Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Performance of Large Language Models in Nursing Licensure Examinations: A Systematic Review and Meta-Analysis

Isaac Amankwaa; PhD, MSN, RN

Isaac.amankwaaa@aut.ac.nz

School of Clinical Sciences, Auckland University of Technology, Auckland, New Zealand
https://orcid.org/0000-0002-7879-4731

Alex Odoom

2. Department of Medical Microbiology, University of Ghana Medical School, P. O. Box KB 4236, Korle Bu, Accra, Ghana.

Adams Kasim

3. Nursing and Midwifery Training College, Techiman, Krobo, Ghana

Emmanuel Kobiah

4. Department of Nursing, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Maximous Diebieri

5. Nursing and Midwifery Training College, Kpembe, Salaga, Ghana

Edward Appiah Boateng

4. Department of Nursing, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Sebastian Gyamfi

6. Faculty of Nursing, University of Windsor, Windsor, Ontario, Canada

Caz Hales

7. Faculty of Health, School of Nursing, Midwifery, and Health Practice, Victoria University of Wellington, Wellington, New Zealand

---

Systematic Review

Keywords: Artificial Intelligence, Large Language Models, Licensure, Nursing, Educational Measurement, Systematic Review, Nursing

Posted Date: January 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8671859/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Performance of Large Language Models in Nursing Licensure Examinations: A Systematic Review and Meta-Analysis

Authors

Isaac **Amankwaa**[1*] Alex **Odoom**[2] Adams **Kassim**[3] Emmanuel **Kobiah**[4] Maximous **Diebieri**[5] Edward Appiah **Boateng**[4] Sebastian **Gyamfi**[6] Caz **Hales**[7]

1. Department of Nursing, Auckland University of Technology, Auckland, New Zealand
2. Department of Medical Microbiology, University of Ghana Medical School, P. O. Box KB 4236, Korle Bu, Accra, Ghana.
3. Nursing and Midwifery Training College, Techiman, Krobo, Ghana
4. Department of Nursing, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
5. Nursing and Midwifery Training College, Kpembe, Salaga, Ghana
6. Faculty of Nursing, University of Windsor, Windsor, Ontario, Canada
7. Faculty of Health, School of Nursing, Midwifery, and Health Practice, Victoria University of Wellington, Wellington, New Zealand
8. University of Canberra, Nursing Department, University of Canberra, Bruce ACT 2617, Australia

Declaration of Interest Statement
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding
This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

***Corresponding author:**
Isaac Amankwaa, PhD, MSc, RN
Senior Nursing Lecturer
Department of Nursing
Auckland University of Technology
Auckland, New Zealand
Email: Isaac.amankwaa@aut.ac.nz
Ph (work): +64 - 09 921 9999 **ext** 31229

Registration

Registered with OSF (Doi: https://doi.org/10.17605/OSF.IO/RUJX2)

CRediT authorship contribution statement

Isaac Amankwaa: Conceptualisation, Writing -original draft & editing, Formal analysis, Data curation. **Alex Odoom:** Methodology, Writing, formal analysis, review and editing. **Adam Kassim**: Methodology, Writing – review and editing. **Emmanuel Kobiah:** Methodology, reviewing and editing. **Maximous Diebieri**: Methodology, Writing – review and editing. **Edward Boateng:** Writing -review and editing, supervision **Sebastian Gyamfi:** Writing -review and editing, supervision. **Caz Hales:** Writing -review and editing, supervision

Performance of Large Language Models on Nursing Licensure Examinations: A Systematic Review and Meta-Analysis

Abstract

Objectives: This systematic review and meta-analysis assessed the performance of large language models (LLMS) in nursing licensure examinations. Despite the increasing use of LLMS in healthcare education, their capabilities in nursing licensure examinations remain uncertain. This study provides evidence on the accuracy and limitations of LLMs to help guide their integration into nursing education and licensure.

Design: The systematic review and meta-analysis adhered to PRISMA 2020 guidelines.

Data sources: PubMed, CINAHL, PsycINFO, EMCARE, and ERIC were searched from April to June 2025.

Eligibility criteria: Studies were eligible if they evaluated LLMs (e.g., GPT-4, ChatGPT, Qwen-2.5) using multiple-choice nursing licensure questions under exam-like conditions and reported quantitative accuracy. Open-ended items were excluded from the meta-analysis but narratively synthesised

Review methods: Two reviewers independently screened, extracted data, and appraised the risk of bias. A random-effects meta-analysis estimated pooled accuracy; subgroup and meta-regression analyses explored heterogeneity.

Results: Twelve studies assessed 13,870 MCQs across five exam systems and eight LLMs. Pooled accuracy was 69.6% (95% CI: 65.6–73.6%) with substantial heterogeneity ($I^2 = 98\%$). GPT-4 outperformed GPT-3.5 (77.2% vs. 60.4%, six studies); domain-customised and newer models reached 93.6%. LLMs excelled in general medicine and pharmacology but underperformed in ethics and psychosocial integrity. Accuracy differed significantly by exam system ($p < 0.01$), but not by question difficulty ($p = 0.90$) or format ($p = 0.96$). Translated NCLEX-RN items reduced accuracy ($p = 0.03$); CNNLE was the only system with a significant positive effect ($p < 0.001$). Methodological variability and underreporting of model parameters were common.

Conclusions: LLMs show promise for low-stakes educational applications, such as formative assessments within hybrid teaching models; however, they are unsuitable for unmoderated, high-stakes licensure decisions due to inconsistent performance. Regulatory guidelines, equitable access, and nursing-specific model development are needed to ensure fairness and

validity. Research must prioritise standardised frameworks, error analysis, and broader geographic representation to address these limitations.
Key words: Artificial Intelligence, Large Language Models, Licensure, Nursing, Educational Measurement, Systematic Review, Nursing.

1 Introduction

Artificial Intelligence (AI) is transforming healthcare education (Bajwa et al., 2021; Wang et al., 2024) by delivering personalised learning and tailoring content to individual needs, while reducing barriers related to geography and cost (Singh et al., 2025). It manages repetitive administrative tasks, allowing health educators to focus more on building relationships (Amankwaa et al., 2025). Among AI's branches, generative AI (GenAI) marks a key advancement by allowing machines to produce human-like outputs (Bandi et al., 2023). A subset of GenAI includes large language models (LLMs), trained on vast text datasets (Lee, 2025). LLMs such as ChatGPT, Bard, and Qwen have gained attention for their use in summarisation, translation, tutoring, and exam preparation (Schlegel et al., 2025). This review explores LLMs as a distinct class of GenAI tools and evaluates their performance on nursing licensure exams.

Since the release of ChatGPT in late 2022, LLMs have drawn significant attention in nursing education for their potential to support personalised learning, clinical simulations, academic writing, and critical thinking (Amankwaa et al., 2025). Existing reviews highlight a strong interest in LLMs in nursing education (Hobensack et al., 2024; Zhou et al., 2024), with positive feedback on accessibility and affordability, while raising concerns about ethical issues and inconsistent results (Abujaber et al., 2023). These reviews consistently highlight the promising role of LLMs in enhancing nursing education. However, they also emphasise the urgent necessity for comprehensive evaluation methods, alignment with pedagogical goals, and ethical protections to ensure responsible implementation integration.

1.1 Background and rationale

While educational applications are gaining traction, a more pressing and underexplored question is whether LLMs are suitable for high-stakes assessments such as nursing licensure exams. This issue pertains not only to the utility but also to the validity of domain-specific knowledge and higher-order thinking. Licensure exams require structured reasoning, contextual interpretation, and the application of discipline-specific principles, including clinical judgement, metacognitive awareness, and critical thinking (Joseph et al., 2019; Simmons, 2010). These exams assess more than factual recall. They test cognitive processes such as perceiving patient conditions, selecting appropriate actions, and reflecting on outcomes (Mohammadi-Shahboulaghi et al., 2021). This is especially true in nursing, where the epistemological framework emphasises holistic reasoning, patient-centred communication, and complex clinical judgement (Thorne, 2014). These qualities set nursing apart from other health professions, such as medicine or pharmacy, and make it harder to assess LLM competence using general benchmarks or comparisons across disciplines (Gunawan et al., 2024).

Despite these unique requirements, no systematic review to date has specifically evaluated the performance of LLMs in nursing licensure examinations. Existing reviews have focused on medicine, dentistry, pharmacy, and other specialties (Bagde et al., 2023; Bongco et al., 2024; Liu et al., 2024). Only one review (Jin et al., 2024) included nursing, drawing on just two primary studies out of 23. In that review, nursing recorded the lowest LLM accuracy compared to pharmacy, medicine, and dentistry. The reported LLM performance in health licensing exams varies widely, from 18.3% to 100%, with high heterogeneity (I^2 = 87–96%), driven by differences in exam formats, question types, and model versions. These limitations raise concerns about the generalisability of existing findings to nursing, a discipline with unique cognitive, relational, and epistemological demands (Salviano et al., 2016). A targeted

review is needed to provide nursing educators, regulators, and researchers with accurate insights into the strengths and limitations of LLMs in the context of nursing licensure.

Another reason for discipline-specific synthesis is that common AI benchmarks, such as the Massive Multitask Language Understanding (MMLU) dataset and Medical Question Answering (MedQA), are often used to evaluate overall performance across disciplines (Yan et al., 2024). However, they lack alignment with the knowledge structures and reasoning processes needed for nursing licensure exams (Liu et al., 2023). These benchmarks focus on factual recall and broad medical content (Yan et al., 2024), and do not assess the affective, ethical, and situational judgement central to nursing practice (Lingle, 2024). LLM performance assessments must reflect the nursing discipline's pedagogical goals and epistemological foundations to evaluate LLMs meaningfully for nursing licensure use. Field-specific scrutiny is needed to determine whether LLMs meet the safety, empathy, and critical judgment standards required in high-stakes nursing exams.

This systematic review and meta-analysis aimed to evaluate the performance of LLMs on nursing licensure examinations across diverse global contexts. In doing so, it informs educators, policymakers, and regulars about the capabilities, limitations, and readiness of LLMs for integration into nursing programmes aimed at preparing students for high-stakes licensure exams. To the best of our knowledge, this is the first systematic review and meta-analysis to provide a comprehensive, field-specific synthesis of LLM performance and accuracy in nursing licensure contexts. It goes beyond aggregated accuracy to explore variations by model type, exam system, question format, and geographic setting. It lays the foundation for responsible implementation and future model development tailored to the needs of nursing education.

1.2 Aim and Questions

This systematic review and meta-analysis examined the accuracy of LLMs on nursing licensure exams. It focused on: a) pooled accuracy; b) performance across clinical domains; c) sensitivity to question formats; d) sources of heterogeneity by model, exam system, and region. The research questions were:

1. What types of nursing licensure exams and question formats (e.g., multiple-choice questions (MCQs), clinical vignettes) have been used to evaluate LLMs?
2. How accurate are LLMs on nursing licensure content, and how does performance vary by model version, exam system, and region?
3. Which clinical domains consistently show strong or weak LLM performance, and how does accuracy relate to question complexity?
4. How do methodological factors (e.g., sample size, confounder control, scoring thresholds) affect reported outcomes and study comparability?

2 Methodology

2.1 Design

This review followed the 2020 PRISMA guidelines (Page et al., 2021) and was prospectively registered with Open Science Framework (OSF) (blinded for review). Both narrative synthesis and meta-analysis were planned (see Supplementary Files S1a and S1b for the PRISMA checklist).

2.2 Eligibility criteria

This review employed the PICO framework (Eriksen & Frandsen, 2018) to define eligibility criteria. Included studies evaluated LLMs (e.g., GPT-4, ChatGPT, Qwen-2.5) in nursing licensure exams using MCQs. Studies had to simulate exam-like conditions, use LLMs to generate responses, and report quantitative outcomes, primarily the proportion of correct answers. Comparative analyses with other models or human benchmarks were useful but not

required. Studies must report sufficient data for synthesis (e.g., total and correct items). We excluded studies that lacked MCQs, focused on non-licensure uses, or did not report quantifiable performance. Open-ended questions were excluded from the meta-analysis but included in the narrative synthesis if relevant. Non-English studies, reviews, editorials, and those without full-text access were excluded.

