InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol 2(No.2), 13-19 (2025)
DOI: https://doi.org/10.51173/ijds.v2i2.33
ijds.mtu.edu.iq

EVALUATING AI LANGUAGE MODELS IN NEWS RETRIEVAL: A COMPARATIVE STUDY OF CHATGPT-PLUS AND DEEPSEEK (R1)

Omar Mustafa AL-Janabi [1*], Osamah Mohammed Alyasiri [2,3], Elaf Ayyed Jebur [1], Shahad Mohgoob Nafl [1]

[1] *College of Medicine, University of Baghdad, Baghdad, 10047, Iraq*
[2] *Karbala Technical Institute, Al-Furat Al-Awsat Technical University, Karbala, 56001, Iraq*
[3] *School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia*

** Corresponding author E-mail: omar.m@comed.uobaghdad.edu.iq (Omar Mustafa AL-Janabi)*

RESEARCH ARTICLE

ARTICLE INFORMATION

SUBMISSION HISTORY:
Received: 7 February 2025
Revised: 25 May 2025
Accepted: 27 June 2025
Published: 30 June 2025

KEYWORDS:
LLMs;
ChatGPT;
DeepSeek;
Information Retrieval;
News Accessing.

ABSTRACT

The increasing complexity of how humans interact with and process information has demonstrated significant advancements in Natural Language Processing (NLP), transitioning from task-specific architectures to generalized frameworks applicable across multiple tasks. Despite their success, challenges persist in specialized domains such as translation, where instruction tuning may prioritize fluency over accuracy. Against this backdrop, the present study conducts a comparative evaluation of ChatGPT-Plus and DeepSeek (R1) on a high-fidelity bilingual retrieval-and-translation task. A single standardize prompt directs each model to access the Arabic-language news section of the College of Medicine, University of Baghdad, retrieve the three most recent articles, and translate them into English. ChatGPT-Plus fulfilled the prompt successfully, extracting authentic Arabic content and delivering fluent, semantically accurate English translations. DeepSeek (R1), by contrast, failed to retrieve the requested articles and instead produced only generic procedural advice – evidence of its lack of real-time web access and a retrieval-augmented generation (RAG) mechanism.

1. INTRODUCTION

The critical importance of how humans interact with and process information has heightened the need for Artificial Intelligence (AI) as a transformative force capable of redefining how to manage these processes. The prevalence of specialized, task-oriented architectures in Natural Language Processing (NLP) pressed the need for Large Language Models (LLMs) to revolutionize the field into more generalized, versatile frameworks. Contemporary LLMs, equipped with intelligently designed zero-shot and few-shot learning capabilities, exhibit remarkable flexibility and can perform a broad range of linguistic tasks without task-specific training [1-4].

OpenAI's ChatGPT exemplifies a general-purpose framework, demonstrating its proficiency in performing linguistically complex conversational interaction thanks to extensive pre-training and reinforcement learning from human feedback (RLHF) [5,6]. However, operating in specialized fields can pose unique challenges because of the persistent trade-off between fluency and accuracy [7,8].

In specific tasks such as high-fidelity translation, ChatGPT often prioritizes fluency and readability over literal accuracy, particularly in technical or domain-specific texts [9]. Preserving the original meaning in scientific communication or medical documentation, therefore, remains a significant challenge, prompting researchers to refine LLM capabilities through specialized feedback mechanisms [10–12].

While researchers continue to refine LLMs' capabilities [13-18], a Chinese model named DeepSeek (R1) was developed to compete with the traditionally dominant Western AI research ecosystem [13-18]. DeepSeek (R1) has attracted global attention by achieving performance comparable to proprietary models at a significantly lower cost, while simultaneously promoting openness [19]. The model also raises important questions about intellectual property, innovation, accessibility, and competitive dynamics in proprietary AI [20-22]. Despite its broad linguistics competence, DeepSeek (R1) still struggles with information retrieval and other specialized real-time tasks, pressing the need for further enhancement.

To this end, the present study examines the essential components of advanced language models, such as ChatGPT and DeepSeek, that are critical to effective AI-driven information retrieval, including robust data indexing and continuous model training. The findings demonstrate that general-purpose LLMs remain limited in real-time scenarios involving cross-lingual news retrieval.

