ClinDoc Copilot: Enhancing Efficiency and Normalization in Chinese Outpatient Settings

Anonymous ACL submission

Abstract

Clinical documentation in Chinese outpatient settings is a legally regulated and time-intensive task, requiring physicians to produce records that are both efficient and auditable. Existing medical large language models demonstrate strong performance on medical benchmarks. However, their generated content cannot be directly adopted in clinical documentation due to legal risk, as responsibility for medical records cannot be delegated to or assumed by autonomous models. We present ClinDoc Copilot, a physician-centered assistant designed to support outpatient documentation. We evaluate ClinDoc Copilot in realistic simulated outpatient scenarios with licensed physicians and clinical trainees. Results show that the system improves documentation efficiency, terminology standardization, and evidence-grounded traceability. Our work highlights the value of interactive, physician-led AI assistance for clinical documentation in regulated healthcare settings. All code is available at an anonymous repository: https://anonymous.4open.science/r/ClinDoc-Copilot/README.md.

1 Introduction

In Chinese outpatient settings, clinical documentation is a strictly regulated professional activity. Outpatient medical records are **legally binding documents** subject to routine audits by the national health insurance reimbursement system. Physicians bear full legal responsibility for the accuracy, completeness, and clinical validity of all documented content. As a result, every diagnosis and treatment order must be explicitly supported by documented evidence and remain traceable throughout the record.

At the same time, outpatient documentation is completed under **severe time constraints**. Physicians are expected to conduct patient interviews, perform clinical reasoning, and finalize medical records within a limited consultation window

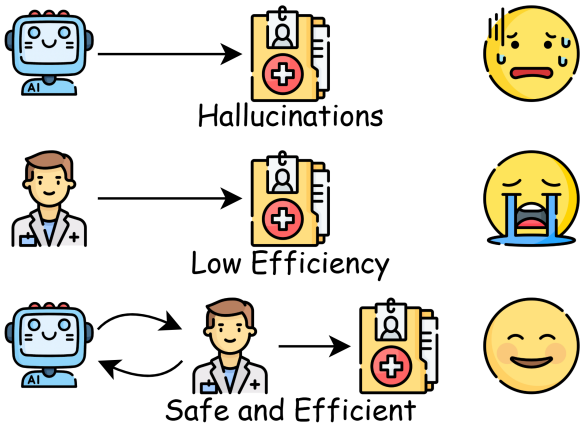

Figure 1: Comparison of Different HCI Modes in Chinese Clinical Documentation.

(Shanafelt et al., 2016). The coexistence of documentation rigor and time pressure places substantial cognitive burden on physicians, particularly during real-time writing.

Motivated by current situations, we introduce **ClinDoc Copilot**, a physician-centered interactive assistant for outpatient clinical documentation. ClinDoc Copilot operates in a human-in-the-loop manner, providing real-time, non-binding assistance during the writing process while ensuring that the physician remain the sole legal owner of medical records. By prioritizing physician control and accountability, the system aims to improve documentation efficiency and writing quality without compromising regulatory requirements.

We evaluate ClinDoc Copilot in realistic simulated outpatient scenarios involving both licensed physicians and clinical trainees. Experimental results show that the system substantially reduces documentation effort while improving terminology standardization and evidence-grounded traceability of diagnoses and treatment orders. A user study further indicates that clinicians perceive ClinDoc Copilot as practical, trustworthy, and well aligned with real outpatient documentation workflows.

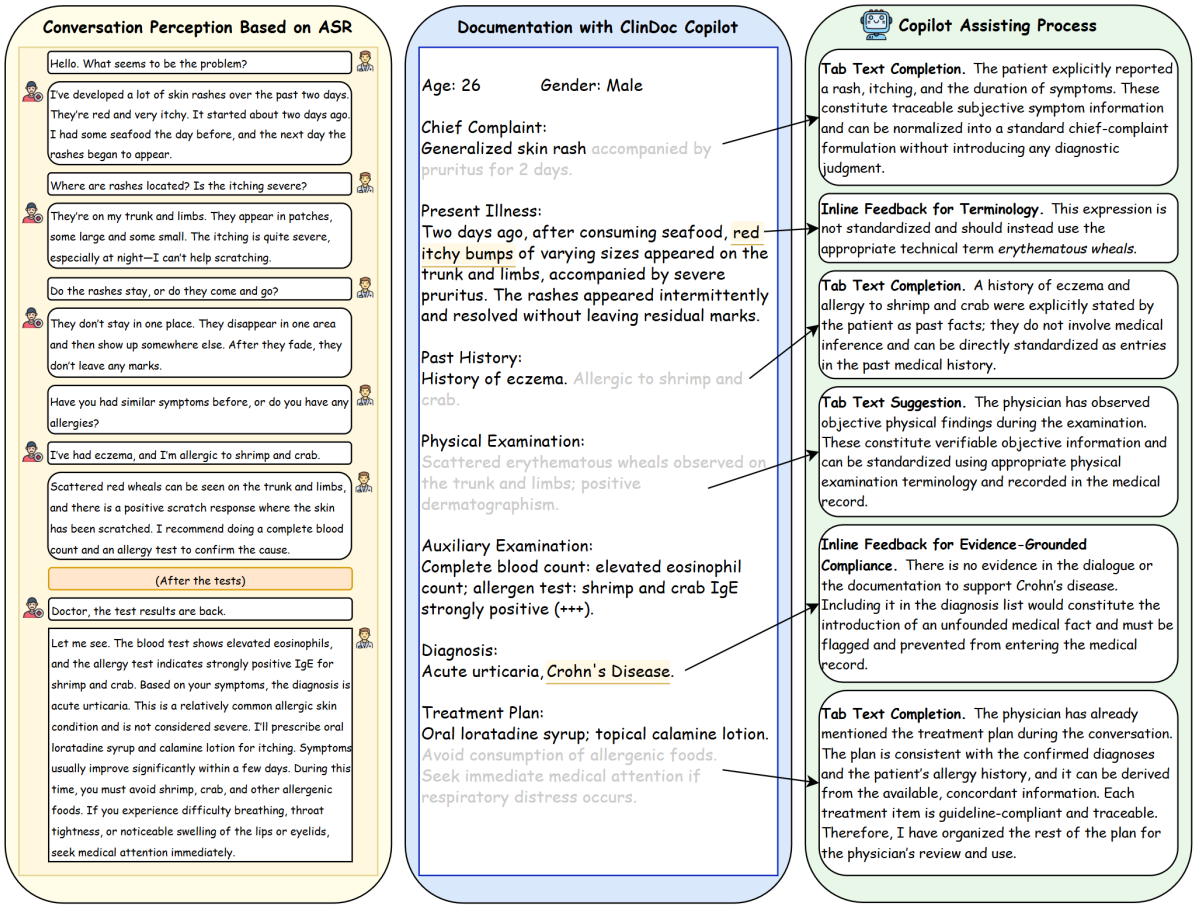

Figure 2: Overview of the ClinDoc Copilot Workflow. Black text represents content entered by the physician, while gray text indicates suggested completions. For readers' convenience, the illustration is in English.

