TN-Eval: Rubric and Evaluation Protocols for Measuring the Quality of Behavioral Therapy Notes

Raj Sanjay Shah[1]*, Lei Xu[2], Qianchu Liu[2], Jon Burnsky[2], Drew Bertagnolli[3], Chaitanya Shivade[2]
[1] Georgia Institute of Technology, [2] AWS AI Labs, [3] OneMedical
rajsanjayshah@gatech.edu
{ leixx, liufqian, jburnsky, shivadc } @amazon.com
abertagnolli@onemedical.com

Abstract

Behavioral therapy notes are important for both legal compliance and patient care. Unlike progress notes in physical health, quality standards for behavioral therapy notes remain underdeveloped. To address this gap, we collaborated with licensed therapists to design a comprehensive rubric for evaluating therapy notes across key dimensions: *completeness*, *conciseness*, and *faithfulness*. Further, we extend a public dataset of behavioral health conversations with therapist-written notes and LLM-generated notes, and apply our evaluation framework to measure their quality. We find that: (1) A rubric-based manual evaluation protocol offers more reliable and interpretable results than traditional Likert-scale annotations. (2) LLMs can mimic human evaluators in assessing completeness and conciseness but struggle with faithfulness. (3) Therapist-written notes often lack completeness and conciseness, while LLM-generated notes contain hallucinations. Surprisingly, in a blind test, therapists prefer and judge LLM-generated notes to be superior to therapist-written notes. As recruiting therapists for annotation is expensive, we release the rubric, therapist-written notes, and expert annotations to support future research.[1]

1 Introduction

Automated medical note generation using large language models (LLMs) has the potential to enhance clinicians' efficiency by reducing the time spent on electronic health records, allowing them to focus more on patient care. However, applying LLMs to behavioral health notes presents unique challenges (Hua et al., 2024). In therapy, the conversation itself is the treatment; therefore, techniques like motivational interviewing used in a session may not be explicitly stated. Furthermore, sessions cover various topics, making it crucial to discern significant details from less relevant information. Given the high-stakes nature of behavioral health, using LLMs to generate notes must be rigorously evaluated to ensure they capture key information at an appropriate level of detail.

Evaluating the quality of talk therapy notes, however, is not straightforward. Traditionally, human evaluation has been the primary method for assessing their quality, making it resource intensive and costly. Moreover, a lack of standardized reference notes and the limited literature on what constitutes an effective behavioral health therapy note further complicates the evaluation process. Therapists and healthcare providers often have their own styles and preferences, leading to subjective assessments and considerable variation. Without clear standards and evaluation protocols, it becomes difficult to determine the quality of LLM-generated therapy notes.

In this work, we focus on the SOAP (Subjective, Objective, Assessment, Plan) format of therapy notes and propose an evaluation framework for notes (TN-Eval). The framework includes (1) a comprehensive, fine-grained, section-wise rubric that outlines the key components and characteristics of a therapy note and (2) both human and automatic evaluation protocols. The rubric, which we co-designed with 5 licensed therapists, details the relevant items for each of the four SOAP sections and their respective levels of importance **(Section 3)**. We then design a human evaluation protocol – TN^{H}-Eval – in which 9 licensed behavioral health therapists from diverse backgrounds assess notes along three dimensions: completeness, conciseness, and faithfulness **(Section 4.1)**. The completeness and conciseness are scored with reference to the rubric to improve the consistency of the evaluation, while faithfulness is evaluated at the sentence level with source attribution (Rashkin et al., 2023). Finally, we explore

*Work done during internship at Amazon.
[1] https://github.com/amazon-science/TN-Eval

the potential of LLMs to emulate expert evaluations, introducing an automatic evaluation protocol called TN^A-Eval **(Section 4.2)**.

Our experimental results show that our proposed human evaluation protocol – TN^H-Eval achieves higher Inter-Annotator Agreement (IAA) compared to conventional Likert-scale human evaluation, making it more reliable. We additionally show that using the automatic evaluation protocol – TN^A-Eval, we can achieve a better correlation with TN^H-Eval on completeness and conciseness evaluation when compared to N-gram-based metrics like ROUGE (Lin, 2004) or conventional LLM-as-a-Judge (Zheng et al., 2023), making it a quick and cost-effective solution for evaluation. When compared to expert-written notes, we find that LLM-generated notes achieve around 10% higher scores in completeness and conciseness but show relatively lower faithfulness.

Deployment considerations: Our TN^A-Eval is a *deployable* and *scalable* framework for assessing therapy notes with fine-grained, human-like judgments, which is designed by domain experts and has been evaluated on available datasets. Integrating this evaluation into clinical workflows and EHR systems enables: (1) Automated review that flags low-quality notes; (2) Automated scoring systems that assist therapists in refining notes before submission, reducing post-session documentation workload; and (3) Cost-effective, scalable quality assessments in standardized documentation practices. Refer appendix section I for workflow integration suggestions.

2 Related Work

AI in Mental Health Care: Recently, interest in using LLMs for mental health care has grown (Greer et al., 2019; Peng et al., 2020; Srivastava et al., 2022; Luo et al., 2025), with research focusing on three main directions. First, to classify therapeutic methods used by clinicians, assess the effectiveness of treatments, and predict the quality of service (Saha and Sharma, 2020; Chikersal et al., 2020; Liu et al., 2021; Shah et al., 2022). Second, virtual counselors emulate human behavior in chatbot-like environments (Shen et al., 2020; O'neil et al., 2023), but ethical and legal concerns (Woodnutt et al., 2024; Stade et al., 2024) have shifted research toward augmenting therapists with suggestions to enhance their responses (Saha et al., 2022; Sharma et al., 2023a). Third, AI tools train novice counselors by providing automatic feedback (Chaszczewicz et al., 2024; Lin et al., 2024) and simulating client personas for role-play (Stapleton et al., 2023; Wang et al., 2024; Louie et al., 2024, 2025). Despite growing interest in AI for mental health support, LLMs for behavioral therapy note generation remain underexplored.

Automated clinical note generation: Generation of medical documentation has been shown to improve clinician efficiency (Joshi et al., 2020), with research primarily focused on physical health using role-play or anonymized conversations and human-written notes (Papadopoulos Korfiatis et al., 2022; Ben Abacha et al., 2023; Yim et al., 2023). Early work fine-tuned lightweight transformer models (Sharma et al., 2023b; Michalopoulos et al., 2022; Milintsevich and Agarwal, 2023; Yuan et al., 2024), while recent studies explore LLM prompting for summarization (Ben Abacha et al., 2023; Mathur et al., 2023).

Automatic evaluations for summarization: Reference-based metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020) are widely used to measure lexical similarity between generated and reference summaries. Recent work has expanded to fact-checking-based evaluators (Honovich et al., 2022; Zha et al., 2023; Laban et al., 2022) and LLM-as-a-Judge protocols (Zheng et al., 2023; Wang et al., 2023), which rely on general-purpose models to holistically score summaries. Benchmarks like HealthBench (Arora et al., 2025) further incorporate physician-created rubrics with LLM-based graders to evaluate model utility and safety in clinical tasks. However, previous methods are usually developed for general text summarization tasks and do not account for the challenges of therapy notes, where obtaining high-quality reference summaries is complex, and evaluations require substantial domain knowledge. In contrast, our TN^A-Eval adapts LLM-based evaluation to operate over a structured, domain-specific rubric grounded in behavioral health practice, enabling scalable yet clinically grounded assessment.

