AI-Generated Responses to Sensitive Pesantren Cases for Musyrif/Musyrifah Counseling Communication Training: A Snapshot Study

Hengki Tri Hidayatullah[1], Arbin Janu Setiyowati[2], Muslihati[3]
Guidance and Counseling Study Program, Faculty of Education,
Universitas Negeri Malang, Indonesia[1]
Guidance and Counseling Study Program, Faculty of Education,
Universitas Negeri Malang, Indonesia[2]
Guidance and Counseling Study Program, Faculty of Education,
Universitas Negeri Malang, Indonesia[3]
E-mail: hengki.tri.2401118@students.um.ac.id[1], arbin.janu.fip@ac.id[2],
muslihati.fip@um.ac.id[3]
Correspondent Author: Arbin Janu Setiyowati, arbin.janu.fip@um.ac.id
Doi: 10.31316/g-couns.v10i02.8646

Abstrack

Pesantren increasingly face psychosocial issues such as bullying and harassment, requiring sensitive interventions. Musyrif/musyrifah often serve as first responders but lack formal counseling or Psychological First Aid (PFA) training. This snapshot study examines the feasibility of large language models (LLMs) as simulators for developing counseling communication skills in pesantren contexts. Five LLMs: Gemini, DeepSeek, ChatGPT, Claude, and Meta AI, were tested using a culturally relevant case scenario. Their counseling responses were evaluated through a seven-skill rubric: attending, empathy, reflection, clarification, paraphrasing, summarizing, and nonverbal sensitivity. Five raters independently scored the outputs. Gemini achieved the highest mean (M = 4.89), followed by DeepSeek (M = 4.71) and ChatGPT-5 (M = 4.57). Claude (M = 3.86) and Meta AI (M = 3.20) scored lower, especially in nonverbal sensitivity. The study concludes that advanced LLMs, particularly Gemini and DeepSeek, show strong potential as scalable, culturally adaptive tools to strengthen counseling readiness in pesantren.
Keywords: AI-assisted counseling, psychological first aid counseling communication, pesantren mental health, safe learning environment

Abstrak

Pesantren semakin menghadapi masalah psikososial seperti bullying dan pelecehan, yang membutuhkan intervensi sensitif. Musyrif/musyrifah sering berfungsi sebagai responden pertama tetapi tidak memiliki konseling formal atau pelatihan Pertolongan Pertama Psikologis (PFA). Studi snapshot ini mengkaji kelayakan model bahasa besar (LLM) sebagai simulator untuk mengembangkan keterampilan komunikasi konseling dalam konteks pesantren. Lima LLM: Gemini, DeepSeek, ChatGPT, Claude, dan Meta AI diuji menggunakan skenario kasus yang relevan secara budaya. Tanggapan konseling mereka dievaluasi melalui rubrik tujuh keterampilan: menghadiri, empati, refleksi, klarifikasi, parafrase, meringkas, dan kepekaan nonverbal. Lima penilai secara independen menilai hasil. Gemini mencapai rata-rata tertinggi (M = 4,89), diikuti oleh DeepSeek (M = 4,71) dan ChatGPT-5 (M = 4,57). Claude (M = 3,86) dan Meta AI (M = 3,20) mendapat skor lebih rendah, terutama dalam sensitivitas nonverbal. Studi ini menyimpulkan bahwa LLM tingkat lanjut, khususnya Gemini dan DeepSeek, menunjukkan potensi yang kuat sebagai alat yang dapat diskalakan dan adaptif secara budaya untuk memperkuat kesiapan konseling di pesantren.
Kata kunci: konseling berbantuan AI, komunikasi konseling pertolongan pertama psikologis, pesantren kesehatan mental, lingkungan belajar yang aman

Article info
Received September 2025, Revised November 2025, Accepted December 2025, Published January 2026

Published by: Guidance and Counseling Study Program
Faculty of Teacher Training and Education
Universitas PGRI Yogyakarta

INTRODUCTION

Pesantren, as both traditional and modern Islamic educational institutions, serve not only as centers for the transmission of religious knowledge but also as social entities that shape the character, identity, and psychological resilience of their students(Ardianto & Ansori, 2024; Cibro et al., 2023; Rahmatullah, 2021). In recent years, pesantren have increasingly come under media scrutiny due to the emergence of sensitive cases, ranging from bullying to harassment, raising serious questions about the adequacy of protection systems and psychosocial support within these institutions (Fatmawati et al., 2024; Khusumadewi, 2022). Such public attention highlights that pesantren face contemporary challenges that extend beyond the domain of religious curricula, encompassing students' emotional and psychological well-being.

The image of pesantren as educational institutions that uphold moral values and noble character has inevitably been affected by these cases (Fatmawati et al., 2024; Nurtawab & Wahyudi, 2022). While segments of society perceive these problems as undermining the social legitimacy of pesantren as safe and civilized learning environments, the majority still maintain that such incidents are isolated acts committed by individuals rather than a reflection of pesantren culture as a whole. This tension between negative public perceptions and the enduring trust of the broader community creates a discursive space that calls for strengthening internal systems within pesantren to ensure their continued relevance and credibility in the public sphere (Wulandari, 2020; Zainal et al., 2022).

The primary challenge faced by pesantren in addressing psychosocial problems lies in the limited availability of professional human resources in psychology and counseling (Attarwiyah et al., 2025; Kirsh & Gewurtz, 2012). The disproportionate ratio between the large number of students and the scarcity of counselors or psychologists makes it difficult to adequately and sustainably respond to sensitive issues. Inadequate handling of such problems may result in more serious psychological consequences for students, including trauma, reduced learning motivation, and the loss of a sense of safety within the educational environment (Arifin et al., 2025; O'Neill, 2010).

