Are LLMs reliable? An exploration of the reliability of large language models in clinical note generation

Kristine Ann M. Carandang, Jasper Meynard P. Araña,
Ethan Robert A. Casin, Christopher P. Monterola,
Daniel Stanley Y. Tan, Jesus Felix B. Valenzuela, Christian M. Alis
Analytics, Computing & Complex Systems Laboratory
Asian Institute of Management

Abstract

Due to the legal and ethical responsibilities of healthcare providers (HCPs) for accurate documentation and protection of patient data privacy, the natural variability in the responses of large language models (LLMs) presents challenges for incorporating clinical note generation (CNG) systems, driven by LLMs, into real-world clinical processes. The complexity is further amplified by the detailed nature of texts in CNG. To enhance the confidence of HCPs in tools powered by LLMs, this study evaluates the reliability of 12 open-weight and proprietary LLMs from Anthropic, Meta, Mistral, and OpenAI in CNG in terms of their ability to generate notes that are string equivalent (consistency rate), have the same meaning (semantic consistency) and are correct (semantic similarity), across several iterations using the same prompt. The results show that (1) LLMs from all model families are stable, such that their responses are semantically consistent despite being written in various ways, and (2) most of the LLMs generated notes close to the corresponding notes made by experts. Overall, Meta's Llama 70B was the most reliable, followed by Mistral's Small model. With these findings, we recommend the local deployment of these relatively smaller open-weight models for CNG to ensure compliance with data privacy regulations, as well as to improve the efficiency of HCPs in clinical documentation.

1 Introduction

The capability of LLMs to produce text similar to human writing has led to research on their potential role in aiding clinical documentation. This led to the development of clinical note generation (CNG) tools designed to address extended working hours and healthcare provider (HCP) fatigue (Balloch et al., 2024; Biswas and Talukdar, 2024; Giorgi et al., 2023; Heilmeyer et al., 2024; Moramarco et al., 2022; Tung et al., 2024), issues which have persisted despite the adoption of electronic health records (Wu et al., 2024; Zhang et al., 2022; Ghatnekar et al., 2021; Maas et al., 2020; Momenipour and Pennathur, 2019; Quiroz et al., 2019). Considering the legal and ethical responsibility of HCPs to write accurate clinical documentation (McCoy et al., 2024), the reliability of these tools is critical.

LLM reliability is typically assessed using *inter-prompt stability* which checks the consistency of responses when subjected to a variety of prompts designed to elicit the same response (Azimi et al., 2025; Cheng et al., 2024; Dentella et al., 2023; Kozaily et al., 2024; Li et al., 2024; Luo et al., 2024; Wang et al., 2024c). An alternative is to evaluate *intra-prompt stability* by checking the consistency of responses in several iterations using the same prompt (Atil et al., 2024; Barrie et al., 2024; Dentella et al., 2023; Savage et al., 2024; Saxena et al., 2024; Yim et al., 2024; Zhao et al., 2024). However, assessing LLM reliability is more challenging for natural language generation tasks, especially long-form text generation such as CNG. Evaluation typically requires reference texts so that comparisons can be made using automatic evaluation metrics, and involves human evaluation as it remains the gold standard (Giorgi et al., 2023; Moramarco et al., 2022).

Although there exist studies that evaluated LLM performance in CNG from transcripts of provider-patient conversations (Balloch et al., 2024; Chen and Hirschberg, 2024; Giorgi et al., 2023), only Kernberg et al. (2024) evaluated LLM reliability in CNG in terms of intra-prompt stability. While Kernberg et al. (2024) evaluated only one proprietary LLM, their findings show the variability in LLM responses. This may raise concerns on reliability if integrated in the clinical setting (Kernberg et al., 2024), similar with incorporating other healthcare tools developed using LLMs or artificial intelligence in general (Tucci et al., 2021; Wang et al., 2024b).

Additionally, no study exploring the reliability

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 1413–1422

of open-weight LLMs in CNG was found. Using open-weight LLMs over proprietary ones is a typical consideration for healthcare applications due to data privacy concerns related to protected and sensitive health information (Giorgi et al., 2023; Heilmeyer et al., 2024; Wang et al., 2024a).

In this study, we sought to determine whether LLMs are reliable in CNG by evaluating how consistent and correct their generated notes are when using the same prompt in multiple iterations. We focus our evaluation on the CNG task of producing a clinical note based on a transcript of a conversation between a healthcare provider and a patient using an LLM. Four (4) proprietary models and eight (8) open-weight models from Anthropic, Meta, Mistral, and OpenAI were evaluated. This is done with the intention of providing evidence to HCPs on the reliability of LLMs. By doing so, we aim to enhance the body of knowledge regarding the design of reliable tools, crucial for industries such as healthcare, which would benefit from incorporation into real clinical workflows.

More concretely, our findings contribute to our continuous efforts to validate and improve the LLM-powered CNG component of SINTA (*Scalable Intelligent Note-taking and Teaching-learning Assistant*), a system that we have developed to alleviate the workload of HCPs. Supported by an innovation grant from a government-run tertiary training hospital, we are currently evaluating our system with the goal of integrating it into their clinical workflows.

