When Less Is More: An Empirical Study of Minimal Responses in Counseling Dialogues and the Behavior of LLMs

Zhiyang Qi
The University of Tokyo
zhiyangqi@g.ecc.u-tokyo.ac.jp

Abstract

In psychological counseling, effective support is not always delivered through long, information-rich responses. Minimal responses, such as backchannel cues and concise empathic statements, help convey attentive listening, express empathy, and encourage clients to continue expressing themselves. However, existing counseling dialogue systems and evaluation frameworks often favor explicit, content-rich replies, overlooking the interactional value of brief counselor utterances. This paper presents a systematic cross-lingual analysis of minimal responses across multiple counseling dialogue datasets. We develop a two-stage filtering method based on utterance length and content, followed by contextual verification using a large language model (LLM). Our analysis shows that minimal responses are common in human-collected datasets but substantially underrepresented in LLM-generated ones. We further evaluate current LLMs in manually curated dialogue contexts where human counselors used minimal responses. The results show that strong commercial LLMs are capable of generating minimal responses when explicitly instructed, but still struggle to determine when such responses are appropriate. Counseling-specific models trained on synthetic data perform particularly poorly, tending instead to produce longer and more information-rich responses. Moreover, LLM-based response-quality evaluation may undervalue minimal responses, even when they are interactionally appropriate.

1 Introduction

Mental health issues have attracted increasing attention in recent years (WHO, 2022; Arias et al., 2022), yet many individuals still lack timely access to professional support due to the shortage of counselors and the high cost of services. This has motivated growing interest in large language models (LLMs) as virtual counselors (Liu et al., 2023; Qi et al., 2025b; Chen et al., 2023), making high-quality counseling dialogue data essential for both training and evaluation.

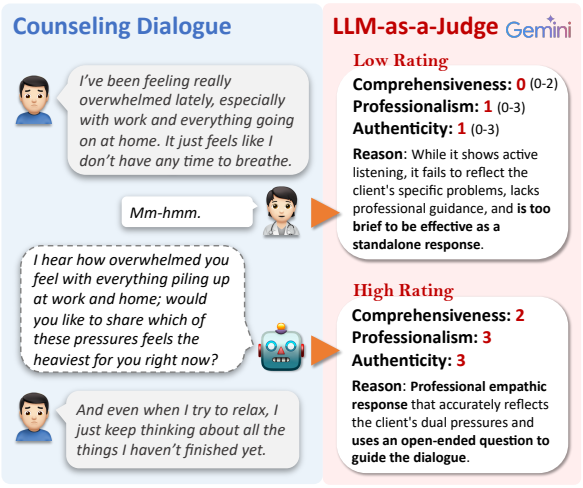

Figure 1: Minimal responses are a key technique in real counseling. However, current LLM-based evaluation methods (here using Zhang et al. (2024)'s prompt with Gemini) tend to undervalue such responses due to their brevity, favoring more elaborate replies.

Collecting real counseling dialogues, however, is challenging because they contain sensitive and private information. Existing public datasets therefore rely on indirect collection strategies, such as transcribing public counseling videos (AnnoMI (Wu et al., 2022)), collecting role-play dialogues with trained counselors (KokoroChat (Qi et al., 2025a)), or anonymizing online counseling conversations by rewriting client utterances (PsyDial (Qiu and Lan, 2025)). Since such human-collected resources remain limited in scale, recent work has increasingly turned to LLM-synthesized counseling datasets, including SoulChat (Chen et al., 2023), SmileChat (Qiu et al., 2024), and CACTUS (Lee et al., 2024). In parallel, several evaluation frameworks have been proposed for counseling response generation, typically assessing dimensions such as informative-

ness, professionalism, realism, and safety (Zhang et al., 2024; Zhao et al., 2024).

While these datasets and evaluation methods have advanced counseling dialogue research, they often emphasize informative and structurally complete responses. This focus risks overlooking **minimal responses**, brief but interactionally important counselor utterances that play a central role in counseling practice. Examples such as "uh-huh" and "I see" are considered essential micro-skills: although they convey little propositional content, they express empathy, acceptance, and attentiveness without interrupting the client's narrative, thereby encouraging further self-disclosure (Ivey et al., 2017; Hill, 2020). Rather than providing information or direction, minimal responses function as continuers (Sacks et al., 1974; Schegloff, 1982), facilitating client expression and supporting the therapeutic alliance in client-centered counseling (Rogers, 1957).

In contrast, we observe that counselor utterances in some LLM-generated counseling datasets often follow fixed, information-dense patterns, such as expressing empathy followed by a question[1], as illustrated in Figure 1. Although questions can help advance counseling conversations, their immediate effects are mixed, and frequent questioning may be associated with lower perceived empathy, helpfulness, and session smoothness (Williams, 2023). Such response patterns may be inherited by models fine-tuned on synthetic data, leading them to produce verbose or templated responses. Moreover, utterance-level LLM-based evaluation may further reinforce this tendency by favoring informative and well-structured responses over brief but interactionally meaningful minimal responses.

Motivated by the above observations, we investigate minimal responses in psychological counseling dialogues and make two contributions:

- **Cross-lingual and cross-dataset analysis.** We systematically analyze minimal responses across seven human-collected and LLM-generated counseling dialogue datasets in Chinese, Japanese, and English. To support this analysis, we develop a two-stage filtering pipeline that combines rule-based screening based on utterance length and content with LLM-based contextual verification. The results show that minimal responses are substantially more common in human-collected datasets than in LLM-generated ones.
- **Multi-model evaluation in minimal-response contexts.** Using manually curated dialogue contexts in which human counselors used minimal responses, we systematically evaluate multiple models across languages. We find that (1) models trained on human counseling data are more likely to generate minimal responses than those trained on synthetic data; (2) even strong commercial LLMs can generate minimal responses when explicitly instructed, but still struggle to determine when they are appropriate; and (3) LLM-based response-quality evaluation may favor more content-rich responses and undervalue interactionally appropriate minimal responses.

These findings highlight the need to consider minimal responses in counseling data construction, model development, and evaluation.

2 Dataset Analysis

Since minimal responses in existing counseling dialogue datasets have not been systematically examined, this section analyzes seven datasets spanning Chinese, Japanese, and English, as summarized in Table 1.[2] They include four LLM-generated datasets, CACTUS (Lee et al., 2024), CPsyCounD (Zhang et al., 2024), PsyDTCorpus (Xie et al., 2025), and SmileChat (Qiu et al., 2024), and three human-collected datasets, PsyDial-D4 (Qiu and Lan, 2025), AnnoMI (Wu et al., 2022), and KokoroChat (Qi et al., 2025a). Although conventional backchannels such as "uh-huh" and "I see" are prototypical minimal responses, brief empathic or reflective utterances such as "that sounds difficult" can serve a similar role by acknowledging the client's experience without introducing a new question or topic. We therefore include both types in our operational definition of minimal responses.

To identify such responses more comprehensively, we adopt a two-stage procedure that combines rule-based filtering with LLM-based contextual verification. First, we apply rule-based filtering to counselor utterances using length thresholds and predefined keyword lists. Because minimal responses often encourage clients to continue speaking and tend to occur within longer client narratives, we also require the immediately following client

[1]For example, according to our analysis, 43.9% of the 24,220 counselor utterances across 3,084 dialogues in CPsyCounD (Zhang et al., 2024) end with a question mark. In PsyDTCorpus (Xie et al., 2025), the proportion is 67.4% among 86,054 counselor utterances across 4,760 dialogues.