3 SOAP Note and Rubric creation

In TN-Eval, we look at a popularly used therapy note documentation format: SOAP, an acronym for Subjective, Objective, Assessment, Plan, with

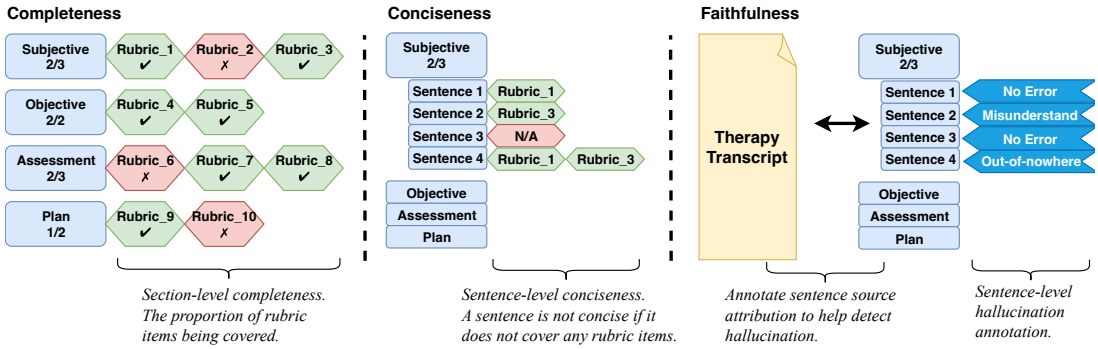

Figure 1: The TN^{H}-Eval human evaluation protocol.

each letter representing a section of the note (Weed, 1964).

At a high level, in SOAP notes, the *subjective* component consists of insights about the client's presenting problem from the client's viewpoint and that of significant others. In contrast, the *objective* component includes the counselor's observations. The *assessment* section shows how the subjective and objective data are being analyzed, interpreted, and considered, and the *plan* section outlines the treatment approach (Cameron and Turtle-Song, 2002). While there exist other therapy note formats, we use SOAP notes because they are widely referenced in behavioral therapy (Berghuis et al., 2014; Reiter and Sabo, 2023), standardized in major electronic health records (Podder et al., 2022; Gao et al., 2023), and provide a representative framework for developing better evaluation protocols.

In practice, the exact definitions and information present in each of the sections are determined by the healthcare provider organization and its record management practices. Therefore, a fine-grained set of consistent rubrics is necessary to complement the definition. In addition to the generally underspecified definitions, there is a lack of consistent clinical psychology literature for best practices in writing therapy notes and key characteristics that determine the quality of a note. To determine what *high-quality means to domain experts*, we work with five therapists to **co-design a rubric consisting of the different section-wise dimensions of note quality.**

3.1 Domain Experts

To develop the rubric, we collaborated with **Therapist A**[2] who has over 20 years of clinical experience. Additionally, we worked with four other therapists from diverse professional backgrounds who hold a Psy.D., Ph.D. in counseling psychology, or licensed clinical social work, and have experience with multiple healthcare providers, as well as training new therapists in therapeutic techniques and note-writing.

3.2 Rubric Creation Procedure

Each rubric item captures a key characteristic expected in each section of a SOAP note. These characteristics reflect clinical best practices and are annotated with their relative importance (Mandatory, Recommended, etc). We developed the rubric in a two-step co-design process. In the first step, we conduct three hour-long sessions with Therapist A to identify key characteristics of each SOAP note section, assign their relative importance, and refine the rubric through iterative feedback and example notes, including general section-agnostic guidelines.

In the second step, we ask four more therapists to verify the section appropriateness and the relative importance of each key characteristic. We also ask the therapists to suggest key characteristics that may be missed from the first step. The process is completed in an annotation tool shown in appendix figure 2. After the annotation, we consolidate the rubric by taking the majority vote. The final definitions of the SOAP note and the corresponding section-wise key characteristics are presented in appendix A. We also validate the final rubric with ($N = 17$) external therapists employed in note writing ($N = 8$) and evaluation ($N = 9$), as mentioned in appendix C.

Rubric Quality: We observe perfect IAA among 5 experts for the appropriate section for each key characteristic and observe high agreements for the relative importance of each key characteristic – Fleiss' κ: 0.68, Krippendorff's α: 0.73. Detailed IAAs by section is shown in appendix Table 9.

[2]Therapist A is a co-author of this paper

4 Evaluation Protocols

In this section, we introduce human and automatic evaluation protocols using the rubric, denoted as TN^H-Eval and TN^A-Eval, respectively. Both focus on three dimensions:

Completeness: This dimension evaluates whether each rubric element appears in its corresponding note section (e.g., the chief complaint in subjective). The score is computed as the ratio of covered elements aggregated across sections.

Conciseness: This metric measures whether each sentence contributes to a rubric item. Annotators (human or automated) label sentences accordingly, and the score is the ratio of necessary sentences in a section.

Faithfulness: This evaluation checks whether a note's content is factually grounded in the therapy session. Errors are categorized into hallucination types, ensuring a granular assessment.

Why these three dimensions? The evaluation dimensions were chosen based on practical considerations and therapist feedback. Since no standardized framework exists for grading therapists' notes, completeness is crucial to meet regulatory requirements. Therapists also emphasized conciseness, noting concerns about AI-generated verbosity. Lastly, faithfulness was included to mitigate hallucinations in LLM-generated text, ensuring accuracy and reliability.

4.1 Human Evaluation Protocols

Our TN^H-Eval relies on our rubric design to break down each dimension into more objective, simpler, and cost-effective tasks. Figure 1 illustrates the human evaluation protocol. The left panel shows completeness and conciseness annotations, where sentences are labeled with associated rubric items. The right panel illustrates faithfulness evaluation via sentence-to-transcript alignment and hallucination labeling.

To find the **completeness**, a therapist reviews a note section and marks covered rubric elements. The full note score is computed as a micro average, weighted by section rubric elements. This design minimizes annotators' effort in reviewing lengthy therapy transcripts (45 minutes). For **conciseness**, annotators label sentences with relevant rubric items or mark them as unnecessary. This annotation is done separately from completeness to prevent biased coverage assessment. In the case of **faithfulness**, annotators cross-check sentences against the therapy transcript, selecting supporting content from the source and categorizing hallucinations into (1) Out-of-nowhere, (2) Misinterpreted Information, or (3) No Error. Given the session length, this is the most costly evaluation. *While non-experts could perform this task, all annotations in our study are conducted by licensed U.S. therapists to ensure accuracy and reliability.*

4.2 Automatic Evaluation Protocols

We use LLMs to mimic human annotators to get the completeness and conciseness evaluation. For **completeness**, we present a note and one rubric item to an LLM and ask if the item appears in the summary. For **conciseness**, we break down the note into sentences, and for each sentence, we verify if a rubric element is covered in the sentence. We use AlignScore (Zha et al., 2023) for the **faithfulness** evaluation.

5 Data collection and note generation

Dataset: We conducted experiments on therapy conversations from the AnnoMI dataset (Wu et al., 2023). Due to the cost of recruiting expert therapists for annotation, we chose the first 50 conversations from the high-quality split of AnnoMI (the median conv. len. = 1067 words/ 42 turns).

Human Note Collection: The notes were written by the $N = 5$ internal therapists involved in the rubric design, and we also recruited ($N = 8$) therapists to write notes for these 50 conversations. The cost to collect each note was $206.

LLM Note Generation: We prompted several off-the-shelf LLMs to generate notes, including Claude (Anthropic, 2024), Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023), and also use two clinical and therapy domain adapted LLMs – MentalLlama(Yang et al., 2024) and OpenBioLLM (Ankit Pal, 2024). Appendix F shows the prompt we used for note generation. The prompt is simple and not carefully optimized for any particular LLM to achieve a fair comparison between LLMs.

