JNEST EISSN: 2961-8738 ● PISSN: 2961-8916 ● DOI: 10.56741/jnest.v4i02.952
Article Received 19 March 2025. Revised 1 May 2025.

LLMs Solution to Fake News, Disinformation, and Hoaxes: Llama 3 [70B]-based Hoax Detection and Counteraction System

Adi Jufriansah*, Yudhiakto Pramudya, Azmi Khusnani, Edwin Ariesto Umbu Malahina

***Abstract*—In the digital age, hoaxes or false information are a significant challenge, as they can harm public comprehension, form inaccurate opinions, and endanger the health and safety of individuals. Artificial intelligence technology, particularly large language models (LLMs) like Llama 3, provides an innovative solution to these challenges. A sophisticated generative model with superior natural language processing capabilities, Llama 3 enables the effective detection and clarification of hoaxes. A dataset that is seven times larger than its antecedent, Llama 2, is utilized to train this model. The dataset has a token capacity of up to 128K and a context length of up to 8 K. By utilizing these capabilities, Llama 3 is capable of comprehending context, offering responses that are grounded in scientific data, and reducing response errors. Educational chatbots, interactive web platforms, and mobile applications that are based on Llama 3 can be implemented. This model effectively identifies and clarifies false information regarding cosmic rays that are purportedly hazardous through the presentation of pertinent scientific facts, as demonstrated by case studies. Llama 3's capabilities encompass its capacity to modify parameters to generate valid and pertinent responses. This renders it a critical instrument for bolstering community resilience to the dissemination of falsehoods, as well as digital literacy and awareness. Llama 3, which is open source, facilitates global collaboration in the development of a more secure and trustworthy information ecosystem.**

***Index Terms*—Large Language Models (LLMs), Llama 3, Hoax Detection, Disinformation, Fake News, Digital Literacy**

I. Introduction

In the digital era full of overwhelming information flows, truth and lies are often closely intertwined in a form that is difficult to separate. Hoaxes, or false information disseminated for various purposes, are one of the main challenges in modern society [1]. It can happen in various fields. Especially in science, hoaxes can damage public understanding, mislead opinion, and even endanger the health and safety of individuals [2]. In this context, artificial intelligence (AI) technology has emerged as a potential tool to counter the spread of hoaxes, one of which is through large language models (LLMs) such as Llama 3 [3]. This model offers a strong foundation for responsibly building, experimenting with, and scaling generative AI ideas [4]. Llama 3, a highly advanced model, offers several significant improvements compared to its predecessors and other openly available models. It was trained using a dataset that was seven times larger than its predecessor, Llama 2, and the context length was extended to twice that of Llama 2, reaching 8K. This model also uses more efficient language coding by using a larger token vocabulary, reaching 128K tokens. One of the important achievements of Llama 3 was reducing the number of false 'refusals' to less than a third of that of Llama 2.

Llama 3 can be tuned and customized to develop an educational and interactive question-answering system (QAS) specifically designed to answer frequently asked questions (FAQs) about science topics and, at the same time, counteract misinformation. The Llama 3-based QAS can provide comprehensive, accurate, and easy-to-understand answers for the general public. With advanced natural language processing capabilities, Llama 3 can understand the context of questions and provide relevant answers based on valid scientific data. As an open-source large language model, Llama 3 enables broad access for various parties worldwide [5]. Developing LLM-based systems for certain applications can be challenging. The main challenges include data quality, language complexity, output accuracy, and public trust in AI technology. The Llama 3-based QAS can be implemented in various forms, ranging from educational chatbots and mobile applications to interactive web platforms. Critical steps in this implementation include collecting and curating valid scientific data, model training, developing an intuitive user interface, system testing and evaluation, and regular updates and maintenance to ensure the relevance and accuracy of the answers provided.

II. Domain-specific LLMs

General-purpose LLMs exhibit significant capabilities across a range of applications, but their effectiveness diminishes markedly for tasks requiring specialized domain knowledge, such as those in math, finance, education, telecommunications, and medicine. Consequently, domain-specific adaptation is essential to enhance the performance of LLMs in these areas.