2.3 Search Strategy and Study Retrieval

The search strategy aimed to identify empirical studies evaluating LLMs in nursing licensure exams. It combined three concept groups using Boolean operators: (1) LLM and AI terms (e.g., ChatGPT, GPT-4, Qwen, Bard); (2) nursing licensure terms (e.g., NCLEX, NNLE, nurse certification); and (3) performance outcomes (e.g., accuracy, evaluation, explanation quality). Controlled vocabulary (e.g., MeSH terms) was used where applicable, and strategies were adapted to each database's indexing system.

Searches were conducted in PubMed, PsycINFO, CINAHL, EMCARE, and ERIC between April and June 2025, limited to studies published from January 2022 onward to reflect the emergence of ChatGPT (De Angelis et al., 2023). No filters were applied during the database search; exclusions (e.g., non-English, reviews) occurred during screening. Weekly alerts were set up to capture new studies. All records were imported into Covidence for deduplication, screening, full-text review, and data extraction. The strategy was developed with a health sciences librarian, peer-reviewed, piloted, and detailed in Appendix A.

2.4 Study selection

Study screening and selection followed the PRISMA guidelines and were managed using Covidence. Two reviewers (MD and AK) independently screened all titles and abstracts for relevance, followed by full-text reviews of the potentially eligible studies. Any disagreements were resolved through discussion with a third reviewer (EK). The dual-review process was blinded and systematic, ensuring methodological transparency and minimising

bias (Stoll et al., 2019). Documentation of decisions, identification of conflicts, and adherence to inclusion criteria were facilitated by Covidence (Cleo et al., 2019).

2.5 Data extraction

Data were extracted using a pre-defined and piloted form within Covidence (Supplementary File S2). Two reviewers (SG, EB) independently and blindly extracted data. The primary outcome was the proportion of correct MCQ responses, reported as published. For studies assessing multiple models or conditions, data were extracted separately. Secondary outcomes included domain-specific accuracy, benchmark attainment, explanation quality, and error types. Open-ended or non-quantifiable outputs were excluded from meta-analysis but included in narrative synthesis.

Additional variables extracted included study year, LLM version, exam system, country, prompt strategy (e.g., zero-shot, chain-of-thought), input format (batch vs. individual), question source (official vs. practice), exam alignment (real vs. simulated), question type, and performance thresholds. Missing values (e.g., temperature, prompt type) were marked as 'not reported'; no assumptions or imputations were made. Covidence flagged discrepancies, which were resolved by consensus. If unresolved, a third reviewer (CH) adjudicated. All required data were available in the reports, so no authors were contacted.

2.6 Study risk of bias assessment

We assessed the risk of bias using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Analytical Cross-Sectional Studies. The tool consists of eight domains, each scored as 1 (criterion met) or 0 (not met, unclear, or not applicable), with a maximum possible score of 8 (Ma et al., 2020). Two reviewers (IA and AO) independently assessed each study. Discrepancies were resolved through discussion and consensus with a third reviewer (CH). Studies were rated as high risk if two or more domains scored 0; moderate risk if one domain

scored 0 or if 2 or more were unclear; and low risk if no domains were rated high risk. These ratings guided our interpretation and conclusions, but were not used to exclude studies. s.

2.7 Statistical analysis

A random-effects meta-analysis (DerSimonian and Laird method) was used to estimate pooled LLM accuracy on nursing licensure exams. The primary outcome was the proportion of correct responses on MCQs, reported with 95% confidence intervals. Subgroup differences, prediction intervals, and p-values supported interpretation. Heterogeneity was assessed using I^2, Tau^2, and Cochran's Q, with I^2 >75% indicating substantial heterogeneity.

All included studies reported MCQ-based accuracy and were eligible for pooling. Outcomes not suited to meta-analysis were synthesised narratively. To support subgroup comparisons, MCQ formats and subject domains were recoded using structured frameworks (Supplementary File S3), including a four-level cognitive complexity scheme (A1–A4) and a three-tier clinical reasoning model (Levels 1–3), informed by NCLEX-RN blueprints and educational taxonomies (Brady, 2019; Edwards, 2015).

Analyses included subgroup comparisons (e.g., model type, exam system, year, question format, domain, real vs simulated), meta-regression of study-level covariates, and leave-one-out sensitivity analysis. Reporting bias was assessed using funnel plots and Egger's test. All analyses were conducted in R (v4.3.3) using the meta and metafor packages. Results are presented in forest plots and subgroup figures, risk of bias, and study characteristics. No formal GRADE assessment was performed. Secondary outcomes (e.g., domain accuracy, model comparisons) were reported narratively.

3 Results

3.1 Search results

We identified 117 potential studies from the database and other source searches (**Error! Reference source not found.**). After removing 12 duplicates and screening for relevance, 105

studies were screened. Of these, 39 were sought for retrieval and assessed for eligibility. Twenty-seven studies were excluded during the eligibility assessment for the following reasons: incorrect study design (n = 12), being only abstracts or conference papers (n = 7), not involving nursing students or nurse educators (n = 6), and incorrect intervention (n = 2). Supplementary file S4 summarises the characteristics of excluded studies. Twelve studies were included in the final review and meta-analysis (García-Rudolph et al., 2024; García-Rudolph et al., 2025; Huang, 2023; Kaneda et al., 2023; Krumsvik, 2024; Miao, 2024; Su et al., 2024; Taira et al., 2023; Wu et al., 2024; Zhao et al., 2025; Zhu et al., 2025).

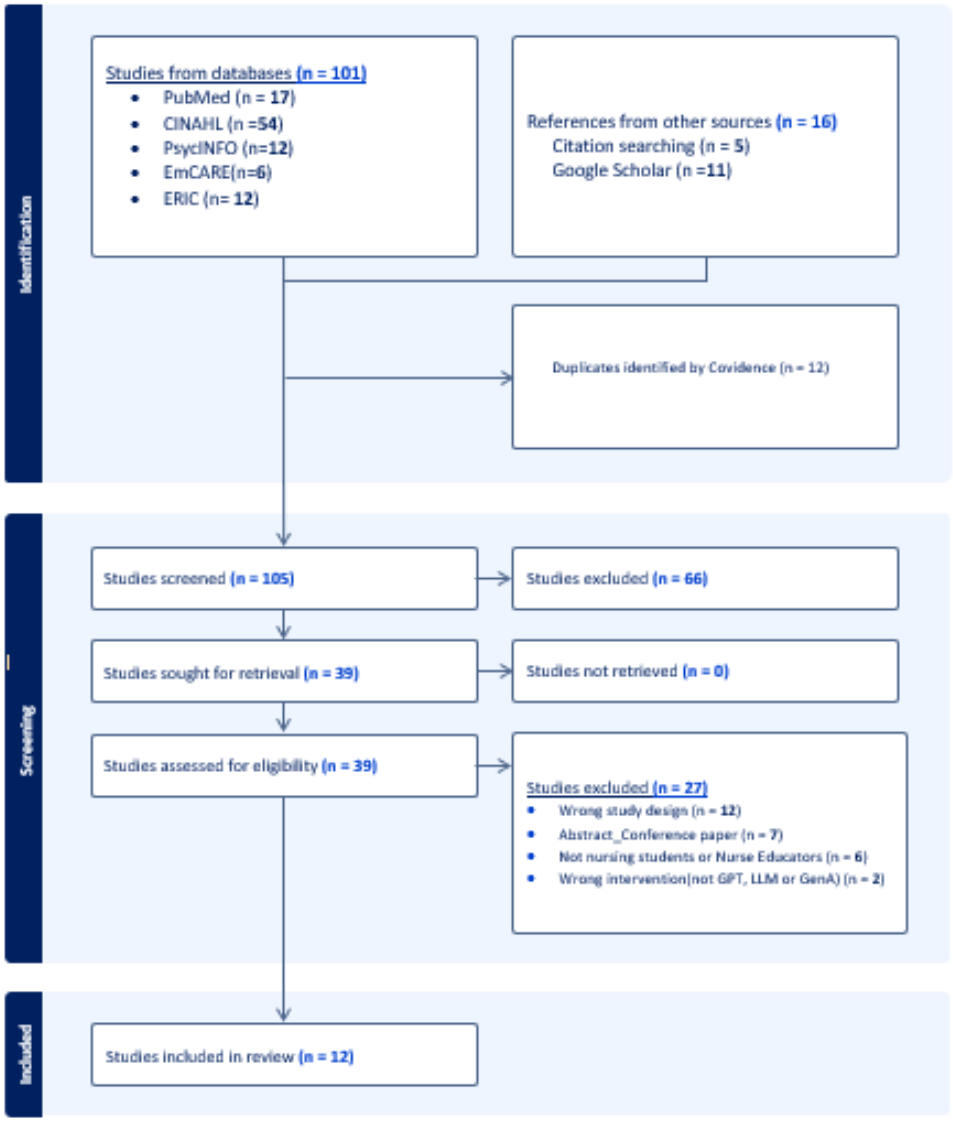


Figure 1: PRISMA flow diagram of study selection process

3.2 Risk of bias assessment (JBI)

The risk of bias assessment, using the JBI Critical Appraisal Checklist for Analytical Cross-Sectional Studies (Joanna Briggs Institute, 2017). Of the 12 studies, four studies (Miao, 2024; Su et al., 2024; Wu et al., 2024; Zhu et al., 2025) scored eight and were classified as having a low risk of bias. Six studies (Huang, 2023; Kaneda et al., 2023; Krumsvik, 2024; Taira et al., 2023; Zong et al., 2024) scored between 6 and 6.5 and were deemed to have a moderate risk. Two studies (García-Rudolph et al., 2024; García-Rudolph et al., 2025) scored 5 and were judged as high risk due to the absence of confounder management and reliance on descriptive statistics. The most frequent methodological weaknesses included failure to identify or adjust for confounding variables and limited statistical sophistication across several studies (Supplementary File S5: Quality assessment of included studies).

3.3 Characteristics of included studies

Table 1 summarises the 12 studies evaluating LLM performance. The studies showed geographic clustering, with nine from East Asia (China, Japan, Taiwan) and limited Western representation (Spain, Norway, two USA-linked studies). Cross-sectional designs were the most common (n = 4), followed by comparative validation (n = 2), comparative evaluation (n = 2), retrospective evaluation (n = 1), quantitative evaluation (n = 2), and one experimental study. Sample sizes ranged from fewer than 100 items (Krumsvik, 2024) to over 6,000 (Zong et al., 2024). Most studies (n = 8) assessed fewer than 1,000 items, with only two studies (Zhu et al., 2025; Zong et al., 2024) including over 1,000 MCQs. Collectively, the studies evaluated over 13,000 multiple-choice questions, though item counts varied markedly.

OpenAI models were predominant, with GPT-3.5 (n=8) and GPT-4 (n=4) being the most frequently assessed. Non-OpenAI models appeared less often: Google models featured in three studies, while Chinese-developed models (Qwen-2.5, ERNIE, SPARK) appeared only in Zhu et al. (2025) multi-model comparison. Model comparison rigour ranged from Zhu et al.

(2025), a systematically controlled evaluation of seven LLMs under standardised conditions, to single-model assessments without comparators, such as those by Krumsvik (2024) and Huang (2023), where no benchmarking against other models or human performance was performed. Human-model comparison was rare, with only a few studies (e.g., Su et al. (2024) and Wu et al. (2024)) including reference to human performance or conducting formal expert evaluations. Reporting of model version, session setup, and prompting parameters was inconsistent across both modes. Temperature settings were reported in only two studies, with values ranging from 0 to 0.7, where available. Some studies (e.g., Kaneda et al. (2023)) preserved contextual integrity in scenario-based questions by submitting grouped items within shared sessions.

