I'm interested in the use of large language models (e.g., GPT-4, Claude) for summarizing scientific research, especially in the biomedical field. Please find peer-reviewed studies that evaluate how factually accurate these models are when generating abstracts or summaries of full-text articles. Focus on papers that include benchmarking, error analysis, or human expert evaluation

A review of 26 studies shows that while large language models achieve high accuracy in some biomedical summarization tasks (up to 100% for systematic reviews), they still struggle with factual accuracy and hallucinations in more complex specialized tasks (as low as 25%).

Abstract

Large language models such as GPT-4, GPT-3.5, Claude, and others yield biomedical summaries that are promising yet imperfect. Li et al. (2024) report that a knowledge-guided GPT-4 summary improved task success by 13.1% but still suffered from factual inconsistencies. Chen et al. (2023) observed that GPT-4, when fine-tuned, achieved around 80% macro F1 in biomedical natural language processing, despite hallucinations and incomplete outputs. In emergency medicine discharge summaries, Williams et al. (2024) recorded 33% error-free outputs for GPT-4 compared with 10% for GPT-3.5-turbo, while Hake et al. (2024) noted a median accuracy of 92.5% in general medicine summaries.

Across 26 studies spanning oncology, systematic reviews, clinical notes, and coding tasks, performance metrics ranged widely—from accuracy as low as 25–37% in some complex or specialized tasks (e.g., neurodegenerative disease summaries and medical coding) to sensitivity and precision scores near 1.0 in systematic review screening. Common error types include hallucinations, omissions, and domain-specific inaccuracies. Retrieval-augmented methods and fine-tuning approaches improved factual accuracy in studies such as Dai et al. (2024) and Sanghera et al. (2025). These findings indicate that while models like GPT-4 routinely outperform earlier versions and some human benchmarks in sensitivity and overall accuracy, challenges in preserving full factual integrity remain evident in biomedical summarization.

Paper search

Using your research question "I'm interested in the use of large language models (e.g., GPT-4, Claude) for summarizing scientific research, especially in the biomedical field. Please find peer-reviewed studies that evaluate how factually accurate these models are when generating abstracts or summaries of full-text articles. Focus on papers that include benchmarking, error analysis, or human expert evaluation", we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 500 papers most relevant to the query.

Screening

We screened in sources that met these criteria:

- **LLM Biomedical Focus**: Does the study evaluate large language models (LLMs) specifically for summarizing biomedical or healthcare-related scientific literature?

- **Accuracy Assessment**: Does the study include a quantitative or qualitative assessment comparing the accuracy of LLM-generated summaries to the source material?
- **Validation Method**: Does the study employ at least one validated evaluation methodology (expert human evaluation, automated benchmarking, or systematic error analysis)?
- **Study Type**: Is the paper an original research article or systematic review presenting empirical evaluation data?
- **Peer Review**: Was the study published in a peer-reviewed journal or conference proceedings?
- **Empirical Evidence**: Does the study present actual evaluation data rather than theoretical discussion or opinion?
- **LLM Technology**: Does the study focus on large language models rather than traditional extractive summarization methods or smaller language models?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **Study Design**:

  Identify the specific type of study design used to evaluate large language models (LLMs) in scientific research summarization. Look for terms like:

- Systematic review
- Benchmarking study
- Comparative analysis
- Performance evaluation
- Human expert validation study

If multiple design elements are present, list all that apply. If the design is not clearly stated, note "Design not explicitly specified" and provide any contextual details from the methods section that describe the research approach.

- **Language Models Evaluated**:

  List ALL specific large language models tested in the study, including:

- Full model name
- Version number (if provided)
- Parameter size (if mentioned)

Examples might include:

- GPT-4
- Claude 3 Opus
- Gemini 1.5 Pro
- PaLM

If models were compared, note the specific comparisons made. If multiple models were used in an ensemble, list all models and indicate their role in the ensemble.

- **Evaluation Metrics**:

  Identify ALL quantitative and qualitative metrics used to assess the language models' performance:

- Quantitative metrics (e.g., accuracy, precision, sensitivity)
- Qualitative assessment criteria (e.g., factuality, comprehension, reasoning)
- Specific scoring or ranking methods

Record the exact metrics used, their numerical values if provided, and the context of their application (e.g., objective questions, open-ended questions, diagnostic accuracy).

If multiple evaluation approaches were used, list them in order of prominence or as described in the study.