Human Evaluation: For evaluation, we recruited $N = 9$ external therapists who are different from those who wrote the notes. The cost to collect a single human evaluation related to one note is $190. We followed the TN^H-Eval protocol described in Section 4.1, and collected two independent annotations for each note. We also collect

Note	Completeness		Conciseness		Faithfulness		Acceptance
	TN^H-Eval	Likert	TN^H-Eval	Likert	TN^H-Eval	Likert	Likert
Human	29.5 (±12.4)	2.85 (±1.09)	75.6 (±14.9)	4.28 (±0.89)	87.0 (±12.6)	4.43 (±0.81)	2.34 (±0.75)
Llama 3.1 70B	39.7 (±7.9)	3.80 (±0.79)	84.0 (±12.1)	4.83 (±0.35)	68.5 (±15.1)	4.68 (±0.50)	3.34 (±0.61)
Mistral Large V2	38.1 (±7.5)	4.01 (±0.70)	91.5 (±7.1)	4.88 (±0.35)	71.8 (±14.0)	4.90 (±0.34)	3.73 (±0.70)

Table 1: Human evaluation results using TN^H-Eval and Likert human evaluations. The values in brackets show the standard deviation over the 50 examples. "Acceptance" refers to whether the therapist would accept the note for clinical use, rated on a 5-point Likert scale. This table shows aggregated scores for the full note. See Table 2 for a breakdown by sections.

Section	Note	Completeness		Conciseness		Faithfulness	
		TN^H-Eval	Likert	TN^H-Eval	Likert	TN^H-Eval	Likert
Subjective	Human	41.7 (±22.8)	3.28 (±1.11)	84.7 (±20.8)	4.49 (±0.72)	92.0 (±15.0)	4.64 (±0.67)
	Llama 3.1 70B	46.0 (±12.4)	3.86 (±0.74)	90.8 (±17.8)	4.81 (±0.35)	95.0 (±10.9)	4.66 (±0.52)
	Mistral Large V2	47.8 (±13.6)	4.14 (±0.61)	88.7 (±15.4)	4.91 (±0.24)	97.9 (±5.7)	4.87 (±0.40)
objective	Human	21.8 (±18.3)	2.51 (±1.06)	65.9 (±29.9)	4.10 (±0.86)	85.1 (±23.2)	4.40 (±0.74)
	Llama 3.1 70B	36.0 (±8.8)	3.56 (±0.87)	81.8 (±27.0)	4.82 (±0.36)	49.0 (±30.0)	4.75 (±0.39)
	Mistral Large V2	39.6 (±7.8)	3.95 (±0.64)	89.0 (±14.7)	4.90 (±0.36)	60.4 (±28.9)	4.94 (±0.26)
Assessment	Human	26.9 (±16.1)	2.94 (±1.02)	83.0 (±23.8)	4.35 (±0.81)	85.4 (±22.9)	4.57 (±0.62)
	Llama 3.1 70B	34.1 (±10.6)	3.72 (±0.71)	94.7 (±12.2)	4.82 (±0.37)	80.9 (±22.7)	4.70 (±0.52)
	Mistral Large V2	30.4 (±9.9)	3.97 (±0.68)	95.5 (±11.8)	4.80 (±0.44)	84.8 (±21.4)	4.90 (±0.35)
Plan	Human	26.2 (±19.9)	2.67 (±1.03)	68.4 (±35.9)	4.17 (±1.11)	78.2 (±33.1)	4.13 (±1.08)
	Llama 3.1 70B	42.5 (±19.4)	4.05 (±0.76)	72.9 (±25.2)	4.87 (±0.33)	46.6 (±34.2)	4.61 (±0.55)
	Mistral Large V2	37.2 (±19.3)	3.97 (±0.85)	94.4 (±10.1)	4.89 (±0.32)	43.8 (±34.4)	4.88 (±0.33)

Table 2: Section-wise human evaluation results using TN^H-Eval and Likert-style human evaluations.

annotations for *5-point Likert-scale baseline* on three aspects – Completeness, Conciseness, and Faithfulness. Experts also annotate the *overall acceptance* of a note on a scale of 1 to 5. Due to the high cost of human annotation, we only conducted the TN^H-Eval on human notes and 2 LLM-generated notes – Llama 3.1 (70B) and Mistral Large V2.

Automatic Evaluation: We followed the protocol in Section 4.2 to conduct automatic evaluation. We also explored a Likert-style automatic evaluation similar to LLM-as-a-judge (Zheng et al., 2023) (refer to appendix E.1 for corresponding prompts). We compared TN-Eval with conventional reference-based evaluation, such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020), and we find the efficacy of our automatic evaluation protocol by correlating the automatic metric with human annotations at the note-level.

6 Experiments

Q1: How reliable is TN^H-Eval compared to the conventional Likert-based approach?

Table 3 shows the IAA between two annotators for each type of human rating we collect. We found that Krippendorff's α for TN^H-Eval is significantly higher than that of the Likert-style evaluation for all three dimensions, showing that TN^H-Eval can achieve more consistent annotations across two independent annotators and is thus more reliable. Furthermore, TN^H-Eval provides distinct variance in outputted scores as compared to expert Likert scale judgments (refer figures 3, 4).

Dimension	TN^H-Eval		Likert	
	Raw Agg.	K-α	MSE	K-α
Completeness	77.6	0.52	2.72	0.08
Conciseness	85.5	0.49	1.01	0.16
Faithfulness	85.9	0.62	0.86	0.18
Acceptance	-	-	2.24	0.15

Table 3: IAA of human evaluations. We show raw agreement and Krippendorff's α (K-α) for rubric annotations and mean squared error (MSE) and K-α for Likert annotations. "Acceptance" refers to the overall acceptance annotated on a Likert scale. TN^H-Eval appears to have better annotation consistency compared to Likert annotations.

Table 1 shows human-annotated scores for human-written notes and 2 LLM-generated notes. Table 2 shows the corresponding breakdown of scores by section. Note that the sources are revealed to the annotators. **It is surprising to see,**

Evaluator	Note Source	Completeness		Conciseness		Faithfulness	
		TN^{A}-Eval	Likert	TN^{A}-Eval	Likert	TN^{A}-Eval	Likert
Claude 3 Sonnet (AlignScore for Faithfulness)	Human	26.4 (±12.4)	2.85 (±0.41)	63.4 (±22.2)	3.15 (±0.49)	73.2 (±14.9)	4.11 (±0.57)
	Claude 3 Sonnet	34.8 (±7.3)	3.39 (±0.27)	86.0 (±10.3)	**3.86 (±0.21)**	74.0 (±10.1)	4.73 (±0.36)
	Claude 3 Haiku	36.8 (±8.7)	3.46 (±0.27)	**87.6 (±12.1)**	3.81 (±0.21)	69.9 (±10.1)	4.70 (±0.38)
	Llama 3.1 (70B)	35.1 (±8.0)	3.33 (±0.36)	84.8 (±12.5)	3.52 (±0.27)	69.0 (±11.6)	4.49 (±0.50)
	Llama 3.1 (8B)	35.0 (±6.8)	3.22 (±0.27)	85.9 (±8.6)	3.55 (±0.27)	70.2 (±11.5)	4.63 (±0.44)
	Mistral Large V2	36.8 (±8.2)	3.50 (±0.32)	84.3 (±9.1)	3.83 (±0.20)	75.8 (±8.8)	4.91 (±0.20)
	Mistral (7B)	**37.7 (±8.6)**	**3.58 (±0.28)**	81.2 (±10.5)	3.85 (±0.17)	75.2 (±9.5)	**4.93 (±0.19)**
	MentaLlama (13B)	24.5 (±10.2)	2.86 (±0.33)	77.0 (±20.8)	3.42 (±0.40)	**80.4 (±9.9)**	4.50 (±0.60)
	OpenBioLLM (70B)	24.6 (±9.9)	3.19 (±0.42)	72.9 (±13.9)	3.72 (±0.45)	80.0 (±11.0)	4.76 (±0.55)

Table 4: TN^{A}-Eval and Likert-style automatic evaluation. We show the results using Claude 3 Sonnet as the evaluator. Note that the TN^{A}-Eval faithfulness evaluation is conducted using AlignScore, not LLM-grading.

according to Likert-style scores, that experts judge LLM-generated notes to be superior to human-written notes across all dimensions – completeness, conciseness, faithfulness, and overall acceptance. Our TN^{H}-Eval shows the same order for completeness and conciseness, however, for faithfulness, TN^{H}-Eval shows higher scores for human-generated notes, which shows the advantage of using our rubric, breaking down each section into smaller, more objective annotation tasks.