2 Background

This section provides background for the design of ClinDoc Copilot. Section 2.1 discusses why existing medical large language models are unsuitable. Section 2.2 summarizes key user needs in Chinese outpatient documentation. Next, section 2.3 reviews existing documentation assistance systems and their limitations. Finally, section 2.4 outlines principles from responsible and human-centered AI that inform the system design.

2.1 Limitations of Medical LLMs for Clinical Documentation

Recent advances in medical large language models demonstrate near-human or even super-human performance on exam-style benchmarks (Singhal et al., 2025; Kung et al., 2023; Pal et al., 2022; Xie et al., 2025). Such capabilities make LLMs effective for medical education (Gilson et al., 2023), knowledge dissemination (Thirunavukarasu et al., 2023), and general-purpose decision support (Oniani et al., 2024). However, clinical documentation poses fundamentally different requirements. Beyond medical correctness, documentation demands efficiency under time pressure, normalized medical expression, and clear attribution of clinical responsibility. Recent studies highlight that benchmark-driven LLMs remain insufficient for legally accountable documentation, due to risks such as hallucination, reasoning errors, and misalignment with real clinical workflows (Nori et al., 2023; Asgari et al., 2025; Vladika et al., 2025).

2.2 User Needs in Chinese Outpatient Documentation

Interviews. To better understand real-world outpatient documentation challenges, we conducted semi-structured interviews with licensed physicians across both modern Western medicine and traditional Chinese medicine (TCM). Despite differences in clinical experience and medical systems, participants reported a highly consistent set of documentation pain points, with these issues being particularly pronounced among trainees.

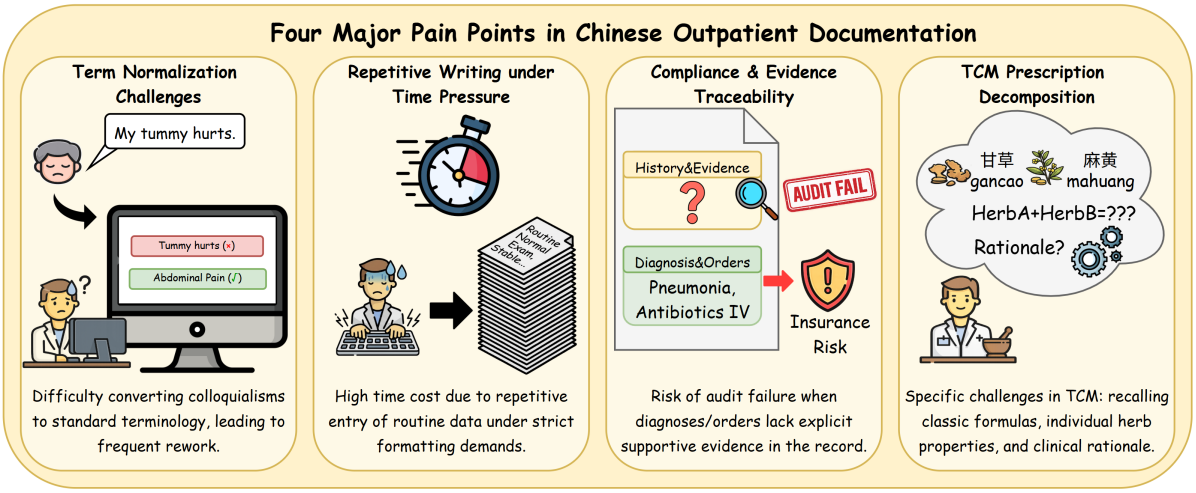

Figure 3: Four major pain points in Chinese Outpatient Documentation

Pain points. Four recurring pain points in Chinese outpatient clinical documentation that hinder efficiency and compliance are identified:

- **Terminology normalization.** Participants frequently reported difficulties in translating patients' colloquial descriptions into standardized medical terminology, resulting in records that require substantial post hoc revision.
- **Repetitive writing under time pressure.** Physicians emphasized the significant time cost associated with repeatedly drafting routine documentation sections while adhering to strict formatting requirements.
- **Compliance and evidence-grounded traceability.** Many trainees struggled to ensure documentation compliance, particularly the traceability requirement of diagnoses and treatment orders. This requirement is closely tied to national health insurance auditing practices.
- **TCM prescription decomposition.** Physicians practicing modern traditional Chinese medicine highlighted domain-specific challenges in prescription writing, including recalling classical formulas, individual herbal functions, and the clinical rationale underlying herb combinations.

Taken together, these pain points reveal two fundamental needs in Chinese outpatient documentation: improving writing efficiency by reducing repetitive manual effort, and enforcing writing norms that ensure terminology accuracy and evidence-grounded traceability.

2.3 Existing Systems

2.3.1 Automated Generation

A substantial body of prior work has framed clinical documentation as an automatic text generation or summarization task. Representative approaches include dialogue-to-SOAP note generation pipelines (Krishna et al., 2021; Ben Abacha et al., 2023). Subsequent studies have explored modular summarization, retrieval augmentation, and large language models for generating summaries and discharge notes from clinical dialogues or transcripts (Giorgi et al., 2023; Xie et al., 2023).

While these systems demonstrate promising performance on benchmarks, they typically emphasize autonomous note generation. Such designs implicitly treat documentation as an output artifact rather than a physician-controlled legal record, making them less suitable for outpatient settings where clinicians remain fully accountable for every documented statement. In regulated environments, silently generated content may introduce unacceptable risks when diagnoses or treatment decisions are insufficiently grounded in explicitly documented evidence.

2.3.2 Commercial Systems

In parallel with academic research, commercial systems based on ambient clinical intelligence have been deployed to support clinical documentation (Healthcare, 2025). Despite demonstrated efficiency benefits, these commercial systems are proprietary, English-centric, and optimized for healthcare systems with documentation norms that differ from Chinese outpatient practice.