Musyrif and musyrifah, who serve as daily companions for students, ultimately become the frontline in dealing with these problems. They often act as the "first listeners" for students experiencing difficulties, despite most of them lacking formal training in psychology or counseling. As a result, their responses tend to take the form of moral lectures or normative advice rather than empathetic listening supported by basic counseling skills such as attending, reflection, or clarification (Kawakip, 2020; Dougherty et al., 2015). This practice illustrates a significant competence gap between the ideal role expected of musyrif/musyrifah and the actual capacity they possess.

This competency gap has serious implications for students' learning experiences and daily lives in pesantren. Members of Generation Z and Generation Alpha, who currently constitute the majority of the student population, are characterized by open and expressive communication styles, along with a strong need for safe spaces to share emotional experiences (Chan & Lee, 2023; Danilova, 2023; Sakdiyakorn et al., 2021; Zahra et al., 2025). When their psychosocial needs are left unaddressed, pesantren risk being perceived as unsupportive environments, potentially undermining their effectiveness as institutions of moral, spiritual, and social development.

Access to basic counseling training, particularly in the form of Psychological First Aid (PFA), remains highly limited for musyrif/musyrifah as well as ustadz/ustadzah. A culturally grounded PFA framework is needed to equip educators with practical skills to

provide initial support to students facing emotional crises. The limited availability of PFA training in pesantren further underscores the urgency of developing a competency model that is adaptive, applicable, and aligned with the cultural foundations of Islamic education.

The potential for innovation in enhancing the competencies of musyrif and musyrifah can be directed through the integration of Artificial Intelligence (AI). AI has become a prominent trend in global academia and education, particularly as a medium for simulation, assessment, and case-based learning (Marimekala & Lamb, 2024). Through prompt engineering, AI can be guided to respond to counseling scenarios in accordance with established rubrics of basic counseling skills. This technology offers the opportunity to use AI as a consistent, measurable, and accessible training partner for musyrif and musyrifah to strengthen their practical counseling skills. However, no previous studies have directly compared the counseling communication quality of multiple LLMs within pesantren contexts.

This research gap highlights the need to examine how AI can serve as a case-based counseling training tool in sensitive pesantren contexts. This study aims to (1) compare the counseling communication competence among five LLMs, and (2) identify the most suitable AI for use in musyrif/musyrifah counseling training modules. Theoretically, this study contributes to the growing field of AI-based counselor training; practically, it supports pesantren leadership in enhancing PFA-based communication

METHOD

This study employed a snapshot prospective design to evaluate the effectiveness and accuracy of large language models (LLMs) in responding to sensitive pesantren counseling cases. The primary objective was to assess the quality of AI-generated responses in terms of basic counseling skills, proportionality of response length, and the ability to construct an atmosphere through culturally appropriate nonverbal cues. Each case was presented once to every model without repetition, clarification, or iterative interaction, thereby simulating the real conditions in which musyrif or musyrifah may encounter spontaneous consultations requiring immediate responses.

The sensitive scenario used in this study concerned a senior student engaging in behavior with sexual undertones, expressed both verbally and nonverbally, toward a junior student**.** This situation was chosen because it reflects a highly relevant and delicate issue in pesantren counseling practice, where counselees often present with discomfort, anxiety, fear, and uncertainty about whether to disclose their experiences. The case was considered appropriate for testing the extent to which LLMs could generate responses that demonstrate basic counseling skills and cultural sensitivity in addressing high-risk, sensitive issues.

Five large language models were selected to generate responses: DeepSeek**,** Claude Sonnet 4 (Anthropic)**,** Meta AI**,** ChatGPT-5 (OpenAI)**,** and Gemini 2.5 Pro (Google)**.** The selection of five LLMs was based on their accessibility, popularity, and representativeness of current AI architectures. All models received the same case prompt, formulated in counseling terminology adapted to the pesantren context. The responses were then evaluated using an AI-based automated rating system through prompt engineering. The prompt-engineered rubric was reviewed by three counseling experts to ensure content validity. This system, termed the Evaluator of Basic Counseling Skills, was designed in JSON format and functioned as an automatic rater with strict instructions for scoring, word count calculation, and the detection of cultural sensitivity in nonverbal scene-setting.

In this mechanism, the same five LLMs also functioned as evaluators. All responses were anonymized and coded so that the evaluator LLMs could not identify the generating model. The evaluation prompt instructed each evaluator to assign scores from 1 to 5 across eight dimensions: attending, empathy, reflection, clarification, paraphrasing, summarizing, nonverbal atmosphere, and word count (Table 1) (Ahmad et al., 2025; Ilyas & Chalidaziah, 2022; Violin & Basuki, 2024). A strict scoring policy was applied, with maximum scores awarded only if all key indicators were fulfilled, including the accurate description of professional, culturally sensitive, and ethically appropriate nonverbal cues. Word count was scored separately to monitor the proportionality of response length, with the ideal range set between 51 and 120 words. For benchmarking, each case was accompanied by a criterion standard response representing evidence-based counseling practices and pesantren cultural norms, which evaluators used as a reference for comparison.

Table 1.
The Dimensions of Evaluation

Aspect	Assessment Indicators (Score 1–5)	Special Focus
Attending	Score 1: no attention, off-topic. Score 3: mentions the main issue but inconsistently. Score 5: full attention, emotional & cultural details captured.	Focus on attentiveness, relevance, and cultural sensitivity.
Empathy	Score 1: no empathy, rigid. Score 3: shallow empathy with generic phrases. Score 5: strong empathy, warmth, validation of emotions + culture.	Focus on warm tone, emotional validation, and non-judgmental stance.
Reflection	Score 1: no reflection. Score 3: partial reflection, missing the core meaning. Score 5: accurate reflection, capturing the essence of emotions & content.	Focus on ability to capture the core of feelings & content.
Clarification	Score 1: absent/creates confusion. Score 3: general/superficial question. Score 5: open-ended, precise, culturally sensitive questions that clarify intent.	Focus on exploratory questioning & clarity.
Paraphrasing	Score 1: no paraphrase. Score 3: partial paraphrase. Score 5: concise, accurate, contextually aligned with emotional & cultural aspects.	Focus on restating core content in different words.
Summarizing	Score 1: no summary. Score 3: partial/unclear summary. Score 5: concise, coherent summary integrating content + emotions.	Focus on coherence & completeness of the summary.
Atmosphere & Nonverbal	Score 1: no scene-setting. Score 3: 1–2 general gestures (e.g., soft tone). Score 5: ≥3 relevant nonverbal cues, culturally sensitive (e.g., sitting sideways, professional smile, respectful distance).	Focus on scene-setting & nonverbal communication in line with pesantren etiquette.