2 Related Work

With the ability of LLMs to generate texts, recent work explored the performance of various ChatGPT models (i.e., ChatGPT 3.5 Turbo, ChatGPT 4) in CNG from transcripts of provider-patient conversations to assess the potential of using an LLM in ambient clinical documentation (Balloch et al., 2024) or to compare the performance of fine-tuned pretrained encoder-decoder or decoder-only language models with at least an LLM (Chen and Hirschberg, 2024; Giorgi et al., 2023). Kernberg et al. (2024) assessed not only the correctness of notes generated from ChatGPT 4, but also the reliability of its responses through three repeated runs for each input, although they did not alter model parameters to make the model more deterministic. In addition, they used standardized assessments rated by human experts to evaluate the quality of responses, without using automatic evaluation metrics. Aside from ChatGPT, we evaluate various open-weight and proprietary LLMs in generating clinical notes from provider-patient dialogues.

Some studies (Atil et al., 2024; Savage et al., 2024; Yim et al., 2024) that evaluated LLM reliability also set model parameters that influence the determinism of LLMs, $temperature$, top_p and top_k, to a value of or close to 0 to make the model deterministic. We also set the model parameters to make them more deterministic.

Assessing LLM performance in CNG usually requires reference notes against which LLM outputs are matched via automated evaluation metrics that assess string overlap or semantic similarity. These evaluations are frequently supplemented by human judgment (Giorgi et al., 2023; Moramarco et al., 2022). Giorgi et al. (2023); Moramarco et al. (2022) found BERTScore (Zhang et al., 2020), an automatic evaluation metric that checks the similarity of two texts in the embedding space, to be the most appropriate embedding-based metric for the task of CNG. We use BERTScore to measure semantic consistency across responses per prompt and semantic similarity of the responses with the notes generated by experts.

In addition to semantic consistency and semantic similarity, we also measure consistency rate to reflect how much of its responses are string equivalent. Consistency of responses was usually measured by considering string equivalence (*total agreement rate for raw model response* (Atil et al., 2024)) or by semantic equivalence (*consistency rate* (Zhao et al., 2024) or *sample consistency* (Savage et al., 2024)) in reliability evaluation studies. String equivalence was noted as a strict measure of reliability while evaluating whether responses contextually mean the same is specifically important in CNG due to stylistic differences of HCPs in documenting their sessions (Moramarco et al., 2022).

3 Method

We evaluate the reliability of LLMs according to their *intra-prompt stability* and their *correctness* following the process illustrated in Figure 1. Each transcript was incorporated into a user prompt template which instructs the LLM to generate a clinical note, with specified headings, from the transcript, for k iterations. Evaluation was then done by using automatic evaluation metrics to determine consis-

tency rate (CR) and semantic consistency (SC) as measures of intra-prompt stability, and correctness through its semantic similarity (SS).

3.1 Dataset

We use **aci-bench** (Yim et al., 2023), a benchmark dataset for automatic visit note generation, licensed under the Creative Commons Attribution 4.0 International Licence (CC BY). We select the aci subset comprising 112 transcripts of natural conversations in English between a patient and a doctor taken during a role-play of a session, as this reflected the real-world scenario the most. Each data point contains the consultation session ID, the corrected transcript of the dialogue (*transcript*), and the corresponding clinical note (*ground truth note*). Figure 5 in Appendix A shows an example of said transcript and ground truth note.

3.2 Clinical Note Generation

Clinical notes were generated based on said *transcripts* using the same prompt in multiple iterations using several LLMs configured to maximize determinism.

3.2.1 User Prompt

A user prompt template was used across all iterations to have a consistent format for the input. This template (Figure B), contains (1) the task, (2) the list of note headings present in the dataset, (3) the *transcript*, and (4) other specific instructions. When using Llama models, modifications to this user prompt had to be made to align with the required format (see Appendix C).

3.2.2 Models & their Configurations

Various versions of open-weight LLMs (i.e, models from Meta and Mistral) and proprietary LLMs (i.e., models from Anthropic and OpenAI) were explored (Table 1). These were accessed using AWS Bedrock API requests through the AWS SDK for Python (Boto3), except for OpenAI's models, which required the use of its API from its platform.

At the minimum, for each model family, the smallest and largest models that took in multilingual text as input were included. Smaller models generally cost less than larger ones. For open-weight LLMs, smaller models also require less compute and storage resources than larger ones when deployed locally. Local deployment is an important option for CNG as this involves processing sensitive personal information which must be kept confidential in accordance with data privacy laws (Giorgi et al., 2023; Heilmeyer et al., 2024; Wang et al., 2024a). However, larger models were still considered, as they were generally reported to perform better in a variety of tasks than smaller models.

