Elicit Prompt for Literature Review Research Question How accurate are large language models (LLMs) for summarizing scientific papers in microbiology and microbial genomics, and what methodologies exist for systematically evaluating AI-generated summaries of scientific literature? Literature Search Objectives Primary Focus: Find studies that evaluate the accuracy, reliability, and quality of AI-generated summaries of scientific papers, particularly in: Microbiology and microbial genomics Biomedical sciences and life sciences (as broader context) Scientific literature processing and analysis Secondary Focus: Identify methodological frameworks for: Comparing multiple AI models/services Human-AI hybrid evaluation approaches Metrics and assessment criteria for summary quality Systematic evaluation designs for NLP tasks in scientific domains Specific Search Parameters Keywords and Concepts LLM evaluation, GPT evaluation, language model assessment Scientific paper summarization, biomedical text summarization AI accuracy in scientific literature, automated literature review Human-AI evaluation, mixed-methods assessment Microbiology informatics, genomics text mining Summary quality metrics, ROUGE scores, semantic similarity Comparative AI model studies, benchmarking NLP systems Paper Types Sought Empirical studies evaluating LLM performance on scientific texts Methodological papers describing evaluation frameworks for AI summarization Comparative studies testing multiple AI models or services Domain-specific evaluations in life sciences, biology, or medicine Reviews or surveys of AI applications in scientific literature processing Time Frame Focus on papers from 2020-2025 (covering the modern LLM era), but include seminal earlier works on text summarization evaluation if highly relevant. Quality Criteria Peer-reviewed publications preferred Studies with clear methodology and metrics Papers with quantitative evaluation results Research from reputable journals in AI/NLP, bioinformatics, or microbiology Expected Outputs Please provide papers that will help answer: What evaluation metrics are most appropriate for scientific summary assessment? How have other researchers structured comparative AI model studies? What sample sizes and selection criteria are used in similar evaluations? What are the current benchmarks for LLM performance in scientific domains? How do human experts typically participate in AI evaluation studies? Target: 20-30 highly relevant papers that will

the need for domain-specific evaluation frameworks in microbiology and genomics.

Abstract

Large language models (LLMs) such as GPT-4 and fine-tuned open-source models show strong performance in summarizing scientific literature in biomedical and clinical domains. Several studies report that single-document summaries achieve high quality, with some metrics indicating domain accuracy above 90% (e.g., GPT-4 variants scoring up to 91.3% and quality ratings near 90/100) and ROUGE improvements by as much as 15.8 points. Yet in tasks that require multi-document synthesis or involve specialized subjects (for example, antibiotic discovery), some models yield lower factual accuracy (as low as 43.6%) and exhibit challenges with consistency and reference fidelity.

Evaluation methods for AI-generated summaries fall into three broad categories. Nine studies rely on human expert assessments that rate relevance, completeness, and factual accuracy; 11 studies combine human judgment with automated metrics such as ROUGE (all variants), METEOR, BERTScore, and F1 scores; and two studies use solely automatic evaluations. Although most available papers focus on biomedical, clinical, and broader life-science contexts, they consistently stress that domain-specific evaluation—potentially tailored to microbiology and microbial genomics—is essential to capture the nuances in summary quality.

Paper search

Using your research question "Elicit Prompt for Literature Review Research Question How accurate are large language models (LLMs) for summarizing scientific papers in microbiology and microbial genomics, and what methodologies exist for systematically evaluating AI-generated summaries of scientific literature? Literature Search Objectives Primary Focus: Find studies that evaluate the accuracy, reliability, and quality of AI-generated summaries of scientific papers, particularly in:

Microbiology and microbial genomics Biomedical sciences and life sciences (as broader context) Scientific literature processing and analysis

Secondary Focus: Identify methodological frameworks for:

Comparing multiple AI models/services Human-AI hybrid evaluation approaches Metrics and assessment criteria for summary quality Systematic evaluation designs for NLP tasks in scientific domains

Specific Search Parameters Keywords and Concepts

LLM evaluation, GPT evaluation, language model assessment Scientific paper summarization, biomedical text summarization AI accuracy in scientific literature, automated literature review Human-AI evaluation, mixed-methods assessment Microbiology informatics, genomics text mining Summary quality metrics, ROUGE scores, semantic similarity Comparative AI model studies, benchmarking NLP systems

Paper Types Sought

Empirical studies evaluating LLM performance on scientific texts Methodological papers describing evaluation frameworks for AI summarization Comparative studies testing multiple AI models or services Domain-specific evaluations in life sciences, biology, or medicine Reviews or surveys of AI applications in scientific literature processing

Time Frame Focus on papers from 2020-2025 (covering the modern LLM era), but include seminal earlier works on text summarization evaluation if highly relevant. Quality Criteria

Peer-reviewed publications preferred Studies with clear methodology and metrics Papers with quantitative evaluation results Research from reputable journals in AI/NLP, bioinformatics, or microbiology

Expected Outputs Please provide papers that will help answer:

What evaluation metrics are most appropriate for scientific summary assessment? How have other researchers structured comparative AI model studies? What sample sizes and selection criteria are used in similar evaluations? What are the current benchmarks for LLM performance in scientific domains? How do human experts typically participate in AI evaluation studies?

