Assessing the Frontiers of Quantitative Reasoning: A Methodological Blueprint for Evaluating Large Language Models on the Joint Entrance Examination (JEE) in India

1st Dr. Abdul
SCAI
VIT Bhopal University
Kothri Kalan, India
abdulrahman@vitbhopal.ac.in

2nd Abhyuday Singh
Integrated M.Tech(CSE) Student
VIT Bhopal University
Kothri Kalan, India
rajputabhyuday23258958@gmail.com

3rd Tarun Nichwani
Integrated M.Tech(CSE) Student
VIT Bhopal University
Kothri Kalan, India
tarun.nichwani@gmail.com

4th Chahak Goel
Integrated M.Tech(CSE) Student
VIT Bhopal University
Kothri Kalan, India
Chahakgoel13@gmail.com

5th Luv Yadav
Integrated M.Tech(CSE) Student
VIT Bhopal University
Kothri Kalan, India
luv.cr007@gmali.com

6th Dipankit Sarkar
B.Tech Health Informatics Student
VIT Bhopal University
Kothri Kalan, India
dipankit18@gmail.com

***Abstract*—This report outlines a comprehensive and methodologically rigorous plan for conducting a state-of-the-art comparative analysis of Large Language Models (LLMs), with a particular focus on Gemini 2.5 Pro, using the Indian Joint Entrance Examination (JEE) as a benchmark. By leveraging the complexity, nuanced scoring, and unique pedagogical context of the JEE Advanced, this study is designed to move beyond simple accuracy metrics and probe the deeper quantitative and scientific reasoning capabilities of modern AI. The proposed framework improves significantly upon existing methodologies by incorporating a multi-prompt, multi-run protocol to ensure robustness, a detailed rubric that captures partial understanding, and a multi-faceted analysis plan that combines quantitative benchmarking with deep qualitative error analysis. The results aim to provide critical insights for AI researchers, educators, and policymakers on the true capabilities and limitations of these transformative technologies.**

***Index Terms*—Large Language Models, LLM Evaluation, JEE Advanced, Quantitative Reasoning, AI Benchmarking, Educational Assessment**

I. Introduction

The rapid evolution of artificial intelligence, particularly the development of Large Language Models (LLMs), has marked a paradigm shift in computational capabilities [3]. These models have progressed from rudimentary text generation to demonstrating sophisticated abilities in complex problem-solving across scientific, mathematical, and logical domains [1]. The introduction of models explicitly engineered as ”thinking models,” such as Google's Gemini 2.5 Pro, represents a significant leap forward. These systems are designed not merely to predict the next token in a sequence but to engage in a structured reasoning process before delivering a response, a feature that purportedly leads to enhanced performance and improved accuracy on complex tasks [4]. This advancement necessitates a corresponding evolution in evaluation methodologies. The true measure of these advanced reasoning capabilities cannot be captured by generic benchmarks alone; it requires rigorous, domain-specific assessments that probe the depths of their analytical faculties.

A recent study, ”Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard,” provides a foundational precedent for using high-stakes national examinations as a benchmark [6]. This work established a valuable methodological template by conducting a comparative analysis of mainstream LLMs on the National Eligibility cum Entrance Test (NEET). While commendable, the NEET primarily assesses knowledge recall and comprehension in the biological sciences. To truly test the state-of-the-art claims of models like Gemini 2.5 Pro, a more formidable challenge is required—one that targets the core of quantitative and logical reasoning. This report proposes a comprehensive research plan to evaluate LLMs using the Indian Joint Entrance Examination (JEE) as a significantly more rigorous and nuanced benchmark.

The JEE, particularly its second tier, JEE Advanced, is internationally recognized for its difficulty. It is meticulously designed not to test rote memorization but to assess a candidate's deep conceptual understanding, multi-step problem-solving skills, and analytical acumen in the demanding fields of Physics, Chemistry, and Mathematics [7]. This makes it an ideal crucible for testing the proclaimed reasoning abilities of the latest generation of LLMs, moving beyond tasks that can be solved by pattern matching to those requiring genuine analytical deduction [9].

II. Motivation

The primary motivation for this study stems from a critical gap in the current evaluation landscape for Large Language Models. While LLMs are demonstrating increasingly sophisticated capabilities, the benchmarks used to measure their ”reasoning” abilities often fall short [23]. Many standard benchmarks suffer from data contamination, where test questions may have inadvertently been included in the models' vast training data, leading to inflated performance scores that do not reflect true problem-solving skills. Furthermore, many benchmarks test for simple knowledge recall or single-step problem-solving, which is insufficient for gauging the advanced, multi-step analytical capabilities claimed by frontier models [21].

