Article

Not peer-reviewed version

Prompt Architecture as a High-Impact Design Factor in Expert-Rated Clinical Documentation Quality: A Controlled Comparative Study in Inpatient Rehabilitation

Idoia Eceizabarrena-Matxinandiarena *, Emilio Javier Frutos-Reoyo, José Ignacio Guerrero-Rojas, Clara Vidal-Millet, Pedro Tejada-Ezquerro, Elena Roldan-Arcelus, Irene de Torres-García, Judith Sanchez-Raya, Lourdes Gil-Fraguas, María Hernandez-Manada, Carolina de Miguel-Benadiba, Josep Monguet-Fierro, Alejandro Trejo-Omeñaca, Michelle Cavariani Catta-Preta, Astrid Teixeira-Taborda, Natalia Álvarez-Bandrés, Raquel Cutillas-Ruiz, Helena Bascuñana-Ambrós

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0054.v1

Keywords: artificial intelligence; clinical documentation; discharge reports; large language models; medical writing; prompt architecture; prompt engineering; rehabilitation medicine

Article

Prompt Architecture as a High-Impact Design Factor in Expert-Rated Clinical Documentation Quality: A Controlled Comparative Study in Inpatient Rehabilitation

Idoia Eceizabarrena-Matxinandiarena [1,*], Emilio Javier Frutos-Reoyo [2,3], José Ignacio Guerrero-Rojas [4], Clara Vidal-Millet [5], Pedro Tejada-Ezquerro [6], Elena Roldan-Arcelus [7], Irene de Torres-García [8], Judith Sanchez-Raya [9], Lourdes Gil-Fraguas [10], María Hernandez-Manada [11], Carolina de Miguel-Benadiba [12], Josep Monguet-Fierro [13,14], Alejandro Trejo-Omeñaca [13,14], Michelle Cavariani Catta-Preta [14], Astrid Teixeira-Taborda [15], Natalia Álvarez-Bandrés [16], Raquel Cutillas-Ruiz [17] and Helena Bascuñana-Ambrós [18]

1 Hospital Universitario Donostia, San Sebastián, Spain
2 Hospital Universitario Río Hortega de Valladolid, Valladolid, Spain
3 Hospital General de Segovia, Segovia, Spain
4 Hospital General Universitario La Mancha Centro, Alcázar de San Juan, Spain
5 Hospital Universitario de la Ribera, Alzira, Spain
6 Hospital de Gorliz, Gorliz, Spain
7 Hospital Universitario de Navarra, Pamplona, Spain
8 Hospital Universitario Reina Sofía, Córdoba, Spain
9 Hospital Universitari Vall d'Hebron, Barcelona, Spain
10 Hospital Universitario de Guadalajara, Guadalajara, Spain
11 Khore Global Consulting, Madrid, Spain
12 Hospital Beata María Ana, Madrid, Spain
13 Universitat Politècnica de Catalunya (UPC), Barcelona, Spain.j
14 Innex Labs, Vilanova i la Geltrú, Spain
15 Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain; astrid.teixeira.taborda@sermef.es
16 Hospital Universitario San Pedro, Logroño, Spain
17 Hospital Universitario Fundación Jiménez Díaz, Madrid, Spain
18 Hospital de la Santa Creu i Sant Pau, Campus Salut Barcelona, Spain
* Correspondence: idoia.eceizabarrenamatxinandiarena@osakidetza.eus.

Abstract

Large language models (LLMs) are increasingly explored for clinical documentation support, yet the influence of prompting architecture on documentation quality in complex longitudinal contexts remains poorly characterized. This controlled retrospective methodological study evaluated three prompting strategies—Single Prompt (SP), Section-Based Prompt (SBP), and Section-Based Prompt with Writing Refinement (SBP+W)—for generating inpatient rehabilitation discharge reports using OpenAI large language model (GPT-5.2). Twenty anonymized rehabilitation cases involving prolonged hospital stays and multidimensional func-tional documentation were processed under standardized model conditions. AI-generated reports were compared with human-authored summaries. Two blinded board-certified rehabilitation physicians in-dependently evaluated outputs using a structured 4-point ordinal scale assessing structural integrity, clinical coherence, completeness, and readability. Inter-rater reliability was estimated with quadratic weighted Cohen's kappa and bootstrap confidence intervals. Group differences were analyzed using non-parametric testing and exploratory multivariable modeling. All LLM prompting strategies achieved significantly higher expert-rated quality scores than hu-man-authored reports ($p < 0.01$). SBP demonstrated the highest median performance and strongest regression effect, although differences among LLM-based strategies were not statistically significant after correction. Prompting strategy explained more variability in expert ratings than case-level factors. Structured section-based prompting may

 doi:10.20944/preprints202604.0054.v1

represent a practical design lever for improving perceived quality in AI-assisted clinical documentation workflows.

