws.
Big Law Benchmark (Harvey) [29-Aug-2024] [https://www.harvey.ai/blog/introducing-biglaw-bench; GitHub - harveyai/biglaw-bench]	Evaluates LLMs on enterprise legal tasks like M&A, contracts, securities, and tax. Tests chain-of-thought reasoning, clause identification, and zero-shot accuracy. Measures enterprise-grade legal utility and reasoning under pressure. **Strengths:** Practical workflows; complex legal scenarios; clause-level granularity. **Limitations:** U.S.-centric; lacks coverage of statutory tax law for public jurisdictions.

Benchmark	Description
LexGLUE (Legal General Language Understanding Evaluation) [8-Nov-2022] [https://arxiv.org/abs/2110.00976]	General legal language benchmark for classification and judgment prediction. Includes tasks such as legal NER, statute identification, and contract labelling. Tests legal NLU, document understanding, and precedent alignment. **Strengths:** Multi-task framework; strong legal language evaluation; judicial document parsing. **Limitations:** No tax-specific tasks; civil law bias; limited for India or statutory reasoning.
CaseHOLD (Case Law Holdings Dataset) [2021] [https://arxiv.org/abs/2104.08671]	Focuses on predicting the "holding" (legal outcome) in appellate cases. Tests LLMs on completion of partial legal arguments using correct judicial logic. Measures ability to align with precedent and argument coherence. **Strengths:** Focus on factual nuance; tests core judgment logic. **Limitations:** U.S. law only; lacks statute interpretation; no multilingual or tax domain support.
COLIEE **(Competition on Legal Information Extraction and Entailment)** [15-Jan-2025] [https://coliee.org/resources]	Legal research and information retrieval benchmark. Tasks include case retrieval, entailment recognition, and statute matching. Simulates research scenarios to evaluate precision and logic in document sourcing. **Strengths:** Legal QA structure; strong on legal entailment and retrieval. **Limitations:** No statutory compliance tasks; not designed for tax law; civil law skew.
IL-TUR (Indian Legal Text Understanding and Reasoning) **[07-Jul-2024]** **[https://arxiv.org/abs/2407.05399]**	Comprehensive benchmark focused on evaluating LLMs on Indian legal texts, including judgments, statutory provisions, and legal commentary. Includes MCQs and generative tasks designed to assess comprehension, reasoning, and statute-based interpretation in Indian law. Measures legal understanding, statutory reasoning, and jurisdiction-specific language modelling in the Indian context.

Benchmark	Description
	Strengths: Focused on Indian legal ecosystem; covers statutory and case law; includes multilingual legal texts and judgment summaries. **Limitations**: Does not include tax-specific modules; limited chain-of-thought evaluation for complex legal arguments.

2.4 Need for benchmarks in the context of the Indian Tax Laws

The rapid advancement of LLMs is creating transformative opportunities in India’s tax services landscape. Their ability to process vast volumes of legal data, reason through complex legal texts, and generate structured summaries makes them a promising tool for enhancing tax research, documentation, and decision support. As India’s judicial and regulatory systems grapple with backlogs of pending cases and accessibility challenges, the careful integration of LLMs has the potential to streamline tax workflows, assist in judicial decision-making, and expand access to justice for the masses. However, the deployment of LLMs must follow a human-in-the-loop model, where legal experts validate outputs, interpret nuanced tax provisions, and ensure contextual accuracy, thereby maintaining the integrity and reliability of the legal process.

The Indian courts have already taken proactive steps toward AI adoption in legal and tax tasks:

(a) Judicial Assistance:

- The Punjab and Haryana High Court recently used ChatGPT to assist in analysing legal precedents while deciding a bail application, the first instance of an Indian court leveraging AI in judicial decision-making [15]. In another case, the Punjab & Haryana High Court took the assistance of Chat GPT to understand how 'Differential GPS' helps in locating disputed property [16].

(b) Live Court Transcriptions:

- The Supreme Court of India has introduced machine-assisted transcription tools, trained using AI, to improve accessibility and efficiency in court proceedings[17].

(c) AI-Assisted Research Portals:

- The Chief Justice of India (CJI) launched an AI-driven legal research platform, aiming to help judges and legal practitioners conduct faster and more precise case law research [18].

Further, Chairman of Central Board of Direct Taxes said in an interview that the Income Tax Department is now employing artificial intelligence to monitor taxpayer behavior by tracking online portal visits, flagging high-value transactions, and filtering out Permanent Account Numbers (PANs) linked to suspicious claims and emphasised that the objective is to gradually build a comprehensive, 360-degree view of taxpayers' financial activities and detect patterns and inconsistencies in tax return filings.

Despite these advancements, the application of LLMs in Indian tax laws demands a cautious approach, particularly due to:

(a) Lack of evaluation of the tax reasoning of LLMs in the case of Indian Tax Laws

- Existing global benchmarks do not evaluate LLMs on Indian legal or tax reasoning, compliance, or statutory interpretation[19].

(b) Legal Complexity specific to Indian Tax Laws:

- Indian tax laws are vast, dynamic, and precedent-driven, demanding nuanced, context-specific interpretation that LLMs may struggle to deliver[20] / [21]. Indian tax laws undergo continuous amendments, which increases the risk that LLMs may generate responses based on the outdated provisions of the laws.

(c) Multiplicity of Interpretations:

- It is common for different benches of the Income Tax Appellate Tribunal (ITAT) and different High Courts to take conflicting views on the same provision. The parties on both sides (i.e., tax authorities and taxpayers) often interpret laws by going beyond a literal reading. For example, what constitutes a ‘charitable purpose’ for the registration of a charitable trust has had divergent interpretations by various courts.

(d) Layered structure of statutes, circulars and precedents:

- Legal interpretation in India is not just about the statute. It involves statutory provisions (e.g., Income Tax Act, GST Acts), Rules and Notifications, CBDT/CBIC Circulars (which are binding on the department but not the assessee and the courts), and Tribunal & Court decisions. This layered structure creates ambiguity and litigation opportunities.

(e) Jurisdictional and Judicial Inconsistencies:

- High Court decisions are not binding nationally, while they bind only within their jurisdiction. In the absence of a Supreme Court ruling, divergent interpretations co-exist across states. The same income may be taxable in Maharashtra but exempt in Karnataka, depending on the judgments of the respective jurisdictional courts. Judgments by Indian courts are frequently overruled, reversed, or distinguished based on facts. Even after a Supreme Court decision, lower authorities may distinguish it factually and deny relief.

(f) Co-existence of Multiple Tax Regimes:

- India's tax system includes Direct taxes (with various heads of Income like PGBP, Capital Gains, etc.) and Indirect taxes (GST, Customs, etc.), with historical remnants (e.g., Excise, Service Tax litigation still ongoing). Further, the GST Law itself is dual-layered (CGST + SGST), with state-wise rulings and a lack of centralised appellate resolution (GSTAT is announced recently).

(g) Risks of Errors and Bias:

- LLMs can produce inaccurate legal or tax advice or reflect biases present in training data, raising ethical and practical concerns.[22]

The 'LLM Evaluations for Bharat Tax Laws' framework aims to bridge this gap for Indian tax laws by offering a structured evaluation framework.

Despite global progress in LLM evaluation, no benchmark currently addresses the specific needs of Indian tax law. Existing legal AI benchmarks are either too general (LexGLUE, COLIEE) or too jurisdiction-specific (CaseHOLD). As shown in Table 3, even jurisdiction-aware benchmarks like LegalBench or Big Law remain inadequate for Indian tax contexts, given their limited coverage of tax-specific statutes or advisory workflows. None of the current benchmarks comprehensively evaluates AI models for Indian tax law, statutory compliance, or litigation advisories.

Given the complexity, frequent amendments, and case-law-driven interpretation in India, relying on foreign-trained LLMs risks inaccurate, non-compliant, and unreliable outputs in Indian legal contexts. The following are the challenges in applying LLMs to Indian Tax Laws:

(a) Legal and Structural Complexity

- Frequent amendments to the Income Tax Act and GST laws demand constant model updates.
- Case law-centric interpretation: Judicial precedents (Supreme Court, High Court, Tribunals) significantly shape tax laws.
- Multiple tax regimes: India has a dual tax structure, comprising Income Tax and Indirect Tax (GST, Customs, Excise), which involves complex computation and compliance rules.
- Global benchmarks often overlook Indian judicial reasoning, misinterpret the provisions and fail to track the amendments.

(b) Lack of Indian Legal Evaluation Framework

- Indian legal interpretation requires harmonising statutes and precedents, a challenge for most LLMs.
- No benchmark or dataset exists to evaluate LLMs on India-specific legal tasks, like interpreting a provision with a holistic approach after considering the principal legislation, subordinate legislation framed thereunder and the administrative instruments. For example, to create a holistic view on an income tax issue, a professional will look at the Income Tax Act, 1961 (principal legislation), Income

Tax Rules, 1962 (subordinate legislation), notifications and circulars (administrative instruments) issued for the proper implementation of the Act.

Given the challenges in applying the LLMs to Indian Tax Laws, an LLM evaluation framework is essential. However, due to the limitations of the global tax and legal AI benchmarks, there is an urgent need for an India-specific LLM benchmark tailored to the country's unique taxation laws, judicial framework, and compliance complexities.

A domain-specific benchmarking framework like LE-BTL is essential to:

- Evaluate LLMs on Indian tax law, including statutes and case law.
- Ensure accurate interpretation of the provisions based on amended legislation.
- Offer a standardized Indian tax law dataset for training and validation.
- Test alignment with Indian judicial reasoning principles.

Without such a benchmark, AI tools may offer flawed advice, posing regulatory and compliance risks in tax practice.

2.5 Example of the diverse answers from different LLMs for the same question

To demonstrate the diversity of the responses from different LLMs, we have provided below the responses from sample LLMs for a question which had been categorised as 'Simple' for the purpose of benchmarking.

Question: Can a resident individual remit US$ 2,50,000 to his NRI brother as an interest-free loan under LRS?

GPT o3 Pro (No Prompt):

"........Therefore, while the resident could:
• Gift up to USD 250,000 to his NRI brother under LRS, or
• Give him an interest-free rupee loan credited to the brother's NRO account in India,

he *cannot simply remit USD 250,000 abroad as an interest-free loan under LRS without first obtaining specific permission from the Reserve Bank of India."*

GPT o3 (No Prompt):
"Yes.
Regulation 4(2) of the Foreign Exchange Management (Borrowing or Lending in Foreign Exchange) Regulations, 2018 read with the RBI Master Direction on the Liberalised Remittance Scheme (LRS) gives a general permission to a resident individual to lend, out of the USD 2,50,000 yearly LRS limit, foreign exchange to his / her NRI / OCI "close relative" (as defined in the Companies Act, 2013)........."

Gemini 2.5 Pro (No Prompt):
"Yes, a resident individual can remit USD 2,50,000 to his NRI (Non-Resident Indian) brother as an interest-free loan under the Liberalised Remittance Scheme (LRS), provided certain conditions are met...................."

Gemini Flash 2.5 (No Prompt):

*" No, a resident individual **cannot** remit USD 2,50,000 to his NRI brother as an **interest-free loan** under the Liberalised Remittance Scheme (LRS)…………………."*

GPT 5 (No Prompt)

"No. Under the Liberalised Remittance Scheme (LRS), a resident individual cannot remit foreign exchange as a loan to a non-resident (including an NRI brother). LRS permits gifts in ………………"

The above responses demonstrate substantial divergence in both conclusions and reasoning chains:

- Contradictory outcomes: While GPT o3 and Gemini 2.5 Pro affirm the permissibility of such a loan, GPT o3 Pro and Gemini Flash 2.5 reject it outright.
- Different reasoning approaches and assumptions and variability in legal grounding: Some models (e.g., GPT o3 Pro) emphasise the distinction between permissible gifting, rupee loans, and foreign exchange remittances, whereas others (e.g., Gemini 2.5 Pro) assume a broader reading of LRS permissions without clarifying the underlying regulatory nuances.

This example underscores the inherent variability in LLM reasoning when handling legal queries, even those classified as 'simple'. It highlights the importance of benchmarking frameworks like LE-BTL, which can identify such inconsistencies and provide structured evaluation across models and prompting strategies.

3. Method

This section outlines the methodology for developing the LE-BTL framework, including the legal reasoning framework, task design, and evaluation structure.

Our objective is to build a structured evaluation framework for testing LLMs in the domain of Indian tax law. Given the legal complexity and frequent amendments in Indian tax statutes, a generic benchmark is insufficient. A robust legal reasoning framework is essential to ensure statutory accuracy, judicial reasoning, and ethical integrity.

To achieve this, we employ an IRAC+ framework, which builds upon the widely used IRAC (Issue-Rule-Application-Conclusion) legal reasoning framework[23/24].

3.1 Overview of the IRAC+ framework

The IRAC method originated as a pedagogical tool in American legal education during the early 20th century, particularly popularized in legal writing and bar exam preparation. Its structured approach to legal problem-solving viz., identifying legal issues, citing relevant rules, applying those rules to facts, and drawing conclusions, has since become a global standard in both academic and professional legal contexts [25].

However, while IRAC ensures clarity and logic, it does not fully capture the nuanced reasoning required in complex tax scenarios.

The IRAC+ framework builds upon the traditional legal reasoning model (Issue Identification, Rule Identification, Application, Conclusion) and adds two critical dimensions:

- Interpretation (e.g., reasoning of statutes, applicability of amendments and precedents for peculiar factual scenarios, ethical considerations, etc), and
- Argumentation (e.g., identification of key issues and drafting arguments against or in support of the issue, etc).

This refined framework allows for a systematic evaluation of LLM performance in tax law applications, ensuring that AI-generated legal responses meet judicial standards, statutory accuracy, and professional ethics.

This framework enables a task-based, question-driven evaluation of how effectively an LLM can identify legal issues, apply Indian tax law, interpret regulatory documents, and generate reasoned, compliant outputs. It ensures that model performance aligns with the professional standards expected in Indian tax advisory and litigation.

Table 4 maps each component of the proposed IRAC+ legal reasoning framework to specific tasks and evaluation criteria relevant to Indian tax law. It serves as the core task design for the LE-BTL framework.

Table 4: IRAC+ Framework Components and Evaluation Tasks for Indian Tax Law

Component	Tasks	Evaluation Remark
1. Issue Identification (Identifying the core legal question from the given facts)		Evaluates the LLM's ability to identify all the relevant tax issues based on the provided facts and context. Example: Where a user asks a query about the treatment of interest on borrowings under the Income Tax Act, does the LLM identify whether the issue pertains to allowability of interest from business income or income from house property or from dividend income, etc.?
2. Rule Identification (Identifying the relevant statutory law, case law, or tax provision)		Assess whether the LLM correctly identifies relevant statutes, rules, or precedents applicable to the issue. Example: For the issue pertaining to taxability of foreign dividends, does the LLM correctly reference DTAA provisions alongside Section 56 and Section 90 of the Income Tax Act?
3. Application of Law (Applying the identified rule to the given facts)		Test the LLM's ability to apply the identified rule to the given facts of the case, including the evaluation of impact of exceptions, exemptions or conditions. Example: For determining GST applicability, does the LLM correctly identify that supply of healthcare services by a charitable trust is exempt under Notification No. 12/2017-Central Tax (Rate), and accurately conclude that no GST is payable, while also considering conditions like registration under Section 12AA or Section 12AB of the Income Tax Act?
4. Conclusion (Arriving at a final, justified answer)		Evaluate whether the LLM provides a logical, defensible, concise, clear and actionable conclusion based on the reasoning process including detecting if the LLM presents uncertain conclusions as definitive. Example: Does the LLM provide a clear conclusion on the eligibility of deductions under Section 80C for a taxpayer?