2.4 Responsible AI in Healthcare

A growing body of research has emphasized that deploying AI systems in healthcare requires not only high predictive or generative performance, but also careful consideration of responsibility, accountability, and human oversight. Unlike many consumer-facing AI applications, clinical systems operate in safety-critical environments where errors may carry legal, ethical, and clinical consequences, and where responsibility for decisions cannot be delegated to opaque automated systems (Wiens et al., 2019; Sendak et al., 2020).

Prior work in responsible and human-centered medical AI consistently highlights several recurring principles. First, clinical AI systems should preserve clinician agency and decision authority, ensuring that human professionals remain accountable for all medically relevant actions and documentation (Amershi et al., 2019). Second, automation should be carefully scoped to avoid over-reliance, silent failure modes, or the unintentional introduction of unsupported or fabricated information into clinical workflows (Bansal et al., 2021; Nori et al., 2023). Third, AI assistance must be aligned with existing clinical workflows, regulatory constraints, and auditing practices, rather than optimizing solely for technical performance metrics (Sculley et al., 2015).

These considerations are particularly salient for clinical documentation, which serves not only as a communication artifact but also as a legally binding record subject to post hoc review and insurance auditing. In such contexts, even fluent and factually plausible AI-generated text may be problematic if it cannot be clearly traced to documented clinical evidence or explicit physician intent.

Taken together, this line of work motivates a design stance in which AI systems support, rather than replace, clinician-led documentation. For ClinDoc Copilot, these insights translate into concrete system requirements: AI assistance must be non-binding, explicitly invoked, and grounded in the encounter context; any content incorporated into the medical record must result from deliberate physician action; and the system must avoid introducing new clinical facts or decisions beyond those already established by the clinician. By operationalizing these responsible AI principles at the interaction and workflow level, ClinDoc Copilot aims to reconcile efficiency gains with the accountability demands of regulated outpatient clinical practice.

3 System Overview

ClinDoc Copilot is a physician-centered, human-in-the-loop assistant designed to support outpatient clinical documentation under strict regulatory constraints in Chinese clinical settings. Rather than autonomously generating medical records, the system assists physicians during documentation writing by providing optional, real-time support grounded in the evolving conversation and draft.

3.1 Design Goals

Based on interviews with licensed physicians and clinical trainees, we identify two primary design goals for ClinDoc Copilot.

First, the system should **improve documentation efficiency** by reducing repetitive and time-consuming manual writing. In particular, routine documentation sections that follow stable structural patterns should be completed with minimal friction, allowing physicians to focus on clinically meaningful decisions including medical orders.

Second, the system should **support documentation norms required** in regulated clinical practice. This includes helping ensure that medical terminology is standardized and that diagnoses and treatment orders are supported by, and derivable from, information documented earlier in the record. Such evidence-grounded traceability is especially critical in the context of national health insurance auditing. Importantly, both goals must be achieved without weakening physician authority or accountability. ClinDoc Copilot is designed as a supportive tool rather than an autonomous author, with all assistance remaining optional and under explicit physician control.

3.2 Core Components

ClinDoc Copilot consists of three integrated components. Together, they address both efficiency and writing norm requirements while preserving physician authorship and responsibility.

Conversation-Grounded Ghost Text and Tab Completion. ClinDoc Copilot supports fast drafting through ghost text preview and explicit tab-based completion. During outpatient encounters, the system relies on audio recording and transcription to construct a conversational context. This context, together with the surrounding draft text, is used to generate ghost text suggestions that appear as low-contrast inline previews.

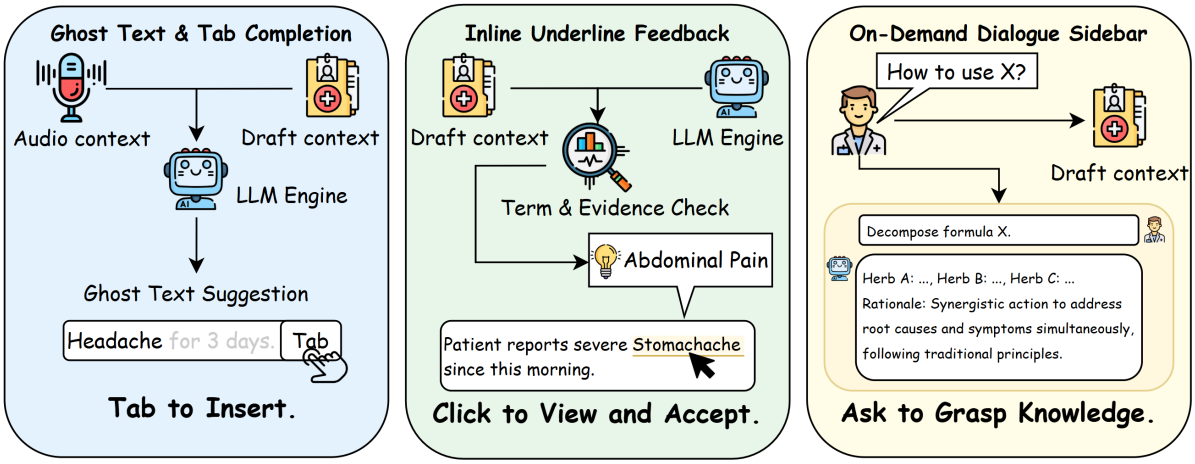

Figure 4: Overview of the ClinDoc Copilot Components.

The ghost text **can adapt dynamically to user input**. And physicians may explicitly accept a suggestion by pressing the tab key, at which point the preview text is inserted and becomes fully editable. If a suggestion is ignored, it disappears without affecting the draft. In this way, conversational grounding enables relevant suggestions, while the ghost-text-and-tab interaction ensures that insertion remains intentional and physician-controlled.

Inline Underline Feedback for Terminology and Evidence-Grounded Compliance. To support writing norms, ClinDoc Copilot provides inline feedback using visual underlines. When colloquial expressions in the physician's draft should be normalized into standard medical terminology, the system highlights the relevant text spans with a yellow underline. Upon clicking for details, ClinDoc Copilot presents localized revision suggestions that illustrate how the highlighted expression can be rewritten using appropriate clinical terminology. These suggestions are scoped to the underlined fragment and do not trigger automatic rewriting, allowing physicians to decide whether and how to apply the proposed changes.