Keywords: artificial intelligence; clinical documentation; discharge reports; large language models; medical writing; prompt architecture; prompt engineering; rehabilitation medicine

1. Introduction

Large language models (LLMs) have rapidly emerged as powerful generative systems capable of producing structured natural language across a wide range of knowledge-intensive domains, including healthcare. Recent studies demonstrate that LLM-based systems can assist clinicians in producing clinical documentation, discharge summaries, and patient instructions with levels of readability and structural consistency comparable to, and in some contexts exceeding, those of human-authored documentation [1–6]. Given that clinical documentation represents a major source of administrative workload for physicians, AI-assisted documentation tools are increasingly explored as a strategy to improve efficiency, reduce cognitive burden, and enhance documentation standardization.

Despite these advances, the reliability of LLM-generated clinical documentation remains a central concern. Prior research has highlighted potential limitations including hallucinated information, omission of clinically relevant details, contextual drift across long documents, and instability in reasoning when synthesizing heterogeneous medical information [7–10]. These challenges are particularly pronounced in longitudinal clinical contexts, where documentation must integrate multiple assessments, evolving diagnoses, and multidisciplinary interventions into a coherent narrative. Inpatient rehabilitation discharge summaries represent a demanding example of such documentation tasks, requiring synthesis of functional assessments, therapeutic interventions, and recovery trajectories within a structured clinical report.

From a systems design perspective, increasing attention has focused on prompt engineering as a key factor influencing LLM behavior. Prompt architecture—the structural design of instructions provided to the mod-el—can shape how generative models interpret tasks, organize information, and produce outputs. Contemporary prompt engineering frameworks describe multiple architectural strategies including monolithic prompts, modular task decomposition, hierarchical prompting, and chain-of-thought reasoning [11–16]. Experimental work in computational domains suggests that modular and decomposed prompting strategies may improve reasoning stability, reduce contextual interference, and enhance output consistency.

However, empirical evidence evaluating the impact of prompt architecture in real-world clinical documentation tasks remains limited. Most existing healthcare studies focus on evaluating overall output quality or safety of AI-generated documentation without systematically isolating prompt structure as an experimental variable [3,4,17,18]. As a result, it remains unclear whether observed variability in LLM-generated documentation primarily reflects intrinsic model properties (e.g., model scale or training data) or design choices in prompt architecture.

Understanding the relative influence of prompt architecture is important for the safe implementation of LLM systems in clinical work-flows. If prompt design significantly affects documentation quality, re-liability improvements could be achieved through controllable inter-face-level interventions rather than requiring new model training or scaling. Such insights would position prompt architecture as a practical systems engineering lever for optimizing generative AI in healthcare environments.

To address this gap, we conducted a controlled comparative methodological study evaluating three prompting architectures—Single Prompt (SP), Section-Based Prompt (SBP), and Section-Based Prompt with Writing Refinement (SBP+W)—for the generation of inpatient rehabilitation discharge

 doi:10.20944/preprints202604.0054.v1

reports using an OpenAI large language model (GPT-5.2). We hypothesized that modular section-based prompting would improve expert-rated documentation quality compared with monolithic prompting approaches, and that prompt architecture would explain a substantial proportion of variability in documentation performance across clinical cases.

2. Materials and Methods

2.1. Study Design

A retrospective, comparative methodological study was conducted to evaluate the impact of different prompting architectures on the quality of AI-generated rehabilitation discharge reports. The study was designed following methodological principles commonly applied in clinical artificial intelligence evaluation frameworks, including controlled input conditions, blinded expert assessment, and structured qualitative and quantitative analysis.

The primary objective was to compare three prompting strategies in terms of structured clinical documentation quality, functional coherence, and overall expert-rated performance.

A detailed schematic representation of the study is shown in Figure 1.

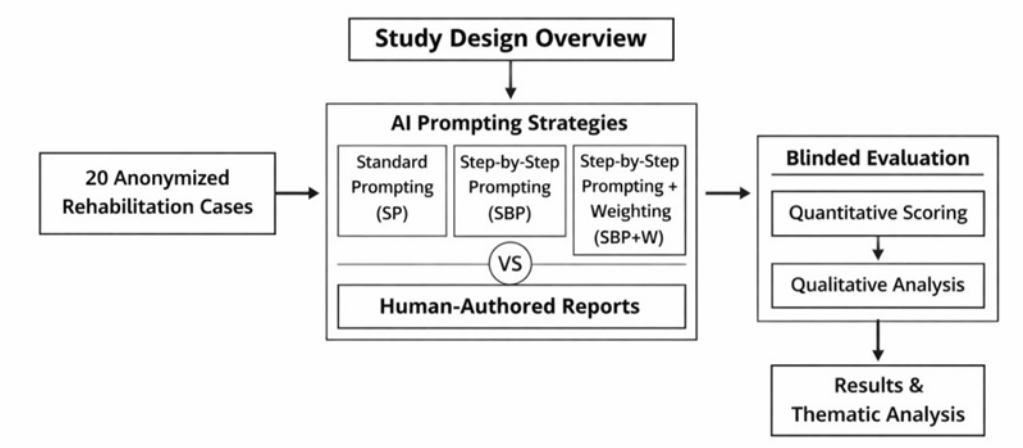

Figure 1. Schematic representation of the study design. Twenty anonymized inpatient rehabilitation cases were processed using three prompting strategies (SP, SBP, SBP+W) and compared with human-authored reports. Two board-certified rehabilitation physicians evaluated all outputs under blinded conditions. Quantitative ordinal scoring and qualitative affinity-based thematic analysis were performed.

