Automatic Generation of Discharge Summaries Using Large Language Models: A Systematic Literature Review

Lucas Molino Piñar, Manuel Carlos Díaz Galiano, María Teresa Martín Valdivia
SINAI Research Group, Universidad de Jaén, Spain
{lmolino, mcdiaz, maite}@ujaen.es

Abstract

Discharge summaries are critical documents for continuity of care, yet their manual creation imposes significant burdens on clinical staff. This systematic literature review examines current approaches to automatic generation of discharge summaries using Natural Language Processing (NLP) and Large Language Models (LLMs). Following the Kitchenham guidelines for systematic reviews in software engineering, we searched Scopus and PubMed databases for studies published between 2023 and 2026, identifying 9 primary studies from an initial pool of 102 papers. Our analysis reveals that GPT-4 and its variants dominate current research (appearing in 6 of 9 studies), while open-source alternatives like LLaMA show promise for privacy-preserving deployments. Evaluation primarily relies on automatic metrics (ROUGE, BLEU) combined with human expert assessment. Key challenges include hallucination rates ranging from 33% to 64%, information omission, integration with Electronic Health Record (EHR) systems, and context window limitations. Studies addressing factuality employ human-in-the-loop validation, prompt engineering techniques, and knowledge graph-based correction mechanisms. Despite these challenges, recent implementations demonstrate clinical feasibility, with one study achieving a 94.35% System Usability Score. This review provides a comprehensive synthesis of the state-of-the-art and identifies opportunities for future research in this rapidly evolving field.

Keywords: Large Language Models, Discharge Summaries, Natural Language Processing, Clinical Documentation, Systematic Review

1. Introduction

Discharge summaries are essential clinical documents that communicate critical patient information between healthcare providers, ensuring continuity of care after hospitalization (Kripalani et al., 2007). These documents typically include diagnoses, treatments administered, hospital course, and discharge recommendations. However, their manual creation represents a significant burden for clinicians, contributing to documentation fatigue and reducing time available for direct patient care (Sinsky et al., 2016).

The emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023), and their clinical adaptations has created unprecedented opportunities for automating clinical documentation tasks. These transformer-based architectures have demonstrated remarkable capabilities in text generation, summarization, and domain adaptation, making them promising candidates for addressing the discharge summary automation challenge.

1.1. Motivation and Scope

The motivation for this systematic review stems from three key observations. First, the clinical documentation burden continues to escalate, with physicians spending up to 50% of their time on documentation tasks. Second, recent advances in LLM technology (2023-2026) have dramatically improved the feasibility of automatic text generation in specialized domains. Third, there is a critical need to understand the current state of research, including both achievements and limitations, to guide future development efforts in clinical NLP.

This review focuses specifically on the automatic generation of discharge summaries intended for healthcare professionals (physician-to-physician communication). We explicitly exclude studies focusing on patient-friendly simplification of existing summaries, as this represents a distinct NLP task with different objectives. Our temporal scope (2023-2026) captures the most recent developments in LLM-based clinical text generation, particularly following the release of GPT-4 and similar advanced models.

1.2. Objectives and Research Questions

The primary objective of this systematic literature review is to identify, analyze, and synthesize current approaches for automatic generation of discharge summaries using NLP and LLMs. Specifically, we aim to catalog the NLP models and techniques employed for discharge summary generation, identify evaluation methodologies and metrics used to assess generated summary quality, document technical and software engineering challenges encountered during implementation, and examine how issues of factual accuracy, hallucinations, and patient

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 73–82

data privacy are addressed.

To guide this systematic review, we formulated four research questions addressing complementary aspects of automatic discharge summary generation. **RQ1** asks what language models and NLP techniques are currently used for automatic generation of discharge summaries. **RQ2** investigates what evaluation metrics are employed to validate the clinical and linguistic quality of generated reports. **RQ3** explores the main technical and software engineering challenges reported in implementing these systems. Finally, **RQ4** examines how issues of factuality, hallucinations, and medical data privacy are addressed in these systems. These questions were designed to cover the complete lifecycle of discharge summary generation systems, from model selection (RQ1), through quality assessment (RQ2), to practical implementation considerations (RQ3 and RQ4).

The remainder of this paper is organized as follows: Section 2 describes the systematic review protocol, including the search strategy and selection criteria. Section 3 presents the results organized by research question, analyzing models, evaluation metrics, technical challenges, and factuality issues. Section 4 discusses our findings and their implications for practice. Finally, Section 5 concludes with limitations and future work directions.