Word Count	Score 1: <10 words (very brief). Score 3: 26–50 words (adequate). Score 5: 81–120 words (elaborative, focused, coherent).	Focus on proportionality of response length.

The evaluation data were analyzed descriptively by calculating the mean, median, minimum, maximum, standard deviation, standard error, and coefficient of variation. Reliability across AI evaluators was examined using Cronbach's alpha and the intraclass correlation coefficient, while correlations between evaluators' scores were analyzed using Pearson's *r* and Spearman's rho. Differences in the average scores among models were tested using Friedman's test, Kruskal–Wallis test, and the Wilcoxon signed-rank test. All analyses were conducted using IBM SPSS version 29.0, with the Exact Tests module based on Monte Carlo simulation. The significance level was set at $\alpha = 0.05$ ($p \leq 0.05$).

RESULT AND DISCUSSION

Result

This study evaluated the counseling responses generated by five large language models (LLMs): DeepSeek, ChatGPT-5, Claude Sonnet 4, Gemini 2.5 Pro, and Meta AI, based on a sensitive pesantren counseling scenario. Each response was scored across eight dimensions: attending, empathy, reflection, clarification, paraphrasing, summarizing, nonverbal atmosphere, and word count. The results are presented in three layers: descriptive summaries, reliability tests, and comparative analyses.

Table 2.
Comparative Evaluation Scores of LLM Responses

Evaluator	Model	A	E	R	C	P	S	NV	N= (35)
ChatGPT Plus	Gemini 2.5 Pro	5	5	5	5	5	5	5	35
	DeepSeek v3.1	5	5	5	5	5	5	3	33
	ChatGPT-5	5	5	4	5	4	5	4	32
	Claude Sonnet 4	5	4	4	4	5	5	2	29
	Meta AI	4	4	4	4	4	3	2	25
DeepSeek	DeepSeek v3.1	5	5	5	4	5	4	4	32
	ChatGPT-5	4	4	4	3	4	4	3	28
	Claude Sonnet 4	4	4	3	2	4	4	1	22
	Meta AI	3	3	2	4	3	1	1	17
	Gemini 2.5 Pro	5	5	5	5	5	5	5	35
Claude	DeepSeek v3.1	5	5	5	5	5	5	2	32
	Gemini 2.5 Pro	4	5	5	4	5	5	3	31
	Claude Sonnet 4	4	4	4	4	5	5	1	27
	Meta AI	2	3	3	3	3	1	1	16
Gemini	ChatGPT-5	5	5	5	5	5	5	5	35
	Gemini 2.5 Pro	5	5	5	5	5	5	5	35
	DeepSeek v3.1	5	5	5	5	5	5	4	34
	Claude Sonnet 4	5	5	5	4	5	5	1	30
	Meta AI	4	3	4	3	3	1	1	19
Meta	ChatGPT-5	5	5	5	5	5	5	5	35
	DeepSeek v3.1	5	5	5	5	5	5	5	35
	Gemini 2.5 Pro	5	5	5	5	5	5	5	35
	Claude Sonnet 4	5	5	5	5	5	5	4	34

Meta AI	4	4	4	5	4	3	3	27

Notes: A = Attending; E = Empathy; R = Reflection; C = Clarification; P = Paraphrasing; S = Summarizing; NV = Nonverbal

Table 2 presents the comparative evaluation scores for five large language models (LLMs): Gemini 2.5 Pro, DeepSeek v3.1, ChatGPT-5, Claude Sonnet 4, and Meta AI, based on assessments by five independent evaluators (ChatGPT Plus, DeepSeek, Claude, Gemini, and Meta). Each response was scored on seven core counseling skills: attending (A), empathy (E), reflection (R), clarification (C), paraphrasing (P), summarizing (S), and nonverbal sensitivity (NV), with a maximum possible score of 35.

Across evaluators, Gemini 2.5 Pro consistently achieved the highest scores, with several evaluators awarding perfect scores of 35/35 (ChatGPT Plus, DeepSeek, Gemini, and Meta) and the lowest being 31/35 (Claude evaluator). This indicates a strong consensus that Gemini's superior performance in embodying counseling skills, particularly in empathy, reflection, and summarizing.

DeepSeek v3.1 also showed robust performance, receiving scores ranging from 32/35 (evaluators DeepSeek and Claude) to a near-perfect 35/35 (Meta evaluator). While DeepSeek's performance was consistently strong across cognitive and affective skills, a recurrent pattern of slightly reduced nonverbal sensitivity (scores of 2-4) prevented it from achieving the same level of perfection as Gemini. ChatGPT-5 demonstrated stable but slightly lower performance**,** with evaluator scores ranging between 28/35 (DeepSeek evaluator) and 35/35 (Gemini and Meta evaluators). Most evaluations placed ChatGPT-5 in the 32–35 range, suggesting that the model is reliable but somewhat less nuanced than Gemini or DeepSeek, particularly in aspects such as reflection and clarification. Claude Sonnet 4 showed more variation across evaluators. Scores ranged from a low of 16/35 (Meta evaluator) to a high of 30/35 (Gemini evaluator). Strengths were consistently observed in paraphrasing and summarizing, while major weaknesses appeared in nonverbal sensitivity (often rated 1–2), indicating a limited ability to simulate contextual cues relevant to counseling communication.

