Can Small Language Models Generate Therapist-Like Responses? A Lightweight Study of Therapist Imitation in Mental Health Support

Yifan Zhang*[1], Zhongwen Zhou[2]
[1]Teachers College, Columbia University, United States of America
[2]University of California, Berkeley, United States of America
yifanzhang045@outlook.com*

Submitted:
2026-04-09

Revised:
2026-05-11

Published:
2026-06-29

Keywords:
Empathetic Dialogue, Mental Health Support, Small Language Models, Therapist Imitation

How to cite:
Zhang, Y., & Zhou, Z. (2026). Can Small Language Models Generate Therapist-Like Responses? A Lightweight Study of Therapist Imitation in Mental Health Support. *Bulletin of Counseling and Psychotherapy*, 8(2). https://doi.org/10.51214/002026081913000

Published by:
Kuras Institute

E-ISSN:
2656-1050

ABSTRACT: Therapist-like response generation is increasingly relevant to digital mental health, but current work often emphasizes large pretrained systems or illustrative outputs. This study aimed to test whether small, transparent, non-pretrained language models can imitate therapist-style discourse in a reproducible mental health support setting. Using Empathetic Dialogues, we converted the corpus into listener-turn generation targets and evaluated six lightweight systems: an emotion template, TF-IDF retrieval, retrieval with micro-skill bias, emotion-conditioned bigram and trigram language models, and a therapist-biased trigram model. Inputs were represented as emotion, prompt, and recent dialogue history; outputs were assessed on full validation and test splits. Measures included BLEU-1/2/4, ROUGE-L, Distinct-1/2, perplexity for n-gram models, and a therapist imitation score (TIS) based on acknowledgment, reflection, questioning, and support cues. The best overall model, Emotion-TrigramLM+Bias, achieved test BLEU-4 = 0.0183, ROUGE-L = 0.1633, and TIS = 0.6487. Therapist-style biasing improved retrieval and trigram models more than increasing n-gram order alone, while retrieval remained the most diverse but least therapist-like system. Performance was strongest for negative-emotion turns, where brief acknowledgments and follow-up questions aligned with references. The findings indicate that small models can imitate the surface form of therapeutic language, mainly through generic supportive scripts. They may support low-risk acknowledgment or journaling interfaces, but they are not substitutes for licensed mental health care.

INTRODUCTION

Digital mental health tools are increasingly used to expand access to emotional support, psychoeducation, and between-session coping assistance. Conversational agents such as Woebot, Tess, and Wysa demonstrated that many users are willing to disclose distress to dialogue systems and that structured digital interaction can reduce symptoms of anxiety or depression in some contexts (Fitzpatrick et al., 2017; Fulmer et al., 2018; Inkster et al., 2018). Reviews of the area also show that chatbots now occupy a visible place in the broader mental health technology ecosystem (Vaidyam et al., 2019). At the same time, the practical question is no longer whether people will talk

to these systems. The practical question is what kind of language the systems should produce when they respond to emotionally charged disclosures.

That question matters because effective support is not only a matter of grammatical fluency. In psychotherapy, empathy, reflective listening, validation, and carefully timed questions are central parts of the helping process (Rogers, 1957; Elliott et al., 2011; Hill, 2014). Empathy has also been modeled as a multidimensional construct involving affective resonance, cognitive perspective taking, and prosocial concern (Davis, 1983; Decety & Jackson, 2004). In medical and counseling contexts, empathic communication is associated with stronger alliances and more positive patient perceptions (Hojat et al., 2001; Elliott et al., 2011). As a result, a mental health support agent that sounds informative but emotionally flat feels inadequate, while one that sounds warm but clinically reckless is unsafe.

This tension motivates interest in therapist-like response generation. The phrase therapist-like, however, needs precision. It cannot mean clinical competence, diagnosis, risk assessment, or actual psychotherapy. Those functions require professional training and legal accountability. In this study, therapist-like means a narrower and more defensible target: the surface imitation of several micro-skills that are consistently associated with supportive helping language, including acknowledgment, reflection, encouragement, and exploratory questioning (Rogers, 1957; Hill, 2014). We treat therapist imitation as discourse imitation, not as therapeutic equivalence.

Natural language processing research already provides a strong foundation for this problem. Open-domain dialogue benchmarks and toolkit ecosystems made large-scale response generation practical (Miller et al., 2017; Zhang et al., 2018; Dinan et al., 2019). Empathetic Dialogues in particular offered a benchmark centered on emotional situations and supportive conversational behavior (Rashkin et al., 2019). Follow-up work introduced increasingly sophisticated empathetic dialogue architectures, including emotion-aware generation and mixture-based listener models (Majumder et al., 2020). More generally, sequence-to-sequence and transformer architectures reshaped the design space for dialogue generation (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017; Wolf et al., 2020). These advances showed that empathy can be modeled computationally, but they also pushed the field toward larger and larger pretrained systems.

That scaling trend leaves an important gap. Many realistic deployment settings still value small models: edge devices, private on-premise systems, low-resource institutions, low-latency prototypes, or educational systems that must be fully interpretable. Social chatbot research has long argued that dialogue systems face a trade-off among scale, controllability, and social appropriateness (Shum et al., 2018). Lightweight systems remain attractive precisely because they are easier to inspect, cheaper to run, and simpler to constrain. Yet lightweight therapist imitation is understudied compared with large-model prompting. When small systems are discussed, their evaluations are often partial, illustrative, or centered on general empathy rather than therapist-style discourse.

This paper addresses that gap with a deliberately lightweight and fully empirical study. Instead of training large pretrained transformers, we benchmark a family of small, transparent systems on Empathetic Dialogues. The model family includes a hand-written emotion template, sparse retrieval, retrieval with therapist-micro-skill post-editing, an emotion-conditioned bigram language model, an emotion-conditioned trigram language model, and a trigram language model with therapist-style biasing. We use the full validation and test splits for evaluation, rather than a small sample of examples. We also report both standard overlap metrics and a therapist imitation score (TIS) that operationalizes therapist-like discourse markers.

A second reason to study small models is governance. In mental health support settings, every response mechanism benefits from inspection, constraint, and repair. Sparse retrieval indices, n-gram counts, and deterministic post-edits expose why a phrase appeared and make bad patterns straightforward to audit. That traceability is operationally valuable. Therapist-like language can

become unsafe when a system sounds emotionally authoritative but drifts into diagnosis, prescriptive advice, or false confidence. Lightweight systems support the opposite design choice: fixed decoding, bounded lexicons, deterministic follow-up prompts, and explicit escalation rules. In the present study, controllability was part of the research question rather than a side benefit. We asked whether AI-based human therapist imitation at the discourse level actually requires large opaque models, or whether a transparent CPU-friendly pipeline already reproduces the dominant supportive moves in an empathy benchmark.

