npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00935-4

Benchmarking large language models GPT-4o, llama 3.1, and qwen 2.5 for cancer genetic variant classification

Kuan-Hsun Lin[1,2], Tzu-Hang Kao[3], Lei-Chi Wang[3,4], Chen-Tsung Kuo[1,2], Paul Chih-Hsueh Chen[3,4], Yuan-Chia Chu[1,2,5] ✉ & Yi-Chen Yeh[3,4] ✉

Classifying cancer genetic variants based on clinical actionability is crucial yet challenging in precision oncology. Large language models (LLMs) offer potential solutions, but their performance remains underexplored. This study evaluates GPT-4o, Llama 3.1, and Qwen 2.5 in classifying genetic variants from the OncoKB and CIViC databases, as well as a real-world dataset derived from FoundationOne CDx reports. GPT-4o achieved the highest accuracy (0.7318) in distinguishing clinically relevant variants from variants of unknown clinical significance (VUS), outperforming Qwen 2.5 (0.5731) and Llama 3.1 (0.4976). LLMs demonstrated better concordance with expert annotations for variants with strong clinical evidence but exhibited greater inconsistencies for those with weaker evidence. All three models showed a tendency to assign variants to higher evidence levels, suggesting a propensity for overclassification. Prompt engineering significantly improved accuracy, while retrieval-augmented generation (RAG) further enhanced performance. Stability analysis across 100 iterations revealed greater consistency with the CIViC system than with OncoKB. These findings highlight the promise of LLMs in cancer genetic variant classification while underscoring the need for further optimization to improve accuracy, consistency, and clinical applicability.

In the era of precision medicine, cancer genetic testing has become essential in guiding the treatment of cancer patients. Multigene next-generation sequencing (NGS), and even comprehensive genomic profiling using large NGS panels, are now recommended to thoroughly assess genomic alterations in tumors[1–3]. Evidence suggests that patients managed with NGS-informed targeted treatments achieve superior outcomes across multiple tumor types[4]. Consequently, tumor genomic profiling using NGS has increasingly become a standard of care in clinical practice.

One significant challenge in implementing NGS testing is the complexity of interpreting NGS results. Large NGS panels, such as comprehensive genomic profiling panels, often detect numerous genomic alterations, many of which are rare and unfamiliar to clinicians or oncologists. To enhance the understanding of NGS results and maximize the utility of NGS testing, it is recommended that identified genomic alterations be annotated with their pathogenicity and clinical actionability in the clinical reports[5,6]. For pathogenicity classification, the five-tier system recommended by ClinGen, CGC, and VICC categorizes variants based on their biological and functional impact as pathogenic, likely pathogenic, variant of unknown significance, likely benign, or benign group[7]. For clinical actionability, the AMP/ASCO/CAP classification categorizes variants into four tiers: Tier I for variants of strong clinical significance, Tier II for variants of potential clinical significance, Tier III for variants of unknown clinical significance (VUS), and Tier IV for benign or likely benign variants. Other widely used frameworks include the ESMO Scale for Clinical Actionability of Molecular Targets (ESCAT), the OncoKB level of evidence classification, and the Clinical Interpretation of Variants in Cancer (CIViC) evidence levels[5,8–10]. These classification systems play a crucial role in standardizing the assessment of genomic alterations in oncology. By integrating these frameworks into NGS reports, genomic alterations can be presented in a structured and standardized format, improving report readability and facilitating interpretation by clinicians and oncologists.

[1]Department of Information Management, Taipei Veterans General Hospital, Taipei, Taiwan, ROC. [2]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan, ROC. [3]Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC. [4]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, ROC. [5]Big Data Center, Taipei Veterans General Hospital, Taipei, Taiwan, ROC. ✉e-mail: xd.yuanchia@gmail.com; lordaaa@gmail.com

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

Although variant classification systems are effective and widely used in clinical practice, accurate classification remains a complex challenge, requiring a comprehensive evaluation of biological and functional evidence, medical literature, clinical trial data, treatment guidelines, and FDA approvals. The process is highly expertise-dependent and labor-intensive. Variant knowledge databases, whether proprietary or public, can significantly aid this process. However, maintaining up-to-date databases with the latest literature is an equally demanding task, requiring expert input and manual effort. This is exemplified by the OncoKB and CIViC databases, which rely on specialized committees and expert crowdsourcing, respectively[8,9]. Furthermore, variant classification is inherently subjective, with studies showing significant interobserver variability among experts[11–13].

Artificial Intelligence (AI), particularly large language models (LLMs), holds significant potential in enhancing the variant classification process. AI has already proven its value across various medical domains, including disease diagnosis, clinical decision support, medical image analysis, and the automation of medical record processing[14,15]. LLMs excel at managing and analyzing vast quantities of unstructured data, such as medical literature, patient records, and clinical reports. Their strength lies in their ability to process large volumes of information quickly, enabling efficient understanding and summarization of key insights. In the field of variant classification, LLMs can be leveraged to analyze vast amounts of medical literature, clinical trial data, and treatment guidelines to assess the clinical significance of genetic alterations. This capability holds great potential for automating and continuously updating genetic variant classifications.

Several previous studies have explored the use of LLMs in genetic variant interpretation. For instance, Lu et al. utilized LLMs for variant annotation and demonstrated that retrieval-augmented generation (RAG) and fine-tuning can enhance performance[16]. Paoli et al. developed a generative AI assistant, VarChat, which facilitates efficient literature retrieval and summarization from PubMed based on gene symbols and genomic variants[17]. While these studies have highlighted the potential of LLMs in supporting variant annotation and interpretation, their utility in assisting genetic variant classification, particularly in terms of clinical significance or actionability scale, remains largely unexplored. To address this gap, this study aims to evaluate the effectiveness of LLMs in classifying genetic variants based on clinical significance or actionability, compare the performance of different models, and analyzes their respective strengths and limitations.

Results

Performance of LLMs in distinguishing between clinically relevant variants and VUS

To assess the ability of LLMs to distinguish clinically relevant variants from VUS, we analyzed 10,506 genetic variants from NGS testing reports of 612 patients who underwent FoundationOne CDx testing at our hospital. In these reports, genetic variants are categorized as either clinically relevant (listed in the Genomic Findings section of the reports, $n = 5240$) or (VUS, listed in the APPENDIX: Variants of Unknown Significance section of the reports, $n = 5266$).

We instructed the LLMs to classify genetic variants using the CIViC classification system, following the system prompts detailed in Supplementary Table 3 (Variant classification (CIViC level of evidence system)—Basic prompt). Variants classified within CIViC levels A to E were grouped as clinically relevant, while those categorized as "VUS" were considered VUS.

The performance of the LLMs in classifying variants from the FoundationOne dataset is summarized in Table 1 and Fig. 1. GPT-4o demonstrated significantly higher accuracy (0.7318) compared to Qwen 2.5 (0.5731) and Llama 3.1 (0.4976). Figure 2A presents the confusion matrix of LLM classifications compared to the ground truth annotations from the FoundationOne CDx report. The figure highlights distinct behavioral differences among the LLMs. Both Llama 3.1 and Qwen 2.5 tended to overcall VUS as clinically relevant variants, while GPT-4o did not exhibit this bias. Conversely, while GPT-4o accurately classified 94.1% of VUS variants, it misclassified nearly half of the clinically relevant variants as VUS, suggesting a more conservative approach in classifying variants as clinically relevant. The distribution of LLM classifications based on the CIViC level of evidence system for ground truth clinically relevant variants and VUS is shown in Supplementary Fig. 1. Among VUS misclassified as clinically relevant by LLMs, most were assigned to CIViC levels C to E, which correspond to weak or indirect clinical evidence.