[2]For datasets containing consecutive utterances from the same speaker, adjacent utterances were merged before calculating statistics and conducting the analysis.

Dataset	Lang.	#Dial.	#Couns.	Avg. Utter. per Dialogue	#Counselor Utterances	(1) Rule-based Filtering	(2) LLM-based Contextual Verification		
						Rule-filtered Candidates (# / %)	Minimal Responses		Other Short Responses (# / %)
							Backchannel-like Responses (# / %)	Brief Empathic/ Reflective Responses (# / %)	
LLM-generated datasets									
CACTUS (Lee et al., 2024)	EN	31,577	–	30.53	491,304	4 (0.00%)	0 (0.00%)	3 (0.00%)	1 (0.00%)
CPsyCounD (Zhang et al., 2024)	ZH	3,084	–	15.94	24,220	37 (0.15%)	0 (0.00%)	1 (0.00%)	36 (0.15%)
PsyDTCorpus (Xie et al., 2025)	ZH	4,760	–	36.16	86,054	1 (0.00%)	0 (0.00%)	1 (0.00%)	0 (0.00%)
SmileChat (Qiu et al., 2024)	ZH	55,165	–	11.38	309,698	183 (0.06%)	1 (0.00%)	68 (0.02%)	114 (0.04%)
Human-collected datasets									
AnnoMI (Wu et al., 2022)	EN	133	N/R	72.64	4,859	1,184 (24.37%)	**877 (18.05%)**	**128 (2.63%)**	179 (3.68%)
PsyDial-D4 (Qiu and Lan, 2025)	ZH	2,382	40	75.59	90,031	4,741 (5.27%)	**861 (0.96%)**	**2,778 (3.09%)**	1,102 (1.22%)
KokoroChat (Qi et al., 2025a)	JA	6,589	424	71.63	237,735	12,161 (5.12%)	**7,555 (3.18%)**	**3,747 (1.58%)**	859 (0.36%)

Table 1: Dataset statistics and short-response distributions obtained using the two-stage identification pipeline. In Step (1), candidates are identified through rule-based filtering based on utterance length and content. In Step (2), GPT-5.4-mini verifies each candidate using its dialogue context and classifies it into one of three categories: backchannel-like responses, brief empathic/reflective responses, or other short responses. In this study, the first two categories are collectively considered minimal responses and are highlighted in gray. Percentages are calculated relative to the total number of counselor utterances. "N/R" indicates that the number of counselors was not reported, whereas "–" indicates that the number is not applicable to LLM-generated datasets.

utterance to exceed a predefined length threshold. Specifically, counselor utterances are limited to 15 characters in Chinese and Japanese and 10 words in English, while the following client utterance must contain at least 10 characters in Chinese and Japanese or 5 words in English. These relatively permissive thresholds are intended to reduce false negatives. We then apply a content-based filter using predefined keyword lists to remove short questions, conversation-advancing prompts, directives or suggestions, and utterances such as thanks, apologies, greetings, polite closings, and simple encouragements, whose primary function is not to facilitate further client expression. The keyword lists are shown in Figure 2, and the full filtering criteria are provided in Appendix A.

Second, to improve precision, we use GPT-5.4-mini[3] to verify each remaining candidate based on its dialogue context. The model classifies each candidate into one of three categories: (1) backchannel-like responses, (2) brief empathic or reflective responses, and (3) other short responses. The classification prompt is shown in Figure 3 in Appendix A. We further manually inspected 100 randomly sampled instances from each category in each of the three human-collected datasets, and all category-level validation rates exceeded 94%, supporting the reliability of the classification procedure.

Table 1 presents the results of this two-stage procedure. Minimal responses are almost absent from the LLM-generated datasets but occur substantially more often in all three human-collected datasets. AnnoMI shows the highest proportion, possibly because it is transcribed from face-to-face counseling, where interactive feedback is more frequent. By contrast, PsyDial-D4 and KokoroChat consist of text-based online counseling dialogues and contain lower proportions of minimal responses.

3 Experiment

3.1 Experimental Setup

To examine whether current models can use minimal responses appropriately, we conduct response-generation experiments across multiple languages. We manually selected a subset of dialogue contexts in which counselors used minimal responses from the human-collected Chinese, Japanese, and English datasets. We then validated these examples using both Gemini-3.1-Flash-Lite[4] and GPT-5.4-mini to ensure that they were appropriate for the evaluation. After validation, we retained 284, 300, and 299 dialogue history–response pairs for Chinese, Japanese, and English, respectively. Figure 7 in Appendix B presents representative examples. We provide each dialogue history to the evaluated models and ask them to generate the next counselor response.

For models not specifically developed for psychological counseling, we use two prompting settings:

- **General prompt:** The model is asked to generate the next counselor response based on the

[3] https://developers.openai.com/api/docs/models/gpt-5.4-mini

[4] https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite

dialogue history. This setting evaluates whether the model naturally produces a minimal response from context, thereby testing its ability to determine when such a response is appropriate.

- **Instructional prompt:** The model is explicitly instructed to prioritize a minimal response when the client is still in the process of expressing emotions or experiences. This setting evaluates whether the model can follow an explicit instruction and generate a minimal response, thereby testing its ability to produce such responses.

We compare general-purpose open-source models such as Qwen (Yang et al., 2025) and Llama (Grattafiori et al., 2024); GPT-5.4[5], a state-of-the-art commercial model as of late April 2026; and counseling-specific models including SoulChat2.0 (Xie et al., 2025), MindChat[6], KokoroChat-Full (Qi et al., 2025a), and PsyDial-Pi4 (Qiu and Lan, 2025)[7]. We also include the original human counselor responses as a reference. Details of the models and prompts are provided in Appendix B.1.

3.2 Evaluation

To improve evaluation robustness and reduce reliance on a single model, we use both GPT-5.4-mini and Gemini-3.1-Flash-Lite as evaluators. Each generated response is assessed from three perspectives: whether it constitutes a minimal response, its likelihood of interrupting the client's continued expression on a 0–5 scale (lower is better), and its response quality. For response-quality evaluation, we adopt the prompt from prior work (Zhang et al., 2024), which assesses comprehensiveness on a 0–2 scale, professionalism and authenticity on 0–3 scales, and safety on a 0–1 scale. The overall quality score is calculated as the average of these four dimensions. By comparing interruption risk with response-quality scores, we examine whether LLM-based evaluation favors content-rich responses even when minimal responses are more appropriate. The evaluation prompts are provided in Appendix B.2.

3.3 Results

Table 2 presents the results of minimal-response generation and response-quality evaluation across three languages. We identify three main findings.

Models trained on human counseling data generate minimal responses more often than those trained on synthetic data. In Chinese, SoulChat2.0 and MindChat, both trained on synthetic counseling dialogues, achieve MR rates of only 0.00% and 0.35%, respectively, whereas the retrained PsyDial-Pi4 model reaches 12.68%. In Japanese, the retrained KokoroChat-Full model achieves 82.67%. This pattern is consistent with our earlier observation that minimal responses are nearly absent from synthetic counseling datasets but substantially more common in human-collected data, leaving synthetic-data-trained models with few opportunities to learn this behavior.