- **Sample Characteristics**:

  Document the following details about the evaluation sample:

- Total number of scientific articles or texts used
- Total number of questions or tasks
- Domains of scientific literature (e.g., biomedical, clinical research)
- Source of texts (e.g., PubMed, specific journal collections)

If multiple datasets or sources were used, list each separately with its specific characteristics. If sampling methodology is described, include brief details about how texts were selected.

- **Performance Outcomes**:

  Extract the primary performance results for each language model:

- Specific accuracy percentages
- Comparative performance rankings
- Statistically significant differences between models
- Any notable limitations or error patterns identified

Focus on results directly related to factual accuracy in scientific research summarization. Include numerical results and any qualitative assessments of model performance.

If human expert performance is compared, include those comparative results as well.

Results

Characteristics of Included Studies

Study	Study Design	Evaluation Method	Biomedical Domain	Large Language Models Evaluated	Full text retrieved
Li et al., 2024	Comparative analysis, Performance evaluation	Benchmarking, Human evaluation	Evidence-based medicine, Biomedical	GPT-4, 6 others (No mention found)	No

Study	Study Design	Evaluation Method	Biomedical Domain	Large Language Models Evaluated	Full text retrieved
Chen et al., 2023	Benchmarking, Comparative analysis, Performance evaluation, Human expert validation	Quantitative & qualitative benchmarking, Human review	Biomedical natural language processing	GPT-3.5, GPT-4, LLaMA 2, PMC LLaMA	Yes
Rydzewski et al., 2024	Comparative analysis, Benchmarking, Performance evaluation	Large-scale question answering, Human comparison	Oncology	LLaMA 1, PaLM 2, Claude-v1, GPT-3.5, GPT-4, GPT-4 Turbo, Gemini 1.0 Ultra, Mixtral 8×7B, LLaMA 2	Yes
Williams et al., 2024	Cross-sectional, Performance evaluation, Human expert validation, Comparative analysis	Discharge summary generation, Expert review	Emergency medicine	GPT-4, GPT-3.5-turbo	Yes
Hake et al., 2024	Performance evaluation, Human expert validation, Comparative analysis	Summarization, Physician rating	General medicine, Multispecialty	ChatGPT-3.5	Yes
Bianchi et al., 2025	Benchmarking, Human expert validation, Performance evaluation	Question and answer benchmark, Expert annotation	Neurodegenerative diseases	Claude-3.5-Sonnet, ChatGPT-4o, 5 others (No mention found)	No
Wang et al., 2025	Systematic review, Network meta-analysis	Meta-analysis, Bayesian ranking	Clinical research	ChatGPT-4o, Aeyeconsult, ChatGPT-4, Claude 3 Opus, Gemini	No

Study	Study Design	Evaluation Method	Biomedical Domain	Large Language Models Evaluated	Full text retrieved
Singhal et al., 2022	Benchmarking, Human expert validation, Comparative analysis, Performance evaluation	MultiMedQA, Human evaluation	General medicine, Consumer health	PaLM, Flan-PaLM, Med-PaLM	Yes
Ben Abacha et al., 2024	Benchmarking, Performance evaluation, Comparative analysis, Human expert validation	Error detection/correction, Physician comparison	Clinical notes	Phi-3-7B, Claude 3.5 Sonnet, Gemini 2.0 Flash, ChatGPT, GPT-4, GPT-4o, GPT-4o-mini, o1-mini, o1-preview	Yes
Sanghera et al., 2024	Performance evaluation, Benchmarking, Comparative analysis, Human expert validation	Abstract screening, Human/LLM ensemble	Systematic reviews	GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o, Llama 3 70B, Gemini 1.5 Pro, Claude Sonnet 3.5	Yes
Sanghera et al., 2025	Comparative analysis, Performance evaluation, Human expert validation, Benchmarking	Retrieval-augmented question and answer, Clinician panel	Clinical medicine	Almanac, ChatGPT-4, Bing, Bard, Vicuna-7B, BERT, BioBERT, RoBERTa, SapBERT, QA-GNN	Yes
Waldock et al., 2024	Performance evaluation, Benchmarking, Comparative analysis, Human expert validation	Abstract screening, Human/LLM ensemble	Systematic reviews	GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o, Llama 3 70B, Gemini 1.5 Pro, Claude Sonnet 3.5	Yes