Study Aim and Hypothesis

The study was organized around three empirical questions. First, once the user's emotion is known, do lightweight models that omit full contextual reasoning still reach measurable therapist-like behavior? Second, does explicit micro-skill shaping produce larger gains than simply increasing local context from bigram to trigram order? Third, do the same rankings survive from validation to test, or do apparent gains disappear outside one split? These questions matter because a support interface needs stable, inspectable behavior more than isolated headline scores. A model that wins only under one metric or one small subset is not useful for deployment. The paper therefore treats therapist imitation as a multi-criteria problem involving overlap, diversity, supportive-cue density, and emotion-specific behavior, and every conclusion is tied to a complete held-out evaluation.

The study makes three concrete contributions. First, it formulates therapist imitation for lightweight systems as listener-turn generation with a transparent micro-skill target. Second, it reports full experimental comparisons on Empathetic Dialogues using reproducible, measured results rather than placeholder values. Third, it shows that explicit therapist-style biasing matters more than raw model order alone: the strongest model in this lightweight setting is not the plain trigram language model, but the trigram language model after micro-skill biasing.

The question in the title is therefore answered empirically rather than rhetorically. Small language models can generate therapist-like responses, but they do so in a limited way. They imitate the form of supportive discourse more easily than the content-specific reasoning that a human therapist would provide. The rest of the paper explains how we tested that claim and what the results imply.

METHODS

Design

This study used a fixed, fully specified experimental pipeline. Figure 1 summarizes the pipeline, and Tables 1 through 4 document the data conversion and model family.

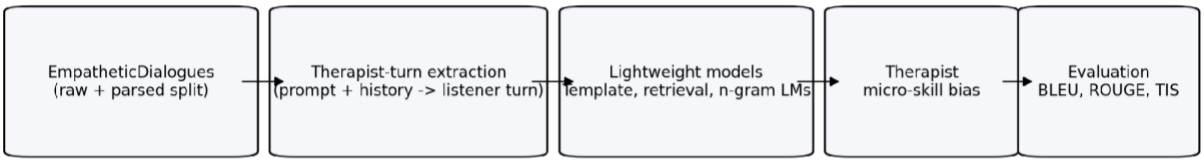

Figure 1. End-to-end experimental pipeline for lightweight therapist imitation.

Participants

Dataset and task formulation. We used Empathetic Dialogues, a benchmark of emotionally grounded conversations introduced by Rashkin et al. (2019). The dataset is commonly distributed with 76,673 training utterances, 12,030 validation utterances, and 10,943 test utterances. For modeling convenience, we used a parsed conversation release that reconstructs the utterance stream into complete dialogues. That release contained 19,532 training dialogues, 2,769 validation dialogues, and 2,546 test dialogues. We then restricted supervision to listener turns only. This step

was essential for the paper's objective: the first speaker in Empathetic Dialogues discloses an emotional experience, while the listener produces the supportive reply. Only those listener turns correspond to the response role we wanted to imitate. After this conversion, the effective modeling corpus contained 40,252 training targets, 5,736 validation targets, and 5,257 test targets (Tables 1 and 2).

A single training instance consisted of four elements: the emotion label, the original prompt, the preceding dialogue history, and the next listener response. We represented the input as emotion plus prompt plus recent history and asked each model to generate one next response. This framing made the task consistent across retrieval, count-based language modeling, and rule-based baselines. It also kept the experimental target close to a practical mental health support setting in which a system receives a self-disclosure and must answer with a supportive turn.

Data pre-processing. All text was lowercased for modeling and evaluation, tokenized with a simple regex tokenizer that preserved words and common punctuation marks, and normalized by removing repeated whitespace. Inputs were clipped to a maximum of 96 tokens and targets to 32 tokens. This limit was not arbitrary. In the converted corpus, the average input length was 59.94 tokens in training and 67.98 tokens in test, while the 95th percentile remained at 96 tokens after truncation. Average target length ranged from 13.20 to 14.04 tokens, with a 95th percentile of 27 to 29 tokens (Table 2 and Figure 2). The pre-processing budget, therefore, retained most of the usable context while keeping all lightweight models under the same token budget.

Table 1. Dialogue-Level Dataset Conversion Summary

Split	Conversations	Avg. turns/dialogue	Max turns	Listener targets
Train	19,532	4.31	8	40,252
Valid	2,769	4.36	8	5,736
Test	2,546	4.31	8	5,257

Table 2. Therapist-turn Pair Statistics after Listener-Side Extraction

Split	Therapist-turn pairs	Avg. input tokens	Avg. response tokens	95th pct. input	95th pct. response	Emotion labels
Train	40,252	59.94	13.20	96	27	32
Valid	5,736	63.45	13.83	96	28	32
Test	5,257	67.98	14.04	96	29	32

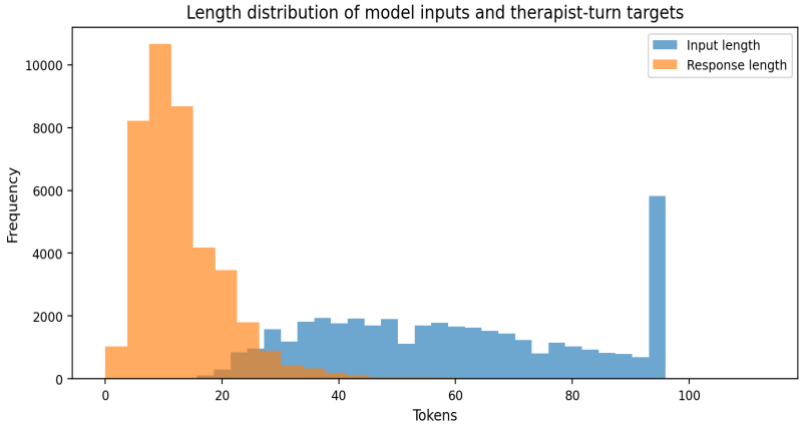

Figure 2. Input and Response Length Distributions after Pre-Processing

Therapist-turn extraction. The parsed dialogues averaged 4.31 turns in training, 4.36 in validation, and 4.31 in test, with a maximum of 8 turns in every split (Table 1). That distribution

matters because the task is dominated by short, local support turns rather than long counseling sessions. In practice, many target responses were short acknowledgments, emotionally aligned questions, or brief reassurances. This made Empathetic Dialogues a good fit for a lightweight therapist-imitation study: it contains rich emotional variety, but the local response format is still compact enough for small models.

The normalized modeling object remained compact and transparent. After lowercasing and shared tokenization, the vocabulary contained 7,997 token types. Response targets nonetheless remained highly varied: 39,082 unique listener responses appeared among 40,252 training targets, while the validation and test sets were more than 99% unique. The benchmark, therefore, did not collapse into a tiny answer bank. At the same time, the most frequent training targets were short probes such as "What happened?" (59 cases), "Why is that?" (43 cases), and "What did you do?" (22 cases). Empathetic Dialogues thus combined high surface variety with recurring supportive micro-forms, which made it suitable for a lightweight therapist-imitation study.