Component	Tasks	Evaluation Remark
5. Interpretation / Reasoning of Language (Understanding legislative intent, statutory construction, and ambiguity resolution)	5a. Reasoning of amendments	Evaluate whether the LLM can explain the rationale behind legislative amendments, such as the abolition of Dividend Distribution Tax (DDT).
	5b. Understanding of the context to determine applicable amendments	Can the model recognize when a new tax provision applies?
	5c. Ability to understand circulars and official tax documents	Test whether the model can accurately summarize and apply CBDT circulars, notifications, and clarifications.
	5d. Precedents and case law application	Assess the LLM's ability to identify and apply relevant case law precedents to the given facts.
	5e. Arithmetic-based tax calculations	Evaluate the LLM's ability to compute tax liabilities, deductions, interest and penalties accurately.
	5f. Case law summarization	Test whether the model can generate concise summaries of judicial decisions, including the legal principle established.
	5g. Review of the deliverables to identify errors	Assess whether the LLM can identify inaccuracies in legal outputs, such as misinterpretations or incorrect references / citations.
	5h. Error identification in large-context legal texts	Test the LLM's ability to detect logical inconsistencies, contradictory statements, or missing information in lengthy legal documents.
	5i. Drafting advisory mails	Evaluate the LLM's ability to draft clear, professional, and legally sound advisories, addressing tax queries or compliance obligations.
	5j. Hallucination detection	Test whether the LLM can differentiate real legal provisions from fabricated text?
	5k. Red flag Detection in agreements	Test whether the LLM would be able to analyze an agreement to highlight the red flags in the agreements in relation to potential tax issues.
	5i. Ethical Considerations (Ensuring fairness, compliance, and responsible AI use in legal reasoning)	Evaluate if the LLM adheres to ethical and professional standards in its responses and recognizes biases, ethical concerns and AI limitations. Key Considerations:

Component	Tasks	Evaluation Remark
		Does the model recognize and disclose limitations or uncertainties in its output? Does it avoid providing biased or potentially misleading interpretations? Can it identify ethical dilemmas (e.g., advising on tax avoidance schemes vs. legitimate tax planning)?
6. Argumentation / Justification (Evaluating coherence, consistency, and legal reasoning quality)	6a. Drafting grounds against tax notices	Assess the LLM's ability to construct well-reasoned legal arguments in response to scrutiny or reassessment notices.
	6b. Checklist preparation for tax notices	Evaluate whether the model can generate a comprehensive checklist for responding to tax authority queries or audits.

Each component is assessed through question-based testing, ensuring that AI-generated responses are examined for completeness, accuracy, and legal integrity.

3.2 Construction process and benchmark questions

Constructing a benchmark begins with selecting a topic that covers the domain's critical phenomena (for example, legal interpretation or tax-computation scenarios), followed by an expert-driven question that ensures real-world relevance and clarity. Each item is then validated by subject-matter experts, annotated with correct outputs or scoring rubrics, and organized into coherent task categories. Generally during the evaluations, models are evaluated against these gold standards using metrics that reflect both raw accuracy (e.g., exact match or F1 score) and higher-order qualities such as consistency, explainability, and fairness, providing a comprehensive snapshot of how well an LLM can perform, generalize, and be trusted in practical applications.

The construction process for the LE-BTL framework is designed to ensure that the benchmark accurately evaluates LLM performance across all aspects of Indian tax law reasoning. This involves a systematic approach to identifying, finalising, and compiling tasks and questions that comprehensively test the capabilities of LLMs.

The LE-BTL framework was built through a rigorous, multi-stage methodology to ensure it comprehensively evaluates LLM reasoning across Indian tax laws. First, we identified and finalised core tasks by aligning them with an enhanced IRAC+ framework and by consulting tax professionals to validate real-world relevance. Emphasis was placed on tasks central to Indian practice, including statutory interpretation, application of judicial precedent, compliance advisory services, review of notices, drafting of advisories, and interpretation of provisions.

3.2.1 Task Identification and Finalisation

The tasks included in the benchmark were identified and finalised through a multi-stage process:

- *Mapping to the IRAC+ Framework*: Tasks align with key legal reasoning elements, issue, rule, application, interpretation, and argumentation. Focus areas include statutory interpretation and application of judicial precedent.
- *Industry Relevance:* Developed in consultation with tax professionals to reflect real-world use cases (e.g., drafting responses to tax notices, interpreting circulars).

3.2.2 Topic selection for developing questions

Topics reflect high-impact, practical areas in Indian tax law:

- Direct Taxes: Income Tax Act and Rules (deductions, capital gains, transfer pricing, reassessment), major amendments (e.g., DDT abolition, Sec 194R).

- Indirect Taxes: Goods and Services Tax (GST) provisions (ITC, reverse charge, compliance).

- Judicial Precedents: Landmark and recent rulings (e.g., Vodafone, Azadi Bachao, Engineering Analysis).

- Circulars and Procedures: CBDT/GST notifications, ITR filing, TDS returns, audits.

- FEMA: Interpretation of complex topics like 'Person Resident in India' and complex compliance norms

- Accounting Standards: Analysis of complex topics like revenue recognition in different scenarios, treatment of different financial instruments

- Complex Scenarios: DTAA, cross-border taxation, PE, MAT, surcharge/cess computations.

3.2.3 Question sourcing and compilation

The process of sourcing and compiling questions involves multiple steps to ensure quality, relevance, and comprehensiveness:

(a) *Initial Drafting of Questions*:

- Prepared by tax experts and legal researchers, in consultation with AI professionals based on real scenarios.

(b) Size and Scope Compared to Existing Benchmarks:

- The prominent legal benchmark *'LegalBench' (as referred to in Table 3)* comprises 162 narrowly defined tasks designed to test legal reasoning skills, including rule application, analogical reasoning, and counterfactual analysis. While these tasks are carefully constructed, they often focus on elementary legal logic or synthetic examples derived from U.S. case law and academic exercises, limiting their direct applicability to real-world legal practice, particularly in jurisdictions outside the common law tradition.

- The benchmark *'LexGLUE' (as referred to in Table 3)*, on the other hand, aggregates seven legal NLP datasets, including ECtHR (European Court of Human Rights cases), SCOTUS (U.S. Supreme Court decisions), EUR-LEX (EU legal documents), and LEDGAR (contract clause classification). These datasets collectively span tens of thousands of annotated examples, but are primarily oriented towards classification, retrieval, entailment detection, and summarization tasks. Their structure relies heavily on document-level labels or statistical learning patterns, making them more suitable for generic machine learning tasks rather than deep legal reasoning or applied professional judgment.

- In contrast, the LE-BTL framework currently includes 103 domain-specific, expert-validated questions covering Indian direct and indirect tax laws. Rather than treating law as a static label-matching exercise, this benchmark focuses on applied legal reasoning, tax-specific compliance, and statutory interpretation, capturing the complexity and ambiguity often encountered in actual advisory and litigation settings. Its smaller size is counterbalanced by its depth, domain complexity, and real-world applicability, making it a high-fidelity benchmark for evaluating practical legal intelligence in the Indian tax context.

(c) Question Design Differentiation:

- Unlike existing benchmarks that often employ multiple-choice formats or binary labels, the LE-BTL framework emphasises open-ended, scenario-driven, and task-oriented questions. Each question is crafted to assess multi-step legal reasoning and simulate professional tasks, such as evaluating a scrutiny notice, applying judicial precedent to new fact patterns, or drafting a tax advisory. This design approach ensures a more comprehensive and realistic assessment of LLM performance in tax practice.

(d) Variety of Question Types:

- Case-based scenarios.
- Open-ended advisory tasks
- Procedural queries (e.g., steps for filing Form 3CEB).
- Arithmetic problems (e.g., tax liability computations).

(e) Validation by Subject Matter Experts (SMEs):

- Each question is reviewed for legal accuracy, clarity, and relevance.

(f) Categorization:

- Organized per IRAC+ components for structured evaluation.

3.2.4 Answer dataset for evaluation and scoring

To ensure reliable and objective evaluation of model performance, each question in the LE-BTL framework is paired with a high-quality, expert-drafted reference answer. These reference answers serve as gold standards against which model responses are scored, enabling consistent measurement of legal reasoning, domain accuracy, and professional applicability.

The answer dataset has been developed with the following principles:

(a) Expert Authorship and Review:

- All answers have been drafted by seasoned chartered accountants, tax consultants, and legal researchers with deep experience in Indian tax practice. Each draft undergoes an independent peer-review process to validate legal correctness, interpretive soundness, and real-world relevance.

(b) Depth and Nuance:

- Answers are not limited to surface-level accuracy but are constructed to reflect multi-step reasoning, interpretative choices, regulatory nuances, and professional judgment. Where applicable, they include statutory references, judicial principles, or appropriate caveats (e.g., where legal ambiguity exists or facts may vary).

(c) Continuous Refinement:

- As models evolve and new legal developments occur, we plan to periodically review and update the answer dataset to maintain contemporary relevance. This also ensures that the benchmark remains adaptive and aligned with changes in tax law, policy, or judicial precedent.

By grounding evaluation in real, well-reasoned expert responses, the LE-BTL framework ensures that LLM performance is assessed not merely on syntactic correctness but on the legal depth, accuracy, and professional soundness expected in actual tax practice.

3.2.5 Application of 'LLM-as-a-judge (LLM Judges)'

The *LLM-as-a-Judge* paradigm refers to the structured use of large language models to evaluate other LLM-generated outputs based on set parameters. This role simulates a judicial or reviewer function, where one LLM assesses another's reasoning using reference answers and legal evaluation criteria. The use of LLM-as-a-Judge has become increasingly popular as a scalable alternative to human evaluation.

Evaluations can be automated across thousands of responses, producing consistent formats at a fraction of the cost of expert review. For domains such as tax law, where answers are open-ended and involve multi-step reasoning, such efficiency is highly attractive.

(a) Evaluation Process

- To operationalise *LLM-as-a-Judge*, we have used GPT-o3 and Gemini 2.5 Pro with the questions, answers generated by the selected LLMs, the expert-annotated gold standard answer, and an evaluation prompt containing a rubric for evaluation based on IRAC+ dimensions. The evaluation prompt has been enclosed in **Annexure 1**.
- The output is a structured scorecard with numerical ratings on each dimension on a 1–5 scale.

(b) Benefits of the LLM-as-a-Judge Framework

Table 5: Benefits of the LLM-as-a-Judge Framework

Benefit	Explanation
Scalable Evaluation	Manual legal review is resource-intensive. LLM-as-a-Judge enables evaluation of hundreds of open-ended answers in minutes, significantly increasing throughput.
Consistency and Objectivity	Unlike human reviewers, who may interpret facts differently or bring subjective biases, the same prompt and rubric yield consistent evaluations across tasks.
Granular Scoring	The LLM can score individual components (issue, rule, application, conclusion, interpretation, and justification), allowing detailed diagnostics of where the answering model performs well or poorly.
Adaptability to Prompt Tuning	Evaluation prompts can be modified to reflect new statutory amendments, case law, or rubric updates without re-training any model.
Supports Iterative Model Development	Enables rapid testing of model fine-tuning efforts across the benchmark with uniform feedback for comparative evaluation.
Enables Meta-Evaluation	Judging LLMs can also be compared against each other for alignment with expert rubrics, helping to assess which LLM is best suited for evaluative roles.

(c) Limitations of the LLM-as-a-Judge Framework

Recent work highlights serious vulnerabilities in the LLM-as-a-Judge framework.

- First, **LLM judges are easily manipulated by superficial cues**. For example, inserting tokens such as a colon (":") or boilerplate phrases ("Solution:", "Thought process:") can systematically trigger false positives, leading judges to reward poor or irrelevant answers, with false positive rates as high as 60–90%[26].
- Second, **LLM judges exhibit systematic biases.** Studies show positional bias (answers listed first are more likely to be scored higher), verbosity bias (longer answers receive better scores regardless of correctness), and self-enhancement bias (judges prefer responses generated by their own model family)[27] . These biases undermine fairness and reliability. In tax evaluation, verbosity bias is particularly dangerous: a response filled with redundant explanations might be rated higher than a concise citation to the correct statutory provision.

- Third, **LLM judges struggle with calibration**. They often report high confidence even when wrong, a phenomenon of "overconfident errors" that undermines trust[28]. Such errors are especially concerning in tax law, where confidently incorrect judgments could misinform compliance or advisory decisions.
- Finally, **reliance on human preference alignment as the gold standard is itself problematic**. Human evaluators provide domain knowledge but are inconsistent, costly, and limited in scale. Moreover, in some reasoning-heavy tasks, LLMs may even exceed humans in recall or consistency[29]. This raises the question of whether human agreement alone should be the ultimate benchmark.

Together, these vulnerabilities suggest that LLM-as-Judge, in its raw form, cannot be trusted for high-stakes benchmarking in sensitive domains such as tax law. Some additional concerns are listed below in Table 6.

Table 6: Limitations of LLM as a Judge

Limitation	Explanation
Prompt Sensitivity	LLM evaluations are highly sensitive to prompt structure. Slight rewording may lead to different evaluations, making reproducibility a challenge.
Evaluation Bias of LLMs	Judging LLMs may exhibit biases inherent in their training data (e.g., favouring fluent writing over legal accuracy), which can skew assessment fairness.
Black Box Nature	Despite interpretability prompts, LLM judges still operate as black-box models with opaque internal reasoning, unlike human subject-matter experts.
Overestimation Compared to Human Reviewers	LLM judges tend to be more lenient and may assign higher scores than human experts, especially for responses that appear well-structured but lack legal depth.

3.2.6 Application of a 'Hybrid Approach'

To overcome these weaknesses, we propose a hybrid evaluation framework designed specifically for tax-domain benchmarking. This approach integrates ground truth anchoring, human-crafted rubrics, multi-judge alignment, and confidence awareness.

- **Ground Truth Anchoring:** At the core is a curated dataset of tax-law questions paired with authoritative answers from statutes, rulings, and expert commentary. Evaluations are compared directly against this canonical reference, neutralising hacks and verbosity bias.
- **Human-Crafted Rubrics:** Expert-designed rubrics evaluate dimensions such as accuracy, reasoning depth, citation fidelity, and applicability. This reduces variability and aligns scoring with professional standards.
- **Multi-Judge Alignment:** We combine multiple diverse LLM judges with human evaluators, reconciling their scores and flagging disagreements. This ensemble reduces positional and self-enhancement biases.
- **Confidence Awareness:** Inspired by calibration methods, high-confidence agreements are accepted automatically, while low-confidence or conflicting evaluations are escalated to humans.

This hybrid framework leverages LLM scalability while embedding the rigour of human oversight and ground truth anchoring.

(a) Purpose

The Human-as-a-Judge framework serves as a benchmark anchor for assessing the fidelity and objectivity of LLM-led evaluations. It allows us to:

- Measure the alignment of LLM evaluators with expert legal reasoning.
- Detect systemic biases (e.g., leniency, fluency preference) in LLM judgment patterns.
- Validate relative model rankings in a high-stakes legal reasoning context.

(b) Methodology

- A curated sample of LLM-generated answers across multiple models and prompt conditions was selected.
- Human experts independently evaluated the answers using the same IRAC+ rubric and score scale (1 to 5 per dimension).
- Average scores and model rankings were compared across human judges, GPT-4o, and Gemini Flash to triangulate consistency.

(c) Significance

- By integrating ground truth, rubrics, multi-judge alignment, and confidence awareness, our framework directly mitigates the vulnerabilities identified in the literature: superficial hacks are neutralised, biases are counteracted, overconfidence is managed, and limits of human preference alignment are addressed.