Target: 20-30 highly relevant papers that will form the foundation for a systematic evaluation study comparing ChatGPT, Claude, Elicit, and local models for microbiology paper summarization.", 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 papers that met these criteria:

- **AI Summarization Focus**: Does the study evaluate LLM or AI performance in summarizing scientific literature specifically in microbiology, microbial genomics, or broader life sciences content?
- **Quantitative Evaluation**: Does the study include quantitative evaluation metrics for assessing summary quality?
- **Expert Validation**: Does the study incorporate validation or evaluation by domain experts or subject matter specialists?
- **Evaluation Metrics**: Does the study either use established evaluation metrics (e.g., ROUGE, BLEU, semantic similarity) or present comparative analyses of multiple AI models?
- **Sample Size**: Does the study analyze a sample size of 50 or more scientific papers?
- **Methodology Documentation**: Does the study provide clear, reproducible methodology documentation?
- **Summarization Approach**: Does the study evaluate AI systems capable of both extractive and abstractive summarization?

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 Type**:

  Identify the primary type of study design used. Look in the methods section for explicit description of the study approach. Categorize as:

- Empirical evaluation study
- Comparative analysis
- Methodological framework development
- Systematic review/meta-analysis

If multiple design elements are present, list in order of prominence. If unclear, note "Mixed/Hybrid design". Prioritize the study's primary methodological approach for scientific literature/LLM evaluation.

- **Evaluation Methodology for LLM Summarization**:

  Extract the specific methodological approach used to evaluate LLM performance. Look in methods and results sections. Capture:

- Evaluation metrics used (e.g., ROUGE scores, METEOR, CHRF)
- Comparison approach (zero-shot, few-shot, fine-tuned)
- Evaluation types (automatic metrics, human evaluation, hybrid)
- Specific benchmarking techniques

Record exact metrics and scoring methods. If multiple evaluation approaches are used, list all in order of prominence. Include confidence intervals or statistical significance if reported.

- **LLM Models Evaluated**:

  List all specific large language models examined in the study. Include:

- Model names (e.g., ChatGPT-4, Claude 3 Opus, LongT5)
- Model type (proprietary/open-source)
- Model size or version if specified

If models are compared, note the specific comparison methodology. If models are fine-tuned, describe the fine-tuning approach and dataset used.

- **Performance Outcomes**:

  Extract quantitative performance results. Prioritize:

- Accuracy scores
- Summarization quality metrics
- Comparative performance between models
- Error rates or limitations identified

Record numerical values with their corresponding metrics (e.g., "9.89 increase in ROUGE-L"). Include confidence intervals or statistical significance if reported. Note any significant variations in performance across different domains or contexts.

- **Domain and Context of Evaluation**:

  Identify the specific scientific domain of the study:

- Primary domain (e.g., medical, microbiology)
- Subdomain or specific research area
- Type of scientific literature summarized

If multiple domains are examined, list in order of focus. Capture any domain-specific challenges or considerations mentioned in the study.

Results

Characteristics of Included Studies

Study	Study Focus (Evaluation/Framework)	Domain Area	Models Evaluated	Primary Metrics Used	Full text retrieved
Li et al., 2024	Methodological framework development; Empirical evaluation study	Biomedical (cancer immunotherapy, Large Language Models (LLMs) in medicine)	GPT-4 turbo, ChatGPT-4, ScholarAI, Gemini	Human expert evaluation (accuracy, comprehensiveness, reference integration), statistical significance	No
Zhang et al., 2024	Empirical evaluation study, Comparative analysis	Medical evidence summarization	PRIMERA, LongT5, Llama-2 (open-source, fine-tuned), GPT-3.5-turbo, GPT-4	Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-L, Metric for Evaluation of Translation with Explicit Ordering (METEOR), Character n-gram F-score (CHRF), human/GPT-4 evaluation, Population, Intervention, Comparison, Outcome (PICO) extraction	Yes