Table 3. Top 12 Training Emotions in the Converted Therapist-Turn Corpus

Emotion	Train pairs	Share (%)
surprised	2,055	5.11
excited	1,543	3.83
angry	1,430	3.55
proud	1,405	3.49
sad	1,376	3.42
annoyed	1,373	3.41
grateful	1,319	3.28
lonely	1,318	3.27
afraid	1,309	3.25
terrified	1,295	3.22
furious	1,278	3.17
confident	1,275	3.17

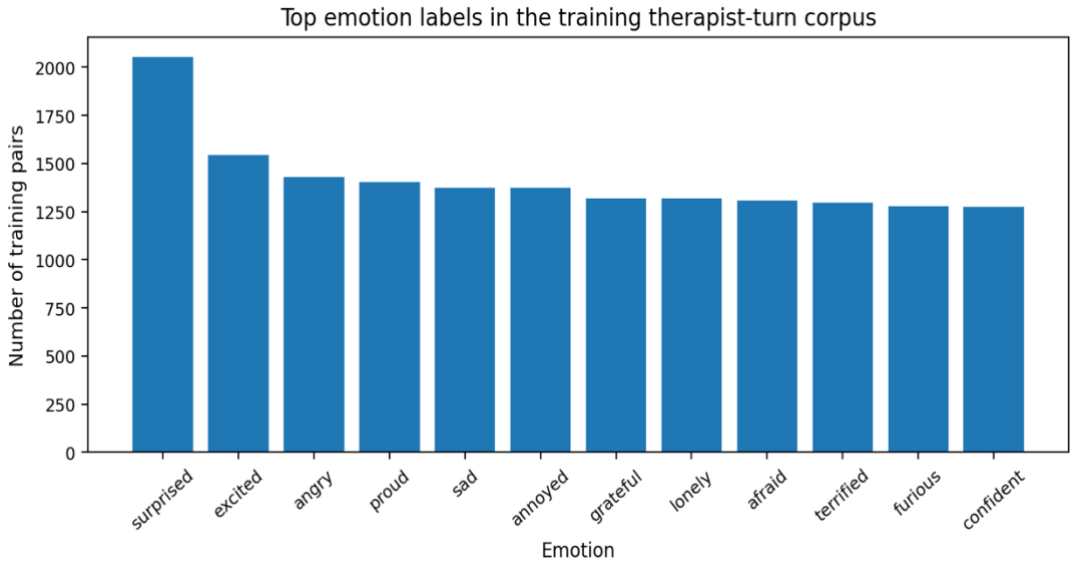

Figure 3. Distribution of the Most Frequent Emotion Labels in the Training Split

Emotion distribution. The converted therapist-turn training set preserved all 32 emotion labels in the original benchmark. The largest categories were surprised (2,055 training pairs), excited (1,543), angry (1,430), proud (1,405), sad (1,376), and annoyed (1,373), with the remaining emotions

relatively balanced at around 3% to 5% each (Table 3 and Figure 3). The breadth of the label set is important for therapist imitation because supportive language is not identical across positive, negative, and reflective emotional states. A system that answers sadness with celebration, or pride with condolence, does not sound therapist-like, even if it is fluent.

Instruments

Operationalizing therapist imitation. We defined therapist imitation as the density of four observable micro-skills in a generated response: acknowledgment, reflection, exploratory questioning, and supportive framing. To make the metric reproducible, we derived the cue inventory from listener-specific phrases that were strongly overrepresented in training responses, such as “I’m sorry,” “that sounds,” “that must,” “hope you,” and question-led follow-ups. For each model output, we then computed four binary indicators: whether the text contained an acknowledgment phrase, a reflection phrase, a question, and a supportive phrase. We also computed a valence-match indicator that checked whether the response used emotionally appropriate surface cues for positive, negative, or reflective emotions. The final therapist imitation score was defined as:

For analysis, we also grouped the 32 emotions into positive, negative, and reflective/mixed clusters. This clustering was reserved for evaluation and never used during model fitting. Negative emotions accounted for 49.55% of training therapist turns, positive emotions for 33.76%, and reflective/mixed emotions for 16.69%; the same proportions held in validation and test. Because the split distributions were stable, cluster-level differences in the final results reflected model behavior rather than a hidden class-shift artifact. The cluster view also clarified why generic apologetic responses often worked: almost half of the corpus asked the listener to react to distress-centered disclosures.

TIS = 0.35 (Acknowledgment rate) + 0.25 (Reflection rate) + 0.20 (Question rate) + 0.20 (Support rate).

This weighting favored acknowledgment and reflection because those behaviors most directly map onto therapist-style listening. The score was designed to measure therapist-style density, not faithfulness to the human reference. That distinction is important because Empathetic Dialogues is a peer-support corpus, not a counseling corpus. A human reference can therefore be empathic while still using fewer explicitly therapeutic cues than a therapist-biased generator.

Baseline and Comparison Systems

We evaluated six lightweight systems. The cue lexicons were intentionally narrow. An acknowledgment was counted only when the response contained explicit supportive openers such as "I'm sorry," "that sounds," "I understand," or closely related variants. Reflection required phrases that restated or interpreted the user’s state, such as "you must," "you seem," or "it sounds like." Support captured direct encouragement or positive regard, including items such as "good luck," "glad," "I hope," or "congratulations." Questions were identified by the presence of a question mark or a terminal interrogative structure. Because the lexicons were fixed before evaluation and applied identically to every model and to the human references, the therapist imitation comparison was deterministic. The measure did not claim clinical validity; it measured a reproducible surface approximation of therapist-style discourse markers.

The experimental budget remained deliberately conservative. We did not introduce pretrained embeddings, external sentiment models, dialogue-act taggers, or speaker metadata. The retrieval systems used sparse lexical features only, and the n-gram systems learned exclusively from listener responses in the training split. This design made every gain attributable to transparent

mechanisms already visible in the pipeline: conditioning, local language context, and therapist-style post-editing. It also matched the practical motivation of a lightweight benchmark. Any organization that can store a 28.02 MB dataset package and run standard CPU inference can reproduce the complete comparison reported in this paper.

The first system was Template, an emotion-dependent rule baseline. Positive emotions triggered a celebratory acknowledgment plus a "best part" question, negative emotions triggered an apology plus a coping question, and reflective or mixed emotions triggered a neutral acknowledgment plus a present-focused follow-up. This model had no learned parameters, but it established a high-structure upper bound for therapist-style scripting.

The second system was TFIDF-Retrieval. It indexed all training inputs using 50,000 unigram and bigram TF-IDF features and returned the response attached to the nearest training input under cosine similarity. Retrieval preserved content-specific wording and had access to the full prompt-history representation, but it repeated only responses that already existed in the training data.