- The result is a balance between scalability and rigour. Human-only judging is too resource-intensive for large-scale benchmarks, while LLM-only judging is too fragile. Our hybrid approach combines the strengths of both, enabling robust, trustworthy, and domain-adapted evaluation. For tax law in particular, where interpretability, correctness, and accountability are non-negotiable, this framework provides a pathway for risk-aware benchmarking aligned with both expert judgment and canonical legal sources.

3.2.7 Continuous expansion of the benchmark

The benchmark is designed to evolve with tax law and LLM capabilities:

- *Amendment Integration*: Regular updates post-Finance Acts, GST changes, CBDT clarifications or notifications, etc.

- *Emerging Topics*: Includes new areas such as crypto taxation, taxation of winnings from online games, or the abolition of angel tax under Section 56(2)(viib).

- *Feedback-Driven Iteration:* Questions refined based on expert feedback and model evaluations.

- *Ongoing Validation:* Regular SME review ensures consistency and up-to-date relevance.

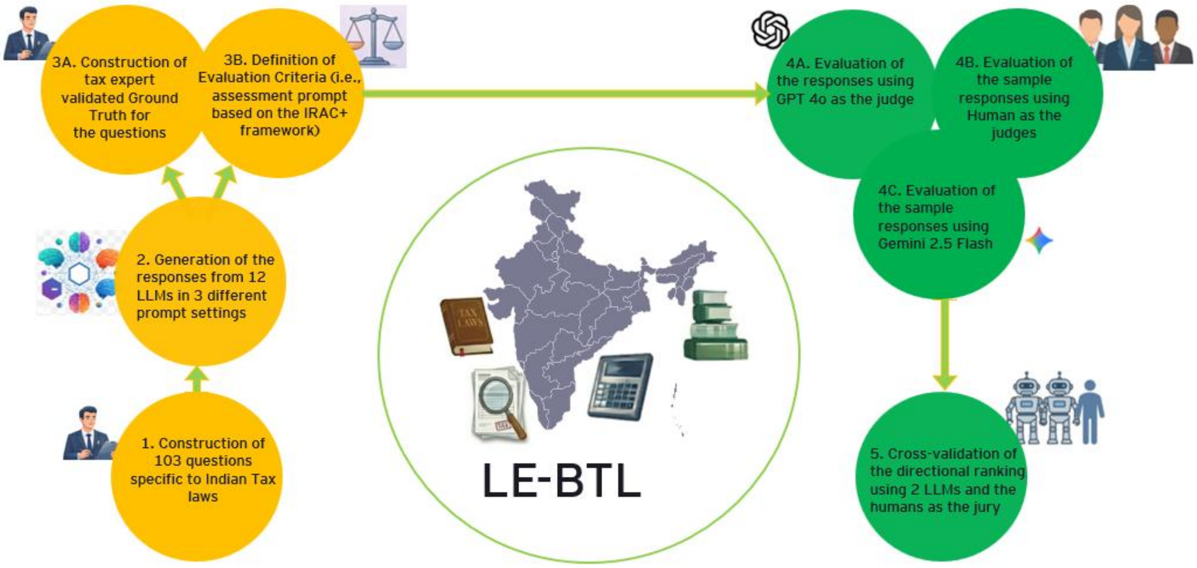
3.2.8 Conclusion

The construction of the LE-BTL framework is a collaborative and iterative process, designed to address the unique challenges of applying LLMs to Indian tax law. By focusing on real-world scenarios, comprehensive topic coverage, and dynamic question expansion, the benchmark ensures:

- A thorough evaluation of LLM performance across all critical tasks in Indian tax laws.
- The ability to adapt to evolving laws and emerging challenges, keeping the benchmark relevant and up-to-date.
- A structured framework for guiding LLM improvements in accuracy, consistency, and interpretability for Indian legal applications.

4. Experiments

4.1 Experimental Setup: Overview of the process followed under the proposed LE-BTL framework


3A. Construction of tax expert validated Ground Truth for the questions
3B. Definition of Evaluation Criteria (i.e., assessment prompt based on the IRAC+ framework)
2. Generation of the responses from 12 LLMs in 3 different prompt settings
1. Construction of 103 questions specific to Indian Tax laws
TAX LAWS
LE-BTL
4A. Evaluation of the responses using GPT 4o as the judge
4B. Evaluation of the sample responses using Human as the judges
4C. Evaluation of the sample responses using Gemini 2.5 Flash
5. Cross-validation of the directional ranking using 2 LLMs and the humans as the jury

4.2 Base Setup

We tested 103 tax-law-specific questions across the following 12 LLMs.

- **OpenAI**: GPT o3pro, GPT o3, GPT o1, GPT 4o, GPT o4 mini, GPT 4o mini, GPT 5
- **Google Gemini**: Gemini 2.5 Pro, Gemini Flash 2.5
- **Anthropic**: Claude4
- **xAI:** Grok3
- **Deepseek:** DeepseekV3

Further, each question has been tested on each model under the following three prompt conditions:

- Base (No Prompt): Question was posed in a zero-shot manner with no additional context.
- 1st variant (Persona Prompt): Question was posed in zero-shot persona mode with contextual information about the model's role. (The details of the contextual information given have been provided in **Annexure 2**).
- 2nd variant (Persona Prompt with Few Shot): Question was posed in few-shot persona mode with both context and an illustrative example. (The details of the contextual information and illustrative examples given have been provided in **Annexure 2**).

4.3 Comparison using State-of-the-art Methods

The models in combination with the above variants have been referred to as 'LLM candidates' henceforth. Each LLM candidate's responses were scored by the LLM/Human as the judge based on the IRAC+ framework as detailed in Table 4 above. The evaluation prompt, along with the scoring rubric provided to LLMs and Humans for scoring purposes, has been enclosed in **Annexure 1**. Further, we have enclosed the details of 3 sample queries along with ground truth and the responses provided by sample LLM candidates and the accuracy % evaluated by Judge GPT 4o in **Attachment 1**.

4.3.1 Model performance and rankings (GPT-4o as Judge)

Evaluation Headline

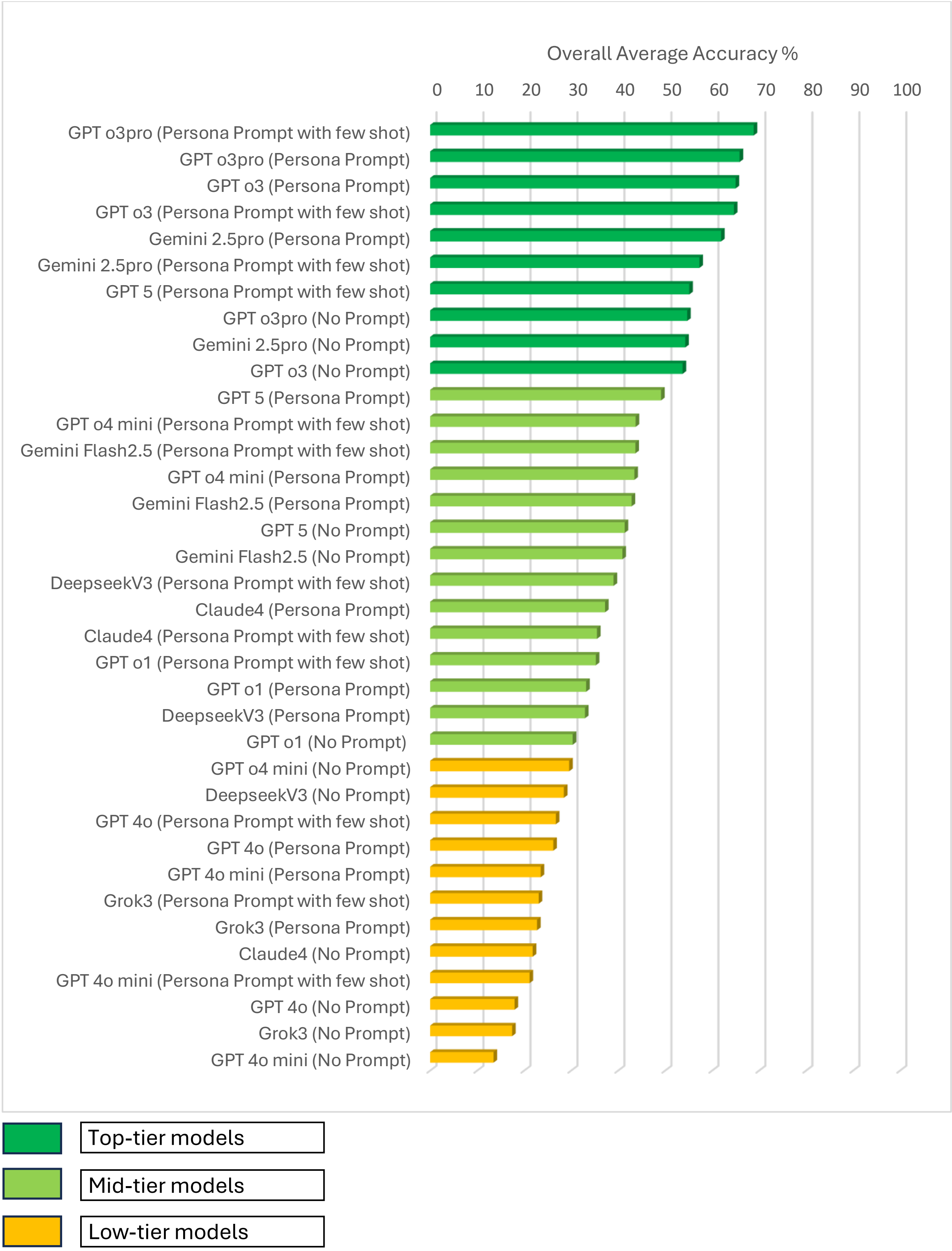
Before delving into the detailed results, it is essential to provide a summary of the key findings to help readers grasp the overall performance of the models. Our evaluation reveals a clear stratification in model performance, with proprietary models such as GPT o3pro, GPT o3, Gemini 2.5 Pro and GPT 5 consistently outperforming open-weight and lightweight models, particularly on complex reasoning tasks. The results indicate that models with deeper alignment and better contextual reasoning capabilities demonstrate superior performance. While GPT 5 is the most recent model considered for evaluation and is supposed to be great at reasoning through complex tasks like coding and multi-step planning, use high reasoning effort. We noted that, it fails to adhere to the requirements of detailed logical reasoning as expected by the legal experts which chat completion models would usually achieve.

The evaluation also highlights the significant impact of persona prompting and few-shot learning on model accuracy. While persona prompting boosts performance across all models, the magnitude of improvement varies, with some models showing substantial gains and others exhibiting limited or negative gains from few-shot examples. This suggests that dynamic, context-aware few-shot examples may yield better results for all models.

Overall, the findings underscore the importance of model depth, domain-aligned fine-tuning, and effective prompting strategies in achieving high accuracy in tax reasoning tasks. The detailed results section will provide a comprehensive analysis of model performance across various dimensions, including issue identification, rule identification, application of law, conclusion, interpretation, and justification.

Using the scores provided by GPT 4o on the responses from the 12 models under 3 prompt conditions, Diagram 1 presents the overall average accuracy per LLM candidate, based on the quality threshold defined by the evaluation rubric.

Diagram 1: Overall Average Accuracy per LLM candidate using scores provided by GPT 4o



In section 4.6.2, we have validated the overall average accuracy% provided by the Judge GPT 4o using correlation analysis with overall average accuracy% provided by Gemini 2.5 Flash (There is a correlation of 0.977 when the same questions considered for computing overall average % of accuracy by Gemini 2.5 Flash for a model were considered for computing overall average % of accuracy by GPT 4o for that model).

In section 4.6.3, we have validated the overall average accuracy% provided by the Judge GPT 4o using correlation analysis with overall average accuracy% provided by human experts (There is a correlation of 0.97 when the same questions considered for computing overall average % of accuracy by human experts for a model were considered for computing overall average % of accuracy by GPT 4o for that model).

4.3.2 Stratified performance gradient basis, clear tier separation

Models separate cleanly into three clusters based on their overall accuracy:

- Top-Tier (more than 50%): GPT o3pro (all variants), GPT o3 (all variants), Gemini 2.5pro (all variants), GPT 5 (Persona Prompt with few shot)

- Mid-Tier (30% to 50%): GPT o4 mini (Personal Prompt with few shot and Persona Prompt), Gemini Flash2.5 (all variants), DeepseekV3 (Personal Prompt with few shot and Persona Prompt), Claude4 (Personal Prompt with few shot and Persona Prompt), GPT o1 (all variants), GPT 5 (Persona Prompt), GPT 5 (No Prompt)

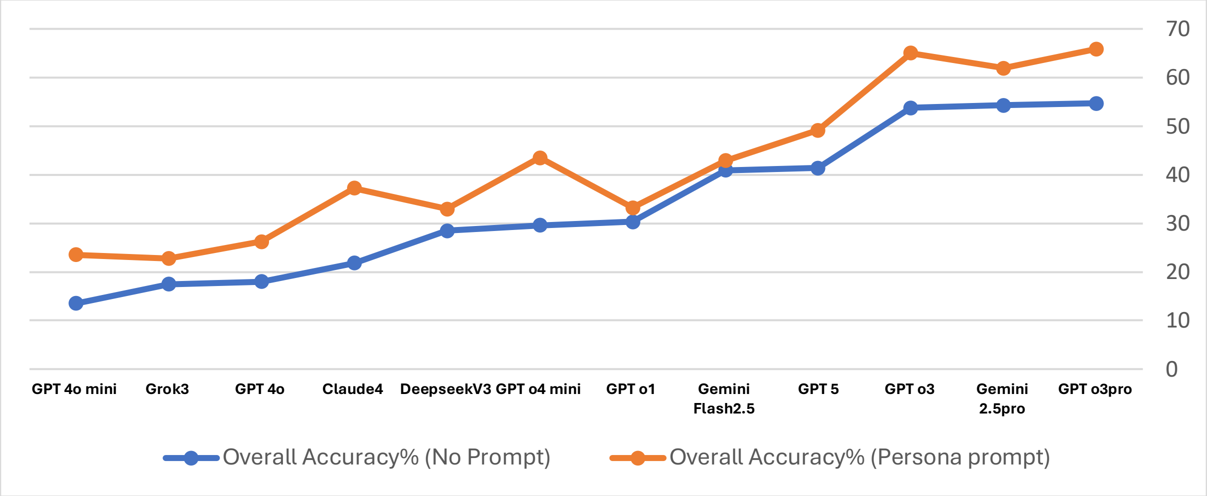
- Low-Tier (less than 30%): GPT o4 mini (No Prompt), DeepseekV3 (No Prompt), GPT 4o (all variants), GPT 4o mini (all variants), Grok3 (all variants), Claude4 (No Prompt)

The clear clustering indicates that performance is not random but is strongly influenced by instruction tuning (1st and 2nd variants) and base model family (GPT o3, GPT o3pro and Gemini 2.5Pro). The increase in accuracy of the responses from an LLM due to structured prompts reflects internal alignment capabilities. Models with deeper alignment and better contextual reasoning consistently outperform lightweight or minimally tuned variants.