Study	Study Design	Evaluation Method	Biomedical Domain	Large Language Models Evaluated	Full text retrieved
Soroush et al., 2024	Systematic review, Meta-analysis	Meta-analysis, Accuracy quantification	Medical exams	ChatGPT	No
Montenegro et al., 2025	Benchmarking, Performance evaluation, Human expert validation	Medical code querying, Manual grading	Medical coding	GPT-3.5 Turbo, GPT-4, Gemini Pro, Llama2-70b Chat	Yes
Liu et al., 2025	Systematic review, Benchmarking, Comparative analysis, Performance evaluation	Review, Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), Benchmarking	Synthetic medical text	No mention found	No
Dai et al., 2024	Systematic review, Meta-analysis, Comparative analysis	Meta-analysis, Retrieval-augmented generation (RAG) vs. baseline	Biomedicine	No mention found	No
Gartlehner et al., 2025	Diagnostic study, Comparative analysis, Performance evaluation, Human expert validation	Literature screening, Human comparison	Thoracic surgery	ChatGPT-4o, Claude-3.5 sonnet, Gemini-1.5 pro	No
Alkalbani et al., 2025	Study Within a Review (SWAR), Comparative analysis, Performance evaluation	Data extraction, Human adjudication	Systematic reviews	Claude 2.1, 3.0 Opus, 3.5 Sonnet	Yes

Study	Study Design	Evaluation Method	Biomedical Domain	Large Language Models Evaluated	Full text retrieved
Ray et al., 2025	Systematic review, Comparative analysis, Human expert validation	Review, Human validation	Medical specialties	GPT-4, GPT-3.5, BERT-based, CancerBERT, medBERT.de, SurgicBERTa	Yes
Sushil et al., 2024	Cross-sectional, Comparative analysis, Performance evaluation, Human expert validation	Translation, Equivalence testing	Patient instructions	GPT-4o	No
Mudrik et al., 2024	Comparative analysis, Benchmarking, Performance evaluation	Pathology report classification	Breast cancer pathology	GPT-4, GPT-3.5, Starling, ClinicalCamel	No
Zurita et al., 2025	Systematic review, Performance evaluation, Comparative analysis	Review, Narrative synthesis	Hematology	GPT-3.5, GPT-4, Bard (Gemini), Bing	Yes
Wu et al., 2024	Performance evaluation, Comparative analysis, Human expert validation	Information extraction, Human comparison	Oncology electronic health records	GPT-3.5-turbo-1106, GPT-4-1106-preview	No
Guo et al., 2022	Benchmarking, Comparative analysis, Performance evaluation	Multiple choice question answering, Subspecialty	Nephrology	Llama2-70B, Koala 7B, Falcon 7B, Stable-Vicuna 13B, Orca-Mini 13B, GPT-4, Claude 2	No
Yue Guo et al., 2022	Performance evaluation, Comparative analysis	Lay language generation, Retrieval augmentation	Biomedical	Llama 2, GPT-4	No

Study Design and Evaluation Methods:

- Comparative analysis and performance evaluation:Most common, each used in 21 studies.
- Benchmarking:Used in 14 studies.
- Human expert validation:Reported in 14 studies.
- Systematic review or meta-analysis:Used in 6 and 4 studies, respectively.
- Other designs:Cross-sectional (2 studies), diagnostic study (1 study), and Study Within a Review (SWAR, 1 study).
- Human evaluation, review, or comparison:Described in 8 studies.
- Other evaluation methods:Each in 1–2 studies, including abstract screening, summarization, question and answer benchmarking, discharge summary generation, error detection/correction, translation, information extraction, pathology report classification, medical code querying, data extraction, and lay language generation.

Biomedical Domains:

- Systematic reviews:Most common specific domain (3 studies).
- General medicine:Addressed in 2 studies; biomedical (general) in 2 studies.
- Other domains:Each in 1 study, including multispecialty, evidence-based medicine, biomedical natural language processing, oncology, oncology electronic health records, emergency medicine, neurodegenerative diseases, clinical research, consumer health, clinical notes, clinical medicine, medical exams, medical coding, synthetic medical text, biomedicine, thoracic surgery, medical specialties, patient instructions, breast cancer pathology, hematology, and nephrology.