Study	Study Focus (Evaluation/Framework)	Domain Area	Models Evaluated	Primary Metrics Used	Full text retrieved
Akyon et al., 2024	Methodological framework development	Medical (observational studies, epidemiology)	GPT-3.5-Turbo, GPT-4-0613, GPT-4-1106, PaLM 2, Claude v1, Gemini Pro	Human evaluation (Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist), accuracy, statistical tests	Yes
Green et al., 2023	Methodological framework development; Empirical evaluation study	Life sciences (RNA science, non-coding RNAs)	Commercial Large Language Model (LLM) (unspecified)	Human/manual assessment, automated metrics (no mention found)	No
Evans et al., 2024	Empirical evaluation study	Scientific synthesis (multi-domain)	GPT-4 Turbo, Mistral-7B	ROUGE, Spearman's , human evaluation	Yes
Sarwal et al., 2023	Methodological framework development; Empirical evaluation study	Bioinformatics (multiple tasks)	GPT-4, Bard, LLaMA	ROUGE-1/2/L, proficiency scores, accuracy	Yes
Low et al., 2024	Empirical evaluation study	Medical (clinical questions, evidence-based medicine)	ChatGPT-4, Claude 3 Opus, Gemini Pro 1.5, OpenEvidence, ChatRWD	Human evaluation (relevance, reliability, actionability), inter-rater agreement	Yes
Shaib et al., 2023	Empirical evaluation study	Biomedicine (Randomized Controlled Trials (RCTs))	GPT-3	Human expert evaluation (factual accuracy)	No

Study	Study Focus (Evaluation/Framework)	Domain Area	Models Evaluated	Primary Metrics Used	Full text retrieved
Hake et al., 2024	Empirical evaluation study	Medical (clinical abstracts, multiple specialties)	ChatGPT-3.5	Human evaluation (quality, accuracy, bias, relevance), statistical analysis	Yes
Karotia and Susan, 2024	Methodological framework development, Empirical evaluation study	Biomedical (lay summarization)	Bidirectional and Auto-Regressive Transformers (BART), Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS), T5-small, BioBART	ROUGE-1/2/L, Bidirectional Encoder Representations from Transformers (BERT) Score, readability/factuality metrics	Yes
Asai et al., 2024	Methodological framework development, Empirical evaluation study, Comparative analysis	Scientific literature synthesis (multi-domain)	GPT4o, Llama 3.1 (8B/70B), PaperQA2, Perplexity Pro, OpenScholar	Correctness, citation accuracy, human/LLM evaluation, SCHOLAR-QABENCH	Yes
Hijazi et al., 2024	Empirical evaluation study, Comparative analysis	Biomedical (query-focused summarization)	GPT-3.5, GPT-4	Spearman correlation, quadratic kappa, ROUGE, Sci-BERTscore, BioASQ data	Yes
Liang et al., 2023	Empirical evaluation study, Comparative analysis, Methodological framework development	Biomedicine, Artificial Intelligence (AI) (peer review feedback)	GPT-4	F1 scores (extraction/matching), overlap with human feedback, user study	Yes

Study	Study Focus (Evaluation/Framework)	Domain Area	Models Evaluated	Primary Metrics Used	Full text retrieved
Tang et al., 2024	Empirical evaluation study	Medical (information extraction from abstracts)	GPT-3.5, GPT-4.0	BERTScore, ROUGE-1, GPT-4.0 evaluator, Analysis of Variance (ANOVA)	Yes
Wysocka et al., 2023	Empirical evaluation study	Biomedical (antibiotic discovery)	ChatGPT, GPT-4, Llama 2, BioGPT-large	Factual accuracy, fluency, semantic coherence	No
Chen et al., 2024	Methodological framework development	Scientific summarization (general)	No mention found	Facet-aware Metric (FM), facet-level annotations	No
Xie et al., "Biomedical Text Summarization"	Systematic review/meta-analysis	Biomedical text summarization	No mention found	No mention found (review of metrics, datasets, approaches)	No
Ying et al., 2024	Comparative analysis; Empirical evaluation study	Medical (drug labeling, drug safety)	ChatGPT	Human evaluation (similarity to expert summaries), Food and Drug Administration (FDA) labeling as benchmark	No
Mohsin et al., 2024	Comparative analysis	Medical (hematology, plasma cell disorders)	Claude 3.5, ChatGPT 4.0, Gemini, Llama-3.1	Likert scales (accuracy, completeness, hallucinations), 95% Confidence Interval (CI)	No
Blasingame et al., 2024	Empirical evaluation study, Comparative analysis	Medical (clinical evidence synthesis)	GPT-4 (aiChat)	Human evaluation (3-point Likert), reference verification	Yes

