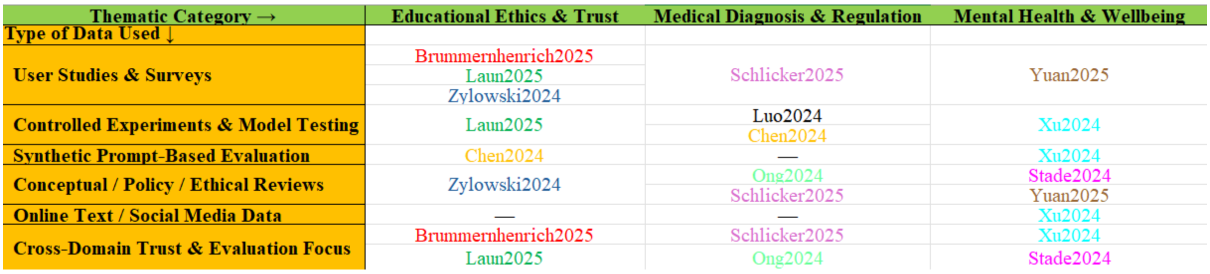
Reviewing Previous research for Evaluating the Decision Making of LLMs for Student Support Scenarios

Amena Inamdar, 20639678

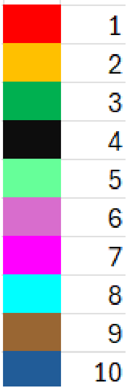
1. Introduction and Motivation:

LLMs have evolved rapidly and are increasingly applied in healthcare, educational, and mental wellbeing decision-support systems. A highly underexploited yet highly applicable application is "student support decision-making", supporting students with academic, emotional, or health-related choices such as accessing wellbeing resources, getting to a GP, applying for an EC extension, or seeking a support plan. These kinds of decisions are time-critical, emotionally nuanced, and highly personal, so it is critical to analyze the precision and appropriateness of advice produced by LLM. Because this is a nascent, cross-disciplinary issue, there is not a body of literature discussing it. To bridge this gap, this review cross-leverages three areas of overlap: (a) LLM decision-making and trusting in educational and moral contexts, (b) LLM use for general health decision-making, and (c) LLM use for mental health and wellbeing support. All these areas supply context on the way LLMs tackle delicate tasks and on how users receive their recommendations. For example, LLMs outperform students but are less reliable (paper 3); medical trust relies on human reasoning congruence and risk communication (papers 2,6); and in mental health, LLMs show empathetic potential but raise regulatory and ethical concerns (papers 7,8,9). Overall, these threads discuss how LLMs may be able to perform in tasks involving students and pose significant issues regarding their appropriate use.



Thematic Category →	Educational Ethics & Trust	Medical Diagnosis & Regulation	Mental Health & Wellbeing
Type of Data Used ↓			
User Studies & Surveys	Brummernhenrich2025 Laun2025 Zylowski2024	Schlicker2025	Yuan2025
Controlled Experiments & Model Testing	Laun2025	Luo2024 Chen2024	Xu2024
Synthetic Prompt-Based Evaluation	Chen2024	—	Xu2024
Conceptual / Policy / Ethical Reviews	Zylowski2024	Ong2024 Schlicker2025	Stade2024 Yuan2025
Online Text / Social Media Data	—	—	Xu2024
Cross-Domain Trust & Evaluation Focus	Brummernhenrich2025 Laun2025	Schlicker2025 Ong2024	Xu2024 Stade2024