Additionally, we evaluated accuracy in cases where all three LLMs agreed on the classification versus those with discordant results. Each LLM's answer was determined by its most frequently selected response across all iterations. Among the 10,506 genetic variants, 2766 (26.3%) had identical classifications across all three LLMs. In these cases, the three-model consensus achieved a high accuracy of 0.9732 (Supplementary Table 4).

The stability of the LLMs' responses in the FoundationOne dataset across 100 iterations is shown in Fig. 2B. All three LLMs exhibited high consistency ratios across queries, with the majority of queries achieving a consistency ratio above 90%, indicating that the same answer was provided in over 90% of iterations.

To compare LLM results with human experts, we randomly selected 100 variants from the FoundationOne dataset and asked three pathologists to classify them as either clinically relevant variants or VUS. The agreement between FoundationOne annotations, LLM responses, and pathologists' classifications is visualized in Supplementary Fig. 2. The high inter-pathologist agreement suggests strong consistency among human experts. Notably, GPT-4o aligns more closely with the pathologists than other AI models. In contrast, Llama 3.1 and Qwen 2.5 show lower agreement with both the pathologists and FoundationOne, indicating greater variability in their classifications. These findings suggest that GPT-4o provides more concordant assessments, whereas other AI models deviate more significantly from both human judgment and the FoundationOne annotations.

Performance of LLMs in classifying genetic variants into evidence tiers: analysis of variants from the OncoKB database

The performance of the LLMs in classifying variants from the OncoKB database based on the OncoKB level of evidence system is summarized in Table 1 and Fig. 1. Among the three LLM models (GPT-4o, Llama 3.1, and Qwen 2.5), GPT-4o achieved the highest top-1 accuracy of 0.3393, while Llama 3.1 showed the lowest top-1 accuracy at 0.3066. In cases where all three LLMs agreed on the top-1 classification (43.7% of variants), the accuracy increased to 0.4286. (Supplementary Table 4).

In terms of top-2 and top-3 accuracy, Qwen 2.5 performed the best, with scores of 0.4357 and 0.4567, respectively. In contrast, GPT-4o and Llama 3.1 showed only modest improvements in top-2 and top-3 accuracy. (Supplementary Fig. 3 and Supplementary Table 5). This can be attributed to the tendency of both GPT-4o and Llama 3.1 to provide only a single answer in 87.2% and 93.9% of cases, respectively, compared to just 5.2% for Qwen 2.5, despite explicit instructions in the system prompts to generate up to three answers.

Figure 3A shows the confusion matrix of LLM classification (top-1 prediction only) compared with the ground truth expert annotation in the OncoKB database. As depicted in the figure, we observed that for variants with the highest clinical evidence (OncoKB level 1), the LLMs provided more accurate classifications. In contrast, for variants with weaker clinical evidence (OncoKB levels 2, 3, and 4), the LLMs exhibited greater discordance with the ground truth expert annotation. However, most of the discordance fell within one degree of difference (e.g., misclassifying level 2 variants as level 3). For OncoKB levels associated with drug resistance (R1 and R2), the LLMs often misclassified these variants as levels 1–4 instead of R1 or R2. Notably, GPT-4o showed better performance in accurately classifying more R1 variants compared to other LLMs.

The stability of the LLMs' responses across 100 iterations is shown in Fig. 3B. As illustrated, Qwen 2.5 demonstrated significantly higher stability compared to GPT-4o and Llama 3.1. For most queries, Qwen 2.5 achieved a consistency ratio above 90%, whereas GPT-4o and Llama 3.1 exhibited a broader distribution of consistency ratios across queries.

Table 1 | Classification accuracy of LLMs (GPT-4o, Llama 3.1, Qwen 2.5) based on top-1 answer in the FoundationOne, OncoKB, and CIViC datasets

	GPT-4o		Llama 3		Qwen 2.5		
	Mean accuracy	95% CI	Mean accuracy	95% CI	Mean accuracy	95% CI	*p* value
Foundation one	0.7318	0.7307–0.7329	0.4976	0.4974–0.4978	0.5731	0.5725–0.5736	<0.001
OncoKB	0.3393	0.3369–0.3417	0.3066	0.3041–0.309	0.3328	0.3316–0.334	<0.001
CIViC	0.1865	0.1857–0.1874	0.1212	0.1205–0.1219	0.2485	0.2477–0.2492	<0.001

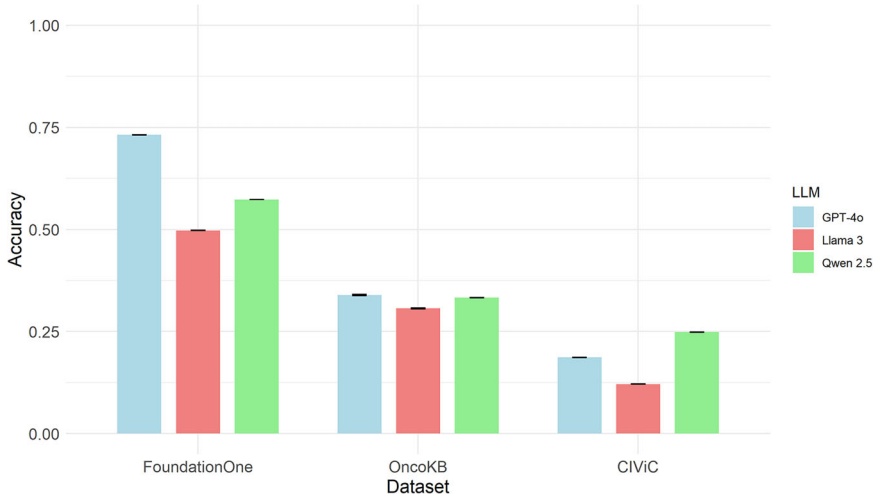

Fig. 1 | Classification accuracy comparison among GPT-4o, Llama 3.1, and Qwen 2.5 across the FoundationOne, OncoKB, and CIViC datasets.