Large Language Models Evaluated:

- GPT-4:Most frequently evaluated, included in 12 studies.
- GPT-4o:Evaluated in 7 studies.
- GPT-3.5 and variants:Appeared in 8 studies.
- LLaMA and variants:Evaluated in 8 studies.
- Claude models:Appeared in 7 studies.
- Gemini and variants:Evaluated in 7 studies.
- ChatGPT and variants:Appeared in 9 studies.
- Other models:Each evaluated in 1 study, including PaLM, Flan-PaLM, Med-PaLM, PMC LLaMA, Mixtral 8×7B, Starling, ClinicalCamel, Almanac, Aeyeconsult, BERT-based, CancerBERT, medBERT.de, SurgicBERTa, BERT, BioBERT, RoBERTa, SapBERT, QA-GNN, Koala 7B, Falcon 7B, Stable-Vicuna 13B, Orca-Mini 13B, o1-mini, o1-preview, 3.0 Opus, 3.5 Sonnet, Phi-3-7B.
- No mention found:We didn't find mention of the specific large language models evaluated in 4 studies.

Factual Accuracy Metrics and Benchmarks

Study	Evaluation Metric	Average Performance	Error Types	Key Findings
Li et al., 2024	Percentage improvement, task success, human evaluation	+13.1% (PICO, GPT-4, knowledge-guided)	Factual inconsistencies, domain inaccuracies	Li et al., 2024, found that large language models performed well in summarization, but were below state-of-the-art in named entity recognition; factual errors persisted.
Chen et al., 2023	Macro F1, entity-level F1, cost	~80% (GPT-4, USMLE), 20% gain (MedQA)	Inconsistencies, hallucinations, incompleteness	Chen et al., 2023, found GPT-4 performed best overall; state-of-the-art fine-tuning outperformed zero- or few-shot approaches; cost-performance tradeoff noted.
Rydzewski et al., 2024	Accuracy, confidence, human benchmark	GPT-4: 68.7%, others lower	Overconfidence, fixed false beliefs	GPT-4 was the only model above the 50th percentile compared to humans; high error rates remained.
Williams et al., 2024	Error-free percentage, error counts, readability	GPT-4: 33% error-free, GPT-3.5: 10%	Hallucinations, omissions, inaccuracies	GPT-4 was more accurate, but omissions and hallucinations were common.
Hake et al., 2024	Quality, accuracy, bias (0-100)	Accuracy median 92.5	Minor inaccuracies, rare hallucinations	High-quality summaries, rare serious errors, modest relevance classification.
Bianchi et al., 2025	Response Quality Rate, Safety Rate	Claude-3.5: 25%/76%, GPT-4o: 37%/31%	Unsafe, low-quality responses	State-of-the-art large language models had fundamental gaps in complex biomedicine.

Study	Evaluation Metric	Average Performance	Error Types	Key Findings
Wang et al., 2025	SUCRA (Surface Under the Cumulative Ranking), accuracy	ChatGPT-4o: 0.92 (objective questions)	No mention found	Large language models excelled in some question types, but humans were best in diagnosis.
Singhal et al., 2022	Accuracy, consensus, harm, comprehension	Flan-PaLM: 67.6% (MedQA)	Harm, bias, incomplete	Med-PaLM was closer to clinicians, but still inferior.
Ben Abacha et al., 2024	Error detection/correction accuracy	Claude 3.5: 70.2% (flag), o1-preview: 0.698 (correction)	Hallucinated errors, low precision	Large language models performed well but not as well as physicians; error detection was rare in pretraining.
Sanghera et al., 2024	Sensitivity, precision, F1, kappa	Sensitivity: 1.0, Precision: 0.93 (development set)	Low precision (large set), review variation	Large language models outperformed humans in sensitivity, but precision dropped with class imbalance.
Sanghera et al., 2025	Accuracy, factuality, completeness	Vicuna-7B+retrieval: 48.5%	No mention found	Retrieval improved accuracy; more facts led to better performance.
Waldock et al., 2024	Sensitivity, precision, kappa	Sensitivity: 1.0, Precision: 0.93 (development set)	Precision drop (large set)	Large language models outperformed humans in sensitivity and precision; workload reduction observed.
Soroush et al., 2024	Accuracy, sensitivity, precision	Large language models: 0.61, ChatGPT: 0.64	No mention found	Large language models were promising, but not at human level.
Montenegro et al., 2025	Exact match, similarity, CodeSTS (Code Semantic Textual Similarity)	GPT-4: 45.9% (ICD-9), 33.9% (ICD-10)	Fabricated, imprecise codes	Large language models performed poorly at coding, not ready for clinical use.