2. Review Protocol

This systematic review follows the guidelines proposed by Kitchenham and Charters (2007) for conducting systematic literature reviews in software engineering. The review was conducted using the Parsifal tool[1] for protocol management and study tracking. Data extraction sheets and quality assessment scores are available from the authors upon request.

2.1. PICOC Framework

We defined the scope of our review using the PICOC framework (Petticrew and Roberts, 2006), an acronym for Population, Intervention, Comparison, Outcome, and Context. The **Population** comprises software systems for automatic generation of clinical documentation, specifically discharge summaries. The **Intervention** encompasses NLP techniques, large language models (LLMs such as GPT, BERT, LLaMA), natural language generation methods, sequence-to-sequence architectures, and transformer-based approaches. For **Comparison**, we considered traditional template-based methods, manual generation by clinicians, rule-based systems, different model architectures (transformer vs. RNN), and extractive vs. abstractive approaches. The expected **Outcome** includes quality of generated reports (linguistic and clinical), evaluation metrics (ROUGE, BLEU, BERTScore, human assessment), technical feasibility, implementation viability, and clinical acceptance. The **Context** covers hospital environments, electronic health record (EHR) systems, clinical documentation workflows, medical software engineering, and healthcare informatics.

2.2. Search Strategy

We selected two complementary databases for our search: **Scopus** as the primary source for its comprehensive coverage of computer science and medical informatics literature, and **PubMed/MEDLINE** as a secondary source for its specialized coverage of biomedical literature. We initially considered IEEE Xplore (3 results) and ACM Digital Library (14 results, only 1-2 within temporal scope), but excluded them due to insufficient coverage of our topic. ACL Anthology and related computational linguistics venues (EMNLP, BioNLP, LOUHI workshops) were not included as separate sources; however, Scopus indexes proceedings from these venues, ensuring their coverage in our search.

Table 1 presents the keywords derived from the research questions and their synonyms used to construct the search string.

Table 1: Keywords and synonyms used in search strategy.

Keyword	Synonyms
discharge summary	discharge report, hospital summary, clinical summary, discharge letter, discharge note
automatic generation	automated generation, AI generation, machine-generated, computer-generated
large language model	LLM, GPT, BERT, language model, neural language model, transformer
natural language processing	NLP, text processing, computational linguistics, automatic summarization
deep learning	machine learning, artificial intelligence, neural networks

The search string was constructed using Boolean operators following the structure: `(keyword1 OR synonym1) AND (keyword2 OR synonym2)`. The following search string was applied to Scopus on January 6, 2026:

```
TITLE-ABS-KEY(("discharge summary" OR "discharge report" OR
```

[1] https://parsif.al

```
"hospital summary" OR "clini-
cal summary") AND ("automatic
generation" OR "automated gen-
eration" OR "AI generation" OR
"LLM" OR "large language model"
OR "natural language generation"
OR "text generation") AND ("NLP"
OR "natural language processing"
OR "deep learning" OR "machine
learning" OR "artificial intel-
ligence")) AND PUBYEAR > 2022
AND (LIMIT-TO(DOCTYPE, "ar") OR
LIMIT-TO(DOCTYPE, "cp"))
```

This search returned 94 papers. The equivalent PubMed query, executed on January 7, 2026, was:

```
("discharge sum-
mary"[Title/Abstract] OR "dis-
charge report"[Title/Abstract]
OR "clinical sum-
mary"[Title/Abstract])
AND ("large language
model"[Title/Abstract] OR
"LLM"[Title/Abstract] OR
"natural language genera-
tion"[Title/Abstract] OR
"GPT"[Title/Abstract]) AND
("2023"[Date - Publication] :
"2026"[Date - Publication])
```

This search returned 8 papers. The total initial pool was 102 papers.

2.3. Selection Criteria

Studies were included if they met all of the following criteria: published between 2023 and 2026, written in English, focused on discharge summaries for healthcare professionals, presented specific methods for automatic discharge summary generation, included empirical evaluation with real or synthetic clinical data, appeared in a peer-reviewed venue or high-quality preprint server with subsequent peer-reviewed publication, constituted a full paper (minimum 6 pages, not extended abstract), and addressed NLP, AI, or software engineering aspects of the problem.

Conversely, studies were excluded if they were secondary studies (systematic reviews, literature reviews, surveys), tertiary studies (reviews of reviews), abstract-only or limited access without full text, only mentioned discharge summaries tangentially, had unclear or absent methodology, represented grey literature without peer review, were duplicate publications, addressed the wrong domain (not discharge summaries), or focused on patient-friendly simplification only.

