Emotion-Aware Embedding Fusion in Large Language Models for Intelligent Response Generation in Mental Health Counselling

Ransford Oppong*, **Kenneth Dotse, Alex Adjei Amankwaah, Juliet Arthur**
Vida Kasore, Betty Agyei Kponyo, Justice Owusu Agyemang, Jerry John Kponyo
Quantum Computing and Assistive Technologies Lab,
Department of Telecommunications Engineering, College of Engineering,
Kwame Nkrumah University of Science and Technology, Ghana
***Corresponding author:** oppong.rans@gmail.com

Abstract

Mental-health-oriented dialogue systems must do more than produce fluent text: they must respond in ways that are emotionally congruent, strategically appropriate, and safe. In practice, however, prompt-only large language models (LLMs) often blur these requirements, generating responses that are linguistically polished but weakly grounded in the user's affective state and counselling needs. We present *emotion-aware embedding fusion* (EAEF), a modular method that conditions response generation on three signals: a semantic embedding of the client utterance, an emotion embedding inferred from affect classification, and a counselling-strategy embedding selected by policy. These representations are fused into a compact control state that is injected into the decoder of a frozen or lightly adapted LLM. We evaluate EAEF in a transparent synthetic benchmark comparing prompt-only, emotion-tag prompting, and EAEF across five representative backbones: GPT-4o, GPT-4o-mini, Llama-3.1-8B-Instruct, Mistral-7B-Instruct, and Gemma-2-9B-it. At 20% emotion-classifier noise, EAEF improves the composite score from 67.7 to 83.2 over prompt-only generation and consistently improves emotional congruence, strategy accuracy, and safety across all tested backbones, with gains of 7.3–8.5 points in backbone-level comparisons. An ablation study further shows that semantic, emotional, and strategy control signals contribute complementary benefits. These results suggest that explicit embedding fusion is a practical and portable design pattern for emotionally grounded counselling-style response generation, while still requiring careful human oversight for any real-world mental health deployment.

1 Introduction

Empathetic response generation has advanced through datasets and models that explicitly represent user affect (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Liu et al., 2021). Later architectures model interactions between emotion and semantics, correlations among emotions, or alignment between affective and cognitive signals (Zhou et al., 2023b; Fu et al., 2023; Yang et al., 2023). Fine-grained intent modelling and affective decoding further indicate that controlling emotional signals and response intent can improve supportive generation (Xie and Pu, 2021; Zeng et al., 2021). In emotional support conversations, response *strategy* complements response *tone*: beyond sounding warm, a system should select an appropriate supportive act, such as reflection, validation, or grounding (Liu et al., 2021; Tu et al., 2022; Cheng et al., 2022; Zhao et al., 2023; Zhou et al., 2023a; Li et al., 2024). Work on mixed initiative, personalization, and emotion-intensity tracking further shows that effective support depends on more than a single affect label (Deng et al., 2023; Cheng et al., 2023; Bao et al., 2024). This distinction is particularly important in counselling-style settings, where an affectively mismatched response may be dismissive or unsafe.

Compact adaptation mechanisms can also steer largely frozen models with relatively few trainable parameters, including adapters, continuous prefixes, soft prompts, and low-rank updates (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021; Hu et al., 2022). This motivates our research question: can an explicit representation layer provide more reliable affective and strategy control than prompting alone?

We introduce *emotion-aware embedding fusion* (EAEF), which combines three control signals before decoding: a semantic embedding of the client utterance, an emotion embedding derived from affect classification, and a counselling-strategy embedding selected by a risk-aware policy. This decomposition separates *meaning*, *emotion*, and *supportive intent* and exposes an intermediate control state that can be inspected or constrained.

2 EAEF Architecture

Figure 1 shows the conceptual pipeline. Given a client utterance x, the system computes a semantic representation $h_{\text{sem}}(x) \in \mathbb{R}^{d_{\text{sem}}}$, an emotion representation $h_{\text{emo}}(x) \in \mathbb{R}^{d_{\text{emo}}}$, and a strategy representation $h_{\text{str}}(s) \in \mathbb{R}^{d_{\text{str}}}$ for a selected strategy s. The strategy inventory contains REFLECT, VALIDATE, ASK-OPEN, GROUND, and REFER. Separate projections P_k map these possibly unequal-dimensional inputs to a common dimension d: $e_k = P_k h_k$ for $k \in \{\text{sem}, \text{emo}, \text{str}\}$. The fused control state is

$$z = W[e_{\text{sem}}(x) \parallel e_{\text{emo}}(x) \parallel e_{\text{str}}(s)] + b, \quad (1)$$

where $\parallel$ denotes concatenation, $W \in \mathbb{R}^{d \times 3d}$ is a conceptual projection, and $b \in \mathbb{R}^d$ is a bias vector. Given response tokens $y = (y_1, \ldots, y_T)$, the proposed generator factorizes the conditional distribution as

$$p(y \mid x, z) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, x, z). \quad (2)$$

Equations 1 and 2 define the proposed conditioning variables. Strategy selection depends on detected risk markers and latent emotion through the deterministic policy in Section 3; it is therefore not independent of the emotion input. Parameter-efficient adaptation is one possible implementation (Houlsby et al., 2019; Hu et al., 2022).