In addition, when diagnoses or treatment orders appear insufficiently supported by information documented earlier in the record, the system similarly applies underline cues to prompt physician review. These inline underline feedbacks are triggered not only by non-standard terminology, but also by insufficient logical support within the documentation. In all cases, the underline serves as a reminder rather than an enforcement mechanism, preserving physician-led correction and decision-making.

On-Demand Right-Side AI Dialogue Sidebar. ClinDoc Copilot includes an AI dialogue sidebar that appears only when the physician explicitly invokes it. This design reflects the clinical requirement that assistance must be physician-initiated and non-intrusive. When invoked, the sidebar supports clarification and just-in-time assistance during documentation. Typical use cases include: asking how to properly phrase a section in compliant clinical language, requesting terminology clarification, or, in modern traditional Chinese medicine practice, querying prescription decomposition such as rationale for herb combinations. Outputs in the sidebar are advisory. They do not automatically insert content into the medical record and must be manually incorporated by the physician if desired.

3.3 Implementation Details

ClinDoc Copilot adopts a modular system architecture that cleanly separates speech understanding, language modeling, and interaction logic. For speech recognition, the system employs SenseVoice (An et al., 2024) to transcribe outpatient doctor–patient conversations in real time. For language understanding and generation, ClinDoc Copilot uses GPT-4o (OpenAI, 2024) as the backend large language model engine.

3.4 Accountability and Control

Preserving clear responsibility boundaries is a central design principle of ClinDoc Copilot. The system does not introduce new clinical facts as committed record content, infer diagnoses, or modify treatment decisions. By relying on explicit user actions for any insertion and by keeping the dialogue interface physician-invoked, ClinDoc Copilot en-

sures that physicians remain the sole authors and legal owners of medical records. This design reduces the risk of uncontrollable hallucination being silently committed into documentation and supports accountable, auditable practices aligned with outpatient regulatory requirements.

4 Evaluation

We evaluate ClinDoc Copilot in realistic simulated outpatient documentation scenarios. The evaluation addresses three research questions: (1) whether ClinDoc Copilot improves documentation efficiency, (2) whether it improves documentation quality and evidence-grounded traceability without introducing hallucination risks, and (3) whether clinicians perceive the system as usable and suitable for real clinical practice.

4.1 Experimental Setup

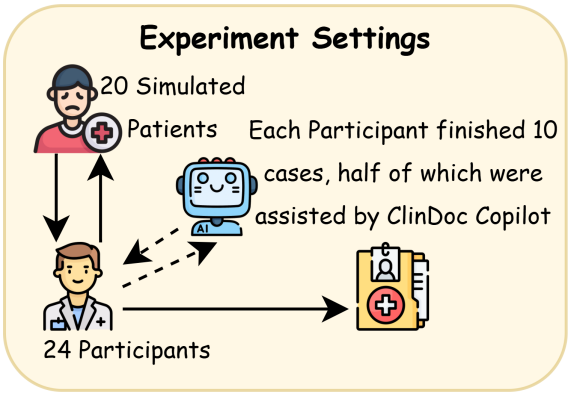

Figure 5: Overview of Evaluation Settings.

Participants. We recruited a total of 24 participants, including 12 licensed outpatient physicians and 12 clinical trainees. All physicians had comparable outpatient experience within their respective clinical systems, and all trainees were at a similar stage of clinical training. Participants covered two independent clinical systems: Western medicine (WM) and traditional Chinese medicine (TCM).

Tasks and Scenarios. The evaluation was conducted using realistic simulated outpatient encounters derived from twenty de-identified clinical cases (zhangmingme, 2025; SylvanL, 2025). All use data are in Chinese. Ten cases were from WM system and another ten were from TCM system. For each case, a trained actor played the role of the patient following a standardized script, ensuring every participant would receive identical information. Each participant completed documentation for ten outpatient encounters within their corresponding clinical system. Because simulated patients cannot fully reproduce all clinical details, a controlled mechanism was adopted for auxiliary examinations. When a physician requested laboratory tests or imaging studies during the encounter, the simulated patient provided the corresponding results if such examinations were present in the original medical record. If the requested examination did not exist in the real case, the physician was explicitly informed that the examination could not be performed.

Conditions. Participants were randomly assigned five cases under **Copilot** condition, in which ClinDoc Copilot was enabled, and five cases under **Baseline** condition, in which participants used the same documentation interface without Copilot assistance. This fully randomized assignment mitigates potential confounding effects arising from fixed case–condition pairing, ensuring that observed differences are not attributable to particular case characteristics. For each participant, the ten documentation tasks were arranged in an interleaved manner, alternating between the Copilot and Baseline conditions rather than being completed in separate blocks. This cross-condition ordering helps reduce fatigue- and learning-related biases that could otherwise skew efficiency and quality measurements. This setting resulted in a balanced $2\times2\times2$ design. Table 1 summarizes the participant groups and task distribution.

Group	Cond.	#Part.	Cases
WM / P.	Cop.	6	30
WM / T.	Cop.	6	30
TCM / P.	Cop.	6	30
TCM / T.	Cop.	6	30
WM / P.	Base.	6	30
WM / T.	Base.	6	30
TCM / P.	Base.	6	30
TCM / T.	Base.	6	30

Table 1: Participant groups and documentation tasks. P. denotes Physician; T. denotes Trainee; Cop. denotes Copilot condition; Base. denotes Baseline condition; Cond. denotes Condition; #Parts. denotes number of participants.

Ethical Considerations. All participants provided informed consent prior to the study. Audio recordings were collected solely for documentation assistance in simulated scenarios and did not involve real patient care. No identifiable patient information was used or stored, and all case materi-

Group	Cond.	Time↓	Keys↓	Edit↓	TN↑	TR↑	SH↑
WM / P.	Cop.	485.50 (68.5)	45.50 (12.5)	0.22 (0.05)	4.75 (0.42)	4.68 (0.45)	4.92 (0.28)
WM / T.	Cop.	530.00 (75.2)	68.00 (15.8)	0.18 (0.06)	4.62 (0.48)	4.55 (0.50)	4.83 (0.38)
TCM / P.	Cop.	550.50 (72.0)	52.50 (14.0)	0.25 (0.07)	4.70 (0.45)	4.65 (0.48)	4.96 (0.20)
TCM / T.	Cop.	610.00 (85.5)	85.00 (18.5)	0.15 (0.05)	4.58 (0.52)	4.42 (0.55)	4.79 (0.41)
WM / P.	Base.	618.75 (58.4)	88.25 (12.5)	–	4.38 (0.49)	4.46 (0.51)	–
WM / T.	Base.	793.33 (75.6)	123.67 (18.2)	–	3.67 (0.61)	3.54 (0.66)	–
TCM / P.	Base.	715.00 (68.2)	110.00 (15.3)	–	4.13 (0.50)	4.21 (0.51)	–
TCM / T.	Base.	920.00 (85.4)	145.50 (20.1)	–	3.21 (0.59)	3.13 (0.61)	–
Overall Effect Size	d	1.94	2.23	–	1.35	1.12	–
Significance	p	<0.001	<0.001	–	<0.001	<0.001	–