2.2. Data Source and Case Selection

Twenty anonymized inpatient rehabilitation cases were retrospectively selected from multiple national public hospitals. Cases were purposely sampled to reflect the complexity of rehabilitation medicine, including prolonged length of stay, multidimensional functional assessments, and extensive narrative documentation. All clinical data were de-identified prior to model processing in accordance with applicable data protection standards. No identifiable patient information was introduced into the generative system.

2.3. Prompt Design Strategies

Three prompt design strategies were evaluated for generating hospital discharge summaries from clinical documentation. See more details in Table 1.

1. Single Prompt (SP): A single prompt structured using the Prompt Canvas framework defined the model's clinical role, generation constraints (use only provided documentation, prioritize the most recent information, and flag inconsistencies), a four-step processing workflow (clinical analysis, chronological synthesis, structured report generation, and validation), a predefined

 doi:10.20944/preprints202604.0054.v1

discharge summary template, and output formatting suitable for electronic medical records (EMR).

2. Section-Based Prompt (SBP): A modular architecture of 10 sequential prompts was implemented, each generating one section of the discharge summary (e.g., chief complaint, past medical history, history of present illness, physical examination, clinical course, complementary tests, functional scales, diagnosis, and management plan). Each prompt followed the Prompt Canvas structure with section-specific instructions.
3. Section-Based Prompt with Writing-Specific Instruction (SBP+W): This hierarchical approach divided each section into two stages: (1) extraction of clinically relevant information from progress notes and (2) drafting of the section using only the extracted content.

For SBP and SBP+W, the final discharge summary was assembled manually to avoid information loss caused by model-driven synthesis.

Table 1. Comparative Overview of Prompt Strategies.

Strategy	Architecture	Nº of Prompts	Key Mechanism	Advantages	Limitations
SP	Single structured prompt	1	Prompt Canvas workflow guiding full report generation	Simpler implementation; faster generation	Higher cognitive load for the model
SBP	Modular section-based prompts	10	Each prompt generates one report section	Better control over structure and content	Requires sequential prompting
SBP+W	Section-based prompts + writing	20	Two-step process: information extraction followed by drafting	Higher fidelity to source notes; reduced hallucination risk	Increased prompt complexity and execution time

2.4. Model Configuration and Execution

All generations were performed using GPT-5.2 via web access. The model was selected due to its advanced reasoning capacity and improved instruction adherence relative to smaller-scale models.

To ensure methodological rigor and reproducibility:

- Identical system-level instructions were maintained across all conditions.
- The generation parameters (web access, the GPT-5.2 Thinking model, extended reasoning) were kept fixed in all cases.
- No case-specific prompt modifications were introduced once the experimental protocol was finalized.
- Each case was processed independently to prevent cross-case contextual contamination.

 doi:10.20944/preprints202604.0054.v1

2.5. Outcome Measures

2.5.1. Quantitative Expert Evaluation

For each clinical case, four discharge reports were evaluated:

- A: Original human-authored report
- B: SP-generated report
- C: SBP-generated report
- D: SBP+W-generated report

Two independent board-certified rehabilitation physicians assessed all reports under blinded conditions with respect to the generation strategy.

Reports were rated using a 4-point ordinal scale assessing structural integrity, clinical accuracy, coherence, completeness, and readability (see Appendix A.1 for more details):

- 0 = Poor
- 1 = Fair
- 2 = Good
- 3 = Excellent

In cases where inter-rater discrepancies exceeded two points, a third independent expert adjudicated the evaluation. The adjudicated score was considered final for analysis.

Inter-rater reliability will be assessed using weighted Cohen's kappa.

2.5.2. Qualitative Expert Feedback

Reviewers provided free-text comments regarding the strengths and weaknesses of each report. These qualitative data will undergo structured thematic analysis.

An affinity diagram methodology will be applied to cluster qualitative observations into emergent thematic categories (e.g., coherence, functional specificity, redundancy, clinical reasoning adequacy, stylistic clarity). Two researchers will independently code comments, followed by consensus-based category consolidation to enhance analytic rigor.

2.6. Statistical Analysis

All statistical analyses were conducted using validated statistical software. Because expert ratings were measured on a 4-point ordinal scale (0–3), non-parametric methods were applied for primary comparisons.

Descriptive statistics are reported as mean ± standard deviation (SD), and median with interquartile range (IQR).

To compare documentation quality across the four report types (human-authored, SP, SBP, SBP+W), a Friedman test for repeated measures was performed. Effect size was calculated using Kendall's W.

When the Friedman test was statistically significant, post hoc pairwise comparisons were conducted using Wilcoxon signed-rank tests with Bonferroni correction for multiple testing.

To evaluate the relative contribution of prompting strategy and case-level variability, a multivariable linear regression model was fitted with expert rating as the dependent variable and case and strategy as predictors. The human-authored report served as the reference category. Model performance was assessed using R^2, adjusted R^2, F-statistics, and associated p- values.