Developer	Model	Model Configurations			
		max output tokens	temperature	top_p	top_k
Anthropic	Claude 3.5 Haiku	8192	0	0	1
	Claude 3.5 Sonnet v2	8192	0	0	1
Meta	Llama 3.1-8B-Instruct	2048	0	0	N/A
	Llama 3.1-70B-Instruct	2048	0	0	N/A
	Llama 3.1-450B-Instruct	8192	0	0	N/A
	Llama 3.2-1B-Instruct	8192	0	0	N/A
	Llama 3.2-3B-Instruct	8192	0	0	N/A
Mistral	Large-2407 123B	8192	0	0	N/A
	Small-2402 22B	8192	0	0	1
	Mixtral-8x7B-Instruct	4096	0	0	1
OpenAI	ChatGPT-4o	8192	0	0	N/A
	ChatGPT-40-mini	8192	0	0	N/A

Table 1: **Models used and their parameters**. At least two LLM versions per developer was selected for use in this study - their smallest and their largest models. Maximum output tokens was set to 8192, unless otherwise specified due to model limitation. Other parameters were set accordingly to maximize determinism.

For Meta's Llama 3.1 models, its 70B model was also considered in this study, as its largest model (405B) may be impractical to deploy in low resource settings. Llama 3.2 1B and 3B models were also included as they can be run locally on edge devices, which could more conveniently facilitate compliance with data privacy protection.

Additionally, for the Mistral family, also included is their Mixtral model as this showcases a sparse mixture of experts model, which is said to improve computational efficiency compared with its counterpart LLMs. Not included in this study are the edge models of Mistral - Ministral 3B and 8B - as these were not available in AWS Bedrock at the time of the study.

To maximize determinism of these models during CNG, three parameters known to influence model determinism, $temperature$, top_p and top_k, were configured when relevant as enumerated in Table 1. We also set the maximum output tokens to the respective maximum capacity of each model.

3.3 Reliability Evaluation

Reliability was assessed in terms of *intra-prompt stability* and *correctness* on ten (10) iterations to show how consistent an LLM generates notes across multiple runs using the same prompt, and how well an LLM generates notes compared to those made by experts, respectively.

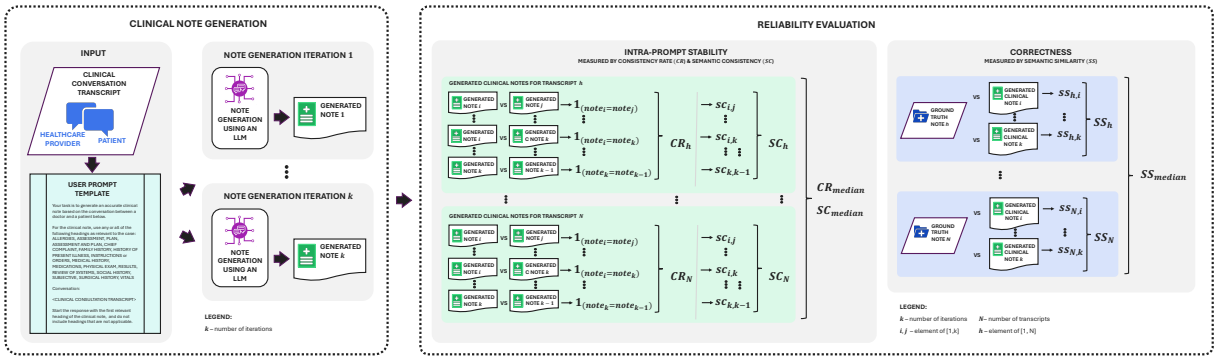

Figure 1: **Large Language Model (LLM) Reliability Evaluation Framework for the Task of Clinical Note Generation (CNG).** This has two phases, CNG and reliability evaluation, which are executed for each *transcript* that has a corresponding clinical note made by an expert (*ground truth note*). (1) *CNG* starts with said transcript being incorporated into a *user prompt template*, which then serves as an *input* to an LLM. The LLM response is the *generated note*. For each transcript, CNG is executed for k iterations, resulting in k generated notes. (2) *Reliability Evaluation* is then done to assess LLM reliability according to its *intra-prompt stability* and *correctness*. *Intra-prompt stability* is measured by consistency rate (CR) and median semantic consistency (SC_{median}), whereas *correctness* is evaluated by its median semantic similarity (SS_{median}). Once done for all transcripts, model performance is calculated by taking the median of these scores.

3.3.1 Intra-Prompt Stability

Intra-prompt stability is measured using the following metrics:

- ***Consistency Rate (CR)*** is measured by calculating the percentage of the number of pairs across the total number of iterations where a pair of generated notes are identical (i.e., string equivalent) over all possible combinations of pairs regardless of whether the outputs are correct. This strict measure of intra-prompt stability was first calculated per transcript (CR_h) as follows:

$$CR_h = \frac{\sum_{i,j \in \binom{k}{2}} 1_{i=j}}{\binom{k}{2}} * 100 \qquad (1)$$

  where k is the number of iterations and $h \in [1, N]$. Model performance was then calculated taking the median consistency rate (CR_{median}) from all CR scores.