Study	Study Focus (Evaluation/Framework)	Domain Area	Models Evaluated	Primary Metrics Used	Full text retrieved
Ahad et al., 2024	Methodological framework development	Medical/meta-analysis (applied sciences)	Llama-2 (7B), Mistral-v0.1 (7B)	Bilingual Evaluation Understudy (BLEU), ROUGE, cosine similarity, human evaluation	Yes
Aliod, 2023	Empirical evaluation study	Biomedical (query-focused multi-document summarization)	GPT-3.5	ROUGE-SU4 F1, human evaluation (recall, precision, readability)	Yes
Salazar-Lara et al., 2024	Empirical evaluation study	Health literacy (biomedical texts, Plain Language Summaries (PLS))	GPT 3.5, GPT 4	BERTScore, human expert ratings, precision/recall/F1	Yes
Craig and Draghici, 2024	Methodological framework development, Empirical evaluation study, Comparative analysis	Medical (genetics, gene expression)	GPT-3, GPT-4, Llama2, PaLM2, BioGPT	No mention found (comparison with LLMs, hallucination avoidance)	No
Tang et al., 2023	Empirical evaluation study	Medical (evidence summarization, clinical domains)	GPT-3.5, ChatGPT	No mention found (automatic/human evaluation, error analysis)	Yes

Study Focus:

- Empirical evaluation studies: 20
- Methodological framework development: 10
- Comparative analyses: 8
- Systematic review/meta-analysis: 1

Domain Area:

- Medical/biomedical: 18
- Scientific synthesis/literature: 3
- Other (bioinformatics, health literacy, life sciences, AI/peer review): 1 each

Models Evaluated:

- GPT-4 (including variants): 13
- GPT-3.5 (including ChatGPT-3.5): 8
- ChatGPT (all versions): 7
- Llama/Llama-2/Llama-3.1: 6
- Gemini (all versions): 4
- Claude (all versions): 3
- Other models: 1-2 mentions each
- No mention found: 3

Primary Metrics Used:

- Human evaluation: 15
- ROUGE (all variants): 8
- BERTScore or Sci-BERTscore: 3
- F1/precision/recall: 3
- Likert scales: 2
- Statistical tests: 4
- Reference/citation accuracy/integration/verification: 3
- Readability, factuality, completeness, or actionability: 4
- Other metrics: 1 mention each
- No mention found: 3

Evaluation Frameworks and Metrics

Automated Evaluation Metrics

Study	Framework Type	Key Metrics	Validation Method	Applicability Score
Li et al., 2024	Empirical/human expert	No mention found	Human expert review, statistical significance	Rated as high applicability for biomedical domain by study authors
Zhang et al., 2024	Empirical/automatic & human	ROUGE-L, METEOR, CHRF, PICO extraction	Human, GPT-4 simulated, Confidence Interval (CI)	Rated as high applicability for medical evidence by study authors
Akyon et al., 2024	Human/hybrid	Accuracy (STROBE Qs)	Human expert, statistical tests	Rated as high applicability for medical research by study authors

Study	Framework Type	Key Metrics	Validation Method	Applicability Score
Green et al., 2023	Hybrid	No mention found	Manual & automated (not detailed)	Rated as moderate applicability for RNA science by study authors
Evans et al., 2024	Hybrid	ROUGE, Spearman's	Human/LLM correlation	Rated as moderate applicability for multi-domain by study authors
Sarwal et al., 2023	Automatic	ROUGE-1/2/L, proficiency, accuracy	Automatic, task-based	Rated as high applicability for bioinformatics by study authors
Low et al., 2024	Human	Relevance, reliability, actionability	Human expert, inter-rater	Rated as high applicability for clinical domain by study authors
Shaib et al., 2023	Human	Factual accuracy	Human expert	Rated as moderate applicability for biomedicine by study authors
Hake et al., 2024	Human	Quality, accuracy, bias, relevance	Human expert, stats	Rated as high applicability for medical domain by study authors
Karotia and Susan, 2024	Automatic	ROUGE, BERTScore, readability, factuality	Validation/test sets	Rated as high applicability for biomedical lay summarization by study authors
Asai et al., 2024	Hybrid	Correctness, citation accuracy	SCHOLARQABENCH, Prometheus v2	Rated as high applicability for multi-domain by study authors
Hijazi et al., 2024	Hybrid	Spearman, kappa, ROUGE, Sci-BERTscore	BioASQ, human	Rated as high applicability for biomedical QA by study authors
Liang et al., 2023	Hybrid	F1 (extraction/matching), overlap	Human, user study	Rated as high applicability for peer review by study authors