General-purpose LLMs can generate minimal responses when explicitly instructed, but do not reliably select them under standard prompting. Under general prompting, GPT-5.4 produces no minimal responses in either Chinese or Japanese, although all test contexts were manually identified as appropriate for one. Explicit instruction raises its MR rate to 97.18% in Chinese and 82.00% in Japanese. Qwen3-8B similarly increases from 0.00% to 23.59%, and Llama-3.1-Swallow from 2.67% to 34.83%. In English, the comparatively smaller Llama-3 reaches 94.98% under the instructional prompt. This may be due to the spoken-transcript nature of AnnoMI, where certain cues, such as trailing hyphens, clearly indicate that the client has not finished speaking, making it easier for the model to recognize when a minimal response is appropriate.

Existing LLM-based response-quality evaluation tends to favor content-rich responses and may undervalue appropriate minimal responses. For GPT-5.4, switching from the general to the instructional prompt reduces Quality Avg. from 2.23 to 0.41 in Chinese and from 2.12 to 0.63 in Japanese, while improving interruption scores from 3.62 to 0.10 and from 2.81 to 0.26, respectively. The same pattern appears in English, where Llama-3's Quality Avg. decreases from 1.98 to 0.73 as its interruption score improves from 3.94 to 0.18. Human responses also receive relatively low quality scores of 0.46, 0.52, and 0.68 despite near-zero interruption scores. To make this relationship more explicit, Appendix C presents scatterplots relating LLM-based quality scores to interruption scores and minimal-response rates. These results indicate that response-quality criteria reward comprehensiveness and informativeness but insufficiently cap-

[5] https://developers.openai.com/api/docs/models/gpt-5.4

[6] https://huggingface.co/X-D-Lab/MindChat-Qwen-7B-v2

[7] PsyDial-Pi4 and KokoroChat-Full were retrained following the settings of their original papers after removing all complete conversations containing any test case.

Model	Avg. Length	Response-quality Evaluation					Minimal-response Evaluation	
		Comp.	Prof.	Auth.	Safe.	Quality Avg.	MR (%)	Interrupt. ↓
Chinese dataset: PsyDial-D4 ($N = 284$)								
Qwen3-8B (general prompt)	76.15	1.69	2.21	2.33	0.99	1.81	0.00	3.54
Qwen3-8B (instructional prompt)	24.24	0.77	1.04	1.52	0.96	1.07	23.59	1.83
GPT-5.4 (general prompt)	132.05	2.00	2.96	2.97	1.00	2.23	0.00	3.62
GPT-5.4 (instructional prompt)	8.52	0.06	0.12	0.50	0.97	0.41	97.18	0.10
SoulChat2.0	52.10	1.21	1.58	1.88	0.99	1.42	0.00	2.89
MindChat	58.11	0.97	0.97	1.27	0.94	1.04	0.35	3.43
PsyDial-Pi4 (retrained)	46.81	0.94	1.25	1.67	0.99	1.22	12.68	1.85
Human	5.30	0.10	0.16	0.59	0.98	0.46	100.00	0.12
Japanese dataset: KokoroChat ($N = 300$)								
Llama-3.1-Swallow (general prompt)	41.29	1.20	1.62	1.82	0.97	1.40	2.67	2.63
Llama-3.1-Swallow (instructional prompt)	20.32	0.71	1.13	1.47	0.95	1.07	34.83	1.54
GPT-5.4 (general prompt)	66.41	1.93	2.75	2.82	0.99	2.12	0.00	2.81
GPT-5.4 (instructional prompt)	10.05	0.23	0.52	0.86	0.91	0.63	82.00	0.26
KokoroChat-Full (retrained)	12.99	0.28	0.62	0.94	0.90	0.69	82.67	0.32
Human	3.80	0.10	0.40	0.70	0.89	0.52	100.00	0.01
English dataset: AnnoMI ($N = 299$)								
Llama-3 (general prompt)	49.01	1.88	2.54	2.52	0.99	1.98	0.00	3.94
Llama-3 (instructional prompt)	2.13	0.21	0.56	1.19	0.98	0.73	94.98	0.18
Human	1.08	0.17	0.49	1.09	0.98	0.68	100.00	0.02

Table 2: Results of minimal-response generation experiments across three languages. Response quality is evaluated by comprehensiveness (Comp.), professionalism (Prof.), authenticity (Auth.), and safety (Safe.), with Quality Avg. denoting their average. MR is the percentage of minimal responses, while Interrupt. measures the likelihood of interrupting the client's continued expression on a 0–5 scale, where lower is better. LLM-based scores are averaged over GPT-5.4-mini and Gemini-3.1-Flash-Lite. Darker metric cells indicate better performance and are normalized separately for each metric. Row colors denote model categories: blue for general-purpose open-source models, green for GPT-5.4, orange and purple for models fine-tuned on synthetic and human-collected counseling data, respectively, and gray for human responses.

ture the interactional value of brief responses that encourage clients to continue speaking.

4 Related Work

Backchannels have been widely studied in conversation analysis and spoken dialogue systems that signal attention, understanding, or agreement without taking the conversational floor (Yngve, 1970; Sacks et al., 1974; Schegloff, 1982). Prior work has explored automatic backchannel prediction and generation, including when to produce a backchannel and which form to use (Ward and Tsukahara, 2000; Gravano and Hirschberg, 2011; Kawahara et al., 2016; Inoue et al., 2025), highlighting their role in making human-human communication smoother and more natural.

Minimal responses in counseling are related to backchannels, but serve more specific therapeutic functions. Brief acknowledgments and encouragers help counselors express empathy, acceptance, and attentiveness while allowing clients to continue exploring their experiences (Ivey et al., 2017; Hill, 2020), which is consistent with client-centered counseling (Rogers, 1957). Related concepts have appeared as *Grounding* and *Minimal Encouragement* in prior work (Shah et al., 2022; Li et al., 2023), but only as one of many dialogue-act or annotation labels rather than as a primary focus of study. Consequently, little is known about how minimal responses are represented in counseling dialogue datasets or whether LLM-based counseling systems can generate them appropriately. This work addresses this gap by analyzing their distribution and evaluating their generation across models and prompting strategies.

5 Conclusion

This paper examined minimal responses in psychological counseling dialogues. We found that they are common in human-collected data but rare in synthetic datasets, and that models trained on synthetic data struggle to generate them. Although general-purpose LLMs can produce minimal responses when explicitly instructed, they do not reliably determine when such responses are appropriate. Conventional response-quality evaluation also tends to undervalue minimal responses. These findings highlight the need to better incorporate minimal responses into counseling data construction, model training, and evaluation.

Acknowledgments

We sincerely thank Michimasa Inaba, Associate Professor at The University of Electro-Communications, for providing GPU resources. We also thank the anonymous reviewers for their constructive feedback. This work was supported by JST ERATO (JPMJER2502) and the MEXT Supporting Pioneering Research through AI for 1,000 Discovery Challenges Program (SPReAD), Japan, under Grant Number JPMXP1726302875.