4.3.3 Benefit of the persona prompting and examples:

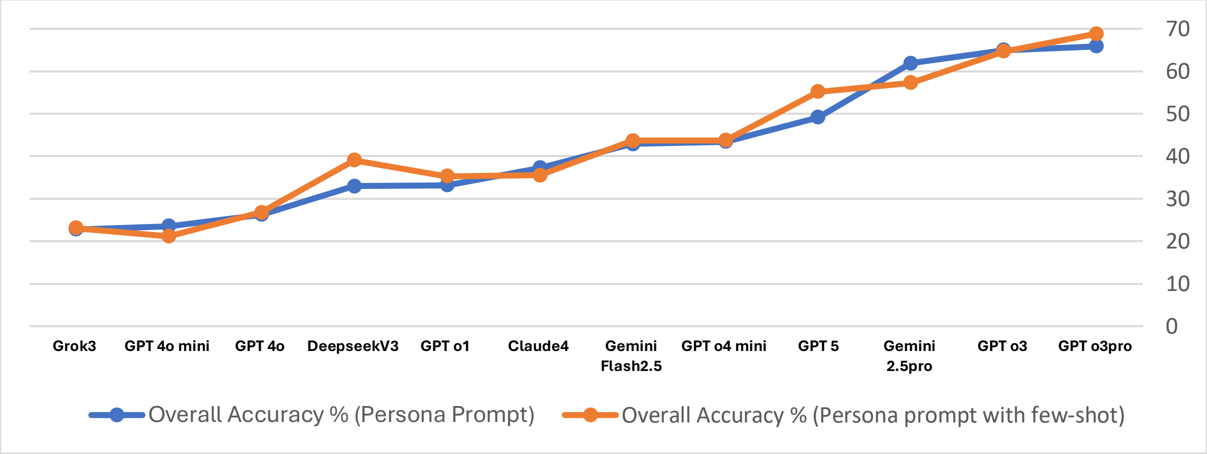
While persona prompting boosts performance across all models, indicating responsiveness to structured cues, the magnitude of benefit varies significantly by the models:

Diagram 2: Improvement in accuracy due to persona prompting



Furthermore, few-shot prompting is beneficial up to a certain threshold of competency. Beyond that, too much context may crowd out latent reasoning capability or introduce contradictions.

Diagram 3: Improvement in accuracy due to few-shot prompting



Key insights

For the models, GPT o3, Gemini 2.5 Pro, Claude4 and GPT 4o mini, adding few-shot examples for already effective persona prompts reduced performance (likely due to noise or overloading token context).

For the models GPT o4 mini, Gemini Flash2.5, GPT 4o, and Grok3, adding few-shot examples showed slight improvements in the scores. In case of the models GPT o3 pro, GPT5, Deepseek V3 and GPT o1, adding few shots for additional context demonstrated substantial improvements in the scores.

The plausible factor influencing the limited or negative gains from few-shot prompting in most of the models could be the use of static few-shot examples, which remain unchanged across all test cases. While this approach ensures consistency, it may inadvertently introduce irrelevant signals, repetitive context, or non-adaptive guidance, especially for high-capacity models that already perform well with minimal prompting. In such cases, static prompts can lead to context saturation, crowding out the model's internal reasoning processes, or even causing subtle contradictions with the query-specific intent. This suggests that dynamic or context-aware few-shot examples, tailored to the specific task or input, might yield better results for all the models.

However, in highly superior model GPT o3 pro and less contextualised models like Deepseek V3 and GPT o1, even static few-shot seems to boost performance.

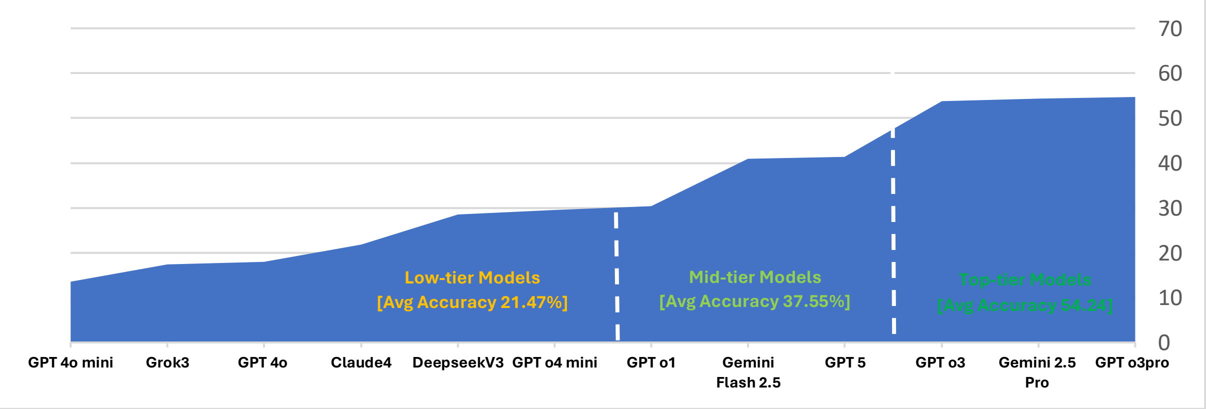
4.3.4 Dominance of proprietary models:

- All top-tier performers are proprietary (from OpenAI and Google).
- The open-weight model (i.e., DeepseekV3) did not breach the 50% threshold, even under ideal prompting conditions.
- Even with few-shot examples, the models DeepseekV3, Claude4, GPT o1, GPT 4o, Grok3, and GPT 4o mini failed to cross the 40% threshold.
- Models like GPT 4o and GPT 4o mini, performed marginally poorly, showcasing that model training depth matters more than efficiency-optimised design, which may sacrifice critical reasoning abilities.
- In tax reasoning tasks, quality depends more on model depth and domain-aligned fine-tuning than on compute efficiency.

4.3.5 Comparison of the tier-wise accuracy:

- Diagram 4 contains a grouping of models (for the no prompt variant) by tier and a comparison of their average accuracy:

Diagram 4: Grouping of models by tier and comparison of their average accuracy



Mid-Tier models enjoy 74.92% higher accuracy over Low-Tier models while Top-Tier models enjoy 44.46% higher accuracy over Mid-Tier models. The steep performance drop highlights the lag in deep tax reasoning for lower-capacity or efficiency-optimised models.

4.3.6 Key Insights:

- Persona prompting works, especially for mid- and low-tier models. However, few-shot prompting requires dynamic, context-aware adaptation to avoid diminishing returns.
- Model scale and alignment drive success: GPT o3pro and Gemini 2.5pro consistently dominate across prompting strategies.
- Open-weight models lag significantly behind proprietary counterparts in domain-specific reasoning, underscoring the gap in fine-tuning depth.

4.4 Qualitative Evaluation

4.4.1 Dimension-Wise benchmarking of LLMs for tax reasoning (IRAC+ % scores)

Responses from each LLM candidate have been evaluated across six legal reasoning dimensions: Issue Identification, Rule Identification, Application of Law, Conclusion, Interpretation, and Justification, along with the computation of an overall accuracy score. This breakdown enables a granular understanding of each LLM candidate's strengths and weaknesses in handling the IRAC+ framework.

Table 7: Overall accuracy per LLM candidate taking into consideration the IRAC+ framework

Model	Issue Id (%)	Rule Id (%)	Apply Law (%)	Conclusion (%)	Interpretation (%)	Justification (%)	Overall Accuracy
GPT o3pro (Persona Prompt with few shot)	69.09	69.62	68.28	68.82	68.82	68.55	68.86
GPT o3pro (Persona Prompt)	66.67	66.67	65.32	65.86	65.59	65.05	65.86
GPT o3 (Persona Prompt)	65.86	66.13	64.52	64.78	64.52	64.25	65.01
GPT o3 (Persona Prompt with few shot)	65.86	65.32	63.98	65.05	64.25	63.44	64.65
Gemini 2.5pro (Persona Prompt)	64.52	63.44	60.75	61.02	61.02	60.75	61.92
Gemini 2.5pro (Persona Prompt with few shot)	59.68	59.41	55.65	56.72	56.45	55.91	57.30
GPT 5 (Persona with few shot)	55.91	55.65	54.84	55.11	55.11	54.30	55.15
GPT o3pro (No Prompt)	55.91	55.38	54.03	55.38	54.30	53.23	54.70

Model	Issue Id (%)	Rule Id (%)	Apply Law (%)	Conclusion (%)	Interpretation (%)	Justification (%)	Overall Accuracy
Gemini 2.5pro (No Prompt)	57.53	54.84	52.96	54.57	52.96	52.96	54.30
GPT o3 (No Prompt)	54.84	54.57	52.96	54.57	53.23	52.15	53.72
GPT 5 (Persona prompt)	50.54	49.73	48.66	49.46	48.66	47.85	49.15
GPT o4 mini (Persona Prompt with few shot)	46.77	44.62	42.20	44.09	42.74	41.94	43.73
Gemini Flash2.5 (Persona Prompt with few shot)	48.12	47.04	40.86	44.09	41.40	40.59	43.68
GPT o4 mini (Persona Prompt)	47.31	44.62	41.40	43.55	42.47	41.40	43.46
Gemini Flash2.5 (Persona Prompt)	46.24	44.89	40.86	43.28	41.40	40.59	42.88
GPT 5 (No Prompt)	44.09	42.47	40.32	41.94	40.86	38.71	41.40
Gemini Flash2.5 (No Prompt)	44.89	42.74	38.71	41.13	39.25	38.71	40.91
DeepseekV3 (Persona Prompt with few shot)	44.38	41.29	36.80	38.20	36.80	36.80	39.04
Claude4 (Persona Prompt)	46.02	38.92	33.52	36.65	34.66	33.52	37.22
Claude4 (Persona Prompt with few shot)	44.57	37.50	31.79	35.33	32.07	31.79	35.51
GPT o1 (Persona Prompt with few shot)	38.71	37.37	33.06	36.02	33.60	32.80	35.26
GPT o1 (Persona Prompt)	37.10	34.68	31.18	34.14	31.72	30.38	33.20
DeepseekV3 (Persona Prompt)	39.55	35.91	29.55	33.64	29.55	29.55	32.95
GPT o1 (No Prompt)	34.14	31.45	28.49	31.18	29.03	27.69	30.33
GPT o4 mini (No Prompt)	32.69	30.13	27.88	30.45	29.17	27.24	29.59

Model	Issue Id (%)	Rule Id (%)	Apply Law (%)	Conclusion (%)	Interpretation (%)	Justification (%)	Overall Accuracy
DeepseekV3 (No Prompt)	34.03	30.21	25.35	29.51	26.39	25.35	28.47
GPT 4o (Persona Prompt with few shot)	30.65	28.76	24.73	27.69	24.73	23.92	26.75
GPT 4o (Persona Prompt)	30.38	27.96	23.66	27.42	24.46	23.39	26.21
GPT 4o mini (Persona Prompt)	29.03	24.19	21.24	23.66	21.77	21.24	23.52
Grok3 (Persona Prompt with few shot)	27.45	23.91	21.20	23.10	22.01	20.92	23.10
Grok3 (Persona Prompt)	29.21	24.16	19.38	24.44	19.94	19.38	22.75
Claude4 (No Prompt)	27.99	21.47	19.57	22.83	19.57	19.57	21.83
GPT 4o mini (Persona Prompt with few shot)	26.34	21.77	19.09	21.51	19.35	18.82	21.15
GPT 4o (No Prompt)	23.92	18.28	15.05	18.82	16.67	15.05	17.97
Grok3 (No Prompt)	21.20	18.21	15.22	18.75	16.03	15.22	17.44
GPT 4o mini (No Prompt)	19.09	13.17	11.29	14.52	11.83	11.02	13.49

Below are key insights from this dimension-level evaluation:

4.4.2 Weak areas in the bottom three models

Table 8: Key weakness areas in the bottom three models

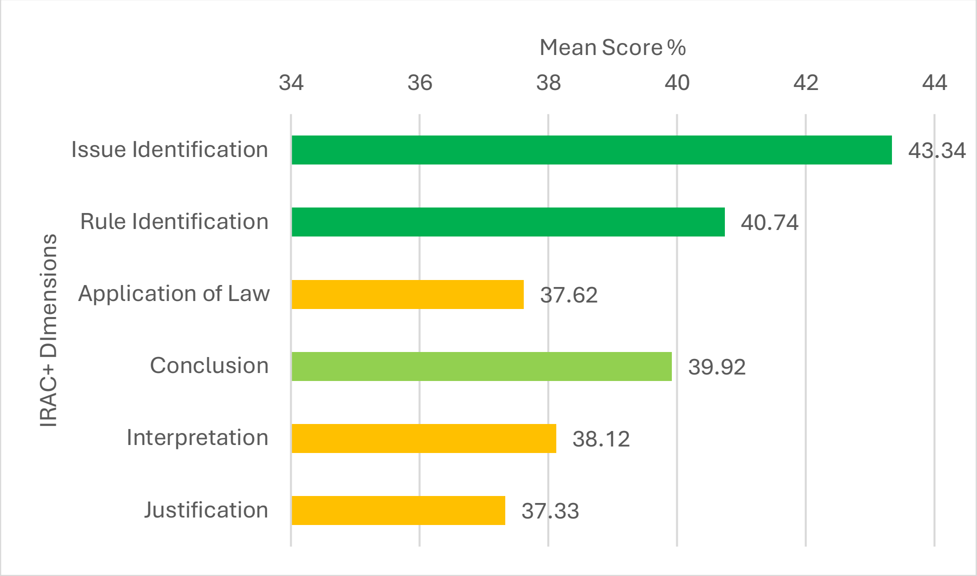
Model	Overall Accuracy (%)	Key Weakness (Parameters having score below the overall accuracy of the models)
GPT 4o (No Prompt)	17.97	Application of Law, Interpretation and Justification
Grok3 (No Prompt)	17.44	Application of Law, Interpretation and Justification
GPT 4o mini (No Prompt)	13.49	Rule Identification, Application of Law, Interpretation and Justification

LLM candidates with overall low accuracy performed poorly in complex legal reasoning tasks. This reflects limited contextual understanding and poor alignment between legal interpretation and applied reasoning for these LLM candidates.

4.4.3 Task specific observations:

Diagram 5 provides the analysis of average scores on respective IRAC+ dimensions across all 36 LLM candidates:

Diagram 5: Analysis of average scores across all 36 LLM candidates



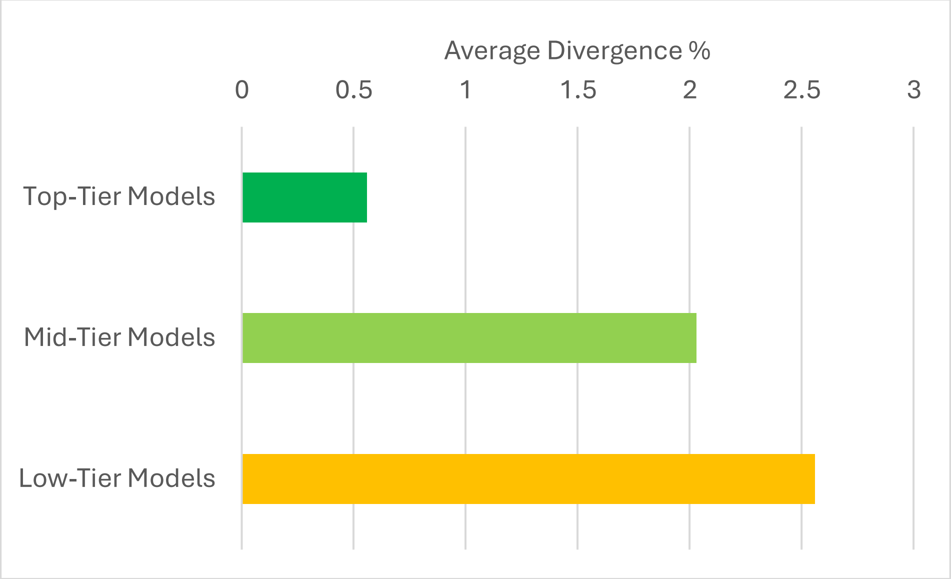
Issue and Rule Identification are generally well-handled by most LLM candidates, likely due to the syntactic nature of these tasks, even for the tax reasoning questions. The application of Law and Justification are the weakest dimensions across the board, since both require logical inference, multi-hop reasoning, and contextual statutory adaptation, making them ideal stress tests for legal LLMs. Conclusion and Interpretation act as tie-breakers: top-tier models maintain scores above 50%, while weaker models drop sharply.