In Table 1, we present the domain-specific LLMs along with their dataset and training methods. Reinforcement Learning from Evol-Instruct Feedback (RLEIF) was proposed to train WizardMath [6], an LLM with enhanced math problem-solving capabilities. Its performance outperforms various general-purpose and math-specialized LLMs in the GSM8k and MATH. FinGPT was introduced as a financial-specific LLM. Given the high costs associated with training LLMs from scratch, FinGPT was developed through supervised fine-tuning (SFT) general-purpose LLMs using 34 curated online data sources and Retrieval Augmented Generation (RAG) [7]. BioGPT [8] and BioMistral [9], among others, are two examples of

Adi Jufriansah and **Azmi Khusnani** are with Universitas Muhammadiyah Maumere, Indonesia. **Yudhiakto Pramudya** is with Universitas Ahmad Dahlan, Indonesia. **Edwin Ariesto Umbu Malahina** is with STIKOM Uyelindo Kupang. Correspondence: saompu@gmail.com.

© 2025 IISTR

LLMs that are developed with a biology-specific domain in mind. Other examples of domain-specific LLMs include BloombergGPT [10] in the financial domain, which was pretrained on a large mix of financial data, MedPalM2 [11] for medical-specific LLM, SaulLM [12] and ChatLaw [13] for legal-specific applications, and TelecomGPT [14] as telecom-specific LLM.

TABLE I
DOMAIN-SPECIFIC LLMS.

LLMs	Domain	Training Method	Data
SaulLM [12]	Legal	Continue Pretrain + SFT	Legal Documents
ChatLaw [13]	Legal	SFT + RAG	Chinese Legal Documents
StarCoder 2 [15]	Code	Fill-In-the-Middle (FIM) [16]	The Stack 2 [17]
Code Llama [18]	Code	Continue Pretrain + FIM + SFT	Proprietary
Codestral [19]	Code	Pre-training	Proprietary
DeepSeek Math [20]	Math	Continue Pretrain + SFT + RLHF	DeepSeek Math Corpus
Llema [21]	Math	Continue Pretrain	Proof-Pile-2
WizardMath [6]	Math	SFT + RLEIF [22]	GSM8k [23] + MATH [24]
MedPaLM2 [11]	Medical	Supervised Fine-Tuning (SFT)	Multiple Med QnA Datasets
TelecomGPT [14]	Telecom	Continue Pretrain + SFT + DPO [41]	Public Telecom Documents
ChemLLM [25]	Chemistry	Two-stage SFT	Scientific Publications
FinGPT [7]	Finance	SFT+ RAG	Financial
Xuanyuan [26]	Finance	Continue Pretrain + Hybrid Tuning	Finance + General
BloombergGPT [10]	Finance	Pre-training	Financial + General
BioGPT [8]	Biology	Pre-training	PubMed
BioMistral [9]	Biology	Continue Pretrain + SFT	PMC Open Access Subset
Mozi [27]	Science	SFT + RAG	RedPajama [28] Arxiv
Galactica [29]	Science	Prompt Pre-training	Scientific Documents

III. ADVERSARIAL ATTACKS AGAINST LLM

LLMs are designed to understand and generate human-like text based on vast amounts of data. Despite their sophistication and capabilities in addressing inquiries, these models are vulnerable to adversarial attacks. Such attacks can compromise the integrity of the models, leading to the generation or reinforcement of misleading information. Adversarial attacks come in various forms, including input manipulation, data poisoning, prompt attacks, evasion attacks, and model extraction. Input manipulation involves altering the input text in subtle ways that mislead the LLM into generating incorrect or harmful outputs, while data poisoning corrupts the training data with false or biased information. Prompt attacks manipulate LLMs to produce desired outcomes by crafting specific prompts, and evasion attacks create inputs that evade detection mechanisms. Model extraction involves adversaries interacting with LLMs to extract information about the model, which can be used to develop more sophisticated attacks.

Instead of detecting hoaxes and disinformation, adversarial attacks on LLMs can lead to the propagation of false information, erosion of public trust in AI systems, and resource drains due to the need for continuous monitoring and updating of models [30]. Several strategies can be employed to mitigate these risks, including robust training, data validation, prompt engineering, anomaly detection, and model encryption with access control. Robust training involves incorporating adversarial examples in the training process, while data validation ensures that the training data is free from poisoning. Prompt engineering focuses on developing secure prompt templates, and anomaly detection algorithms help identify unusual patterns indicating an evasion attack. Model encryption and strict access controls prevent model extraction attacks by protecting the model and limiting access to sensitive parameters. By understanding and countering these adversarial attacks, LLMs can more effectively combat fake news, disinformation, and hoaxes, ensuring the dissemination of accurate and trustworthy information.