NCLEX-RN/PN (n=3) and CNNLE/NNLE (n=5) were the most frequently evaluated exam systems, followed by RNLE and JNNE, with Western examinations remaining underrepresented (four instances in total: three NCLEX-RN/PN, one Norwegian AFB, and one Spanish general nursing exam). MCQs dominated (95%), with subtypes varying in format, including single-answer, multiple-response, and situation-setup formats. Language handling also varied: for example, Wu et al. (2024) tested translated NCLEX-RN and NNLE items using ChatGPT; Krumsvik (2024) translated the Norwegian AFB exam questions prior to testing. Authors employed diverse prompting strategies: zero-shot (n = 5) was the most common, followed by chain-of-thought (n = 2), step-by-step, and role-specified prompts. Input procedures ranged from individual item entry to batch processing.

Several studies sourced their questions from official licensing exams (e.g., NNLE, CNNLE, JNNE). Others relied on practice materials or web-based databases (e.g., Nurseslabs in Wu et al. (2024)). Studies also showed varied validation rigour. Some studies (García-Rudolph et al., 2024; García-Rudolph et al., 2025; Huang, 2023; Kaneda et al., 2023; Krumsvik, 2024; Miao, 2024; Taira et al., 2023; Wu et al., 2024; Zhao et al., 2025; Zhu et al.,

2025; Zong et al., 2024) used official answer keys, while others (Krumsvik, 2024; Su et al., 2024) applied expert review, and some (Zhao et al., 2025; Zhu et al., 2025) incorporated automated scoring or statistical comparison.

Table 1. Characteristics of included studies

Author, Year	Country	Study Design	LLM Model	Exam name	Source of questions	Exam format	No. of questions	Prompting & Input Procedure
García-Rudolph et al. (2024)	Spain	Comparative validation	GPT-3.5, GPT-4, Gemini	NCLEX-RN	NCLEX Practice Questions Exam	MCQ	250	Items were input individually to Gemini, GPT-3.5, and GPT-4; responses recorded in Excel. Zero-shot* (MCQ format); temp not reported
García-Rudolph et al. (2025)	Spain	Comparative validation	GPT-3.5, Gemini	NCLEX-style practice (RN & PN); CONVALIDATE-EU-SPAIN	Nurseslabs & Spanish Ministry of Health exams	MCQ	360 (240 English + 120 Spanish)	Zero-shot format; each item was input individually to both LLMs without additional prompting; answers were recorded in Excel. Temp not reported.
Huang (2023)	Taiwan	Descriptive quantitative evaluation	GPT-3.5	RNLE	Official RNLE past exams	MCQ (single-answer only)	~1,540	Randomly sampled, formatted MCQ were input without prompt engineering; no model comparisons conducted. Prompt strategy not specified; GPT-3.5; temp not reported.
Kaneda et al. (2023)	Japan	Comparative evaluation	GPT-3.5, GPT-4	JNNE	2023 MHLW exam bank	compulsory, general, scenario, conversation	237	Manually entered items into ChatGPT-3.5 and GPT-4 using new sessions; scenario sets answered jointly for context. Image-based items excluded; accuracy validated against MHLW key and model comparison performed. Zero-shot prompting; temp not reported.
Krumsvik (2024)	Norway	Controlled experimental evaluation	GPT-4	Norway AFB	Official 2023 AFB exam; translated by author	MCQ (text and image-based)	53	Tested GPT-4 on 53 items using chain-of-thought** prompts (temperature*** 0.7) to simulate an exam taker; scored per official guidelines and validated by two researchers; no model or human comparisons. Chain-of-thought prompt; temp = 0.7.
Miao (2024)	China	Retrospective evaluation	GPT-4	NNLE	NNLE archives + People's Health of China	MCQ + open-ended case prompts	720 MCQ	Inputs standardised via prompt engineering with new sessions per question. MCQs scored against keys; open-ended responses evaluated for similarity, logic, and quality. Direct MCQ were input; role-assigned prompts for open-ended items; temp not reported
Su et al. (2024)	Taiwan	Cross-sectional evaluation	GPT-4	Taiwan RNLE	Tawian MoE 2022 item bank	MCQ (single and multi-response)	400	Evaluated items using prompts with 'think step by step' and 'show me the answer'; responses generated at temp = 0 & assessed in two phases: explanation & expert comparison. Chain-of-thought; API access
Taira et al. (2023)	Japan	Quantitative evaluation study	GPT-3.5	JNNE	Official 2019–2023 JNNE exams; images excluded	MCQ (simple + situation-setup)	~1,160	Evaluated items (excluding images) using two standardised prompts; accuracy calculated separately for basic and general questions using official scoring criteria. Prompt strategy not specified; temp not reported
Wu et al. (2024)	USA/ China	Cross-sectional evaluation	GPT-4, GPT-3.5, Bard	NCLEX-RN, NNLE	NCLEX: Nurseslabs;	MCQ (4-option, single-answer)	630	Compiled NCLEX-RN and NNLE MCQs; translated items via ChatGPT-3.5. Original and translated versions tested on GPT-4, GPT-3.5, and Google Bard. Prompt type and temp not reported

					NNLE: Baidu 2021–2022			
Zhao et al. (2025)	China	Cross-sectional descriptive	Custom GPT, GPT-4	CNLE	2024 CNLE mock exams	MCQ (A1–A4 types)	720	Custom GPT and ChatGPT-4 assessed on 3 Chinese mock exams using prompt-engineered**** inputs; responses scored for accuracy and compared by question type using Z-tests in R. Role-specified prompt ('choose and explain'); temp not reported; web interface
Zhu et al. (2025)	China	Retrospective cross-sectional study	GPT-3.5/4/4o, Qwen-2.5, Copilot, ERNIE, SPARK	CNNLE	CNNLE official exams (2019–2023	MCQ (A1–A4: brief, shared, knowledge)	1,200	Seven LLMs tested on questions under identical conditions using separate sessions & devices; responses validated against official answer keys. Zero-shot; temp not reported
Zong et al. (2024)	China	Quantitative comparative evaluation	GPT-3.5	NNLE, NPLE, NMLE	Official 2017–2021 archives; non-text items excluded	MCQ (single + multiple-answer)	~ 6,600	Structured MCQs with context were input to ChatGPT; responses were clinician-evaluated against the answer key. Zero-shot; temp not reported.

Abbreviations
NCLEX: National Council Licensure Examination (USA); **NNLE:** National Nursing Licensing Examination (China); **JNNE:** Japanese National Nursing Examination; **RNLE:** Registered Nurse Licensure Examination (Taiwan); **CNNLE:** Chinese National Nurse Licensure Examination; **CNLE:** Chinese Nursing Licensing Examination; **AFB:** Anatomy, Physiology, and Biochemistry exam (Norway); **PN:** Practical Nurse; MoE: Ministry of Education (Taiwan); **MHLW:** Ministry of Health, Labour and Welfare (Japan); **CONVALIDATE-EU**: European Union project on cross-border nursing qualification recognition.

Terminologies
***Zero-shot:** Model receives only the question, without examples or prior context.
****Chain-of-thought:** Prompting that encourages step-by-step reasoning;
*****Temp (Temperature):** A model parameter controlling randomness, lower values yield more consistent responses.
******Prompt-engineering:** Customising input phrasing to improve output quality;

3.4 Narrative Synthesis of Findings

Analysis of the 12 studies revealed marked variability in LLM performance, which was shaped by differences in model architecture, examination systems, item formats, language context, and content domains. Overall, GPT-4 consistently outperformed earlier versions such as GPT-3.5 and other LLMs, including Google Bard and ERNIE Bot, yet performance remained context-dependent.

3.4.1 Accuracy benchmarks and model comparison

Across the 12 studies, reported accuracy ranged from 49.5% (GPT-3.5, China) to above 90% (Custom GPT, China). GPT-4 often met or exceeded the minimum passing scores required for national licensure examinations (national thresholds), such as 92.4% in Spain (García-Rudolph et al., 2024), 80.75% in Taiwan (Su et al., 2024) and 79.7% in Japan (Kaneda et al., 2023). However, it fell below benchmarks in certain years or components (Taira et al., 2023). In China, ChatGPT failed to meet the 60% threshold across five years (Zong et al., 2024). Newer-generation models showed substantial gains. Qwen-2.5 achieved the highest overall accuracy (88.9%) among six models and excelled in practical skills (Zhu et al., 2025). Wu et al. (2024) reported that ChatGPT-4.0 outperformed ChatGPT-3.5 and Google Bard on NCLEX-RN practical items with 88.7% in English and 79.3% in translated Chinese items. Krumsvik (2024) documented high overall accuracy (84.9%–94.5%) for GPT-4 on anatomical and biochemical content.

3.4.2 Domain-specific trends and question complexity

The studies highlighted marked variability by domain. Performance was high in general medicine (88.75%), basic nursing knowledge (75.1%), as well as in clinical epidemiology, dermatology, and nutrition. Consistent weaknesses were observed in psychosocial integrity (13.3%), pharmacology, social welfare, humanistic nursing, parasitology, anatomy, and fields such as fundamental nursing, nursing administration, maternal and paediatric nursing, and

medical-surgical nursing (Huang, 2023). LLM accuracy also varied with the complexity of the questions. Complex formats, such as clinical vignettes and scenario-based items, lowered performance. Su et al. (2024) reported significant effects for clinical vignettes ($p = .007$) and complex MCQs ($p = .049$). Zong et al. (2024) found accuracy differences between single- and multiple-choice questions ($p < .0001$). Kaneda et al. (2023) reported a scenario-based performance range of 51.7% to 80.0%, while odds ratios for incorrect answers in complex formats ranged from 2.19 to 2.37.

3.4.3 Customisation, Language Effects, and Limitations

Customised models performed best. Zhao et al. (2025) reported a domain-specific GPT model scoring over 90% in all six parts of a Chinese exam, surpassing GPT-4 across Professional Practice and Practical Abilities. Wu et al. (2024) demonstrated a statistically significant drop in translated NCLEX-RN performance ($p = .03$). However, no such effect was found for NNLE items ($p = 0.92$). Language-related challenges were further noted by Huang (2023) including hallucinations, bias, and content misalignment. Krumsvik (2024) reported wide accuracy variability in multimodal tasks (62.5% to 100%), highlighting difficulties with image-based content. Miao (2024) noted low performance (56.3%) on visual questions. Other limitations included logical inconsistency, misinterpretation of clinical cases, and unreliable outputs under ambiguity (Huang, 2023; Miao, 2024).

3.4.4 Author-reported model limitations

The handling of multimodal content remained a critical constraint. Due to this limitation, several studies excluded multimodal items such as images and diagrams (Huang, 2023; Taira et al., 2023). Among those that included visuals, Miao (2024) and Krumsvik (2024) reported poor or variable outcomes (as low as 56.3%). Language and cultural mismatches also emerged, particularly for models trained primarily in English contexts (Miao, 2024; Wu et al., 2024). Prompt engineering and implementation fidelity introduced further variability. Zong et al.

(2024) identified issues such as "zero-shot learning may not be optimal" and "prompt variations affect responses." The evolving nature of LLMs was recognised by Kaneda et al. (2023), cautioning against static benchmarking of rapidly updating tools.