Meta AI received the lowest overall scores, ranging from 16/35 (Claude evaluator) to 27/35 (Meta evaluator). Meta AI occasionally demonstrated adequate attention and clarification, but consistently underperformed in summarization and nonverbal sensitivity, producing an uneven, weaker profile compared to the other models. Taken together, the results highlight a clear hierarchy of performance: Gemini 2.5 Pro and DeepSeek v3.1 emerged as the strongest models**,** consistently rated highly by multiple evaluators, while ChatGPT-5 occupied an intermediate position with generally reliable but less nuanced responses. In contrast, Claude Sonnet 4 and Meta AI were evaluated as weaker models**,** limited by low nonverbal expression and inconsistent performance across evaluators. This pattern indicates that only certain advanced LLMs currently meet the requirements for effective counseling simulation in the pesantren context.

Table 3.
Comparative Evaluation Scores of LLM Responses

Model AI	A	E	R	C	P	S	NV	Total Score	Mean Score
Gemini 2.5 Pro	4.8	5.0	5.0	4.8	5.0	5.0	4.6	34.2 / 35	4.89
DeepSeek v3.1	5.0	5.0	5.0	4.8	5.0	4.8	3.6	33.2 / 35	4.74
ChatGPT-5	4.8	4.8	4.4	4.6	4.6	4.8	4.4	32.4 / 35	4.63
Claude Sonnet 4	4.4	4.4	4.0	3.8	4.6	4.6	1.2	27.0 / 35	3.86
Meta AI	3.4	3.6	3.2	4.0	3.6	2.6	2.0	22.4 / 35	3.20

Notes: A = Attending; E = Empathy; R = Reflection; C = Clarification; P = Paraphrasing; S = Summarizing; NV = Nonverbal

To reduce evaluator-specific variation and present a holistic picture, the scores were averaged across all five evaluators. Table 3 reports the mean results for each counseling skill and the aggregated total scores. Gemini 2.5 Pro achieved the highest aggregate score (34.2/35; M = 4.89), showing near-perfect competence in empathy, reflection, paraphrasing, and summarizing (all means = 5.0), and a strong performance in nonverbal sensitivity (mean = 4.6). DeepSeek v3.1 ranked second (33.2/35; M = 4.74), with perfect means in attending, empathy, and reflection, and high scores in clarification and paraphrasing (means = 4.8 each), though nonverbal sensitivity was lower (mean = 3.6). ChatGPT-5 placed third (32.4/35; M = 4.63), showing balanced competence across domains but somewhat weaker reflection and clarification. Claude Sonnet 4 scored 27.0/35 (M = 3.86), moderate in paraphrasing and summarizing (means = 4.6 each) but weaker in attending, empathy, and reflection (means = 4.0–4.4). Its lowest performance was in nonverbal sensitivity (mean = 1.2). Meta AI obtained the lowest aggregate score (22.4/35; M = 3.20), with slightly better clarification (mean = 4.0) but poor empathy, reflection, summarizing, and nonverbal sensitivity. These results establish a hierarchy: Gemini and DeepSeek as top performers, ChatGPT as a strong intermediate, and Claude and Meta AI as weaker models with limited training utility

Table 4.
Descriptive statistics of LLM

Model AI	Mean	SD	Min	Max	SE	CV (%)
Gemini 2.5 Pro	4.89	0.35	3	5	0.13	7.1
DeepSeek v3.1	4.71	0.76	3	5	0.29	16.2
ChatGPT-5	4.57	0.53	4	5	0.20	11.6
Claude Sonnet 4	3.86	1.46	1	5	0.55	37.8
Meta AI	3.86	0.64	3	5	0.24	16.6

Table 4 provides descriptive statistics of LLM performance, including mean, standard deviation (SD), minimum, maximum, standard error (SE), and coefficient of variation (CV%). These indicators highlight not only overall performance but also consistency across evaluators. Gemini 2.5 Pro recorded the highest mean score (M = 4.89) with very low variability (SD = 0.35; CV = 7.1%), confirming its superiority as both the strongest and most stable model. Four evaluators gave perfect scores, while one evaluator rated slightly lower on nonverbal sensitivity, creating minimal variance. DeepSeek v3.1 followed with M = 4.71 (SD = 0.76; CV = 16.2%), showing strong competence but more variability across evaluators, particularly in nonverbal aspects. ChatGPT-5 obtained M = 4.57 (SD = 0.53; CV = 11.6%), reflecting dependable performance with limited variation, though consistently below Gemini and DeepSeek.

Claude Sonnet 4 showed the greatest inconsistency, with M = 3.86, but SD = 1.46 and CV = 37.8%. The wide spread of scores (range 1–5) highlights evaluator disagreement, particularly due to very low nonverbal ratings. Meta AI also averaged M = 3.86, but with smaller variability (SD = 0.64; CV = 16.6%), reflecting consistently modest performance rather than erratic ratings. This analysis emphasizes that stability is as critical as high mean scores**.** Gemini emerges as both the highest-performing and most consistent model, DeepSeek and ChatGPT are strong but moderately variable, while Claude and Meta AI are weaker—Claude due to inconsistency and Meta AI due to consistently low performance. Figure 1 shows the total score comparison across models, while Figure 2 illustrates a radar chart of performance on each counseling skill dimension.

These visualizations highlight Gemini's superior consistency and the relative weaknesses of Claude and Meta AI.