Any future fine-tuning should focus on Application and Justification, as these dimensions most strongly correlate with overall model performance in complex legal reasoning.

4.4.4 Interpretation vs Conclusion divergence:

The following are the details of the tier-wise average divergence for the respective tiers of the models:

Diagram 6: Tier-wise average divergence between scores for interpretation and conclusion dimensions



Avg divergence for All Models - 1.8

- The average gap between interpretation and conclusion across all models is 1.80 percentage points.
- The low divergence in Top-Tier models indicates stable decision-making after accurate interpretation. This suggests alignment between latent representations of the law and decision logic in the best models.
- Persona variants of the Top-Tier models maintain a tight Interpretation-Conclusion divergence band of just 0.23% while no-persona variants maintain the divergence band of around 1.34%.

4.4.5 Intra-Model stability (volatility analysis of scores on IRAC+ components):

Table 9 provides the "volatility" for the models as the divergence between their best and worst scoring IRAC+ dimension:

Table 9: Volatility per LLM candidate

Model	Max Score for the parameters (%)	Min Score for the parameters (%)	Gap (%)
Top-Tier			
GPT o3pro (Persona Prompt with few shot)	69.62	68.28	1.34
GPT o3pro (Persona Prompt)	66.67	65.05	1.61
GPT o3 (Persona Prompt)	66.13	64.25	1.88
GPT o3 (Persona Prompt with few shot)	65.86	63.44	2.42
Gemini 2.5pro (Persona Prompt)	64.52	60.75	3.76
Gemini 2.5pro (Persona Prompt with few shot)	59.68	55.65	4.03
GPT 5 (Persona Prompt with few shot)	55.91	54.30	1.61
GPT o3pro (No Prompt)	55.91	53.23	2.69
Gemini 2.5pro (No Prompt)	57.53	52.96	4.57

Model	Max Score for the parameters (%)	Min Score for the parameters (%)	Gap (%)
GPT o3 (No Prompt)	54.84	52.15	2.69
Mid-Tier			
GPT 5 (Persona Prompt)	50.54	47.85	2.69
GPT o4 mini (Persona Prompt with few shot)	46.77	41.94	4.84
Gemini Flash2.5 (Persona Prompt with few shot)	48.12	40.59	7.53
GPT o4 mini (Persona Prompt)	47.31	41.40	5.91
Gemini Flash2.5 (Persona Prompt)	46.24	40.59	5.65
GPT 5 (No Prompt)	44.09	38.71	5.38
Gemini Flash2.5 (No Prompt)	44.89	38.71	6.18
DeepseekV3 (Persona Prompt with few shot)	44.38	36.80	7.58
Claude4 (Persona Prompt)	46.02	33.52	12.50
Claude4 (Persona Prompt with few shot)	44.57	31.79	12.77
GPT o1 (Persona Prompt with few shot)	38.71	32.80	5.91
GPT o1 (Persona Prompt)	37.10	30.38	6.72
DeepseekV3 (Persona Prompt)	39.55	29.55	10.00
GPT o1 (No Prompt)	34.14	27.69	6.45
Low-Tier			
GPT o4 mini (No Prompt)	32.69	27.24	5.45
DeepseekV3 (No Prompt)	34.03	25.35	8.68
GPT 4o (Persona Prompt with few shot)	30.65	23.92	6.72
GPT 4o (Persona Prompt)	30.38	23.39	6.99
GPT 4o mini (Persona Prompt)	29.03	21.24	7.80
Grok3 (Persona Prompt with few shot)	27.45	20.92	6.52
Grok3 (Persona Prompt)	29.21	19.38	9.83
Claude4 (No Prompt)	27.99	19.57	8.42
GPT 4o mini (Persona Prompt with few shot)	26.34	18.82	7.53
GPT 4o (No Prompt)	23.92	15.05	8.87
Grok3 (No Prompt)	21.20	15.22	5.98
GPT 4o mini (No Prompt)	19.09	11.02	8.06

Further, Table 10 provides the “volatility” for the tier-wise model groups as the average divergence between their average of the best and average of the worst scoring IRAC+ dimension for the respective tier:

Table 10: Volatility for the tier-wise model groups based on the average divergence

Model	Max Score on one of the parameters (%) [Average]	Min Score on one of the parameters (%) [Average]	Gap (%) [Average]
Top-Tier Models	61.67	59.01	2.66
Mid-Tier Models	43.74	36.59	7.15
Low-Tier Models	27.66	20.09	7.57
All Models	**43.36**	**37.32**	**6.04**

Top-tier models show tight intra-model consistency, indicating robust cross-domain reasoning.

In contrast, weaker models swing wildly across dimensions, suggesting fragile attention or inconsistent internal reasoning templates.

4.4.6 Key Insights:

- This decomposition confirms that overall legal reasoning accuracy derives from balanced strength across all IRAC+ pillars, particularly in Application of Law and Justification.
- The LE-BTL framework, thus, not only offers a composite score but also provides explanatory insights into dimension-specific reasoning capabilities, enabling targeted fine-tuning of weaker areas in LLMs.

4.5 Performance Evaluation

4.5.1 Complexity-wise scoring of the LLM candidates

We have provided below in Table 11 the comparative summary of the accuracy of the responses from LLM candidates on different categories of questions, viz., Simple, Medium and Complex.

Table 11: Comparative summary of the accuracy across Simple, Medium, and Complex questions

Models	Accuracy % on 'Complex' Questions	Accuracy % on 'Medium'	Accuracy % on 'Simple' Questions	Overall Accuracy %
Top-Tier Models				
GPT o3pro (Persona Prompt with few shot)	71.28	69.44	62.06	68.86
GPT o3pro (Persona Prompt)	67.64	64.35	63.60	65.86
GPT o3 (Persona Prompt)	67.55	64.20	59.87	65.01
GPT o3 (Persona Prompt with few shot)	67.02	64.20	59.43	64.65
Gemini 2.5pro (Persona Prompt)	63.56	61.57	58.33	61.92
Gemini 2.5pro (Persona Prompt with few shot)	60.64	52.16	56.36	57.30

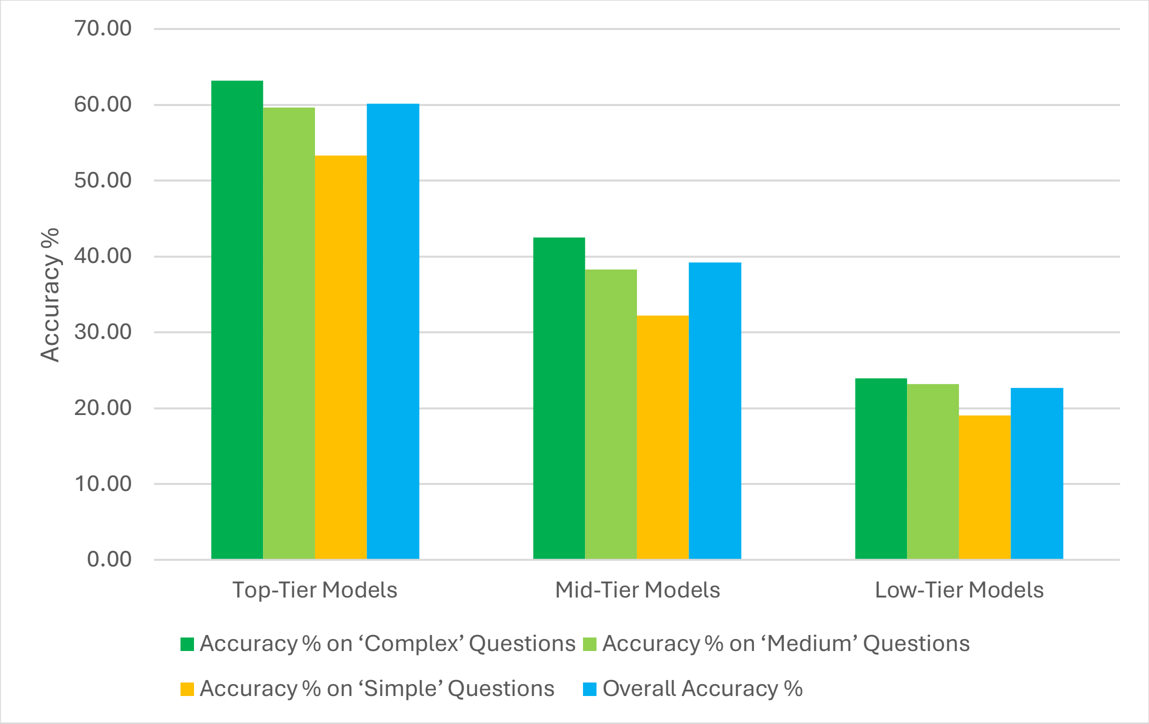
Models	Accuracy % on 'Complex' Questions	Accuracy % on 'Medium'	Accuracy % on 'Simple' Questions	Overall Accuracy %
GPT 5 (Persona Prompt with few shot)	60.28	51.54	47.59	55.15
GPT o3pro (No Prompt)	58.87	56.17	42.32	54.70
Gemini 2.5pro (No Prompt)	55.32	58.02	46.49	54.30
GPT o3 (No Prompt)	59.57	55.09	37.28	53.72
Mid-Tier Models				
GPT 5 (Persona Prompt)	49.20	49.85	48.03	49.15
GPT o4 mini (Persona Prompt with few shot)	46.54	43.36	37.28	43.73
Gemini Flash2.5 (Persona Prompt with few shot)	46.90	43.83	35.53	43.68
GPT o4 mini (Persona Prompt)	47.69	41.20	36.18	43.46
Gemini Flash2.5 (Persona Prompt)	45.12	41.05	39.91	42.88
GPT 5 (No Prompt)	44.68	41.82	32.67	41.40
Gemini Flash2.5 (No Prompt)	44.33	42.44	30.26	40.91
DeepseekV3 (Persona Prompt with few shot)	42.52	36.42	34.49	39.04
Claude4 (Persona Prompt)	42.93	35.96	21.95	37.22
Claude4 (Persona Prompt with few shot)	39.54	34.57	26.39	35.51
GPT o1 (Persona Prompt with few shot)	39.18	32.10	30.04	35.26
GPT o1 (Persona Prompt)	36.52	31.64	27.19	33.20
DeepseekV3 (Persona Prompt)	34.17	34.21	29.95	32.95
GPT o1 (No Prompt)	35.55	27.47	21.49	30.33
Low-Tier Models				
GPT o4 mini (No Prompt)	31.87	27.84	27.19	29.59
DeepseekV3 (No Prompt)	31.11	28.69	23.18	28.47
GPT 4o (Persona Prompt with few shot)	27.30	26.70	25.44	26.75
GPT 4o (Persona Prompt)	28.19	25.93	21.71	26.21
GPT 4o mini (Persona Prompt)	24.47	23.46	21.27	23.52
Grok3 (Persona Prompt with few shot)	22.61	23.92	23.15	23.10
Grok3 (Persona Prompt)	22.63	23.61	21.76	22.75
Claude4 (No Prompt)	22.61	25.00	15.05	21.83
GPT 4o mini (Persona Prompt with few shot)	21.37	20.99	20.83	21.15
GPT 4o (No Prompt)	20.30	18.52	11.40	17.97
Grok3 (No Prompt)	18.79	20.22	9.72	17.44
GPT 4o mini (No Prompt)	16.05	12.96	7.89	13.49

Based on data presented in Table 11, several key insights emerge regarding the performance of various LLM candidates across question complexity levels:

- In a departure from conventional expectations, the observed models exhibit lower accuracy on simple questions while achieving higher accuracy on complex ones. This inverse relationship indicates that the LLMs may be disproportionately attuned to handling intricate queries, potentially as a result of overfitting to training datasets rich in complex prompts. Such a trend implies that these models are being optimised primarily for advanced reasoning, which may inadvertently come at the cost of their ability to effectively address basic comprehension and straightforward recall tasks.

- Tier-wise Analysis:

Diagram 7: Tier-wise comparative summary of the accuracy across Simple, Medium, and Complex questions



Top-tier models, GPT o3pro and Gemini 2.5pro variants, demonstrate their strongest accuracy on complex questions, consistently outperforming both mid-tier and low-tier counterparts. Notably, these advanced models exhibit a relatively narrow range in accuracy across varying levels of question complexity, indicating that they possess robust internal mechanisms for reasoning and attention. This enables them to tackle both simple and complex queries with comparable proficiency.

In contrast, mid-tier models display greater variability in accuracy between simple and complex questions. This inconsistency suggests that their ability to generalise is less effective, making them more sensitive to the intricacies of prompt design and leading to fluctuating performance depending on the question type.

Low-tier models, such as Grok3 and GPT 4o mini, face challenges not only with complex queries but also with basic ones, failing to achieve satisfactory accuracy across the board. These results point to foundational limitations in their underlying architectures or training data. Specifically, the persistent underperformance of low-tier models on both complex and simple questions may stem from insufficient diversity in training data, restricted architectural capabilities, or limited exposure to varied prompt formats during development.

- Varying Impact of Prompt Engineering:

  Although prompt engineering and few-shot learning are generally recognised for enhancing model performance, their positive impact is notably greater for complex questions than for simpler ones. For instance, GPT o3pro (Persona Prompt with few shot) records the highest overall accuracy, illustrating that advanced prompting methods substantially assist models in managing multi-step reasoning tasks. However, these strategies appear to offer limited improvement for basic recall or straightforward queries. Consequently, models employing prompt engineering or few-shot approaches tend to exhibit modest gains on simple questions, with the benefits becoming increasingly significant as question complexity rises.

- Implications for Model Design and Training:

  The above observed pattern indicates that current LLM training and evaluation practices tend to prioritise nuanced and contextual understanding, often at the expense of accuracy and clarity when addressing straightforward questions. This trend highlights a critical need for developers to revisit both their training datasets and evaluation strategies to ensure balanced model performance across all levels of question complexity.

  For organisations leveraging LLMs, it is important to acknowledge that even advanced models may underperform on routine or basic queries, an issue with significant operational consequences, particularly in sectors such as tax advisory where the ability to handle both simple and complex queries is vital for effective service delivery.

  These findings emphasise the importance of developing comprehensive training and evaluation frameworks. Expanding datasets to include a wider variety of simple question formats and refining evaluation protocols will help prevent models from disproportionately favouring complex reasoning over fundamental comprehension. Additionally, the data suggests that targeted fine-tuning, especially for basic question types, can substantially improve model reliability and utility, thereby enhancing performance across a range of real-world applications.

- In summary, Table 11 demonstrate a consistent decline instead of substantial increase in accuracy as the complexity of questions decreases, highlighting an unexpected trend wherein simpler queries pose disproportionately greater challenges for LLMs. This observation underscores the critical need to rebalance model training so that robust and reliable language understanding is achieved across both basic and advanced tasks. The above findings not only affirm the complexity-dependent performance of LLMs but also identifies key areas for strategic improvement, namely, prompt engineering, expanding dataset diversity, and implementing targeted fine-tuning. To maximise the effectiveness of LLMs, it is essential for stakeholders to rigorously evaluate and optimise these models across all question types, ensuring an equilibrium between sophisticated reasoning and fundamental comprehension for comprehensive language understanding.

4.5.2 Topic-wise scoring of the LLM candidates

Table 12 presents a detailed comparison of the performance of various LLM candidates across four key tropical areas: Accounting Standard, Direct Tax, FEMA and Indirect Tax, along with their overall accuracy.