2. BACKGROUND

Artificial Intelligence (AI) has evolved from rule-based expert systems to advanced machine learning approaches, culminating in Large Language Models (LLMs) that can understand and generate human-like text [24-28]. A key breakthrough was the introduction of transformer-based architectures (e.g., the Transformer [5]) that enabled training extremely large language models on vast corpora. By 2020, models like GPT-3 (175 billion parameters) demonstrated remarkable zero-shot and few-shot learning abilities – they could perform diverse NLP tasks without task-specific training, simply by being prompted with examples or instructions. Subsequent LLMs pushed these limits further: for instance, Google's PaLM (with 540 billion parameters) showed that scaling model size and data leads to even stronger language understanding and generation capabilities. These models significantly influenced Natural Language Processing (NLP) by achieving state-of-the-art results across tasks with minimal or no task-specific supervision. An important refinement in LLM development was reinforcement learning from human feedback (RLHF), used in training ChatGPT to fine-tune its responses to be more helpful and aligned with user intent.

The release of DeepSeek (R1) in 2024-2025 disrupted the AI landscape by showing that high-performance LLMs can be produced outside of big tech companies and shared openly. This has sparked discussions on intellectual property and AI governance, as organizations reconsider strategies in response to the rise of powerful open models. In summary, modern AI is increasingly defined by large language models – massive neural networks trained on internet-scale data, which have become central to advancing NLP and are now at the forefront of international AI research and competition.

3. RECENT WORK

Recent works evaluated the translation performance of GPT-3.5 and GPT-4 against traditional Machine Translation (MT) systems, using human assessments and prompt-based techniques to address low-resource and domain-specific challenges.

Hendy et al. [29] and Wang et al. [30] evaluated ChatGPT on formal news and casual speech translations and observed quality approaching that of top MT systems. In many cases, when prompts are carefully crafted (providing instructions or context), ChatGPT can output translations that human evaluators rate as highly as those from dedicated MT models. This suggests that prompt engineering and interactive feedback can mitigate some weaknesses of LLMs in translation tasks. Moreover, independent evaluation by Rahman et al. [31] compared DeepSeek-V2 with GPT-4 and commercial MT engines across multiple translation scenarios, finding that DeepSeek-V2 offered robust performance in preserving semantic fidelity and stylistic nuance, especially in newswire and conversational domains. Their results highlighted DeepSeek's strength in maintaining discourse coherence across sentence boundaries, a challenge where many traditional MT systems often falter. Another study by Zhang [19] further reported that DeepSeek's fin-grained contextual modeling contributed to higher adequacy and fluency ratings in human evaluations, particularly when translating between structurally dissimilar language pairs (e.g., English-Korean, Chinese-German) [32].

These findings collectively indicate a shift in the MT landscape, where emerging open-source LLMs like DeepSeek not only rival closed commercial models in translation tasks but also democratize access to high-performance multilingual NLP. The growing body of research suggests that with targeted training and architectural improvements, LLMs are becoming increasingly viable as general-purpose translation engines in both academic and industrial contexts.

These studies (among others) form the foundation of current knowledge on large language models in translation. They collectively show an evolution from demonstrating the raw potential of LLMs as translators [2], to refining their performance and comparing them with industry MT systems (2023 studies), and even comparing them with human translation quality (Huang et al. 2024). Furthermore, the comparative research involving DeepSeek and ChatGPT (2025) extends this line of inquiry into real-time translation of web content, pointing toward future systems that combine LLMs with retrieval for on-demand multilingual information access.

Therefore, the consistent theme across recent work is that LLMs have become increasingly effective at translation tasks, and ongoing research is actively exploring how to address their remaining weaknesses (such as handling niche domains, low-resource languages, and real-time knowledge updates) to fully realize their potential in both professional translation workflows and interactive, user-driven translation services.