- ***Semantic Consistency*** *(SC)* denotes whether the generated notes contextually mean the same regardless of how they were written across all iterations per transcript. This was measured by calculating the BERTScore which is an automatic text generation evaluation metric that calculates the cosine similarity between a pair of notes in the contextual embedding space (Zhang et al., 2020), using the implementation in Hugging Face. The pairs of notes refer to all combinations of the ten (10) generated notes per transcript. To determine model performance, the median semantic consistency SC_{median} was then calculated by getting the median of all semantic consistency scores calculated per transcript.

3.3.2 Correctness

Correctness is measured by ***semantic similarity***, which is similar to semantic consistency but the pair of notes compared here were the (1) generated note and (2) ground truth note made by an expert. The model performance (SS_{median}) was then calculated by taking the median of the semantic similarity scores calculated per transcript.

4 Results and Discussion

It took about 36 hours to generate the notes. Generally, we note that having perfect semantic consistency does not require having perfect consistency rate, and having perfect consistency rate and semantic consistency do not correspond to perfect semantic similarity as shown in Table 2.

4.1 Intra-prompt Stability

Figure 2 shows the intra-prompt stability of LLMs in CNG. Meta's Llama 1B and 3B models, as well as Anthropic's Claude Haiku model, demonstrated perfect intra-prompt stability, which means that all outputs from all iterations were exactly the same and thus have the same meaning. Such performances seemed inconsistent with prior work on LLM intra-prompt stability evaluation for multiple

Developer	Model	LLM Reliability ↑		
		Intra-prompt Stability		Correctness
		$CR_{median} \pm IQR$	$SC_{median} \pm IQR$	$SS_{median} \pm IQR$
Anthropic	Claude Haiku 3.5	**100.00** ± 00.00	**100.00** ± 00.00	85.61 ± 0.97
	Claude Sonnet 3.5 v2	0.00 ± 00.00	96.86 ± 1.44	86.52 ± 1.21
Meta	Llama 3.1-8B	35.56 ± 41.11	98.39 ± 6.06	83.71 ± 3.17
	Llama 3.1-70B	80.00 ± 37.78	**100.00** ± 0.00	85.90 ± 1.29
	Llama 3.1-450B	22.22 ± 20.00	96.34 ± 6.30	86.72 ± 1.95
	Llama 3.2-1B	**100.00** ± 00.00	**100.00** ± 0.00	80.49 ± 2.53
	Llama 3.2-3B	**100.00** ± 00.00	**100.00** ± 0.00	83.85 ± 1.39
Mistral	Large-2407 123B	8.89 ± 20.00	97.75 ± 2.83	84.36 ± 1.49
	Small-2402 22B	62.22 ± 33.33	**100.00** ± 0.00	85.72 ± 1.71
	Mixtral-8x7B	4.44 ± 11.11	96.14 ± 4.64	85.55 ± 1.55
OpenAI	ChatGPT-4o	0.00 ± 00.00	97.52 ± 1.42	87.01 ± 1.33
	ChatGPT-4o-mini	0.00 ± 00.00	97.40 ± 1.91	**87.26** ± 1.24

Table 2: **Summary of Model Performances based on Intra-Prompt Stability and Correctness**. In boldface are the best scores across all models while underlined are the best scores per model family. Three LLMs (25%) demonstrated perfect consistency rate, 41.67% (n=5) had perfect semantic consistency, and no model had perfect semantic similarity.

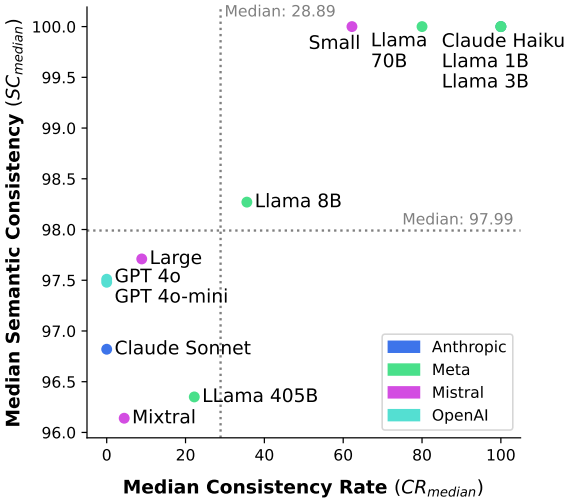

Figure 2: **Intra-prompt stability of LLMs in CNG.** Despite LLMs generating notes written in varied ways, the meaning of these notes were relatively consistent across multiple iterations, implying that these LLMs performed well in terms of intra-prompt stability.

choice question-answering tasks. Although the outputs of such tasks were linguistically controllable and are short-form texts, none of the LLMs studied by Atil et al. (2024) had string equivalent responses in multiple executions using the same prompt in all questions, and no LLM had perfect semantic consistency in the study of Zhao et al. (2024).

On the opposite side of the spectrum, at least in terms of consistency rate alone, the models from OpenAI consistently never produced string equivalent responses. Meta's LLama 3.1 8B model was unstable with the inter-quartile range greater than the median. Interesting to note as well are Meta's Llama 70B model and Mistral's Small model, which had likewise wide variances in their outputs, denoting that there are instances that these models can produce exactly the same results but can also produce responses that are written differently.