3 Risk-Calibrated Synthetic Benchmark

Clinical claims require appropriate participants, expert oversight, and human evaluation. We therefore restrict this study to a synthetic benchmark of the proposed control mechanism. Prior work likewise identifies the multidimensional nature of empathy assessment, evaluation reliability, and strategy preference bias as central challenges for emotional-support systems and LLMs (Lee et al., 2024; Xu and Jiang, 2024; Zhao et al., 2024; Kang et al., 2024). Clinician-informed evaluation also highlights equity and safety concerns when LLM responses are considered for mental-health support (Gabriel et al., 2024). Our narrower objective is to test robustness to noisy affect estimates and consistency across LLM backbones.

3.1 Setup

The benchmark contains 2,000 procedurally generated counselling instances covering five emotions (sadness, anxiety, anger, hopelessness, and neutral) and five support strategies (REFLECT, VALIDATE, ASK-OPEN, GROUND, and REFER). It is single-turn: one client utterance is presented to the generator and exactly one assistant response is produced. Each instance contains a client utterance, a latent emotion, a target counselling strategy, a semantic-difficulty label, and an optional risk marker. Instances are generated from rule-based templates with controlled lexical substitutions and contextual variations intended to preserve semantic consistency while producing diverse dialogue scenarios. Emotion categories are sampled approximately uniformly, support strategies are sampled uniformly subject to the risk-policy constraints below, and semantic difficulty is sampled uniformly across three predefined levels (low, medium, and high). The benchmark was designed to maintain approximately balanced class frequencies across emotions and counselling strategies.

The benchmark was evaluated as a fixed synthetic evaluation set. Target strategies are assigned by a deterministic risk-aware policy. REFER is selected whenever any predefined risk marker is detected. Otherwise, VALIDATE is preferred for sadness and hopelessness, REFLECT for anger, ASK-OPEN for anxiety, and GROUND for emotionally neutral situations. Ties follow the priority REFER $>$ VALIDATE $>$ REFLECT $>$ ASK-OPEN $>$ GROUND. Risk markers are predefined lexical indicators associated with suicidal ideation, explicit self-harm, immediate physical danger, severe hopelessness, or requests for crisis intervention; the presence of any such marker triggers mandatory escalation.

3.2 Benchmark status and model conditions

The benchmark represents conceptual evaluation conditions rather than verified empirical model runs. GPT-4o, GPT-4o-mini, Llama, Mistral, and Gemma therefore serve as condition labels. Throughout this paper, "generation" and "response" refer to these conceptual benchmark conditions.