Img 1: Classification of research papers under three main focus



🟥	1
🟧	2
🟩	3
⬛	4
🟩	5
🟪	6
🟪	7
🟦	8
🟫	9
🟦	10

Img 2: paper Reference for Img 1

1.1. Research Field Challenges

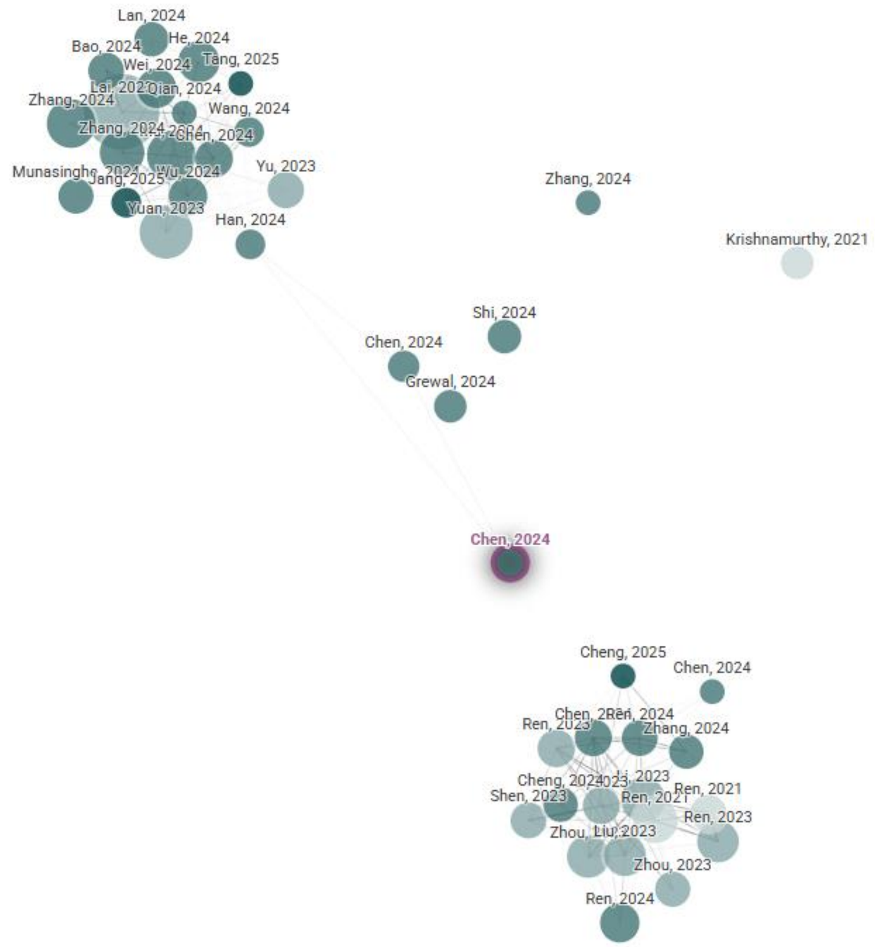
1. **User Trust Mismatch** – Students may under- or over-trust LLM advice due to source bias or algorithm aversion, despite high output quality, complicating use in sensitive scenarios. *(Papers 3, 6)*
2. **Reasoning Misalignment** – LLMs can produce correct answers with flawed logic, making their decision process opaque and hard to trust in support contexts. *(Paper 2)*
3. **Bias and Fairness Risks** – LLMs trained on biased data may generate unfair or inappropriate suggestions, especially for vulnerable student groups. *(Papers 5, 8)*
4. **Lack of Contextual Metrics** – Standard metrics miss emotional, ethical, or real-world relevance; few student-specific evaluation frameworks exist. *(Papers 7, 9)*

1.2. Survey Scope

With an emphasis on situations where students seek advice on matters like mental health, physical health, and academic accommodations (e.g., EC extensions or assistance plans), this survey investigates how Large Language Models (LLMs) make choices in student support scenarios. Because this particular topic is new, the survey incorporates information from three related fields: the use of LLM in diagnosing general health issues, their application in mental health support systems, and LLM trust and decision-making in educational ethics. It addresses topics including bias and fairness, decision-making transparency, trust calibration, and the moral application of LLMs in high-stakes, student-facing settings.

1.3. Search Methodology

To gather relevant literature for this survey, I initially explored research on Large Language Models (LLMs) in healthcare decision-making. I contacted Professor Simon Kent in the last week of March 2025, and we held a meeting in May 2025, he then recommended refining the topic to focus on *LLM decision-making for student support scenarios*. He suggested starting by exploring how LLMs make decisions in sensitive areas so I proceeded with education, mental health, and academic support. I searched through Google Scholar, ACM Digital Library, PubMed, and Semantic Scholar, identifying key papers across these intersecting domains. A central reference in this review is *(paper 2)* which I found through Semantic Scholar. The screenshot below highlights this paper as a key node in the decision-making literature network.