Study	Framework Type	Key Metrics	Validation Method	Applicability Score
Tang et al., 2024	Hybrid	BERTScore, ROUGE-1, GPT-4.0 eval	ANOVA, test set	Rated as high applicability for medical info extraction by study authors
Wysocka et al., 2023	Hybrid	Factual accuracy, fluency, coherence	Task-based, % correct	Rated as moderate applicability for antibiotic discovery by study authors
Chen et al., 2024	Methodological	Facet-aware Metric (FM)	Facet-level annotation	Rated as high applicability for scientific summarization by study authors
Xie et al., "Biomedical Text Summarization"	Review	No mention found	No mention found	Not Applicable
Ying et al., 2024	Human	Similarity to expert summaries	FDA labeling	Rated as high applicability for drug safety by study authors
Mohsin et al., 2024	Human	Likert (accuracy, completeness, hallucinations)	Blinded review, CI	Rated as high applicability for hematology by study authors
Blasingame et al., 2024	Human	3-point Likert, reference verification	Human, stats	Rated as high applicability for clinical evidence by study authors
Ahad et al., 2024	Hybrid	BLEU, ROUGE, cosine similarity	Human, MAD dataset	Rated as high applicability for meta-analysis by study authors
Aliod, 2023	Hybrid	ROUGE-SU4 F1, recall, precision, readability	BioASQ, human	Rated as high applicability for biomedical QA by study authors
Salazar-Lara et al., 2024	Hybrid	BERTScore, precision, recall, F1	Human expert, stats	Rated as high applicability for health literacy by study authors
Craig and Draghici, 2024	No mention found	No mention found	No mention found	Rated as moderate applicability for genetics by study authors

Study	Framework Type	Key Metrics	Validation Method	Applicability Score
Tang et al., 2023	Hybrid	No mention found	Human/automatic, error analysis	Rated as moderate applicability for medical evidence by study authors

Framework Type:

- Human-based: 9 studies
- Hybrid: 11 studies
- Automatic: 2 studies
- Methodological: 1 study
- Review: 1 study
- No mention found: 1 study

Key Metrics:

- ROUGE (any variant): 8 studies
- BERTScore: 3 studies
- F1: 3 studies
- Accuracy: 4 studies
- Likert-type ratings: 2 studies
- Other metrics: 1-2 mentions each
- No mention found: 5 studies

Validation Method:

- Human expert review: 17 studies
- Automatic validation or use of validation/test sets: 11 studies
- LLM-based validation: 4 studies
- Task-based or user study validation: 3 studies
- No mention found: 2 studies

Applicability Score:

- High: 18 studies
- Moderate: 6 studies
- Not Applicable: 1 study

Human-AI Hybrid Evaluation Methods

Our review found that several studies employed hybrid evaluation frameworks. For example, Zhang et al. (2024) and Asai et al. (2024) combined human expert review with LLM-based simulated evaluations, while Hijazi et al. (2024) and Liang et al. (2023) compared LLM-generated scores directly with human judgments. This approach was particularly prevalent in domains where factual accuracy and domain expertise are critical, such as biomedical research and clinical evidence synthesis.

Domain-Specific Assessment Criteria

Study	Framework Type	Key Metrics	Validation Method	Applicability Score
Li et al., 2024	Biomedical, literature recommendation	Relevance, quality, accuracy, comprehensiveness, reference integration	Human expert, stats	Rated as high applicability by study authors
Green et al., 2023	RNA science	Human/manual assessment, automated metrics	RNAcentral resource	Rated as moderate applicability by study authors
Sarwal et al., 2023	Bioinformatics	ROUGE, proficiency, accuracy	Task-based, domain-specific	Rated as high applicability by study authors
Wysocka et al., 2023	Antibiotic discovery	Factual accuracy, fluency, bias	Task-based	Rated as moderate applicability by study authors
Mohsin et al., 2024	Hematology (Plasma Cell Disorders (PCDs))	Likert (accuracy, completeness, hallucinations)	Blinded review	Rated as high applicability by study authors
Salazar-Lara et al., 2024	Health literacy	BERTScore, expert ratings	Plain Language Summaries (PLS), clinical trial protocols	Rated as high applicability by study authors
Craig and Draghici, 2024	Genetics	No mention found	Differential Expression Gene/Pathway Enrichment Analysis (DEG/PEA) analysis	Rated as moderate applicability by study authors

Framework Type by Domain:

- Biomedical/literature recommendation
- RNA science
- Bioinformatics
- Antibiotic discovery
- Hematology
- Health literacy
- Genetics

Key Metrics:

- Human or expert-based metrics: 5 studies
- Automated or statistical metrics: 4 studies

- Both human/expert and automated metrics (hybrid): 3 studies
- No mention found: 1 study

Validation Method:

- Human expert or blinded review: 2 studies
- Resource-based validation: 3 studies
- Task-based validation: 2 studies
- Domain-specific validation: 1 study

Applicability Score:

- High: 4 studies
- Moderate: 3 studies

Comparative Performance Analysis

Accuracy Metrics Across Models

Study	Model Type	Summary Accuracy	Domain Accuracy	Reliability Score
Li et al., 2024	GPT-4 turbo, ChatGPT-4, ScholarAI, Gemini	RefAI outperformed baselines (stat. sig.)	Biomedical	No mention found
Zhang et al., 2024	PRIMERA, LongT5, Llama-2, GPT-3.5-turbo, GPT-4	Fine-tuned LLMs: +9.89 ROUGE-L, +13.21 METEOR, +15.82 CHRF	Medical evidence	95% CI provided
Akyon et al., 2024	GPT-3.5-Turbo, GPT-4-1106, PaLM 2, Claude v1, Gemini Pro, GPT-4-0613	66.9% (GPT-3.5-Turbo), 65.6% (GPT-4-1106)	Medical research	Stat. sig. differences ($P<.001$)
Green et al., 2023	Commercial LLM	Majority rated "extremely high quality"	RNA science	No mention found
Evans et al., 2024	GPT-4 Turbo, Mistral-7B	Weak correlation with human ratings	Multi-domain	Spearman's : 0.710 (human), 0.786 (LLMs)
Sarwal et al., 2023	GPT-4, Bard, LLaMA	GPT-4: 91.3% (domain), 65.32% (ML dev); Bard: 97.5% (math)	Bioinformatics	None >40% ROUGE (summarization)

Study	Model Type	Summary Accuracy	Domain Accuracy	Reliability Score
Low et al., 2024	ChatGPT-4, Claude 3 Opus, Gemini Pro 1.5, OpenEvidence, ChatRWD	ChatRWD: 58% relevant/evidence-based; OpenEvidence: 24%; ChatGPT: 2-10%	Clinical	Actionable: ChatRWD 44%, OpenEvidence 30%
Shaib et al., 2023	GPT-3	Faithful single-article summaries; struggles with multi-doc	Biomedicine	No mention found
Hake et al., 2024	ChatGPT-3.5	Quality: 90/100; Accuracy: 92.5/100; Bias: 0/100	Medical	Serious inaccuracies rare
Karotia and Susan, 2024	BART, PEGASUS, T5-small, BioBART	ROUGE-1: 0.4681 (eLife val), 0.4525 (PLOS val)	Biomedical lay	High readability, factuality lower
Asai et al., 2024	GPT4o, Llama 3.1, PaperQA2, OpenScholar	OpenScholar-8B: +5% over GPT-4o; +7% over PaperQA2	Multi-domain	Citation F1: ~40 (OpenScholar)
Hijazi et al., 2024	GPT-3.5, GPT-4	Spearman: 0.511 (fine-tuned), 0.501 (meta-eval)	Biomedical QA	Pairwise accuracy: 0.61
Liang et al., 2023	GPT-4	Overlap: 30.85% (Nature), 39.23% (ICLR)	Biomed/AI	F1: 0.968 (extraction), 0.824 (matching)
Tang et al., 2024	GPT-3.5, GPT-4.0	GPT-4.0: 0.688-0.964 accuracy; Evaluator: 0.9714	Medical info extraction	Sensitivity: 0.8550 (GPT-4.0)
Wysocka et al., 2023	ChatGPT, GPT-4, Llama 2, BioGPT-large	GPT-4: 70% factual (compounds), 43.6% (relations)	Antibiotic discovery	BioGPT-large: 30%/30%
Chen et al., 2024	No mention found	No mention found	Scientific summarization	No mention found
Xie et al., "Biomedical Text Summarization"	No mention found	No mention found	Biomedical	No mention found
Ying et al., 2024	ChatGPT	"Closely resembled" human summaries	Drug safety	No mention found

Study	Model Type	Summary Accuracy	Domain Accuracy	Reliability Score
Mohsin et al., 2024	Claude 3.5, ChatGPT 4.0, Gemini, Llama-3.1	Claude: 3.92/4.00 (accuracy/completeness); Llama: 1.92/1.67	Hematology	None perfect
Blasingame et al., 2024	GPT-4 (aiChat)	83.3% correct, 16.2% partial, 0.5% incorrect	Clinical evidence	37% references verified
Ahad et al., 2024	Llama-2, Mistral-v0.1	87.6% relevant (fine-tuned); irrelevance: 1.9%	Meta-analysis	BLEU, ROUGE used
Aliod, 2023	GPT-3.5	Top ROUGE-F1 in BioASQ; human eval >4/5	Biomedical QA	Retrieval-augmented best
Salazar-Lara et al., 2024	GPT 3.5, GPT 4	GPT 4: 99.6% classification; BERTScore higher	Health literacy	Expert rating: 4.71 (GPT 4)
Craig and Draghici, 2024	GPT-3, GPT-4, Llama2, PaLM2, BioGPT	No mention found; "free of hallucinations"	Genetics	No mention found
Tang et al., 2023	GPT-3.5, ChatGPT	No mention found; error analysis	Medical evidence	No mention found