Table 2: Performance of LLMs on Nursing Licensure Exams

Author, Year	Performance Metrics	Content Domains/Areas	Benchmark Standards & Results	Performance Threshold	Statistical Analysis & Domain Results	Model Strengths, Limitations & Error Patterns
García-Rudolph et al. (2024)	Accuracy rate, domain-specific performance, error pattern analysis	Safe Effective Care, Health Promotion and Maintenance, Psychosocial Integrity, Physiological Integrity	**Benchmark**: 77% passing grade **Results:** Gemini: 73.2% GPT-3.5: 72% GPT-4: 92.4%	GPT-4 met the threshold (77%) Gemini & GPT-3.5 did not meet the threshold.	**Statistical Analysis:** Not reported **Domain Results:** Psychosocial Integrity: 13.3% Health Promotion & Maintenance: 8.8% Safe Effective Care: 7.7% Physiological Integrity: 6.4%	**Strengths:** GPT-4 strong in Physiological Integrity **Weaknesses:** Weak in Psychosocial Integrity **Error Patterns:** Not specified
García-Rudolph et al. (2025)	Accuracy rate, domain-specific performance, error analysis	NCLEX-RN topics: Physiological Adaptation, Risk Reduction, Health Promotion, etc.; CONVALIDATE-EU domains	**Benchmark:** 77% passing grade Results: GPT-3.5: 69.2% (RN), 67.5% (PN), 76.7% (EU) GEMINI: 65.8% (RN), 67.5% (PN), 76.7% (EU)	GPT-3.5 and GEMINI met the threshold only for the Spanish exam	**Statistical Analysis:** Not reported **Domain Results:** Physiological Adaptation: GPT-3.5 47.8%, GEMINI 52.2% Risk Reduction: GPT-3.5 48.0%, GEMINI 52.0% Health Promotion: GPT-3.5 57.1%, GEMINI 42.9%	**Strengths:** GPT-3.5 in Health Promotion; GEMINI in Physiology **Limitations**: Below threshold in US exams **Error Patterns:** Recurrent issues with pregnancy, legal ethics, paediatric growth
Huang (2023)	Correct percentage; pass rate; advantages/ disadvantages	BMS, FNNA, MSN, MPN, PCN	**Benchmark:** Minimum passing score of 60 **Results:** Average scores ranged from 51.6 to 63.75, with passing rates of 1st place in 2022 and 2nd place in 2023	Performance met the minimum threshold in some assessments	**Statistical Analysis:** Not reported **Domain Results:** Exceptional performance in BMS and PCN; poor performance in FNNA, MSN and MPN	**Strengths:** Strong performance in BMS and PCN **Limitations:** Inadequate medical knowledge, confusion in complex scenarios, hallucinations, language bias **Error Patterns**: Knowledge gaps in specialised areas
Kaneda et al. (2023)	Accuracy rate; question type performance	Basic nursing, adult nursing, geriatric nursing, paediatric nursing, maternal nursing, psychiatric nursing, home care nursing theory, integrated and practical nursing	**Benchmark:** 80% for compulsory questions, >40 points for compulsory questions, and a score of 152 points or more for general & scenario-based questions **Results:** GPT-3.5: Failed to meet passing standards (59.9%) GPT-4: Met passing standards (79.7%)	GPT-4 met the threshold GPT-3.5 did not meet the threshold	**Statistical Analysis**: $p < 0.01$ for overall accuracy, compulsory, and scenario-based questions; $p = 0.014$ for general questions; $p = 0.248$ for conversation questions **Domain Results:** Compulsory: 58.0% to 90.0% General: 64.6% to 75.6% Scenario-based: 51.7% to 80.0%	**Strengths:** Not reported **Weaknesses**: Not reported **Error Patterns:** Performance varied significantly by question type
(Krumsvik, 2024)	Accuracy rate; error rate;	Anatomy, Physiology, and Biochemistry	**Benchmark:** Grading guidelines provided by NOKUT	Not reported	**Statistical Analysis:** Not reported **Domain Results**: Performance	**Strengths:** High overall accuracy **Weaknesses:** Performance variations in

	multimodal performance		**Results:** Overall accuracy: 84.9% to 94.5%		variations by domain/question type: Multimodal questions (62.5% to 100%)	multimodal questions **Error Patterns:** Difficulty with complex visual/multimodal content
(Miao, 2024)	Accuracy rate; cosine similarity; logical consistency; information quality	Respiratory, Circulatory, Haematologic, Endocrine, Urinary system diseases; Professional knowledge unit; Practical knowledge unit	**Benchmark:** Prior GPT-3.5 results **Results:** • Multiple-choice questions: 71.0% • Image-based questions: 56.3%	No statistical difference between the professional and practical knowledge units	**Statistical Analysis:** No significant differences among exam years (P > .05); no statistical difference between professional and practical knowledge units **Domain Results**: No significant differences across units; poor performance on image-based questions (56.3%)	**Strengths:** Clinical reasoning **Weaknesses**: Identifying specific cut-off values, humanistic nursing, prioritising nursing diagnoses **Error Patterns:** Difficulty with visual/image interpretation
Su et al. (2024)	Accuracy rate; domain-specific performance; consistency analysis	BN, GM, MSN, OGN, PCN	**Benchmark:** 60% **Results**: GPT-4: Overall accuracy 80.75%	Model met the threshold	**Statistical Analysis:** Clinical vignettes (p=0.007), complex multiple-choice questions (p=0.049) **Domain Results:** GM: 88.75%, MSN: 80%, PCN: 70%, OGN: 67.5%, BN: 63%	**Strengths:** Clinical vignettes, complex multiple-choice questions **Weaknesses**: Not specified **Error Patterns:** Performance varied by question complexity
Taira et al. (2023)	Accuracy rate, question type performance, subject area performance	Pathology, Anatomy, Physiology	**Benchmark**: ~80% for basic knowledge, ~60% for general questions **Results:** Basic knowledge: 75.1% (SD 3%) General questions: 64.5% (SD 5%)	80% threshold: Basic knowledge met in 2019, did not meet in other years	**Statistical Analysis:** Standard deviation reported **Domain Results:** 75.1% (SD 3%) for basic knowledge questions, 64.5% (SD 5%) for general questions	**Strengths**: Nutrition, pathology **Weaknesses:** Pharmacology, social welfare **Error Patterns:** Difficulty with complex problem-solving
Wu et al. (2024)	Accuracy rate; multilingual performance; professional knowledge assessment	Nursing Education, Clinical Practice	**Benchmark:** Pass threshold NR **Results:** NCLEX-RN practical questions: 88.7% (133/150) Chinese-translated NCLEX-RN practical questions: 79.3% (119/150) NNLE Theoretical MCQs: 71.9% (169/235) NNLE Practical MCQs: 69.1% (161/233)	Not reported	**Statistical Analysis:** p-values: P=.03 for NCLEX-RN English vs. Chinese input; P=.92 for NNLE Theoretical MCQs English vs. Chinese input **Domain Results:** Higher accuracy for NCLEX-RN practical questions in English; ChatGPT 4.0 outperforms ChatGPT 3.5 and Google Bard	**Strengths:** NCLEX-RN practical questions **Weaknesses:** Performance degradation with translation **Error Patterns**: Language-dependent performance variations

Zhao et al. (2025)	Accuracy rate, explanation quality	Six parts of nursing examination (specific domains not detailed)	**Benchmark:** Not reported **Results:** Custom GPT: >90% across all six parts ChatGPT-4: 73% to 89% across all six parts	Custom GPT outperformed ChatGPT-4 in all parts	**Statistical Analysis**: Not reported **Domain Results:** Professional Practice: Custom GPT >85% Practical Abilities: Custom GPT >85% A1 type questions: Custom GPT significantly better A2 type questions: Custom GPT significantly better A3/A4 type questions: Custom GPT significantly better	**Strengths:** Custom GPT superior performance across all question types **Weaknesses:** Not reported **Error Patterns:** Standard ChatGPT-4 showed consistent performance gaps across question types
Zhu et al. (2025)	Accuracy; AUC; sensitivity; specificity; F1-score; PPV; NPV	Nursing-related knowledge in clinical settings, Application of nursing knowledge and skills	**Benchmark:** Not reported Results: Qwen-2.5: 88.9% GPT-4o: 80.7% ERNIE Bot-3.5: 78.1% GPT-4.0: 70.3% SPARK: 65.0% GPT-3.5: 49.5%	Pass threshold: Not reported	**Statistical Analysis:** Not reported Domain Results: Not reported	**Strengths:** Qwen-2.5 better in Practical Skills than Professional Practice **Weaknesses:** Not specified **Error Patterns:** Not reported
Zong et al. (2024)	Accuracy rate, domain-specific performance, error pattern analysis	Clinical Epidemiology, Human Parasitology, Dermatology	**Benchmark:** 0.6 accuracy threshold **Results:** ChatGPT: Failed to meet the 0.6 accuracy threshold across all examination types (5-year period)	Did not meet pass the threshold of 0.6	**Statistical Analysis:** Single vs. multiple-choice questions ($p<0.0001$) Domain Results: No significant differences across units. Significant difference between single/multiple-choice in NPLE	**Strengths:** Clinical epidemiology, dermatology **Weaknesses:** Parasitology, anatomy **Error Patterns:** Performance varied by question format

Note:
The abbreviations used in this table refer to subject domains and performance metrics evaluated across the studies.

- **BN**: Basic Nursing; **GM**: General Medicine; **MSN**: Medical–Surgical Nursing; **OGN**: Obstetrics and Gynecology Nursing; **PCN**: Psychology and Community Nursing.
- **BMS**: Basic Medical Science; **FNNA**: Fundamental Nursing and Nursing Administration; **MPN**: Maternal and Pediatric Nursing.
- **NCLEX-RN**: National Council Licensure Examination for Registered Nurses (USA); **NNLE**: National Nursing Licensing Examination (China); **PN**: Practical Nurse; **CONVALIDATE-EU**: A European project assessing cross-national nursing qualification equivalence; **NPLE**: National Physician Licensing Examination (China).
- **AUC**: Area Under the Curve; **PPV**: Positive Predictive Value; **NPV**: Negative Predictive Value; **F1-score**: A composite metric of precision and recall; **SD**: Standard Deviation.
- **A1/A2/A3/A4-type questions**: Classification used in some national nursing exams, where A1 indicates simple recall, A2 represents understanding and direct application, A3 requires decision-making based on multiple factors, and A4 involves complex clinical judgment or scenario-based reasoning.

3.5 Meta-analysis

3.5.1 Model type/version

The pooled accuracy of the LLM licensure examination questions was 69.6 (95% CI: 65.6–73.6%), with a wide prediction interval (43.8–95.3%) and substantial heterogeneity (I^2 = 98%). Subgroup analysis revealed clear performance differentials across models. ChatGPT-3.5 achieved the lower pooled accuracy at 60.4% (95% CI: 55.8–64.9%), while ChatGPT-4.0 demonstrated a marked improvement with a pooled accuracy of 77.2% (95% CI: 72.4–82.0%). Performance further increased with ChatGPT-4o (80.7%), Qwen 2.5 (88.9%), and a task-specific Custom GPT model, which reported the highest accuracy at 93.6%. Models such as ERNIE Bot 3.5 (78.1%) and Gemini (71.4%) also performed comparably, whereas Google Bard exhibited the lowest subgroup performance (53.3%).