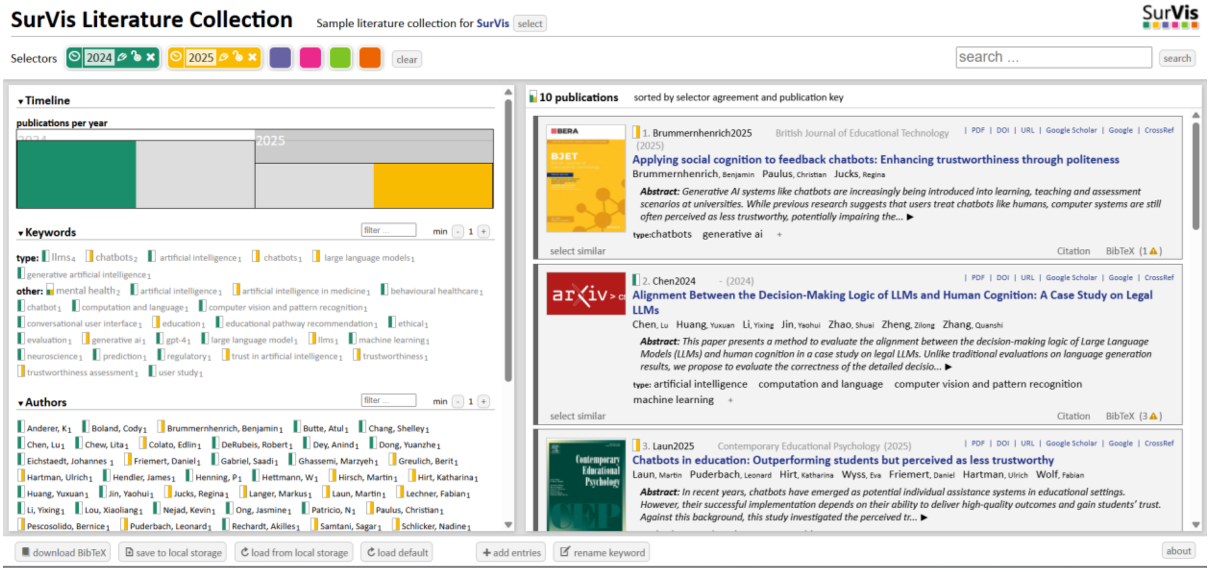


Img 3: ConnectedPaper for (paper 2); Live URL

1.4. Classification of Literature and Organisation

For the literature to be meaningful, classification of papers was done based on three key subsections relevant to the topic being explored: Evaluating the Decision Making of LLMs for Student Support Scenarios. The first subsection explores decision making and trust of LLMs in educational and ethical contexts, as trust is foundational for any decision support system used by students (paper 1, 3). The second subsection focuses on Medical and diagnostic settings where parallels to academic support decisions exist due to high-stakes and context-sensitive nature (paper 5, 6). The third subsection reviews LLMs applied to mental health and wellbeing, which is relevant to student's psychological and emotional support needs (paper 7, 8, 9). They are in the order of logical progression where it begins with general trust in LLMs, moving through clinical decision making and ends with emotional support. A screenshot of my Survis literature map is provided below and can be accessed at:

URL: https://amenainamdar.github.io/researchMethods/src/index.html



SurVis Literature Collection

Sample literature collection for SurVis

Keywords

type: llms4 chatbots2 artificial intelligence1 chatbots1 large language models1 generative artificial intelligence1

other: mental health2 artificial intelligence1 artificial intelligence in medicine1 behavioural healthcare1 chatbot1 computation and language1 computer vision and pattern recognition1 conversational user interface1 education1 educational pathway recommendation1 ethical1 evaluation1 generative ai1 gpt-41 large language model1 llms1 machine learning1 neuroscience1 prediction1 regulatory1 trust in artificial intelligence1 trustworthiness1 trustworthiness assessment1 user study1