Model Types (number of studies):

- GPT-4 (all variants): 13
- GPT-3.5 (all variants): 6
- Llama/Llama-2/Llama-3.1: 6
- ChatGPT (all variants): 3
- Claude (all variants): 3
- Gemini (all variants): 4
- Other models: 1-2 each
- No mention found: 2

Summary Accuracy:

- 23 studies reported quantitative accuracy metrics
- 6 studies included qualitative or human evaluation of summary quality
- 2 studiesreported explicit comparisons to baseline models
- 2 studies reported correlation with human ratings
- No mention found: 3 studies

Domain Accuracy:

- Biomedical domains: 4 studies
- Medical domains: 4 studies

- Clinical or clinical evidence domains: 2 studies
- Medical evidence: 2 studies
- Biomedical QA: 2 studies
- Multi-domain: 2 studies
- Other domains: 1 study each

Reliability Score:

- 2 studies reported statistical significance for accuracy differences
- 1 study reported confidence intervals
- 2 studies reported correlation coefficients
- 10 studies reported quantitative reliability metrics
- 6 studies included qualitative reliability statements
- No mention found: 8 studies

Domain-Specific Performance

Our review found that performance varied by domain and task complexity. For example, Wysocka et al. (2023) reported that in antibiotic discovery, GPT-4 achieved 70% factual accuracy for compounds but only 43.6% for relations. In highly specialized areas (e.g., antibiotic discovery, plasma cell disorders, gene expression analysis), even the best LLMs fell short of expert-level accuracy, particularly in multi-document synthesis or meta-analysis. In more general biomedical summarization tasks, LLMs performed well in single-document settings but struggled with aggregation and factual consistency across multiple sources.

Error Analysis and Limitations

Our review identified several common limitations across studies:

- Persistent hallucinations and fabricated references, especially in general-purpose LLMs
- Lower performance in multi-document summarization and meta-analysis compared to single-document tasks
- Variable correlation between automatic metrics and human expert judgment
- Limited generalizability of findings to microbiology and microbial genomics, as most studies focused on broader biomedical or life science domains
- Rapid evolution of LLMs and evaluation frameworks, leading to potential obsolescence of findings

Summary

Our review of the included studies found that:

- Large Language Models (LLMs), particularly GPT-4 and fine-tuned open-source models, were reported as capable of generating high-quality summaries of scientific literature in biomedical and related domains, according to the metrics used in those studies.
- Accuracy, reliability, and factual consistency remained imperfect, especially in complex or highly specialized tasks.
- Evaluation frameworks were increasingly sophisticated, combining automatic, human, and hybrid methods, with a trend toward domain-specific validation and the development of new metrics.
- Significant challenges persisted, including hallucinations, reference fabrication, and the need for more robust evaluation in microbiology and genomics contexts.

- Human-AI hybrid evaluation methods were prevalent, especially in domains requiring high factual accuracy and domain expertise.
- Performance varied significantly across domains and task complexity, with LLMs generally performing better on single-document tasks compared to multi-document synthesis or meta-analysis.

These findings suggest that while LLMs show promise for scientific literature summarization, their application in specialized fields like microbiology and microbial genomics requires careful evaluation and domain-specific adaptations.

References

Akanksha Karotia, and Seba Susan. "BioLay_AK_SS at BioLaySumm: Domain Adaptation by Two-Stage Fine-Tuning of Large Language Models Used for Biomedical Lay Summary Generation." *Workshop on Biomedical Natural Language Processing*, 2024.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, et al. "OpenScholar: Synthesizing Scientific Literature with Retrieval-Augmented LMs." *arXiv.org*, 2024.

Aleenah Mohsin, Samuel M. Rubinstein, Rahul Banerjee, Sanjay Mishra, Mary L. Kwok, Peter Yang, Jeremy Lyle Warner, and Andrew J Cowan. "Evaluating the Accuracy of Artificial Intelligence(AI)-Generated Synopses for Plasma Cell Disorder Treatment Regimens." *Blood*, 2024.

Andrew Green, C. Ribas, Nancy Ontiveros-Palacios, Anton I. Petrov, A. Bateman, and Blake Sweeney. "LitSumm: Large Language Models for Literature Summarisation of Non-Coding RNAs." *arXiv.org*, 2023.

Carolina Salazar-Lara, Andrés Felipe, Arias Russi, and Rubén Manrique. "Bridging the Gap in Health Literacy: Harnessing the Power of Large Language Models to Generate Plain Language Summaries from Biomedical Texts." *medRxiv*, 2024.