2.4. Quality Assessment

Each candidate study was evaluated using six quality assessment criteria, scored on a 3-point scale of 0 (no), 0.5 (partially), or 1 (yes). The criteria assessed whether results are valid and conclusions supported by evidence (QA1), whether research objectives and questions are clearly defined (QA2), whether methodology is described in sufficient detail to allow replication (QA3), whether experiments or evaluations are reported using real or synthetic data (QA4), whether study limitations and threats to validity are discussed (QA5), and whether the contribution is relevant to software engineering or clinical NLP (QA6).

Studies scoring ≥ 3.5 out of 6 were included in the final analysis. Additional rigor criteria applied: sample size $n < 10$ was considered insufficient; non-blinded evaluation without justification was penalized; purely fictitious data without real-world validation resulted in rejection.

2.5. Data Extraction

For each included study, we extracted 29 data fields organized into five categories. **Bibliographic** data included venue and publication type. **Study Context** (addressing RQ1) covered research goal, domain, dataset, models, architecture, training approach, and input/output formats. **Evaluation** data (addressing RQ2) encompassed metrics, metric values, human evaluation details, baseline comparisons, and best results. **Software Engineering** aspects (addressing RQ3) included technical challenges, system architecture, EHR integration, computational requirements, and software tools. Finally, **Factuality and Ethics** data (addressing RQ4) covered hallucination handling, privacy/security measures, clinical validation, and limitations.

2.6. Study Selection Process

Table 2 illustrates our selection process following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021).

Table 2: PRISMA flow: study selection.

Phase	N	Excluded
Initial search	102	–
Duplicates removed	101	1
Relevance screening	20	81 (off-topic)
Title/abstract	14	6 (no focus)
Full-text + QA	**9**	5 (see text)

From the initial 102 papers, we removed 1 duplicate. During relevance screening, 81 papers were excluded for being off-topic (e.g., general clinical NLP without discharge focus), addressing wrong domains (radiology reports, pathology), or focusing on patient-friendly simplification. Title/abstract

screening excluded 6 additional papers: 3 lacked specific generation methods, 2 were surveys or position papers, and 1 was an extended abstract. During quality assessment, five papers were rejected: two used completely fictitious data without real-world validation, one had sample size $n = 10$ deemed insufficient, one focused exclusively on patient simplification, and one had methodological issues (single non-blinded evaluator). One included study (Williams et al., 2025) was initially available as a medRxiv preprint but has since been accepted for peer-reviewed publication in PLOS Digital Health (June 2025); we retained it given its methodological rigor and subsequent peer review.

3. Results

Table 3 summarizes the 9 primary studies included in this review. It is important to note that while eight studies utilized real clinical records, Grazhdanski et al. (2025) employed entirely synthetic data generated from MSD Manuals. This distinction should be carefully considered when interpreting cross-study comparisons, particularly regarding faithfulness and hallucination metrics.

3.1. RQ1: Models, Architectures, and Training Approaches

Our analysis reveals a clear hierarchy of model preferences in discharge summary generation. GPT-4 and its variants dominate current research, appearing in 6 of 9 studies (67%). Table 4 presents detailed performance comparisons across different models.

Williams et al. (2025) provided the most direct comparison between GPT-4 and GPT-3.5-turbo on identical emergency department data. GPT-4 achieved significantly better results across all error categories: 33% completely error-free summaries versus only 10% for GPT-3.5-turbo, with hallucination rates of 42% versus 64% respectively. GPT-4 also produced more concise outputs (median 235 words vs. 369.5 words) with better readability scores (Flesch-Kincaid Grade Level 10.0 vs. 10.7).

LLaMA 3 appears in three studies, offering a privacy-preserving alternative to commercial APIs. Hains et al. (2025) demonstrated that locally-deployed LLaMA 3 models achieved 61.6-61.9% of human performance using a validated 17-domain scoring tool (Savvopoulos et al.). Notably, the 70B parameter model showed no statistically significant improvement over the 8B model ($p = 0.974$), suggesting diminishing returns at higher parameter counts for this task. Li et al. (2025) compared base and fine-tuned LLaMA 3 variants, finding that fine-tuning with 102 clinical cases produced summaries with length closest to the gold standard, but did not consistently outperform zero-shot GPT-4 on quality metrics. This suggests that prompt engineering may be more cost-effective than fine-tuning for current applications.

It is important to note that this field evolves rapidly; as newer and more capable models emerge (such as GPT-5 or future LLaMA iterations), the model preferences documented here will likely shift accordingly. The general architectural patterns and evaluation methodologies, however, are expected to remain relevant.