Authors

Anderer, K1 Boland, Cody1 Brummernhenrich, Benjamin1 Butte, Atul1 Chang, Shelley1 Chen, Lu1 Chew, Lita1 Colato, Edlin1 DeRubeis, Robert1 Dey, Anind1 Dong, Yuanzhe1 Eichstaedt, Johannes 1 Friemert, Daniel1 Gabriel, Saadi1 Ghassemi, Marzyeh1 Greulich, Berit1 Hartman, Ulrich1 Hendler, James1 Henning, P1 Hettmann, W1 Hirsch, Martin1 Hirt, Katharina1 Huang, Yuxuan1 Jin, Yaohui1 Jucks, Regina1 Langer, Markus1 Laun, Martin1 Lechner, Fabian1 Li, Yixing1 Lou, Xiaoliang1 Nejad, Kevin1 Ong, Jasmine1 Patricio, N1 Paulus, Christian1 Pescosolido, Bernice1 Puderbach, Leonard1 Rechardt, Akilles1 Samtani, Sagar1 Schlicker, Nadine1

10 publications sorted by selector agreement and publication key

1. Brummernhenrich2025 British Journal of Educational Technology (2025)
Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness
Brummernhenrich, Benjamin Paulus, Christian Jucks, Regina
Abstract: *Generative AI systems like chatbots are increasingly being introduced into learning, teaching and assessment scenarios at universities. While previous research suggests that users treat chatbots like humans, computer systems are still often perceived as less trustworthy, potentially impairing the...* ▶
type:chatbots generative ai +

2. Chen2024 - (2024)
Alignment Between the Decision-Making Logic of LLMs and Human Cognition: A Case Study on Legal LLMs
Chen, Lu Huang, Yuxuan Li, Yixing Jin, Yaohui Zhao, Shuai Zheng, Zilong Zhang, Quanshi
Abstract: *This paper presents a method to evaluate the alignment between the decision-making logic of Large Language Models (LLMs) and human cognition in a case study on legal LLMs. Unlike traditional evaluations on language generation results, we propose to evaluate the correctness of the detailed decisio...* ▶
type: artificial intelligence computation and language computer vision and pattern recognition machine learning +

3. Laun2025 Contemporary Educational Psychology (2025)
Chatbots in education: Outperforming students but perceived as less trustworthy
Laun, Martin Puderbach, Leonard Hirt, Katharina Wyss, Eva Friemert, Daniel Hartman, Ulrich Wolf, Fabian
Abstract: *In recent years, chatbots have emerged as potential individual assistance systems in educational settings. However, their successful implementation depends on their ability to deliver high-quality outcomes and gain students' trust. Against this background, this study investigated the perceived tr...* ▶

Img 4: Survis visualization of the classified literature on LLM decision-making, trust, and support contexts.

2. Paper Summaries

This review is divided into three primary components that represent the fundamental aspects of how LLMs are assessed in situations pertaining to student support. Because there isn't much direct literature on this subject, these subsections were picked to span the range of pertinent applications and difficulties. First, with an emphasis on accuracy and trust, LLM Decision-Making and Trust in Educational Contexts investigates how students view and react to academic help produced by LLM. Second, research on the clinical use of LLMs is covered in LLMs in Health Diagnosis and Medical Decision Support. This book provides transferable insights into high-stakes, guidance-driven interactions. Last but not least, LLMs in Mental Health and Well-Being Guidance explores how LLMs function in emotionally charged support settings that are intimately related to the unique characteristics of students' well-being. This order reflects a progression from domain familiarity (education) to critical application (healthcare), and finally to emotionally complex interactions (mental health).

One of the studies looks at how politeness of feedback affects trust in educational chatbots, through a controlled user experiment and questionnaires. Polite (indirect) feedback increases user trust and perceived competence. It uses social cognition and linguistics theories and measures with Likert scales. It doesn't assess decision accuracy or reasoning quality but shows how surface level style affects trust which is important for student facing LLMs. But it doesn't evaluate deeper decision logic, so the research will lean towards filling the gap by looking at alignment between LLM generated support recommendations and student needs. (paper 1)

Another paper, which focusses on legal LLMs, suggests a novel method of extracting and scoring AND-OR interactions to assess how closely the internal decision logic of LLMs matches human reasoning. The study demonstrates that LLMs may produce accurate outputs utilising faulty logic by employing input token mapping and quantitative reliability metrics. Although this study is not in education, it is closely related to the exploration topic because it emphasises how important it is to assess both the underlying reasoning and the results. Its legal focus is the main weakness; this work will expand this framework to include academic support and mental health settings. (paper 2)