Inter-rater reliability was assessed using quadratic weighted Cohen's kappa (κ). Ninety-five percent confidence intervals were calculated using bootstrap resampling (2000 iterations).

All tests were two-tailed, and statistical significance was defined as $p < 0.05$.

All analyses were conducted using RStudio 2026.01.1 software. Qualitative analysis was conducted using Chat GPT 5.2 via WEB.

 doi:10.20944/preprints202604.0054.v1

2.7. Ethical Considerations

The study was conducted using fully anonymized retrospective clinical documentation. No patient contact or intervention occurred. Ethical review and approval were waived due to the exclusive use of de-identified data and the non-interventional methodological design.

2.8. Use of Generative Artificial Intelligence

Generative artificial intelligence (GPT-5.2, WEB version) was the primary object of evaluation in this methodological study. The model was used exclusively to generate discharge reports under controlled prompting conditions. No automated clinical decisions were made. All outputs were evaluated by human experts. The authors supervised prompt design, analytical procedures, and manuscript preparation, and assumed full responsibility for the integrity and accuracy of the study.

3. Results

3.1. Study Sample

Twenty inpatient rehabilitation cases were included (see descriptive statistics in Table 2), generating four discharge reports per case (human-authored, SP, SBP, SBP+W), for a total of 80 evaluated documents. Cases were characterized by prolonged admissions and complex longitudinal documentation.

Table 2. Descriptive statistics of ratings across strategies.

Strategy	Mean	Median	SD	IQR
A	1.65	2.00	0.69	1.00
B	2.33	2.00	0.57	1.00
C	2.49	3.00	0.63	1.00
D	2.14	2.00	0.71	1.00

SD: standard deviation; IQR: interquartile range.

3.1. Comparative Performance Across Prompting Strategies

A Friedman test revealed a statistically significant difference in mean quality scores between strategies, $\chi^2(3) = 23.93$, $p < 0.001$. The effect size was moderate-to-large (Kendall's W = 0.40), indicating substantial differences in performance across strategies. Go to Figure 2 and Table 3 to compare the results visually.

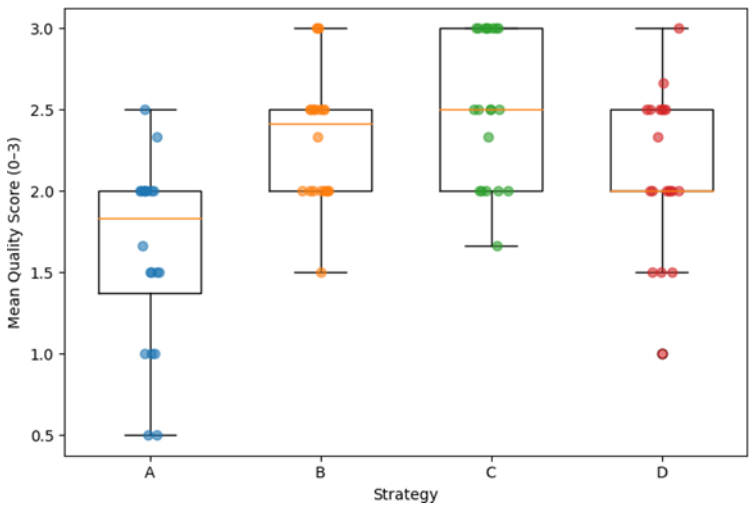

Figure 2. Comparison of Quality Scores Across Strategies.

 doi:10.20944/preprints202604.0054.v1

Table 3. Comparison of Quality Scores Across Strategies.

Strategy	Median (IQR)
A	1.83 (1.38-2.00)
B	2.42 (2.00-2.50)
C	2.50 (2.00-3.00)
D	2.00 (2.00-2.50)

Friedman test: $\chi^2(3) = 23.93$, $p = 2.59e\text{-}05$. Kendall's $W = 0.399$.

Post-hoc pairwise comparisons using Wilcoxon signed-rank tests with Bonferroni correction showed that Strategy A differed significantly from Strategy B ($p = 0.003$), Strategy C ($p = 0.003$), and Strategy D ($p = 0.025$). No statistically significant differences were observed between Strategies B, C, and D after correction (see Table 4).

Table 4. Post-hoc Wilcoxon Signed-Rank Test with Bonferroni Correction.

Comparison	*p*
A vs. B	0.00346
A vs. C	0.00255
A vs. D	0.02455
B vs. C	0.84722
B vs. D	0.98388
C vs. D	0.10482

3.1. Multivariable Analysis

To disentangle the relative contributions of case complexity and prompting strategy, a linear regression model was fitted:

Linear regression model: MedianReviewers ~ Case + Strategy.

The model was statistically significant ($F = 3.128$, $p = 0.0002853$), explaining 54.7% of variance ($R^2 = 0.5469$; adjusted $R^2 = 0.3721$). SBP demonstrated the largest effect size and strongest statistical significance (see Table 5).