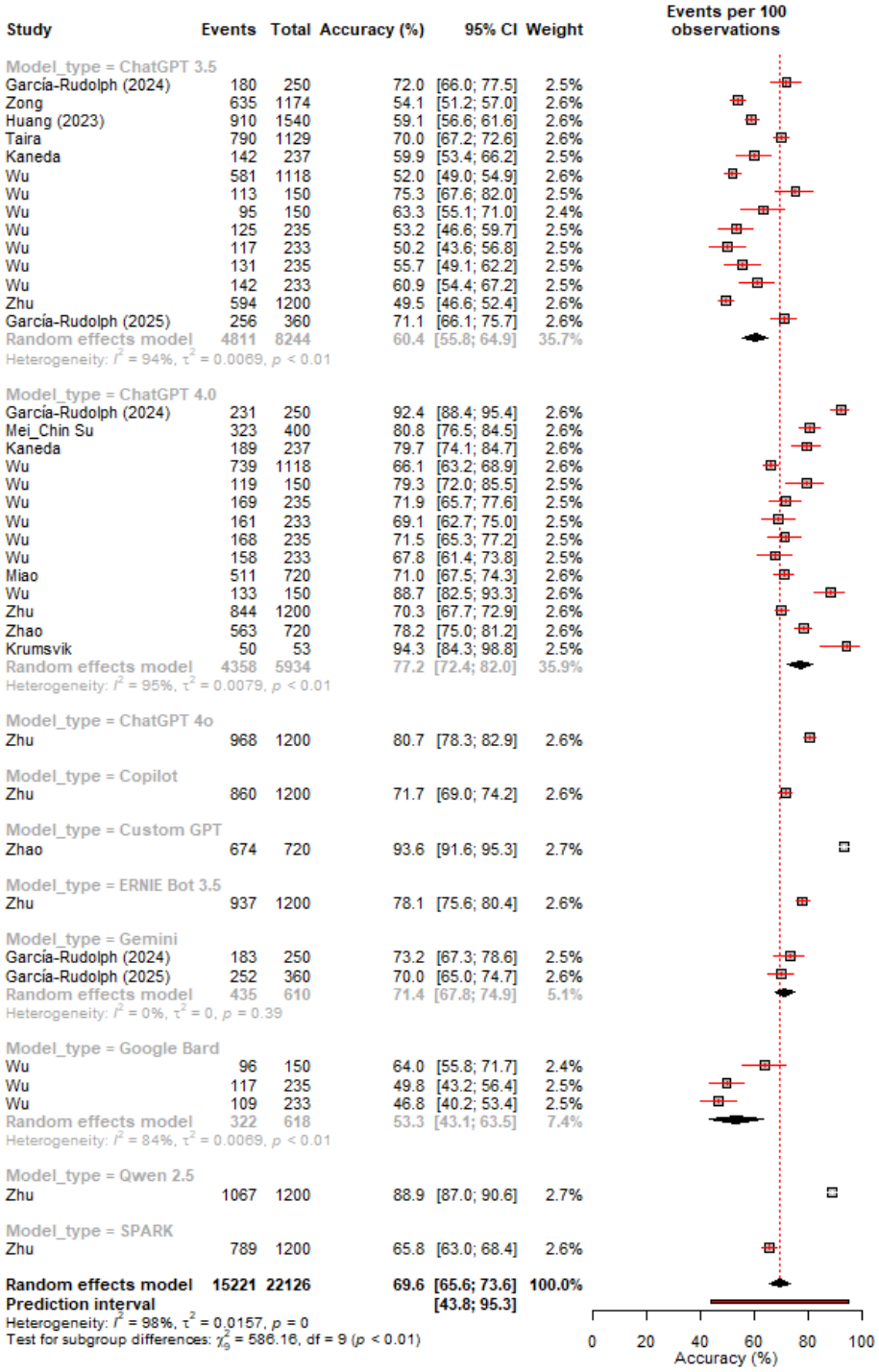


Figure 2 : Accuracy of LLMs (subgroup by model type)

3.5.2 Performance by GPT version

The meta-analysis compared the accuracy of different ChatGPT versions, revealing significant performance differences. ChatGPT 4.0 achieved higher accuracy (77.2%, 95% CI:

72.4–82.0%) than ChatGPT 3.5 (60.4%, 95% CI: 55.8–64.9%), with the Custom GPT performing best (93.6%, 95% CI: 91.6–95.3%). Heterogeneity was high across all models (I^2 = 94–98%), indicating substantial variability between studies. Subgroup differences were statistically significant ($p < 0.01$), confirming that the model version significantly impacted accuracy.

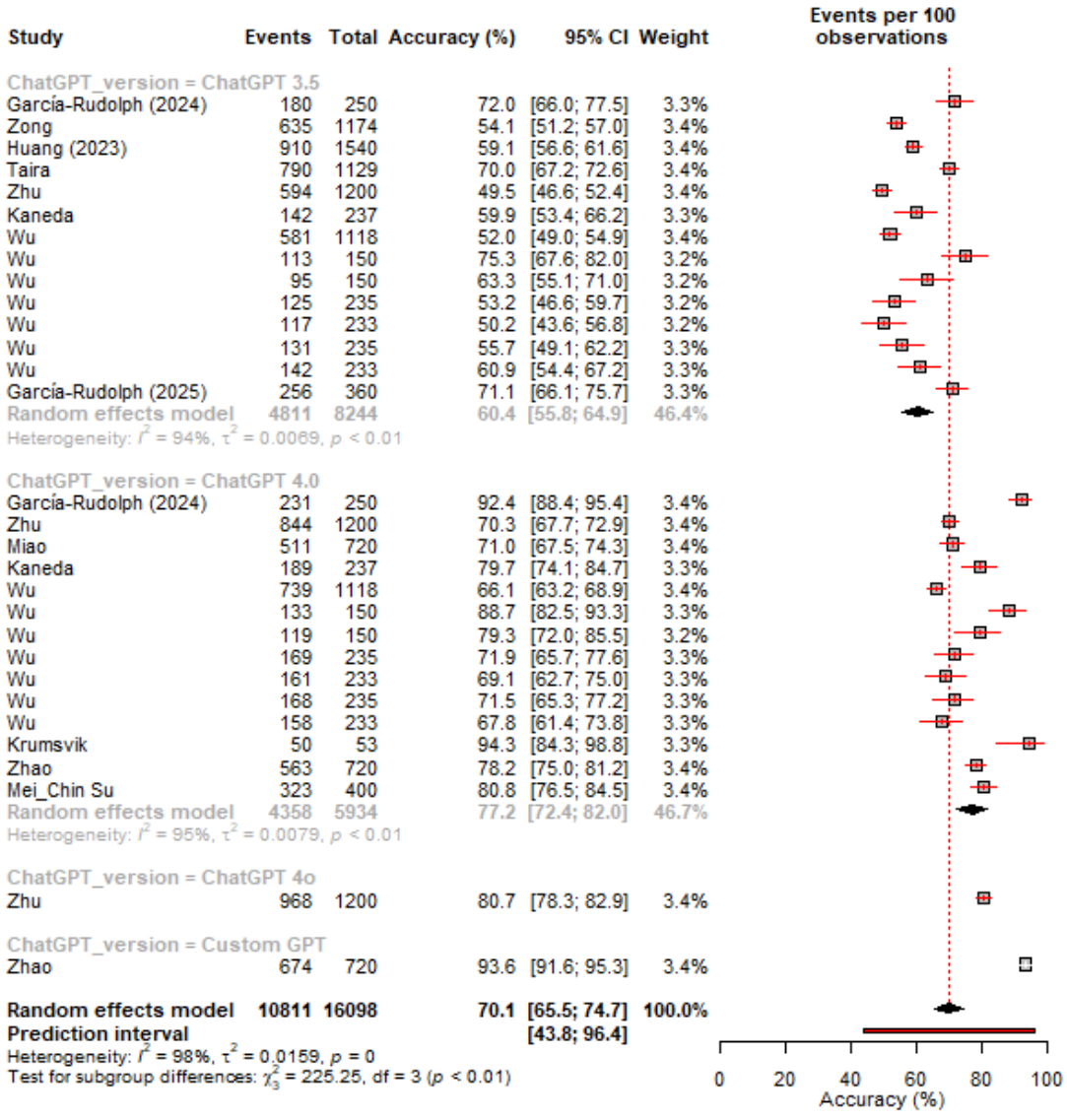


Figure 3: Accuracy of LLMs (subgroup by GPT version)

3.5.3 Performance by Exam system/country

The meta-analysis evaluated the model performance of LLMs across different types of nursing licensure exams, revealing significant variations in accuracy. The highest accuracy was observed for TNNE (80.8%, 95% CI: 76.5–84.5%), while RNLE showed the lowest performance (59.1%, 95% CI: 56.6–61.6%). NCLEX-RN, CNNLE and JNNE demonstrated moderate accuracy at 77.1% (67.8–86.5%), 74.8% (68.3–81.2%) and 69.9% (67.7–72.2%), respectively, with NNLE exhibiting the widest variability (65.5%, 95% CI: 43.1–87.8%). Subgroup differences were statistically significant ($p < 0.01$), suggesting that the exam format had a significant impact on LLMs' performance.

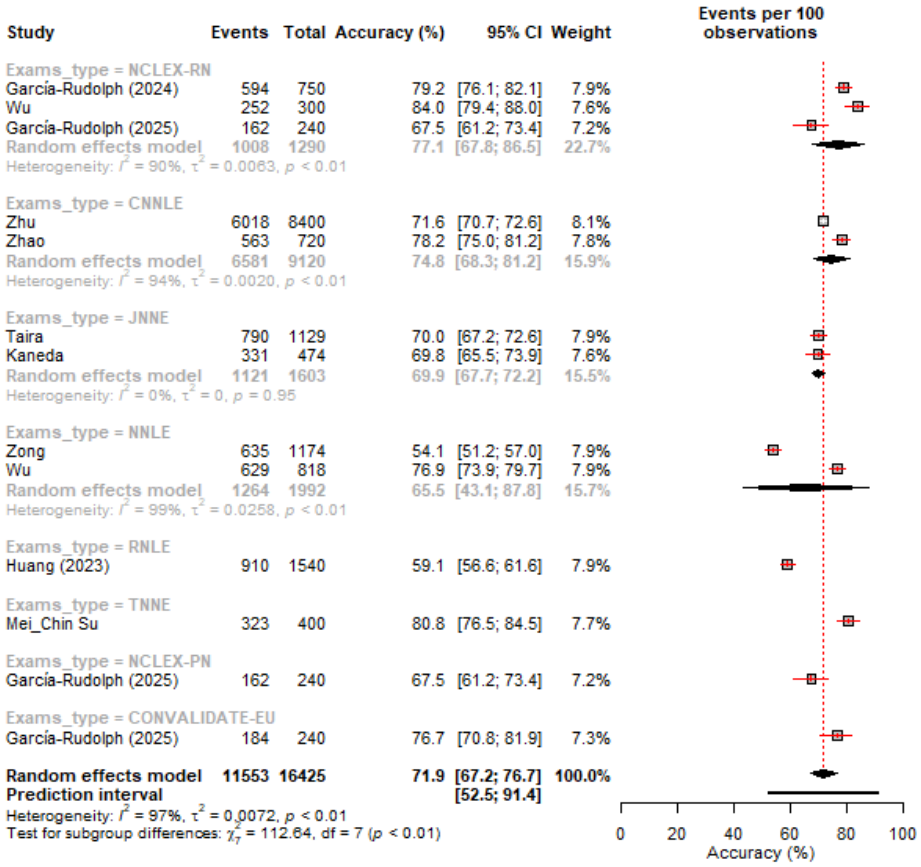


Figure 4: Accuracy of LLMs (subgroup by Exam system/country)

3.5.4 Performance by year group

The meta-analysis revealed no clear temporal trend in LLM accuracy across exam years (2017-2023). Annual fluctuations were observed, and accuracy peaked in 2022 (71.5%) and dipped to its lowest in 2020 (67.4%), however these year-to-year variations were not statistically significant (p=0.62), suggesting that study-specific factors (e.g., model type or exam format) had a greater influence on results than examination year.