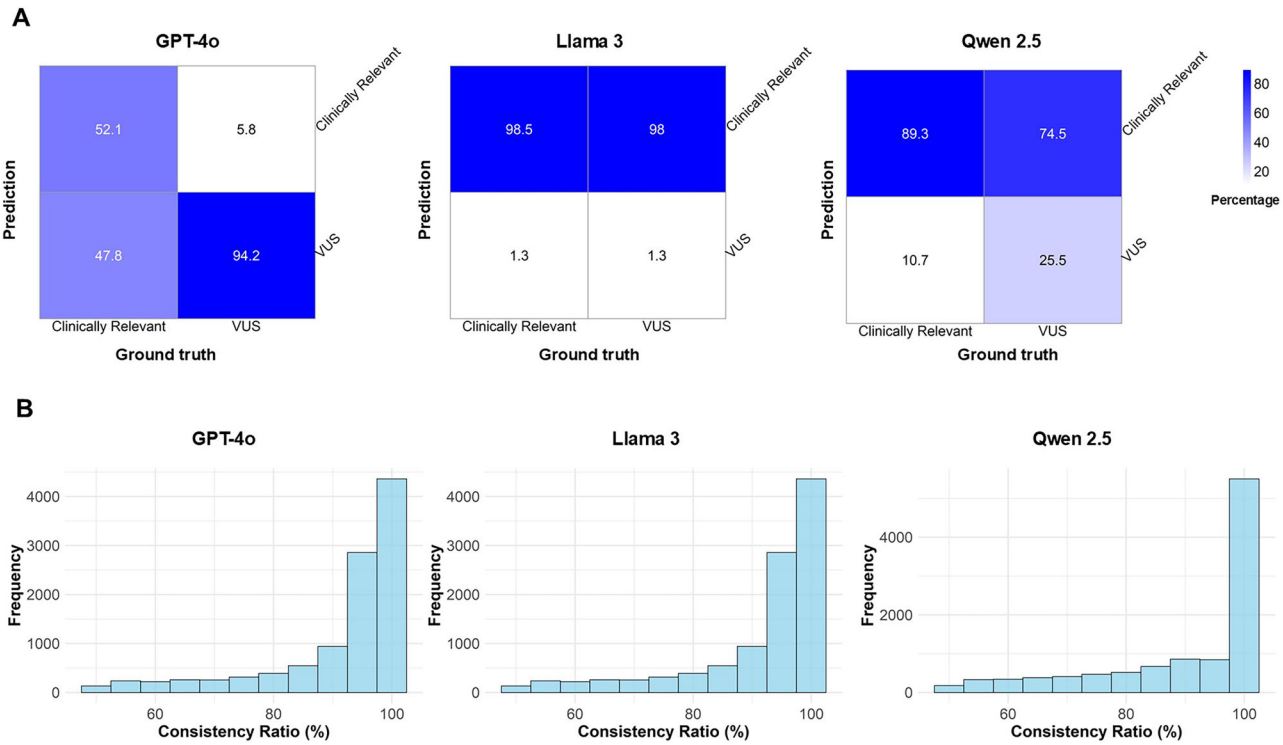

Fig. 2 | Performance and consistency of LLM classification on the FoundationOne dataset. A Confusion matrix of LLM classification compared with the ground truth in the FoundationOne dataset. **B** Consistency ratio of the LLMs' responses in the FoundationOne dataset across 100 iterations.

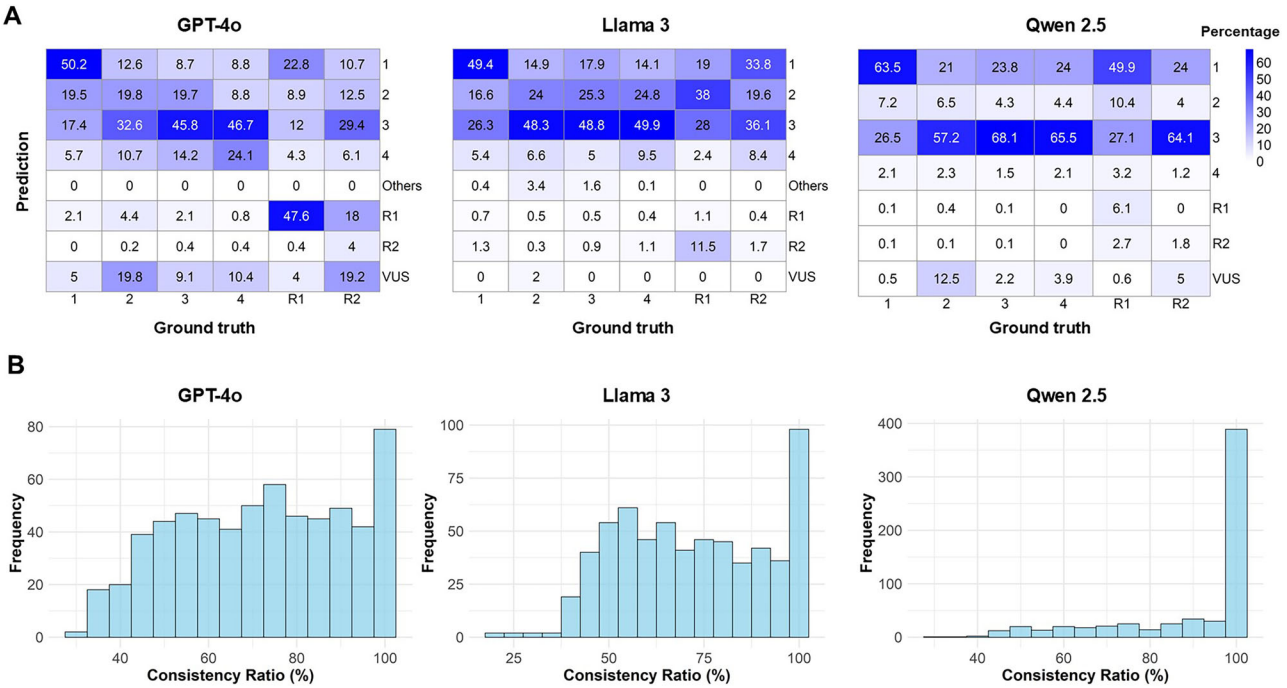

Fig. 3 | Comparing LLM classification against expert annotations in the OncoKB dataset. A Confusion matrix showing the top-1 classification predictions of GPT-4o, Llama 3.1, and Qwen 2.5 compared to expert annotations for genetic variants in the OncoKB dataset. **B** Consistency ratio of LLMs' responses across 100 iterations in the OncoKB dataset.

Performance of LLMs in classifying genetic variants into evidence tiers: analysis of variants from the CIViC database

The performance of LLMs in classifying variants from the CIViC database is summarized in Table 1 and Fig. 1. Among the three models, Qwen 2.5 achieved the highest top-1 accuracy (0.2485), while GPT-4o (0.1865) and Llama 3.1 (0.1212) performed lower in comparison. For top-2 and top-3 accuracy, Qwen 2.5 again outperformed the others, with mean accuracies of 0.5476 and 0.6992, respectively. GPT-4o and Llama 3.1 showed only modest improvements in these metrics. (Supplementary Fig. 4 and Supplementary Table 6) This trend is similar to what was observed in the OncoKB dataset, where both GPT-4o and Llama 3.1 predominantly provided a single response (GPT-4o: 97.2%, Llama 3.1: 91.5%, Qwen 2.5: 6.4%), limiting their top-2 and top-3 accuracy gains. Unlike the trends observed in the FoundationOne and OncoKB datasets, accuracy did not improve when all three LLMs agreed on the classification in the CIViC dataset (Supplementary Table 4).

Figure 4A shows the confusion matrix of LLM classification (top-1 prediction only) compared with the ground truth annotation in the CIViC database. As depicted in the figure, we observed that for variants with the highest clinical evidence (CIViC level A), all three LLMs provided the most accurate classifications. In contrast, for variants with weaker clinical evidence (CIViC levels B, C, D, and E), the LLMs exhibited greater discordance with the ground truth expert annotation. In addition, we observed that LLMs tend to classify variants at higher CIViC levels, such as classifying level B variants as level A, or level C variants as either level B or A.

The stability of the LLMs' responses in the CIViC dataset across 100 iterations is presented in Fig. 4B. Compared to their stability in the OncoKB dataset, the LLMs' responses in the CIViC dataset demonstrated significantly higher consistency ratios across queries, with most queries achieving a consistency ratio above 90%.