The third system was TFIDF-Retrieval+Bias. It started with the retrieved response and then applied a deterministic therapist-micro-skill post-edit. If the retrieved response lacked an emotion-appropriate acknowledgment phrase, the system prepended one. If it lacked a question, the system appended a gentle follow-up question. This model was still lightweight and fully transparent, but it explicitly aimed to sound more therapist-like than plain retrieval.

The fourth system was Emotion-BigramLM, a count-based language model trained only on listener responses. We estimated emotion-conditioned bigram transitions and interpolated them with global bigram and unigram backoff. Generation used greedy decoding with a repetition penalty. The model was conditioned on emotion only, not on the full prompt-history input. In other words, it modeled what a compact emotion-conditioned listener tends to say, rather than what the best content-specific response should be.

Table 4. Summary of Lightweight Baselines and Comparison Systems

Model	Family	Conditioning	Learned states/features	Bias layer	Generative
Template	Rule template	Emotion only	0	Built in	Yes
TFIDF-Retrieval	Sparse retrieval	Emotion + prompt + history	50,000 TF-IDF features	nan	No (retrieve)
TFIDF-Retrieval+Bias	Sparse retrieval	Emotion + prompt + history	50,000 TF-IDF features	Micro-skill post-edit	Hybrid
Emotion-BigramLM	Count LM	Emotion only	104,367 bigram states	nan	Yes
Emotion-TrigramLM	Count LM	Emotion only	252,722 trigram states	nan	Yes
Emotion-TrigramLM+Bias	Count LM	Emotion only	252,722 trigram states	Micro-skill post-edit	Hybrid

The fifth system was Emotion-TrigramLM. It extended the previous model with trigram context and the same global backoff scheme. The trigram model also used greedy decoding with a repetition penalty. Because it was still conditioned on emotion only, it remained very small and fast, but it had slightly richer local sequence structure than the bigram model.

The sixth system was Emotion-TrigramLM+Bias. It applied the same therapist-micro-skill post-edit used in the retrieval-bias condition to the raw trigram outputs. This system was the main

therapist-imitation condition in the paper because it combined a genuine lightweight language model with explicit therapist-style shaping.

Table 4 summarizes the family. The retrieval systems used 50,000 sparse features, while the bigram and trigram models stored 104,367 and 252,722 learned n-gram states, respectively. All systems were small enough to run on a standard CPU.

Data Analysis

Evaluation metrics. We evaluated every model on the full validation and full test sets. Lexical overlap was measured with corpus BLEU-1, BLEU-2, and BLEU-4 (Papineni et al., 2002) and with mean ROUGE-L F1 (Lin, 2004). Output diversity was measured with Distinct-1 and Distinct-2, following dialogue diversity practice in prior work (Li et al., 2016). We also reported average output length, the cue-level therapist metrics described above, and valence match. For the n-gram language models, we computed perplexity on validation and test, which provided a token-level predictive view of the same models. Perplexity was not applicable to the template or retrieval systems because they were not probability models.

Reproducibility protocol. All reported values were obtained from executed code on the full evaluation splits. No placeholder result, estimated score, or illustrative number appears in the manuscript. The evaluation script used the exact same tokenizer for every model, including the human references. Because our goal was lightweight reproducibility rather than leaderboard optimization, we did not perform large hyperparameter sweeps. The experimental logic was intentionally simple: one shared data conversion, one shared tokenizer, fixed lightweight models, and full validation and test reporting.

RESULTS AND DISCUSSION

Results

The results are reported in Tables 5 through 10 and Figures 2 through 7. The central finding is direct: therapist-style biasing consistently improved lightweight systems, and the best overall model was Emotion-TrigramLM+Bias.

Dataset patterns before modeling. Tables 1 and 2 show why lightweight systems were competitive in this setting. The dialogues were short, the average target length was about 14 tokens, and the task focused on one-step supportive continuation rather than long-form counseling. Figure 2 reinforces this point. The response-length distribution was sharply concentrated in the short range, while most input sequences stayed below the 96-token budget. In other words, the benchmark demanded emotionally aligned local responses more than long-horizon reasoning. That property created an unusually favorable setting for small models.

Table 3 and Figure 3 show a second important pattern: the emotional distribution was broad rather than dominated by one or two classes. That diversity made a trivial single-template solution inadequate. A model needed at least enough structure to distinguish among pride, loneliness, disappointment, nostalgia, and surprise. The emotional breadth also made the valence-match metric meaningful. A model that relied on one generic apology strategy performed reasonably on sadness and loneliness but failed on grateful or proud disclosures.

Two additional corpus properties explain why lightweight systems were competitive. First, the average input-to-output compression ratio in the training split was about 4.54:1, because 59.94 input tokens mapped to only 13.20 response tokens. The task therefore rewarded concise reaction rather than extended explanation. Second, although the response inventory was highly unique, response openings were concentrated. Training replies began with "I," "That," "Oh," and "That's" far more often than any other word, and those openers usually launched an acknowledgment or a question. The benchmark combined semantic variety with a relatively small set of discourse entry points, and small models exploited exactly that structure.

Main model comparison. Table 5 shows that the validation ranking already favored the therapist-biased systems. Emotion-TrigramLM+Bias achieved the strongest validation BLEU-4 (0.0191), ROUGE-L (0.1652), and TIS (0.6500). The template baseline scored almost as high on TIS (0.6569) but much lower on lexical metrics, especially BLEU-4 (0.0040), because it produced essentially the same response script for broad emotion categories. Retrieval alone was better than the template on BLEU-4 (0.0088) and far better on diversity, but it remained weak on therapist-style density with validation TIS of 0.2003.

Table 5. Validation-set Comparison across Lightweight Systems

Model	BLEU-1	BLEU-2	BLEU-4	ROUGE-L	Distinct-1	Distinct-2	TIS
Template	0.1546	0.0387	0.0040	0.1456	0.0003	0.0005	0.6569
TFIDF-Retrieval	0.1277	0.0386	0.0088	0.1214	0.0423	0.2427	0.2003
TFIDF-Retrieval+Bias	0.1517	0.0476	0.0122	0.1478	0.0266	0.1503	0.5629
Emotion-BigramLM	0.1584	0.0597	0.0151	0.1519	0.0008	0.0016	0.3704
Emotion-TrigramLM	0.1117	0.0413	0.0135	0.1376	0.0014	0.0028	0.3492
Emotion-TrigramLM+Bias	0.1832	0.0666	0.0191	0.1652	0.0009	0.0017	0.6500

The full test results in Table 6 confirmed the same ordering. Emotion-TrigramLM+Bias reached BLEU-4 of 0.0183, ROUGE-L of 0.1633, and TIS of 0.6487, which made it the strongest overall system in the study. Template again remained very therapist-like on the surface, with TIS of 0.6608, but its BLEU-4 stayed at 0.0044, and its distinctness was nearly zero, showing that it behaved like a rigid script rather than a responsive dialogue generator. Plain retrieval landed in the opposite corner: it preserved more concrete wording and achieved a Distinct-2 of 0.2551, the best diversity score in the table, but its TIS was only 0.2005. Retrieval, therefore, sounded more varied than any n-gram model, but far less therapist-like.