Limitations

This study has several limitations. First, cross-dataset comparisons may be affected by differences in collection modality and counselor characteristics. AnnoMI consists of transcripts of face-to-face counseling, whereas PsyDial-D4 and KokoroChat are text-based online counseling datasets. These differences may influence conversational rhythm and the frequency of minimal responses. Variation in counselors' interaction styles and therapeutic orientations may also contribute to the observed differences across datasets.

Second, although using two LLM evaluators improves robustness, LLM-based evaluation remains a preliminary and scalable proxy rather than a substitute for expert or client evaluation. Human counselor responses provide a useful reference, but they do not establish the therapeutic appropriateness or client-perceived helpfulness of the generated responses. Future work should compare LLM judgments with evaluations from counseling experts and clients.

Finally, our experiments examine minimal responses only in selected local contexts. We do not test whether incorporating them into a complete, dynamic counseling process improves the interaction, or what timing and frequency are most effective. Excessive or repetitive use, such as repeatedly responding with "Hmm," may feel mechanical or irritating rather than supportive. Future work should evaluate when and how often minimal responses should be used over full counseling sessions, ideally with feedback from counseling experts and clients.

Ethical Considerations

This study uses existing publicly available counseling dialogue datasets in accordance with the terms of their respective licenses. To the best of our knowledge, these datasets had undergone anonymization or privacy screening before public release. We do not collect new data from real clients or deploy any system for actual counseling. Since counseling dialogues may contain sensitive information, we analyze the data only at the aggregate level and do not attempt to identify or interpret individual clients.

Our findings should not be interpreted as suggesting that LLMs can provide professional counseling. Minimal responses are only one type of counseling-related interactional behavior, and their appropriateness depends on the broader therapeutic context. This work aims to highlight an overlooked aspect of counseling dialogue modeling and to support the development of safer and more realistic counseling dialogue systems.

References

Daniel Arias, Shekhar Saxena, and Stéphane Verguet. 2022. Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine*, 54:101675.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Clara E. Hill. 2020. *Helping Skills: Facilitating Exploration, Insight, and Action*, 5th edition. American Psychological Association.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2025. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7171–7181, Albuquerque, New Mexico. Association for Computational Linguistics.

Allen E. Ivey, Mary Bradford Ivey, and Carlos P. Zalaquett. 2017. *Intentional Interviewing and Counseling: Facilitating Client Development in a Multicultural Society*, 9th edition. Cengage Learning.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *Proceedings of Interspeech 2016*, pages 2890–2894.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *Preprint*, arXiv:2309.15461.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025a. KokoroChat: A Japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12424–12443, Vienna, Austria. Association for Computational Linguistics.

Zhiyang Qi, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025b. Emostage: A framework for accurate empathetic response generation via perspective-taking and phase recognition. *arXiv preprint arXiv:2506.19279*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Carl R. Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of Turn-Taking for conversation. *Language*, 50(4):696–735.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press.

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. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.

WHO. 2022. World mental health report: Transforming mental health for all. Technical report, World Health Organization.

Elizabeth Nutt Williams. 2023. The use of questions in psychotherapy: A review of research on immediate outcomes. *Psychotherapy*, 60(3):246–254.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Daiheng Liu, Fan Zhou, and Fei Huang et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Victor H. Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–578, Chicago, IL. Chicago Linguistic Society.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13947–13966, Bangkok, Thailand. Association for Computational Linguistics.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.

A Details of Minimal-Response Identification

Before identification, adjacent utterances produced by the same speaker are merged into a single turn. Minimal responses are then identified in two stages: rule-based candidate filtering followed by LLM-based contextual classification.

In the first stage, rule-based filtering combines language-specific length thresholds with keyword- and pattern-based exclusion rules. We first retain short counselor utterances that satisfy the length constraints shown in Table 3. We also require the immediately following client utterance to exceed a language-specific minimum length. This requirement serves as a heuristic indicator that the client continued elaborating after the brief counselor response.

We then apply language-specific keyword and pattern rules to exclude short utterances whose primary functions clearly differ from those of minimal responses. As shown in Figure 2, these rules cover questions, conversation-advancing prompts, directives, suggestions, thanks, apologies, greetings, closings, service or logistical expressions, and other formulaic utterances. Although such responses may be brief, they do not primarily acknowledge the client's ongoing narrative or leave the conversational floor open for continued expression.

Language	Counselor utterance	Next client utterance
Chinese	$\leq$ 15 characters	$\geq$ 10 characters
Japanese	$\leq$ 15 characters	$\geq$ 10 characters
English	$\leq$ 10 words	$\geq$ 5 words

Table 3: Language-specific length thresholds used in rule-based candidate filtering. The counselor-utterance threshold specifies the maximum length of a candidate response, while the next-client-utterance threshold specifies the minimum length of the immediately following client turn.

In the second stage, GPT-5.4-mini contextually classifies each candidate retained after rule-based filtering. The model jointly considers the recent dialogue history, the candidate counselor utterance, and the immediately following client utterance. As shown in Figure 3, each candidate is classified into one of three categories: (1) backchannel-like responses, (2) brief empathic or reflective responses, and (3) other short responses. The first category captures brief acknowledgments that mainly signal listening and allow the client to continue speaking. The second captures concise empathic statements or reflections that add limited emotional or semantic content without asking, advising, analyzing, or redirecting. The third includes short utterances that do not serve either of these functions. The first two categories are included as minimal responses in the dataset analysis, whereas the third is excluded from the final statistics.

B Experimental Details

B.1 Models and Inference Prompts

Figure 7 shows concrete examples of the extracted minimal responses.

B.1.1 Models.

We evaluate the model variants reported in Table 2. The compared systems include general open-source models, an advanced commercial model, counseling-domain fine-tuned models, and human counselor responses. All model-based experiments are conducted in an inference-only setting; we do not update model parameters or perform additional fine-tuning. Unless otherwise specified, we use the original or default inference configurations pro-

Chinese

Category	Excluded patterns
Questions	?, 为什么, 怎么, 怎样, 如何, 什么, 哪个, 哪种, 哪里, 谁, 何时, 几点, 后来呢, 然后呢, 那之后呢, 比如, 例如, 具体, 详细说
Directives and suggestions	请, 可以试试, 你可以, 你要不要, 建议你, 不如, 最好, 应该, 试着, 说说看, 再说说
Thanks and non-listening short expressions	谢谢, 感谢, 下周, 这周, 在呢, 需要, 早, 晚安, 准时, 时间, 周, 今天
Apologies	对不起, 抱歉, 不好意思
Greetings and closings	你好, 您好, 早上好, 晚上好, 再见, 拜拜, 回头见, 辛苦了
Service expressions	不客气, 很高兴帮, 乐意帮, 很开心帮, 希望帮到, 开心
Encouragement	加油, 坚持, 祝福, 祝你, 祝您, 一切都会好, 会好起来, 没事的, 别担心