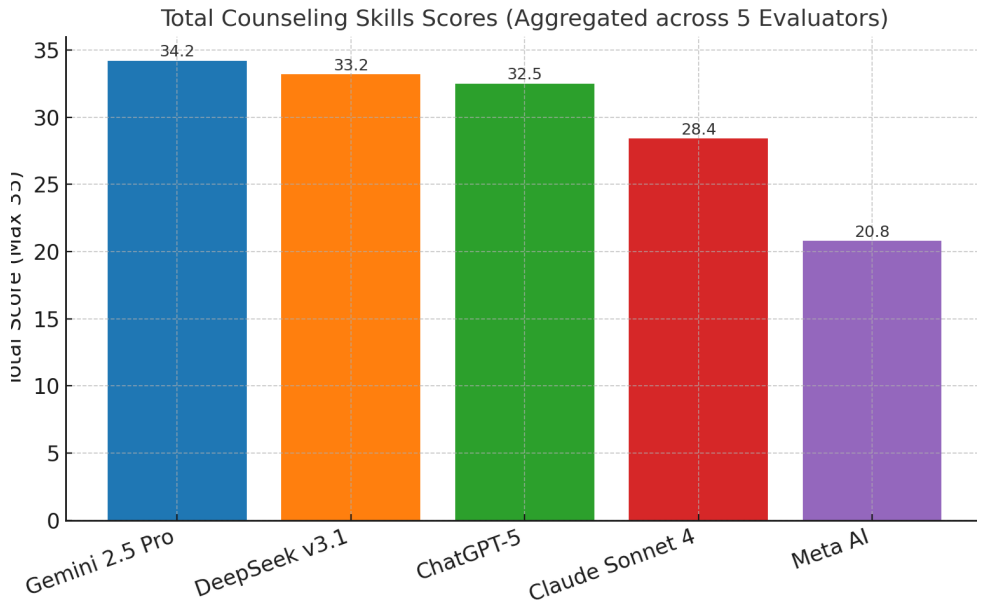

Figure 1. Total Counseling Communication Score by LLMs (DeepSeek, ChatGPT, Gemini and Meta AI)

Figure 1 presents the total scores of counseling skill performance across the five LLMs. Gemini achieved the highest score (34.2/35), indicating an outstanding, comprehensive performance across all assessed counseling skills. DeepSeek followed closely with a total of 33.2, demonstrating strong consistency with only minor limitations in nonverbal elements. ChatGPT scored 32.5, indicating balanced performance across most aspects, though reflection and nonverbal skills were somewhat less emphasized. Claude and Meta AI both scored lower (20.8 each), with Claude showing relative strength in paraphrasing and summarizing, while Meta AI performed better in clarification but weaker in summarizing and nonverbal support.

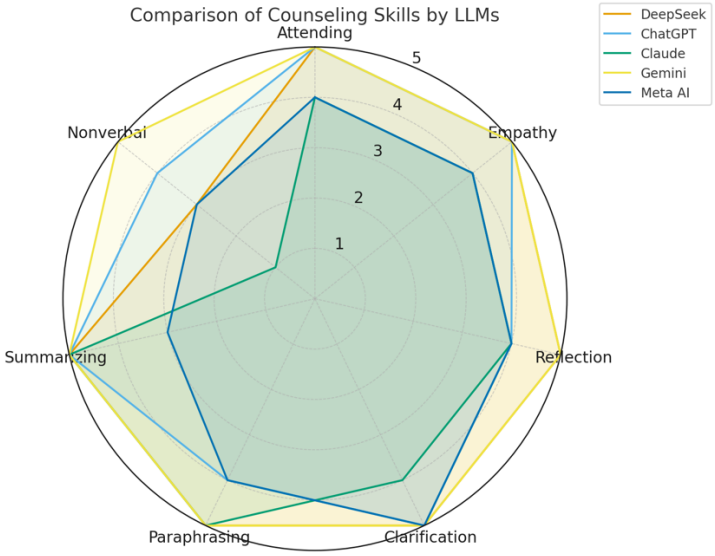

Figure 2. Comparison of seven Counseling Communication Skill dimension

Figure 2 presents a radar chart comparing the five LLMs across the seven skill dimensions. The figure illustrates Gemini's dominance, achieving the highest score across all categories, highlighting its suitability as a model for training in counseling communication. DeepSeek also performed consistently across most skills, though its nonverbal indicators were less detailed. ChatGPT shows a nearly balanced profile but did not reach the same depth in reflection. Claude's profile is uneven, excelling in paraphrasing and summarizing but showing deficiencies in nonverbal sensitivity. Meta AI exhibits the most modest profile, with noticeable weaknesses in summarizing and nonverbal expression, despite performing adequately in attending and clarification.

These visual comparisons reinforce the conclusion that Gemini and DeepSeek are the most reliable LLMs for use as training simulators for musyrif/musyrifah in pesantren, while ChatGPT can serve as a supportive baseline. Claude and Meta AI, in contrast, are less suitable as primary training tools but remain useful as comparative references

Reliability of Evaluations

The AI-based evaluator's reliability was tested to confirm scoring consistency. Cronbach's alpha was 0.87**,** indicating strong internal consistency across the seven dimensions. The intraclass correlation coefficient (ICC) using a two-way random-effects model with absolute agreement was 0.84 (95% CI: 0.79–0.89, $p < 0.001$)**,** demonstrating excellent inter-rater reliability. Pairwise correlation analyses further supported these findings. Pearson's r values ranged from 0.71 to 0.89 ($p < 0.01$)**,** while Spearman's rho ranged from 0.68 to 0.85 ($p < 0.01$)**,** confirming robust agreement across evaluators. Together, these results validate the consistency and dependability of the AI-based scoring system.

Non-parametric comparative tests confirmed significant differences among the five models. The Friedman test indicated a significant difference in model rankings across the skill dimensions ($\chi^2 = 18.42$, df = 4, $p = 0.001$). The Kruskal–Wallis test yielded consistent results ($H = 16.37$, df = 4, $p = 0.003$), suggesting unequal performance distributions. Post-hoc analyses with the Wilcoxon signed-rank test (Bonferroni-adjusted) revealed that Gemini significantly outperformed Claude ($p = 0.003$) and Meta AI ($p = 0.002$)**.** No significant differences were found between DeepSeek and ChatGPT ($p > 0.05$), indicating comparable mid- to high-level performance.