Nevertheless, considering both measures of intra-prompt stability, models from the Meta family generally performed better in terms of both consistency rate and semantic consistency than those from the other model families, whereas the models from the OpenAI family generally performed worse. Interestingly, for Anthropic, Meta and Mistral families, their smaller models performed remarkably better than their larger models. Also worth noting are the performance of Meta's Llama 70B model and Mistral's Small model, which both had perfect semantic consistency despite having an imperfect, but notably high, consistency rate.

In general, all models had a semantic consistency greater than 96% regardless of the consistency rate, which varied greatly between models from 0% to 100%. This implies that despite the models generating clinical notes written in a variety of ways, the meaning of the content of these notes was relatively consistent across multiple iterations. Thus, all models performed well in terms of intra-prompt stability. This implies that intra-prompt stability may be measured using semantic consistency alone than with consistency rate.

4.2 Correctness

Figure 3 shows how close the generated notes were to the ground truth notes, indicating correctness. Generally, all LLMs had a median semantic similarity between 80 and 88. For Anthropic, Meta and OpenAI, their larger models performed better than their smaller models. For Mistral, its Small model performed better than its Large model and its mixture-of-experts model.

A BERTScore of 80 is higher than the reported best performing LLM in the study of Giorgi et al. (2023), which has a BERTScore of 60.8, as validated by senior resident physicians. Although they also used **aci-bench**, they incorporated in-context learning in their implementation with the $temperature$ parameter set to 0.2.

4.3 Overall LLM Reliability

Shown in Figure 4 is the performance of the LLMs in CNG in terms of intra-prompt stability measured by semantic consistency and correctness measured by semantic similarity. Meta's Llama 70B model performed the best considering both semantic con-

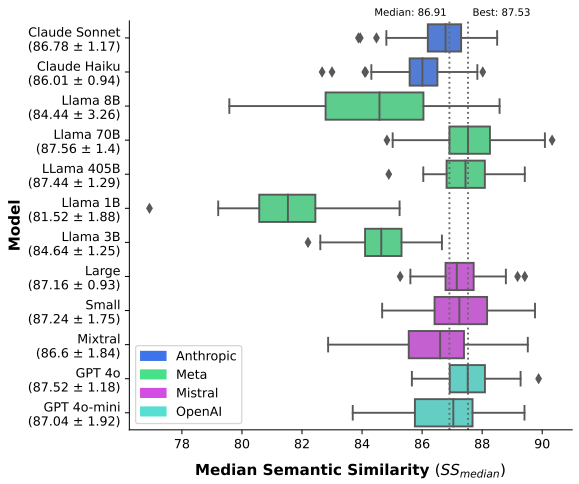

Figure 3: **Correctness of LLMs in CNG.** Except for Mistral, the larger models per model family performed better than their smaller models.

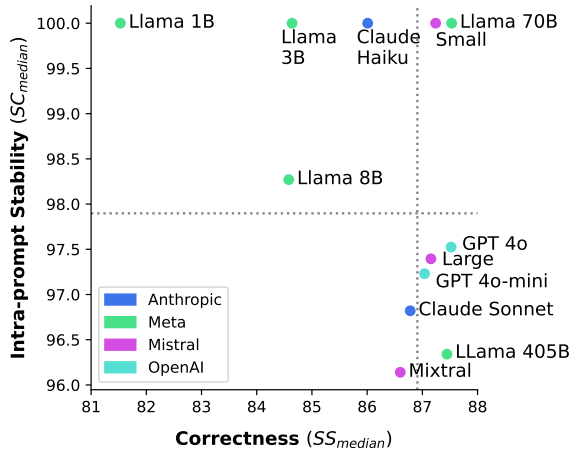

Figure 4: **Comparison of intra-prompt stability and correctness of LLMs in CNG.** Meta's Llama 70B model and Mistral's Small model appear to be among the most reliable models.

sistency and semantic similarity, followed by Mistral's Small model.These two open-weight models outperformed all proprietary models. For proprietary models, Anthropic's Claude Haiku had perfect semantic consistency but outperformed OpenAI's ChatGPT models in terms of semantic similarity.

Using proprietary models through their respective platforms is more accessible to HCPs but may result in a breach of data privacy regulations as the prompts submitted may get added to their database and subsequently used for training. Furthermore, the parameters that maximize determinism cannot be configured in these platforms. With Meta's Llama 70B and Mistral Small outperforming the proprietary models, we can develop CNG tools that use these and make these more accessible to HCPs without data privacy issues.

In addition, the choice of the final model would also depend on the practice setting considering that clinical conversations can vary in length, i.e., from as short as 5 minutes to at least an hour depending on the profession and area of practice. This is of particular concern for settings that deal with longer conversations such as in psychiatry, psychology, and occupational therapy where evaluations can take about an hour because of model limitations in terms of the number of tokens it can process. For such settings, we recommend Mistral's Small model over Meta's Llama 70B model.

5 Conclusion

The potential of LLMs for text generation has led to investigations into their ability to produce clinical notes, with the aim of improving the efficiency of documentation of HCPs. As part of our efforts to incorporate LLM-powered CNG tools into real clinical workflows, we have focused on building trust on these tools by assessing the reliability of LLMs in performing CNG.