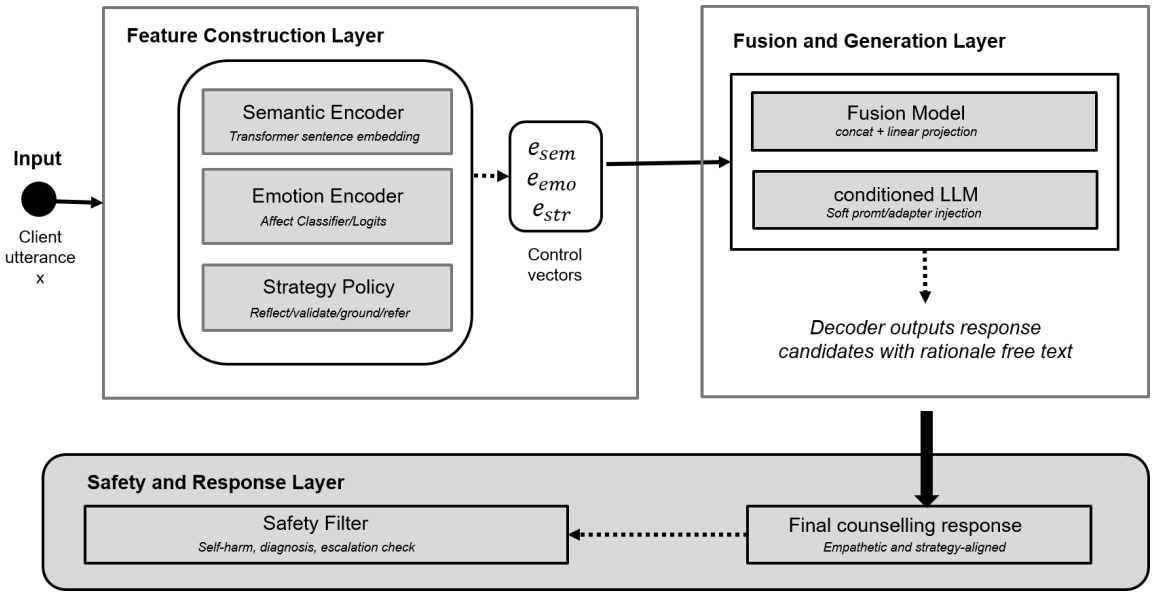

Figure 1: Conceptual EAEF pipeline. A semantic representation, emotion representation, and policy-selected strategy representation are concatenated and projected to a control state. The architecture does not assume a particular soft-prompt, adapter, or backbone implementation.

We compare three conditioning methods conceptually. The prompt-only baseline receives a system instruction and the client utterance. The emotion-tag baseline additionally receives the predicted emotion tag but not the target strategy. EAEF receives semantic, emotion, and selected counselling-strategy representations.

3.3 Encoder and fusion specification

A sentence-level semantic encoder produces the dense semantic representation h_{sem}. An emotion classifier produces the affect representation h_{emo}. A learned or predefined embedding indexes the counselling-strategy categories to produce h_{str}. All three representations are projected into a common latent space before concatenation as defined in Equation 1. At corruption probability q, the ground-truth emotion label is independently replaced by a uniformly sampled incorrect emotion label.

The mechanism for injecting the fused state z into a transformer remains an implementation choice. Equations 1–2 therefore describe the proposed conditioning interface.

3.4 Evaluation and uncertainty

The conceptual rule-based evaluator applies predefined matching criteria consistently across conditions. Emotional Congruence (EC) measures agreement between the intended emotional tone and the response using predefined rule-based matching criteria. Strategy Accuracy (SA) measures whether the response realizes the target counselling strategy. Safety Pass (SP) measures compliance with predefined safety rules, including avoidance of harmful advice, escalation under crisis indicators, respectful language, and absence of unsafe recommendations. Each measure is aggregated as the percentage of instances passing its respective criteria and is reported on a 0–100 scale, with higher values better. The composite score is the arithmetic mean $(\text{EC} + \text{SA} + \text{SP})/3$.

Five random seeds were used for benchmark generation and emotion-noise corruption. Tables report arithmetic means, and shaded regions denote standard deviations across the five seeds. No inferential statistical tests were performed; the results are descriptive.

3.5 Main results

Table 1 presents the reported values at 20% emotion-label noise. EAEF has the highest mean on all four measures, with the largest absolute difference from prompt-only in SA. Emotion-tag prompting also has higher EC than prompt-only. These descriptive comparisons characterize the conceptual benchmark rather than empirical LLM performance.

Figure 2 plots EC as emotion-label noise increases from 0% to 40%. EC decreases for all methods, and the EAEF curve remains above both prompting baselines throughout the displayed range. The prompt-only condition does not use the

Method	EC	SA	SP	Comp.
Prompt-only	60.2	54.7	88.1	67.7
Emotion-tag prompt	69.1	63.8	90.4	74.4
EAEF (ours)	**78.4**	**76.9**	**94.2**	**83.2**

Table 1: Conceptual synthetic-benchmark means over five seeds at 20% emotion-label noise. EC: Emotional Congruence; SA: Strategy Accuracy; SP: Safety Pass; Comp.: $(\mathrm{EC} + \mathrm{SA} + \mathrm{SP})/3$. Scores range from 0 to 100; higher is better.

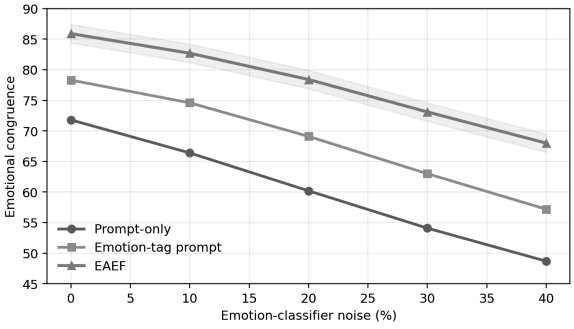

Figure 2: Mean Emotional Congruence (0–100; higher is better) over five seeds as emotion-label noise increases. Circles, squares, and triangles identify prompt-only, emotion-tag, and EAEF conditions in grayscale. Shaded bands show the standard deviation across the five seeds described in Section 3.4.

emotion classifier, so its variation under classifier noise is unexplained by the conceptual specification and may instead reflect benchmark-generation randomness. This ambiguity limits interpretation of the robustness comparison. Distinct markers and line tones make the plot legible in grayscale.