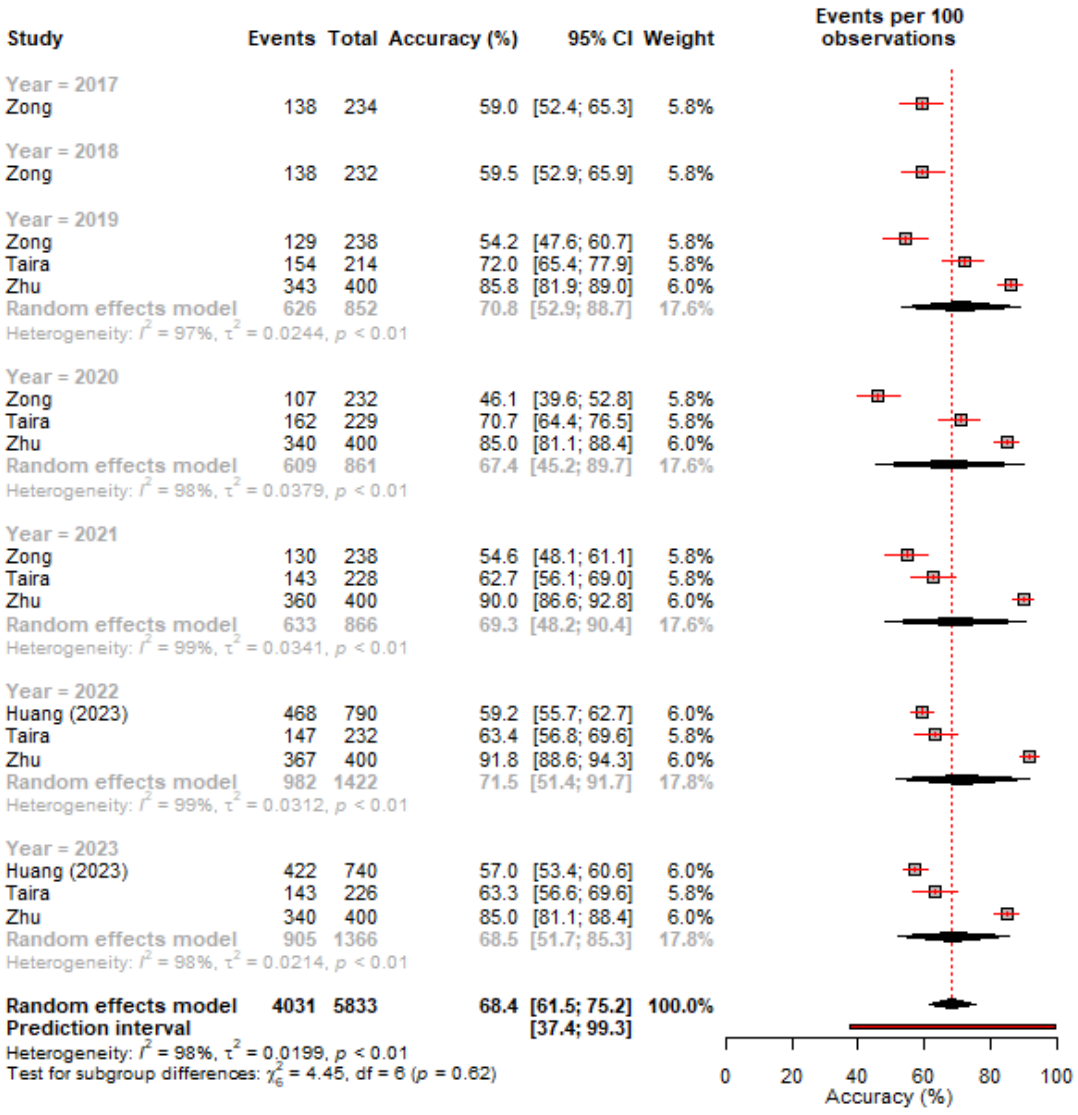


Figure 5: Accuracy of LLMs (subgroup by Year Group)

3.5.5 *Performance by subject area level*

The meta-analysis compared the accuracy of the LLMs across the three subject-area difficulty levels (Levels 1-3), revealing no significant performance differences (p = 0.90) despite varying complexity. Performance remained consistent across levels: Level 1 (70.4%), Level 2 (67.0%), and Level 3 (69.1%). The wide prediction interval (40.6–97.1%) suggests that factors beyond subject-area difficulty (e.g., model type or implementation) likely drive performance variations more than complexity level alone.

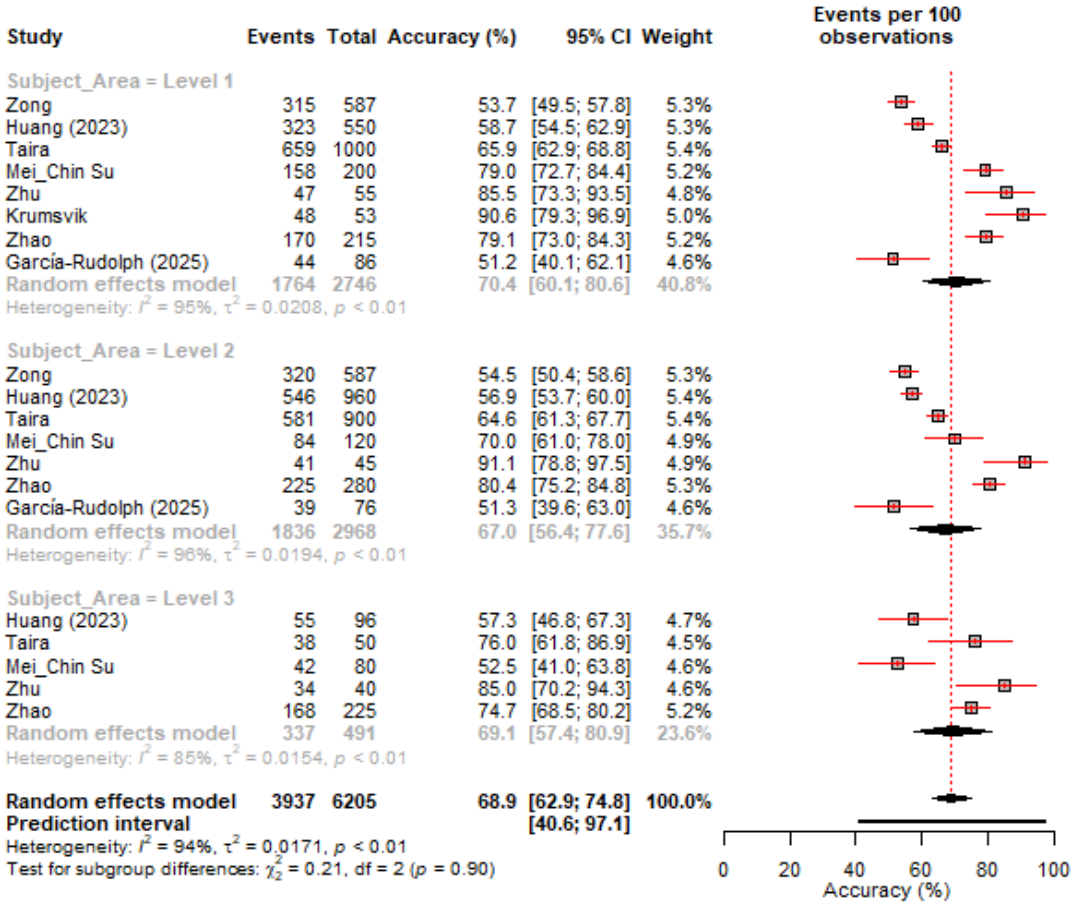


Figure 6: Accuracy of LLMs (subgroup by Subject area level)

3.5.6 *Alignment with national exams (Real exams and simulated).*

The meta-analysis compared the LLMs performance between simulated and real exam conditions, revealing no significant difference in accuracy (p=0.13). Simulated exams showed slightly higher performance (75.5%) than real exams (71.2%). However, real exam results

exhibited extreme heterogeneity (I^2=98%) compared to simulated exams (I^2=87%), indicating substantial variability in real-world conditions.

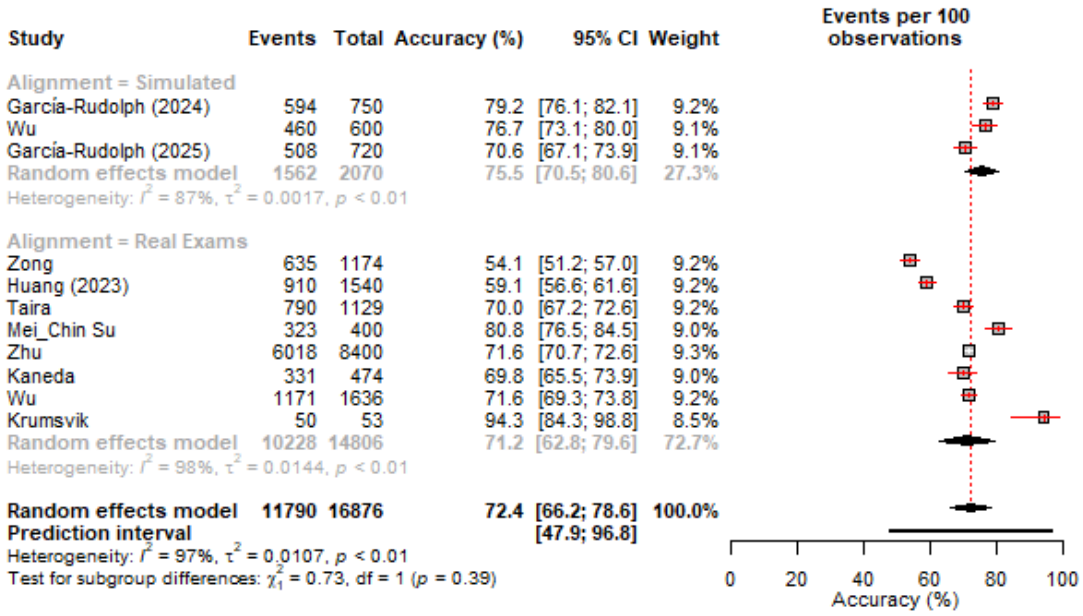


Figure 7: Alignment with national exams (Real exams and simulated)

3.5.7 Question format

The meta-analysis of LLM performance across the four question formats (A1-A4) revealed no significant accuracy differences between formats (p = 0.96). The format-level accuracy ranged narrowly from 77.7% (A4) to 80.4% (A1).

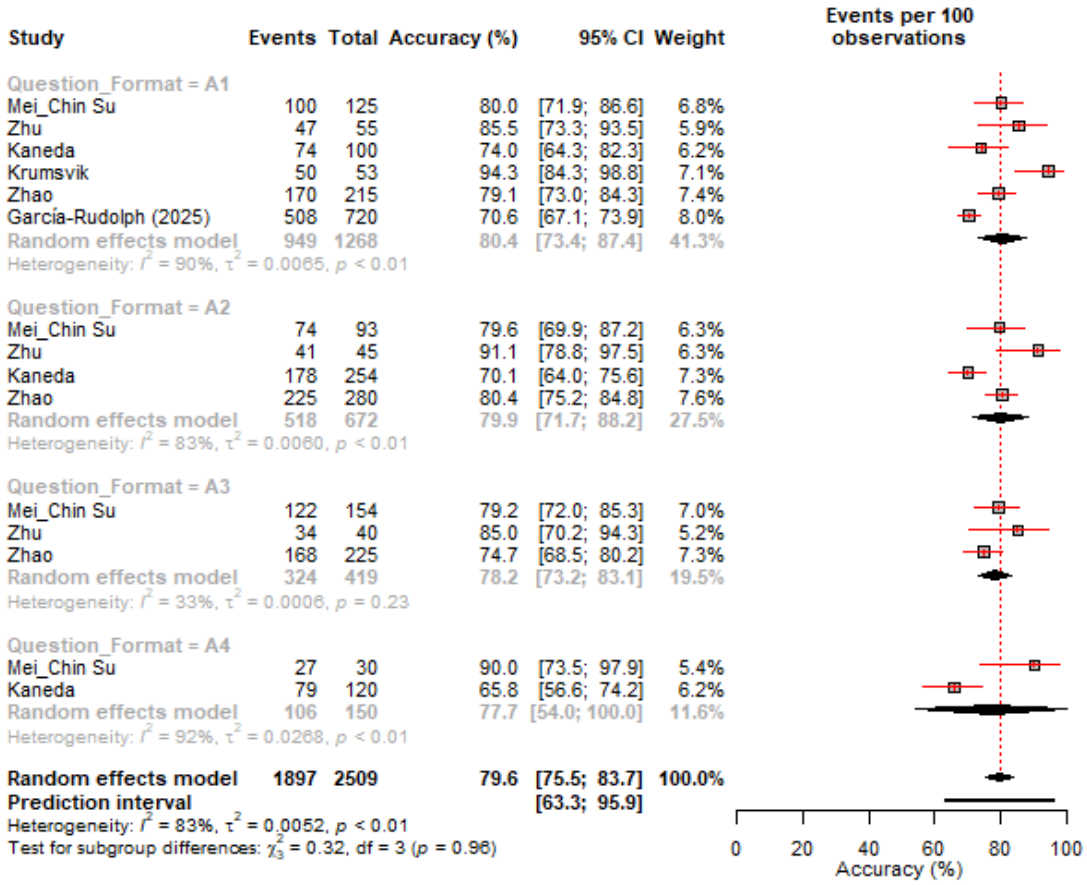


Figure 8: Question format

3.5.8 Publication bias assessment

The funnel plot (Supplementary file S6) displays the standardised mean difference (SMD) on the x-axis and the standard error of the SMD on the y-axis. Visual inspection of the funnel plot suggests a relatively symmetric distribution of effect sizes around the pooled estimate, with no clear evidence of asymmetry or small-study effects. The Eggers' regression test assessed funnel plot asymmetry by regressing the standardised effect size against the standard error. The results of the Egger's test did not reveal any significant asymmetry (p = 0.9161), suggesting that small-study effects or publication bias are unlikely to have substantially influenced the meta-analytic finding.