Our observations indicate that LLMs do not consistently produce string-identical responses when aiming for semantically alike outputs, which are also aligned with annotations crafted by human experts. On multiple runs using the same prompt, we found that Meta's Llama 3.1 70B model was the most reliable, followed by Mistral's small model. Anthropic's Claude Haiku model outperformed OpenAI's ChatGPT 4o and 4o-mini models in terms of semantic consistency while the opposite was true for semantic similarity, but both proprietary models are subpar to Llama 3.1 70B and Mistral Small. With these findings, we recommend local deployment of these relatively smaller open-weight models for CNG to ensure compliance with data privacy regulations. We likewise consider using these models for SINTA as we validate its performance in the real world setting at the tertiary training hospital we are working with.

These findings provide support for the eventual integration of CNG tools powered by LLMs whilst protecting the health information of patients in compliance with data privacy regulations (Giorgi et al., 2023; Heilmeyer et al., 2024; Wang et al., 2024a). In this way, we can contribute to easing the burden of HCPs by providing them with tools that

can help them comply with their documentation requirements more efficiently.

6 Limitations

As this study did not include prompt optimization, future work could involve comparing the same measures across various prompts to check for robustness and, at the same time, identify the prompt most suitable for the task. Metrics that utilize knowledge graphs and sentence parsers can also be used, along with an evaluation by human experts.

Furthermore, our work only used one publicly available dataset that includes data gathered from simulations in English. We believe that it is necessary to conduct clinical validation and utility studies to capture and address contextual nuances before such tools can be fully adopted.

7 Ethical Considerations

The data used includes transcripts of dialogues between HCPs and patients, taken from a publicly available dataset. Protected health information was not used.

Although we used proprietary models in our experiments such that the prompts we submitted may get added to their database and subsequently used for training, caution must be exercised when considering the use of these models in real clinical workflows to avoid any potential breach of data privacy regulations.

Since clinical note generation tools are being developed with the intent of being integrated in real clinical workflows, we recommend conducting clinical validation and clinical utility studies prior to integration to ensure that the tools meet health standards and comply with regulations.

References

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. LLM Stability: A detailed analysis with some surprises. *arXiv preprint*. ArXiv:2408.04667 [cs].

Iman Azimi, Mohan Qi, Li Wang, Amir M. Rahmani, and Youlin Li. 2025. Evaluation of LLMs accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval. *Scientific Reports*, 15(1):1506. Publisher: Nature Publishing Group.

Jasmine Balloch, Shankar Sridharan, Geralyn Oldham, Jo Wray, Paul Gough, Robert Robinson, Neil J. Sebire, Saleh Khalil, Elham Asgari, Christopher Tan, Andrew Taylor, and Dominic Pimenta. 2024. Use of an ambient artificial intelligence tool to improve quality of clinical documentation. *Future Healthcare Journal*, 11(3):100157.

Christopher Barrie, Elli Palaiologou, and Petter Törnberg. 2024. Prompt Stability Scoring for Text Annotation with Large Language Models. *arXiv preprint*. ArXiv:2407.02039 [cs].

Anjanava Biswas and Wrick Talukdar. 2024. Intelligent Clinical Documentation: Harnessing Generative AI for Patient-Centric Clinical Note Generation. *arXiv preprint*. ArXiv:2405.18346.

Yu-Wen Chen and Julia Hirschberg. 2024. Exploring Robustness in Doctor-Patient Conversation Summarization: An Analysis of Out-of-Domain SOAP Notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 1–9, Mexico City, Mexico. Association for Computational Linguistics.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. RELIC: Investigating Large Language Model Responses using Self-Consistency. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120. Publisher: Proceedings of the National Academy of Sciences.

Shilpa Ghatnekar, Adam Faletsky, and Vinod E. Nambudiri. 2021. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. *Health and Technology*, 11(4):803–809.

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.

Felix Heilmeyer, Daniel Böhringer, Thomas Reinhard, Sebastian Arens, Lisa Lyssenko, and Christian Haverkamp. 2024. Viability of open large language models for clinical documentation in german health care: Real-world model evaluation study. *JMIR Med Inform*, 12:e59617.

Annessa Kernberg, Jeffrey A. Gold, and Vishnu Mohan. 2024. Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. *Journal*

of Medical Internet Research, 26(1):e54419. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Elie Kozaily, Mabelissa Geagea, Ecem R. Akdogan, Jessica Atkins, Mohamed B. Elshazly, Maya Guglin, Ryan J. Tedford, and Ramsey M. Wehbe. 2024. Accuracy and consistency of online large language model-based artificial intelligence chat platforms in answering patients' questions about heart failure. *International Journal of Cardiology*, 408:132115.

Taiji Li, Zhi Li, and Yin Zhang. 2024. Improving Faithfulness of Large Language Models in Summarization via Sliding Generation and Self-Consistency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8804–8817, Torino, Italia. ELRA and ICCL.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. Factual consistency evaluation of summarization in the Era of large language models. *Expert Systems with Applications*, 254:124456.