4. METHODOLOGY

The methodology of this study undertakes a comparative evaluation of two large language models: ChatGPT-Plus and DeepSeek (R1) – in the context of information retrieval and cross-lingual translation. The task scenario was designed to reflect a real-world use case: retrieving the latest Arabic-language news articles from the official website of the College of Medicine, University of Baghdad, and translating them faithfully into English. This dual challenge of accurate retrieval from a non-English institutional website and high-fidelity translation offers a rigorous benchmark for evaluating model capabilities. To enable consistent evaluation, both language models were prompted with the same standardized user instruction:

"*Access the Arabic-language official website of the College of Medicine – University of Baghdad (https://comed.uobaghdad.edu.iq/), navigate to the news section, and retrieve the three most recent full-length news articles (excluding announcements or events with only dates). Extract the full text content of each article, including the headline and publication date, and translate them into English with high fidelity, preserving the informational tone and structure of the original.*"

Model performance was evaluated along four dimensions:

1- Accuracy: The degree to which the translation reflected the source content.
2- Timeliness: The model's ability to capture the most recent updates.
3- Relevance: Focus on news items pertinent to the institution.
4- Translation quality: Clarity, readability, and semantic fidelity of the English output.

The comparative assessment emphasized the incorporation of real-time retrieval functionality to perform specialized, multilingual, and time-sensitive tasks effectively.

5. RESULTS

The experimental findings, derived from the proposed methodology, reveal a clear disparity in performance between the two language models when evaluated using the standardized prompting statement described earlier. ChatGPT-Plus demonstrated effective performance in accessing the designated URL, retrieving authentic Arabic-language news content, and translating it fluently into English. This outcome highlights the model's robust integration of real-time web search capabilities with a cross-lingual translation function.

In contrast, DeepSeek R1 either lacked direct access to the web or was architecturally constrained to provide only general suggestions, such as outlining generic steps for information retrieval, rather than performing the retrieval itself. This limitation manifested in two critical deficiencies: 1) the inability to directly access or navigate online content, and 2) the absence of a retrieval-augmented generation (RAG) framework, indicating that the model could not incorporate dynamic information beyond its static training corpus.

The comparative performance of these models was benchmarked against Google Translate, with evaluation criteria encompassing accuracy (alignment with official published content), timeliness (ability to access the most recent updates), relevance (focusing on institution-specific information), and translation quality (linguistic fidelity and coherence) as in Table 1. ChatGPT-Plus, particularly its GPT-4-turbo variant (ChatGPT-o1), demonstrated proficiency across all metrics, as evidenced in Table 2. By contrast, DeepSeek R1, despite its purported optimization for multilingual tasks, was unable to retrieve structured institutional news content from a non-English academic source, ultimately diminishing its reliability in this context, as in Table 3.

Table 1: Comparative performance of ChatGPT-Plus and DeepSeek (R1)

Criteria	ChatGPT-Plus	DeepSeek (R1)
Accuracy (Matching official Site)	High: faithful to source meaning and style; context preserved	Low: did not retrieve specific content; generic responses provided
Timeliness (Latest Articles)	High: retrieved and translated the three latest news articles accurately	Very Low: failed to retrieve current news articles
Relevance (Pertaining to College)	High: targeted retrieval of relevant institutional news	Low: focused on methodological suggestions, not specific content
Translation Quality (Linguistic Coherence and Fidelity)	High: fluent, coherent, and contextually nuanced translations	Not Applicable: did not perform translation task

Table 2: Summary of ChatGPT-Plus performance in retrieval and translation of institutional news

Aspect	Details
Retrieved Articles	- 99th Cohort First-Year Student Welcome Ceremony (November 12, 2024) - 20th International Scientific Conference (November 29, 2024) - Medical Research Collaboration Agreement Signing (December 5, 2024)
Translation Performance	- Accurate reflection of original content - Timely retrieval of recent news - High relevance to the College of Medicine, University of Baghdad - Maintained linguistic coherence and structural fidelity

Table 3: Summary of DeepSeek (R1) performance in retrieval and translation of institutional news

Aspect	Details
Retrieval Capability	- Failed to retrieve specific news articles - Provided methodological guidance instead of actual content
Translation Output	- Suggested translation tools without performing translations - Did not produce English versions of the Arabic articles
Observed Limitations	- Focused on general methodological advice rather than targeted retrieval - Inadequate real-time web indexing - Absence of retrieval-augmented generation (RAG) mechanisms

5. DISCUSSION

This study underscores the necessity of integrating advanced retrieval methods into LLM frameworks, suggesting that future models should prioritize dynamic data access and structured indexing techniques. Additionally, institutions should adapt their online content structures to facilitate easier parsing and utilization by AI-driven retrieval systems. Such strategic improvements can significantly enhance the reliability and utility of LLMs for specialized, real-world translation tasks, particularly in academic and institutional contexts.