Study	Evaluation Metric	Average Performance	Error Types	Key Findings
Liu et al., 2025	ROUGE, BLEU, qualitative	No mention found	Hallucinations, factual errors	Hybrid or retrieval approaches improved accuracy, but errors persisted.
Dai et al., 2024	Odds ratio (retrieval-augmented generation vs. baseline)	Odds ratio: 1.35 (95% CI 1.19-1.53)	No mention found	Retrieval-augmented generation improved large language model performance.
Gartlehner et al., 2025	Sensitivity, specificity, AUC (Area Under the Curve)	Sensitivity: 0.87, Specificity: 0.96, AUC: 0.96	No mention found	Large language models outperformed machine learning tools in screening accuracy.
Alkalbani et al., 2025	Accuracy, recall, precision, F1	Accuracy: 91.0%, Recall: 89.4%, Precision: 98.9%	Missed data	Large language models saved time and had fewer errors than humans.
Ray et al., 2025	Accuracy, F1, recall, Likert	3–90% (varies by task)	Hallucinations, factual errors	GPT-4 performed best, but humans were still superior in many cases.
Sushil et al., 2024	MQM (Multidimensional Quality Metrics, 0-100), error rates	No significant difference vs. humans	Fewer mistranslations	GPT-4o was equivalent to human translators for Spanish instructions.
Mudrik et al., 2024	Macro F1	GPT-4: 0.86, LSTM-Att: 0.75	Inference, task design	GPT-4 outperformed supervised models, especially with label imbalance.
Zurita et al., 2025	Diagnostic accuracy, completeness	GPT-4: 76–88% (varies)	Reference errors, outdated information	GPT-4 outperformed others, but only 59% match with experts.

Study	Evaluation Metric	Average Performance	Error Types	Key Findings
Wu et al., 2024	Sensitivity, precision, McNemar	GPT-4: Sensitivity 96.8%, Precision 96.8%	False positives	GPT-4 outperformed humans in sensitivity, with slightly lower precision.
Guo et al., 2022	Percentage correct multiple choice question	GPT-4: 73.3%, Claude 2: 54.4%	No mention found	GPT-4 outperformed Claude 2 and open-source large language models.
Yue Guo et al., 2022	Qualitative (summary quality)	No mention found	Factuality, simplicity	Retrieval improved quality, but not ideal.

Evaluation Metrics:

- Accuracy or equivalent:Most common metric, used in 12 studies.
- F1 score (macro/entity-level):Used in 5 studies.
- Sensitivity and precision:Each used in 5 studies.
- Recall:Used in 2 studies.
- Completeness, quality (response/summary), and human evaluation/benchmark:Each used in 2 studies.
- Other metrics:Used in 1 study each, including safety, specificity, area under the curve, cost, consensus, harm, comprehension, kappa, bias, confidence, surface under the cumulative ranking, odds ratio, multidimensional quality metrics, ROUGE/BLEU, exact match/similarity/CodeSTS, Likert, and McNemar.
- No mention found:We didn't find mention of evaluation metric information in the available abstracts or full texts for some studies.

Average Performance:

- Accuracy or equivalent values:Reported in 13 studies, with most values between 60% and 90%. The lowest reported was 33.9% (ICD-10 coding), and the highest was 92.5% (accuracy median).
- F1 scores:Reported in 5 studies, with values around 0.80–0.86 where specified.
- Sensitivity and precision:Often high (up to 1.0 in 2 studies), but some studies reported lower values (e.g., 0.61).
- Quality and safety rates:Claude-3.5 had 25% response quality and 76% safety, while GPT-4o had 37%/31%.
- Error-free outputs:One study reported 33% error-free outputs for GPT-4.
- Odds ratio for retrieval-augmented generation vs. baseline:1.35 in one study.
- Area under the curve:0.96 in one study.
- No mention found:We didn't find mention of average performance values in 2 studies.

Error Types:

- Hallucinations:Reported in 6 studies.
- Factual errors or inconsistencies:Reported in 4 studies.
- Incompleteness or omissions:Reported in 4 studies.
- Low precision:Reported in 3 studies.
- Inaccuracy, unsafe/low-quality responses, overconfidence/false beliefs, fabrication/imprecision, harm/bias, and reference/outdated information:Each reported in 2 studies.
- Other error types:Reported in 1 study each, including false positives, mistranslation, simplicity, review variation, and task design/inference.
- No mention found:We didn't find mention of error type details in 6 studies.