Three distinct architectural paradigms emerge from our analysis. The most common approach (5 studies) relies on single-model API-based generation using commercial APIs (GPT-4, GPT-4o), which offers rapid prototyping but raises privacy concerns and limits reproducibility. A second paradigm involves multi-agent orchestration: Croxford et al. (2025) evaluated Microsoft MagenticOne, a multi-agent framework with specialized agents for planning, web search, file handling, and coding, while Trejo Omeñaca et al. (2025) implemented custom multi-agent orchestration with iterative prompt engineering for real-world deployment. The third paradigm, explored by Fu et al. (2025), proposes a hybrid structured-unstructured pipeline combining extractive summarization (TextRank with ClinicalBERT embeddings) for unstructured notes with data-to-text generation for structured EHR data (medications, labs).

Regarding training approaches, the overwhelming majority of studies (7/9) employed zero-shot prompting without model fine-tuning. Schwieger et al. (2024) conducted extensive prompt engineering over 6 weeks with hundreds of iterations, highlighting that prompt optimization is the current preferred approach over fine-tuning. Only Li et al. (2025) performed supervised fine-tuning on clinical data, and Grazhdanski et al. (2025) employed a self-refinement approach based on Madaan et al.'s method. The preference for zero-shot approaches reflects both the strong out-of-box capabilities of modern LLMs and the challenges of acquiring sufficient clinical training data.

3.2. RQ2: Evaluation Paradigms and Metrics

Our analysis reveals a significant paradigm shift in evaluation methodology. Early approaches relied solely on automatic metrics (ROUGE, BLEU), but contemporary studies recognize that clinical quality cannot be captured by lexical overlap metrics alone. Eight of nine studies (89%) now combine automatic metrics with human expert evaluation, reflecting a maturing understanding of the multidimensional nature of clinical text quality.

Table 5 presents the specific metric values re-

Table 3: Summary of included primary studies.

Study	Venue	Domain	Models	N	QA
Williams et al. (Williams et al., 2025)	PLOS Digit. Health	Emergency	GPT-4, GPT-3.5	100	5.0
Croxford et al. (Croxford et al., 2025)	npj Digit. Med.	General	GPT-4o, LLaMA	200	6.0
Schwieger et al. (Schwieger et al., 2024)	Int. J. Med. Inf.	Psychiatry	GPT-4	20	6.0
Challener et al. (Challener et al., 2025)	J. Hosp. Med.	Multi-spec.	GPT-4o (Epic)	100	5.0
Trejo et al. (Trejo Omeñaca et al., 2025)	Computers	Multi-spec.	Multiple LLMs	47	4.5
Hains et al. (Hains et al., 2025)	Intern. Med. J.	General	LLaMA 3	10	4.0
Li et al. (Li et al., 2025)	J. Biomed. Inf.	Oncology	GPT-4, LLaMA 3	50	5.0
Grazhdanski et al.	J. Biomed. Inf.	Synth. Gen.	GPT-based, KG	900	4.5
Fu et al. (Fu et al., 2025)	IEEE BigData 2024	General	BART, PEGASUS	MIMIC	4.0

ported across studies that employed automatic evaluation metrics. Not all studies reported comparable automatic metrics; some relied primarily on human evaluation or used domain-specific instruments.

ROUGE metrics appear in 7 of 9 studies, but reported values are notably low (ROUGE-L: 0.16-0.24). This reflects the abstractive nature of discharge summary generation, good summaries need not share exact phrasing with source documents. Fu et al. (2025) demonstrated that supplementing unstructured summaries with structured data improved ROUGE-L by 5.6 percentage points (from 17.6% to 23.2% for the Lead-3 baseline). Semantic similarity metrics appear as alternatives to lexical overlap: Li et al. (2025) reported semantic similarity scores of 0.83-0.837, with LLaMA 3 achieving the highest similarity to reference summaries despite lower ROUGE scores.

Human evaluation approaches vary significantly in rigor and standardization. Table 6 summarizes the methodologies across studies.

The PDSQI-9 instrument (Provider Documentation Summarization Quality Instrument) emerges as the most rigorously validated tool. Croxford et al. (2025) demonstrated that it achieves excellent psychometric properties (ICC 0.867) and captures nine clinically relevant dimensions: Cited, Accurate, Thorough, Useful, Organized, Comprehensible, Succinct, Synthesized, and Stigmatizing. Schwieger et al. (2024) developed a comprehensive 15-item rubric specifically for psychiatric summaries, evaluating dimensions including text coherence ($\Delta 0.96$ human advantage), summarization quality ($\Delta 1.07$), and specificity ($\Delta 1.19$). Importantly, AI showed advantages in conciseness (low unnecessary information $\Delta 0.07$) and safety/legal aspects ($\Delta 0.01$).