6. CONCLUSION

This study demonstrates that ChatGPT-o1 outperforms DeepSeek R1 in retrieving and translating news from an Arabic website (i.e., College of Medicine-University of Baghdad). The observed discrepancies highlight differences in web accessibility, data indexing, and real-time processing capabilities between the two models. While ChatGPT-o1 effectively retrieved and translated content, DeepSeek R1 struggled to provide actual news articles, instead focusing on generic methodologies for information retrieval.

CONFLICT OF INTEREST

The authors declare that there is *no* conflict *of interest* regarding the publication of this paper.

REFERENCES

[1] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," arXiv (Cornell University), Jan. 2020, doi: 10.48550/arxiv.2012.15723.

[2] X. Chen, T. Liu, P. Fournier-Viger, B. Zhang, G. Long, and Q. Zhang, "A fine-grained self-adapting prompt learning approach for few-shot learning with pre-trained language models," Knowledge-Based Systems, vol. 299, p. 111968, Jun. 2024, doi: 10.1016/j.knosys.2024.111968

[3] W. Lu, R. K. Luu, and M. J. Buehler, "Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities," Npj Computational Materials, vol. 11, no. 1, Mar. 2025, doi: 10.1038/s41524-025-01564-y.

[4] A. Matarazzo and R. Torlone, "A Survey on Large Language Models with some Insights on their Capabilities and Limitations," arXiv (Cornell University), Jan. 2025, doi: 10.48550/arxiv.2501.04040.

[5] L. Zangari, C. M. Greco, D. Picca, and A. Tagarelli, "A survey on moral foundation theory and pre-trained language models: current advances and challenges," AI & Society, Mar. 2025, doi: 10.1007/s00146-025-02225-w.

[6] Z. Cao, K. Wong, and C.-T. Lin, "Weak human preference supervision for deep reinforcement learning," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 12, pp. 5369–5378, Jun. 2021, doi: 10.1109/tnnls.2021.3084198.

[7] R. Lou, K. Zhang, and W. Yin, "Large Language model instruction following: A survey of progresses and challenges," Computational Linguistics, pp. 1–43, Jun. 2024, doi: 10.1162/coli_a_00523.

[8] V. Iyer, P. Chen, and A. Birch, "Towards effective disambiguation for machine translation with large language models," in *Proc. Conf. Mach. Transl.*, 2023, pp. 482–495, doi: 10.18653/V1/2023.WMT-1.44.

[9] A. Toral, S. Castilho, K. Hu, and A. Way, "Attaining the unattainable? Reassessing claims of human parity in neural machine translation," in *Proc. WMT 2018 - 3rd Conf. Mach. Transl.*, vol. 1, pp. 113–123, 2018, doi: 10.18653/V1/W18-6312.

[10] S. A. Al Amer, M. G. Lee, and P. Smith, "Comparative Evaluation of Machine Translation Models Using Human-Translated Social Media Posts as References: Human-Translated Datasets," Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025), pp. 1–9, 2025, doi: 10.18653/v1/2025.loresmt-1.1.

[11] P. Savcı and B. Das, "Enhancing Text Summarization: Evaluating Transformer-Based Models and the Role of Large Language Models like ChatGPT," 2023 4th International Informatics and Software Engineering Conference (IISEC), pp. 1–4, Dec. 2023, doi: 10.1109/iisec59749.2023.10391040.

[12] S. H. Koenig and S. S. Hashemi, "Fine-tuning for Lesson Planning," 2024, Accessed: Jul. 01, 2025. [Online]. Available: https://gupea.ub.gu.se/handle/2077/83633.

[13] Y. K. Dwivedi, T. Malik, L. Hughes, and M. A. Albashrawi, “Scholarly discourse on GenAI’s impact on academic publishing,” *J. Comput. Inf. Syst.*, Dec. 2024, doi: 10.1080/08874417.2024.2435386.

[14] V. Du Preez et al., “From bias to black boxes: understanding and managing the risks of AI – an actuarial perspective,” British Actuarial Journal, vol. 29, p. e6, Apr. 2024, doi: 10.1017/S1357321724000060.