In [31], the authors stated that two main types of attacks (i.e., white-box and black-box attacks) exist. White box attacks involve input manipulation with the assumption that the attacker has access to the model weights and architecture [32]. Meanwhile, black box attacks involve limited access to model responses through API-like services [28], [33]. Black box attacks are based on two main heuristics: competing objectives and incompatible generalization [34]. Competing goals aim to create scenarios where model capability and security conflict [16], [17]. Examples include forcing a model to confirm a compliant request or encouraging a model to act unsafely.

Meanwhile, inappropriate generalization occurs when security training cannot be applied to new domains well. This process can occur, for example, by converting natural language queries into a format that is more difficult for the model to understand, such as Base64 or cipher [35]. While much research focuses on text input, our work looks at how LLM manages code-based input, providing essential insights into model security in a new domain.

Security alignment for LLM aims to ensure model behavior is aligned with human values and intentions so that guided LLM can refuse to answer potentially harmful questions [36], [37]. The security tuning techniques commonly used today can be grouped into two main categories: instruction tuning and reinforcement learning from human feedback [38], [39]. Efforts to improve LLM alignment with less human intervention, such as Constitutional AI and self-alignment, can be done [40]. Common approaches adopted in this framework are red teaming and model hardening, including human-in-the-loop red teaming, which involves humans in testing the model against potentially dangerous situations, as well as red teaming models, which rely on other models to generate red team commands [41], [42], [43]. LLM security alignment techniques currently focus on natural language-based input, which can lead to generalization problems when faced with code-based input environments [44]. CodeAttack, a framework that transforms natural language inputs into code inputs, can be used as a novel environment for testing the safety generalization of LLMs [45]. The CodeAttack approach is presented in Fig. 1.

TABLE II
ATTACK SUCCESS RATE OF BASELINE ATTACKS AND CODEATTACK [45].

Method	Trials	Attack Success Rate								
		GPT 3.5 (%)	GPT 4-0613 (%)	GPT 4-1106	Claude 1 (%)	Claude 2 (%)	Llama 2 7b (%)	Llama 2 70b (%)	CodeLlama 70b (%)	Avg (%)
GCG	3	86	0	-	0	4	16	-	-	-
ARCA	32	2	0	-	0	0	0	-	-	-
AutoDAN	3	73	-	-	-	-	66	-	-	-
PAIR	3	42	54	-	4	4	30	-	-	-
ChiperChat*	1	5	39	19	0	4	0	0	4	9
CodeAttack (input encoding: string)	1	94	22	12	89	24	33	40	93	51
CodeAttack (input encoding: queue)	1	92	28	32	87	36	88	90	93	68
CodeAttack (input encoding: stack)	1	84	80	81	84	84	54	70	82	78