LLM reasoning for classification

To gain a deeper understanding of the LLM's rationale for classifying genetic variants, we also examined its explanations for selected cases. Supplementary Table 7 provides examples of these LLM responses. In these examples, we identified several factors that contributed to the LLMs' misclassifications. For instance, some LLMs appeared unaware of recent FDA approvals. In the case of KRAS G12C in non-small cell lung cancer, GPT-4o recognized its FDA approval and correctly classified it as OncoKB level 1. By contrast, Qwen 2.5 and Llama 3.1 seemed unaware of this approval and misclassified the variant as level 3A, resulting in under-classification. Moreover, LLMs do not always account for the specific details of genetic alterations, as illustrated by the NTRK1 Q570* variant. This variant is best classified as VUS due to limited data on its biological impact and clinical significance, and there is no evidence indicating that it would result in NTRK1 activation, unlike NTRK1 fusions. However, Llama 3.1 mistakenly concluded that the variant leads to constitutive activation of the NTRK1 kinase and misclassified it as CIViC level A. Meanwhile, GPT-4o considered this NTRK1 alteration likely to respond to TRK inhibitors and misclassified it as CIViC level B. Both misclassifications resulted in over-classification. Likewise, although MSH6 K1358fs*2 is a frameshift mutation, it is relatively common in the population and is predicted to be nonpathogenic because it only deletes two amino acids in the C-terminal region, potentially leaving the protein function intact[18]. Nevertheless, all three LLMs classified this variant as pathogenic, assuming it would lead to microsatellite instability (MSI-high), and thus overclassified it as CIViC level B.

Exploring factors affecting LLMs' performance and behavior

Finally, we explored various factors that may influence LLM performance and behavior. First, we refined the original basic system prompts, which simply instructed the LLMs to provide the classification level number. The revised prompts specified the LLM's role, objectives, scope, behavior, and expected input/output format in greater detail (Supplementary Table 3). To assess the impact of these refinements, we compared the performance of the Qwen 2.5 model using basic versus refined prompts (Figs. 5 and 6). In the OncoKB and CIViC datasets, top-1 accuracy decreased with refined prompts (from 0.3328 to 0.2994 in OncoKB and from 0.2485 to 0.1722 in CIViC). However, in the FoundationOne dataset, refined prompts led to a substantial accuracy improvement (from 0.5731 to 0.7246). Additionally, as shown in the confusion matrix, Qwen 2.5 with basic prompts tended to overcall VUS as clinically relevant variants. This tendency was significantly reduced with refined prompts. Instead, nearly half of the clinically relevant

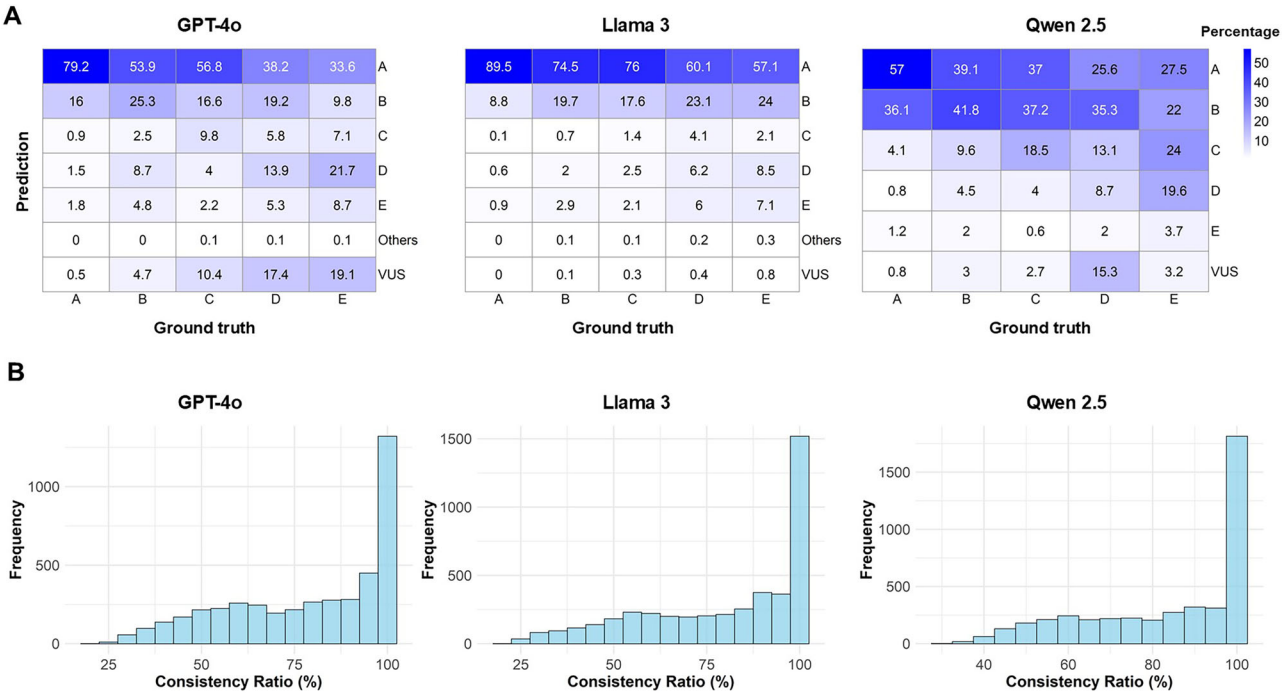

Fig. 4 | Comparing LLM classification against expert annotations in the CIViC dataset. A Confusion matrix showing the top-1 classification predictions of GPT-4o, Llama 3.1, and Qwen 2.5 compared to expert annotations for genetic variants in the CIViC dataset. **B** Consistency ratio of LLMs' responses across 100 iterations in the CIViC dataset.

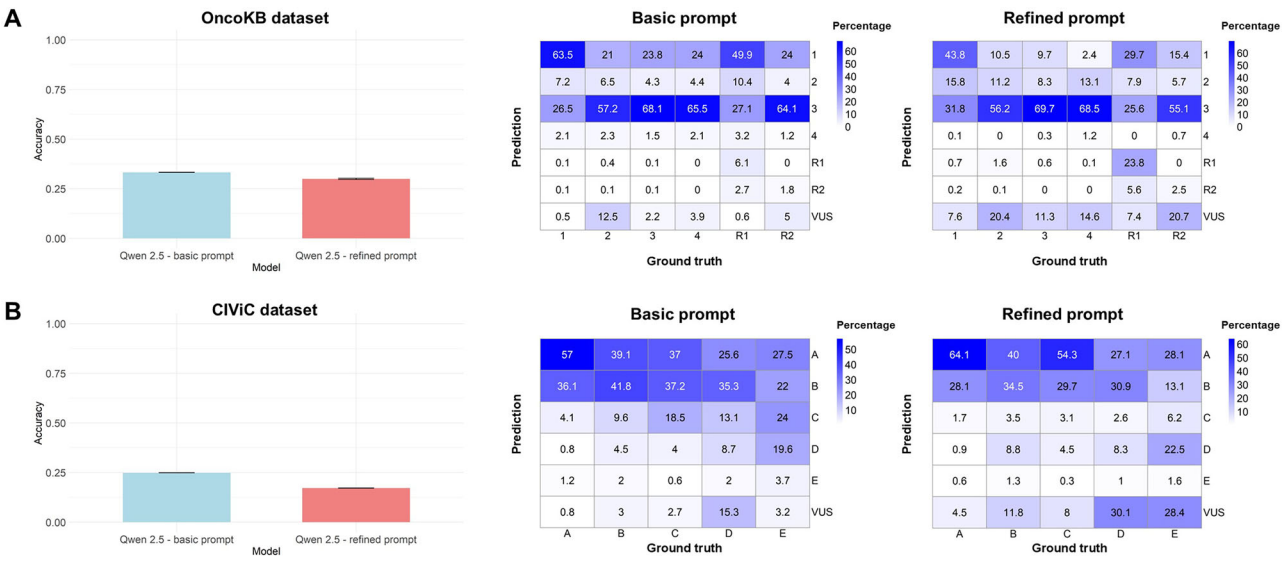

Fig. 5 | Performance comparison of the Qwen 2.5 model using basic and refined prompts. A Results on the OncoKB dataset. **B** Results on the CIViC dataset.

variants were classified as VUS, suggesting that refined prompts encourage a more conservative classification approach.