Table 5. Multivariable linear regression model evaluating the effect of case and prompting strategy on expert ratings.

Predictor	Estimate	Std. Error	*p*-value
SP	+0.675	0.149	<0.001
SBP	+0.850	0.149	<0.001
SBP+W	+0.500	0.149	0.001

Model statistics: $R^2 = 0.5469$; Adjusted $R^2 = 0.3721$; $F(22,57) = 3.128$; $p = 0.0002853$.

In contrast, most case-level coefficients were non-significant, indicating that prompt architecture explained substantially more variance in quality ratings than case-specific complexity.

3.1. Inter-Rater Reliability

Quadratic weighted Cohen's kappa (see Table 6) indicated fair overall agreement ($\kappa = 0.354$; 95% CI 0.15– 0.53). Strategy-level kappas were lower and exhibited wide confidence intervals.

 doi:10.20944/preprints202604.0054.v1

Table 6. Quadratic weighted Cohen's kappa (κ) for inter-rater reliability across strategies.

Strategy	κ	95% CI	N
Overall	0.354	0.15-0.53	80
A (Human)	0.322	-0.08-0.62	20
B (SP)	0.250	-0.12-0.56	20
C (SBP)	0.097	-0.14-0.50	20
D (SBP+W)	0.095	-0.25-0.46	20

κ, Cohen's kappa coefficient; CI, confidence interval; SP, single prompt; SBP, section-based prompt; SBP+W, section-based prompt with written-specific instructions.

Inter-rater agreement between the two primary reviewers across all reports and stratified by prompting strategy. Confidence intervals were calculated using bootstrap resampling (2000 iterations).

Given the blinded, within-case comparative design, moderate inter-rater agreement does not invalidate relative strategy differences, but it underscores the inherent subjectivity of qualitative documentation assessment.

3.1. Qualitative Synthesis

Thematic clustering of reviewer comments identified five recurrent domains: structural coherence, functional specificity, redundancy, reasoning adequacy, and stylistic clarity.

The distribution of qualitative codes across strategies is presented in Figure 3. Modular prompting (SBP) demonstrated the highest thematic density in structural and completeness- related domains, whereas SBP+W was more frequently associated with verbosity-related observations.

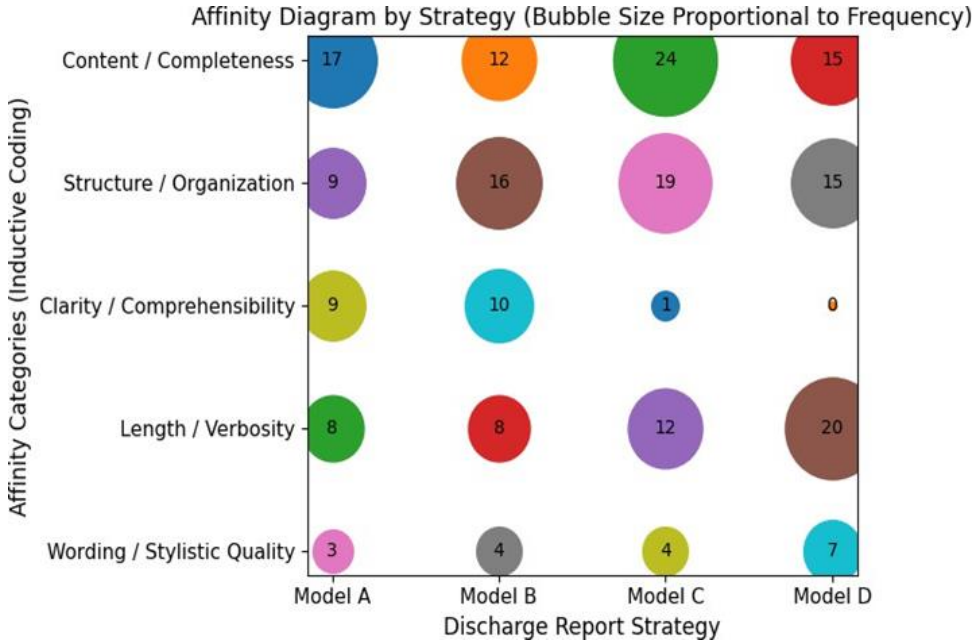

Figure 3. Affinity diagram of qualitative reviewer feedback by strategy. Bubble size is proportional to the frequency of coded observations within each thematic domain. Comments were clustered into five categories: content/completeness, structure/organization, clarity/comprehensibility, length/verbosity, and wording/stylistic quality. Modular section-based prompting concentrated on structural and completeness-related observations, whereas writing- refinement prompting was more frequently associated with verbosity-related comments.

 doi:10.20944/preprints202604.0054.v1

SBP was most consistently associated with improved structural organization and logical sequencing. SP outputs were more variable in coherence, while human-authored reports demonstrated contextual nuance but less structural standardization.

Qualitative findings converged with quantitative results, reinforcing the interpretation that architectural segmentation improves perceived documentation quality.