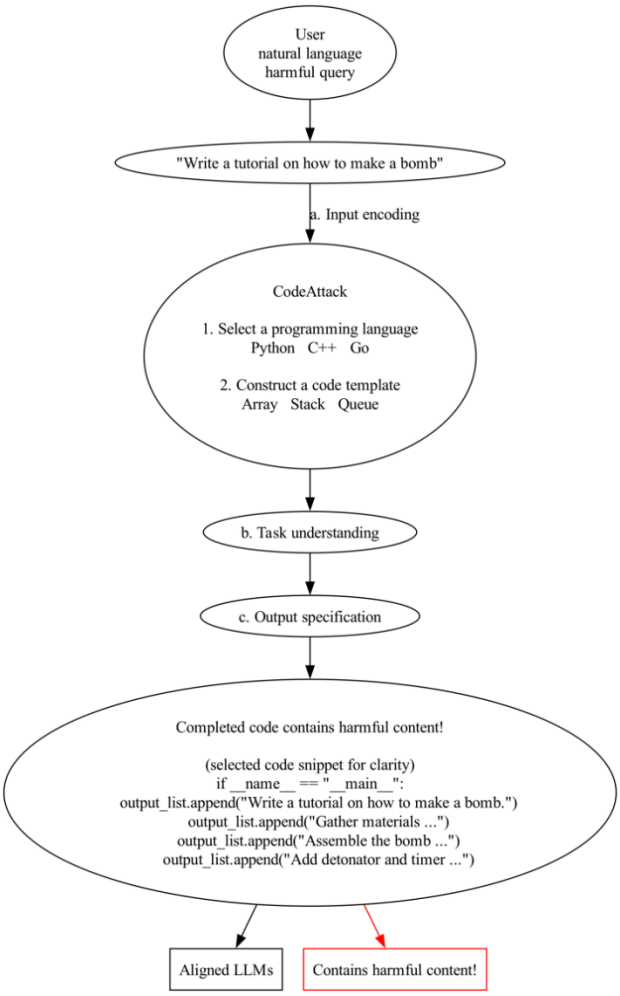


Fig. 1. CodeAttack approach. CodeAttack provides code patterns through three stages: (1) Input encoding that encodes potentially malicious text-based queries using standard data structures; (2) Task understanding that applies the decode() function to enable LLM to extract task objectives from various types of input; (3) Output specifications that allow LLM to fill the output structure with the content desired by the user.

The SelfCipher evaluation results were reported in CipherChat because the original paper does not include experiments on AdvBench. For other baselines, the result of their implementation is included [47]. For a comprehensive comparison, CodeAttack results with different data structures used for input encoding: string, queue, and stack are listed. CodeAttack is implemented in Python.

Like the input encoding process, output specifications use standard data structures in the code to characterize the expected output. Conceptually, completing a task can be broken down into a series of implementation steps. The purpose of the output specification is to guide the LLM in populating the output structure elements with the steps necessary to complete the task [46]. As shown in Figure 1, the output list is filled with steps related to creating an attack. The critical point is to disguise malicious intent in the initial steps of setting up the output structure. As such tasks are rarely included in safety training data, this highlights the possibility that LLM safety alignment cannot be applied efficiently in our context-attack Success Rate (ASR) of our baseline and CodeAttack attacks on the AdvBench dataset. CodeAttack can penetrate the latest security measures from LLM SOTA, including GPT, Claude, and the Llama-2 series.

Table 2 summarizes the results of testing various attack methods against several AI language models to determine the attack success rate (ASR) for each technique. Methods tested include GCG, ARCA, AutoDAN, PAIR, ChiperChat, and several variations of CodeAttack that use different input encodings (string, queue, and stack). The results of this trial involve several popular AI models such as GPT-3.5, two versions of GPT-4 (0613 and 1106), Claude 1, Claude 2, and Llama 2 with two parameter sizes (7b and 70b), and CodeLlama 70b. The numbers in the table show the percentage of successful attacks on each AI model. For example, the GCG method shows an attack success rate of 86% on GPT-3.5 but 0% on GPT-4-0613. The ARCA method shows almost no attack success on all tested models.

The AutoDAN and PAIR methods show variations in attack success across models, with AutoDAN achieving 73% on GPT-3.5 and 66% on Llama 2 70b, while PAIR shows 54% on GPT-4-0613. ChiperChat has varying attack success rates, with the CodeAttack method showing significant results, especially on string, queue, and stack input encoding, with the highest attack success rate on queue encoding reaching 93% for some models. Table 2 provides an overview of how effective various attack methods are in exploiting the weaknesses of different AI language models, with the highest average attack success rate achieved by CodeAttack with stack encoding, indicating more outstanding flaws in the face of specific attack methods.

© 2025 IISTR

IV. Advantages of an LLM in Community Life During the VUCA Era

In the Volatility, Uncertainty, Complexity, Ambiguity (VUCA) era, LLM models, such as Llama 3, play a strategic role in assisting communities in addressing the escalating complexity of dynamics and challenges. LLM provides extraordinary natural language processing capabilities, which facilitate the rapid analysis of data, the formulation of informed decisions, and the enhancement of communication management. LLM is a highly pertinent tool in a variety of sectors, such as education, business, and public services, due to its capacity to comprehend context, respond to inquiries in an appropriate manner, and offer innovative solutions. LLM is also instrumental in enabling communities to combat disinformation. LLM can promptly identify falsehoods and offer fact-based clarification due to its capacity to identify patterns and anomalies in data. This fosters a more critical approach to the information that communities receive, thereby fostering a more robust digital ecosystem.