3.6 Heterogeneity Assessment

3.6.1 Meta regression

The regression analysis revealed that only the CNNLE exam system showed a statistically significant positive effect (Estimate = 3127.805, $p<0.0001$), while all other predictors, including exam systems, LLM model types, question characteristics (alignment, format), subject domains, and year groups demonstrated no significant associations (all $p>0.05$). (Supplementary File S7: Meta-regression)

3.6.2 Sensitivity Analyses

Our sensitivity analysis demonstrates significant variability in model performance across studies (range: 0.465-0.942), complementing the meta-analytic findings of 69.6% overall accuracy (95% CI: 65.6–73.6%). High-performing models demonstrated exceptional sensitivity (Krumsvik: 0.942; Zhao: 0.936; Su: 0.924). The median sensitivity (0.689) closely approximated the point estimate for accuracy, while the interquartile range (0.607-0.781) aligned with the 95% CI bounds. However, the extreme values at both ends of the sensitivity distribution (particularly the lower range) correspond to the wide prediction interval (43.0-96.1%) observed in the accuracy analysis, suggesting that methodological differences across studies may account for more performance variation than model architecture alone (Supplementary File S8: Sensitivity analysis).

4 Discussion

This systematic review and meta-analysis examined the accuracy of LLMs on nursing licensure examinations. The review found substantial heterogeneity in LLM performance, with a pooled accuracy of 69.6% (95% CI: 65.6-73.6). Performance varied significantly by model architecture, with GPT-3.5 achieving 60.4%, GPT-4 reaching 77.2%, and domain-customised models, such as Zhao's version, attaining 93.6%. Among the licensure systems, NCLEX-RN

demonstrated higher accuracy (77.1%) compared to RNLE (59.1%), while performance across subject domains varied, with general medicine reaching 88.75% and consistent underperformance in psychosocial integrity. Question complexity and language also influenced outcomes. Subject-area difficulty levels (Levels 1 to 3) showed no significant accuracy differences ($p = 0.90$), suggesting that LLMs handle gradients of content difficulty comparably. The meta-analysis also found no temporal trend in model accuracy for the exam years 2017 to 2023 ($p = 0.62$), indicating that the observed gains reflect improvements in model design rather than variations due to exam years.

This review aligns with existing reviews that show variability in LLM performance across medical exams, with GPT-4 consistently outperforming GPT-3.5 (Jin et al., 2024; Keshavarz et al., 2024; Liu et al., 2024). The pooled accuracy of 69.6% matches Jin et al.'s 70.1% and falls within the broader 56–70% range reported by Wei et al. (2024), Levin et al. (2023) and Waldock et al. (2024), confirming moderate accuracy across healthcare fields. However, the NCLEX-RN subgroup accuracy of 77.1% challenges Jin et al. (2024) claim that nursing shows the lowest LLM performance, exceeding their reported medical average. This suggests variation within nursing assessments and highlights the value of newer models. This review also adds to prior work by examining exam systems, question complexity, and translation effects, identifying domain-specific models and structured formats (e.g., CNNLE) as performance enhancers. Unlike Levin et al. (2023), who reported underperformance at higher cognitive levels, this review found no significant differences in performance by subject-level difficulty. However, our review noted that accuracy declines on multi-step reasoning tasks, suggesting that item structure is a key factor in this decline. This contrasts with Sumbal et al. (2024), who directly linked accuracy to the difficulty level. The finding may suggest that nursing exams may follow different cognitive patterns. Notably, the consistent underperformance in psychosocial integrity domains across regions reveals a content-specific

limitation not captured in broader reviews. These findings refine our understanding of LLM reliability in high-stakes, nursing-specific contexts and support the need for culturally informed, standardised evaluation tools.

LLM performance demonstrated a clear developmental trajectory, with accuracy progressing from GPT-3.5 through GPT-4 to domain-specific models such as Qwen-2.5 and Zhao's customised version (93.6%). While OpenAI models predominated, comparable performance was observed with ERNIE Bot 3.5 (78.1%) and Gemini (71.4%), contrasting sharply with Google Bard's suboptimal results (53.3%). Meta-regression identified CNNLE as uniquely conducive to LLM performance ($p < 0.0001$), suggesting that specific structural features, such as consistent item phrasing, domain clustering, or high-fidelity answer keys, may optimise AI compatibility and warrant investigation for future assessment design.

Despite architectural improvements, LLMs have consistently underperformed in domains that require complex reasoning, interdisciplinary synthesis, and affective judgment. Lower accuracy in psychosocial integrity, pharmacology, and humanistic nursing persisted across both Western and Asian contexts, contrasting with stronger performance in general medicine and pathology. Performance declined significantly on multi-step reasoning tasks, including clinical vignettes ($p = 0.007$) and multi-response formats ($p < 0.0001$), while question format categories showed no meaningful differences ($p = 0.96$). This pattern suggests that internal cognitive complexity, rather than surface-level formatting, drives performance variability.

The predominance of East Asian studies (9/12) limits global applicability, with translation effects demonstrating variable impact across examination systems. Accuracy dropped significantly for translated NCLEX-RN items ($p = 0.03$) but remained stable for NNLE items ($p = 0.92$). The finding may reflect differential language localisation in training datasets. Methodological inconsistencies further compromise evidence quality: sample sizes

varied dramatically (100-5,000 items), critical model parameters were underreported, and performance thresholds lacked standardisation. The extreme heterogeneity ($I^2 = 98\%$) and wide prediction interval (43.8–95.3%) underscore the urgent need for standardised evaluation frameworks to enhance rigour and reproducibility in this rapidly evolving field.

5 Limitations

This review has several methodological limitations that affect its robustness and generalizability. First, excluding non-English studies may have introduced language bias and limited global representation. Second, reliance on author-reported outcomes, often based on differing definitions of exam-like conditions, may have led to inconsistent exposure settings. Third, while dual screening and extraction were conducted using established tools, the JBI checklist may not fully capture AI-specific risks, such as prompt design or model version drift. Fourth, post hoc classification of subject domains and question formats, though necessary for comparison, may have introduced misclassification bias. Finally, key model variables such as temperature, prompting strategy, and scoring were often unreported, limiting causal inference despite subgroup and meta-regression analyses.

6 Implications for practice, policy and research

Customised LLMs, such as Zhao's GPT, show promise as supplementary tools in nursing education, especially for formative assessment in domains like general medicine and pharmacology, where accuracy can reach 88.75%. However, their limitations in ethical reasoning, psychosocial integrity, and complex clinical judgment require caution. Combining LLM-generated content with culturally grounded, instructor-led teaching and structured prompting may mitigate risks, especially when using translated materials. Integration into simulation-based learning could help foster critical thinking and support responsible use.

For licensure examinations, regulatory bodies should establish clear guidelines distinguishing between educational and high-stakes assessment contexts. Given the high variability in performance (I^2 = 98%), unmoderated use in critical assessments should be avoided. Policies should support culturally and linguistically adapted models to address translation-related accuracy drops, such as the 8–12% decrease observed in NCLEX-RN ($p = 0.03$). Pilot testing AI-informed assessment formats, drawing from effective systems like CNNLE ($p < 0.0001$), will be important. Broader international collaboration is necessary to address geographic research imbalances and promote equitable, values-aligned adoption of AI in nursing.

Addressing methodological and performance limitations requires a standardised research agenda. The adoption of reporting frameworks like CONSORT-AI and PRISMA-AI (Cacciamani et al., 2023) will enhance transparency and reduce heterogeneity. Expanding research to include Western licensure systems is essential for generalisability. Future studies should investigate error patterns, including factual inaccuracies, misinterpretations, and hallucinations, particularly in the domains of ethics and psychosocial issues. Comparative evaluations with human performance, development of nursing-specific benchmarks, and culturally tailored models will guide improvement. Longitudinal studies are also needed to assess whether model upgrades lead to sustained gains, given the lack of temporal trends ($p = 0.62$).

7 Conclusion

This systematic review and meta-analysis demonstrate that LLMs are promising for low-stakes nursing licensure preparation, particularly in formative assessments that incorporate hybrid teaching and simulation-based learning. However, their inconsistent performance and weaknesses in ethical reasoning, interdisciplinary synthesis, and cultural-linguistic adaptation make them unsuitable for unmoderated high-stakes exams. Limitations in ethics, psychosocial

integrity, and the translation of items raise concerns about fairness, validity, and reliability across diverse contexts. Regulatory bodies should set guidelines and ensure equitable access, while exam developers create AI-aligned formats inspired by effective systems. Researchers must focus on standardised frameworks, error analysis, and nursing-specific models to address geographic biases and performance gaps. Progress depends on evidence-based integration, rigorous evaluation, culturally aligned benchmarks, and clinically tailored architectures.

8 References

Abujaber, A. A., Abd-Alrazaq, A., Al-Qudimat, A. R., & Nashwan, A. J. (2023). A Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analysis of ChatGPT Integration in Nursing Education: A Narrative Review. *Cureus*, *15*(11), e48643. https://doi.org/10.7759/cureus.48643

Amankwaa, I., Ekpor, E., Cudjoe, D., Kobiah, E., Fuseini, A.-K. J., Diebieri, M., Gyamfi, S., & Brownie, S. (2025). Patterns, advances, and gaps in using ChatGPT and similar technologies in nursing education: A PAGER scoping review. *Nurse Education Today*, 106822. https://doi.org/https://doi.org/10.1016/j.nedt.2025.106822

Bagde, H., Dhopte, A., Alam, M. K., & Basri, R. (2023). A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon*, e23050. https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e23050

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J*, *8*(2), e188-e194. https://doi.org/10.7861/fhj.2021-0095

Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260.

Bongco, E. D. A., Cua, S. K. N., Hernandez, M., Pascual, J. S. G., & Khu, K. J. O. (2024). The performance of ChatGPT versus neurosurgery residents in neurosurgical board examination-like questions: a systematic review and meta-analysis. *Neurosurg Rev*, *47*(1), 892. https://doi.org/10.1007/s10143-024-03144-y

Brady, C. L. (2019). *Undergraduate Nursing Faculty and Test Development: An Exploration into Their Understanding of Higher Order Thinking Test Questions*. Kent State University.

Cacciamani, G. E., Chu, T. N., Sanford, D. I., Abreu, A., Duddalwar, V., Oberai, A., Kuo, C. C. J., Liu, X., Denniston, A. K., Vasey, B., McCulloch, P., Wolff, R. F., Mallett, S., Mongan, J., Kahn, C. E., Sounderajah, V., Darzi, A., Dahm, P., Moons, K. G. M.,...Hung, A. J. (2023). PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nature Medicine*, *29*(1), 14-15. https://doi.org/10.1038/s41591-022-02139-w

Cleo, G., Scott, A. M., Islam, F., Julien, B., & Beller, E. (2019). Usability and acceptability of four systematic review automation software packages: a mixed method design. *Systematic Reviews*, *8*(1), 145. https://doi.org/10.1186/s13643-019-1069-6

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120.

Edwards, P. A. (2015). *The Effects of a Concept-Based Curriculum on Nursing Students' NCLEX-RN Exam Scores* Walden University].