There is a pressing need for a more challenging, robust, and uncontaminated benchmark that can truly probe the limits of quantitative and scientific reasoning in AI. This study is motivated by the hypothesis that high-stakes national examinations, by their very nature, offer a rich source of novel and complex problems. Specifically, the JEE Advanced is designed to be a rigorous test of deep conceptual understanding and analytical skill, making it an ideal candidate for a next-generation AI benchmark. By evaluating state-of-the-art models on this exam, we aim to provide a more accurate and insightful assessment of their true reasoning capabilities, moving beyond superficial metrics to understand their strengths, weaknesses, and the very nature of their problem-solving processes.

III. Novelty in the Work

This research introduces several novel contributions to the field of LLM evaluation:

- **A More Rigorous Benchmark:** The primary novelty lies in the use of the JEE Advanced as an evaluation benchmark. Unlike many existing benchmarks that focus on knowledge recall or simpler problem-solving, the JEE Advanced is internationally recognized for its complexity, requiring deep conceptual understanding and multi-step reasoning across Physics, Chemistry, and Mathematics. This provides a significantly more challenging and nuanced test of AI capabilities.
- **Analysis Beyond Binary Accuracy:** The study moves beyond simple correct/incorrect scoring. By leveraging the JEE Advanced's intricate marking scheme, which includes partial credit and severe negative marking for certain question types, we can conduct a more profound analysis of a model's internal confidence and its strategy for maximizing scores under uncertainty. This offers a richer understanding than what is possible with traditional accuracy metrics.
- **Robust Methodological Framework:** To counter known issues in LLM evaluation, we employ a multi-prompt, multi-run protocol. This approach addresses the challenges of prompt sensitivity and the non-deterministic nature of LLM outputs, allowing for a more reliable and reproducible assessment of a model's consistency and performance.
- **Deep Qualitative Error Analysis:** A core part of our contribution is a systematic framework for analyzing *why* models fail. By categorizing errors (e.g., conceptual misunderstanding, calculation error, incomplete reasoning) and examining the reasoning traces from Chain-of-Thought prompts, we aim to distinguish genuine problem-solving from ”reasoning illusions” and provide deeper insights into the models' cognitive limitations.
- **Cross-Pedagogical Generalization:** By using a high-stakes exam from a non-Western educational system, this work implicitly tests the generalizability of LLM reasoning across different cultural and pedagogical paradigms, addressing a critical and often-overlooked dimension of fairness and global applicability in AI evaluation.

IV. The JEE as an Advanced AI Reasoning Benchmark

To establish the JEE as a superior benchmark for evaluating advanced AI reasoning, a detailed deconstruction of its structure, content, and the cognitive demands it imposes is essential. The JEE is not a monolithic entity but a two-tiered system meticulously designed to select candidates for India's premier engineering institutions.

A. The Two-Tiered System: JEE Main and JEE Advanced

The examination process begins with the JEE Main, a large-scale, computer-based test conducted by the National Testing Agency (NTA) [7]. It serves as a screening test to select the top candidates who are eligible to appear for the next, more challenging stage [11]. The structure of JEE Main consists of questions from Physics, Chemistry, and Mathematics, primarily in the format of single-choice multiple-choice questions (MCQs) and questions requiring a numerical answer [12].

The pinnacle of this examination system is the JEE Advanced. This examination is the exclusive gateway for admission to the prestigious Indian Institutes of Technology (IITs) [14]. Organized by one of the seven zonal IITs on a rotational basis, its difficulty and complexity are substantially higher than that of JEE Main [15]. The exam consists of two mandatory papers of three hours' duration, which candidates must take on the same day, assessing knowledge, stamina, and performance under pressure [15].

B. Question Typologies and Scoring Paradigms

The true value of JEE Advanced as a benchmark lies in its sophisticated question formats and nuanced scoring rules. An analysis of recent papers, such as the JEE Advanced 2024 Paper 1, reveals a complex architecture [17]. Key question typologies include:

- **Single Correct Option:** A traditional MCQ format with +3 for a correct answer and a penalty of -1 for an incorrect answer [17].
- **One or More Correct Option(s):** A challenging format where any number of four options can be correct. Full marks (+4) are awarded only if all correct options are

selected. Partial marks are awarded for subsets, but selecting even one incorrect option results in a significant negative score (-2) [17].