One of them compares the quality and perceived trust of care plans generated by students, students using ChatGPT (GPT-4), and ChatGPT alone. Despite producing higher quality outputs (expert reviewed) ChatGPT was trusted less when its authorship was disclosed. Evaluation combined expert scoring and controlled student surveys, showed a mismatch between performance and trust. This is directly relevant to the research to be worked on, on decision support: it shows that even accurate LLM guidance may be dismissed by students if perceived as impersonal or untrustworthy. But it doesn't look at complex decision making scenarios like choosing between support options which this study is going to target. (paper 3)

One of the viewpoint articles outlines ethical and governance challenges of Large language models in medicine which also includes hallucinations, IP concerns and privacy risk. It also has a lack of regulatory readiness. The paper also analyses the conceptual framework to propose multilayered safeguards. Even though, not empirical, it does provide a crucial ethical lens for exploring research on student wellbeing. This paper highlights the importance of a responsible design, but it does not provide applied evaluations or student specific cases. (paper 5)

Another case study evaluates chatbot helping students navigating wellbeing services using focus groups, usability surveys and task success rates. It identifies limitations in relevance personalization and tone, and the evaluation is task based and qualitative, this is one of the most relevant studies for the topic being explored as it shows a real example of LLM powered student support system, however it does not assess the needs of the students and accuracy of decision making. (paper 10)

Moving onto the common health problems diagnosed by LLMs, one study explores how people judge the trustworthiness of LLM powered virtual doctors, through qualitative interviews, it highlights that the trust of user depends on heuristics, third party validation and risk perception. The findings are relevant for this exploration on student support scenario where students following a chatbot's advice will be based on a similar perception. (paper 6)

A study benchmarks six large language models including GPT-4 and FLAN-T5 on tasks of mental distress like depression or anxiety detection and it uses an online database (e.g., reddit). It also compares zero or few shots and instruction tuned models where instruction tuning outperforms larger models on balanced accuracy. Evaluation emphasizes om prediction over interpretation and is task based. It demonstrates that LLMs can identify mental health risk, but it does not address the next-step decision making. (paper 8)

A literature review surveys 50+ mental health chatbots, 22 using LLMs which identify gaps in workplace applicability and evaluation consistency. It also proposes an evaluation framework and a research prototype for chatbots aligning with mental health guidelines. The insights from this are well for the related work but this does not test chatbot decision quality or user alignment in student context. (paper 9)

Coming to large language model's decisions, reactions and suggestions on mental health and wellbeing, a conceptual roadmap discusses how LLMs could assist in psychotherapy to identify high risk nuanced tasks like suicide assessments and suggesting safeguards such as interdisciplinary collaboration. The evaluation is supported by real-world use cases but is theoretical. This helps in the scenario of student support based on mental health and wellbeing however it lacks observational validation which can be taken up further. (paper 7)

A benchmark study demonstrates that when it comes to forecasting neuroscience findings, LLMs such as GPT-4 and an optimised BrainGPT perform better than human experts. It illustrates the potential of LLMs in scientific forecasting using a custom dataset (BrainBench) and quantitative metrics for accuracy and calibration. The work supports the notion that LLMs can reason over complex data, which may support their use in student support triage, even though it doesn't concentrate on giving personal advice or making moral decisions. It does not, however, include user interaction studies or scenario-based evaluation, which this work will fill in by emphasising customised student-facing choices. (paper 4)

2.1. Subsections selected under research

This review focuses on three subdomains to explore LLM promise in student support decision-making:

- LLM trust and education ethics, examining the manner in which students understand and trust AI-generated academic advice.
- LLM health diagnostic aid, providing insights into how LLMs generate and justify critical medical decisions.
- LLM mental health and well-being support, highlighting their empathetic potential and its accompanying ethical risks.

These regions overlap in the underlying issues of trust, reasoning alignment, and safety, and offer a relevant foundation to explore how LLMs might responsibly guide students in making academic, emotional, or health-related choices.