Domain-Specific Performance

Study	Medical Domain	Accuracy Range	Common Errors	Success Factors
Li et al., 2024	Evidence-based medicine, Biomedical	No mention found	Factual inconsistencies, domain errors	Prompting, knowledge-guided
Chen et al., 2023	Biomedical natural language processing	~80% (GPT-4, USMLE)	Hallucinations, incompleteness	Fine-tuning, more shots
Rydzewski et al., 2024	Oncology	25.6–68.7%	Overconfidence, bias	Model selection, prompt repetition
Williams et al., 2024	Emergency medicine	10–33% error-free	Hallucinations, omissions	Model choice (GPT-4 over 3.5)
Hake et al., 2024	General medicine	92.5 (median)	Minor inaccuracies	Shorter summaries, high readability
Bianchi et al., 2025	Neurodegenerative diseases	25–37% (quality rate)	Unsafe, low-quality	No mention found
Wang et al., 2025	Clinical research	0.87–0.97 (surface under the cumulative ranking)	No mention found	Model-task fit
Singhal et al., 2022	General medicine	67.6% (MedQA)	Harm, bias	Instruction tuning
Ben Abacha et al., 2024	Clinical notes	65–70% (detection)	Hallucinated errors	No mention found
Sanghera et al., 2024	Systematic reviews	Sensitivity 1.0, Precision 0.93	Low precision (large set)	Prompting, ensemble
Sanghera et al., 2025	Clinical question and answer	48.5% (best)	No mention found	Retrieval, more facts
Waldock et al., 2024	Systematic reviews	Sensitivity 1.0, Precision 0.93	Precision drop	Ensemble, validation
Soroush et al., 2024	Medical exams	0.51–0.64	No mention found	No mention found
Montenegro et al., 2025	Medical coding	1.2–49.8%	Fabricated, imprecise codes	Code frequency, short descriptions

Study	Medical Domain	Accuracy Range	Common Errors	Success Factors
Liu et al., 2025	Synthetic text	No mention found	Hallucinations	Hybrid, structure
Dai et al., 2024	Biomedicine	Odds ratio 1.35	No mention found	Retrieval-augmented generation
Gartlehner et al., 2025	Thoracic surgery	Sensitivity 0.87, Specificity 0.96	No mention found	Prompt revision
Alkalbani et al., 2025	Systematic reviews	Accuracy 91.0%	Missed data	Human verification
Ray et al., 2025	Medical specialties	3–90%	Hallucinations	Fine-tuning, domain large language models
Sushil et al., 2024	Patient instructions	Equivalent to human	Fewer mistranslations	Prompting
Mudrik et al., 2024	Pathology	0.86 (macro F1)	Inference, task design	Zero-shot, label imbalance
Zurita et al., 2025	Hematology	53–88%	Reference errors	Model choice (GPT-4)
Wu et al., 2024	Oncology electronic health records	Sensitivity 96.8%, Precision 96.8%	False positives	Prompting
Guo et al., 2022	Nephrology	17.1–73.3%	No mention found	Model choice (GPT-4)
Yue Guo et al., 2022	Biomedical	No mention found	Factuality	Retrieval

Accuracy Range:

- Reported accuracy or quality rates:Found in 16 studies, with values ranging from as low as 1.2% to as high as 92.5%.
  - 2 studies reported values below 20%.
  - 4 studies reported values between 20–50%.
  - 4 studies reported values between 50–70%.
  - 4 studies reported values between 67–92%.
  - 1 study reported ~80%.
  - 1 study reported macro F1 of 0.86.
  - 1 study described performance as equivalent to humans.
  - 1 study described performance as ”not ideal.”
- Sensitivity, precision, or specificity metrics:Found in 5 studies, with sensitivity ranging from 0.87 to 1.0 and precision from 0.93 to 0.968.
- Odds ratio:Found in 1 study (odds ratio 1.35).
- No mention found:We didn't find mention of accuracy information in 3 studies.

Common Errors:

- Hallucinations:Reported in 5 studies.

- Bias:Reported in 2 studies.
- Factual inconsistency or factuality issues:Reported in 2 studies.
- Incompleteness, omissions, or missed data:Reported in 3 studies.
- Unsafe, low-quality, or harm-related errors:Reported in 3 studies.
- Overconfidence, minor inaccuracies, fabricated/imprecise codes, reference errors, false positives, low precision/precision drop, mistranslations, and inference/task design errors:Each reported in 1–2 studies.
- No mention found:We didn't find mention of common errors in 6 studies.