A notable methodological development is the LLM-as-Judge paradigm, which has gained traction in NLP evaluation since 2023 (Zheng et al., 2023). In the context of clinical summarization, Croxford et al. (2025) systematically evaluated this approach for discharge summary assessment. This paradigm uses LLMs to automatically evaluate clinical summaries, potentially scaling human-quality assessment. Key findings from their systematic evaluation reveal that reasoning models excel, with GPT-o3-mini achieving ICC 0.818 (95% CI: 0.772-0.854), the highest correlation with human experts. DeepSeek R1 achieved ICC 0.762, demonstrating that open-source reasoning models can approach commercial performance. Non-reasoning models underperform: GPT-4o achieved only ICC 0.730, and the multi-agent framework MagenticOne achieved ICC 0.768. The speed advantage is substantial, with LLM evaluation completing in 22 seconds versus 10 minutes for human experts, a 27$\times$ speedup. Reasoning models particularly excelled in attributes requiring deeper analysis (Cited, Organized, Synthesized, Thorough), while simpler attributes showed less differentiation. This paradigm offers a promising path toward scalable, reproducible evaluation of clinical text generation systems.

Two studies assessed readability using established metrics. Williams et al. (2025) found that GPT-4 achieved Flesch-Kincaid Grade Level 10.0 (IQR 9.5-11.1) and Flesch Reading Ease 48.6 (IQR 41.0-52.0), indicating 10th-grade reading level. Challener et al. (2025) reported that human summaries showed higher Flesch Reading Ease (33.11 vs 26.2, $p < 0.05$), indicating simpler language, though LLM summaries scored higher on clinical utility.

Table 4: Detailed model performance comparison across studies.

Model	Type	Study	Score	Notes
GPT-4	Comm. API	Williams et al.	33% err-free	Halluc. 42% vs 64% in 3.5
GPT-4	Comm. API	Schwieger	Mean 3.12/5	40% halluc.; 17% “ready”
GPT-4	Comm. API	Li et al.	Rel. 4.95/5	Factuality 4.40/5
GPT-4o	Epic Integ.	Challener et al.	All superior	Superior all 9 PDSQI-9
GPT-4o	Comm. API	Li et al.	Compl. 4.55/5	Best ROUGE-L (0.16)
GPT-o3-mini	Reasoning	Croxford et al.	ICC 0.818	Best inter-rater reliability
LLaMA 3 8B	Open Source	Hains et al.	19.1/31	Locally deployable
LLaMA 3 70B	Open Source	Hains et al.	19.2/31	No sig. diff vs 8B
Fine-tuned	Open Source	Li et al.	BLEU 0.04	Best length alignment
PEGASUS	Pre-trained	Fu et al.	+5.6% R-L	Best w/ struct. data

Table 5: Automatic metric values reported in studies using quantitative evaluation.

Study	R-1	R-2	R-L	BLEU	Other Metrics
Li et al. (GPT-4o)	0.38	0.11	0.16	–	Sem. Sim.: 0.83
Li et al. (LLaMA)	–	–	0.16	0.04	Sem. Sim.: 0.837
Fu et al. (base)	–	–	0.176	–	–
Fu et al. (+struct)	–	–	0.232	–	+5.6% improvement
Trejo et al.	0.43	0.22	0.24	–	Spanish (es-es), n=35

Table 6: Human evaluation methodologies across studies.

Study	Raters	Blind	Instrument
Schwieger et al.	8	Yes	Custom 15-item
Challener et al.	5	Yes	Mod. PDQI-9
Williams et al.	2	No	Error taxonomy
Croxford et al.	Expert	Yes	PDSQI-9
Grazhdanski et al.	2	Yes	SUS (10 item)
Trejo-Omeñaca et al.	Clinical	No	4-pt Usefulness

3.3. RQ3: Technical Challenges and System Architectures

Hallucinations and information omission represent the most critical barriers to clinical deployment. Table 7 presents detailed error analysis across studies.

Williams et al. (2025) developed a detailed hallucination taxonomy identifying four primary error patterns: redacted information filling (where LLMs inferred content from redacted/de-identified portions), fabricated follow-up (hallucinated specialty appointments not arranged), incorrect precautions (fabricated ED return instructions), and plan conflation (interim plans reported as final discharge plans). Notably, hallucinations were most concentrated in the Plan and Recommendations sections, which paradoxically are most critical for patient safety.