Q2: Does TN^{A}-Eval align with human and reference-based automatic evaluations?

LLM	R1-F	R2-F	RL-F	BERT.
Claude 3 Sonnet	39.8	10.1	20.0	87.9
Claude 3 Haiku	40.7	10.9	20.3	87.9
Llama 3.1 (70B)	41.1	10.6	20.5	88.1
Llama 3.1 (8B)	39.4	10.4	20.2	87.6
Mistral Large V2	40.1	10.3	19.9	87.9
Mistral (7B)	39.9	9.7	19.5	87.9

Table 5: Reference-based evaluation metrics for notes generated by different LLMs, using human notes as a reference. We show F-measure for ROUGE-1/2/L, as well as BERTScore.

We find that all traditional reference-based metrics show similar values, making these n-gram-based metrics insufficient to provide meaningful signals for generation quality (refer table 5 for ROUGE and BERTScore results). Next, table 6 shows the correlation between two sets of automatic evaluation (TN^{A}-Eval and Likert-style LLM-as-a-Judge) and two sets of human evaluation (TN^{H}-Eval and Likert-scores). **Notably, TN^{A}-Eval and TN^{H}-Eval indicated the highest correlation, as shown in Column (A), demonstrating that the fine-grained evaluation achieves higher agreement between human and LLM evaluators.** Column (B) reveals the utility of the LLM-as-a-Judge for the completeness evaluation but presents a poor correlation for conciseness and faithfulness. When comparing automatic evaluations with Human Likert-scale annotations, the correlations appear to be generally poor, suggesting that neither of the automatic evaluations correlates well with human Likert-scale evaluation. Overall, the faithfulness correlation shows significant challenges in hallucination detection. Human and automatic evaluations agree that human-written notes are roughly 10% less complete and 10% less concise compared to LLM-generated notes. For the faithfulness evaluation, humans appear to favor human-written notes, while automatic evaluations favor LLM notes.

	v.s.	TN^{H}-Eval		Human Likert	
	Evaluator	(A) TN^{A}-Eval	(B) Likert	(C) TN^{A}-Eval	(D) Likert
Comp.	Claude 3 Sonnet	**0.58**	0.46	0.24	0.34
	Llama 3.1 (70B)	0.44	0.55	0.23	0.36
	Mistral Large V2	0.48	0.55	0.34	0.36
Conc.	Claude 3 Sonnet	0.36	0.27	0.19	0.26
	Llama 3.1 (70B)	0.39	0.14	0.26	0.11
	Mistral Large V2	**0.40**	0.24	0.21	0.17
Faith.	Claude 3 Sonnet	-	-0.15	-	0.28
	Llama 3.1 (70B)	-	-0.20	-	0.18
	Mistral Large V2	-	-0.22	-	0.19
	AlignScore	**0.34**	-	0.27	-

Table 6: The note-level correlation between automatic metrics and human annotations. Column (A) and (B) compares automatic evaluation with TN^{H}-Eval. TN^{A}-Eval achieves much higher correlation than Likert-style LLM-as-a-Judge. Column (C) and (D) compares automatic evaluation with human Likert-style annotation, where the correlation is generally poor.

Q3: How effectively do LLMs generate notes?

Upon asking experts to rate notes without telling the source of the note (refer table 1), we observe that experts prefer and judge LLM-generated notes

Note Source	S.	O.	A.	P.
Human Notes	76 (±57)	32 (±21)	57 (±41)	29 (±14)
Claude 3 Sonnet	73 (±23)	41 (±10)	64 (±13)	71 (±12)
Claude 3 Haiku	97 (±25)	46 (±11)	77 (±16)	94 (±22)
Llama 3.1 (70B)	65 (±15)	37 (±13)	61 (±11)	75 (±11)
Llama 3.1 (8B)	94 (±25)	56 (±13)	77 (±17)	82 (±15)
Mistral Large V2	88 (±23)	51 (±9)	65 (±12)	74 (±11)
Mistral (7B)	86 (±25)	51 (±10)	66 (±12)	75 (±11)

Table 7: Number of words (and standard deviation) in each section of the note based on source.

to be superior to human-written notes across all dimensions except fine-grained faithfulness evaluation. This highlights the potential of using LLMs for therapy note construction.

Note Length: For further investigation, we examine the length of notes written by the therapists and LLMs (table 7). Human-written notes are generally shorter, and in particular, the "plan" section of human notes is much shorter than LLM notes (Average length of human-plan section notes = 29 words, Average length of LLM-generated plan section = 78.5 words). This is because therapists tend to be very concise, with just one sentence stating the follow-up session, while LLM-generated notes contain more content such as "short-term goals" and "long-term goals" (see table 10). We believe that the natural and fluent English writing from LLMs likely biases human annotators, thus conflating fluency with accuracy (Elangovan et al., 2024). Next, we manually observed some examples (table 10) and found that humans tend to write shorter sentences for the same rubric items. Based on a subsequent conversation with Therapist A, we uncover that therapists spend substantial time with various kinds of documentation and find themselves hard-pressed to write descriptive quality notes (Griswold, 2019).

Section-wise scores: On analyzing the TN^H-Eval score breakdown for each rubric item, we observe that human-written notes show considerably less coverage for some rubric items (refer to table 8). For example, "symptoms" in the Subjective, "mental status" in the Objective, and "future interventions" in the Plan show a large discrepancy (more than 20%).

Automatic evaluation: We show the automatic evaluation results on Human notes and several LLM notes in Table 4. The numbers reflect a similar pattern to the human evaluation, where LLMs, in general, outperform humans in note completeness and conciseness. Among LLMs, Mistral tends to be more conservative, with more faithful content.

Rubric	Human	Llama	Mistral
Subjective			
chief-complaint	78%	75%	78%
symptoms	*56%*	87%	90%
history	59%	56%	59%
goals	33%	40%	42%
homework	1%	1%	3%
quotes	23%	17%	15%
Objective			
observed-behavior	*53%*	96%	98%
mental-status	*22%*	73%	88%
assessment-tools	10%	5%	7%
therapy-activities	12%	4%	4%
interventions	12%	2%	1%
Assessment			
diagnosis	8%	22%	13%
triggers	*19%*	40%	24%
progress	24%	38%	34%
analysis	*72%*	97%	92%
response	39%	30%	32%
overall-progress	8%	11%	11%
goals	4%	4%	3%
stages	41%	31%	34%
Plan			
future-interventions	*39%*	83%	75%
follow-up	31%	45%	41%
adjustment	2%	9%	7%
homework	33%	33%	26%

Table 8: Coverage of key characteristics in the rubric in therapist-written and LLM-generated notes. We highlight rubric items where coverage of human notes is over 20% lower than the best LLM.

7 Conclusion

In this paper, we conducted analyses on quality evaluation strategies for behavioral health therapy notes. By collaborating with domain experts to design a rubric, we designed fine-grained human evaluation and automatic evaluation protocols. We demonstrated the advantage of TN^H-Eval against conventional Likert-style human evaluation. Expert evaluation with TN^H-Eval and conventional Likert-scales shows preference towards LLM-generated notes. Our TN^A-Eval outperformed the conventional LLM-as-a-Judge strategy and showed a higher correlation with human evaluations for completeness and conciseness, while the faithfulness evaluation remains a challenge. Thus, we urge research toward robust and automatic evaluation of therapy notes. Subsequently, we are sharing high-quality note annotations from practitioners, the co-designed rubric, and all annotations we collected in the project to benefit the research community.