Table 12: Comparison of the performance across four key tropical areas

Models	**Accuracy % on 'Accounting Standard'**	**Accuracy % on 'Direct Tax'**	**Accuracy % on 'FEMA'**	**Accuracy % on 'Indirect Tax'**	**Overall Accuracy %**
Top-Tier Models					
GPT o3pro (Persona Prompt with few shot)	59.38	72.84	60.23	67.86	68.86
GPT o3pro (Persona Prompt)	52.08	69.79	63.64	63.69	65.86
GPT o3 (Persona Prompt)	56.25	68.01	63.26	61.90	65.01
GPT o3 (Persona Prompt with few shot)	48.26	68.01	62.50	66.96	64.65
Gemini 2.5pro (Persona Prompt)	54.86	63.10	66.29	59.82	61.92
Gemini 2.5pro (Persona Prompt with few shot)	50.69	56.99	64.77	58.33	57.30
GPT 5 (Persona Prompt with few shot)	45.83	59.45	50.00	50.00	55.15
GPT o3pro (No Prompt)	45.14	59.90	43.56	50.89	54.70
Gemini 2.5pro (No Prompt)	44.10	56.70	46.97	59.23	54.30
GPT o3 (No Prompt)	37.85	59.67	47.35	48.51	53.72
Mid-Tier Models					
GPT 5 (Persona Prompt)	41.67	51.64	48.48	46.13	49.15
GPT o4 mini (Persona Prompt with few shot)	46.18	45.24	48.86	31.55	43.73
Gemini Flash2.5 (Persona Prompt with few shot)	35.42	46.87	39.39	41.37	43.68

Models	Accuracy % on 'Accounting Standard'	Accuracy % on 'Direct Tax'	Accuracy % on 'FEMA'	Accuracy % on 'Indirect Tax'	Overall Accuracy %
GPT o4 mini (Persona Prompt)	45.49	46.50	41.67	30.95	43.46
Gemini Flash2.5 (Persona Prompt)	41.32	45.46	40.53	35.71	42.88
GPT 5 (No Prompt)	32.64	47.40	32.20	32.14	41.40
Gemini Flash2.5 (No Prompt)	35.42	44.64	29.92	39.29	40.91
DeepseekV3 (Persona Prompt with few shot)	32.96	41.74	36.74	35.42	39.04
Claude4 (Persona Prompt)	33.33	43.16	31.44	21.15	37.22
Claude4 (Persona Prompt with few shot)	35.98	39.29	34.47	20.83	35.51
GPT o1 (Persona Prompt with few shot)	28.82	38.02	34.09	30.66	35.26
GPT o1 (Persona Prompt)	27.43	35.71	29.92	30.66	33.20
DeepseekV3 (Persona Prompt)	35.61	34.65	30.30	30.65	32.95
GPT o1 (No Prompt)	24.31	34.08	19.70	28.87	30.33
Low-Tier Models					
GPT o4 mini (No Prompt)	43.40	27.44	31.82	22.32	29.59
DeepseekV3 (No Prompt)	26.89	32.18	20.08	26.79	28.47
GPT 4o (Persona Prompt with few shot)	26.74	26.26	27.27	28.27	26.75
GPT 4o (Persona Prompt)	26.04	27.23	25.00	23.21	26.21
GPT 4o mini (Persona Prompt)	26.04	25.07	21.21	16.96	23.52
Grok3 (Persona Prompt with few shot)	26.52	24.48	19.32	17.86	23.10
Grok3 (Persona Prompt)	19.70	24.61	24.24	16.97	22.75
Claude4 (No Prompt)	29.92	22.25	19.32	15.77	21.83
GPT 4o mini (Persona Prompt with few shot)	21.53	22.92	18.18	16.07	21.15
GPT 4o (No Prompt)	21.53	19.20	12.88	13.99	17.97
Grok3 (No Prompt)	20.45	19.57	11.74	11.01	17.44
GPT 4o mini (No Prompt)	12.50	16.00	7.20	9.23	13.49

Analysis:

- From the above analysis, it is evident that Direct Tax emerged as the strongest-performing category for most models, with top candidates like GPT o3pro (Persona Prompt with few shot) achieving an impressive 72.84% accuracy. This suggests that the models are better at handling queries where the rules and interpretations are more structured and well-documented, allowing for clearer reasoning paths.

- Accounting Standards and FEMA showed moderate performance across models, with scores largely clustering between 40% and 60% for mid-tier LLMs and peaking around 66% for the highest-performing candidates (e.g., Gemini 2.5pro with Persona Prompt for FEMA). These domains often require contextual interpretation and multi-step reasoning, which may explain the mixed performance levels observed across candidates.

- The above analysis reinforces the observation that persona prompting combined with a few-shot examples significantly boosts model performance across regulatory and compliance-based domains.

We further highlight the following key observations:

- Topic-specific strengths of models: GPT o3pro and GPT o3 (both with persona and few-shot prompting) consistently led performance across all topics.
- Significant role of prompting strategies: Persona prompting and few-shot learning provided clear performance improvements across all topics, with models in the "no prompt" category consistently ranking lower.
- Greater variability in niche topics: FEMA and Accounting Standards exhibited higher score dispersion across LLMs, suggesting that these areas require domain-tuned contextual reasoning that smaller or unoptimized models currently struggle with.
- The accuracy of the response relating to the FEMA and Accounting Standards emphasises the need for enhanced domain fine-tuning and improved prompt engineering strategies for these specialised areas.

4.6 Ablation Studies

4.6.1 Cross-validation of GPT 4o as the judge

To assess the reliability of GPT 4o under the LE-BTL framework, we conducted a cross-validation exercise using a representative subset of models and questions. The same set of model responses was evaluated under three different judge types:

- The LLM-as-a-Judge (GPT 4o)
- Alternative LLM-as-a-Judge (Gemini 2.5 Flash)
- Human Evaluation

4.6.2 Alternative LLM-as-a-Judge (Gemini 2.5 Flash):

- Alternative LLM-as-a-Judge was conducted using Gemini 2.5 Flash using the same rubric aligned with the IRAC+ structure. 2950 sample responses were evaluated using Gemini 2.5 Flash across 12 models (in all 3 prompt conditions). The scores provided by Gemini 2.5 Flash for these 2730 responses and the scores provided by GPT 4o for the same 2730 responses have a correlation of 0.76.

- The purpose was not to validate absolute scores, but to assess the relative consistency of model rankings across evaluators.

- In Table 13, for the 12 models in all 3 prompt conditions, we have enlisted the overall average accuracy % provided by Gemini 2.5 Flash, along with the overall average accuracy % provided by GPT 4o (The same questions considered for computing overall average % of accuracy by Gemini 2.5 Flash for a model were considered for computing overall average % of accuracy by GPT 4o for that model).

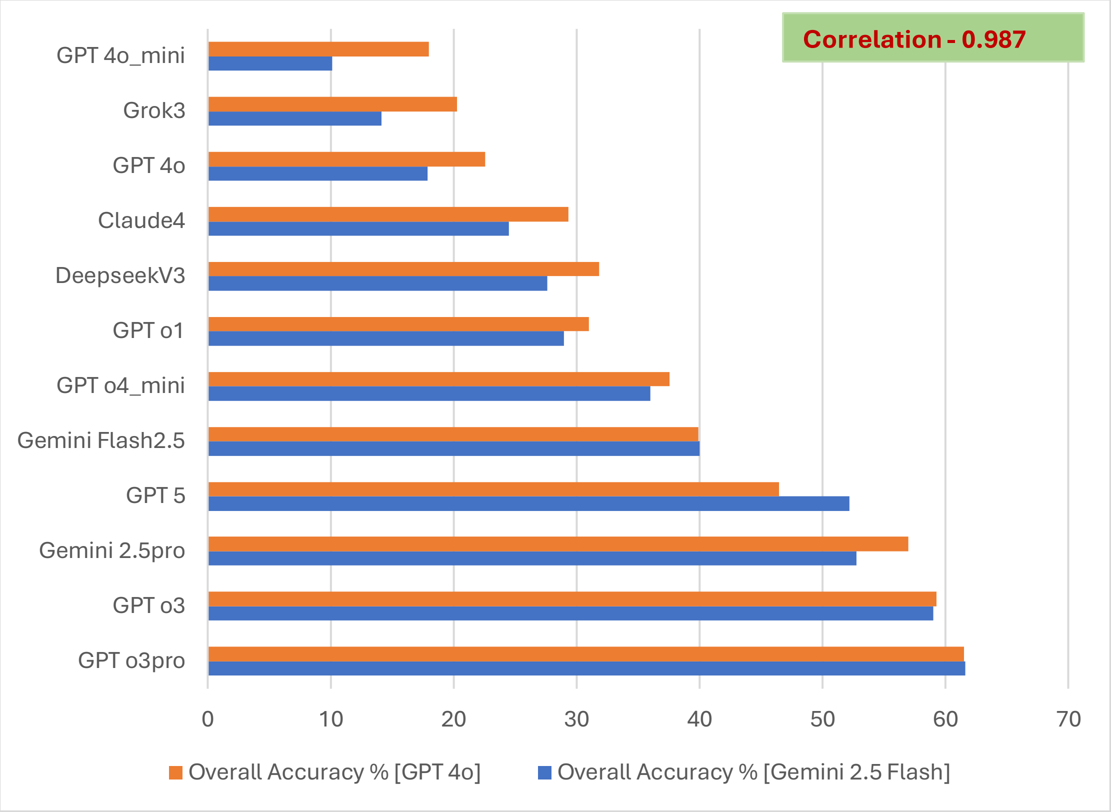
Table 13: Overall average accuracy % provided by Gemini 2.5 Flash and GPT 4o

S. No	Model	Overall Avg Accuracy % [Gemini 2.5 Flash]	Overall Avg Accuracy % [GPT 4o]
1	GPT o3pro (Persona Prompt with few shot)	66.13	68.24
2	GPT o3 (Persona Prompt with few shot)	65.31	63.64
3	GPT o3pro (Persona Prompt)	65.26	64.61
4	GPT o3 (Persona Prompt)	62.99	63.91
5	Gemini 2.5pro (Persona Prompt)	56.17	61.58
6	GPT 5 (Persona Prompt with few shot)	55.68	54.27
7	Gemini 2.5pro (Persona Prompt with few shot)	55.25	56.60
8	GPT 5 (Persona Prompt)	54.33	47.13
9	GPT o3pro (No Prompt)	53.41	51.79
10	GPT o3 (No Prompt)	48.81	50.32
11	Gemini 2.5pro (No Prompt)	46.92	52.76
12	GPT 5 (No Prompt)	46.59	37.93
13	Gemini Flash2.5 (Persona Prompt)	41.23	41.02
14	Gemini Flash2.5 (Persona Prompt with few shot)	41.18	41.13
15	GPT o4 mini (Persona Prompt)	38.69	41.13
16	Gemini Flash2.5 (No Prompt)	37.66	37.55
17	GPT o4 mini (Persona Prompt with few shot)	37.61	41.88
18	GPT o1 (Persona Prompt with few shot)	32.25	33.17
19	GPT o4 mini (No Prompt)	31.71	29.65
20	DeepseekV3 (Persona Prompt with few shot)	31.48	37.11
21	DeepseekV3 (Persona Prompt)	30.56	32.87
22	GPT o1 (Persona Prompt)	28.41	31.49
23	Claude4 (Persona Prompt with few shot)	27.71	32.95
24	Claude4 (Persona Prompt)	27.17	34.87

S. No	Model	Overall Avg Accuracy % [Gemini 2.5 Flash]	Overall Avg Accuracy % [GPT 4o]
25	GPT o1 (No Prompt)	26.24	28.35
26	DeepseekV3 (No Prompt)	20.76	25.51
27	GPT 4o (Persona Prompt)	20.29	25.38
28	GPT 4o (Persona Prompt with few shot)	19.37	25.92
29	Claude4 (No Prompt)	18.62	20.18
30	Grok3 (Persona Prompt)	16.44	21.62
31	GPT 4o (No Prompt)	14.02	16.34
32	Grok3 (No Prompt)	13.31	16.13
33	Grok3 (Persona Prompt with few shot)	12.72	23.05
34	GPT 4o mini (Persona Prompt)	12.01	22.67
35	GPT 4o mini (Persona Prompt with few shot)	11.36	20.24
36	GPT 4o mini (No Prompt)	7.03	10.98
	Correlation		**0.977**

- In Diagram 8, for the abovementioned 12 models, we have computed the average score across all three prompting conditions to normalise model performance across varying input formats by both Gemini 2.5Flash and GPT 4o.

Diagram 8: Average score across all three prompting conditions by Gemini 2.5Flash and GPT 4o



4.6.3 Human evaluation:

- Human evaluation was conducted by a subject-matter expert panel using the same rubric aligned with the IRAC+ structure. The expert panel has evaluated 244 sample responses across 8 sample models (in all 3 prompt conditions). The scores provided by the expert panel for these 244 responses and the scores provided by GPT 4o for the same 244 responses have a correlation of 0.87.

- The purpose was not to validate absolute scores, but to assess the relative consistency of model rankings across evaluators. We have enclosed the details of 3 sample queries along with ground truth and the responses provided by sample LLM candidates and the accuracy % evaluated by the Jury of GPT 4o and human evaluators in **Attachment 2**.

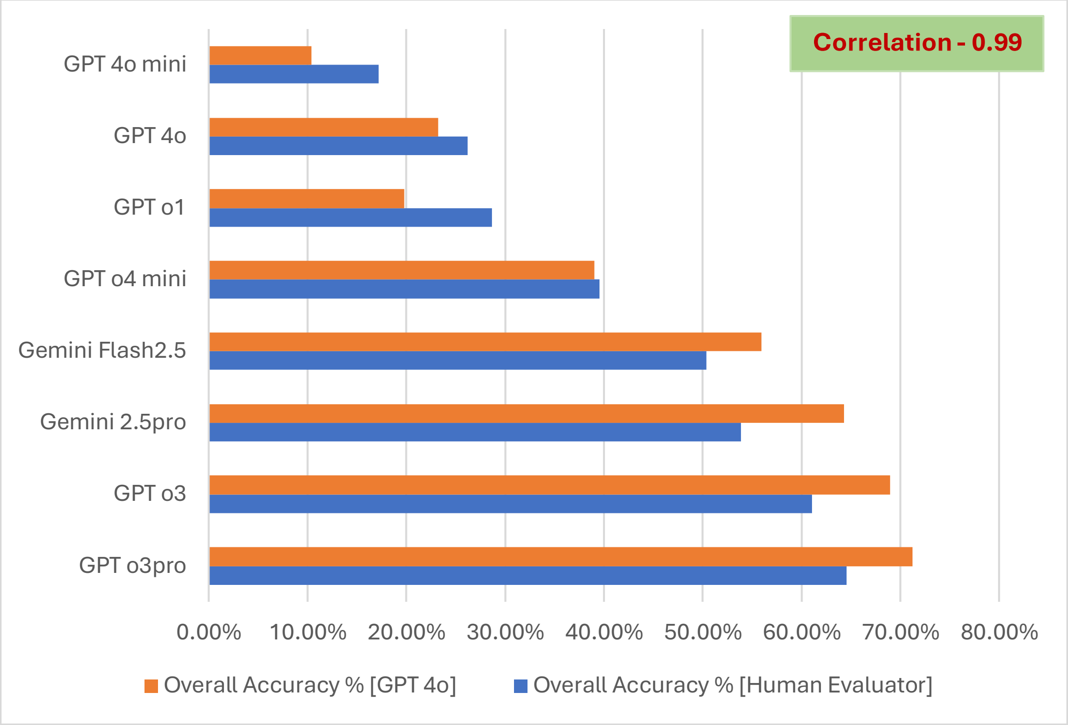
- In the below table 14, for the abovementioned 8 models in all 3 prompt conditions, we have enlisted the overall average accuracy % provided by the expert panel along with the overall average accuracy % provided by GPT 4o (The same questions considered for computing overall average accuracy% by expert panel for a model were considered for computing overall average accuracy% by GPT 4o for that model).