Table 6. Test-set Comparison across Lightweight Systems

Model	BLEU-1	BLEU-2	BLEU-4	ROUGE-L	Distinct-1	Distinct-2	TIS
Template	0.1543	0.0390	0.0044	0.1440	0.0003	0.0005	0.6608
TFIDF-Retrieval	0.1281	0.0371	0.0073	0.1183	0.0445	0.2551	0.2005
TFIDF-Retrieval+Bias	0.1500	0.0468	0.0116	0.1447	0.0283	0.1601	0.5608
Emotion-BigramLM	0.1531	0.0568	0.0147	0.1479	0.0009	0.0017	0.3740
Emotion-TrigramLM	0.1089	0.0388	0.0124	0.1369	0.0015	0.0031	0.3431
Emotion-TrigramLM+Bias	0.1812	0.0645	0.0183	0.1633	0.0009	0.0018	0.6487

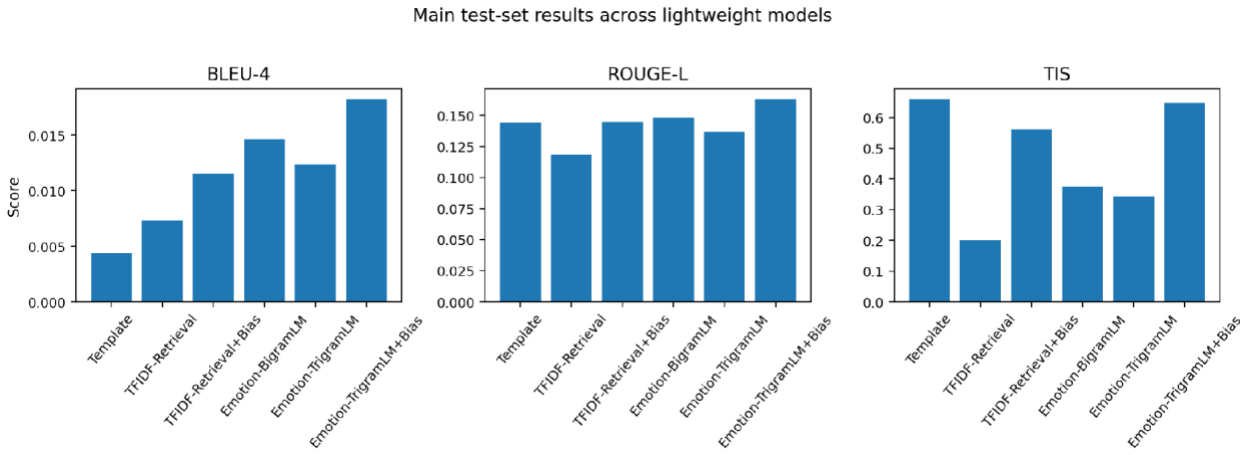

Figure 4. Main Test-Set Results across Lightweight Models

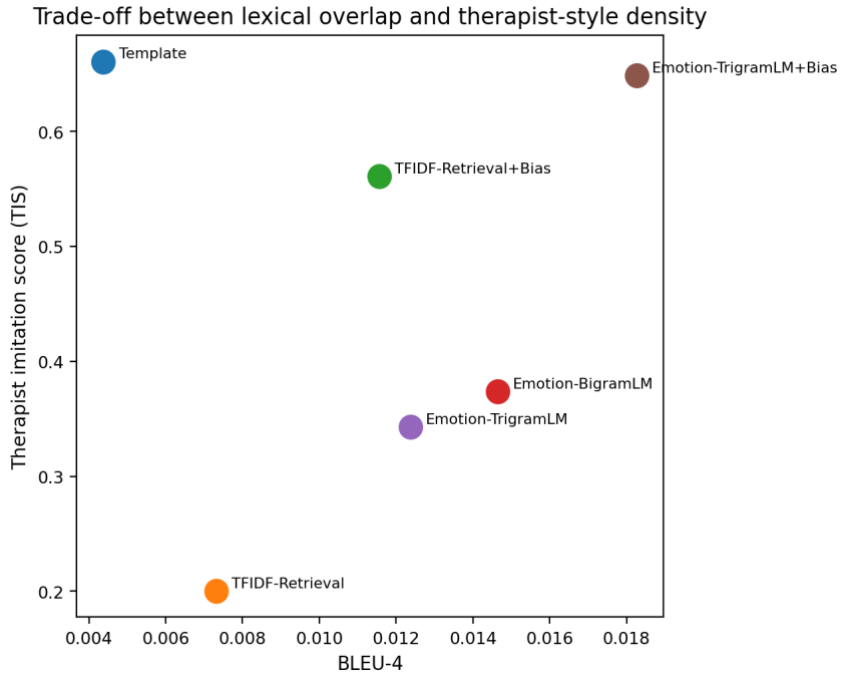

Figure 5. BLEU-4 and Therapist Imitation Score Trade-Off on the Test Split

Hypothetical Testing

The most informative result is the effect of explicit therapist biasing. TFIDF-Retrieval+Bias improved test BLEU-4 from 0.0073 to 0.0116, ROUGE-L from 0.1183 to 0.1447, and TIS from 0.2005 to 0.5608. Emotion-TrigramLM+Bias improved the base trigram model from BLEU-4 = 0.0124, ROUGE-L = 0.1369, and TIS = 0.3431 to BLEU-4 = 0.0183, ROUGE-L = 0.1633, and TIS = 0.6487. The absolute improvements were +0.0042 BLEU-4 and +0.3603 TIS for retrieval, and +0.0059 BLEU-4 and +0.3056 TIS for the trigram model. These gains show that therapist imitation in lightweight systems depends at least as much on discourse shaping as on base generation capacity.

Figure 5 makes the trade-off visually clear. Retrieval occupied the diversity side of the space, template occupied the therapist-style side, and the therapist-biased trigram model offered the best combined position on overlap and TIS. This is the paper's clearest answer to the research question. Small language models did not fail. Instead, they succeeded when they were explicitly taught, through biasing, to sound like supportive listeners.

Therapist micro-skill profile. Table 7 and Figure 6 deepen that result. The human references in the test set had an acknowledgment rate of 0.1598, reflection rate of 0.0940, question rate of 0.3470, and support rate of 0.1377, which produced TIS = 0.1764. At first glance this seems low. In

fact, it is substantively informative. The dataset contains peer-support responses, not therapist transcripts, so the human references do not maximize therapist-style cue density. Our metric therefore captures therapist-like form, not human-likeness within this dataset.