Success Factors:

- Prompting or prompt revision:Identified as a success factor in 5 studies.
- Fine-tuning or instruction tuning:Reported in 3 studies.
- Model choice or selection:Reported in 4 studies.
- Ensemble methods:Reported in 2 studies.
- Retrieval or retrieval-augmented generation approaches:Reported in 3 studies.
- Other factors:Human verification, knowledge-guided approaches, more shots/more facts, validation, shorter summaries/high readability, code frequency/short descriptions, hybrid/structure, zero-shot/label imbalance, domain large language models, and model-task fit were each reported in 1–2 studies.
- No mention found:We didn't find mention of success factors in 3 studies.

Evaluation Methodologies and Expert Assessment

- Combination of automated and human evaluation:Most studies used both automated metrics (such as accuracy, F1 score, surface under the cumulative ranking, multidimensional quality metrics) and human expert validation.
- Role of human review:Human review was used to assess factuality, completeness, and error types, often revealing issues not captured by automated metrics.
- Structured frameworks:Some studies used structured frameworks, such as multidimensional quality metrics for translation, Code Semantic Textual Similarity for coding, and Likert scales for summary quality.
- Expert panel size and expertise:The number and expertise of human reviewers varied, affecting the robustness of expert assessment. Studies with larger, more diverse expert panels (such as Sanghera et al., 2025; Ben Abacha et al., 2024) provided more reliable validation.

Error Patterns and Mitigation Strategies

- Common error patterns:Hallucinations (fabricated facts), omissions (missing key information), and domain-specific inaccuracies (misinterpretation of specialized content) were frequently reported.
- Hallucinations:Particularly problematic in tasks requiring deep reasoning or synthesis.
- Mitigation strategies:
  - Retrieval-augmented and hybrid models (such as retrieval-augmented generation, structured prompting) reduced hallucinations and improved factual accuracy.

  - Human-in-the-loop approaches, ensemble methods, and domain-specific fine-tuning were effective in reducing errors.
- Persistence of errors:Even with these strategies, nontrivial error rates persisted, highlighting the need for ongoing human oversight and further methodological development.

References

A. Alkalbani, Ahmed Salim Alrawahi, Ahmad Salah, Venus Haghighi, Yang Zhang, Salam Alkindi, and Qaun Z. Sheng. “A Systematic Review of Large Language Models in Medical Specialties: Applications, Challenges and Future Directions.” *Information*, 2025.

A. Mudrik, G. Nadkarni, O. Efros, Benjamin S. Glicksberg, E. Klang, and S. Soffer. “Exploring the Role of Large Language Models (LLMs) in Hematology: A Systematic Review of Applications, Benefits, and Limitations.” *medRxiv*, 2024.

Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Y. Barash, Robert Freeman, Alexander W. Charney, G. Nadkarni, and Eyal Klang. “Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying.” *NEJM AI*, 2024.

Amadeo Jesus Wals Zurita, Hector Miras Del Rio, Nerea Ugarte Ruiz de Aguirre, Cristina Nebrera Navarro, Maria Rubio Jimenez, David Muñoz Carmona, and Carlos Miguez Sanchez. “The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis.” *JMIR Medical Informatics*, 2025.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetişgen, Fei Xia, and Thomas Lin. “MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes.” *arXiv.org*, 2024.

C. Y. Williams, J. Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, A. Butte, and A. Kornblith. “Evaluating Large Language Models for Drafting Emergency Department Discharge Summaries.” *medRxiv*, 2024.

Cyril Zakka, R. Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, M. Moor, R. Fong, et al. “Almanac - Retrieval-Augmented Language Models for Clinical Medicine.” *NEJM AI*, 2024.

Jin Li, Yiyan Deng, Qi Sun, Junjie Zhu, Yu Tian, Jingsong Li, and Tingting Zhu. “Benchmarking Large Language Models in Evidence-Based Medicine.” *IEEE Journal of Biomedical and Health Informatics*, 2024.

Joel Hake, Miles Crowley, Allison Coy, D. Shanks, Aundria Eoff, Kalee Kirmer-Voss, Gurpreet Dhanda, and D. J. Parente. “Quality, Accuracy, and Bias in ChatGPT-Based Summarization of Medical Abstracts.” *Annals of Family Medicine*, 2024.