Table 2: Efficiency metrics averaged per medical record and rubric-based documentation quality and safety scores (1–5).Clinical cases were randomly assigned to experimental conditions using a counterbalanced design.

als were derived from de-identified clinical records. All collected data were used exclusively for research purposes and handled in accordance with institutional data protection practices.

4.2 Evaluation Metrics

Efficiency Metrics. Documentation efficiency was evaluated using the following metrics, all reported as averages per medical record:

- **Time** (↓): average seconds needed to complete one outpatient record.
- **Keys** (↓): average number of keyboard inputs per record.
- **Edit** (↓): average ghost text edit rate, defined as the proportion of tab-accepted ghost text suggestions that were subsequently modified.

Quality and Safety Metrics. Documentation quality was assessed using three rubric-based metrics:

- **TN** (↑): average Terminology Normalization.
- **TR** (↑): average Evidence-Grounded Traceability of diagnoses and treatment orders.
- **SH** (↑): average Safety and Hallucination level, with higher scores indicating fewer hallucination issues.

All quality metrics were scored on a five-point scale by two senior physicians in a double-blind manner. Detailed rubric definitions are provided in Appendix B.

4.3 Statistical Analysis and Overall Effects

To assess the overall impact of ClinDoc Copilot on documentation efficiency and record quality,

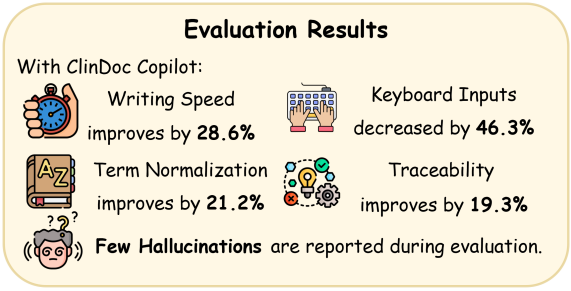

Figure 6: Overview of Evaluation Results.

we conducted a paired-samples t-test on the aggregated per-record metrics across experimental conditions. This analysis compares Copilot-assisted documentation against the baseline condition while controlling for case content via a counterbalanced assignment.

As summarized in the bottom rows of Table 2, ClinDoc Copilot yields statistically significant improvements across all comparative metrics. Beyond statistical significance, we further report Cohen's d to quantify the magnitude of these effects.

Efficiency gains are particularly pronounced. Documentation time exhibits a very large effect size, and keystroke count shows an even stronger effect. These results indicate not merely incremental speedups, but a fundamental shift in documentation workflow, where repetitive manual writing is substantially reduced under Copilot assistance.

In addition to efficiency, documentation quality also improves consistently. Terminology normalization demonstrates a large effect, suggesting that Copilot effectively supports standardized clinical language use. Reasoning traceability similarly shows a substantial effect, indicating improved alignment between documented diagnoses or orders and the supporting clinical evidence recorded earlier in the note.

Meanwhile, the ghost text edit rate in the Copi-

lot condition remained low across groups, indicating that most tab-accepted completions aligned well with physician intent and required minimal post-editing. Importantly, hallucination-related issues are rare under the Copilot condition. This is because all system assistance required explicit physician action and was grounded in the recorded encounter context.

Taken together, these results form a complete evidentiary chain combining descriptive statistics, effect sizes, and significance testing. They demonstrate that ClinDoc Copilot delivers robust efficiency gains while simultaneously enhancing documentation quality.

4.4 User Study

After completing the documentation tasks, all participants who used ClinDoc Copilot completed a post-study questionnaire and a brief semi-structured interview. The questionnaire assessed participants' subjective perceptions along four dimensions: (1) *efficiency improvement* achieved through text auto-completion and AI dialogue (EFF), (2) *documentation quality and standardization*, particularly terminology normalization (QUAL), (3) *trustworthiness of system assistance* grounded in the encounter context (TRUST), and (4) *willingness to adopt the system* in real outpatient practice (ADOPT). All items were rated on a five-point Likert scale with 1 = Strongly Disagree, 5 = Strongly Agree. The full questionnaire is provided in Appendix C.

Table 3 summarizes the questionnaire results, reported as mean Likert scores across participants. Overall, participants gave consistently positive ratings for efficiency improvement as well as documentation quality and standardization, suggesting that ClinDoc Copilot is perceived as practically useful in outpatient documentation. Trust-related ratings were comparatively lower but remained in the positive range, indicating a generally cautious yet favorable attitude toward system assistance during clinical use.

Item	Score (1–5 ↑)
EFF	4.50 (0.50)
QUAL	4.58 (0.51)
TRUST	4.00 (0.59)
ADOPT	4.13 (0.54)

Table 3: User study questionnaire results. Scores denote mean Likert ratings across participants (1 = Strongly Disagree, 5 = Strongly Agree).

5 Discussion

Our results suggest that ClinDoc Copilot provides practical value by aligning AI assistance with the legal and workflow constraints of outpatient clinical documentation. Rather than maximizing automation, the system focuses on supporting physicians while preserving clear responsibility boundaries.

ClinDoc Copilot improves efficiency primarily by reducing repetitive manual writing, rather than by replacing clinical reasoning. At the same time, inline cues help surface terminology and traceability issues that are often overlooked under time pressure. Together, these mechanisms enable faster documentation without sacrificing audit readiness.

More broadly, our findings have implications for the deployment of AI-assisted documentation tools beyond the specific system studied here. As healthcare institutions increasingly explore AI integration, there is a risk that efficiency metrics alone may drive adoption decisions. Our results caution against equating automation with usefulness, particularly in environments where legal accountability and audit readiness are non-negotiable. Interactive, physician-led systems like ClinDoc Copilot may offer a more sustainable path for deployment, especially in early stages of adoption. Over time, this alignment may prove more important than raw generative capability in determining long-term impact.