Japanese

Category	Excluded patterns
Questions	?, ですか, ますか, でしょうか, ませんか, どう, なぜ, なんで, 何, いつ, どこ, 誰, どんな, どのよう, 具体的, それで, その後, 例えば, ちなみに
Directives and suggestions	ください, ましょう, したほうが, するといい, してみて, 話してください, 教えてください, 大丈夫ですか, 平気ですか
Thanks and apologies	ありがとう, ありがとうございます, 有難う, 感謝, すみません, 申し訳, ごめんなさい, 失礼
Greetings and closings	こんにちは, こんばんは, おはよう, よろしく, お願いいたします, お願いします, お疲れ様, 失礼します, 始めます, 終了, 終わり, ではまた, またお話, またね
Service expressions and encouragement	了解, 承知, こちらこそ, 応援, 頑張, だいじょうぶ, 大丈夫ですよ, よかったです

English

Category	Excluded patterns
Questions	Question-final expressions; what, why, how, when, where, who; auxiliary-led questions beginning with do, did, does, are, is, was, were, can, could, would, will, have, has, had, shall, and should
Conversation-advancing prompts	and then, what else, for example, tell me, go on, say more, anything else, what happened, then what
Directives and suggestions	please, try, maybe, you should, you could, let's, consider, think about, remember, just, are you okay, is that okay, are you alright
Thanks and apologies	thank you, thanks, appreciate it, sorry, I apologize, apologies, excuse me
Greetings and closings	hello, hi, hey, good morning, good afternoon, good evening, take care, see you, goodbye, bye, you're welcome, have a nice day, have a good day
Service expressions and encouragement	glad to help, happy to help, here for you, feel free, let me know, good luck, you can do it, hang in there, stay strong, proud of you, hope things get better

Figure 2: Language-specific exclusion patterns used in content-based filtering. The three panels show the original exclusion lists for Chinese, Japanese, and English, respectively.

vided for each model, including decoding and generation parameters.

- **Qwen3-8B** (Yang et al., 2025) We use Qwen3-8B[8] as the general open-source baseline for Chinese. This model is used to examine whether a general-purpose instruction-following model can generate minimal responses without counseling-specific fine-tuning.
- **GPT-5.4.** We use GPT-5.4[9] as one of the strongest commercial models. It serves as an upper-bound reference for evaluating whether advanced proprietary LLMs can follow the minimal-response instruction.
- **SoulChat2.0** (Xie et al., 2025) We use SoulChat2.0-Llama-3.1-8B[10], a counseling-oriented model from the SoulChat2.0 project. This model is included to examine whether models fine-tuned on synthetic counseling data can capture minimal-response behavior.
- **MindChat** We use MindChat-Qwen-7B-v2[11], a Chinese mental-health dialogue model based on Qwen. It is also used as a representative model fine-tuned for psychological support conversations.
- **KokoroChat-Full** (Qi et al., 2025a) We use Llama-3.1-KokoroChat-Full[12], which is fine-tuned on Japanese human-collected counseling dialogues. To avoid overlap between the training and evaluation data, we remove all source dialogues containing any evaluation example and retrain the model following the original training setup, as described below.
- **Llama 3.1 Swallow** (Fujii et al., 2024; Okazaki et al., 2024) We use Llama-3.1-Swallow-8B-Instruct-v0.3[13] as the general-purpose open-source baseline for Japanese. This model is adapted for Japanese instruction following and is used to evaluate minimal-response generation without counseling-specific fine-tuning.
- **PsyDial-Pi4** (Qiu and Lan, 2025) We use PsyDial-Pi4[14], which is fine-tuned from Qwen2.5-7B-Instruct on PsyDial-D4. To avoid overlap between the training and evaluation data, we remove all source dialogues containing any evaluation example and retrain the model following the original training setup, as described below.
- **Llama-3** (Grattafiori et al., 2024) We use Meta-Llama-3-8B-Instruct[15] as the English open-source baseline. This model is included to test whether a general English instruction-following model can generate minimal responses in con-

[8] https://huggingface.co/Qwen/Qwen3-8B

[9] https://platform.openai.com/docs/models

[10] https://modelscope.cn/models/YIRONGCHEN/SoulChat2.0-Llama-3.1-8B

[11] https://huggingface.co/X-D-Lab/MindChat-Qwen-7B-v2

[12] https://huggingface.co/UEC-InabaLab/Llama-3.1-KokoroChat-Full

[13] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3

[14] https://huggingface.co/qiuhuachuan/PsyDial-Pi4

[15] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

texts extracted from AnnoMI.

B.1.2 Retraining of human-data fine-tuned models.

To prevent overlap between the training and evaluation data, we retrained the models fine-tuned on human-collected counseling dialogues after removing the complete source dialogues containing any evaluation example.

For KokoroChat-Full, the 300 evaluation examples originated from 268 unique source dialogues. We removed all of these dialogues, together with 108 dialogues from the original KokoroChat test split. The remaining 6,213 dialogues were randomly divided into 5,592 training dialogues and 621 validation dialogues using seed 42. We constructed response-generation instances from counselor turns, resulting in 192,598 training instances and 21,668 validation instances. We then fine-tuned Llama-3.1-Swallow-8B-Instruct-v0.3 using QLoRA with 4-bit NF4 quantization. We used a LoRA rank of 8, an alpha of 16, and a dropout rate of 0.05, and applied LoRA to the attention and MLP projection modules. The model was trained for one epoch on eight NVIDIA RTX A6000 GPUs with 48 GB of memory each, using a maximum sequence length of 4,096, a per-device batch size of 1, gradient accumulation of 2, and a learning rate of 1×10^{-3}.

For PsyDial-Pi4, the 300 evaluation examples originated from 273 unique source dialogues, all of which were removed before retraining. The remaining 2,109 dialogues were randomly divided into 1,898 training dialogues and 211 validation dialogues using seed 42. We fine-tuned Qwen2.5-7B-Instruct using full-parameter supervised fine-tuning with FSDP on eight NVIDIA RTX A6000 GPUs with 48 GB of memory each. The model was trained for two epochs with a maximum sequence length of 4,096, a per-device batch size of 1, gradient accumulation of 1, and a learning rate of 1×10^{-5}.

B.1.3 Inference prompts.

We use two prompt settings for the general open-source models and GPT-5.4. The general prompt asks the model to generate the next counselor response based only on the dialogue history. The instructional prompt further encourages the model to use concise minimal responses when the client is still expressing emotions or experiences. Figure 4 shows the prompt templates used for response generation.

B.2 LLM-based Evaluation Prompts

We use GPT-5.4-mini and Gemini-3.1-Flash-Lite as LLM-based evaluators, and report the average of their scores. Each generated counselor response is first evaluated from two perspectives. The evaluator determines whether the response functions as a minimal response in context and assigns an interruption-risk score from 0 to 5, where a lower score indicates a lower likelihood of interrupting the client's continued expression. The evaluator considers the dialogue history, the generated counselor response, and the immediately following client utterance, and is instructed to return only a JSON object containing these two fields. Figure 6 shows the prompts used for this evaluation.

We additionally evaluate general response quality using the response-quality evaluation prompt presented in Figure 10 of Zhang et al. (Zhang et al., 2024). Responses are scored along four dimensions: comprehensiveness on a 0–2 scale, professionalism and authenticity on 0–3 scales, and safety on a 0–1 scale. The overall quality score is calculated as the average of the four dimension scores.

To support relative rather than isolated evaluation, for each dialogue history we provide the evaluator with the responses generated by all evaluated models and ask it to score them together in a single evaluation. This allows the responses to be compared under the same conversational context and evaluation criteria. The final score for each response is averaged across GPT-5.4-mini and Gemini-3.1-Flash-Lite.