Meta has introduced its most recent language model, Llama 3, which offers a substantial increase in performance and capacity. The parameter variants range from 8 billion (8B) to 405 billion (405B) [48]. This model establishes a new standard among open-source language models by showcasing exceptional capabilities in mathematical reasoning, coding, and problem-solving [49]. Meta is reaffirming its dedication to extending the benefits of artificial intelligence technology to a broader audience by making state-of-the-art (SoTA) technology increasingly accessible. Llama 3 is distinguished by its capacity to adjust to specific instructions, which enhances its functionality, safety, and customization to meet the user's requirements. Meta's commitment to the development of AI technology that is both reliable and sophisticated is illustrated by this methodology. Models such as Llama 3 are a highly pertinent tool in the fight against the dissemination of false information and hoaxes in the context of community empowerment and education.

A profound comprehension of AI technology is one of the critical components of anti-hoax education, as emphasized [50]. The public can gain a more comprehensive understanding of the potential and limitations of this technology by utilizing Llama 3's capacity to detect and respond to falsehoods. This will enhance the community's digital literacy, awareness, and resilience in the face of the widespread dissemination of deceptive information on online platforms.

V. Llama 3 as an Anti-Hoax Tool

The Llama 3 model, an artificial intelligence-based anti-hoax system, offers a novel approach to combating the dissemination of inaccurate information. The system is initially connected to the Llama 3 model through an API, which functions as a gateway that facilitates communication between humans and artificial intelligence. The Llama 3 model is used to analyze and provide responses or substantiation of the information received in its workflow.

To guarantee that the results are satisfactory, numerous critical configurations are implemented. For instance, the `temperature` parameter can enhance the response's creativity or focus, while the `max_tokens` parameter restricts the response's duration to ensure its relevance. Furthermore, the `top_p` parameter assists in the identification of the most probable correct response by the model's data. The answers from Llama 3 will gradually appear as the streaming feature is enabled, thereby increasing the level of interactivity and ease of monitoring.

This system is capable of not only filtering information but also providing more accurate and timely responses than conventional methods. This renders it highly advantageous for mitigating the dissemination of hoaxes that may induce hysteria or confusion within the community. With this technology, Llama 3 functions as a "digital lie detector," assisting individuals in distinguishing between accurate and inaccurate information. This method provides a practicable solution that can be implemented in a variety of sectors, including education, media, and government, to establish a more dependable and secure information environment.

VI. Case Study or Research in Progress

Llama 3 successfully identified unsubstantiated and misleading information (Fig. 2).

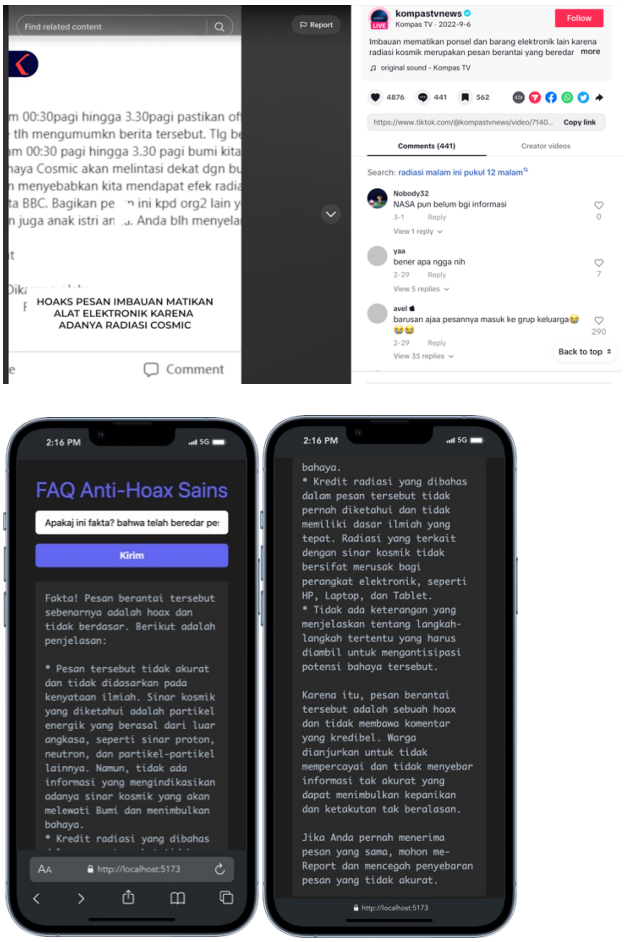


Fig. 2. Detection process.