Information omission emerged as equally problematic: Williams et al. (2025) reported 47% omission rates, while Trejo Omeñaca et al. (2025) found 53% of summaries missing critical information, particularly social context (52%) and temporal details. This suggests that LLMs may prioritize fluency over completeness when generating clinical text.

The 0% hallucination rate reported by Challener et al. (2025) warrants careful interpretation in this context: Epic’s tool uses a proprietary, highly restrictive prompt that excludes structured data (labs, medications, procedures) and existing summaries, limiting the scope for factual errors. Additionally, the study employed blinded physician reviewers using a validated 9-item instrument (PDSQI-9), which may have different sensitivity thresholds than the error taxonomies used in other studies. This methodological heterogeneity makes direct comparisons across studies challenging.

Context window constraints force significant compromises across implementations. Schwieger et al. (2024) limited inputs to 1,000-5,000 tokens for GPT-4’s 8,192 context window, reserving space for output generation. Challener et al. (2025) reported that Epic’s tool requires multi-pass processing when chart data exceeds context limits, with each iter-

Table 7: Detailed hallucination and error analysis across studies.

Study	Hall.	Inacc.	Omiss.	Error Types
Williams et al.	42%	10%	47%	Hallucinated follow-up, redacted info filled, incorrect precautions
Schwieger et al.	40%	30%	–	37.5% relevant; concentrated in recommendations
Trejo et al.	47%	–	53%	Medication dose (56%), social context (52%), temporal inconsistency
Hains et al.	Low	1 case	–	Inc. COVID antiviral; patient/clinician name confusion
Challener et al.	0%[†]	–	–	Restrictive prompts; excludes structured data

ation building on previous outputs. Hains et al. (2025) excluded ICU notes and certain documentation types entirely due to naming convention differences. Trejo Omeñaca et al. (2025) reported lost-in-the-middle effects for lengthy documents, where information in the middle of context windows is poorly utilized.

Two production deployment architectures were evaluated in detail. The Epic Systems Integration described by Challener et al. (2025) uses GPT-4o with a proprietary Epic prompt, automatic extraction of prioritized clinical notes, multi-pass processing for oversized charts, and excludes structured data (labs, procedures) and existing summaries. However, its black-box prompt design limits reproducibility. The Custom Hospital Integration implemented by Trejo Omeñaca et al. (2025) at Hospital General de Granollers, Spain, employs multi-agent orchestration with specialized agents, provides bilingual support (Spanish/Catalan), and followed a three-phase development process: proof-of-concept, prototype, and pilot. During the pilot phase, 47 discharge reports were generated with physicians rating them 2.9/4 average usefulness.

Regarding computational requirements, studies predominantly rely on API-based access, limiting computational transparency. API-based approaches (GPT-4, GPT-4o) dominate due to ease of integration. Local deployment, as demonstrated by Hains et al. (2025), enables privacy preservation but requires local GPU infrastructure. Evaluation with LLM-as-Judge completed in 22 seconds per summary (Croxford et al., 2025).

3.4. RQ4: Factuality, Privacy, and Toward Autonomous Systems

All production-deployed systems currently require physician review. However, our analysis identifies three strategies that may eventually reduce human-in-the-loop dependency. First, knowledge graph validation as proposed by Grazhdanski et al. (2025) enables automated correction using SPARQL queries against medical ontologies (SNOMED, LOINC, RxNorm, DrugBank); notably, their 93.65% faithfulness score was achieved on *synthetic* data generated from MSD Manuals, not real clinical records, and may not generalize to production settings. Second, the LLM-as-Judge framework described by Croxford et al. (2025) demonstrates that reasoning models (GPT-o3-mini, ICC 0.818) can reliably flag problematic summaries for human review, enabling selective rather than universal physician oversight. Third, structured data integration as shown by Fu et al. (2025) reduces hallucination potential for factual content by supplementing free-text generation with verified structured EHR data (medications, labs, procedures).

Table 8 presents comprehensive quality metrics across studies.

Three strategies address privacy concerns in clinical text generation. Local model deployment, as demonstrated by Hains et al. (2025), shows that locally-deployed LLaMA 3 models avoid transmitting patient data to external servers, achieving GDPR/HIPAA compliance by design. Synthetic data training, as employed by Grazhdanski et al. (2025), generated 900 synthetic discharge summaries grounded in MSD Manuals without any real patient data, enabling model training and benchmarking without privacy risks. De-identification preprocessing, used by multiple studies (Williams et al., Schwieger et al.), involves applying pseudonymization before API submission, though this introduces risk of residual identifiers.