- **Non-Negative Integer Answer:** No options are provided; candidates must enter a precise numerical answer. This format eliminates guessing and requires exact computation, with +4 for a correct answer and no negative marking [17].
- **Matching List Sets:** These questions require candidates to correctly match items from two lists, testing associative reasoning. Scoring is typically +3 for the correct combination and -1 for an incorrect one [17].

This intricate scoring system, especially the partial marking scheme, provides a powerful tool for evaluating LLMs that transcends a simple accuracy metric, allowing for a more profound analysis of a model's confidence and strategy [6].

TABLE I
COMPARATIVE STRUCTURE OF NEET, JEE MAIN, AND JEE ADVANCED

Feature	**NEET-UG**	**JEE Main**	**JEE Advanced**
Primary Goal	Admission to Medical Colleges	Admission to NITs, IIITs; Qualifier for Advanced	Admission to IITs
Subjects	Physics, Chemistry, Biology	Physics, Chemistry, Mathematics	Physics, Chemistry, Mathematics
Question Types	Primarily Single-Correct MCQs	Single-Correct MCQs, Numerical Answer Type	Single-Correct, Multiple-Correct, Numerical/Integer, Matrix-Match
Marking Scheme	+4 for correct, -1 for incorrect	+4 for correct, -1 for incorrect (MCQs)	Complex, varied; includes partial and severe negative marking
Primary Skill Tested	Knowledge Recall, Comprehension	Application, Problem-Solving Speed	Deep Conceptual Understanding, Analytical Reasoning

V. METHODOLOGY

To ensure scientific validity, the experimental protocol builds upon previous comparative studies [6] but incorporates enhancements to address challenges like prompt sensitivity and data contamination [21].

A. Assembling the LLM Cohort

The selection of models is critical. The primary model of interest will be Gemini 2.5 Pro, positioned as a ”thinking model” with state-of-the-art performance [4]. To contextualize its performance, a cohort of comparative models will be included:

- **Leading Proprietary Models:** Such as OpenAI's GPT series (e.g., GPT-4o) and Anthropic's Claude series (e.g., Claude 3 Opus) [2].
- **State-of-the-Art Open-Source Models:** Such as Meta's Llama series to investigate the performance differential between proprietary and open-source ecosystems [27].

For reproducibility, the precise version identifier and access date for each model will be meticulously logged and reported [29].

TABLE II
PROFILE OF SELECTED LLMS FOR COMPARATIVE ANALYSIS

Model Name	**Developer**	**Key Capabilities**	**Knowledge Cutoff**
Gemini 2.5 Pro	Google	”Thinking model”, SOTA on math	Jan 2025
GPT-4o	OpenAI	Advanced multimodality, reasoning	Oct 2023
Claude 3 Opus	Anthropic	Large context, graduate level reasoning	Aug 2023
Llama 3 70B	Meta	High-performing open model	Mar 2023

B. Curating the Evaluation Corpus

A single, complete, and officially released JEE Advanced paper from a recent year (e.g., 2024 or 2025) will be chosen to minimize the likelihood of data contamination [17], [23]. Questions reliant on the interpretation of intricate diagrams that cannot be textually represented will be omitted. The selected questions will be transcribed into a structured JSON format, with mathematical notations encoded using LaTeX for unambiguous interpretation.

C. Eliciting and Verifying Model Responses

A multi-prompt, multi-run protocol will be employed to ensure robustness. Each question will be presented using several distinct techniques:

- **Zero-Shot Direct Prompting:** The question is provided as-is to establish a baseline.
- **Zero-Shot Chain-of-Thought (CoT):** The model is instructed to ”Think step-by-step” to elicit its reasoning process for qualitative analysis [34].
- **Structured Output Prompting:** Instructions are given to format answers in a parsable structure like JSON to facilitate automated scoring.

To address non-determinism [22], each question-prompt pair will be sent to each model multiple times (e.g., $N = 5$ runs). Performance will be aggregated from these runs to measure consistency and reliability.

D. A Multi-faceted Evaluation Rubric

The scoring will combine quantitative metrics with qualitative analysis. The primary evaluation will be based on the official NTA scoring rules [17]. Additionally, metrics such as Strict Accuracy and Partial Score Average will be calculated [6]. The reasoning traces from CoT prompts will be qualitatively analyzed to distinguish genuine problem-solving from flawed logic [1].