Next, we evaluated the impact of using a binary classification prompt instead of the original CIViC level of evidence prompt for classifying clinically relevant variants versus VUS in the FoundationOne dataset. Unlike the CIViC prompt, which instructs LLMs to provide a classification based on CIViC evidence levels, the binary classification prompt directly asks the LLMs to categorize each variant as either clinically relevant or VUS (see Supplementary Table 3, Variant classification - Binary classes). As shown in Fig. 6, the binary classification prompt slightly improved accuracy from 0.5731 to 0.6119. Additionally, similar to the refined prompt, it resulted in a more conservative classification approach, with a tendency to classify variants as VUS.

We also explored the potential of RAG to enhance LLM performance by integrating data from the CIViC database and FDA oncology drug approval information. As shown in Fig. 6, implementing RAG significantly improved classification accuracy from 0.5731 to 0.6616 in the FoundationOne dataset.

Finally, we evaluated the impact of model temperature settings on the stability and consistency of LLM responses using Llama 3.1 on the OncoKB dataset. As shown in Fig. 7, lowering the model temperature significantly improved response stability, achieving a 100% consistency ratio when the temperature was set to 0. Interestingly, reducing the model temperature also led to a slight improvement in accuracy, with top-1 accuracy increasing from 0.3066 at a temperature of 0.8 to 0.3178 at 0.4 and 0.3312 at 0 (Fig. 7).

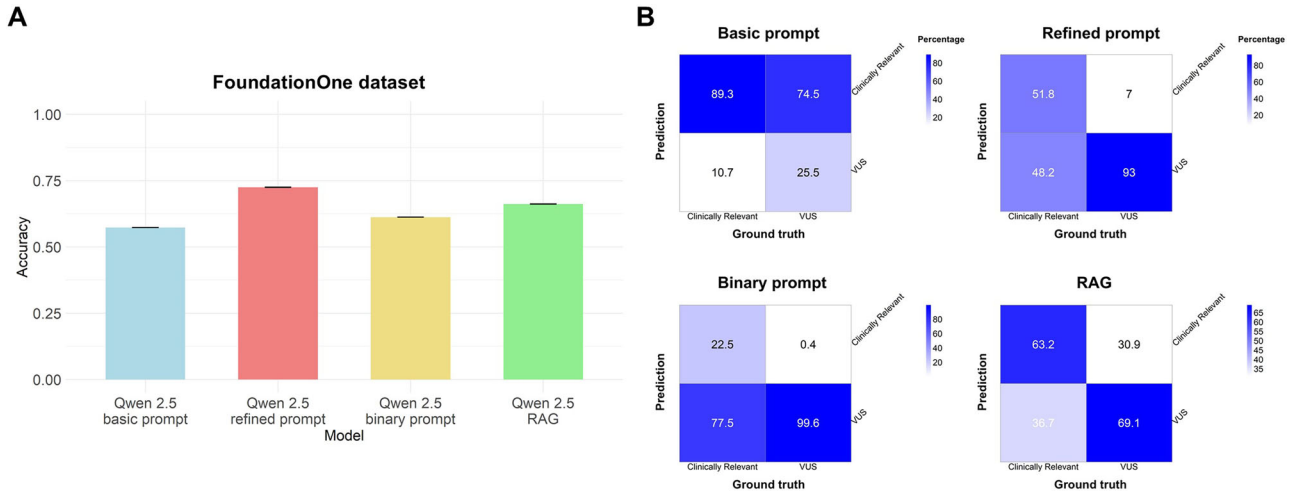

Fig. 6 | Performance comparison of the Qwen 2.5 model across different prompt configurations in the FoundationOne dataset. A Accuracy comparison using basic, refined, and binary prompts, as well as Retrieval-Augmented Generation (RAG). **B** Confusion matrix illustrating the classification outcomes for each prompt condition.

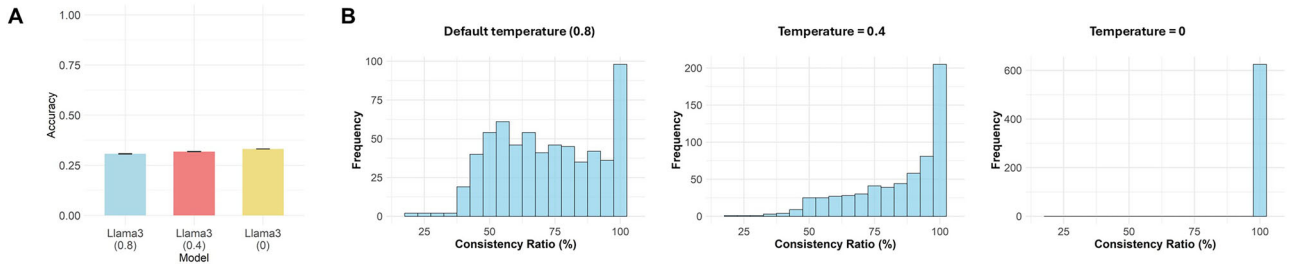

Fig. 7 | Effect of temperature settings on LLM performance and consistency. Accuracy (**A**) and response consistency (**B**) of Llama 3.1 in the OncoKB dataset across 100 iterations. Results are shown for temperature settings of 0.8 (default), 0.4, and 0.

Table 2 | Model accuracy under different prompt configurations and temperature settings

Model	Model conditions	Dataset	Accuracy	Best-in- group
Qwen2.5	Basic prompt + Default temperature (0.8)	FoundationOne	0.5731	
Qwen2.5	Refined prompt + Default temperature (0.8)	FoundationOne	0.7246	✓
Qwen2.5	Binary class prompt + Default temperature (0.8)	FoundationOne	0.6119	
Qwen2.5	RAG + Default temperature (0.8)	FoundationOne	0.6616	
Qwen2.5	Basic prompt + Default temperature (0.8)	OncoKB	0.3328	✓
Qwen2.5	Refined prompt + Default temperature (0.8)	OncoKB	0.2994	
Qwen2.5	Basic prompt + Default temperature (0.8)	CIViC	0.2485	✓
Qwen2.5	Refined prompt + Default temperature (0.8)	CIViC	0.1722	
Llama 3.1	Basic prompt + Default temperature (0.8)	OncoKB	0.3066	
Llama 3.1	Basic prompt + temperature (0.4)	OncoKB	0.3178	
Llama 3.1	Basic prompt + temperature (0)	OncoKB	0.3312	✓

The best-performing configuration for each model–dataset group is indicated with a check mark (✓).

Taken together, these results demonstrate that LLM performance can be substantially influenced by prompt design, classification scheme, and model temperature. Table 2 summarizes the comparative performance of LLMs under these varying configurations across different datasets, highlighting the best-performing conditions for each model–dataset pair.

Discussion

The use of LLMs as a supportive tool for clinical practice demonstrates significant potential[14,19]. Previous studies have shown that LLMs can achieve physician-level performance on medical board examinations[20,21]. However, their capabilities in specialized medical domains requiring high expertise, such as genetic variant classification, remain largely unexplored. This study provides the first systemic analysis of LLMs in the context of classifying cancer genetic variants based on their clinical relevance and actionability. Our primary objective is to assess the baseline performance of LLMs for this task, without any specialized training or reinforcement, aiming to minimize bias. This research highlights the unique strengths and limitations of each model, offering valuable insights into their potential applications in precision medicine.