Based on our analysis, we identify conditional scenarios where reduced human oversight may become acceptable: low-risk sections (automated generation of administrative content such as demographics and medication lists from structured sources with minimal oversight), high-confidence outputs (selective review triggered by LLM-as-Judge quality scores below threshold, e.g., ICC < 0.8), knowledge graph verification (factual claims verified against medical ontologies before inclusion), and dual-LLM verification (independent generation and cross-validation using separate model instances). However, current hallucination rates (33-47%) and the concentration of errors in safety-critical sections (Plan, Recommendations) indicate that full human review remains essential for patient safety.

Table 8: Quality scores achieved across studies.

Study: Key Quality Score	Study: Key Quality Score
Croxford et al.: ICC 0.818 (LLM-as-Judge)	Hains et al.: 19.2/31 (61.9% of human)
Grazhdanski et al.: SUS 94.35%, Faithfulness 93.65%	Trejo-Omeñaca et al.: 2.9/4 usefulness
Challener et al.: LLM superior all 9 PDSQI-9 criteria	Williams et al.: 33% error-free (GPT-4)
Li et al.: Relevance 4.95/5, Completeness 4.55/5	Schwieger et al.: Human 3.78/5 vs AI 3.12/5

4. Discussion

Our review captures a critical inflection point in clinical NLP. Early MIMIC-based work (Johnson et al., 2016) established foundational datasets, while pre-transformer approaches struggled with long clinical documents. The emergence of GPT-3.5/GPT-4 in 2022-2023 marked the first clinically viable LLM capabilities for discharge summary generation.

This review reveals a rapidly maturing field where LLM-based discharge summary generation has progressed from proof-of-concept to production. GPT-4 dominates current research (6/9 studies), though open-source alternatives (LLaMA 3) address privacy concerns. The high hallucination rates (33-64%) represent the most significant barrier to autonomous deployment, though human summaries also contain errors (10%). Information omission rates up to 53% suggest LLMs prioritize fluency over completeness. The evaluation landscape is evolving toward validated instruments (PDSQI-9) and LLM-as-Judge approaches (Croxford et al., 2025).

Regarding study limitations, sample sizes ranged from 10 to 900, with most studies (6/9) using fewer than 100 cases. Only two studies (Croxford et al., 2025; Schwieger et al., 2024) employed blinded evaluation with multiple raters. All studies except one were single-center, and only Trejo Omeñaca et al. (2025) addressed multilingual settings. These factors suggest caution when generalizing findings.

In terms of practical implications, hybrid workflows are essential: LLM-generated summaries should serve as drafts requiring physician review. Furthermore, model selection must be carefully aligned with the clinical context: commercial APIs (e.g., GPT-4o) offer superior reasoning and text generation quality suitable for environments where infrastructure permits, whereas locally deployed open-source models (e.g., LLaMA) provide strict data privacy and regulatory compliance for sensitive EHR integrations. Regarding standardization, we propose minimum evaluation standards for future studies: clinical text generation evaluations must incorporate concurrent human expert assessment using validated instruments (such as the PDSQI-9) alongside automatic metrics, and transparently report inter-rater reliability. EHR-integrated solutions built upon these foundations demonstrate the most viable path to clinical utility.

5. Conclusions

This systematic literature review analyzed 9 primary studies on automatic generation of hospital discharge summaries using LLMs, published between 2023 and 2026. Our findings demonstrate that GPT-4 and its variants currently dominate research, appearing in 67% of studies, while open-source LLaMA alternatives offer privacy-preserving deployment options. Evaluation practices are maturing, with hybrid approaches combining automatic metrics (ROUGE, BLEU) with validated clinical instruments (PDQI, PDSQI-9) and emerging LLM-as-Judge paradigms. Hallucinations remain a critical challenge, with rates ranging from 33% to 64%, necessitating human-in-the-loop validation for clinical deployment. Privacy can be addressed through local model deployment, knowledge graph-based validation, and synthetic data generation methodologies.

Future research should focus on multi-center validation studies, specialty-specific prompt optimization, temporal consistency mechanisms to prevent propagation of outdated information, and robust automated hallucination detection systems. The field is poised for broader clinical adoption, pending resolution of safety and reliability concerns.