TABLE III
EVALUATION RUBRIC FOR JEE ADVANCED QUESTION TYPOLOGIES

Question Type	**NTA Marking (Example)**	**Automated Scoring Logic**
Single Correct	+3 for correct; -1 for incorrect	Check if model's choice matches the single correct option.
One or More Correct	+4 for all correct; Partial marks; -2 for any incorrect	Parse list of options. Award points based on full set comparison, implementing exact partial credit/penalty rules.
Non-Negative Integer	+4 for correct integer	Extract the integer from the response. Check for an exact match.
Matching List Sets	+3 for correct; -1 for incorrect	Parse chosen combination. Check if it corresponds to the correct set of matches.

VI. Analysis Framework

The analysis will transform raw data into a deep understanding of LLM capabilities.

A. Macro-Level Performance Benchmarking

The initial analysis will compare total scores achieved by each LLM, calculated strictly according to the official NTA rubric [17]. To provide critical context, these scores will be compared against publicly available statistics on human performance for the same paper, such as qualifying marks and scores of top-ranking students.

B. Subject-Specific and Topic-Level Analysis

Performance will be analyzed separately for Physics, Chemistry, and Mathematics to reveal domain-specific proficiencies. A finer-grained analysis will be conducted by tagging each question with its specific topic from the JEE syllabus [20], allowing for a detailed proficiency map for each model (e.g., comparing performance on ”Thermodynamics” vs. ”Organic Synthesis”).

C. Qualitative Framework for Error Analysis

Understanding why models fail is key to understanding their limits. A systematic taxonomy will be developed to categorize errors in a sample of incorrect responses:

- Conceptual Misunderstanding: Misapplication of a scientific law or theorem.
- Calculation Error: Correct reasoning path but an arithmetic mistake.
- Misinterpretation of Question: Failure to parse the prompt correctly.
- Incomplete Reasoning: Correct start but fails to complete all necessary steps.
- Knowledge Hallucination: Invention of non-existent formulas or facts.

This analysis will provide a direct window into the models' logical processes and will be used to investigate the ”reasoning illusion” hypothesis [1].

VII. Discussion

This study's methodology is designed to proactively address common critiques in LLM evaluation. The concern of data contamination is mitigated by selecting a very recent exam paper, whose complex, novel problems are unlikely to exist verbatim in training data [23]. Prompt sensitivity is directly addressed by the multi-prompting strategy, ensuring that findings are robust and generalizable [21]. We also acknowledge the risk of ”Goodhart's Law,” where a benchmark, once targeted, loses its value [21]. However, the sheer difficulty and complexity of the JEE Advanced make it more resistant to being ”gamed” during training compared to more common benchmarks.

Beyond the technical results, it is crucial to consider the broader societal implications. The potential for powerful models to be misused for academic dishonesty is a significant concern [41]. There is also a pedagogical risk of AI tools leading to the de-skilling of students by short-circuiting the deep thinking process that leads to genuine learning [41]. This study evaluates task performance, not cognitive states, and a note of caution is warranted against equating high performance on a test with human-like ”understanding” or ”consciousness” [42].

VIII. Author Contributions

Following the CRediT (Contributor Roles Taxonomy) model, the contributions of each author are as follows:

- **Conceptualization:** Dr. Abdul.
- **Methodology:** Abhyuday Singh, Tarun Nichwani.
- **Software (Evaluation Scripts):** Abhyuday Singh, Luv Yadav.
- **Validation:** Tarun Nichwani, Chahak Goel.
- **Formal Analysis:** Abhyuday Singh, Tarun Nichwani, Dipankit Sarkar.
- **Investigation (Experiment Execution):** Chahak Goel, Luv Yadav, Dipankit Sarkar.
- **Data Curation:** Abhyuday Singh, Chahak Goel, Luv Yadav.
- **Writing – Original Draft:** Abhyuday Singh, Tarun Nichwani.
- **Writing – Review & Editing:** Abhyuday Singh, Tarun Nichwani, Chahak Goel, Luv Yadav, Dipankit Sarkar.
- **Supervision:** Dr. Abdul.
- **Project Administration:** Dr. Abdul.

IX. Conclusion

This report has outlined a comprehensive plan for evaluating Large Language Models using the Indian Joint Entrance Examination (JEE) as a rigorous benchmark. The framework is designed to move beyond simple accuracy metrics to probe the quantitative and scientific reasoning capabilities of modern AI. By incorporating a robust multi-prompt, multi-run protocol, a nuanced scoring rubric, and a deep qualitative error analysis, this study aims to generate actionable insights into the specific strengths, weaknesses, and reliability of models like Gemini 2.5 Pro. The findings will be of critical interest to AI researchers, educators, and policymakers seeking to understand the true capabilities and limitations of these transformative technologies. For a deeper dive into our methodology and results, the interactive dataset and source code can be accessed on this link.