The results establish a clear performance hierarchy among the five LLMs. Gemini emerged as the strongest model, excelling across all counseling skill dimensions, particularly in cultural appropriateness and precise nonverbal scene-setting. DeepSeek followed closely, with slight limitations in nonverbal expressiveness. ChatGPT provided stable mid-level performance suitable for baseline training. Claude, while strong in paraphrasing and summarizing, lacked cultural sensitivity and nonverbal realism. Meta AI, though occasionally clear in its clarifications, underperformed in empathy, summarization, and nonverbal sensitivity.

Discussion

The findings of this study provide several important insights into the feasibility of using large language models (LLMs) as training simulators for musyrif/musyrifah in pesantren contexts. Three major themes emerge: (a) model performance hierarchy, (b) skill-specific strengths and limitations, and (c) implications for counseling education and training. First, a clear hierarchy of performance among the evaluated LLMs was established. Gemini 2.5 Pro consistently achieved the highest ratings across evaluators and aggregated results, with a mean score of 4.89 and minimal variability. DeepSeek v3.1

also demonstrated strong performance (M = 4.71), although with slightly higher variability, particularly in nonverbal sensitivity. ChatGPT-5 provided a stable intermediate outcome (M = 4.57), indicating that it can serve as a reliable but less refined option compared to Gemini and DeepSeek. By contrast, Claude (Sonnet 4) and Meta AI showed substantial limitations: Claude displayed high inconsistency across evaluators, while Meta AI performed consistently but modestly. This hierarchy underscores that only the most advanced LLMs currently possess the capability to simulate counseling communication at a level useful for structured training.

Second, the analysis highlights the skill-specific strengths and weaknesses of different models. Gemini and DeepSeek excelled in core verbal counseling skills such as attending, empathy, reflection, and summarizing, indicating their capacity to generate coherent, empathic, and contextually sensitive responses. However, both models showed relatively lower ratings in nonverbal sensitivity, reflecting the inherent challenge of simulating embodied gestures and cultural cues through text-based outputs. ChatGPT-5 demonstrated balanced but less nuanced performance, with moderate weaknesses in reflection and clarification. Claude Sonnet 4, while effective in paraphrasing and summarizing, consistently failed in nonverbal dimensions, suggesting a lack of contextual richness in its generated responses. Meta AI performed poorly across most domains, particularly in empathy and summarizing, which are critical for effective counseling interactions. These differences align with previous literature that stresses the complexity of translating counseling microskills into computational outputs (Skovholt & Rønnestad, 2003; Sue & Sue, 2016).

Third, the findings carry direct implications for counseling education in pesantren. The role of musyrif/musyrifah as daily mentors positions them to respond to sensitive issues among santri, despite their limited background in psychology or counseling. In this context, AI-driven training simulators offer a scalable and accessible alternative to traditional training programs, which are often constrained by limited access to professional counselors and psychologists in pesantren. Gemini and DeepSeek, in particular, could be integrated into training modules as consistent and high-quality response generators, providing musyrif/musyrifah with opportunities to practice attending, empathetic listening, and clarifying skills. ChatGPT-5 may serve as a supplementary tool for practice, while Claude and Meta AI may be more useful as comparative references to illustrate suboptimal responses.

An additional implication is the importance of stability and reliability in model outputs. As highlighted in the descriptive statistics (Table 3), Gemini achieved the highest mean scores and the lowest variability, ensuring consistent training experiences across evaluators. In contrast, Claude's high variability (CV = 37.8%) suggests that its outputs are less predictable, potentially undermining its role in standardized training. This reinforces the argument that consistency is as important as overall competence when integrating AI into counselor training programs.

Finally, the results open pathways for developing an AI-assisted training framework tailored to the pesantren context. Such a framework could combine case-based simulations with prompt-engineered evaluation rubrics based on counseling microskills. Musyrif/musyrifah could engage in structured dialogues with AI models, receive immediate feedback on their responses, and build foundational skills in Psychological First Aid (PFA) adapted to pesantren culture. This aligns with broader trends in digital education that emphasize interactive, adaptive, and context-sensitive learning tools (UNESCO, 2023). By positioning AI as both a training partner and an evaluative

instrument (Beghetto et al., 2024; Rangarajan et al., 2024; Sapci & Sapci, 2020), pesantren could enhance their capacity to address psychosocial challenges among santri, thereby reinforcing their legitimacy as holistic educational institutions (Cibro et al., 2023; Nikmatullah et al., 2023; Zainal et al., 2022).