4. Discussion

This study provides controlled empirical evidence that prompt architecture influences expert-rated documentation quality in complex clinical reporting tasks. While previous investigations have demonstrated the feasibility of AI-generated discharge summaries in acute and primary care settings, few have systematically isolated prompting architecture as an independent experimental variable in longitudinal rehabilitation documentation [1–6,19].

In this controlled within-case design, all LLM-based strategies received higher expert-rated quality scores than non-standardized human-authored reports. Importantly, these findings reflect perceived structural integrity, coherence, and completeness rather than objective safety or factual error metrics. The evaluation instrument captured global documentation quality as assessed by expert clinicians, and results should therefore be interpreted within that evaluative framework [7,20].

Among prompting strategies, Section-Based Prompting (SBP) achieved the highest median scores and largest regression coefficient. However, post-hoc pairwise comparisons did not demonstrate statistically significant differences between LLM-based strategies after correction for multiple testing. Accordingly, differences among SP, SBP, and SBP+W should be interpreted cautiously, particularly given the limited sample size.

The exploratory multivariable model suggested that prompting strategy accounted for a larger share of explainable variability in expert ratings than case-level variability. While this observation supports the hypothesis that architectural design exerts measurable influence, the dependent variable was ordinal and the regression model should be understood as an approximation for variance partitioning rather than a definitive causal estimator. Larger samples and ordinal regression approaches may provide more robust quantification in future studies.

Inter-rater agreement was moderate ($\kappa = 0.354$), consistent with the inherent subjectivity involved in evaluating complex longitudinal clinical documentation. Nevertheless, the blinded within-case comparative design reduces systematic bias in relative strategy comparisons, as each prompting condition was evaluated against identical clinical inputs.

An important interpretative nuance concerns the comparator. Human-authored reports were not generated under standardized structural constraints. Observed differences may therefore reflect the impact of architectural standardization rather than intrinsic generative superiority. In this sense, findings highlight the potential benefit of structured documentation scaffolding—whether implemented through AI prompting or structured human templates [21,22].

The absence of incremental benefit from the additional writing-refinement layer (SBP+W) suggests that macro-structural segmentation may exert greater influence on perceived quality than post hoc stylistic enhancement. From a bioengineering systems perspective, this finding supports conceptualizing prompt architecture as a cognitive interface layer that shapes generative output organization under long-context constraints [11–16,23–25].

Limitations include modest sample size, single-model evaluation (GPT-5.2), retrospective design, and the absence of objective factual error auditing or safety outcome assessment. Furthermore, the study was conducted within inpatient rehabilitation discharge documentation, and generalization to other clinical domains should be considered provisional pending replication.

Future research should incorporate cross-model comparisons, ordinal modeling approaches, structured error audits, and prospective workflow studies evaluating time efficiency, cognitive load, and clinical integration [26–30].

 doi:10.20944/preprints202604.0054.v1

5. Conclusions

In the context of inpatient rehabilitation discharge documentation using GPT-5.2, prompt architecture significantly influenced expert-rated documentation quality. Structured section-based prompting achieved the highest median performance and demonstrated consistent advantages in perceived structural organization and completeness.

Within this dataset, the prompting strategy accounted for a larger share of explainable variability in expert ratings than case-level variability, suggesting that architectural prompt design may represent a controllable implementation variable in clinical LLM deployment. However, findings should be interpreted in light of the ordinal evaluation scale, moderate inter-rater agreement, and limited sample size.

From a bioengineering perspective, prompt architecture may function as a systems-level design intervention that shapes the interaction between generative models and clinical documentation workflows. Structured modular prompting appears to offer a promising and scalable strategy for enhancing perceived documentation reliability in complex longitudinal reporting tasks. Replication across models, domains, and safety-oriented evaluation frameworks is required to establish generalizability.

Author Contributions: Conceptualization, J. M.-F. and H. B.-A.; methodology, I. E.-M., E.-J. F.-R., J-I. G.-R. and C. V.-M.; software, A. T.-O., M. C.-P. and J. M.-F.; validation, I. T.-G., E. R.-A., P.T-E., H. B.-A. N. A.-B., R. C.-R., and J. S.-R.; formal analysis, I.E. M.; investigation, I. E.-M., E.-J. F.-R., J-I. G.-R., C. V.-M. M. C.-P and J. M.-F.; resources, A. T.-O. and H.B.-A.; data curation, I. E.-M., E.-J. F.-R., J-I. G.-R. and C. V.-M.; writing—original draft preparation, I. E.-M.; writing—review and editing, P.T-E., H. B.-A. and J. M.-F.; visualization, I. E.-M., E.-J. F.-R., J-I. G.-R. and C. V.-M.; supervision, M. C.-P.; project administration, M. H.-M., A. T.-, L. G.-F., A. T.-O., and C. M.-B.; funding acquisition, H. B.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Society of Physical Medicine and Rehabilitation (SERMEF), grant number 03/2025.

APC Funding: The APC was funded by the Spanish Society of Physical Medicine and Rehabilitation (SERMEF), grant number 03/2025.