Acknowledgement

We sincerely thank the licensed therapists from OneMedical who contributed their expertise and time to the creation of the evaluation rubric, note generation, and annotation process.

Ethical Considerations

The organization's review protocols approved the current study. We do not advocate for fully automated LLM-generated notes; rather, we propose augmenting therapist workflows by providing an LLM-generated draft as a starting point. Furthermore, all therapy transcripts used in this work are from an open-source dataset – AnnoMI (Wu et al., 2023). Lastly, to ensure the appropriate stakeholder inclusion and the generalizability of findings, ($N = 22$) therapists were consulted in this study in the following capacities:

1. Therapist involvement in the co-design process. We work with $N = 5$ senior therapists from one of the largest behavioral healthcare networks in the country.

2. Therapist involvement in note construction (annotation). For this step, we engage with $N = 5$ of the therapists mentioned above and additionally with $N = 8$ therapists from another organization.

3. Therapist involvement in note evaluation. For this step, we worked with $N = 9$ new therapists, who were separate from the previous two groups. These nine therapists evaluated SOAP notes with the help of the rubric.

Deployment: Automated behavioral health note generation and evaluation tools in real-world settings necessitate compliance with HIPAA (Health Insurance Portability and Accountability Act) regulations and privacy-preserving AI practices. Thus, we include open-source models for both – note generation and evaluation, so as to show results for models which can be run on private compliant servers.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

David J Berghuis, L Mark Peterson, William P McInnis, and Arthur E Jongsma Jr. 2014. *The Adolescent Psychotherapy Progress Notes Planner*, volume 300. John Wiley & Sons.

Susan Cameron and Imani Turtle-Song. 2002. Learning to write case notes using the soap format. *Journal of Counseling & Development*, 80(3):286–292.

Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-level feedback generation with large language models for empowering novice peer counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4130–4161, Bangkok, Thailand. Association for Computational Linguistics.

Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E. Palacios, Derek Richards, and Anja Thieme. 2020. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. Association for Computational Linguistics.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, Ozlem Uzuner, and Majid Afshar. 2023. Progress note understanding—assessment and plan reasoning: Overview

of the 2022 n2c2 track 3 shared task. *Journal of biomedical informatics*, 142:104346.

Stephanie Greer, Danielle E. Ramo, Yin-Juei Chang, Michael Fu, Judith Tedlie Moskowitz, and Jana Haritatos. 2019. Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR mHealth and uHealth*, 7.

Barbara Griswold. 2019. How much time do we spend writing notes? Website: Navigating the Insurance Maze.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, Andrew Beam, and John Torous. 2024. Large language models in mental health care: a scoping review. *Preprint*, arXiv:2401.02984.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv e-prints*, pages arXiv–2310.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Inna Lin, Ashish Sharma, Christopher Rytting, Adam Miner, Jina Suh, and Tim Althoff. 2024. IMBUE: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–840, Bangkok, Thailand. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483, Online. Association for Computational Linguistics.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Ryan Louie, Ifdita Hasan Orney, Juan Pablo Pacheco, Raj Sanjay Shah, Emma Brunskill, and Diyi Yang. 2025. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. *arXiv preprint arXiv:2505.02428*.

Ye Luo, Bonnie L Stice, and A Stephen Lenz. 2025. Mental health apps for depression: A meta-analysis. *Journal of Counseling & Development*, 103(1):25–38.

Yash Mathur, Sanketh Rangreji, Raghav Kapoor, Medha Palavalli, Amanda Bertsch, and Matthew Gormley. 2023. SummQA at MEDIQA-chat 2023: In-context learning with GPT-4 for medical summarization. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 490–502, Toronto, Canada. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kirill Milintsevich and Navneet Agarwal. 2023. Calvados at MEDIQA-chat 2023: Improving clinical note generation with multi-task instruction fine-tuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 529–535, Toronto, Canada. Association for Computational Linguistics.

Emma O'neil, João Sedoc, Diyi Yang, Haiyi Zhu, and Lyle Ungar. 2023. Automatic reflection generation for peer-to-peer counseling. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 62–75, Singapore. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

V Podder, V Lew, and S Ghassemzadeh. 2022. Soap notes.[updated 2022 aug 29]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Michael D Reiter and Kayleigh Sabo. 2023. Writing progress notes. In *A Therapist's Guide to Writing in Psychotherapy*, pages 18–41. Routledge.

Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):590–601.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2650–2656.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023a. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Ashwyn Sharma, David Feldman, and Aneesh Jain. 2023b. Team cadence at MEDIQA-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 228–235, Toronto, Canada. Association for Computational Linguistics.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20.

Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Logan Stapleton, Jordan Taylor, Sarah Fox, Tongshuang Wu, and Haiyi Zhu. 2023. Seeing seeds beyond weeds: Green teaming generative ai for beneficial uses. *arXiv preprint arXiv:2306.03097*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, et al. 2024. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Lawrence L Weed. 1964. Medical records, patient care, and medical education. *Irish Journal of Medical Science (1926-1967)*, 39:271–282.

Samuel Woodnutt, Chris Allen, Jasmine Snowden, Matt Flynn, Simon Hall, Paula Libberton, and Francesca Purvis. 2024. Could artificial intelligence write mental health nursing care plans? *Journal of Psychiatric and Mental Health Nursing*, 31(1):79–86.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

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.

Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeff Ward. 2024. A continued pretrained llm approach for automatic medical note generation. *arXiv preprint arXiv:2403.09057*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Appendix

A Definitions of SOAP note sections

A.1 Subjective

Definition: In this section, document the subjective reports from the client, their family members, and past medical records. Include how the client describes their feelings and current symptoms.

Key Characteristics:

- **Chief Complaint:** The reason why the client is seeking therapy. Could also be a description of what symptoms the client is experiencing. **Importance:** Mandatory
- **Symptoms (as the client is talking about it):** The client's own description of their feelings, thoughts, and behaviors along with the severity. **Importance:** Mandatory
- **History:** Relevant background information, including any past medical, therapy, or behavioral issues. **Importance:** Mandatory
- **Client's Goals:** What the client hopes to achieve through therapy. **Importance:** Highly recommended
- **Homework from Previous Sessions:** Reviewing homework from the previous sessions and noting the client's compliance. **Importance:** Highly recommended
- **Quotes:** Direct quotes from the client can be particularly useful to capture their exact words and emotional tone. **Importance:** Highly recommended

A.2 Objective

Definition: This section is for recording objective observations made during the session. Note any factual, observable information, such as the client's appearance, behavior, mood, affect, and speech patterns. Avoid including any subjective statements or self-reported information from the client.

Key Characteristics:

- **Client's Observed Behavior:** The therapist's observations of the client's behavior, mood, appearance, and affect during the session. **Importance:** Mandatory
- **Mental Status:** Observations regarding the client's appearance, speech, thought processes, and orientation. **Importance:** Mandatory
- **Assessment Tools:** Results from any standardized assessments or scales used during the session. **Importance:** Highly recommended
- **Therapy Activities:** Description of specific interventions or activities conducted during the session. **Importance:** Highly recommended
- **Interventions [A]:** Applied interventions and treatment plans (MI, Cognitive Restructuring, DBT, etc.). **Importance:** Highly recommended
- **Interventions [B]:** Focus on describing active interventions provided rather than passive ones. **Importance:** Highly recommended

A.3 Assessment

Definition: In this section, integrate the subjective and objective information to provide a comprehensive analysis of the client's current condition. Summarize the clinical impressions and hypotheses regarding the client's issues.