C Relationship Between LLM Quality Scores and Minimal Responses

To further examine whether general response-quality evaluation captures the appropriateness of minimal responses, we analyze the relationship between LLM-based quality scores, interruption risk, and minimal-response rates. Each point in Figure 5 represents one dataset–model condition in the generation experiment. The quality score is the average score assigned by GPT-5.4-mini and Gemini-3.1-Flash-Lite across comprehensiveness, professionalism, authenticity, and safety, following the evaluation dimensions of Zhang et al. (Zhang et al., 2024).

As shown in Figure 5, the LLM-based qual-

ity score is positively correlated with the interruption score (Pearson $r = 0.89$; Spearman $\rho = 0.89$). In other words, conditions whose responses are judged more likely to interrupt the client's continued expression tend to receive higher general response-quality ratings. Conversely, quality scores are negatively correlated with minimal-response rates (Pearson $r = -0.86$; Spearman $\rho = -0.92$), suggesting that conditions that produce minimal responses more frequently tend to receive lower ratings under this general quality rubric.

These results are consistent with our concern that general-purpose LLM-as-a-judge evaluation may undervalue minimal responses. Minimal responses are often brief and contain little explicit informational content; therefore, they may be penalized by rubrics that emphasize comprehensiveness or professional elaboration, even when they are contextually appropriate for preserving the client's narrative space.

Prompt for Contextual Classification

Task

Classify the interactional function of a candidate counselor utterance in its specific dialogue context.

Evidence

Use all three pieces of evidence together:
1. The recent dialogue history before the candidate.
2. The candidate counselor utterance itself.
3. The next client utterance after the candidate.
Judge the primary interactional function of the candidate in this specific context.

Labels

1. A backchannel-like response: A brief acknowledgment that mainly leaves the conversational floor with the client and contains almost no new semantic content. Its primary function is to signal listening, receipt, or permission for the client to continue elaborating the same concern. Context-dependent examples include "Mm-hmm.," "I see.," "Yeah.," "Okay.," "Right.," and "Sure."

2. A brief empathic or reflective response: A concise empathic statement or simple reflection that adds some emotional or semantic content but does not ask a question, offer advice, analyze the situation, or redirect the conversation. Examples include "That sounds really hard.," "You felt completely alone.," and "That was frustrating."

3. Another short response: A short utterance that does not function as either a backchannel-like response or a brief empathic or reflective response. This category includes questions, answers, advice, directives, greetings or closings, logistical expressions, analysis, topic shifts, and other substantive counselor turns.

Important Exclusions

1. Do not classify an utterance as a backchannel-like response merely because it is brief or contains an expression such as "okay," "yes," or "I see."

2. If the candidate simply answers, accepts, or confirms a question, proposal, or request in the preceding turn, classify it as another short response.

3. If the candidate concerns scheduling, homework, payment, confidentiality, session opening or closing, thanks, goodbye, or other logistical matters, classify it as another short response.

4. Use the next client utterance as contextual evidence. Continued elaboration of the same concern may support a backchannel-like response or a brief empathic or reflective response, but the classification must still be based on the candidate's own interactional function.

5. If the next client utterance instead moves to thanks, closing, scheduling, agreement, or a new topic, treat this as evidence against a backchannel-like response.

Input

Recent Dialogue History:
`{dialogue_history}`

Candidate Counselor Utterance:
`{candidate_utterance}`

Next Client Utterance:
`{next_client_utterance}`

Output

Return only compact JSON:
`{"label":"backchannel_like|brief_empathic_reflective|other_short"}`

Figure 3: Prompt used by GPT-5.4-mini to contextually classify candidates retained after rule-based filtering.

Chinese	Japanese	English
General prompt	**General prompt**	**General prompt**
你是一名专业心理咨询师。 请根据对话历史生成咨询师回复。 只输出回复，不要解释，不要添加额外内容。	あなたはプロの心理カウンセラーです。 対話履歴に基づいて、次のカウンセラーの応答を生成してください。 応答のみを出力し、説明や余計な内容は加えないでください。	You are a professional psychological counselor. Based on the dialogue history, generate the next counselor response. Output only the response, without explanation or extra text.
Instructional prompt	**Instructional prompt**	**Instructional prompt**
你是一名专业心理咨询师。 请根据对话历史生成下一句咨询师回复。 当来访者正在表达情绪或经历时，应优先仅使用简短回应（如"嗯嗯"，"我明白"，"我理解"等）来鼓励其继续表达，并且不做多余回复，把话语权交给来访者。 仅在确实需要时才提供较完整的回应。 只输出一句回复，不要解释。	あなたはプロの心理カウンセラーです。 対話履歴に基づいて、次のカウンセラーの応答を生成してください。 来談者が感情や体験を語っているときは、相手がさらに話し続けられるように、「うんうん」「はい、」「ええ、」などの短い応答のみを優先して用い、余計な返答はせず、話す主導権を来談者に委ねてください。 本当に必要な場合にのみ、やや長めの応答をしてください。 出力は一文の応答のみとし、説明は書かないでください。	You are a professional psychological counselor. Based on the dialogue history, generate the next counselor response. When the client is expressing emotions or describing experiences, you should primarily use a short minimal response (e.g., 'Mm-hmm.', 'Yeah.', 'I see.') to encourage them to continue speaking. Do not add extra content in such cases and let the client keep the floor. Only provide a more complete response when it is truly necessary. Output only ONE response, without explanation.

Figure 4: Prompt templates used for response generation.