Chantal Shaib, Millicent Li, Sebastian Antony Joseph, I. Marshall, Junyi Jessy Li, and Byron Wallace. "Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success)." *Annual Meeting of the Association for Computational Linguistics*, 2023.

Diego Mollá Aliod. "Large Language Models and Prompt Engineering for Biomedical Query Focused Multi-Document Summarisation." *arXiv.org*, 2023.

Douglas B. Craig, and Sorin Draghici. "What's the Data Say? An LLM-Based System for Interrogating Experimental Data." *IEEE International Conference on Bioinformatics and Biomedicine*, 2024.

Gongbo Zhang, Qiao Jin, Yiliang Zhou, Song Wang, B. Idnay, Yiming Luo, Elizabeth Park, et al. "Closing the Gap Between Open-Source and Commercial Large Language Models for Medical Evidence Summarization." *arXiv.org*, 2024.

Hashem Hijazi, Diego Molla, Vincent Nguyen, and Sarvnaz Karimi. "Using Large Language Models to Evaluate Biomedical Query-Focused Summarisation." *Workshop on Biomedical Natural Language Processing*, 2024.

Jawad Ibn Ahad, Rafeed Mohammad Sultan, Abraham Kaikobad, Fuad Rahman, Mohammad Ruhul Amin, Nabeel Mohammed, and Shafin Rahman. "Empowering Meta-Analysis: Leveraging Large Language Models for Scientific Synthesis." *BigData Congress [Services Society]*, 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.

Julia Evans, Jennifer D'Souza, and Sören Auer. "Large Language Models as Evaluators for Scientific Synthesis." *Conference on Natural Language Processing*, 2024.

Lan Ying, Zhichao Liu, Hong Fang, Rebecca Kusko, Leihong Wu, Stephen Harris, and Weida Tong. "Text Summarization with ChatGPT for Drug Labeling Documents." *Drug Discovery Today*, 2024.

Liyan Tang, Z. Sun, B. Idnay, J. Nestor, A. Soroush, P. A. Elias, Z. Xu, et al. "Evaluating Large Language Models on Medical Evidence Summarization." *medRxiv*, 2023.

M. Wysocka, Oskar Wysocki, M. Delmas, V. Mutel, and Andre Freitas. "Large Language Models, Scientific Knowledge and Factuality: A Systematic Analysis in Antibiotic Discovery." *arXiv.org*, 2023.

Mallory N. Blasingame, Taneya Y. Koonce, Annette M. Williams, Dario A Giuse, Jing Su, Poppy A Krump, and N. Giuse. "Evaluating a Large Language Model's Ability to Answer Clinicians' Requests for Evidence Summaries." *medRxiv*, 2024.

Qianqian Xie, Zheheng Luo, Benyou Wang, and S. Ananiadou. "A Survey for Biomedical Text Summarization: From Pre-Trained to Large Language Models," 2023.

Şeyma Handan Akyon, F. C. Akyon, Ahmet Sefa Camyar, Fatih Hızlı, Talha Sari, and Ş. Hızlı. "Evaluating the Capabilities of Generative AI Tools in Understanding Medical Papers: Qualitative Study." *JMIR Medical Informatics*, 2024.

V. Sarwal, Viorel Munteanu, Timur Suhodolschi, Dumitru Ciorba, E. Eskin, Wei Wang, and S. Mangul. "BioLLMBench: A Comprehensive Benchmarking of Large Language Models in Bioinformatics." *bioRxiv*, 2023.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, et al. "Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis." *NEJM AI*, 2023.

Xiuying Chen, Tairan Wang, Qingqing Zhu, Taicheng Guo, Shen Gao, Zhiyong Lu, Xin Gao, and Xiangliang Zhang. "Rethinking Scientific Summarization Evaluation: Grounding Explainable Metrics on Facet-Aware Benchmark." *arXiv.org*, 2024.

Y. Low, Michael L. Jackson, Rebecca J. Hyde, Robert E. Brown, Neil M. Sanghavi, Julian D. Baldwin, C. W. Pike, et al. "Answering Real-World Clinical Questions Using Large Language Model Based Systems." *arXiv.org*, 2024.

Y. Tang, Z. Xiao, X. Li, Q. Zhang, E. W. Y. Chan, I. C. K. Wong, and Research Data Collaboration Task Force. "Large Language Model in Medical Information Extraction from Titles and Abstracts with Prompt Engineering Strategies: A Comparative Study of GPT-3.5 and GPT-4." *medRxiv*, 2024.

Yiming Li, Jeff Zhao, Manqi Li, Yifang Dang, Evan Yu, Jianfu Li, Zenan Sun, et al. "RefAI: A GPT-Powered Retrieval-Augmented Generative Tool for Biomedical Literature Recommendation and Summarization." *J. Am. Medical Informatics Assoc.*, 2024.