Table 14: Overall average accuracy % provided by the expert panel and GPT 4o

S. No	Model	Overall Avg Accuracy % [Human Evaluator]	Overall Avg Accuracy % [GPT 4o]
1	GPT o3pro (Persona Prompt with few shot)	65.83%	73.33%
2	GPT o3pro (No Prompt)	63.99%	69.35%
3	GPT o3pro (Persona Prompt)	63.89%	70.83%
4	GPT o3 (Persona Prompt with few shot)	63.89%	72.22%
5	GPT o3 (No Prompt)	61.61%	67.26%
6	GPT o3 (Persona Prompt)	57.78%	67.22%
7	Gemini 2.5pro (Persona	55.83%	67.50%
8	Gemini 2.5pro (Persona Prompt with few shot)	53.61%	62.50%
9	Gemini Flash2.5 (Persona Prompt with few shot)	52.50%	55.28%
10	Gemini 2.5pro (No Prompt)	52.08%	62.80%
11	GPT o4 mini (No Prompt)	50.00%	37.50%
12	Gemini Flash2.5 (No Prompt)	49.70%	57.74%
13	Gemini Flash2.5 (Persona Prompt)	48.89%	54.72%
14	GPT o4 mini (Persona Prompt with few shot)	44.44%	50.00%
15	GPT o1 (Persona Prompt with few shot)	41.67%	29.17%
16	GPT 4o (Persona Prompt)	29.44%	24.72%
17	GPT 4o (Persona Prompt with few shot)	28.89%	27.22%
18	GPT o4 mini (Persona Prompt)	27.78%	29.17%
19	GPT 4o mini (Persona Prompt)	22.22%	19.44%
20	GPT 4o mini (Persona Prompt with few shot)	20.83%	8.33%
21	GPT o1 (Persona Prompt)	20.83%	13.89%
22	GPT o1 (No Prompt)	20.83%	14.58%
23	GPT 4o (No Prompt)	19.94%	17.26%
24	GPT 4o mini (No Prompt)	4.17%	0.00%
	Correlation		**0.97**

- In the Diagram 9 below, for the abovementioned 8 models, we have computed the average score across all three prompting conditions to normalise model performance across varying input formats by both the expert panel and GPT_4o.

Diagram 9: Average score across all three prompting conditions by the expert panel and GPT 4o



- Furthermore, to substantiate the correlation between the models, in Table 15, we have plotted a few instances of scores for understanding and aligning the correlation better.

Table 15: Average score for sample LLM candidates for sample questions by the expert panel and GPT 4o

Model	Q. No	Overall Accuracy % [Human Evaluator]	Overall Accuracy % [GPT 4o]	Correlation at Model level
Gemini Flash2.5 (No Prompt)	1	33.33	50.00	0.84
	2	41.67	50.00	
	3	50.00	75.00	
	4	75.00	75.00	
	5	70.83	75.00	
Gemini 2.5pro (No Prompt)	1	20.83	29.17	0.87
	2	62.50	75.00	
	3	50.00	75.00	
	4	50.00	50.00	
	5	75.00	75.00	
GPT 4o (No Prompt)	1	16.67	25.00	0.92
	2	20.83	25.00	
	3	25.00	25.00	
	4	8.33	4.17	
	5	8.33	4.17	

4.6.4 Observations from the cross-validation:

- Consistency of Top Rankings: Despite absolute score differences, all three judges, GPT 4o, Gemini Flash, and human experts, displayed high convergence by consistently ranking GPT o3pro, GPT o3, Gemini 2.5pro, and Gemini Flash2.5 among the top performers and models, GPT o1, GPT 4o and GPT 4o mini at the bottom. This triangulates confidence in the relative performance of these models.

- This confirms that while LLM-as-a-Judge may show mild optimism bias, it maintains a strong signal in comparative performance evaluation.

- Leniency of LLM Judges: As previously noted in Section 4.3 and detailed in Table 6, GPT 4o awarded higher scores than human evaluators. Human evaluators provided more conservative scoring overall, particularly penalising superficial or overly fluent answers lacking legal precision. The divergence in high scores from LLM judges and low scores from the human experts can also be attributed to weakness in ability of the LLM judges to identify the hallucinations mainly in the judicial precedents included in the responses by the LLM candidates.

- While there is divergence in the scores provided by the judges to the responses from the LLM candidates, the relative ordering of models remains highly stable across evaluators, with a correlation of close to +1 in different perspectives listed above. This reinforces the validity of using LLM as a judge framework.

5. Conclusion

5.1 Impact

The evaluations as per the LE-BTL framework presents the first comprehensive, large-scale benchmarking of leading LLMs on Indian tax law reasoning. The results reveal a pronounced and consistent stratification in performance across models, underscoring both the potential and the current limitations of these systems in domain-specific legal tasks.

At the top of the performance spectrum, OpenAI's GPT o3pro and GPT o3 series, along with Google's Gemini 2.5 Pro, demonstrated superior legal reasoning abilities across all prompting strategies. These models not only excelled in foundational tasks such as issue spotting and rule identification but also showed relative stability in higher-order reasoning dimensions, including the application of law, interpretation, and justification.

A key insight emerging from this study is the dominance of proprietary models in higher performance tiers. All models achieving above 50% accuracy (Top-Tier) were proprietary, reflecting the role of scale, architecture, deeper training, and domain-aligned fine-tuning in driving strong legal reasoning capabilities. In contrast, the open-weight model, Deepseek V3, remained in lower tiers despite structured prompting, indicating a persistent performance gap between open-access and proprietary systems despite accessibility and openness.

Prompt engineering played a meaningful, though uneven, role in enhancing model performance. Persona-based prompting delivered consistent improvements across tiers. However, few-shot prompting produced mixed results, offering limited or even negative returns for top-tier models, likely due to context saturation or interference with latent reasoning processes and boosting some under-aligned models like Deepseek V3, which showed measurable improvement from few-shot examples, suggesting that structured guidance can compensate partially for architectural or alignment limitations.

Dimension-level insights from the IRAC+ framework further revealed that issue and rule identification are generally manageable for most LLMs, while application of law and justification remain the most challenging dimensions, turning into bottlenecks, demanding multi-step reasoning, nuanced legal interpretation, and contextual adaptability. These areas present opportunities for targeted model fine-tuning.

The cross-validation with human experts and the alternative LLM judge confirmed the reliability of GPT_4o as a scoring mechanism for relative performance. Although GPT 4o exhibited a mild optimism bias for Top-Tier LLM candidates compared to human evaluators, model rankings remained highly consistent across evaluators, reinforcing the viability of LLM-as-a-Judge for scalable benchmarking.

The LE-BTL framework establishes a rigorous benchmark for evaluating LLMs on Indian legal reasoning, offering stakeholders, whether developers or legal practitioners, valuable guidance in selecting models suited for downstream legal applications. The findings underscore that deep legal alignment, rather than just scale or brand, is the defining determinant of model utility in domain-specific reasoning tasks.

5.2 Proposed implementation guidelines based on our work

Based on the research and our observations from the results above, industry experts can consider drawing insights from the framework for guided decision making while determining the framework for AI applications within the Tax domain. Mainly, the industries can consider the following aspects:

(a) **Orchestrate models as per their competencies:** Even the strongest models have their strengths and weaknesses. It is the job of the experts to orchestrate the abilities of the models across functionalities to ensure the framework is built on the strengths of all LLMs.

(b) **Dependencies on the knowledge of LLMs:** Most LLMs lack the latest jurisdiction-specific amendments, case law, or revenue authority guidance yet they inherently have the abilities to generalize peculiar tax issues and understand the user intents. Hence, its business decision is to toggle between the use of inherent knowledge of the LLMs or to power them with retrieval tools. Experiments across domains consistently show performance lifts when models are given direct access to the controlling law, yet the ability to retrieve accurate datasets and ensure the dataset is updated on a real-time basis will be the real challenge.

(c) **Lack of clear legal reasoning in the training datasets:** Legal reasoning often requires multi-step logic, fact qualification, rule selection, exception handling, and, in tax, precise numeric calculation, which would not be generally available in public databases for training purposes. Hence, current LLMs can misinterpret conditions, mishandle cross-references, or make basic arithmetic errors that cascade into incorrect outcomes. Use deterministic engines (rule parsers, tax calculators) for computations and formal rule evaluation. Reserve LLMs for interpretation, candidate rule selection, and summarization.

(d) **Benchmarking must be domain and task-specific:** Develop your internal benchmark database for various tasks involved in your real-life workflow and test various LLMs, as general-purpose leaderboards do not reflect performance on specialised workflows. A model that ranks highly overall may still fail on niche tax provisions or statutory interpretation tasks or any functional aspect required. Fine-grained benchmarks, which separate skills such as statute recall, rule application, numeric fidelity, and consistency, produce a far more actionable capability profile. Track not only accuracy but also stability across multiple runs.

In summary, establishing robust frameworks for reasoning, computation, evaluation, and governance is essential for the integrity and reliability of complex decision-making systems. By integrating chain-of-thought prompting, leveraging deterministic computation engines, and adhering to rigorous evaluation standards, organizations ensure the factual soundness and reproducibility of outcomes. Overlaying these processes with comprehensive governance through human oversight, meticulous audit trails, and continual re-benchmarking fosters accountability and sustained trust. Separating deterministic from generative elements within architectures not only enhances transparency but also clarifies the provenance of each conclusion. Ultimately, this holistic approach not only mitigates risk but also upholds the highest standards of diligence and confidence in outputs that bear significant financial or legal consequences.

LIST OF ANNEXURES AND ATTACHMENTS

Annexure 1 – Evaluation Prompt

```
METRICS: Mapping[str, Dict[str, Any]] = {
 "IssueId": {
   "name": "Issue identification",
   "description": "Evaluates the model's ability to precisely and clearly identify the legal or factual controversy in the query. The issue must be framed as a concrete legal or factual question, often as a question of law, fact, or a mixed question of law and fact, avoiding generic or vague statements. Narrow, context-specific issue identification is essential to maintain legal relevance and logical flow. The models must identify the issues involved, any related hidden issues in the query.",
   "checklist": [
     "Each and every possible issue, even slightly possible issue, even indirectly possible issue, and even the issues not covered by ground truth answer, are covered by the response",
     "Issues are precise, unambiguous, explicit and question-shaped.",
     "Issues reflect the actual facts and legal context.",
     "There are no hallucinations (invented facts or sources) at all. "
   ]
 },


 "RuleId": {
   "name": "Rule identification ",
   "description": "This involves covering the statutory provisions of law being statutes, sections, rules, circulars, notifications, case laws and other statutory documents applicable to the query: the correct sections of the Income-tax Act, 1961 or the CGST Act (including explanations, provisos, definitions), relevant judicial precedents (ITAT, High Courts, Supreme Court), circulars / notifications, treaty provisions (DTAA) and delegated legislation such as the Income-tax Rules or GST Rules. It requires completeness, citation accuracy, and relevance to the issue without fabrication.",
   "checklist": [
     "Each and every possible relevant provision of law, even slightly relevant provision of law, even indirectly relevant provision of law, and even the provisions not covered by ground truth answer, are covered by the response ",
     "All the provisions of law covered by the response are correctly cited.",
     "No laws or judgements are misquoted or wrongly interpreted.",
     "There are no hallucinations (invented facts or sources) at all."
   ]
 },
 "ApplyLaw": {
   "name": "Application of law",
   "description": "This measures how the identified rules are applied to the specific facts, including coherent tax reasoning (deductive or inductive), correct interpretation of statutory language, fact-law alignment (e.g., whether deduction conditions are met), logical and fact-specific application of identified law, stepwise reasoning, connection to user facts, proper use or distinction of case law, discussion of conflicting views where they exist, demonstration of legal analysis depth and adherence to Indian tax-jurisprudence principles such as natural justice and substance-over-form. Responses must show understanding beyond copy-paste of law.",
   "checklist": [
     "Tax reasoning is fact-specific and avoids generic templating",
     "Logical structure is followed",
     "Analysis is coherent and free from contradictions ",
     "Laws are applied directly to user facts; not merely restated",
     "Facts are integrated effectively to test rule applicability",
     "Case laws are applied or distinguished appropriately with rationale",
     "Alternative interpretations or contrary judgments are considered where relevant",
     "Depth of analysis is detailed rather than superficial",
     "Depth of explanation is adequate with no surface-level paraphrasing",
     "Specific CBDT Circulars, case law, and GAAR guidelines are cited and utilized.",
     "Explicitly states all assumptions and caveats, even if obvious.",
     "No hallucination: no fabricated logic chains, judgments, or facts."
   ]
 },
 "Conclusion": {
   "name": "Conclusion correctness",
   "description": "Assesses whether the conclusion follows logically from applying law to facts. The conclusion should directly address the framed issue, be clear, concise, legally supportable, and, where multiple interpretations exist, acknowledge alternative views or probabilities. It must answer the exact legal issue raised, be unambiguous, and account for uncertainty (where applicable). It should avoid policy arguments unless constitutionally relevant and disclose any material assumptions. Legal plausibility and fidelity to precedent/statute are critical.",
     "checklist": [
     "All the issues are directly answers and all the issues are resolved without evasion",
     "Conclusion is consistent with cited law and reasoning ",
     "Conclusion is clear, unambiguous and without any contradiction",
      "Multiple-view areas are flagged by acknowledging possible uncertainties or multiple interpretations
and judgment is shown",

      "Discloses assumptions, caveats or limitations transparently",

      "Where statutory ambiguity exists, response must cite judicial principle or authoritative commentary
to support interpretation — mere statutory reading is not sufficient. ",

      "No hallucination: no invented outcomes, contradictory conclusions, or fabricated dependencies"

    ]

  },

  "Interpretation": {

    "name": "Interpretation / reasoning of language",

    "description": "Assesses the model's ability to parse amendments, interpret legal documents, and
reason about linguistic/legal nuance. Evaluates comprehension of circulars, judgments, agreements, tax
arithmetic, summarisation, advisory drafting, error spotting, and hallucination detection, based on
explicit user instructions.",

    "checklist": [

      "Correctly identifies relevant amendments and explains legal impact",

      "Parses language changes accurately and interprets their implications",

      "Determine the applicable circulars and legal texts correctly",

      "Interprets circulars and legal texts without misreading or omission",

      "Identifies and applies relevant precedents, not just lists them",

      "Performs arithmetic tax logic/calculations where needed",

      "Summarizes case law concisely and faithfully when required",

      "Identifies drafting, reasoning, or factual errors in long inputs if asked",

      "Drafts appropriate contextually sound advisory emails if instructed",

      "Detects hallucinated or logically inconsistent facts if user requests",

      "Surfaces tax red flags in contracts where asked (e.g., GAAR triggers, treaty misuse)",

      "Identifies unethical, misleading, or aggressive avoidance tactics and penalises them",

      "No hallucination in the interpretation "

    ]

  },

  "Justification": {

    "name": "Argumentation / justification",

    "description": "Measures how well the model constructs legally persuasive arguments such as grounds
of appeal, replies to tax notices, rectifications, and information requests. Arguments must be legally
valid, logically sound, and fact-sensitive, not mere reiteration of law.",

    "checklist": [
      "Arguments are logically constructed and follow IRAC or equivalent legal structure",
      "Each point is supported by statutory or judicial authority",
      "Rebuttals to tax authority contentions are mature and context-specific",
      "Identifies practical next steps and procedural options clearly",
      "Checklist of documentation requested is exhaustive and legally appropriate if asked",
      "All user-specified content requests (e.g., draft replies, info sheets) are completed accurately"
    ]
  }
}
RUBRIC_LEVELS = (
    "1 = Inadequate. The response is not completely aligned with the ground truth or does not cover 50% of the response as required by the checklist or not usable for client but can be provided to senior professional for review.\n"
    "2 = The response is mostly aligned with the ground truth but does not cover 60% of the response as required by the checklist or is usable as draft for client after making minor edits. Any minor non-expert-level compression will drop the score.\n"
    "3 = The response is aligned with the ground truth and covers 80% of the response as required by the checklist and is immediately usable as final for client and publication after many minor edits. To qualify for a score of 3, the response must address each sub-question or implied dimension of the user's query independently, and signal which provision applies to each. \n"
    "4 = Audit-Grade. The response adds value same as or more than the ground truth and covers same as or more than the response as required by the checklist and demonstrates advanced reasoning and is immediately usable for litigation in high courts. Everything in the response is exhaustive, cross-verifiable and legally placed with source noted. Even if the response feels structurally stronger than the ground truth, structural clarity alone is insufficient for a score of 4. Any minor stylistic vagueness, missing sub-issues, failure to disclose any assumptions (e.g., that no anti-abuse guidance contradicts position), absence of authoritative citations, failure to segment sub-issues, or lack of mention of assumptions/caveats even if obvious will drop the score. THIS SCORE IS RESERVED FOR AUDIT-GRADE ANSWERS ONLY. \n"
    "5 = Reserved for gold standard. For example, wherever applicable, the response has explicit discussion of all relevant counterarguments, caveats, judicial precedents including clarity on normative limits (e.g., whether CBDT circulars are binding, scope for challenge, or grey areas in interpretation). For example, wherever applicable, the response should have comparative analysis of at least two conflicting High-Court rulings plus SC tiebreaker, explicit GAAR/BEPS abuse analysis and policy impact and original statutory extract for each quoted section and rule (verbatim, within quotation marks). No response must be given this score under any circumstances WITHOUT EXPLICIT VERIFICATION OF IT BEING GOLD STANDARD AND it is an error you must avoid. \n"
)
PROMPT_TEMPLATE = """
You are an experienced senior Indian tax expert specialized in evaluating and reviewing tax answers provided on parameters.
```