Lientje Maas, Mathan Geurtsen, Florian Nouwt, Stefan Schouten, Robin van de Water, Sandra van Dulmen, Fabiano Dalpiaz, Kees van Deemter, and Sjaak Brinkkemper. 2020. The Care2Report System: Automated Medical Reporting as an Integrated Solution to Reduce Administrative Burden in Healthcare. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Liam G. McCoy, Faye Yu Ci Ng, Christopher M. Sauer, Katelyn Edelwina Yap Legaspi, Bhav Jain, Jack Gallifant, Michael McClurkin, Alessandro Hammond, Deirdre Goode, Judy Gichoya, and Leo Anthony Celi. 2024. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review. *BMC Medical Education*, 24(1):1096.

Amirmasoud Momenipour and Priyadarshini R. Pennathur. 2019. Balancing documentation and direct patient care activities: A study of a mature electronic health record system. *International Journal of Industrial Ergonomics*, 72:338–346.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Juan C. Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *npj Digital Medicine*, 2(1):1–6. Number: 1 Publisher: Nature Publishing Group.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H. Chen. 2024. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association: JAMIA*, page ocae254.

Yash Saxena, Sarthak Chopra, and Arunendra Mani Tripathi. 2024. Evaluating Consistency and Reasoning Capabilities of Large Language Models. *arXiv preprint*. ArXiv:2404.16478 [cs].

Victoria Tucci, Joan Saary, and Thomas E. Doyle. 2021. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. *Journal of Medical Artificial Intelligence*, 5(0).

Joshua Yi Min Tung, Sunil Ravinder Gill, Gerald Gui Ren Sng, Daniel Yan Zheng Lim, Yuhe Ke, Ting Fang Tan, Liyuan Jin, Kabilan Elangovan, Jasmine Chiat Ling Ong, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Tsung Wen Chong. 2024. Comparison of the quality of discharge letters written by large language models and junior clinicians: Single-blinded study. *J Med Internet Res*, 26:e57721.

Hanyin Wang, Chufan Gao, Bolun Liu, Qiping Xu, Guleid Hussein, Mohamad El Labban, Kingsley Iheasirim, Hariprasad Korsapati, Chuck Outcalt, and Jimeng Sun. 2024a. Adapting open-source large language models for cost-effective, expert-level clinical note generation with on-policy reinforcement learning. *Preprint*, arXiv:2405.00715.

Leyao Wang, Zhiyu Wan, Congning Ni, Qingyuan Song, Yang Li, Ellen Clayton, Bradley Malin, and Zhijun Yin. 2024b. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. *Journal of Medical Internet Research*, 26(1):e22769. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024c. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1):1–9. Publisher: Nature Publishing Group.

Yuxuan Wu, Mingyue Wu, Changyu Wang, Jie Lin, Jialin Liu, and Siru Liu. 2024. Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use: Systematic Review and Meta-Analysis. *JMIR Medical Informatics*, 12(1):e54811. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics

Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586. Publisher: Nature Publishing Group.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, and Meliha Yetisgen. 2024. To Err Is Human, How about Medical Large Language Models? Comparing Pre-trained Language Models for Medical Assessment Errors and Reliability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16211–16223, Torino, Italia. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zhan Zhang, Karen Joy, Richard Harris, and Sun Young Park. 2022. Characteristics and Challenges of Clinical Documentation in Self-Organized Fast-Paced Medical Work. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):386:1–386:21.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Improving the Robustness of Large Language Models via Consistency Alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8931–8941, Torino, Italia. ELRA and ICCL.

A Sample Data from aci-bench

Each data point of the **aci** subset of **aci-bench** contains a corrected transcript of a natural conversation between a patient and a doctor (*clinical conversation transcript*), together with its corresponding clinical note which serves as the *ground truth note* for this study.

B User Prompt Template

Figure 6 shows the user prompt template used as input to the evaluated LLMs, except for Llama models. This template contains (1) the task, (2) the list of note headings present in the dataset, (3) the *transcript*, and (4) other specific instructions.

C Formatted Prompt Template for Llama Models

Llama models expect a certain format for the prompt, as shown in Figure 7.