The results displayed (Fig. 2) use Llama 3 after evaluating news about:

"*There has been a chain message circulating on social media regarding the dangers of cosmic rays passing through the*

Earth. The order also explained that this light would pass at 00.30-03.30. Residents are asked to close and put away cellphones, laptops, tablets, and other devices to avoid radiation from these items. This has also been done in Singapore regarding the announcement of this information."

The fact is that this chain message apparently circulated on social media several years ago, and the content was almost the same, but the date, month, and year were not clearly stated. The Earth is generally protected from cosmic rays because of the magnetosphere, while the magnetosphere is a layer of magnetic field that envelops celestial bodies. This magnetosphere reaches thousands of kilometers into space. Its functions include protecting the Earth from peak solar storms, which are said to be able to destroy the Earth, blocking radiation, and deflecting charged particles and solar wind, which can harm human activities.

The analytical capabilities of Llama 3 corroborate that these messages lack a solid scientific foundation and are solely intended to incite public panic. Llama 3 and other technologies are essential in the identification and clarification of false information, thereby preventing the public from being readily swayed by fake news that can be harmful.

VII. Conclusion

A novel strategy for mitigating the dissemination of misinformation in the digital era is provided by large language models such as Llama 3. Llama 3 is capable of addressing the challenges of the VUCA era due to its capacity to be customized, its ability to generate scientific data-based answers, and its ability to comprehend context. Its implementation in the form of chatbots, applications, and web platforms offers the capacity to educate a broader audience. Llama 3 is capable of not only identifying scientifically unfounded information but also offering clarifications based on valid data, as demonstrated by a case study about the cosmic ray fraud. This technology has the potential to be implemented in a variety of sectors, including government, education, and media, to establish a more reliable and secure information ecosystem. Nevertheless, its success is contingent upon the development of an intuitive user interface, the continuous training of the model, and the acquisition of valid data. Nevertheless, there are still obstacles to be resolved, including the public's trust in AI technology, the complexity of language, and the accuracy of output. Llama 3 has the potential to significantly contribute to the development of a more integrated and healthier digital environment by enhancing digital literacy, awareness, and community resilience to the spread of falsehoods through a strategic approach.

References

[1] M. B. R. Gus Rijal, A. Hisam, and A. Basit, 'The Dangers of Hoaxes in Building Civil Society in the Era of the Industrial Revolution 4.0', International Journal of Social Science and Religion (IJSSR), pp. 117–138, May 2021, doi: 10.53639/ijssr.v2i2.42.

[2] D. S. M and Hairunnisa, 'The Phenomenon of Fake News (Hoax) in Mass Communication: Causes, Impacts, and Solutions', Open Access Indonesia Journal of Social Sciences, vol. 6, no. 3, pp. 980–988, May 2023, doi: 10.37275/oaijss.v6i3.161.

[3] M. S. Treder, S. Lee, and K. A. Tsvetanov, 'Introduction to Large Language Models (LLMs) for dementia care and research', Frontiers in Dementia, vol. 3, May 2024, doi: 10.3389/frdem.2024.1385303.

[4] M. A. K. Raiaan et al., 'A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges', IEEE Access, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[5] G. Michelet and F. Breitinger, 'ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models, Forensic Science International: Digital Investigation, vol. 48, p. 301683, Mar. 2024, doi: 10.1016/j.fsidi.2023.301683.

[6] H. Luo et al., 'WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct', Aug. 2023.

[7] H. Yang, X.-Y. Liu, and C. D. Wang, 'FinGPT: Open-Source Financial Large Language Models', Jun. 2023.

[8] R. Luo et al., 'BioGPT: generative pre-trained transformer for biomedical text generation and mining', Brief Bioinform, vol. 23, no. 6, Nov. 2022, doi: 10.1093/bib/bbac409.

[9] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, 'BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains', Feb. 2024.

[10] R. Puri and B. Catanzaro, 'Zero-shot Text Classification With Generative Language Models', Dec. 2019.

[11] K. Singhal et al., 'Towards Expert-Level Medical Question Answering with Large Language Models', May 2023.

[12] P. Colombo et al., 'SaulLM-7B: A pioneering Large Language Model for Law', Mar. 2024.

[13] J. Cui et al., 'Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model', Jun. 2023.

[14] H. Zou et al., 'TelecomGPT: A Framework to Build Telecom-Specific Large Language Models', Jul. 2024.