6 Conclusion

We presented ClinDoc Copilot, a physician-centered, human-in-the-loop assistant for outpatient clinical documentation in Chinese clinical settings. By combining conversation-grounded ghost text completion, inline compliance cues, and an on-demand AI dialogue sidebar, ClinDoc Copilot improves documentation efficiency while supporting terminology standardization and evidence-grounded traceability.

Through a controlled evaluation with licensed physicians and clinical trainees, we showed that ClinDoc Copilot reduces documentation effort, improves documentation quality, and is perceived by clinicians as practical and trustworthy. Our results highlight the importance of interactive, physician-led AI assistance for clinical documentation, particularly in settings where legal accountability and audit requirements are central. We believe ClinDoc Copilot represents a promising direction for deploying AI systems that meaningfully support, rather than replace, clinicians in real-world practice.

Limitations

This study has several limitations. First, our evaluation was conducted in simulated outpatient scenarios using trained actors rather than in live clinical environments. While this design allowed for controlled comparison across conditions, real-world deployment may introduce additional variability, such as institution-specific workflows.

Second, the participant pool, while balanced across roles and clinical systems, was limited in size. Larger-scale studies across more specialties and hospitals are needed to assess generalizability.

Third, ghost text relies on audio recording and transcription of outpatient encounters. Although recordings were collected with informed consent in this study, privacy considerations and institutional policies may affect adoption in some settings.

Ethical Considerations

Human Subjects and Informed Consent. All evaluations in this study were conducted in simulated outpatient scenarios using trained actors rather than real patients. Licensed physicians and clinical trainees participated voluntarily and provided informed consent prior to the study. Audio recordings were collected solely for the purpose of supporting documentation assistance within the simulated tasks and did not involve real patient care or identifiable patient data.

Patient Safety and Clinical Responsibility. ClinDoc Copilot is not designed to diagnose, recommend treatments, or make clinical decisions. All generated suggestions are non-binding and require explicit physician action to be incorporated into the medical record. Physicians remain the sole authors and legal owners of all documentation. This design minimizes the risk of unverified or hallucinated content being silently introduced into clinical records and preserves clear responsibility boundaries consistent with regulated outpatient practice.

Data Privacy and Confidentiality. The study used de-identified clinical case scripts for simulation. No real patient records were accessed or stored. Audio recordings and written records were handled in accordance with institutional data protection practices and were used exclusively for research evaluation. The system design assumes deployment within secure hospital IT environments, where access control, logging, and data governance policies are enforced by the institution. In this study, all data, code, and models we use and we release are under the CC-BY 4.0 license or Apache license 2.0. We have verified that their usage complies with the original license agreements and access conditions. Furthermore, participants involved in the human evaluation phase were explicitly informed that their evaluation results would be used solely for academic purposes. A total of 46 annotators from one country participated in this study, and each annotator was equipped with at least basic medical knowledge. Our study complies with ethical standards and regulations.

Risk of Misuse and Deployment Considerations. While AI-assisted documentation tools may improve efficiency, improper deployment could risk over-reliance or inappropriate automation. ClinDoc Copilot mitigates these risks by requiring explicit user invocation for all assistance and by avoiding end-to-end autonomous note generation. We emphasize that the system is intended as a supportive writing aid rather than a substitute for clinical judgment, and should be deployed only in settings that maintain physician oversight and regulatory compliance.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *Preprint*, arXiv:2407.04051.

Elham Asgari, Nina Montaña Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings*

of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA. Association for Computing Machinery.

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, and 1 others. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. WangLab at MEDIQA-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334, Toronto, Canada. Association for Computational Linguistics.

Microsoft Healthcare. 2025. Microsoft dragon copilot. AI clinical assistant to streamline documentation and workflows.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. 2024. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 694–702. IEEE.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "the human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 99–109, New York, NY, USA. Association for Computing Machinery.

Tait D. Shanafelt, Lotte N. Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P. West. 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings*, 91(7):836–848.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

SylvanL. 2025. Traditional chinese medicine dataset pretrain. Dataset for text pretraining in Traditional Chinese Medicine.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Juraj Vladika, Annika Domres, Mai Nguyen, Rebecca Moser, Jana Nano, Felix Busch, Lisa C. Adams, Keno K. Bressem, Denise Bernhardt, Stephanie E. Combs, Kai J. Borm, Florian Matthes, and Jan C. Peeken. 2025. Improving reliability and explainability of medical question answering through atomic fact checking in retrieval-augmented llms. *Preprint*, arXiv:2505.24830.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. 2025. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine*, 8(1):141.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *Preprint*, arXiv:2304.08763.

zhangmingme. 2025. Medical record diagnosis chinese. Apache License 2.0.

A Human Subjects Information

A.1 User Interface for Human Subjects

As show in Fig. 7, the interface supports structured documentation during the initial patient encounter and consists of two components. The medical record editing interface provides free-text fields for patient demographics, chief complaint, history of present illness, past medical history, physical examination, laboratory and imaging studies, diagnosis, and physician orders. The assistant interface includes a speech transcription assistant and a conversational question-answering assistant. With authorization, the speech assistant transcribes physician–patient conversations to support medical record generation. The conversational assistant enables physicians to interactively query and clarify medical record–related annotations during documentation.

A.2 Instructions Provided to Participants

Our study involved two categories of participants. To approximate real-world clinical consultations and medical documentation workflows, medical students were instructed to act as simulated patients and interact with licensed physicians during system evaluation. Accordingly, we describe the instructions provided to participants separately by participant category.

A.2.1 Instructions for Medical Student Participants Acting as Simulated Patients

Role and Study Purpose. You are invited to participate in a research study evaluating *ClinDoc Copilot*, an AI-assisted clinical documentation system. In this study, you will act as a **simulated patient** and interact with a licensed physician during a simulated outpatient consultation. The purpose of your participation is to approximate realistic doctor–patient interactions so that we can study clinical documentation workflows. You are not being evaluated on medical knowledge, acting performance, or communication skills.

Case Materials and Interaction Guidelines. You will be provided with a predefined clinical case script prior to each session. These case materials are derived from de-identified clinical scenarios and do not contain any real patient information. During the consultation, you should respond strictly according to the provided script and should not introduce additional symptoms, medical history, or test results beyond what is specified. If the physician requests laboratory tests, imaging studies, or examinations, you should provide results only if they are included in the case materials; otherwise, state that the requested information is unavailable. You should communicate naturally but avoid improvisation that deviates from the script.