Key Characteristics:

- **Diagnosis/Symptoms:** Any formal diagnoses made based on DSM-5 criteria or other diagnostic tools. **Importance:** Mandatory
- **Identifying Triggers:** Any triggers shown by the client.
- **Progress:** Evaluation of the client's progress toward their therapeutic goals. **Importance:** Highly recommended
- **Analysis:** The therapist's interpretation of how the client's subjective report and objective observations relate to their overall condition. **Importance:** Highly recommended
- **Response to Interventions. Importance:** Highly recommended
- **Overall/High-Level Progress. Importance:** Highly recommended
- **Treatment Goals:** Specific, measurable, achievable, relevant, and time-bound (SMART) goals for the client. Adjustments to the treatment goals. **Importance:** Highly recommended
- **Stages of Change:** For interventions like Motivational Interviewing, note the client's stage of change (Pre-contemplation, Contemplation, Action, Maintenance, etc.). **Importance:** Highly recommended

A.4 Plan

Definition: Outline the next steps for the client's treatment. Include both short-term and long-term goals, specifying what will be addressed in the next session as well as overall treatment objectives.

Key Characteristics:

- **Future Interventions:** Planned therapeutic techniques or strategies to be used in future sessions. **Importance:** Mandatory
- **Follow-Up:** Scheduling of the next session and any referrals to other professionals if needed. Note the date for the next appointment if decided upon. **Importance:** Mandatory
- **Adjustment of Medication/Intervention Choice. Importance:** Mandatory in certain circumstances
- **Homework:** Assignments or activities for the client to work on between sessions. **Importance:** Highly recommended

A.5 General Items

Key Characteristics:

- Clearly reflect that the practitioner assessed for and addressed any safety concerns (e.g., suicide risks, self-harming behaviors, homicidal ideation, etc.). **Importance:** Mandatory
- Evidence of treatment being provided in a culturally competent manner. **Importance:** Highly recommended
- **Professionalism Importance:** Highly recommended
  - Never describe other clients and staff in a derogatory manner.
  - Avoid using slang.
  - Descriptions of the patient's presenting problem should be used rather than presumptuous adjectives.

B Limitations

Real data availability : Because of the sensitive nature of behavioral health, real doctor-patient conversations are confidential. The public data we use in this study appears to be shorter and less complex than a real therapy session.

The scale of study: This study is relatively small due to the cost of recruiting licensed therapists, involving only two open-weight LLMs and human-written notes across a limited number of conversations.

Annotator bias may also affect results. While differences between human and LLM notes are clear, the gap between the two LLMs is small, and their ranking may vary across different datasets.
LLM performance: We used simple prompts in this study, focused on evaluating the framework rather than optimizing LLM performance. The results could likely improve with more advanced prompt engineering.

C Annotator Qualification and Cost

Human Note Writing: Human notes were written by the $N = 5$ internal therapists involved in the rubric design, as well as $N = 8$ external therapists. All external therapists hold either a Master's or Ph.D. degree in clinical psychology, and are licensed therapists or clinical social workers in the United States, with experience ranging from 3 to 18 years. The cost to collect each note was $206.
Note Evaluation: Human evaluation was conducted by $N = 9$ external therapists who are different from those who wrote the notes. All evaluators are licensed therapists or clinical social workers with a Master's or Ph.D. degree in a related field. The cost to collect a single human evaluation related to one note is $190.
Total cost: Annotating a large number of conversations with highly specialized experts is time-consuming and costly. The cost of collecting one note for each conversation was $206, making the cost of the dataset creation to be $10,300. We incur additional costs in the human evaluations ($190 for each, 150 evaluations total). This makes our total cost to be $38800, limiting the size of the dataset to 50 conversations.

D Human Rubric Creation Details

Figure 2 shows the interface of the tool used to build the rubric.

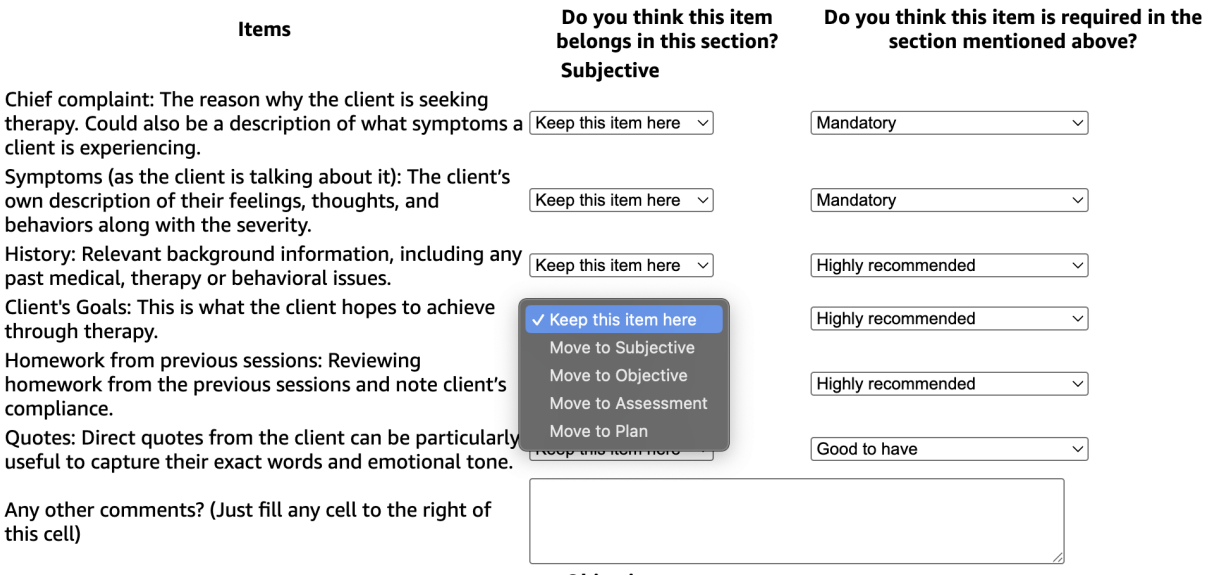

Figure 2: Rubric annotation tool. For each rubric, a therapist would read it and annotate (1) if the section is appropriate and (2) the importance level.

E Automatic Evaluation Details

E.1 Prompts for TN^A-Eval

Rubric-based Completeness Evaluation

```
Below is a behavioral therapy progress note segment. The rubric item outlines one of the necessary
components for the note. Verify if the rubric item presents in the progress note segment.

## Note Segment
{note_segment}
```

```
## Rubric Item (an item that should present in the note segment)
{rubric_item}

Does the note segment contain the rubric item? Response in [Yes, No] with no other content:
```

Rubric-based Conciseness Evaluation

```
Below is a sentence from a behavioral therapy progress note. The rubrics outlines the necessary components
for the note. Verify if the note sentence fit in one of the rubric items.

## Note Sentence
{note_sentence}

## Rubrics (a list of items that should present in the note segment)
{rubrics}

Does the note sentence fit in one of the rubric items? Response in [Yes, No] with no other content:
```

E.2 Prompts for Likert-style automatic evaluation

Completeness

```
Below is a behavioral therapy conversation along with a corresponding progress note segment. The rubrics
outline the necessary components for the note. Based on the conversation and rubrics, evaluate the
completeness of the note segment.

## Conversation
{conversation}

## Note Segment
{note_segment}

## Rubrics (a list of items that should present in the note segment)
{rubrics}

## Rating Codebook
1: The note segment is missing most of the key information from the conversation.
2: The note segment includes some important details but is significantly incomplete.
3: The note segment contains a moderate amount of important information.
4: The note segment captures most of the key information from the conversation.
5: The note segment comprehensively captures all the key information.

Using the 1 to 5 scale from the rating codebook, rate the completeness of the note segment. Output only the
rating [1, 2, 3, 4, 5]:
```