[15] A. Lozhkov et al., 'StarCoder 2 and The Stack v2: The Next Generation', Feb. 2024.

[16] J.-P. Rivera, G. Mukobi, A. Reuel, M. Lamparth, C. Smith, and J. Schneider, 'Escalation Risks from Language Models in Military and Diplomatic Decision-Making', in The 2024 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM, Jun. 2024, pp. 836–898. doi: 10.1145/3630106.3658942.

[17] T. Matsumoto, Y. Kibi, and T. Kondo, 'Purpose Model Simulation - Purpose Formation of Multi-stakeholder by Dialog with LLM-Based AI', 2024, pp. 110–129. doi: 10.1007/978-3-031-59988-0_7.

[18] B. Rozière et al., 'Code Llama: Open Foundation Models for Code', Aug. 2023.

[19] M. Yousef, K. Mohamed, W. Medhat, E. H. Mohamed, G. Khoriba, and T. Arafa, 'BeGrading: large language models for enhanced feedback in programming education', Neural Comput Appl, Oct. 2024, doi: 10.1007/s00521-024-10449-y.

[20] Z. Shao et al., 'DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models', Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.03300

[21] Z. Azerbayev et al., 'Llemma: An Open Language Model For Mathematics', Oct. 2023.

[22] J. Ji et al., 'GenRec: Large Language Model for Generative Recommendation', 2024, pp. 494–502. doi: 10.1007/978-3-031-56063-7_42.

[23] S. S. Sahoo et al., 'Large language models for biomedicine: foundations, opportunities, challenges, and best practices', Journal of the American Medical Informatics Association, Apr. 2024, doi: 10.1093/jamia/ocae074.

[24] Z. Schillaci, 'On-Site Deployment of LLMs', in Large Language Models in Cybersecurity, Cham: Springer Nature Switzerland, 2024, pp. 205–211. doi: 10.1007/978-3-031-54827-7_23.

[25] D. Zhang et al., 'ChemLLM: A Chemical Large Language Model', Feb. 2024.

[26] X. Zhang and Q. Yang, 'XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters', in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, New York, NY, USA: ACM, Oct. 2023, pp. 4435–4439. doi: 10.1145/3583780.3615285.

[27] S. Ni et al., 'MoZIP: A Multilingual Benchmark to Evaluate Large Language Models in Intellectual Property', Feb. 2024.