Audio Recording and Data Collection. The consultation will be audio recorded to support transcription and documentation assistance. Audio recordings are collected solely for research purposes, including evaluation of documentation workflows and expert assessment of documentation quality. No video will be recorded. You will not be asked to provide any personal identifying information, including your real name, student ID, contact details, or personal medical history.

Data Use, Privacy, and Consent. All collected data will be de-identified and used exclusively for academic research and system evaluation. De-identified data may be released publicly for research purposes under a CC BY 4.0 license. Participation is entirely voluntary, and you may withdraw from the study at any time without penalty. By signing the consent form, you confirm that you understand the study purpose, consent to audio recording under the described conditions, and agree to the use of de-identified data for research and publication.

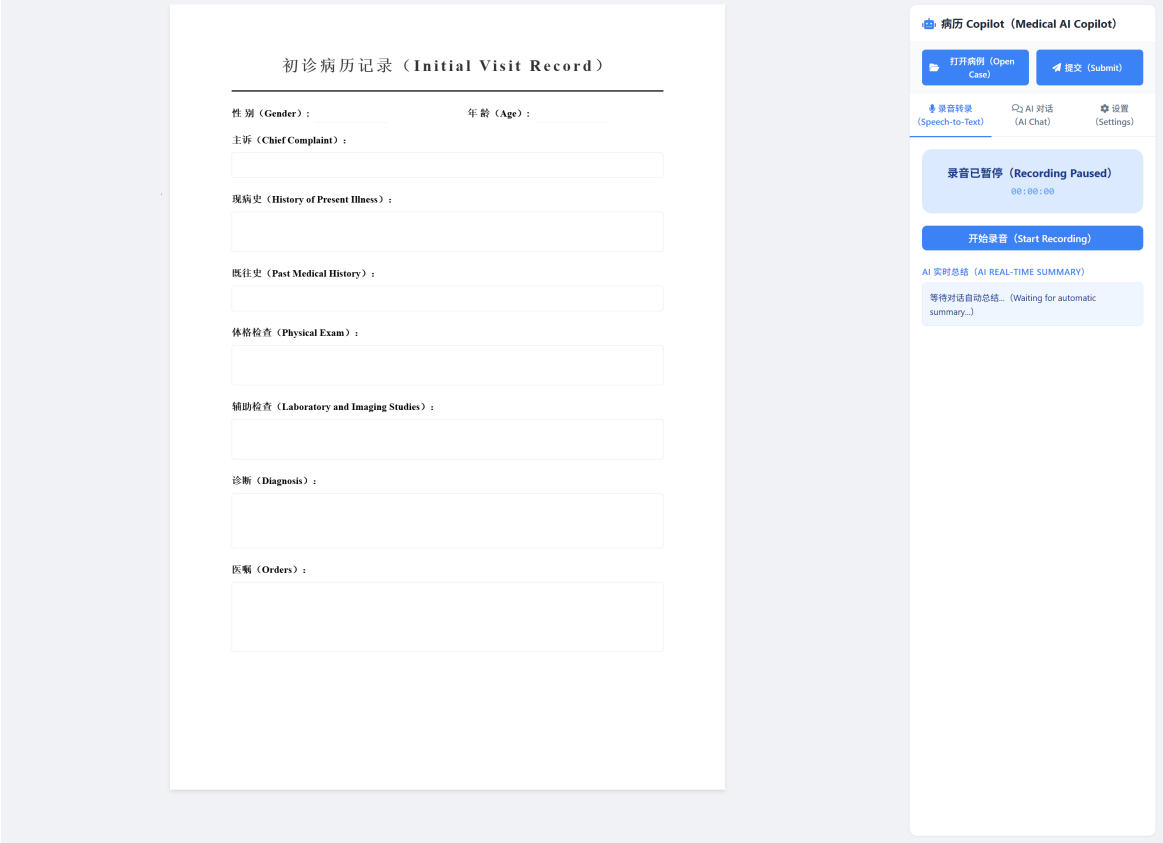

Figure 7: UI of ClinDoc Copilot

Important Notes. This study does not involve real patient care. No clinical decisions are made based on your responses, and the AI system does not diagnose or treat patients.

A.2.2 Instructions for Licensed Physicians and Clinical Trainees

Study Purpose and Your Role. You are invited to participate in a research study evaluating *ClinDoc Copilot*, a physician-centered AI assistant designed to support outpatient clinical documentation. You will conduct simulated outpatient consultations with medical students acting as simulated patients and complete the corresponding medical records using the provided documentation interface. Some cases will be completed with AI assistance enabled, while others will be completed without assistance. The study focuses on documentation behavior rather than diagnostic accuracy or clinical performance.

Use of ClinDoc Copilot. When enabled, ClinDoc Copilot may provide optional ghost text suggestions, inline cues highlighting terminology or traceability issues, and an on-demand conversational assistant. The system does not diagnose, recommend treatments, or make clinical decisions. All AI-generated suggestions are non-binding and require explicit physician action to be incorporated into the medical record. You remain the sole author and legal owner of all documentation and may ignore or modify AI assistance at any time.

Documentation and Interaction Guidelines. You should document each encounter as you would in routine outpatient practice, based solely on information stated during the consultation and results explicitly provided in the case materials. The AI system should not be relied upon to introduce new clinical facts or unsupported information.

Audio Recording and Data Handling. Doctor–patient conversations will be audio recorded to support transcription and research evaluation. No real patient data is involved, and no personal identifying information beyond role and training level will be collected. All data are de-identified and handled in accordance with institutional data protection practices.

Voluntary Participation and Informed Consent. Participation is voluntary, and you may withdraw from the study at any time without consequence.

By signing the consent form, you acknowledge that you understand the study procedures, consent to audio recording under the stated conditions, and agree to the use of de-identified data for research and academic publication.

Clarifications. ClinDoc Copilot is a documentation support tool rather than a clinical decision system. The study does not evaluate your medical competence, and the simulated setting is designed to approximate real workflows while ensuring safety and regulatory compliance.

A.3 Recruitment and Compensation

A.3.1 Compensation

For professional doctor participants, we based our compensation on the average annual salary of physicians in the region where the study was conducted, which is approximately 490,000 RMB. Assuming a standard workload of 2,000 working hours per year, this corresponds to an average hourly wage of about 245 RMB. Given that participation in our study required sustained concentration, we provided compensation at more than twice the average hourly rate, amounting to 800 RMB per hour.