Regarding stability of responses, we observed that LLMs may exhibit different behaviors depending on the variant classification system used, i.e., OncoKB versus CIViC. Both GPT-4o and Llama 3.1 displayed significantly higher variability (lower consistency ratios) and tended to provide differing responses across query iterations when using the OncoKB classification system compared to the CIViC system. One possible explanation is that the system prompt for the CIViC classification system includes examples for each variant category in addition to the category definitions themselves (see Supplementary Table 3). This additional context may provide clearer guidance for LLMs, enabling them to better determine the appropriate category for a given variant and reduce uncertainty in classification. However, it is noteworthy that Qwen 2.5 does not exhibit this variability even when using the OncoKB system prompt, highlighting inherent differences between the LLMs.

One potential factor influencing this variability is the model temperature setting. The default temperature values set by each LLM API may have influenced the observed differences in classification consistency. Since higher temperatures introduce more randomness in model outputs, future studies could investigate whether adjusting temperature settings improves classification stability while maintaining accuracy. Systematic evaluation of model behavior at different temperature settings would help identify the optimal balance between stability and flexibility in clinical applications.

In terms of response accuracy in classifying variants into different tiers of clinical significance, we observed that LLMs demonstrated higher accuracy for variants with the strongest clinical evidence, while showing lower accuracy for those with weaker clinical evidence. This finding aligns with a previous multi-institutional interrater agreement evaluation among human experts in cancer genetic variant classification, which revealed that the greatest disagreement occurred in distinguishing between AMP/ASCO/CAP variant categorization tiers II (Variants of Potential Clinical Significance) and III (VUS)[13]. Notably, we also observed that when using the CIViC classification system, all three LLMs tended to assign variants to higher levels of evidence, such as classifying level B variants as level A, or classifying level C variants as either level B or A. This is an important observation, as it suggests a tendency toward overclassification, which may impact the clinical interpretation and actionability of the variants.

In the real-world dataset derived from FoundationOne CDx NGS reports, it is noteworthy that GPT-4o achieved an accuracy of 0.7318 in distinguishing between clinically relevant variants and VUS, without any specialized training or reinforcement. In comparison, the recent AMP VITAL Somatic Challenge, a variant interpretation challenge involving 134 human expert participants, reported that 86% (range: 54%–94%) of responses correctly distinguished clinically significant variants from other variants[12]. Although the current accuracy of LLMs is not yet comparable to that of human experts, further adjustments or training could enhance their performance. Previous studies have demonstrated that RAG and fine-tuning can enhance LLM performance in genetic variant annotation and interpretation, with RAG outperforming fine-tuning[16]. RAG combines the strengths of both pre-trained language models and external knowledge sources (such as PubMed), enabling the LLMs to retrieve and incorporate relevant information from external databases or documents[22]. This approach is particularly well-suited for specialized tasks requiring advanced domain knowledge, such as genetic variant classification. In our study, we also evaluated the impact of RAG on LLM performance and found that implementing RAG in Qwen 2.5 significantly improved classification accuracy in the FoundationOne dataset. Further studies are needed to more comprehensively evaluate the effectiveness of RAG in improving LLM performance, particularly in terms of accuracy, consistency, and the ability to integrate dynamic, up-to-date scientific knowledge and clinical evidence.

Our study also demonstrated that system prompt design significantly influences LLM performance and behavior in cancer genetic variant classification. In the FoundationOne dataset experiments, refined prompts with detailed instructions led to a substantial accuracy improvement compared to basic prompts. Moreover, refined prompts resulted in a more conservative classification approach, with a tendency to classify variants as VUS. In contrast, the basic prompt was more likely to overcall VUS as clinically relevant variants. A potential explanation is that the refined prompt explicitly defined the LLM's role as an expert assistant specializing in cancer genetic variant classification, reinforcing a more cautious and evidence-based decision-making approach.

By examining the LLM's rationale for classifying genetic variants, we noted some LLMs appeared unaware of recent FDA approvals. One potential factor that may contribute to the observation is the data cut-off of each LLM. Since LLMs rely on fixed training datasets, their ability to accurately classify newly validated genetic variants or FDA-approved targeted therapies is inherently constrained by the recency of their training data. Among the models evaluated, GPT-4o's training data cut-off was in October 2023, Llama 3.1's was in December 2023, while Qwen 2.5's exact cut-off date has not been disclosed. However, these data cut-off dates do not fully explain our case study findings on KRAS G12C in non-small cell lung cancer. GPT-4o recognized its FDA approval and correctly classified it as OncoKB Level 1, while Llama 3.1 misclassified the variant as a lower evidence tier, seemingly unaware of its approval. This discrepancy is unexpected, as the FDA approved KRAS G12C for non-small cell lung cancer in December 2022, well before the training data cut-off dates of both GPT-4o and Llama 3.1. Therefore, factors beyond training data recency—such as differences in data sources, weighting of biomedical knowledge, or reasoning mechanisms—may contribute to variations in LLM performance.

Updating data sources remains a significant challenge in clinical genomics, as genetic variant interpretation continuously evolves with new research findings and regulatory updates. To address this, integrating external knowledge retrieval and leveraging frequently updated models can help bridge the gap between static training data and dynamic, real-world clinical knowledge. One promising approach involves utilizing tools like Deep Research (https://openai.com/index/introducing-deep-research/), which can perform in-depth, multi-step research across the internet to gather the latest information and generate comprehensive summaries. These tools could enable the autonomous updating of knowledge bases with the most current evidence, refining variant classifications based on emerging literature, clinical guidelines, and regulatory changes. This would help mitigate the lag between static datasets and real-world advancements in genomics. Additionally, implementing quality control mechanisms—such as cross-referencing newly retrieved data against established databases (e.g., CIViC, OncoKB) or expert review—is critical to ensuring reliability. Ultimately, a hybrid approach that combines AI-assisted knowledge retrieval, expert evaluation, and structured database updates could help maintain accurate, clinically relevant, and up-to-date genomic interpretations.

Our study has several limitations. First, the datasets used do not cover the full spectrum of cancer genetic variants. The OncoKB and CIViC datasets include only clinically relevant variants, while the FoundationOne dataset contains clinically relevant variants and VUS. Neither dataset includes benign or likely benign variants. Consequently, we were unable to assess LLM performance in distinguishing benign or likely benign variants from other classifications. Future studies should address this limitation by incorporating cancer genetic variant datasets with ground truth annotations for benign and likely benign somatic variants. Second, in evaluating LLM performance in classifying variants from the OncoKB and CIViC databases, it cannot be excluded that current models were at least partially trained on these publicly available datasets, potentially influencing the results. Nevertheless, our findings indicate that even for variants documented in these datasets, current LLMs still have substantial room for improvement in accurately stratifying variants into evidence tiers. Third, due to the lack of large-scale ground truth datasets, we did not evaluate the performance of LLMs using two other widely adopted cancer genetic variant classification systems: the AMP/ASCO/CAP variant categorization and the ESCAT framework. As our findings suggest that the performance of LLMs may vary depending on the classification system used, it is imperative that future efforts focus on assessing the capabilities of LLMs within these alternative frameworks. Lastly, the three LLMs evaluated in our study are general-purpose models rather than biomedical-specialized LLMs. Investigating

whether biomedical-specialized models, such as BioMedLM and OpenBioLLM, could further enhance performance in genetic variant classification would be highly valuable[23,24].