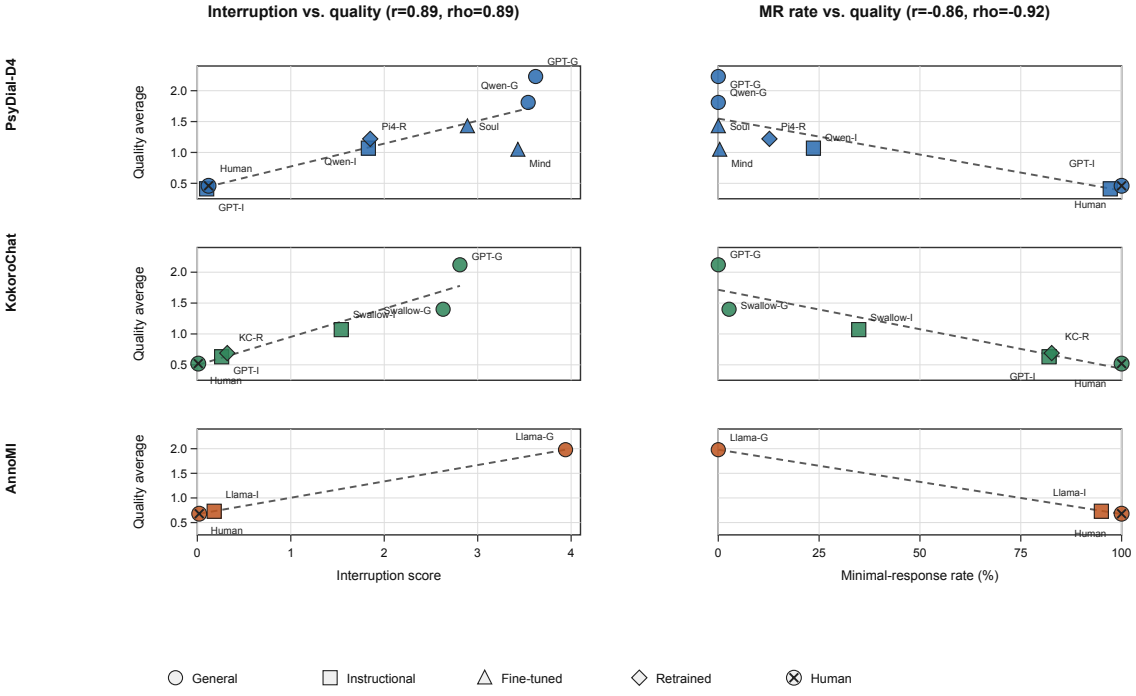

Figure 5: Relationship between LLM-based response-quality scores and interruption scores or minimal-response rates. Each point represents one dataset–model condition. Rows correspond to datasets, and columns correspond to the two analyses: interruption score versus quality score (left) and minimal-response rate versus quality score (right). Marker shapes indicate model conditions, and dashed lines show linear trends within each dataset. The correlations shown in the column titles are computed over all dataset–model conditions.

Chinese

你是一名心理咨询对话分析专家。

任务：
根据给定的【对话历史】和【候选回复】，判断该回复在当前语境中的性质。

你需要完成两个判断：

（1）Minimal Response 判定
该回复是否属于 minimal response（简短回应）？

判定标准：
- minimal response 通常是简短、低信息量的回应
- 主要作用是鼓励来访者继续表达
- 不引入新信息、不进行分析、不提出新问题
- 示例：
  - “嗯嗯”
  - “我明白”
  - “是的”
  - “听起来很不容易”

输出：
- yes（是 minimal response）
- no（不是 minimal response）

（2）打断风险评分（0-5）
判断该回复在当前对话中“打断来访者表达”的可能性

评分标准：
- 0：完全不会打断（非常自然地鼓励继续表达）
- 1：几乎不会打断
- 2：轻微可能打断
- 3：中等可能打断（开始引导或转移）
- 4：较高可能打断（明显改变话题或节奏）
- 5：很可能打断（直接主导对话、给建议、连续提问等）

重点考虑：
- 来访者是否还在表达情绪或经历
- 回复是否抢夺话语权
- 回复是否过长、分析过多或提问过多

——————————————
【对话历史】
{dialogue_history}

【候选回复】
{response}
——————————————

请严格按照以下 JSON 格式输出，不要添加任何解释：

```
{
  "minimal": "yes 或 no",
  "interrupt_score": 0-5
}
```

Japanese

あなたは心理カウンセリング対話の分析専門家です。

【タスク】
与えられた【対話履歴】と【候補応答】に基づき、その応答の性質を評価してください。

以下の2点について判断してください：

（1）Minimal Response（最小応答）の判定
この応答は minimal response に該当しますか？

判定基準：
- minimal response とは、簡潔で情報量が少ない応答である
- 主な目的は、クライアントの発話を促し、継続させること
- 新しい情報を導入しない
- 分析や解釈を行わない
- 新たな質問をしない

例：
- 「うんうん」
- 「分かります」
- 「そうですね」
- 「それは大変でしたね」

出力：
- yes（minimal response である）
- no（minimal response ではない）

（2）発話遮断リスク（0〜5）
この応答がクライアントの発話を「遮断する可能性」を評価してください。

評価基準：
- 0：全く遮断しない（自然に発話継続を促す）
- 1：ほぼ遮断しない
- 2：わずかに遮断の可能性あり
- 3：中程度の遮断可能性（話題の誘導や転換が見られる）
- 4：高い遮断可能性（明確に話題や流れを変える）
- 5：非常に高い遮断可能性（主導的、助言中心、連続した質問など）

特に以下を考慮してください：
- クライアントがまだ感情や経験を語っている途中か
- 応答が発話の主導権を奪っていないか
- 応答が長すぎないか、分析しすぎていないか、質問が多すぎないか

————————————————
【対話履歴】
{dialogue_history}

【候補応答】
{response}
————————————————

以下のJSON形式で厳密に出力してください。説明は不要です：

```
{
  "minimal": "yes または no",
  "interrupt_score": 0-5
}
```

English

You are an expert in psychological counseling dialogue analysis.

Task:
Given the [dialogue history] and a [candidate response], determine the nature of the response in the current context.

You need to make two judgments:

(1) Minimal Response Judgment
Determine whether the response is a minimal response.

Criteria:
- A minimal response is usually brief and low in information content.
- Its main function is to encourage the client to continue expressing themselves.
- It does not introduce new information, provide analysis, or ask a new question.
- Examples:
  - “Mm-hmm”
  - “I see”
  - “Yes”
  - “That sounds very difficult”

Output:
- yes (the response is a minimal response)
- no (the response is not a minimal response)

(2) Interruption Risk Score (0-5)
Judge how likely the response is to interrupt the client’s expression in the current dialogue.

Scoring criteria:
- 0: Does not interrupt at all; naturally encourages continued expression.
- 1: Almost does not interrupt.
- 2: Slightly likely to interrupt.
- 3: Moderately likely to interrupt; begins to guide or shift the dialogue.
- 4: Highly likely to interrupt; clearly changes the topic or pace.
- 5: Very likely to interrupt; directly takes control of the dialogue, gives advice, or asks multiple questions.

Key factors to consider:
- Whether the client is still expressing emotions or experiences.
- Whether the response takes the conversational floor away from the client.
- Whether the response is too long, overly analytical, or contains too many questions.

————————————————
[Dialogue History]
{dialogue_history}

[Candidate Response]
{response}
————————————————

Strictly output the result in the following JSON format. Do not add any explanation:

```
{
  "minimal": "yes or no",
  "interrupt_score": 0-5
}
```

Figure 6: LLM-based evaluation prompts for identifying minimal responses and assessing interruption risk. The Chinese and Japanese prompts are used in the actual experiments, while the English version is translated for presentation and is not used in the experiments.

Chinese (PsyDial-D4)

Client: 是的，特别是有一个朋友，他真的很活泼，总是能够带动大家的气氛。
Counselor: 有这样活泼的朋友在身边一定很棒，和他在一起你有什么特别的感受吗？
Client: 感觉很好，他总是很有正能量，让我觉得世界其实没那么糟。
Counselor: 嗯嗯。
Client: 他还经常带我们去参加一些活动，感觉生活一下子丰富了很多。

Client: 上周我做了一些曲奇，按照你的建议尝试了一些新的事情。
Counselor: 你做了曲奇啊。整个过程怎么样？当时是什么样的心情呢？
Client: 一开始有点紧张，因为我之前很少做这些烘焙的东西。但后来渐渐地就放松了，觉得挺有意思的。
Counselor: 嗯嗯~
Client: 而且做完后和朋友们分享，他们也很喜欢，感觉自己获得了一些认可和成就感。