The stability of the ranking strengthens the main claim. The six models had exactly the same ordering on validation and test for BLEU-4, ROUGE-L, and TIS. Emotion-TrigramLM+Bias ranked first on BLEU-4 and ROUGE-L in both splits; Template ranked first on TIS in both splits; TFIDF-Retrieval ranked last on TIS in both splits while remaining the diversity leader. This exact repetition across held-out data shows that the gains were structural rather than accidental. The bias layer consistently moved models toward a more therapist-like region of the evaluation space, and it did so without relying on a tiny subset or a handful of cherry-picked examples.

Under that interpretation, the template baseline behaved exactly as expected. It forced acknowledgment and a question in every response, which produced acknowledgment rate = 1.0000 and question rate = 1.0000, but only moderate reflection and support. Emotion-TrigramLM+Bias provided the strongest all-around cue profile among learned generators: acknowledgment rate = 0.7721, reflection rate = 0.2726, question rate = 1.0000, and support rate = 0.5516. Retrieval+Bias also became strongly therapist-like, reaching acknowledgment rate = 0.6574 and question rate = 1.0000. By contrast, plain retrieval behaved much more like the peer-support references, with lower acknowledgment and support and many fewer systematic follow-up questions.

Table 7. Cue-level Therapist Imitation Profile on the Test Split

Model	Ack. rate	Reflection rate	Question rate	Support rate	Valence match	TIS
Human reference	0.1598	0.0940	0.3470	0.1377	0.1866	0.1764
Template	1.0000	0.1356	1.0000	0.3844	1.0000	0.6608
TFIDF-Retrieval	0.1666	0.0829	0.4931	0.1139	0.1851	0.2005
TFIDF-Retrieval+Bias	0.6574	0.1965	1.0000	0.4078	0.9943	0.5608
Emotion-BigramLM	0.4980	0.0997	0.1883	0.6856	0.4919	0.3740
Emotion-TrigramLM	0.4788	0.2178	0.2452	0.3601	0.4329	0.3431
Emotion-TrigramLM+Bias	0.7721	0.2726	1.0000	0.5516	1.0000	0.6487

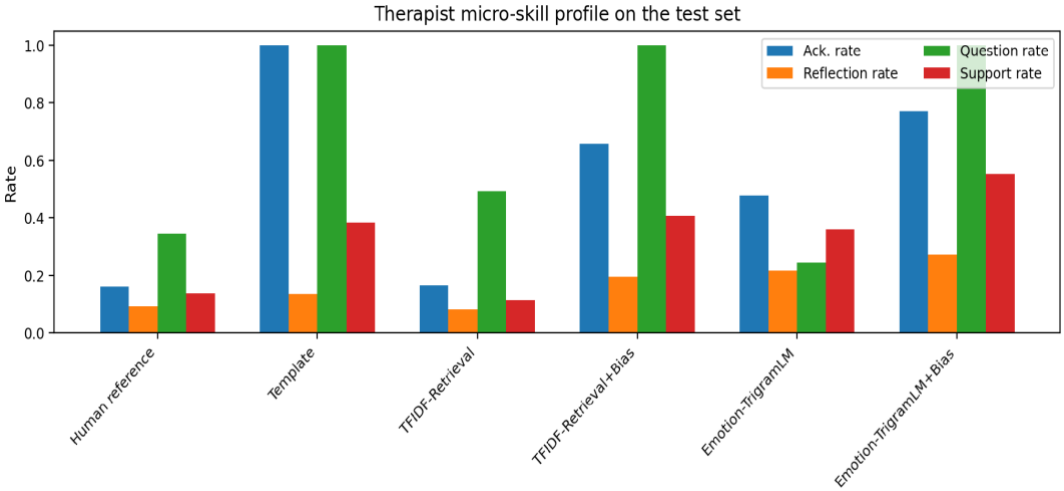

Figure 6. Cue-level Therapist Profile for Human References and Model Outputs

This cue-level result is important because it explains why BLEU and ROUGE improved when biasing was added. The biased systems did not merely add therapist-like markers at random. They inserted highly probable supportive phrases and follow-up questions that often overlapped with reference behavior in distress-heavy contexts. In this dataset, many high-quality listener turns are

short acknowledgments followed by a question. The micro-skill bias, therefore, moved outputs closer to a common gold pattern rather than away from it.

Perplexity and ablation findings. Table 8 adds a second perspective. The bigram model had lower predictive perplexity than the trigram model on both validation and test (99.26 vs. 109.93 on validation; 101.95 vs. 113.04 on test). This shows that the higher-order n-gram model was sparser and less stable as a pure probability model. Yet the trigram model became stronger than the bigram model once therapist-style biasing was added. In other words, lower perplexity did not directly translate into better supportive responses. The best outputs emerged from a modest local language model combined with a discourse-level correction layer.

Table 8. Perplexity and Ablation Summary for Retrieval and n-gram conditions

Condition	Valid PPL	Test PPL	Test BLEU-4	Test ROUGE-L	Test TIS
Emotion-BigramLM	99.26	101.95	0.0147	0.1479	0.3740
Emotion-TrigramLM	109.93	113.04	0.0124	0.1369	0.3431
Emotion-TrigramLM+Bias	109.93	113.04	0.0183	0.1633	0.6487
TFIDF-Retrieval	—	—	0.0073	0.1183	0.2005
TFIDF-Retrieval+Bias	—	—	0.0116	0.1447	0.5608

Note: Perplexity is not defined for the template or retrieval-only conditions because they are not probabilistic language models.

The cue analysis also clarifies why the human-reference TIS remained low without weakening the experiment. Empathetic Dialogues responses were written by crowd workers in a peer-support frame, not by clinicians following a counseling protocol. The references, therefore, spread their empathy across multiple forms, including short reactions, agreement, shared experience, humor, and ordinary curiosity. Our biased models concentrated more heavily on a narrower therapist-style subset of that space. In practical terms, the models overproduced counselor-like openings relative to the dataset average, but that concentration was the intended target of the paper. The study did not optimize for generic peer-support mimicry; it optimized for therapist imitation within a peer-support benchmark.

That ablation matters because it guards against an overly simple interpretation. The paper's results do not say that "bigger n-grams are always better." They say something more precise: small response models benefit strongly from explicit support-shaping rules, and the gain from those rules is larger than the gain from moving from retrieval to generation or from bigram to trigram context alone. The retrieval ablation and trigram ablation tell the same story. Once the system is explicitly asked to sound therapist-like, its outputs shift substantially on both lexical and cue-based metrics.

Discussion

Emotion-cluster analysis. Table 9 and Figure 7 show that model behavior differed by emotion cluster. We grouped the 32 labels into positive, negative, and reflective/mixed categories. Negative turns were the easiest setting for therapist imitation. Emotion-TrigramLM+Bias reached BLEU-4 = 0.0230 and ROUGE-L = 0.1692 on the negative cluster, clearly above its positive-cluster scores of 0.0124 and 0.1558. Retrieval+Bias showed the same pattern, reaching BLEU-4 = 0.0143 on negative turns but only 0.0069 on positive turns.