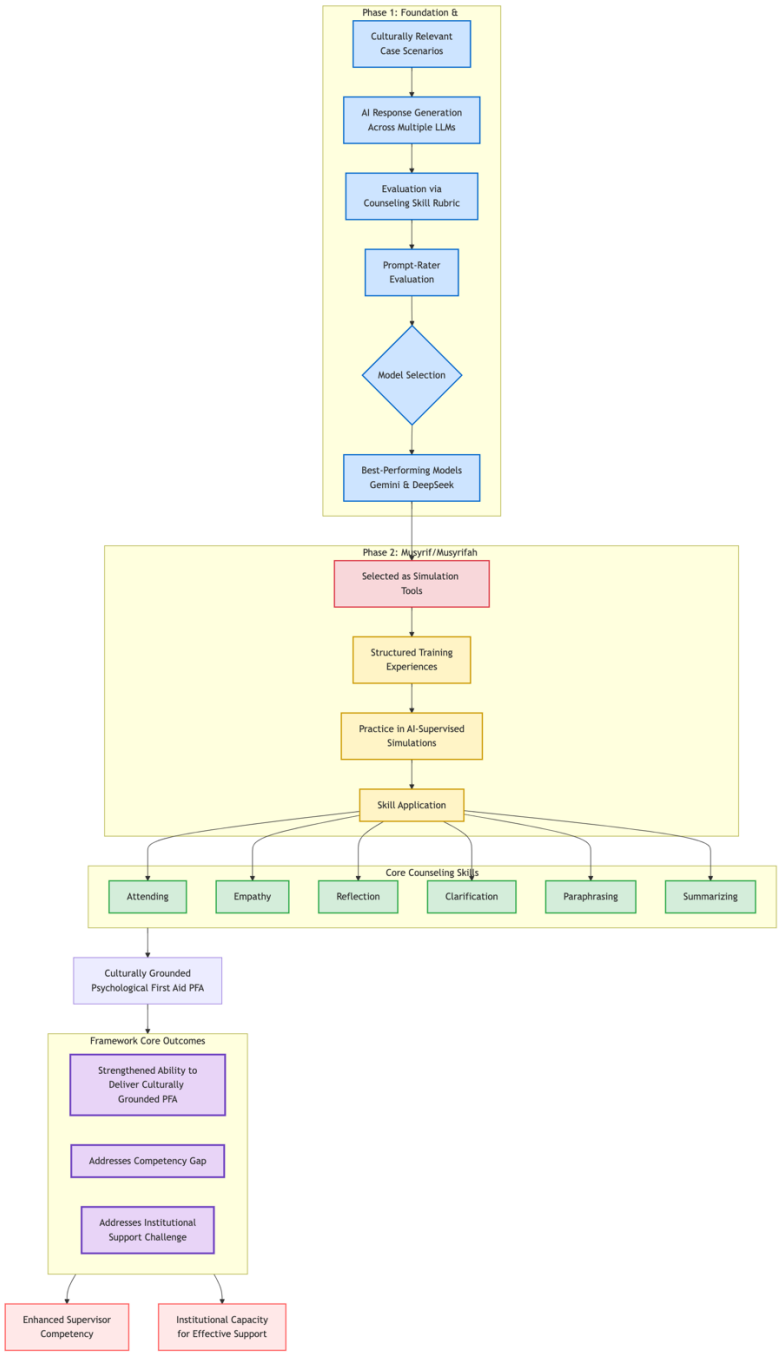

Figure 3. Framework For AI-Assisted Counseling Training In Pesantren

Building on these findings, the study proposes a framework for AI-assisted counseling training in pesantren (Figure 3). The framework begins with culturally relevant case scenarios, followed by AI-generated responses across multiple LLMs. Responses are evaluated using the counseling skill rubric via prompt-rater. The best-

performing models (Gemini and DeepSeek) are selected as simulation tools, providing structured training experiences for musyrif/musyrifah. Through these simulations, supervisors practice attending, empathy, reflection, clarification, paraphrasing, and summarizing, ultimately strengthening their ability to deliver culturally grounded PFA. The framework thus addresses both the competency gap among musyrif/musyrifah and the institutional challenge of providing effective psychosocial support in pesantren.

This study has several limitations that should be acknowledged. First, the evaluation was conducted using a single case scenario of a sensitive pesantren issue. Although the scenario was culturally relevant, broader generalization requires testing multiple case variations, including diverse psychosocial contexts such as academic stress, peer conflict, and grief. Second, the evaluation design was cross-sectional, relying on one-time model outputs. This limits the ability to capture longitudinal changes in model performance, especially as LLMs are frequently updated. Third, the assessment relied on prompt-engineered rubrics for counseling microskills, which, although structured, may not fully capture the complexity of human communication dynamics, such as tone, affect, and embodied presence. Finally, the study did not include direct feedback from human counselors or musyrif/musyrifah, which would have strengthened ecological validity.

Future research should expand its scope by incorporating multiple scenarios that represent a broader range of issues encountered in pesantren life. Comparative studies across languages (e.g., Indonesian, Arabic, English) could provide insight into the linguistic-cultural adaptability of LLMs. Longitudinal designs are recommended to track model updates and consistency over time. Furthermore, participatory research involving musyrif/musyrifah in co-designing AI-assisted training modules could enhance cultural grounding and practical applicability. Finally, integrating multimodal AI that combines text, voice, and gesture recognition could address the current gap in nonverbal sensitivity, offering a more holistic simulation of counseling interactions.

CONCLUSION

This study set out to analyze the responses of five LLMs, ChatGPT, Claude, DeepSeek, Gemini, and Meta AI, towards a sensitive pesantren case using prompt-engineering evaluation based on core counseling skills. The findings reveal substantial differences in quality across models. Gemini achieved perfect scores across all assessed skills, demonstrating high levels of empathy, reflection, clarification, and nonverbal sensitivity. DeepSeek followed closely with similarly strong performance, while ChatGPT provided stable but less nuanced responses. In contrast, Claude and Meta AI yielded lower scores, with limited emotional depth and cultural sensitivity. Based on these results, Gemini is recommended as the primary AI model for counseling training simulations in pesantren, with DeepSeek as a strong alternative and ChatGPT as a supportive baseline. Claude and Meta AI may serve as comparative references but are not ideal as primary training tools. This conclusion directly addresses the study's objective by identifying the most reliable AI models for enhancing musyrif/musyrifah competencies in providing culturally grounded Psychological First Aid (PFA). Future work should expand case scenarios, involve human expert raters, and implement pilot training programs to further validate and refine this framework in real pesantren contexts

ACKNOWLEDGMENT

This research was supported by Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM) – Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik

Indonesia. Grant Scheme for Higher Education Research (PTM). The authors also gratefully acknowledge the institutional support provided by Universitas Negeri Malang, which facilitated the implementation of this study.