Conciseness

```
Below is a behavioral therapy conversation along with a corresponding progress note segment. The rubrics
outline the necessary components for the note. Based on the conversation and rubrics, evaluate the
conciseness of the note segment.

## Conversation
{conversation}

## Note Segment
{note_segment}

## Rubrics (a list of items that should present in the note segment)
{rubrics}

## Rating Codebook
1: The note segment includes substantial non-important information that detracts from the main points.
2: The note segment includes non-important information that needs to be reduced.
3: The note segment includes some non-important information but does not heavily detract from the main
points.
4: The note segment includes minor non-critical information.
5: The note segment includes no non-important information, making it concise and focused.

In the scale of 1 to 5, rate the conciseness of the note segment following the rating codebook. Output only
the rating [1, 2, 3, 4, 5]:
```

Faithfulness

```
Below is a behavioral therapy conversation along with a corresponding progress note segment. Verify the
faithfulness of the note segment based on the conversation.

## Conversation
{conversation}

## Note Segment
{note_segment}
```

```
## Rating Codebook
1: The note segment contains significant inaccuracies or false information.
2: The note segment contains several inaccuracies or false information.
3: The note segment may contain some inaccuracies or false information.
4: The note segment contains minor non-critical inaccuracies or false information.
5: The note segment contains no inaccuracies or false information.

In the scale of 1 to 5, rate the faithfulness of the note segment following the rating codebook. Output only
 the rating [1, 2, 3, 4, 5]:
```

F Prompt for Note Generation

```
In emotional support conversations, two primary roles exist: the therapist (individual providing support)
and the client (individual seeking support). Your task is to summarize an emotional support conversation
into client progress notes. These notes are usually in the SOAP format. The SOAP is a standardized form of
recording a client's progress. It stands for:

- Subjective: In this section, document the subjective reports from the client, their family members, and
past medical records. Include how the client describes their feelings and current symptoms.
- Objective: This section is for recording objective observations made during the session. Note any factual,
 observable information, such as the client's appearance, behavior, mood, affect, and speech patterns. Avoid
 including any subjective statements or self-reported information from the client.
- Assessment: In this section, integrate the subjective and objective information to provide a comprehensive
 analysis of the client's current condition. Summarize your clinical impressions and hypotheses regarding
the client's issues.
- Plan: Outline the next steps for the client's treatment. Include both short-term and long-term goals,
specifying what will be addressed in the next session as well as overall treatment objectives. Be clear and
specific about your expectations and the client's goals for the duration of treatment.

Output Dictionary template:
{
"Subjective": "...",
"Objective": "...",
"Assessment": "...",
"Plan": "..."
}
Generate notes for the provided conversation in the above Dictionary style template.

{Conversation}

SOAP Note:
```

G Human label distribution

Figures 3 and 4 highlight the differences in evaluation methodologies using the visualization method in Elangovan et al. (2025). Despite both methods being expert annotations, TN^{H}-Eval 's structured rubric-based approach leads to a broader distribution of scores, capturing nuances in note quality. In contrast, Likert-scale ratings tend to cluster, potentially overlooking finer distinctions.

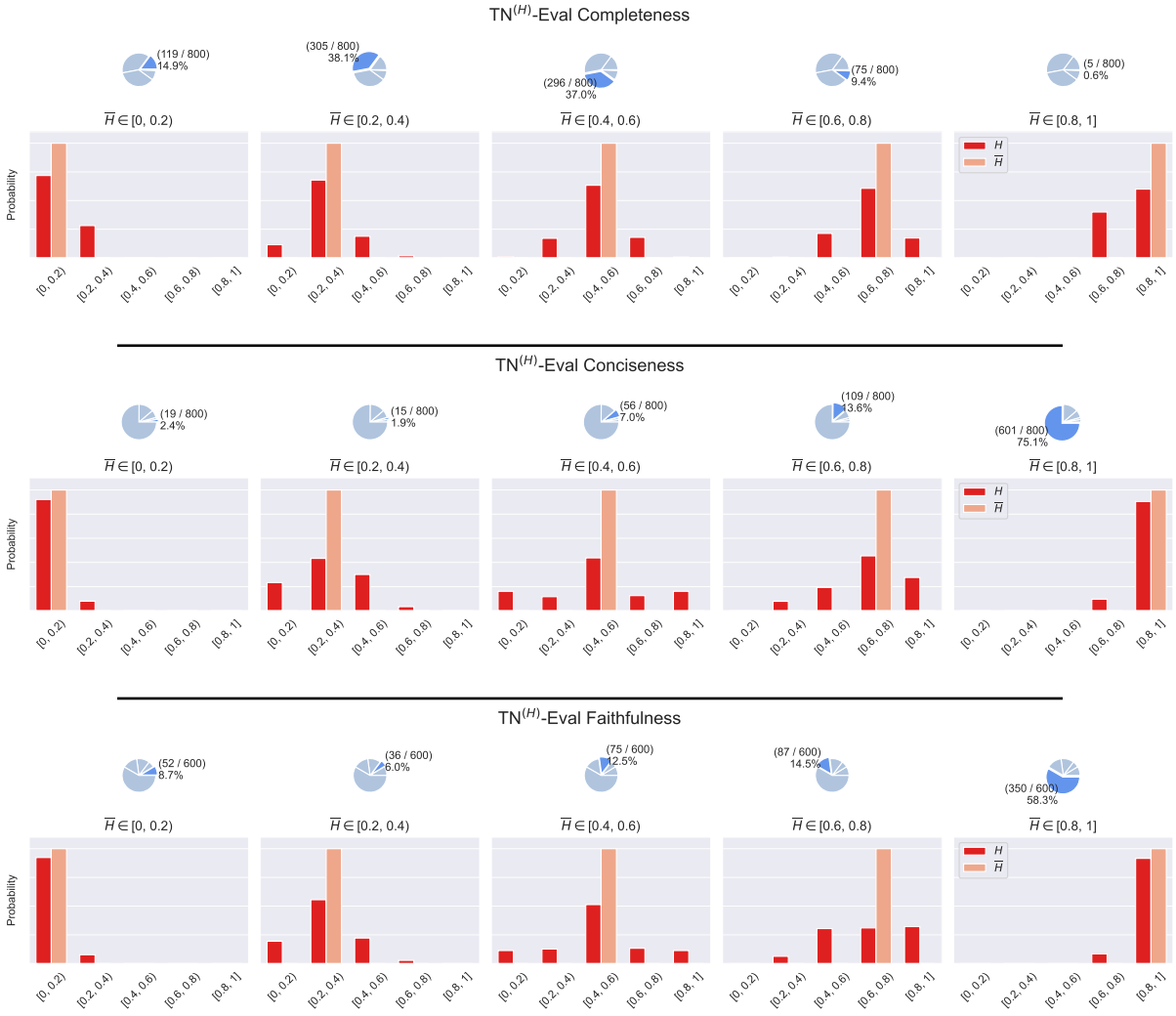

Figure 3: Human label distribution for TN^{H}-Eval annotations.

Figure 4: Human label distribution for Likert style annotations.

H Additional Results

H.1 Traditional reference-based metrics

Table 5 shows the results for traditional reference-based metrics. Notably, all values look similar, making these n-gram-based metrics insufficient to distinguish LLM performance and provide any meaningful signals for note generation quality. The primary reason is that all notes follow a similar structure, with the same section names and fairly standard sentence structure, such as "The client reports/appears ...". This structural similarity dominates the n-gram-based metric computation. Therefore, they fail to detect the nuances.

H.2 Inter-Annotator Agreement on importance of rubric items

Table 9 presents inter-annotator agreement scores among five expert annotators regarding the importance of key characteristics in therapy notes. It includes Krippendorff's alpha (α) and Fleiss' kappa (κ) for four main sections—Subjective (S), Objective (O), Assessment (A), and Plan (P)—as well as an overall agreement score. The importance of each characteristic was categorized into five levels: Mandatory, Mandatory in certain circumstances, Highly recommended, Good to have, and Optional. The high agreement scores indicate strong reliability in expert judgments, supporting the structured rubric-based evaluation framework.