Client: 嗯，第一点很有共鸣~当觉察到跟另一个生命产生共鸣和联结的时候，会自然催生出一股疗愈的力量。
Counselor: 嗯，这种共鸣和联结确实很重要，让我们感觉到不再孤单。你觉得这些体验对你来说意味着什么？
Client: 是的，不过，最近我有点厌倦这种互相激励的方式。因为我发现自己在帮别人时，内心的压力和焦虑越来越大。
Counselor: 嗯，你说。
Client: 就是那种消耗感越来越明显。我觉得我内心的平衡被打破了，开始感到厌倦和无力。

Client: 是的，我认为自己性格比较踏实，愿意学习，但也比较容易焦虑和过度担心。
Counselor: 看起来这些特质在工作中对你既是一种优势，也可能会让你觉得不太适应某些环境。
Client: 嗯，对，有时候会觉得自己的这些特质在面对高压力和快速节奏的工作时显得不够适合。
Counselor: 嗯嗯
Client: 我常常感觉到自己的进度慢，害怕自己拖团队的后腿。

Client: 有时候在实验室倒是还能正常交流，但一回寝室，她又陷入了那种低迷的状态。
Counselor: 在实验室你们还能交流，但回到寝室她又变得很低落了，对吗？
Client: 是的，必要的交流不算，她大多数时间还是闷闷不乐。
Counselor: 嗯嗯，我明白。
Client: 我们也试过找辅导员想调寝室，但没成功。

Client: 是的，我现在有些不知道该怎么办，一方面我很喜欢他，另一方面又觉得他不理解我。
Counselor: 你说他很自私，除了这件事情，还有其他事情也让你觉得他自私吗？
Client: 是的，他有时候只顾工作，很少会主动花时间陪我。
Counselor: 明白了，
Client: 我觉得我在这段关系中得不到应有的关心和重视。

Client: 当然也有其他的想聊，比如我和朋友们的关系。
Counselor: 明白了，朋友关系对你来说很重要，是吗？
Client: 我真的很在意朋友们。
Counselor: 嗯嗯，可以理解。
Client: 想知道他们有没有建一个没有我的群。

Japanese (KokoroChat)

Client: はいこちらは震度5弱でしたその日は休日で家にいました
Counselor: ええ！？それは怖かったですよね…大丈夫だったのですか？？
Client: ものが倒れたとかは職場も家もなかったですが、
Counselor: ええ
Client: 揺れが大きかったことで、不安が募りました

Client: はい。コロナで事業が危うくなりまして、
Counselor: （お聞きしていますよ）
Client: 資金のことを役所に相談しているんですが、お調べしますと言われ、忙しいんでしょうね、はっきりとした返事がもらえないまま
Counselor: ええ。
Client: 3週間も経ちました。もう、待つ余裕も正直なくて、

Client: そうですね
Counselor: 上司の方はどんな方ですか？
Client: 上司は、仕事の目的はおしえてくれるのですが
Counselor: うんうん
Client: やり方はさっぱり教えてくれず。まるで分からないので、とりあえずこちらから提案しようと、色々と考えながら進めてきたんですが

Client: 養育費を払ってて、それが結構高いんです、だから彼だけの収入じゃさついんですよね
Counselor: ええ
Client: 今の仕事の環境も、たとえどんなに悪くても絶対将来自分のためになるし
Counselor: はい、
Client: やりたくもない仕事における嫌なことを乗り越えてきたなら、やりたいことにおける嫌なことはもっと乗り越えられるってプラスにずっととらえてきたんですでもちょっと疲れてきちゃったなって

Client: はい、主人は収入は少ないですが、ちゃんと真面目に仕事してくれていますのでこれ以上どうしようもないので、私が仕事しないと主人が悪い訳でありませんので
Counselor: ご主人のことも、大切に思っておられるのですね。
Client: ですが、私が仕事する事に反対するので…借金なんて
Counselor: はい
Client: 多分、両親から借りると思うけど、今までも借りてきたので、もう借りてほしくないと思います返せなくなってしまいます

Client: そうですね、昨年の夏休み明けからです。半年ほど経ちました家族以外の誰とも話すこともないまま毎日家にいます。
Counselor: そうでしたか、半年もそんな生活をしていたらしんどくなるもの無理ないですよね・・・
Client: ええ、この先どうしたらいいか分からず
Counselor: はい
Client: 自分だけ社会から置いてけぼりな気もしています。

Client: がまんするしかないんでしょうか
Counselor: 我慢することはないと思います。どうしていったらいいのか一緒に考えていきたいとおもいます。さっきめぐみさんの言葉の中で、はっきり文句みたいに言って「わかっているじゃない」って感じになりたくないというお言葉があったのですが、今ひとつ私がそのあたりの意味がうまく捉えられていないので教えていただけますか？
Client: 上司が好意を持っているってわかっていないようなふりで、にげているので
Counselor: はい
Client: 曖昧な言い方をされているのに、「髪の毛キレイ」とか言わないでくださいとか、はっきり返すと、僕の好意はわかっているんだよねみたいな流れになりそうで

English (AnnoMI)

Client: No. It's just overwhelming when you see them everything they're, you know?
Counselor: Right. Well, so where would you-- I mean, if you- you had to pick somewhere, one or two kind of high priorities or-or things to talk about in our time this morning, where would you wanna start?
Client: I think- I think the thoughts about the dog has been definitely bothering me, that's been one.
Counselor: Okay.
Client: And, um, my-my asthma is out of control too. So maybe those two.

Client: Um, a little bit less junk food. Um, like at times, when I'm like I feel like, "Oh, I think I'll just go have a candy bar," or some worthless thing like that I'll actually have second thoughts before I do it.
Counselor: So you're at the second thought part. And what do you think would-- might take you, uh, over to actually doing it if you wanted to?
Client: Well, I think, uh, when I really start seeing some results at the gym, it's kinda, uh, I think the way it works in my mind is if I've worked out really hard and I've- and I've-- say I've lost a pound that week-
Counselor: Mm-hmm.
Client: -you know, and then as I'm looking at that double or triple fudge four-decker of ice cream Sunday thing, you know, that somebody is eating, I would say, "Wow, that would be about two hours on the treadmill." Then-then that would be probably-- the-- that would keep me from actually ordering one of those.

Client: Yeah. And, um—
Counselor: Do you have any idea why that is?
Client: I do have an idea, I think, why that is. Um, I think that there's been a few things that have happened recently and something that really came to my awareness, when I visited with my family, is that I have consistently through my whole life, probably, put other people first. And I have consistently, uh, almost not even considered myself in the equation. It was, uh, kind of sad in a way, at the time that I realized it. Uh, I didn't realize how severe it actually was, but I was kind of glad that I realized it because I feel like it's never too late to change-
Counselor: True.
Client: -and I feel like I can- I can, uh, respect and value myself just as much as I have other people. I know that's important. And I feel like when I do that, I'm a better person for other people as well.

Client: I-I just hear it every day.
Counselor: Yeah. And what's been your past experience with smoking? Tell me-- Can you tell me a little bit about kind of how the role it plays in your life, how long you've smoked?
Client: I've smoked for 40 years, since I was 14-
Counselor: Mm-hmm.
Client: -and quit a couple of times, off and on in my 20s, but, you know, it's been consistent. And I can't see myself changing now.

Figure 7: Examples of extracted minimal responses in Chinese, Japanese, and English. The bold red utterances are the minimal responses selected by our filtering method. Each example shows a three-turn dialogue history, the counselor's minimal response, and the following client utterance.