In conclusion, this study shows that LLMs hold promise for assisting in the classification of cancer genetic variants. Each LLM has unique strengths and weaknesses that must be considered for clinical use. Future research should focus on refining these models to improve accuracy and consistency across various datasets, thereby enhancing their applications in clinical genomics.

Methods

Dataset

We included three datasets to evaluate the performance of cancer genetic variant classification by LLMs. The first two datasets were obtained from public available database, OncoKB and CIViC. The third dataset was compiled from the real-world NGS testing reports of FoundationOne CDx assay.

For the OncoKB dataset, we downloaded the variant clinical implications data table from the OncoKB website (https://www.oncokb.org/actionable-genes, last accessed: 2024/11/20). The table consists of five columns: Level, Gene, Alterations, Cancer Types, and Drugs. The Level column indicates the evidence level of a specific variant, which is listed in the Gene (e.g., ABL1) and Alterations (e.g., BCR-ABL1 Fusion) columns, within a particular cancer type (e.g., B-Lymphoblastic Leukemia/Lymphoma) as annotated by the expert committee according to the OncoKB classification of evidence levels[8]. In total, there are 625 variant associations, including 182, 154, 114, 80, 34, and 61 associations with evidence levels of Level 1, 2, 3, 4, R1, and R2, respectively. (Supplementary Table 1)

For the CIViC dataset, we downloaded the Clinical Evidence Summary data table from the CIViC website (https://civicdb.org/releases/main, last accessed: 2024/11/20). This table provides detailed information on the clinical impact of variants, evidence statements, and data sources. We retrieved three columns of data: molecular_profile, disease, and evidence_level. The molecular_profile column contains the gene and alterations, such as JAK2 V617F. The disease column specifies the cancer type, such as Lymphoid Leukemia. The evidence_level column includes the CIViC evidence level, annotated through expert crowdsourcing and reviewed by expert editors[9]. In total, there are 4426 variant associations, including 166, 1465, 1547, 1216, and 32 associations with evidence levels of A, B, C, D, and E, respectively. (Supplementary Table 2)

To evaluate the performance of LLMs in genetic variant classification within real-world clinical settings, we extracted variant data from FoundationOne CDx assay NGS reports for 612 patients at our hospital. The variants were categorized into two groups: (a) Clinically Relevant: Genetic alterations listed in the Genomic Findings section of the report. (b) Variants of Unknown Clinical Significance (VUS): Genetic alterations listed in the APPENDIX: Variants of Unknown Significance section of the report. In total, the dataset included 10,506 genetic alterations, comprising 5,240 clinically relevant alterations and 5266 VUS.

Model selection

The models selected for this study included Qwen 2.5 (72B), Llama 3.1 (70B), and GPT-4o (version 2024-05-13). Qwen 2.5 and Llama 3.1 were chosen as they represent state-of-the-art open-source models, widely recognized for their performance and accessibility in the research community[25,26]. GPT-4o, representing proprietary models, was included to facilitate a comparison with non-open-source models that excel in diverse language tasks[27]. This selection enables a comprehensive evaluation of open-source and closed-source models, emphasizing their differences in classification accuracy and clinical genomics applicability. Among the models tested, GPT-4o's training data cut-off is in October 2023, while Llama 3.1's data extends to December 2023. Qwen 2.5 (72B) has not publicly disclosed a precise data cut-off, with only its release date of September 2024 available.

Model temperature settings followed the default values specified by each LLM API: 0.8 for Llama 3.1 and Qwen 2.5, and 1.0 for GPT-4o. Additionally, to assess the impact of temperature settings, we conducted extra tests on Llama 3.1 in the OncoKB dataset with the temperature set to 0.4 and 0, and compared the results with its default setting.

System prompts

To ensure consistency, system prompts were provided to each LLM, as detailed in Supplementary Table 3. The system prompts included definitions of the levels of evidence for the OncoKB and CIViC systems, along with an additional category, "VUS." This was added alongside the classification categories from the OncoKB and CIViC levels of evidence to evaluate whether the LLMs could accurately distinguish clinically relevant variants from VUS.

For the CIViC system, the CIViC website provided a variant example for each evidence level category (https://civic.readthedocs.io/en/latest/model/evidence/level.html, last accessed: 2024/11/20). These examples were incorporated into the system prompts. In contrast, the OncoKB website (https://www.oncokb.org/therapeutic-levels, last accessed: 2024/11/20) does not provide example variants for evidence levels; thus, these were not included in the prompts for OncoKB.

To examine the impact of different system prompt designs, we tested multiple prompt sets. The basic system prompts simply instructed the LLMs to provide the corresponding classification level number, without additional context. In contrast, the refined system prompts included detailed instructions specifying the LLM's role, objectives, scope, behavior, and expected input/output format. Both versions explicitly instructed the LLMs to assign classification levels (e.g., 1, 2, 3, 4, R1, R2 for OncoKB; A, B, C, D, E for CIViC), allowing up to three levels per response. Additionally, for the FoundationOne dataset, we tested a binary classification prompt in which LLMs classified gene variants as either "Clinically Relevant" or "VUS", instead of assigning detailed evidence levels.

Testing framework design

We developed a structured testing framework that applies a consistent method across all datasets. Queries were generated based on the OncoKB, CIViC, and FoundationOne datasets, using gene names, alterations, and tumor types as contextual inputs. Each query was formulated as a natural language prompt, providing sufficient context for the LLMs to classify genetic variants into predefined levels (e.g., A, B, C).

Example query: "Given the gene EGFR, with alteration L858R in the context of non-small cell lung cancer, what is the appropriate classification?"

All queries were automatically generated using customized Python scripts to maintain a standardized format. We sent each query to the LLMs via APIs, including GPT-4o (via Azure OpenAI), Qwen 2.5 (72B), and Llama 3.1 (70B). The responses were systematically recorded and subsequently analyzed to assess the models' classification accuracy.

To assess the robustness and variability of LLM responses, we conducted multiple iterations for each experiment—100 iterations for basic system prompt experiments and 10 iterations for all others. This approach allowed us to evaluate each model's stability and consistency across different query iterations.

Execution environment and hardware specifications

To support the batch testing of LLMs, we established a high-performance hardware and software environment. Four NVIDIA A100 GPUs were utilized to host the Ollama server, enabling high-performance computing for LLM inference tasks related to Qwen 2.5 (72B) and Llama 3.1 (70B). The Azure OpenAI GPT-4o API was used to evaluate GPT-4o, providing direct cloud-based inference capabilities. Python 3.10.12 was used to implement the testing framework, ensuring compatibility and stability for the batch processing scripts.

Retrieval-augmented generation (RAG) implementation

We implemented RAG to assess its potential in improving the accuracy and performance of LLMs for cancer genetic variant classification. RAG integrates retrieval mechanisms with generative AI, allowing LLMs to

dynamically access external knowledge during inference rather than relying solely on their pre-trained corpus. This approach enhances the model's ability to classify genetic variants by incorporating newly retrieved evidence.

To build an institutional knowledge base, we retrieved datasets from the *Clinical Evidence Summary*, *Features Summary*, and *Molecular Profiles* tables from the CIViC database (https://civicdb.org/releases/main, last accessed: 2024/11/20), as well as FDA *Oncology/Hematologic Malignancies Approval* information (https://www.fda.gov/drugs/resources-information-approved-drugs/oncology-cancerhematologic-malignancies-approval-notifications, last accessed: 2025/2/13). These datasets were converted into high-dimensional vector representations using Nomic-Embed-Text[28], allowing for efficient retrieval and contextual augmentation. When an LLM processes a genetic variant classification query, the system retrieves relevant information from the knowledge base and supplies it as additional context, improving the model's classification accuracy. To evaluate the impact of RAG, we compared the performance of Qwen 2.5 (72B) on the FoundationOne dataset with and without RAG.