TRANSCRIPT OF DIALOGUE

[doctor] hi russell how are you what's been going on
[patient] well i've been having this sharp pain on the right side of my abdomen below my ribs for the last several days
[doctor] i saw my doctor and they ordered a cat scan and said i had a kidney stone and sent me to see a urologist okay well does the pain move or or or go anywhere or does it stay right in that same spot yeah it feels like it goes to my lower abdomen in into my groin okay and is the pain constant or does it come and go it comes and goes when it comes it's it's pretty it's pretty bad i feel like i ca n't find a comfortable position okay and do you notice any any pain when you urinate or when you pee
[patient] yeah it kinda burns a little bit
[doctor] okay do you notice any blood i do n't think there is any you know frank blood but the urine looks a little dark sometimes okay and what have you taken for the pain i have taken some tylenol but it has n't really helped okay and do you have any nausea vomiting any fever chills i feel nauseated but i'm not vomiting okay is anyone in your in your family had kidney stones yes my father had them and have you had kidney stones before yeah so i i've i've had them but i've been able to pass them but this is taking a lot longer okay well i'm just gon na go ahead and do a physical examination i'm gon na be calling out some of my exam findings and i'm going to explain what what those mean when i'm done okay
[patient] okay
[doctor] okay so on physical examination of the abdomen on a abdominal exam there is no tenderness to palpation there is no evidence of any rebound or guarding there is no peritoneal signs there is positive cva tenderness on the right flank so essentially what that means russell is that you know you have some tenderness over your over your right kidney and that just means that you might have some inflammation there so i i reviewed the results of the ct scan of your abdomen that the primary care doctor ordered and it does show a . five centimeter kidney stone located in the proximal right ureter so this the ureter is the duct in which urine passes between the kidney and the bladder there's no evidence of what we call hydronephrosis this means you know swelling of the kidney which is good means that things are still able to get through so let's talk a little bit about my assessment and my plan okay so for your first problem of this acute nephrolithiasis or kidney stone i i wan na go ahead and recommend that you push fluids to help facilitate urination and peeing to help pass the stone i'm going to prescribe oxycodone five milligrams every six to eight hours as needed for pain you can continue to alternate that with some tylenol i'm going to give you a strainer that you can use to strain your urine so that we can see it see the stone when it passes and we can send it for some some tests if that happens i'm also gon na order what we call a basic metabolic panel a urinalysis and a urine culture now i wan na see you again in one to two weeks and if you're still having symptoms we'll have to discuss further treatment such as lithotripsy which is essentially a shock wave procedure in which we sedate you and use shock waves to break up the stone to help it pass we could also do what we call a ureteroscopy which is a small telescope small camera used to go up to to the urethra and bladder and up into the ureter to retrieve the stone so let's see how you do over the next week and i want you to contact me if you're having worsening symptoms okay okay sounds good thank you

CLINICAL NOTE

CHIEF COMPLAINT
Right-sided abdominal pain

MEDICAL HISTORY
Patient reports history of kidney stones.

FAMILY HISTORY
Patient reports his father has a history of kidney stones.

MEDICATIONS
Patient reports use of Tylenol.

REVIEW OF SYSTEMS
Gastrointestinal: Reports right-sided abdominal pain and nausea. Denies vomiting
Genitourinary: Reports dysuria and dark colored urine. Denies hematuria.

PHYSICAL EXAM
Gastrointestinal
- Examination of Abdomen: No masses or tenderness to palpation. No rebound or guarding. No peritoneal signs. Positive CVA tenderness on the right flank.

RESULTS
Previous CT scan of the abdomen ordered by the patient's PCP is reviewed and demonstrates a 0.5 cm kidney stone located in the proximal right ureter. There is no evidence of hydronephrosis.

ASSESSMENT AND PLAN
1. Acute nephrolithiasis.
- Medical Reasoning: The patient presents with complaints of right-sided abdominal pain. His previous CT scan was reviewed and demonstrates a 0.5 cm kidney stone located in the proximal right ureter without evidence of hydronephrosis.
- Medical Treatment: I have recommended that he push fluids in order to help facilitate urination to help pass the stone. He will be provided with a strainer to allow us to potentially test the stone if he is able to pass it. I have also prescribed oxycodone 5 mg every 6 to 8 hours as needed for pain. He can continue to alternate oxycodone with Tylenol. A basic metabolic panel, urinalysis, and urine culture will also be ordered.

INSTRUCTIONS
He will follow up in 1 to 2 weeks. If he is still having symptoms at that time, we will discuss further treatment such as lithotripsy or ureteroscopy. He is to contact me if he is having worsening symptoms over the next week.

Figure 5: **Sample data from the *aci-bench* dataset.** An example of the corrected transcript of a natural conversation between a patient and a doctor (*clinical conversation transcript*), together with its corresponding clinical note which serves as the *ground truth note* for this study.

User Prompt Template
Your task is to generate an accurate clinical note based on the conversation between a doctor and a patient below.
For the clinical note, use any or all of the following headings as relevant to the case: ALLERGIES, ASSESSMENT, PLAN, ASSESSMENT AND PLAN, CHIEF COMPLAINT, FAMILY HISTORY, HISTORY OF PRESENT ILLNESS, INSTRUCTIONS or ORDERS, MEDICAL HISTORY, MEDICATIONS, PHYSICAL EXAM, RESULTS, REVIEW OF SYSTEMS, SOCIAL HISTORY, SUBJECTIVE, SURGICAL HISTORY, VITALS
Conversation:
$< clinical\ conversation\ transcript >$
Start the response with the first relevant heading of the clinical note, and do not include headings that are not applicable.

Figure 6: **User Prompt Template.** This was used to keep the format consistent across all models.

<|begin_of_text|>
<|start_header_id|>user<|end_header_id|>
user_prompt
<|eot_id|>

Figure 7: **Formatted Prompt Template.** This was used to keep the format consistent across all Llama models. $user\_prompt$ here refers to the input which contains the transcript included in the User Prompt Template (Figure 6).