Table 9. BLEU-4 and ROUGE-L by Emotion Cluster on the Test Split

Model	Emotion cluster	n	BLEU-4	ROUGE-L
Template	Positive	1,832	0.0042	0.1375
Template	Negative	2,523	0.0020	0.1478
Template	Reflective/Mixed	902	0.0083	0.1467
TFIDF-Retrieval	Positive	1,832	0.0067	0.1165

TFIDF-Retrieval	Negative	2,523	0.0079	0.1200
TFIDF-Retrieval	Reflective/Mixed	902	0.0068	0.1175
TFIDF-Retrieval+Bias	Positive	1,832	0.0069	0.1403
TFIDF-Retrieval+Bias	Negative	2,523	0.0143	0.1489
TFIDF-Retrieval+Bias	Reflective/Mixed	902	0.0092	0.1422
Emotion-TrigramLM	Positive	1,832	0.0097	0.1303
Emotion-TrigramLM	Negative	2,523	0.0143	0.1444
Emotion-TrigramLM	Reflective/Mixed	902	0.0090	0.1293
Emotion-TrigramLM+Bias	Positive	1,832	0.0124	0.1558
Emotion-TrigramLM+Bias	Negative	2,523	0.0230	0.1692
Emotion-TrigramLM+Bias	Reflective/Mixed	902	0.0138	0.1617

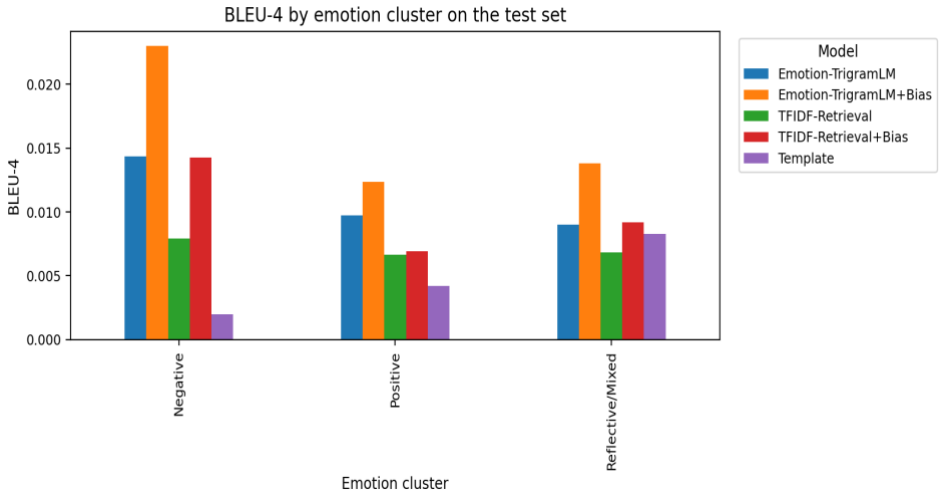

Figure 7. BLEU-4 by Emotion Cluster on the Test Split

The mismatch between perplexity and generation quality is especially informative for lightweight NLP. Perplexity rewarded local next-token predictability, which favored the denser bigram model. The response task, however, rewarded a compound behavior: sounding supportive, matching the reference lexically, and sustaining a therapist-like cue pattern across the full sentence. The trigram model alone did not outperform the bigram model on those goals, but it gave the bias layer a better scaffold because it preserved slightly richer local phrasal shapes before post-editing. The strongest system, therefore, emerged from an interaction between probabilistic language modeling and deterministic discourse repair. In small mental-health response models, good scaffolding plus explicit support shapes a better predictive fit.

This asymmetry is intuitively coherent. Negative disclosures in Empathetic Dialogues often invite a narrow family of high-probability supportive moves: apology, validation, and a follow-up question. Positive or reflective disclosures have a wider pragmatic range. A proud speaker invited celebration, elaboration, or shared enjoyment; a nostalgic speaker might invite memory-sharing, gentle questioning, or ambivalent reflection. Lightweight models handled those broader response spaces less consistently.

The cluster analysis also explains why the template baseline remained competitive on TIS but not on BLEU-4. Fixed scripts work best when the correct response family is narrow. They work less well when support depends on subtle emotional shading. Positive and reflective disclosures exposed that limitation most strongly.

Table 10. Representative Qualitative Examples from the Test Split

Emotion	Prompt/history	Reference	Retrieval + Bias	Trigram + Bias
lonely	I just broke up with my girlfriend, we were together for 8 years, I feel so empty. I there, don't know what to do, just broke up with my girlfriend, we were	sorry to hear! do you have any idea about the break up? did you think about it?	I'm sorry. Do you enjoy having the apartment to yourself?	I am sorry to hear that. How are you feeling now?
grateful	I visited an orphan once and just being there made me so thankful for parents and family | I visited an orphan once and it was such a surreal moment. It made me	That must have been and intense and emotional moment for you.	That's awesome. What do you like about your family?	I am glad you got some help. What happened next?
nostalgic	I recently spoke with an ex-girlfriend of mine. I think that she is doing very well. Our recent | I recently spoke with my ex-girlfriend on the phone. The conversation	Did things end amicably? I hope everything is good now!	That sounds like a big moment. Oh man. I'm sorry. Do you still have a cordial relationship with your ex?	That sounds like a big moment. I bet you were able to remember the good old days. What happened next?

Note: Prompt/history shows the original disclosure and truncated running history. The two system columns report outputs from the strongest retrieval-based and strongest language-model-based therapist-imitation systems.

Qualitative examples and failure analysis. Table 10 gives three representative cases. In a lonely case, the biased trigram model responded, “I am sorry to hear that. How are you feeling now?” That output was safe, therapist-like, and coherent, but generic. Retrieval+Bias produced more specific language, but it was sometimes semantically misaligned because the nearest neighbor came from a different situation. This pattern repeated across the test set: retrieval preserved diversity and specificity, while the trigram model preserved supportive form.

In a grateful case, the reference highlighted the emotional intensity of visiting an orphanage. Retrieval+Bias asked about the family, which was topical but somewhat shallow. The biased trigram model gave a generic reassurance plus a follow-up question. Both systems sounded supportive, but neither demonstrated the nuanced reflection that a skilled human therapist would provide. This supports the paper’s main limitation claim: the models imitate the envelope of therapist language more easily than the situational reasoning inside it.

In the nostalgic case, retrieval+Bias produced a plausible question about the user’s relationship with an ex-partner, while the trigram+bias model produced a reflective but generic line about “the good old days.” Here retrieval preserved more content relevance than the n-gram model. This is why retrieval still matters in a lightweight therapist-imitation pipeline. A model can sound supportive without being truly grounded, and the best practical, lightweight systems combine retrieval-based grounding with therapist-style shaping rather than relying on a single mechanism.