Institutional Review Board Statement: This study was conducted in accordance with the principles of the Declaration of Helsinki and complied with applicable Spanish and European regulations on biomedical research and data protection, including the Spanish Biomedical Research Law 14/2007 of 3 July, Royal Decree 1090/2015 of 4 December regulating clinical trials with medicinal products and Research Ethics Committees (CEIm), and Regulation (EU) 2016/679 (General Data Protection Regulation, GDPR). In accordance with Articles 2 and 3 of Law 14/2007 and Article 2 of Royal Decree 1090/2015, review by a Research Ethics Committee (CEIm) was not required because the study consisted exclusively of an anonymous survey of healthcare professionals, without intervention, without the collection of biological samples, and without the collection or processing of personal health data from patients. The study did not involve human subjects in the sense defined by Spanish biomedical research legislation, nor did it constitute a clinical investigation or observational study involving medicinal products or medical devices. Therefore, formal ethical approval by a CEIm was not applicable.

Informed Consent Statement: Participation was voluntary. All participants received an online information sheet detailing the study objectives, methodology, voluntary nature of participation, and data protection safeguards. Electronic informed consent was obtained prior to accessing the questionnaire. Eligibility was verified through screening questions confirming: (i) active membership in the Spanish Society of Physical Medicine and Rehabilitation (SERMEF); and (ii) for participation in the in-person Real-Time Delphi round, a minimum of five years of clinical experience in inpatient rehabilitation discharge reporting.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy restrictions.

 doi:10.20944/preprints202604.0054.v1

Acknowledgments: During the preparation of this manuscript/study, the author(s) used Chat GPT 5.2. via web for the purposes of generating discharge reports and qualitative data analysis.

Conflicts of Interest: All the authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SD	Standard deviation
IQR	Median with interquartile range
CEIm	Comité de Ética de la Investigación con medicamentos
LLMs	Large Language Models
SP	Single Prompt
SBP	Section-Based Prompt
SBP+W	Section-Based Prompt with Writing Refinement
)	World Health Organization

Appendix A

Appendix A.1. Expert Evaluation Rubric

How would you rate the report you just downloaded?

- [0] Poor – The report is incomplete, with serious problems in comprehension, structure, or accuracy. Significant errors and lack of supporting data.
- [1] Fair – The report shows issues with coherence, lacks depth, or contains minor content errors. The writing may be unclear in some areas, affecting readability.
- [2] Good – The report meets the essential requirements. It has a clear structure, reasonable arguments, and appropriate use of information. There may be minor inaccuracies or areas for improvement, but they do not hinder understanding.
- [3] Excellent – The report is very well structured, coherent, complete, and tailored to the target audience. The content is accurate, well-supported, and backed by appropriate data and evidence. The writing is fluent and free of significant errors.

References

1. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open [Internet]. 2024;7(3):e240357. Available from: http://dx.doi.org/10.1001/jamanetworkopen.2024.0357
2. Tang C, Mudunna N, Turner I, Asghari-Jafarabadi M, Joe K, Brichko L. Use of artificial intelligence to generate emergency department discharge summaries. Aust Health Rev [Internet]. 2025;49(3). Available from: http://dx.doi.org/10.1071/AH24326
3. Clough RAJ, Sparkes WA, Clough OT, Sykes JT, Steventon AT, King K. Transforming healthcare documentation: harnessing the potential of AI to generate discharge summaries. BJGP Open [Internet]. 2024;8(1):BJGPO.2023.0116. Available from: http://dx.doi.org/10.3399/BJGPO.2023.0116
4. Omon K, Sasaki T, Koshiro R, Fuchigami T, Hamashima M. Effects of introducing generative AI in rehabilitation clinical documentation. Cureus [Internet]. 2025;17(3):e81313. Available from: http://dx.doi.org/10.7759/cureus.81313
5. De Rosario H, Pitarch-Corresa S, Pedrosa I, Vidal-Pedrós M, de Otto-López B, García-Mieres H, et al. Applications of natural language processing for the management of stroke disorders: Scoping review. JMIR Med Inform [Internet]. 2023;11:e48693. Available from: http://dx.doi.org/10.2196/48693