Human evaluation of cancer genetic variants

To compare the classification of cancer genetic variants between LLMs and human experts, we randomly selected 100 variants from the FoundationOne dataset. Three pathologists—who regularly sign out molecular reports and actively participate in the molecular tumor board—independently classified each variant as either "Clinically Relevant" or "VUS". The concordance between pathologists, LLMs, and the ground truth annotations in the FoundationOne dataset was then evaluated.

Data analysis and statistics

The accuracy of LLMs in classifying genetic variants was assessed by comparing the LLMs' classifications with the ground truth annotations in these datasets. Top-N accuracy measured how often the correct answer appears within a model's N choices. Top-1 accuracy required the correct answer to be the top choice, while top-2 and top-3 accuracy considered a response correct if the correct answer appeared within the top two or three choices, respectively. This metric is particularly useful when multiple plausible answers exist. Mean accuracy, along with the 95% confidence intervals, was calculated. Accuracy comparisons between different LLM models were analyzed using an ANOVA test, with a p-value of <0.05 considered statistically significant.

The stability of LLM-generated responses across multiple iterations is assessed using the consistency ratio. This metric quantifies the uniformity of responses to a given query by calculating the proportion of times the most frequently selected answer appears relative to the total number of responses. A higher consistency ratio indicates greater stability, while a lower ratio suggests higher variability. For this analysis, only Top-1 answers were used to compute the metric.

Data availability

The OncoKB and CIViC datasets used for the present study are available through OncoKB (https://www.oncokb.org/) and CIViC (https://civicdb.org/) websites, respectively. The FoundationOne dataset used for the present study is not publicly available but is available from the corresponding author on reasonable request.

Code availability

The underlying code used for conducting LLM experiments in this study is available on GitHub and can be accessed via this link. (https://github.com/gslin1224/LLMs-CancerVariant).

Received: 18 January 2025; Accepted: 2 May 2025;
Published online: 15 May 2025

References

1. Mosele, M. F. et al. Recommendations for the use of next-generation sequencing (NGS) for patients with advanced cancer in 2024: a report from the ESMO Precision Medicine Working Group. *Ann. Oncol.* **35**, 588–606 (2024).
2. Chakravarty, D. et al. Somatic genomic testing in patients with metastatic or advanced cancer: ASCO provisional clinical opinion. *J. Clin. Oncol.* **40**, 1231–1258 (2022).
3. Mosele, F. et al. Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: a report from the ESMO Precision Medicine Working Group. *Ann. Oncol.* **31**, 1491–1505 (2020).
4. Gibbs, S. N. et al. Comprehensive review on the clinical impact of next-generation sequencing tests for the management of advanced cancer. *JCO Precis Oncol.* **7**, e2200715 (2023).
5. Li, M. M. et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J. Mol. Diagn.* **19**, 4–23 (2017).
6. van de Haar, J. et al. ESMO Recommendations on clinical reporting of genomic test results for solid cancers. *Ann. Oncol.* **35**, 954–967 (2024).
7. Horak, P. et al. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for Cancer Consortium (VICC). *Genet Med.* **24**, 986–998 (2022).
8. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol*. **17**, 00011 (2017).
9. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
10. Mateo, J. et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Ann. Oncol.* **29**, 1895–1902 (2018).
11. Lebedeva, A. et al. Multi-institutional evaluation of interrater agreement of biomarker-drug pair rankings based on the ESMO scale for clinical actionability of molecular targets (ESCAT) and sources of discordance. *Mol. Diagn. Ther*. **29**, 91–101 (2024).
12. Li, M. M. et al. Assessments of somatic variant classification using the Association for Molecular Pathology/American Society of Clinical Oncology/College of American Pathologists Guidelines: a report from the Association for Molecular Pathology. *J. Mol. Diagn.* **25**, 69–86 (2023).
13. Sirohi, D. et al. Multi-institutional evaluation of interrater agreement of variant classification based on the 2017 Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists Standards and guidelines for the interpretation and reporting of sequence variants in cancer. *J. Mol. Diagn.* **22**, 284–293 (2020).
14. Pressman, S. M. et al. Clinical and surgical applications of large language models: a systematic review. *J. Clin. Med*. **13**, 3041 (2024).
15. Yu, P. et al. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. *Healthcare* **11**, 2776 (2023).
16. Lu, S. & Cosgun, E. Boosting GPT models for genomics analysis: generating trusted genetic variant annotations and interpretations through RAG and Fine-tuning. *Bioinform. Adv.* **5**, vbaf019 (2025).
17. De Paoli, F. et al. VarChat: the generative AI assistant for the interpretation of human genomic variations. *Bioinformatics* **40**, btae183 (2024).
18. Hirotsu, Y. et al. Multigene panel analysis identified germline mutations of DNA repair genes in breast and ovarian cancer. *Mol. Genet. Genom. Med.* **3**, 459–466 (2015).
19. Omar, M. et al. Large language models in medicine: a review of current clinical trials across healthcare applications. *PLOS Digit Health* **3**, e0000662 (2024).
20. Katz, U. et al. GPT versus resident physicians — a benchmark based on official board scores. *NEJM AI.* **1**, AIdbp2300192 (2024).

21. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* **2**, e0000198 (2023).
22. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NIPS'20: In *Proc*. 34th *International Conference on Neural Information Processing Systems* 9459–9474 (Curran Associates Inc., 2020).
23. Ankit, Pal M. S. OpenBioLLMs: advancing open-source large language models for healthcare and life sciences. *Hugging Face Repository*. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B (2024).
24. Bolton, E. et al. BioMedLM: a 2.7B parameter language model trained on biomedical Text. arXiv:2403.18421; (2024).
25. Yang, A. et al. Qwen2 technical report. arXiv preprint arXiv:2407.10671. (2024).
26. Dubey, A. et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783. (2024).
27. Achiam, J. et al. Gpt-4 technical report. arXiv preprint arXiv:*2303.08774*. (2023).
28. Nussbaum, Z., Morris J. X., Duderstadt, B., Mulyar, A. Nomic embed: training a reproducible long context text embedder. arXiv preprint arXiv:2402.01613. (2024).

Acknowledgements

This study was funded by National Science and Technology Council, Taiwan [grant numbers NSTC 112-2320-B-075-004-MY3] and Taipei Veterans General Hospital, Taiwan [grant number V114E-004-3]. We also appreciate the computational resources provided by TVGH Cloud 1. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

K.H.L. conceived and planned the project, set up the system, performed the experiments, and was a major contributor in writing the manuscript. T.H.K. and L.C.W. conducted the pathologists' evaluation of cancer genetic variants. C.T.K. contributed to the system setup. C.H.C. contributed to the study design and data collection. Y.C.C. contributed to the study design, system setup, and funding acquisition. Y.C.Y. conceived and planned the project, acquired funding, performed data analysis, conducted pathologists' evaluation of cancer genetic variants, and was a major contributor to writing the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This study was approved by the Taipei Veterans General Hospital Institutional Review Board (2024-12-010BC), which waived the requirement for informed consent. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00935-4.

Correspondence and requests for materials should be addressed to Yuan-Chia Chu or Yi-Chen Yeh.