6. Limitations

This review has several limitations. We searched Scopus and PubMed; while Scopus indexes ACL/EMNLP proceedings, direct searching of ACL Anthology might have identified additional work. We included only English-language studies. Our temporal scope (2023-2026) excludes earlier foundational work, though we provided historical context. Quality assessment was performed by a single reviewer; multiple reviewers with inter-rater reliability (Cohen's κ) would strengthen validity. Publication bias may overrepresent positive results, and some reported metrics (0% hallucination, 93.65% faithfulness) may reflect methodological choices rather than generalizable performance. Furthermore, a formal meta-analysis or quantitative synthesis was not attempted because the extreme

heterogeneity of the selected studies, in terms of evaluation methodologies, mixed automatic and human evaluation instruments, and underlying data sources, precludes direct, reliable numerical comparison.

7. Ethical Considerations

The deployment of LLMs for clinical documentation raises significant ethical concerns. Patient safety remains paramount: hallucination rates of 33-64% documented in this review underscore that autonomous generation without human oversight is currently unacceptable. All production systems reviewed require physician verification before clinical use.

Privacy and data protection present ongoing challenges. Studies using commercial APIs (GPT-4) transmit patient data to external servers. Local deployment of open-source models (Hains et al., 2025) offers a privacy-preserving alternative, though with performance trade-offs. De-identification preprocessing mitigates but does not eliminate re-identification risks.

Transparency and accountability require that generated summaries be clearly marked as AI-assisted, allowing receiving clinicians to apply appropriate scrutiny. The "black-box" nature of commercial LLMs complicates error attribution and quality assurance.

Regarding research ethics, all included studies either obtained IRB approval, used publicly available de-identified data (MIMIC), or employed synthetic data. Future research should ensure appropriate ethical oversight and consider potential biases in training data that may perpetuate health disparities.

8. Acknowledgements

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project CONSENSO (PID2021-122263OB-C21) Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852:) funded by Junta de Andalucía

9. Bibliographical References

Douglas Challener, Shant Ayanian, Alexander Ryu, John O'Horo, and Heather Heaton. 2025. Quality assessment of artificial intelligence-generated versus human-written hospital summaries evaluating detail, usefulness, and continuity of care. *Journal of Hospital Medicine*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M. Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1).

Jiaojiao Fu, Bowen Yang, Yi Guo, Yangfan Zhou, and Xin Wang. 2025. Automated clinical summary generation via integrating structured and unstructured data. *Communications in Computer and Information Science*, 2301:290 – 304.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170.

Lewis Hains, Oliver Kleinig, Ashwin Murugappa, Samuel Gluck, Jarrod Marks, Toby Gilbert, and Stephen Bacchi. 2025. Large language model discharge summary preparation using real-world electronic medical record data shows promise. *Internal Medicine Journal*, 55(7):1188 – 1192.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Keele University and Durham University.

Sunil Kripalani, Frank LeFevre, Christopher O. Phillips, Mark V. Williams, Preetha Basaviah, and David W. Baker. 2007. Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care. *JAMA*, 297(8):831–841.

Yiming Li, Fang Li, Na Hong, Manqi Li, Kirk Roberts, Licong Cui, Cui Tao, and Hua Xu. 2025. A comparative study of recent large language models on generating hospital discharge summaries for lung cancer patients. *Journal of Biomedical Informatics*, 168.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjorn Hrobjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Mark Petticrew and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences: A Practical Guide*. Blackwell Publishing, Oxford, UK.

Arne Schwieger, Katrin Angst, Mateo de Bardeci, Achim Burrer, Flurin Cathomas, Stefano Ferrea, Franziska Grätz, Marius Knorr, Golo Kronenberg, Tobias Spiller, David Troi, Erich Seifritz, Samantha Weber, and Sebastian Olbrich. 2024. Large language models can support generation of standardized discharge summaries – a retrospective study utilizing chatgpt-4 and electronic health records. *International Journal of Medical Informatics*, 192.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Alex Trejo Omeñaca, Esteve Llargués Rocabruna, Jonny Sloan, Michelle Catta-Preta, Jan Ferrer i Picó, Julio Cesar Alfaro Alvarez, Toni Alonso Solis, Eloy Lloveras Gil, Xavier Serrano Vinaixa, Daniela Velasquez Villegas, Ramon Romeu Garcia, Carles Rubies Feijoo, Josep Maria Monguet i Fierro, and Beatriu Bayes Genis. 2025. Leave as fast as you can: Using generative ai to automate and accelerate hospital discharge reports. *Computers*, 14(6).

Christopher Y.K. Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, Atul J. Butte, and Aaron E. Kornblith. 2025. Evaluating large language models for drafting emergency department encounter summaries. *PLOS Digital Health*, 4(6):e0000899.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

10. Language Resource References

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.