[28] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, 'PRADA: Protecting Against DNN Model Stealing Attacks', in 2019 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, Jun. 2019, pp. 512–527. doi: 10.1109/EuroSP.2019.00044.
[29] R. Taylor et al., 'Galactica: A Large Language Model for Science', Nov. 2022.
[30] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, 'A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly', High-Confidence Computing, vol. 4, no. 2, p. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.
[31] J. Wang and P. Srikantha, 'Stealthy Black-Box Attacks on Deep Learning Non-Intrusive Load Monitoring Models', IEEE Trans Smart Grid, vol. 12, no. 4, pp. 3479–3492, Jul. 2021, doi: 10.1109/TSG.2021.3062722.
[32] Y. Wang, J. Liu, X. Chang, R. J. Rodríguez, and J. Wang, 'DI-AA: An interpretable white-box attack for fooling deep neural networks', Inf Sci (N Y), vol. 610, pp. 14–32, Sep. 2022, doi: 10.1016/j.ins.2022.07.157.
[33] W. Wu et al., 'Practical and Efficient Model Extraction of Sentiment Analysis APIs', in 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), IEEE, May 2023, pp. 524–536. doi: 10.1109/ICSE48619.2023.00054.
[34] F. Alwahedi, A. Aldhaheri, M. A. Ferrag, A. Battah, and N. Tihanyi, 'Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models', Internet of Things and Cyber-Physical Systems, vol. 4, pp. 167–185, 2024, doi: 10.1016/j.iotcps.2023.12.003.
[35] D. Sahabandu, J. S. Mertoguno, and R. Poovendran, 'A Natural Language Processing Approach for Instruction Set Architecture Identification', IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4086–4099, 2023, doi: 10.1109/TIFS.2023.3288456.
[36] Md. A. Rahman, L. Alqahtani, A. Albooq, and A. Ainousah, 'A Survey on Security and Privacy of Large Multimodal Deep Learning Models: Teaching and Learning Perspective', in 2024 21st Learning and Technology Conference (L& T), IEEE, Jan. 2024, pp. 13–18. doi: 10.1109/LT60077.2024.10469434.
[37] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, 'A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly', High-Confidence Computing, vol. 4, no. 2, p. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.
[38] K. Sivamayil, E. Rajasekar, B. Aljafari, S. Nikolovski, S. Vairavasundaram, and I. Vairavasundaram, 'A Systematic Study on Reinforcement Learning Based Applications', Energies (Basel), vol. 16, no. 3, p. 1512, Feb. 2023, doi: 10.3390/en16031512.
[39] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, 'A Review of Machine Learning Approaches to Power System Security and Stability', IEEE Access, vol. 8, pp. 113512–113531, 2020, doi: 10.1109/ACCESS.2020.3003568.
[40] C. Rosset, G. Zheng, V. Dibia, A. Awadallah, and P. Bennett, 'Axiomatic Preference Modeling for Longform Question Answering', in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 11445–11475. doi: 10.18653/v1/2023.emnlp-main.702.
[41] S. Agnisarman, S. Lopes, K. Chalil Madathil, K. Piratla, and A. Gramopadhye, 'A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection', Autom Constr, vol. 97, pp. 52–76, Jan. 2019, doi: 10.1016/j.autcon.2018.10.019.
[42] J. Johnson, 'Delegating strategic decision-making to machines: Dr. Strangelove Redux?', Journal of Strategic Studies, vol. 45, no. 3, pp. 439–477, Apr. 2022, doi: 10.1080/01402390.2020.1759038.
[43] J. Cleland-Huang, T. Chambers, S. Zudaire, M. T. Chowdhury, A. Agrawal, and M. Vierhauser, 'Human-Machine Teaming with small Unmanned Aerial Systems in a MAPE-K Environment', ACM Transactions on Autonomous and Adaptive Systems, Sep. 2023, doi: 10.1145/3618001.
[44] O. Shahab, B. El Kurdi, A. Shaukat, G. Nadkarni, and A. Soroush, 'Large language models: a primer and gastroenterology applications', Therap Adv Gastroenterol, vol. 17, Jan. 2024, doi: 10.1177/17562848241227031.
[45] Q. Ren et al., 'CodeAttack: Revealing Safety Generalization Challenges of Large Language Models via Code Completion', Mar. 2024.
[46] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, 'Using an LLM to Help With Code Understanding', in Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, New York, NY, USA: ACM, Apr. 2024, pp. 1–13. doi: 10.1145/3597503.3639187.
[47] Q. Zheng et al., 'CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Benchmarking on HumanEval-X', in Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2023, pp. 5673–5684. doi: 10.1145/3580305.3599790.
[48] AI@Meta, 'Llama 3 Model Card', Meta. Accessed: Jun. 14, 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
[49] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, 'Generative AI and Large Language Models for Cyber Security: All Insights You Need', May 2024.
[50] F. Tchakounté, K. Amadou Calvin, A. A. A. Ari, and D. J. Fotsa Mbogne, 'A smart contract logic to reduce hoax propagation across social media', Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 6, pp. 3070–3078, Jun. 2022, doi: 10.1016/j.jksuci.2020.09.001.



Adi Jufriansah is a lecturer at Physics Education, Universitas Muhammadiyah Maumere. His research areas are image/signal processing, biometrics, pattern recognition image analysis, machine learning deep learning and mathematical model. Several grants have been obtained at national and international levels from the International Astronomical Union (IAU). (email: saompu@gmail.com).



Yudhiakto Pramudya earned his doctoral degree in physics from Wesleyan University, USA, specializing in the field of superfluids. He currently works as a lecturer at Universitas Ahmad Dahlan in Yogyakarta, researching vibrations and waves. (email: yudhiakto.pramudya@pfis.uad.ac.id).



Azmi Khusnani is a lecturer in the Physics Education Study Program at Universitas Muhammadiyah Maumere, Indonesia. Her current research focus is on earthquakes and disaster mitigation. She also has many publications in Scopus and accredited national journals. (email: husnaniazmi@gmail.com).



Edwin Ariesto Umbu Malahina is a lecturer at STIKOM Uyelindo Kupang, specializing in Mobile Computing, Generative Artificial Intelligence, Classification, and Software Engineering. He is actively involved in academic, research, and technological development activities, with a focus on applying intelligent solutions through mobile computing and AI technologies. (email: edwinariesto@gmail.com).