RFERENCES

Ahmad, M., Bakar, A. N. I., Rani, M. N. H., Mohd Rusdin, N., Mustafa, M. B., & Abd Mukti, T. (2025). Enhancing Communication Skills For Integrating Spiritual And Religious Elements In Marriage And Family Counseling In Malaysia. International Journal Of Advanced And Applied Sciences, 12(2), 150–157. Https://Doi.Org/10.21833/Ijaas.2025.02.017

Ardianto, R. A., & Ansori, I. (2024). Dinamika Radikalisme Di Pesantren: Tinjauan Terhadap Isu Dan Tantangan. Tsaqofah. Https://Doi.Org/10.58578/Tsaqofah.V4i1.2533

Arifin, S., Yohandi, & As'ad. (2025). Konseling Berbasis Pesantren Untuk Meningkatkan Kesehatan Mental Dan Kesejahteraan Psikologis Santriwati Baru. Hisbah: Jurnal Bimbingan Konseling Dan Dakwah Islam. Https://Doi.Org/10.14421/Hisbah.2024.212-09

Attarwiyah, N. M., Chotib, M., & Subakri, S. (2025). Spiritual Leadership And Mental Wellbeing: The Role Of Kiai In Maintaining Santri Mental Health. Qalamuna: Jurnal Pendidikan, Sosial, Dan Agama. Https://Doi.Org/10.37680/Qalamuna.V17i1.6395

Beghetto, R., Ross, W., Karwowski, M., & Glăveanu, V. (2024). Partnering With Ai For Instrument Development: Possibilities And Pitfalls. New Ideas In Psychology. Https://Doi.Org/10.1016/J.Newideapsych.2024.101121

Cibro, A. N., Salminawati, S., & Usiono, U. (2023). Modern Pesantren: The Politics Of Islamic Education And Problems Of Muslim Identity. Al Qalam: Jurnal Ilmiah Keagamaan Dan Kemasyarakatan. Https://Doi.Org/10.35931/Aq.V17i2.1956

Dougherty, M. A., Curtis, R., & Robertson, P. (2015). Boundary Issues In School Counseling: Managing Role Conflicts In School Counseling. In Boundary Issues In Counseling: Multiple Roles And Responsibilities: Third Edition (Pp. 208–213). American Counseling Association. Https://Doi.Org/10.1002/9781119221586.Ch9

Fatmawati, Z., Barir, B., Sarmin, S., & Putra, K. W. R. (2024). Education On The Impact Of Bullying On The Mental Health Of Adolescents In Pesantren Village, Tembelang District, Jombang Regency. Journal Of Indonesian Public Health Service. Https://Doi.Org/10.60050/Jiphs.V1i2.44

Ilyas, S. M., & Chalidaziah, W. (2022). Culture Based Counseling Communication Skill Development Module. Jurnal Neo Konseling, 4(4), 6. Https://Doi.Org/10.24036/00685kons2022

Kawakip, A. (2020). Globalization And Islamic Educational Challenges: Views From East Javanese Pesantren. Ulumuna. Https://Doi.Org/10.20414/Ujis.V24i1.385

Khusumadewi, A. (2022). Identification Of Student (Santri) Problems On Islamic Boarding School (Pondok Pesantren). Proceedings Of The International Joint Conference On Arts And Humanities 2021 (Ijcah 2021). Https://Doi.Org/10.2991/Assehr.K.211223.173

Kirsh, B., & Gewurtz, R. (2012). Promoting Mental Health Within Workplaces. In Handbook Of Occupational Health And Wellness (Pp. 243–265). Springer Us. Https://Doi.Org/10.1007/978-1-4614-4839-6_12

Nikmatullah, C., Wahyudin, W., Tarihoran, N., & Fauzi, A. (2023). Digital Pesantren: Revitalization Of The Islamic Education System In The Disruptive Era. Al-Izzah: Jurnal Hasil-Hasil Penelitian. Https://Doi.Org/10.31332/Ai.V0i0.5880

Nurtawab, E., & Wahyudi, D. (2022). Restructuring Traditional Islamic Education In Indonesia: Challenges For Pesantren Institution. Studia Islamika. Https://Doi.Org/10.36712/Sdi.V29i1.17414

O'neill, L. K. (2010). Mental Health Support In Northern Communities: Reviewing Issues On Isolated Practice And Secondary Trauma. Rural And Remote Health, 10(2), 1369. Https://Www.Scopus.Com/Inward/Record.Uri?Eid=2-S2.0-77958171416&Partnerid=40&Md5=Aca4bb61c43c3c78614bbd91269e7bd8

Rahmatullah, A. S. (2021). Kyai's Psychological Resilience In The Perspective Of Pesantren: Lesson From Indonesia. Jurnal Pendidikan Islam. Https://Doi.Org/10.14421/Jpi.2021.102.235-254

Rangarajan, K., Manivannan, V. V., Singh, H., Gupta, A., Maheshwari, H., Gogoi, R., Gogoi, D., Das, R. J., Hari, S., Vyas, S., Sharma, R., Pandey, S., Seenu, V., Banerjee, S., Namboodiri, V., & Arora, C. (2024). Simulation Training In Mammography With Ai-Generated Images: A Multireader Study. European Radiology. Https://Doi.Org/10.1007/S00330-024-11005-X

Sapci, A., & Sapci, H. (2020). Artificial Intelligence Education And Tools For Medical And Health Informatics Students: Systematic Review. Jmir Medical Education, 6. Https://Doi.Org/10.2196/19285

Violin, A. F. T., & Basuki, A. (2024). Communication Skills And Apprehension In Individual Counseling Practices. Konselor, 13(1), 29–40. Https://Doi.Org/10.24036/0202413265-0-86

Wulandari, A. (2020). Pendidikan Islam Berasaskan Moderasi Agamadi Pondok Pesantren Nurul Ummahat Kotagede Yogyakarta. Https://Consensus.App/Papers/Pendidikan-Islam-Berasaskan-Moderasi-Agamadi-Pondok-Wulandari/0ef6c7f8885c5362a7862815029987f1/

Zainal, S., Prasetyo, M. A. M., & Yaacob, C. M. A. (2022). Adopting Pesantren-Based Junior High School Programs: The Pesantren Change Its Educational System Without Conflict. Jurnal Ilmiah Islam Futura. Https://Doi.Org/10.22373/Jiif.V22i2.13525