For student participants, compensation was determined in accordance with common stipend practices for graduate students in the region where the study was conducted. Medical student participants received a stipend of 3,000 RMB, disbursed on a monthly basis. Although the experimental period did not exceed one month, the full monthly stipend was provided in consideration of the workload involved.

Overall, participant compensation in this study exceeded typical income benchmarks in the local context and was designed to be reasonable and commensurate with the effort, attention, and expertise required, without constituting undue inducement.

A.3.2 Recruitment

Recruitment was conducted primarily through open calls issued by physician participants involved in this study within their respective institutions. In addition, several physician participants were affiliated with teaching hospitals and medical schools; these individuals publicly recruited student participants through their associated medical schools. All recruitment processes were conducted in an open and transparent manner.

A.4 Data Use Authorization and Informed Consent Protection

During public recruitment, it was clearly communicated that code collected in this study would be released under the CC BY 4.0 license. Upon enrollment, all participants signed hard-copy consent forms. Through this process, participants gave informed consent to the collection and public release of their data for research and academic purposes.

B Rubric Definitions for Documentation Quality Evaluation

This appendix details the rubric-based evaluation criteria used to assess documentation quality in our study. All rubric scores were assigned by senior physicians based on the final written medical records, with reference to the recorded outpatient conversations. The evaluation focuses on documentation quality rather than medical decision correctness. Two senior medical experts from medical institute are hired to judge all the documentation with payment of 1000 RMB per hour.

B.1 Term Normalization

Core consideration. This dimension evaluates the system's ability to assist physicians in converting colloquial, patient-facing language into standardized clinical terminology. The assessment reflects how effectively informal expressions in the encounter are transformed into appropriate written medical language using the system's internal language knowledge.

Scoring guidelines (1–5).

- **5 (Excellent):** All colloquial expressions are consistently normalized into standard medical terminology. The record is written entirely in professional clinical language without unnecessary patient phrasing.
- **4 (Good):** Most colloquial expressions are properly normalized, with only minor residual informal phrasing that does not affect readability or professionalism.
- **3 (Acceptable):** Key medical terms are correct, but multiple patient-style expressions remain in the text and would require manual revision for formal documentation.
- **2 (Poor):** Colloquial language is frequently retained, or medical terminology is used inconsistently or imprecisely.

- **1 (Very Poor):** The documentation largely resembles a conversational transcript and fails to meet basic standards of clinical written language.

B.2 Reasoning and Consistency

Core consideration. This dimension assesses whether diagnoses and treatment orders documented in the record are sufficiently supported by information stated earlier in the documentation and in the recorded conversation. The focus is on internal consistency and traceability, rather than on reconstructing explicit reasoning chains or verifying clinical correctness.

Scoring guidelines (1–5).

- **5 (Excellent):** All diagnoses and treatment decisions are clearly supported by documented symptoms, physical findings, or examination results stated earlier in the record or the encounter.
- **4 (Good):** Diagnostic and treatment decisions are generally well supported, with minor omissions or vague references that do not undermine overall traceability.
- **3 (Acceptable):** Diagnoses and treatments are plausible but lack explicit supporting evidence in the documentation, potentially raising concerns during audit or review.
- **2 (Poor):** One or more diagnoses or treatment orders lack clear supporting information or appear weakly connected to documented findings.
- **1 (Very Poor):** Diagnoses or treatment decisions are inconsistent with the documented encounter or contradict earlier information in the record.

B.3 Safety and Hallucination

Core consideration. This dimension evaluates whether the documented content faithfully reflects the recorded outpatient conversation. In the absence of external retrieval constraints, the assessment focuses on whether unsupported or fabricated information appears in the record without being explicitly grounded in the encounter.

Scoring guidelines (1–5).

- **5 (Excellent):** All documented information is fully supported by the recorded conversation or by explicit physician input. No fabricated or unsupported content is present.
- **4 (Good):** Minor inferential phrasing is present, but no clinically significant information is introduced beyond what is supported by the encounter.
- **3 (Acceptable):** The record contains minor unsupported details that do not affect clinical interpretation but would require careful review.
- **2 (Poor):** Clinically relevant information appears that is not supported by the recorded conversation.
- **1 (Very Poor):** The record includes fabricated or contradictory information that could lead to incorrect clinical interpretation.

C User Study Questionnaire

This appendix presents the questionnaire used in the post-study user evaluation of ClinDoc Copilot. The questionnaire was administered after participants completed all documentation tasks. All items were rated on a five-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree*. The questions were designed to assess perceived efficiency improvement, documentation quality, trustworthiness of system assistance, and willingness to adopt the system in real clinical practice.

C.1 Questionnaire Items

1. **Perceived Efficiency Improvement.** "The ClinDoc Copilot system effectively improves efficiency by significantly reducing manual operations through text auto-completion and AI dialogue functions, thereby shortening the overall time required for medical documentation."
2. **Perceived Documentation Quality and Standardization.** "The ClinDoc Copilot system effectively ensures documentation quality by generating professional text that adheres to clinical standards, while the underline function automatically identifies and helps correct colloquial or non-standard medical terminology."
3. **Trustworthiness and Clinical Rationality.** "Based on the real-time context of the doctor–patient dialogue, the medical text suggestions,

underline correction hints, and AI explanations generated by ClinDoc Copilot are trustworthy and demonstrate clinical logical rationality."

4. **Willingness for Real-World Adoption.** "Given ClinDoc Copilot's advantages in improving documentation efficiency and ensuring standardization, I would be willing to use it as a regular auxiliary tool in my daily outpatient practice if deployed in the hospital."

C.2 Response Scale

All questions were answered using the following five-point Likert scale:

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

The questionnaire focuses on subjective user perception rather than objective performance and complements the quantitative efficiency and expert-based quality evaluations reported in the main paper.

D Use of AI Assistants

AI assistants were employed in a strictly supportive capacity during the manuscript preparation process. Their use was limited to assisting with English language editing, including improvements to grammar, wording, and overall clarity in selected sections of the text.

All substantive aspects of the work—including the formulation of research questions, methodological development, experimental design, implementation, evaluation, and interpretation of results—were entirely carried out by the authors. AI assistants did not contribute to analysis, result interpretation, or any form of scientific decision-making.