K. Singhal, Shekoofeh Azizi, T. Tu, S. Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, et al. “Large Language Models Encode Clinical Knowledge.” *Nature*, 2022.

Larissa Montenegro, Luis M. Gomes, and José Machado. “What We Know About the Role of Large Language Models for Medical Synthetic Dataset Generation.” *Applied Informatics*, 2025.

Ling Wang, Jinglin Li, Boyang Zhuang, Shasha Huang, Meilin Fang, Cunze Wang, Wen Li, Mohan Zhang, and Shurong Gong. “Accuracy of Large Language Models When Answering Clinical Research Questions: Systematic Review and Network Meta-Analysis.” *Journal of Medical Internet Research*, 2025.

Madhumita Sushil, T. Zack, Divneet Mandair, Zhiwei Zheng, Ahmed Wali, Yan-Ning Yu, Yuwei Quan, D. Lituiev, and A. Butte. “A Comparative Study of Large Language Model-Based Zero-Shot Inference and Task-Specific Supervised Classification of Breast Cancer Pathology Reports.” *J. Am. Medical Informatics Assoc.*, 2024.

MD Mph Gerald Gartlehner, Mlis Shannon Kugley, PhD Karen Crotty, PhD Meera Viswanathan, MD PhD Andreea Dobrescu, PhD Barbara Nussbaumer-Streit, Bsph Graham Booth, et al. “AI-Assisted Data

Extraction with a Large Language Model: A Study Within Reviews." *medRxiv*, 2025.

Mondira Ray, Daniel J. Kats, Joss Moorkens, Dinesh Rai, Nate Shaar, Diane Quinones, Alejandro Vermeulen, et al. "Evaluating a Large Language Model in Translating Patient Instructions to Spanish Using a Standardized Framework." *JAMA Pediatrics*, 2025.

N. Rydzewski, Deepak Dinakaran, Shuang G. Zhao, E. Ruppin, B. Turkbey, D. E. Citrin, and Krishnan R. Patel. "Comparative Evaluation of LLMs in Clinical Oncology." *NEJM AI*, 2024.

Owen Bianchi, Maya Willey, Chelsea X. Alvarado, Benjamin P. Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, et al. "CARDBiomedBench: A Benchmark for Evaluating Large Language Model Performance in Biomedical Research." *bioRxiv*, 2025.

Qingyu Chen, Jingcheng Du, Yan Hu, V. Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Huan Xu. "Benchmarking Large Language Models for Biomedical Natural Language Processing Applications and Recommendations." *Nature Communications*, 2023.

Rohan Sanghera, A. J. Thirunavukarasu, Marc El Khoury, Jessica O'Logbon, Yuqing Chen, Archie Watt, Mustafa Mahmood, Hamid Butt, George Nishimura, and Andrew A S Soltan. "High-Performance Automated Abstract Screening with Large Language Model Ensembles." *J. Am. Medical Informatics Assoc.*, 2025.

Rohan Sanghera, Arun J. Thirunavukarasu, Marc El Khoury, Jessica O'Logbon, Yuqing Chen, Archie Watt, Mustafa Mahmood, Hamid Butt, George Nishimura, and Andrew A S Soltan. "High-Performance Automated Abstract Screening with Large Language Model Ensembles." *arXiv.org*, 2024.

Sean Wu, Michael Koo, L. Blum, Andy Black, Liyo Kao, Zhe Fei, Fabien Scalzo, and Ira Kurtz. "Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology." *NEJM AI*, 2024.

Siru Liu, Allison B. McCoy, and Adam Wright. "Improving Large Language Model Applications in Biomedicine with Retrieval-Augmented Generation: A Systematic Review, Meta-Analysis, and Clinical Development Guidelines." *J. Am. Medical Informatics Assoc.*, 2025.

William J Waldock, Joe Zhang, Ahmad Guni, Ahmad Nabeel, A. Darzi, and H. Ashrafian. "The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research*, 2024.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and T. Cohen. "Retrieval Augmentation of Large Language Models for Lay Language Generation." *Journal of Biomedical Informatics*, 2022.

Zhang-Yi Dai, Fu-Qiang Wang, Cheng Shen, Yan-Li Ji, Zhi-Yang Li, Yun Wang, and Qiang Pu. "Accuracy of Large Language Models for Literature Screening in Thoracic Surgery: Diagnostic Study." *Journal of Medical Internet Research*, 2024.