[15] J. C. L. Chow and K. Li, “Ethical Considerations in Human-Centered AI: Advancing Oncology Chatbots Through Large Language Models.,” JMIR Bioinform Biotech, vol. 5, no. 1, p. e64406, Nov. 2024, doi: 10.2196/64406.

[16] G. Fragiadakis, C. Diou, G. Kousiouris, and M. Nikolaidou, “Evaluating Human-AI Collaboration: A review and Methodological framework,” arXiv (Cornell University), Jul. 2024, doi: 10.48550/arxiv.2407.19098.

[17] S. Mirza et al., “Global-Liar: Factuality of LLMs over Time and Geographic Regions,” ArXiv, p. arXiv:2401.17839, Jan. 2024, doi: 10.48550/ARXIV.2401.17839.

[18] Z. Qi *et al.*, “AI and cultural context: An empirical investigation of large language models’ performance on Chinese social work professional standards,” *J. Soc. Social Work Res.*, Dec. 2024, doi: 10.1086/735590.

[19] DeepSeek-AI et al., “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,” Jan. 2025, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2501.12948.

[20] G. Mondillo *et al.*, “Comparative evaluation of advanced AI reasoning models in pediatric clinical decision support: ChatGPT O1 vs. DeepSeek-R1,” *medRxiv*, p. 2025.01.27.25321169, Jan. 2025, doi: 10.1101/2025.01.27.25321169.

[21] A. David Mikhail *et al.*, “Performance of DeepSeek-R1 in ophthalmology: An evaluation of clinical decision-making and cost-effectiveness,” *medRxiv*, p. 2025.02.10.25322041, Feb. 2025, doi: 10.1101/2025.02.10.25322041.

[22] . Neha and D. Bhati, “A Survey of DeepSeek Models,” Authorea Preprints, Feb. 2025, doi: 10.36227/TECHRXIV.173896582.25938392/V1.

[23] J. Zheng et al., “Fine-tuning Large Language Models for Domain-specific Machine Translation,” Feb. 2024, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2402.1506.

[24] O. M. Alyasiri, Y. N. Cheah, H. Zhang, O. M. Al-Janabi, and A. K. Abasi, “Text classification based on optimization feature selection methods: A review and future directions,” *Multimed. Tools Appl.*, pp. 1–47, Jul. 2024, doi: 10.1007/S11042-024-19769-6.

[25] O. M. Al-Janabi, N. H. A. H. Malim, and Y. N. Cheah, “Aspect categorization using domain-trained word embedding and topic modelling,” *Lect. Notes Electr. Eng.*, vol. 619, pp. 191–198, 2020, doi: 10.1007/978-981-15-1289-6_18.

[26] O. M. Al-Janabi, N. H. A. H. Malim, and Y.-N. Cheah, “Unsupervised model for aspect categorization and implicit aspect extraction,” *Knowl. Inf. Syst.*, vol. 64, no. 6, pp. 1625–1651, 2022, doi: 10.1007/s10115-022-01678-5.

[27] O. M. Alyasiri, Y. -N. Cheah, A. K. Abasi and O. M. Al-Janabi, "Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review," in IEEE Access, vol. 10, pp. 39833-39852, 2022, doi: 10.1109/ACCESS.2022.3165814.

[28] H. N. Abosaooda, S. B. Ariffin, O. M. Alyasiri, and A. A. Noor, “Evaluating the Effectiveness of AI Tools in Mathematical Modelling of Various Life Phenomena: A Proposed Approach,” InfoTech Spectrum: Iraqi Journal of Data Science , vol. 2, no. 1, pp. 16–25, Jan. 2025, doi: 10.51173/ijds.v2i1.16.

[29] A. Hendy et al., "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation," Feb. 2023, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2302.09210.

[30] T. Wang et al., "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?," Proc Mach Learn Res, vol. 162, pp. 22964–22984, Apr. 2022, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2204.05832.

[31] A. Rahman et al., "Comparative Analysis Based on DeepSeek, ChatGPT, and Google Gemini: Features, Techniques, Performance, Future Prospects," Feb. 2025, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2503.04783.

[32] W. Lai, M. Mesgar, and A. Fraser, "LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback," Jun. 2024, Accessed: Jul. 01, 2025. [Online]. Available: https://arxiv.org/pdf/2406.01771.