García-Rudolph, A., David, S.-P., Mira, C. F., Sandra, C., Eloy, O., Elena, H.-P., & David, S.-P. (2024). How Chatbots Respond to NCLEX-RN Practice Questions: Assessment of Google Gemini, GPT-3.5, and GPT-4. *Nursing Education Perspectives*.

García-Rudolph, A., Sanchez-Pinsach, D., Fernandez-Mira, C., Cunyat, S., Opisso, E., & Hernandez-Pena, E. (2025). Accuracy analysis of AI chatbots GPT-3.5 and GEMINI on English NCLEX-style and Spanish EU general nursing multiple choice questions: challenges and performance insights. *Teaching and Learning in Nursing*, *20*(3), e730-e735. https://doi.org/https://doi.org/10.1016/j.teln.2025.02.013

Gunawan, J., Aungsuroch, Y., Marzilli, C., Nazliansyah, Chaerani, E., & Montayre, J. (2024). Artificial Intelligence Chatbot as Perceived by Nursing Students: A Qualitative Study. *SAGE Open*, *14*(4), 21582440241303453.

Hobensack, M., Von Gerich, H., Vyas, P., Withall, J., Peltonen, L.-M., Block, L. J., Davies, S., Chan, R., Van Bulck, L., & Cho, H. (2024). A rapid review on current and potential uses of large language models in nursing. *International Journal of Nursing Studies*, 104753.

Huang, H. (2023). Performance of ChatGPT on Registered Nurse License Exam in Taiwan: A Descriptive Study. *Healthcare (Basel)*, *11*(21). https://doi.org/10.3390/healthcare11212855

Jin, H. K., Lee, H. E., & Kim, E. (2024). Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. *BMC Medical Education*, *24*(1), 1013. https://bmcmededuc.biomedcentral.com/counter/pdf/10.1186/s12909-024-05944-8.pdf

Joanna Briggs Institute. (2017). *The Joanna Briggs Institute Critical Appraisal Tools for Use in JBI Systematic Reviews: Checklist for Analytical Cross Sectional Studies*. *Joanna Briggs Institute, Adelaide*.

. https://jbi.global/sites/default/files/2019-05/JBI_Critical_Appraisal-Checklist_for_Analytical_Cross_Sectional_Studies2017_0.pdf

Joseph, B., William, J. M., Doyoung, K., N, J., & P, D. (2019). Building a Method for Writing Clinical Judgment Items for Entry-Level Nursing Exams.

Kaneda, Y., Takahashi, R., Kaneda, U., Akashima, S., Okita, H., Misaki, S., Yamashiro, A., Ozaki, A., & Tanimoto, T. (2023). Assessing the Performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. *Cureus*, *15*(8), e42924. https://doi.org/10.7759/cureus.42924

Keshavarz, P., Bagherieh, S., Nabipoorashrafi, S. A., Chalian, H., Rahsepar, A. A., Kim, G. H. J., Hassani, C., Raman, S. S., & Bedayat, A. (2024). ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging*, *105*(7-8), 251-265. https://doi.org/10.1016/j.diii.2024.04.003

Krumsvik, R. J. (2024). Artificial intelligence in nurse education–A new sparring partner? GPT-4 capabilities of formative and summative assessment in National

Examination in Anatomy, Physiology, and Biochemistry. *Nordic Journal of Digital Literacy*(3), 172-186.

Lee, R. (2025). Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI). In R. Lee (Ed.), *Natural Language Processing : A Textbook with Python Implementation* (pp. 241-273). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-3208-4_10

Levin, G., Horesh, N., Brezinov, Y., & Meyer, R. (2023). Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. *Bjog*. https://doi.org/10.1111/1471-0528.17641

Lingle, W. M. (2024). *Faculty Experiences of Adaptation to the Updated Measures of Clinical Judgment on the Next Generation National Council Licensure Examination* (Publication Number 31329728) [Ph.D., William Carey University]. ProQuest One Academic. United States -- Mississippi.

Liu, J., Liu, F., Fang, J., & Liu, S. (2023). The application of Chat Generative Pre-trained Transformer in nursing education. *Nurs Outlook*, *71*(6), 102064. https://doi.org/10.1016/j.outlook.2023.102064

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res*, *26*, e60807. https://doi.org/10.2196/60807

Ma, L.-L., Wang, Y.-Y., Yang, Z.-H., Huang, D., Weng, H., & Zeng, X.-T. (2020). Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*, *7*(1), 7. https://doi.org/10.1186/s40779-020-00238-8

Miao, Y., Luo, Y., Zhao, Y., Li, J., Liu, M., Wang, H., . (2024). Performance of GPT-4 on Chinese Nursing Examination: Potentials for AI-Assisted Nursing Education Using Large Language Models. *Nurse Educator*, *49*, 10-1097.

Mohammadi-Shahboulaghi, F., Khankeh, H., & HosseinZadeh, T. (2021). Clinical reasoning in nursing students: A concept analysis. *Nurs Forum*, *56*(4), 1008-1014. https://doi.org/10.1111/nuf.12628

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,...Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, *372*, n71. https://doi.org/10.1136/bmj.n71

Salviano, M. E. M., Nascimento, P. D. F. S., Paula, M. A. d., Vieira, C. S., Frison, S. S., Maia, M. A., Souza, K. V., & Borges, E. L. (2016). Epistemology of nursing care: a reflection on its foundations. *Revista brasileira de enfermagem*, *69*, 1240-1245.

Schlegel, K., Sommer, N. R., & Mortillaro, M. (2025). Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, *3*(1), 80. https://doi.org/10.1038/s44271-025-00258-x

Simmons, B. (2010). Clinical reasoning: concept analysis. *Journal of Advanced Nursing*, *66*(5), 1151-1158. https://doi.org/10.1111/j.1365-2648.2010.05262.x

Singh, T. M., Reddy, C. K. K., Murthy, B. R., Nag, A., & Doss, S. (2025). Ai and education: Bridging the gap to personalized, efficient, and accessible learning. In *Internet of*

Behavior-Based Computational Intelligence for Smart Education Systems (pp. 131-160). IGI Global.

Stoll, C. R. T., Izadi, S., Fowler, S., Green, P., Suls, J., & Colditz, G. A. (2019). The value of a second reviewer for study selection in systematic reviews. *Res Synth Methods*, *10*(4), 539-545. https://doi.org/10.1002/jrsm.1369

Su, M.-C., Lin, L.-E., Lin, L.-H., & Chen, Y.-C. (2024). Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: Insights from Taiwan's nursing licensing exam. *International Journal of Nursing Studies*, 104717.

Sumbal, A., Sumbal, R., & Amir, A. (2024). Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. *J Med Educ Curric Dev*, *11*, 23821205241238641. https://doi.org/10.1177/23821205241238641

Taira, K., Itaya, T., & Hanada, A. (2023). Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study. *JMIR Nurs*, *6*, e47305. https://doi.org/10.2196/47305

Thorne, S. (2014). What constitutes core disciplinary knowledge? *Nursing Inquiry*, *21*(1), 1-2. https://doi.org/10.1111/nin.12062

Waldock, W. J., Zhang, J., Guni, A., Nabeel, A., Darzi, A., & Ashrafian, H. (2024). The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis. *J Med Internet Res*, *26*, e56532. https://doi.org/10.2196/56532

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*, 124167. https://doi.org/https://doi.org/10.1016/j.eswa.2024.124167

Wei, Q., Yao, Z., Cui, Y., Wei, B., Jin, Z., & Xu, X. (2024). Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J Biomed Inform*, *151*, 104620. https://doi.org/10.1016/j.jbi.2024.104620

Wu, Z., Gan, W., Xue, Z., Ni, Z., Zheng, X., & Zhang, Y. (2024). Performance of ChatGPT on Nursing Licensure Examinations in the United States and China: Cross-Sectional Study. *JMIR Med Educ*, *10*, e52746. https://doi.org/10.2196/52746

Yan, L. K., Niu, Q., Li, M., Zhang, Y., Yin, C. H., Fei, C., Peng, B., Bi, Z., Feng, P., & Chen, K. (2024). Large language model benchmarks in medical tasks. *arXiv preprint arXiv:2410.21348*.

Zhao, Q., Wang, H., Wang, R., & Cao, H. (2025). Deriving insights from enhanced accuracy: Leveraging prompt engineering in custom GPT for assessing Chinese Nursing Licensing Exam. *Nurse Educ Pract*, *84*, 104284. https://doi.org/10.1016/j.nepr.2025.104284

Zhou, Y., Li, S.-J., Tang, X.-Y., He, Y.-C., Ma, H.-M., Wang, A.-Q., Pei, R.-Y., & Piao, M.-H. (2024). Using ChatGPT in Nursing: Scoping Review of Current Opinions. *JMIR Med Educ*, *10*, e54297. https://doi.org/10.2196/54297

Zhu, S., Hu, W., Yang, Z., Yan, J., & Zhang, F. (2025). Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: Retrospective Cross-Sectional Comparative Study. *JMIR Med Inform*, *13*, e63731. https://doi.org/10.2196/63731

Zong, H., Li, J., Wu, E., Wu, R., Lu, J., & Shen, B. (2024). Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination

evaluation study for physicians, pharmacists and nurses. *BMC Med Educ*, *24*(1), 143. https://doi.org/10.1186/s12909-024-05125-7

9 Appendix

9.1 Appendix A: Search Strategy

Database	Search strategy	limits	Date	Hits
Pubmed	(ChatGPT OR GPT-4 OR Qwen OR Bard OR OpenAI OR "large language model" OR "generative AI" OR "natural language processing") AND ((nursing AND licensure) OR (nursing AND exam) OR NCLEX OR NNLE OR "nurse certification") AND (performance OR accuracy OR "question answering" OR evaluation OR "test results" OR "explanation quality")	2022-2025	04/05	17
PsyInfo	((ChatGPT or GPT4 or GPT3 or Bard or Qwen or Claude or Gemini or "large language model*" or LLM) and nursing and (performance or accuracy or "question answering" or evaluation or exam)).mp. [mp=title, abstract, heading word, table of contents, key concepts, original title, tests & measures, mesh word]	2022-2025	04/05	12
CINAHL	TI (chatgpt or ai or artificial intelligence) AND TI (examination or assessment or test) AND TI examination performance	2022-2025	05/05	54
EmCARE	(artificial intelligence/ or machine learning/ or (ChatGPT or GPT-4 or GPT-3 or "Generative Pre-trained Transformer" or "large language model*" or LLM or Chatbot or "generative AI" or "natural language processing").ti,ab.) and (nursing licensure/ or (NCLEX or NNLE or "nurse certification" or "nursing board exam*" or "nursing licensing exam*" or "RN examination*" or "registered nurse exam*").ti,ab.) and (academic performance or exam* performance or test* score* or question answering or accuracy or evaluation or performance or test results).ti,ab.	2022-2025	1/06	6
ERIC (OVID)	(("artificial intelligence" or "machine learning" or "ChatGPT" or "GPT-4" or "GPT-3" or "Generative Pre-trained Transformer" or "large language model*" or "LLM" or "Chatbot" or "generative AI" or "natural language processing") and ("professional licensure" or "nursing licensure" or "certification" or "NCLEX" or "NNLE" or "nursing board exam*" or "nursing licensing exam*" or "RN examination*" or "registered nurse exam*") and ("academic achievement" or "test performance" or "exam performance" or "student evaluation" or "question answering" or "accuracy" or "test scores" or "evaluation")).mp. [mp=abstract, title, heading word, identifiers]	2022-2025	1/06	12

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 1.SupplementaryFileS1aPRISMA2020checklist.docx
- 1a.SupplementaryFileS1bPRISMA2020abstractchecklist.docx
- 2.SupplementaryfileS2Dataextractiontemplate.docx
- 3.SupplementaryFileS4StandardisedClassification.docx
- 4.SupplementaryFileS3SummaryofStudiesExcluded.docx
- 5.SupplementaryFileS5QualityAssessment.docx
- 6.SupplementaryFileS8Funnelplots.docx
- 7.SupplementaryFileS6Metaregression.docx
- 8.SupplementaryFileS7Sensitivityanalysis.docx