2.2. Survey Paper

"A Survey on LLM-as-a-Judge" discusses in detail using Large Language Models to evaluate tasks, choices, and outcomes, a prospective role by the name of "LLM-as-a-Judge". It is especially useful for student support cases, in which LLMs can have to weigh individual, emotional, or academic issues to recommend action like seeking help from wellbeing services or asking ECs. The survey emphasizes the importance of alignment of reliability and human judgment and categorizes the work conducted thus far into four classes: (1) usage methods (e.g., in-context learning, model choice, post-processing), (2) methods for improving judgment accuracy (e.g., prompt engineering, fine-tuning), (3) evaluation methods (e.g., bias detection, robustness, meta-evaluation), and (4) areas of application like model ranking, data annotation, and reasoning agents. Although the survey does not specifically address emotionally nuanced cases of students, it does offer a general guideline for going about how one would go about designing and testing LLMs to act like evaluators, complementing the goal of generating trustworthy, decision-supportive systems for students.

References

1. (Brummernhenrich et al., 2025) Brummernhenrich C., Piontek C. M., Kerres M., **"Applying social cognition to feedback chatbots: Enhancing trustworthiness through politeness,"** *British Journal of Educational Technology*, 2025. Doi: **https://doi.org/10.1111/bjet.13569**
2. (Chen et al., 2024) Chen L., Huang Y., Li Y., Jin Y., Zhao S., Zheng Z., Zhang Q., **"Alignment Between the Decision-Making Logic of LLMs and Human Cognition: A Case Study on Legal LLMs,"** *arXiv preprint*, arXiv:2410.09083, 2024. doi: https://doi.org/10.48550/arXiv.2410.09083
3. (Laun et al., 2025) Laun M., Puderbach L., Hirt K., Wyss E. L., Friemert D., Hartmann U., Wolff F., **"Chatbots in education: Outperforming students but perceived as less trustworthy,"** *Contemporary Educational Psychology*, vol. 81, 2025, Art. no. 102373. doi: https://doi.org/10.1016/j.cedpsych.2025.102373
4. (Luo et al., 2024) Luo X., Rechardt A., Sun G., et al., **"Large language models surpass human experts in predicting neuroscience results,"** *Nature Human Behaviour*, vol. 9, pp. 305–315, Feb. 2025. doi: https://doi.org/10.1038/s41562-024-02046-9
5. (Ong et al., 2024) Ong J. C. L., Chang S. Y-H., William W., et al., **"Ethical and regulatory challenges of large language models in medicine,"** *The Lancet Digital Health*, vol. 6, pp. e428–e432, Jun. 2024. doi: 10.1016/S2589-7500(24)00061-X
6. (Schlicker et al., 2025) Schlicker N., Lechner F., Wehrle K., Greulich B., Hirsch M. C., Langer M., **"Trustworthy Enough? Examining Trustworthiness Assessments of Large Language Model-Based Medical Agents,"** *Technology, Mind, and Behavior*, vol. 6, 2025. https://doi.org/10.1037/tmb0000164
7. (Stade et al., 2024) Stade E. C., Stirman S. W., Ungar L. H., et al., **"Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation,"** *npj Mental Health Research*, vol. 3, Art. 12, 2024. doi: https://doi.org/10.1038/s44184-024-00056-z
8. (Xu et al., 2024) Xu X., Yao B., Dong Y., et al., "**Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data,"** *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, Art. 31, Mar. 2024. doi: https://doi.org/10.1145/3643540
9. (Yuan et al., 2025) Yuan A., Colato E. G., Pescosolido B., Song H. Y., Samtani S., **"Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots,"** *ACM Transactions on*

Management Information Systems, vol. 16, no. 1, Art. 3, Feb. 2025. doi: https://doi.org/10.1145/370104

10. (Zylowski, 2024) Zylowski M., **"Evaluating the Application of Generative AI in Education: Practical Insights for Ethical Integration,"** *INTED2024 Proceedings*, 2024. Doi: 10.21125/edulearn.2024.1528

Survey: (Gu et al., 2025) Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Yuanzhuo Wang, Lionel Ni, Wen Gao, and Jian Guo, **"A Survey on LLM-as-a-Judge,"** *arXiv preprint arXiv:2411.15594*, Mar. 2025. Available: https://arxiv.org/abs/2411.15594,
doi: https://doi.org/10.48550/arXiv.2411.15594