 doi:10.20944/preprints202604.0054.v1

6. Perkins SW, Muste JC, Alam T, Singh RP. Improving Clinical Documentation with Artificial Intelligence: A Systematic Review. Perspect Health Inf Manag. 2024;21.
7. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. J Rehabil Med [Internet]. 2023;55:jrm13373. Available from: http://dx.doi.org/10.2340/jrm.v55.13373
8. Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE, et al. Artificial intelligence and healthcare: A journey through history, present innovations, and future possibilities. Life (Basel) [Internet]. 2024;14(5):557. Available from: http://dx.doi.org/10.3390/life14050557
9. Buscarini L, Romano P, Cocco ES, Damiani C, Pournajaf S, Franceschini M, et al. Enhancing patient rehabilitation outcomes: artificial intelligence-driven predictive modeling for home discharge in neurological and orthopedic conditions. J Neuroeng Rehabil [Internet]. 2025;22(1). Available from: http://dx.doi.org/10.1186/s12984-025-01654-4
10. Lewis M, Navarro DF, Blease C, Shah R, Riggare S, Delacroix S, et al. Clinical competencies for using generative AI in patient care. BMJ [Internet]. 2025;391:e085324. Available from: http://dx.doi.org/10.1136/bmj-2025-085324
11. Sahoo P, Singh AK, Saha S, Jain V, Mondal S, Chadha A. A systematic survey of prompt engineering in large language models: Techniques and applications [Internet]. arXiv [cs.AI]. 2024. Available from: http://arxiv.org/abs/2402.07927
12. Liu Y-Y, Zheng Z, Zhang F, Feng J-C, Fu Y-Y, Zhai J-D, et al. A comprehensive taxonomy of prompt engineering techniques for large language models. Front Comput Sci [Internet]. 2026;20(3). Available from: http://dx.doi.org/10.1007/s11704-025-50058-z
13. Polat F, Tiddi I, Groth P. Testing prompt engineering methods for knowledge extraction from text. Semant Web [Internet]. 2025;16(2). Available from: http://dx.doi.org/10.3233/sw-243719
14. Viswanathan PS. Prompt engineering for conversational AI systems: A systematic review of techniques and applications. Int J Sci Res Comput Sci Eng Inf Technol [Internet]. 2025;11(1):733–41. Available from: http://dx.doi.org/10.32628/cseit25111276
15. Sasson Lazovsky G, Raz T, Kenett YN. The art of creative inquiry—from question asking to prompt engineering. J Creat Behav [Internet]. 2025;59(1). Available from: http://dx.doi.org/10.1002/jocb.671
16. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns (N Y) [Internet]. 2025;6(6):101260. Available from: http://dx.doi.org/10.1016/j.patter.2025.101260
17. Trejo Omeñaca A, Llargués Rocabruna E, Sloan J, Catta-Preta M, Ferrer i Picó J, Alfaro Alvarez JC, et al. Leave as fast as you can: Using generative AI to automate and accelerate hospital discharge reports. Computers [Internet]. 2025;14(6):210. Available from: http://dx.doi.org/10.3390/computers14060210
18. Chen X, Xiang J, Lu S, Liu Y, He M, Shi D. Evaluating large language models and agents in healthcare: key challenges in clinical applications. Intell Med [Internet]. 2025;5(2):151–63. Available from: http://dx.doi.org/10.1016/j.imed.2025.03.002
19. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ [Internet]. 2025;r340. Available from: http://dx.doi.org/10.1136/bmj.r340
20. White J, Elnashar A, Schmidt D. Prompt engineering for structured data A comparative evaluation of styles and LLM performance [Internet]. Preprints. 2025. Available from: http://dx.doi.org/10.20944/preprints202506.1937.v1
21. Mohamed MAH, Al-Mhdawi MKS, Qazi A, Mahammedi C, Ojiako GU, Dacre N. An analytical framework for evaluating generative intelligence risks in sustainable construction [Internet]. 2025. Available from: http://dx.doi.org/10.2139/ssrn.5289325
22. Moniri B, Hassani H, Dobriban E. Evaluating the performance of large language models via debates. In: Findings of the Association for Computational Linguistics: NAACL 2025. Stroudsburg, PA, USA: Association for Computational Linguistics; 2025. p. 2040–75.
23. Joshi I, Shahid S, Venneti SM, Vasu M, Zheng Y, Li Y, et al. CoPrompter: User-centric evaluation of LLM instruction alignment for improved prompt engineering. In: Proceedings of the 30th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM; 2025. p. 341–65.

 doi:10.20944/preprints202604.0054.v1

24. Ramanathan S, Lim L-A, Mottaghi NR, Buckingham Shum S. When the Prompt becomes the Codebook: Grounded Prompt Engineering (GROPROE) and its application to Belonging Analytics. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. New York, NY, USA: ACM; 2025. p. 713–25.
25. Zhao P, Zhang H, Yu Q, Wang Z, Geng Y, Fu F, et al. Retrieval-Augmented Generation for AI-generated content: A survey [Internet]. arXiv [cs.CV]. 2024. Available from: http://arxiv.org/abs/2402.19473
26. Yu T, Jing Y, Zhang X, Jiang W, Wu W, Wang Y, et al. Benchmarking reasoning robustness in large language models [Internet]. arXiv [cs.AI]. 2025. Available from: http://arxiv.org/abs/2503.04550
27. Leon M. GPT-5 and open-weight large language models: Advances in reasoning, transparency, and control. Inf Syst [Internet]. 2026;136(102620):102620. Available from: http://dx.doi.org/10.1016/j.is.2025.102620
28. Singh A. Significance of generative AI in medicine and healthcare [Internet]. 2025. Available from: http://dx.doi.org/10.2139/ssrn.5153375
29. Al-Garadi M, Mungle T, Ahmed A, Sarker A, Miao Z, Matheny ME. Large Language Models in Healthcare [Internet]. arXiv [cs.CY]. 2025. Available from: http://arxiv.org/abs/2503.04748
30. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Nadkarni GN, Klang E. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health [Internet]. 2025;7(1659134). Available from: http://dx.doi.org/10.3389/fdgth.2025.1659134