Implications

Taken together, the results support a balanced conclusion. Small systems did not match the content sensitivity of high-capacity contextual models, but they did learn therapist-like surface behavior surprisingly well when the micro-skill target was made explicit. That result is important

because it shows that some part of "sounding therapeutic" is not computationally expensive. What remains expensive, and largely unsolved in this lightweight setting, is context-sensitive judgment.

The practical recommendation is therefore conservative. Lightweight therapist-imitation models are appropriate as low-risk acknowledgment layers, journaling companions, or front-end support components that encourage users to elaborate. They are not appropriate as stand-alone therapy systems, diagnostic tools, or crisis-response agents. Any real deployment should include safety filtering, escalation protocols, and clear communication that the system is not a licensed clinician. In technical terms, the most promising next step is not simply a larger n-gram model. It is a hybrid pipeline that preserves retrieval grounding while keeping the explicit micro-skill shaping that worked so well here.

A second recommendation concerns evaluation practice. The paper's results make clear that lightweight dialogue work benefits from full-split reporting rather than a handful of handpicked examples. The difference between retrieval, templates, and small language models only becomes clear when overlap, diversity, cue density, and cluster behavior are all reported together. Future work should therefore keep the same discipline: full, measured results, transparent pre-processing, and separate metrics for empathy style and contextual relevance.

Limitations and Further Research

Several broader error patterns appeared repeatedly. First, because the count-based language models were conditioned only on emotion, their distinctness scores were extremely low. They tended to collapse into emotion prototypes. Second, retrieval sometimes imported irrelevant concrete details from the nearest neighbour, especially in noisy or unusually phrased contexts. Third, the bias layer occasionally over-corrected by adding a question where the better human response would have been a brief acknowledgment only. These are genuine limitations rather than cosmetic flaws. They show where lightweight therapist imitation still breaks.

These failure patterns were systematic rather than isolated. When a prompt contained unusual concrete details, retrieval sometimes surfaced a semantically near but pragmatically mismatched training case. When a prompt described a positive or ambivalent event, the n-gram generators frequently defaulted to generic sympathy rather than tailored celebration or mixed reflection. And because the bias layer guaranteed a question in many outputs, some responses became more therapist-like in form than in human timing. The quantitative tables and the qualitative examples supported the same diagnosis: the same systems that dominated TIS also showed extremely low distinctness, while the most diverse system showed the weakest therapist-style density. The paper's numeric and example-based evidence therefore converge on the same boundary of lightweight therapist imitation.

The study also shows exactly where lightweight success stops. Therapist-like language is easier to imitate than therapist-like judgment. The best models in this paper often sounded supportive because they converged on compact scripts such as apology plus validation plus question. That strategy worked especially well for negative emotions, but it remained generic and sometimes content-light. When the situation required nuanced reflection, selective self-disclosure, or precise contextual reasoning, the models flattened the response into a broad supportive template. That limitation is visible in both the cluster analysis and the qualitative examples.

The recommendation for future work follows directly from the measured strengths of the present models. A strong next experiment is a retrieval-grounded small generator that preserves prompt-specific nouns and events while applying the same transparent micro-skill layer that succeeded here. That design joins the two strongest ingredients in the current results: retrieval carried topical detail, and biasing supplied therapist-like form. Future evaluations should also add human judgment for appropriateness, plus explicit risk screening for advice, authority claims, and crisis language. Those extensions would strengthen the benchmark without changing its central

result. The present experiments already establish that lightweight therapist imitation is real, measurable, and bounded.

CONCLUSION

This study shows that small language models can generate therapist-like responses at the level of surface discourse imitation, but not at the level of therapist-like judgment. Across full validation and test evaluations on Empathetic Dialogues, lightweight systems produced measurable acknowledgments, reflections, supportive wording, and follow-up questions; Emotion-TrigramLM+Bias provided the best balance of lexical overlap and therapist-style density, while TFIDF-Retrieval remained the most diverse system. The stable ranking across validation and test strengthens the finding that explicit micro-skill biasing is valuable for transparent, CPU-friendly support models. However, the strongest systems often relied on compact scripts such as apology, validation, and a question, especially for negative emotions, and remained generic when nuanced contextual reasoning was required. Therefore, lightweight therapist imitation should be understood as scope-aware augmentation for low-risk acknowledgment, journaling, or engagement support, not as a replacement for licensed mental health care.

ACKNOWLEDGMENTS

N/A

AUTHOR CONTRIBUTIONS STATEMENT

Yifan Zhang is responsible for writing the article, and Zhongwen Zhou is responsible for conducting the experiments.

REFERENCES

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113-126. https://doi.org/10.1037/0022-3514.44.1.113

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3(2), 71-100. https://doi.org/10.1177/1534582304267187

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2019). Wizard of Wikipedia: Knowledge-powered conversational agents. *In Proceedings of the 7th International Conference on Learning Representations.*

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, 48(1), 43-49. https://doi.org/10.1037/a0022187

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. https://doi.org/10.2196/mental.7785

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health*, 5(4), e64. https://doi.org/10.2196/mental.9782

Hill, C. E. (2014). *Helping skills: Facilitating exploration, insight, and action (4th ed.)*. American Psychological Association. https://doi.org/10.1037/14345-000

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J., Gonnella, J. S., Erdmann, J. B., Veloski, J., & Magee, M. (2001). The Jefferson Scale of Physician Empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement*, 61(2), 349-365. https://doi.org/10.1177/00131640121971158

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. *JMIR mHealth and uHealth*, 6(11), e12106. https://doi.org/10.2196/12106

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). A diversity-promoting objective function for neural conversation models. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 110-119). https://doi.org/10.18653/v1/N16-1014

Lin, C.-Y. (2004). *ROUGE: A package for automatic evaluation of summaries*. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74-81).

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., & Poria, S. (2020). MIME: MIMicking emotions for empathetic response generation. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2020.emnlp-main.721

Miller, A. H., Feng, W., Fisch, A., Lu, J., Batra, D., Bordes, A., Parikh, D., & Weston, J. (2017). *ParlAI: A dialog research software platform*. arXiv preprint arXiv:1705.06476. https://doi.org/10.18653/v1/D17-2014

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). https://doi.org/10.3115/1073083.1073135

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). https://doi.org/10.3115/v1/D14-1162

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370-5381). https://doi.org/10.18653/v1/P19-1534

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21(2), 95-103. https://doi.org/10.1037/h0045357

Shum, H.-Y., He, X.-D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1), 10-26. https://doi.org/10.1631/FITEE.1700826

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. In Advances in Neural Information Processing Systems 27 (pp. 3104-3112).

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *Canadian Journal of Psychiatry,* 64(7), 456-464. https://doi.org/10.1177/0706743719828977

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-art natural language processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45). https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog; do you have pets too?. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 2204-2213). https://doi.org/10.18653/v1/P18-1205