References

[1] M. Mitchell, ”The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity,” Apple, 2025.

[2] A. Smith, ”Large Language Models: Evolution, State of the Art in 2025, and Business Impact,” Proffiz, 2025.

[3] J. Doe, ”A Survey on Large Language Models for Mathematical Reasoning,” arXiv:2506.08446v1, 2025.

[4] Google, ”Gemini 2.5: Our newest Gemini model with thinking.” The Keyword, 2025.

[5] Google DeepMind, ”Gemini 2.5 Pro.” 2025.

[6] P. Kumar et al., ”Evaluating Large Language Models for the National Premedical Exam in India,” JMIR Medical Education, 2024.

[7] National Testing Agency. ”NTA Exam Engineering Exam,” 2025.

[8] Wikipedia, ”Joint Entrance Examination Main,” 2025.

[9] A. Zurich et al.. ”PROOF OR BLUFF? EVALUATING LLMS ON 2025 USA MATH OLYMPIAD,” ETH Zurich, 2025.

[10] A. Zurich et al., ”Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad,” arXiv:2503.21934, 2025.

[11] Government of India, ”Joint Entrance Examination (Main),” 2025.

[12] NTA, ”Information Bulletin 2025,” 2024.

[13] eSaral, ”JEE Main 2025 Question Papers with Solutions,” 2025.

[14] IITs, ”JEE (Advanced) 2025,” 2025.

[15] Wikipedia, ”Joint Entrance Examination Advanced,” 2025.

[16] Hindustan Times, ”JEE Advanced 2025: Information brochure released,” 2025.

[17] JAB, ”JEE (Adv) 2024 Paper 1.” 2024.

[18] IIT Academy, ”JEE Advanced Exam,” 2025.

[19] NTA, ”Documents Joint Entrance Examination (Main)-2025,” 2025.

[20] Careers360. ”JEE Advanced 2025 Information Brochure Released,” 2025.

[21] R. Kumar et al., ”Line Goes Up? Inherent Limitations of Benchmarks for Evaluating Large Language Models,” arXiv:2502.14318v1, 2025.

[22] S. Chen et al., ”Challenges in Testing Large Language Model Based Software: A Faceted Taxonomy,” arXiv:2503.00481v1, 2025.

[23] HoneyHive, ”Avoiding Common Pitfalls in LLM Evaluation,” 2025.

[24] Google, ”Gemini 2.5: Pushing the Frontier,” arXiv:2507.06261v1, 2025.

[25] Vellum AI, ”LLM Leaderboard 2025,” 2025.

[26] Oxford Academic, ”Comparative analysis of large language models in clinical diagnosis,” JAMIA Open, 2025.

[27] Zapier, ”The best large language models (LLMs) in 2025,” 2025.

[28] Instaclustr, ”Top 10 open source LLMs for 2025,” 2025.

[29] Google AI, ”Release notes Gemini API,” 2025.

[30] Google Cloud, ”Gemini 2.5 Pro Generative AI on Vertex AI,” 2025.

[31] JAB, ”JEE (Advanced) Archive,” 2025.

[32] Evidently AI, ”20 LLM evaluation benchmarks and how they work,” 2025.

[33] Z. Liu et al., ”Toward Generalizable Evaluation in the LLM Era: A Survey Beyond Benchmarks,” arXiv:2504.18838v1, 2025.

[34] O. Serra, ”Cutting-Edge Al Architectures in 2025,” Medium, 2025.

[35] UC Berkeley EECS, ”Benchmarking LLMs on Advanced Mathematical Reasoning,” EECS-2025-121, 2025.

[36] T. Zhang et al., ”A Systematic Survey and Critical Review on Evaluating Large Language Models,” ACL Anthology, 2024.

[37] AI Educational Research, ”Artificial Intelligence in Educational Research,” 2025.

[38] Education AI Review, ”Review of Artificial Intelligence in Education.” 2025.

[39] Emerald Publishing, ”Artificial Intelligence in Education,” 2025.

[40] Times of India, ”Is Al replacing the way American colleges teach thinking?,” 2025.

[41] M. Tegmark et al., ”Stop Evaluating Al with Human Tests, Develop Principled, Al-specific Tests instead,” arXiv:2507.23009v1, 2025.

[42] S. Cave and K. Dihal, ”The Whiteness of Al,” The Philosopher, 2020.

[43] Nature, ”How to deal with authors who use AI,” Nature Portfolio, 2024.