	S	**O**	**A**	**P**	**Overall**
α	0.76	0.68	1.00	0.77	0.73
κ	0.63	0.67	1.00	0.61	0.68

Table 9: Inter-annotator scores among 5 experts on the *importance* of each key characteristic. S/O/A/P stands for four sections. "Importance" has 5 levels: Mandatory, Mandatory in certain circumstances, Highly recommended, Good to have, or Optional. α: krippendorff's α; κ: Fleiss' κ.

H.3 Characteristics of Therapist-Written and LLM-Generated Notes

Table 8 compares the presence of key rubric-based characteristics across therapist-written and LLM-generated notes. It highlights specific rubric items where LLM-generated notes exhibit significantly higher coverage (20% or more) than human-written notes.

H.4 Example human and LLM sourced note for the same conversation

Therapist (Human) written note

Subjective: Client is a male who is presenting to therapy for smoking. *Client reports his wife quit smoking around six months ago due to being on a health kick.* Client reports he feels pressured to quit smoking by his wife. Client reports his wife is worried about his health and continuously saves articles about the negative effects of smoking and secondhand smoke. *In addition, he feels pressure as his children feel negatively about his smoking as well.* Client stated "we used to go out and have a cigarette together at the table. Now I have to go outside in this weather. " Client reports feeling upset due to being forced to change his lifestyle habits. Client also feels torn as he enjoys smoking but does not want to lose his family over it. *Client inquired about his wife possibly attending his next session.*

Objective: Client expressed frustration, anger, and reports feeling isolated due to his smoking. *Client reports his smoking is impacting his family life and his physical health.*

Assessment: Client appears to be in the contemplation stage of change. Although client is upset regarding his cigarette smoking, he appears to see the consequences of his smoking. *In addition, he also appears to have an understanding of consequences regarding losing his family if he continued to smoke.*

Plan: Client will follow up with wife to see her schedule in the hopes that she can attend next session with him. Client will also work on reducing his cigarette use.

Note generated by Llama 3.1 (70B)

Subjective: Client reports being forced to attend therapy by his wife due to her concerns about his smoking. He expresses feelings of anger, frustration, and resentment towards his wife and children, who he feels are ganging up on him. He also mentions concerns about the impact of his smoking on his family, particularly his wife. He does not see his smoking as a major problem, but rather as a habit he has had all his life.

Objective: Client appears resistant and defensive, with a tone of frustration and anger. He maintains a somewhat confrontational demeanor, but also shows signs of vulnerability and concern for his relationship with his wife.

assessment: Based on the client's report and observable behavior, it appears that he is struggling with feelings of resentment and anger towards his wife and children, who he feels are pressuring him to quit smoking. He also seems to be motivated by a desire to maintain his relationship with his wife and avoid conflict. It is likely that he is experiencing some ambivalence about quitting smoking, and may benefit from a gradual approach to cessation.

Plan: Short-term goals include exploring the client's motivations and concerns about quitting smoking, and developing a gradual plan for cessation. It may be helpful to involve the client's wife in future sessions to address her concerns and provide a unified approach to supporting the client's quit attempt. *Long-term goals include reducing the client's symptoms of anger and frustration, improving his relationship with his wife and children, and increasing his overall well-being.*

Table 10: Visualized sentences that are considered not concise in human and Llama notes.

H.5 Automatic evaluation scores for different note sources and evaluators

		Completeness		Conciseness		Faithfulness	
Evaluator	**Note Source**	**TN^A-Eval**	**Likert**	**TN^A-Eval**	**Likert**	**TN^A-Eval**	**Likert**
Mistral Large V2	Human	15.0 (±9.1)	2.23 (±0.27)	73.7 (±15.1)	3.65 (±0.53)	73.2 (±14.9)	4.64 (±0.39)
	Claude 3 Sonnet	21.7 (±6.5)	2.67 (±0.32)	93.6 (±7.8)	4.00 (±0.23)	74.0 (±10.1)	4.99 (±0.05)
	Claude 3 Haiku	21.7 (±6.6)	2.84 (±0.33)	94.4 (±6.8)	3.92 (±0.22)	69.9 (±10.1)	4.92 (±0.21)
	Llama 3.1 (70B)	21.0 (±5.4)	2.53 (±0.25)	92.3 (±7.7)	3.73 (±0.32)	70.2 (±11.5)	4.95 (±0.17)
	Llama 3.1 (8B)	22.0 (±6.9)	2.64 (±0.29)	91.3 (±9.0)	3.56 (±0.27)	69.0 (±11.6)	4.64 (±0.43)
	Mistral Large V2	23.1 (±6.5)	2.92 (±0.35)	92.8 (±7.2)	3.97 (±0.21)	75.8 (±8.8)	4.99 (±0.05)
	Mistral 7B	21.4 (±6.6)	3.00 (±0.34)	90.3 (±6.4)	4.03 (±0.20)	75.2 (±9.5)	5.00 (±0.00)

Table 11: TN-Eval and Likert-style automatic evaluation. We show the results using Mistral Large V2 as the evaluator. Note that the TN-Eval faithfulness is not LLM-based metric, instead it uses AlignScore.

Evaluator	Note Source	Completeness		Conciseness		Faithfulness	
		TN^{A}-Eval	**Likert**	**TN^{A}-Eval**	**Likert**	**TN^{A}-Eval**	**Likert**
Llama 3.1 (70B)	Human	19.7 (±11.1)	1.77 (±0.33)	74.8 (±15.3)	4.68 (±0.40)	73.2 (±14.9)	4.63 (±0.50)
	Claude 3 Sonnet	25.0 (±7.2)	2.25 (±0.33)	92.9 (±8.4)	4.93 (±0.13)	74.0 (±10.1)	5.00 (±0.00)
	Claude 3 Haiku	26.9 (±7.0)	2.56 (±0.38)	93.4 (±7.3)	4.93 (±0.12)	69.9 (±10.1)	4.93 (±0.24)
	Llama 3.1 (70B)	24.3 (±6.5)	2.19 (±0.28)	92.3 (±6.8)	4.86 (±0.22)	70.2 (±11.5)	4.91 (±0.19)
	Llama 3.1 (8B)	25.6 (±7.8)	2.38 (±0.35)	92.0 (±8.5)	4.63 (±0.46)	69.0 (±11.6)	4.67 (±0.46)
	Mistral Large V2	28.0 (±7.3)	2.46 (±0.38)	92.8 (±5.5)	4.92 (±0.15)	75.8 (±8.8)	4.99 (±0.06)
	Mistral 7B	27.8 (±6.7)	2.65 (±0.45)	91.2 (±6.2)	4.93 (±0.14)	75.2 (±9.5)	4.98 (±0.09)

Table 12: TN-Eval and Likert-style automatic evaluation. We show the results using Llama 3.1 (70B) as the evaluator. Note that the TN-Eval faithfulness is not LLM-based metric, instead it uses AlignScore.

I Workflow Integration Proposal

Below, we outline detailed integration steps for embedding the TN-Eval framework within clinical workflows:

1. **Session Completion:** Therapists conduct standard therapy sessions, optionally recording or leveraging speech-to-text tools integrated with the Electronic Health Record (EHR). We propose using HIPAA-certified tools for this task to ensure client privacy.

2. **Note Creation:** After completing a session, therapists either write a note from scratch or receive an initial AI-generated SOAP note draft, which they review and edit in the EHR interface. For AI-generated notes, therapists review and manually edit auto-generated drafts within the EHR interface, making necessary adjustments for accuracy and clinical appropriateness. These notes can be in the EHR provider's preferred format.

3. **TN^{A}-Eval Quality Assessment:** The TN^{A}-Eval framework evaluates the edited note in real-time within the EHR, scoring completeness, conciseness, and faithfulness, while providing rubric-aligned actionable feedback.

4. **Verification and Final Submission:** Therapists review the TN^{A}-Eval quality scorecard and address highlighted concerns before formally submitting notes to the EHR, maintaining final responsibility and clinical oversight.