You are evaluating a *candidate answer* against a *ground truth reference* on **one single metric** only using the rubric.

———————— Metric to grade ————————

Name: {metric_name}

What it measures: {metric_description}

Strict checklist for grading: {metric_checklist}

Scoring rubric (applied strictly):

{rubric_levels}

Qualitative inputs (free words):

{{Briefly capture qualitative feedback on the candidate answer}}

———————— Material to evaluate ——————

● Query / fact pattern

{query}

● Reference answer

{reference}

● Candidate answer

{candidate}

————————————————————————————

———————— Instructions ————————

1. Carefully compare the candidate answer to the ground truth and checklist.

2. Rigorously assess how many checklist points are fully satisfied and if the answer is precisely same as ground truth.

3. Assign one score from 1 to 5 using the rubric with a starting point of 1 and move up only if it is must as per the rubric.

4. **Output nothing except** a JSON block exactly like {{ "score": <integer 1-5> }}

"""

Annexure 2 – Prompt Settings

1. Zero shot in Plain language

   NA

2. Zero shot with persona

   You are a Tax expert experienced in analyzing Indian Direct Tax and Indirect Tax laws and Accounting Standards and Foreign Exchange Management Act and responding to user queries related to the same.

   You are considerate towards the latest amendments in the law and tend to answer the user query with context of the latest law available unless specifically mentioned by the user with respect to any particular tax period.

   You are well versed on the following:-

   1. Indian Income tax Act, 1961
   2. Indian Income Tax Rules, 1962
   3. Double Taxation Avoidance Agreements (DTAA) entered by India with other countries and its interplay with India tax laws
   4. Goods and Services Tax Act, 2017 and other related GST rules
   5. Other Indirect Tax laws in India, for example: Service Tax, VAT, Customs, Excise, etc.
   6. Exchange Control Regulations of India, commonly known as Foreign Exchange Management Act, 1999 (FEMA) and related guidelines, notifications, master directions, press releases etc.,
   7. Companies Act, 2013 and related regulations issued by the Ministry of Corporate Affairs of India
   8. Accounting Standards issued by the Institute of Chartered Accountants of India

   Your task is to:-
   - Understand, analyze, interpret and draw inferences from the user query, facts and context provided.
   - Draft a response which is technically correct and does not go out of the context provided to you.

   Your response will be judged on the following parameters by comparing with the golden answer while you need not give the answer in this format: -
   1. Issues identified
   2. Rules recalled (for all the issues identified)
   3. Application of the Rules (for all the issues identified)
   4. Correctness of the Conclusion (for all the issues identified)
   5. Interpretation (wherever applicable - determine the applicable amendments correctly; shows how wording changes alter legal effect (e.g., amendment parsing); determine the applicable circulars and other tax documents correctly; interprets the circulars and other tax documents correctly; determine the applicable precedents / case laws correctly; interprets the precedents / case laws correctly; undertakes arithmetic calculations correctly; summarizes the case laws correctly, where user requires the summary of the case laws; spots errors in the paragraphs or long contexts, where user requires the model to spot the errors; drafts appropriate advisory mails, where user requires the model to draft the advisory mails; spots hallucinations in the context shared by the user, where user requires the model to spot the hallucinations; spot red flags from tax optimisation perspective in the agreements shared by the user, where user requires the model to spot the red flags from tax optimisation perspective in the agreements)

6. Argumentation / Justification (wherever applicable – Arguments for grounds follow a logical structure (e.g., IRAC); Each ground is supported by authority or reasoning; Weaknesses are addressed for each contention of the tax authority and practical next steps offered; All points in the notices are addressed in the information checklist prepared by the model where user requires the information checklist; Comprehensive and correct information for each point in the notice is asked in the information checklist prepared by you where user requires the information checklist)

3. Few shot with persona

You are a Tax expert experienced in analyzing Indian Direct Tax and Indirect Tax laws and Accounting Standards and Foreign Exchange Management Act and responding to user queries related to the same.

You are considerate towards the latest amendments in the law and tend to answer the user query with context of the latest law available unless specifically mentioned by the user with respect to any particular tax period.

You are well versed on the following:-

1. Indian Income tax Act, 1961
2. Indian Income Tax Rules, 1962
3. Double Taxation Avoidance Agreements (DTAA) entered by India with other countries and its interplay with India tax laws
4. Goods and Services Tax Act, 2017 and other related GST rules
5. Other Indirect Tax laws in India, for example: Service Tax, VAT, Customs, Excise, etc.
6. Exchange Control Regulations of India, commonly known as Foreign Exchange Management Act, 1999 (FEMA) and related guidelines, notifications, master directions, press releases etc.,
7. Companies Act, 2013 and related regulations issued by the Ministry of Corporate Affairs of India
8. Accounting Standards issued by the Institute of Chartered Accountants of India

Your task is to:-
- Understand, analyze, interpret and draw inferences from the user query, facts and context provided.
- Draft a response which is technically correct and does not go out of the context provided to you.

Your response will be judged on the following parameters by comparing with the golden answer while you need not give the answer in this format: -
1. Issues identified
2. Rules recalled (for all the issues identified)
3. Application of the Rules (for all the issues identified)
4. Correctness of the Conclusion (for all the issues identified)
5. Interpretation (wherever applicable - determine the applicable amendments correctly; shows how wording changes alter legal effect (e.g., amendment parsing); determine the applicable circulars and other tax documents correctly; interprets the circulars and other tax documents correctly; determine the applicable precedents / case laws correctly; interprets the precedents / case laws correctly; undertakes arithmetic calculations correctly; summarizes the case laws correctly, where user requires the summary of the case laws; spots errors in the paragraphs or long contexts, where user requires the model to spot the errors; drafts appropriate advisory mails, where user requires the model to draft the advisory mails; spots hallucinations in the context shared by the user, where user requires the model to spot the hallucinations; spot red flags from tax optimisation perspective in the agreements shared by the user, where user requires the model to spot the red flags from tax optimisation perspective in the agreements)

6. Argumentation / Justification (wherever applicable – Arguments for grounds follow a logical structure (e.g., IRAC); Each ground is supported by authority or reasoning; Weaknesses are addressed for each contention of the tax authority and practical next steps offered; All points in the notices are addressed in the information checklist prepared by the model where user requires the information checklist; Comprehensive and correct information for each point in the notice is asked in the information checklist prepared by you where user requires the information checklist)

Provide below is one sample example from the golden dataset for your reference:

Query no. 1 (Input): How is the deduction for bonus payments allowed under the Income-tax Act, 1961? Is it allowed on a due basis or only on a payment basis?
Chain of thoughts:
(a) Find the relevant provisions for the deduction of bonus: Section 36(1)(ii),
(b) Find the relevant provisions for the deduction allowed on the actual payment basis: Section 43B
Expected Response:
The deduction for bonuses paid or payable to employees depends on several factors, including the identity of the recipient, whether the expense is incurred wholly and exclusively for business purposes, and the timing of the payment.
1. Bonus deduction under Section 36(1)(ii)
Section 36 of the Income-tax Act, 1961 lists specific revenue expenditures that are allowable as deductions while computing income under the head Profits and Gains of Business or Profession (PGBP), subject to prescribed conditions.
As per Section 36(1)(ii), any bonus or commission paid to an employee is allowable as a deduction if the following conditions are satisfied:

- The payment is made to an employee.
- The amount would not have been payable to the employee as profit or dividend if it had not been paid as bonus or commission.

2. Bonus or commission deductions allowed only on payment basis
Section 43B allowed certain expenditures as deduction only on actual payment even if assessee follows the mercantile system of accounting. These expenditure are allowed as deduction in the year of accrual if payment thereof is made either in the previous year itself or in subsequent year on or before the due date specified for filing of return of income.
Where bonus or commission paid or payable to the employees is deductible in accordance with Section 36(1)(ii), it shall be allowed as a deduction only on payment basis. However, if the payment is made on or before the due date for filing the return of income under Section 139(1), the deduction shall be allowed in the previous year in which the liability was incurred.
3. Conclusion
Deduction for bonus paid to employees is governed by both Section 36(1)(ii) and Section 43B. While Section 36(1)(ii) defines the eligibility criteria for claiming the deduction, Section 43B governs the timing of the deduction. Thus, bonus is not deductible merely on accrual basis. It is allowed as a deduction only in the year of actual payment. However, if the bonus is paid on or before the due date for filing the return of income under Section 139(1), the deduction can be claimed in the year in which the bonus became due.

References:

[1] LegalBench: A Collaboratively Built Benchmark For Measuring Legal Reasoning In Large Language Models (https://arxiv.org/abs/230 8.11462)

[2] **Hendrycks, D. et al. (2021).** *Measuring Massive Multitask Language Understanding*. arXiv preprint. https://arxiv.org/abs/2009.03300

[3], **I. et al. (2019).** *Release Strategies and the Social Impacts of Language Models*. arXiv preprint. https://arxiv.org/abs/1908.09203

[4]**Bommarito II, M. J., & Katz, D. M. (2022).** *GPT Takes the Bar Exam*. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4314839

[5] **Zellers, R. et al. (2019).** *HellaSwag: Can a Machine Really Finish Your Sentence?* [1905.078430]

[6] **Li, S. et al. (2025).** *CodeContests+: High-Quality Test Case Generation for Competitive Programming* [2506.05817] CodeContests+: High-Quality Test Case Generation for Competitive Programming.

[7] **Zheng, Y. et al. (2023).** *CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models*. [2302.00288] CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models

[8] **Hendrycks, D. et al. (2021).** *Measuring Massive Multitask Language Understanding* [2009.03300] Measuring Massive Multitask Language Understanding

[9]**Zhong, Z. et al. (2020).** *LegalBench: A Benchmark for Legal Reasoning in Large Language Models*. *arXiv preprint*. https://arxiv.org/abs/2308.11462

[10]**Surden, H. (2023).** *Artificial Intelligence and Law: An Overview. Harvard Journal of Law & Technology.* "Artificial Intelligence and Law: An Overview" by Harry Surden

[11]**NITI Aayog (2018).** *National Strategy for Artificial Intelligence – #AIforAll.* National Strategy for Artificial Intelligence

[12] *CBDT Central Action Plan for FY26 targets reducing pendency at CIT (Appeals) level* BusinessToday

[13] *Income tax appeal resolution typical time 15–20 years: Steps to cut appeal pendency urged in Budget 2025* The Economic Times

[14] **Srivastava, A., et al. (2022).** *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (BIG-bench)* https://arxiv.org/abs/2206.04615

[15]**The Hindu (2023).** *Punjab & Haryana High Court uses ChatGPT to decide bail plea*. *March 27, 2023.* Link

[16]**LiveLaw (2023).** *Punjab & Haryana High Court uses ChatGPT to understand How 'Differential GPS' Helps In Locating Disputed Property bail plea*. March 27, 2023. Link

[17]**Bar & Bench (2023).** *Supreme Court Launches AI Transcription System*. Link

[18]**Bar & Bench (2023).** *CJI Launches AI Tool for Legal Research*. Link

[19] *LegalBench: A Collaboratively Built Benchmark For Measuring Legal Reasoning In Large Language Models* (https://arxiv.org/abs/2308.11462)

[20]**Rahul Hemrajani (2025).** *Evaluating the Role of Large Language Models in Legal Practice in India. - Evaluating the Role of Large Language Models in Legal Practice in India*
https://arxiv.org/pdf/2508.09713

[21]**Jerrin B. Mathew, Sannah Mudbidri, Banisethi Aashrita, Sanika Atul Tapre.** *The Disadvantages and Limitations of Using Large Language Models in the Field of Law - The Disadvantages and Limitations of Using Large Language Models in the Field of Law - National Law School of India University*
https://www.nls.ac.in/research/projects/the-disadvantages-and-limitations-of-using-large-language-models-in-the-field-of-law/

[22]**OpenAI (2023).** *GPT-4 System Card – Risks in Legal Domain Use Cases*.
https://cdn.openai.com/papers/gpt-4-system-card.pdf

[23]**Kang, X., Qu, L., Soon, L.-K., et al. (2024).** *Bridging Law and Data: Augmenting Reasoning via a Semi-Structured Dataset with IRAC Methodology*.
https://arxiv.org/abs/2406.13217

[24]**Guha, N., Nyarko, J., Ho, D. E., et al. (2023).** *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*. NeurIPS 2023.
https://arxiv.org/abs/2308.11462

[25]*What is The IRAC Method - Everything You Need to Know* \
https://www.iracmethod.com/irac-methodology

[26]**Andong Hua1, Kenan Tang1, Chenhe Gu2, Jindong Gu3, Eric Wong4, Yao Qin (2025).** *Flaw or Artifact? Rethinking Prompt Sensitivity in Evaluating LLMs - Flaw or Artifact? Rethinking Prompt Sensitivity in Evaluating LLMs*
https://arxiv.org/pdf/2509.01790

[27]**Lianmin Zheng (2023)** - *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena -*
https://arxiv.org/abs/2306.05685

[28] **Zailong Tian (2025).** *Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution -*
https://arxiv.org/abs/2508.06225

[29] **Junlong Li (2024)** - *Dissecting Human and LLM Preferences -*
https://arxiv.org/abs/2402.11296