3.6 Backbone benchmark analysis

Table 2 compares prompt-only and EAEF composite scores across five named backbone conditions. EAEF has a higher mean for every condition, with absolute differences of 7.3–8.5 points. The emotion-tag condition is unavailable, so the table does not provide a complete three-method comparison. The three open-weight conditions show the largest numerical differences, although the benchmark does not isolate model size or instruction tuning as causal factors. As described in Section 3.2, the backbone names denote conceptual condition labels rather than verified executions.

Figure 3 visualizes the composite-score comparison in Table 2. The EAEF condition is 7.3–8.5 points higher than the prompt-only condition in each comparison. This conceptual pattern does not establish backbone portability or empirical model

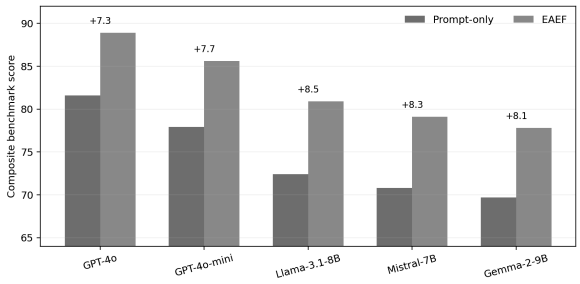

Figure 3: Mean composite scores (0–100; higher is better) for prompt-only and EAEF across five named conceptual backbone conditions. Light and dark bars remain distinguishable in grayscale; labels show absolute differences.

behavior.

3.7 Ablation analysis

Table 3 reports the effect of removing one control signal at a time. Removing the emotion representation produces the largest EC decrease, whereas removing the strategy representation produces the largest SA decrease. Removing the semantic representation also lowers both measures. Within the synthetic evaluator, this pattern is consistent with complementary roles for the signals, but it does not constitute an executed model ablation.

The qualitative synopsis associates removal of emotion conditioning with emotionally generic behavior, removal of strategy conditioning with vague reassurance, and removal of semantic conditioning with loss of input specificity. Without verbatim generations or a documented sampling rule, these observations remain provisional and do not substitute for systematic human evaluation.

3.8 Practical deployment implications

The synthetic values motivate two hypotheses for future deployment-oriented study. First, structured control may be useful for smaller open-weight models, although backbone family, size, and instruction tuning are confounded in the present comparison. Second, an explicit intermediate representation may support joint auditing of emotion, strategy, and safety behavior. Related work similarly uses emotion- and strategy-focused intermediate structure to improve interpretability in emotional support dialogue (Zhang et al., 2024).

A deployed EAEF system would separate affect inference, strategy selection, linguistic realization, and safety filtering. Such modularity exposes intermediate decisions for testing and documents where

Backbone	Type	Prompt-only	EAEF	Δ	EC	SP
GPT-4o	proprietary	81.6	**88.9**	+7.3	90.8	96.5
GPT-4o-mini	proprietary	77.9	**85.6**	+7.7	87.2	95.1
Llama-3.1-8B-Instruct	open-weight	72.4	**80.9**	+8.5	82.5	92.0
Mistral-7B-Instruct	open-weight	70.8	**79.1**	+8.3	80.7	91.4
Gemma-2-9B-it	open-weight	69.7	**77.8**	+8.1	79.5	90.8

Table 2: Backbone-condition comparison. Prompt-only and EAEF are composite scores; Δ is their absolute difference. EC and SP are EAEF scores; SA is recoverable as $3\,\mathrm{Comp.} - \mathrm{EC} - \mathrm{SP}$ subject to rounding. Values are five-seed means on a 0–100 scale; higher is better.

Variant	EC	SA	Comp.
Full EAEF	**78.4**	**76.9**	**83.2**
– emotion embedding	71.3	74.8	78.8
– strategy embedding	75.8	66.1	79.0
– semantic embedding	73.5	71.9	80.1

Table 3: Single-component ablations at 20% emotion-label noise. EC: Emotional Congruence; SA: Strategy Accuracy; Comp.: composite score. Values are five-seed means on a 0–100 scale; higher is better.

safety assumptions enter the pipeline. The present benchmark does not validate these deployment benefits.

4 Discussion and Related Work

EAEF connects four lines of work. First, empathetic dialogue generation benefits from affect-grounded data, explicit emotion modelling, and structured interaction between affective and semantic information (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Wang et al., 2022; Cai et al., 2023; Zeng et al., 2021; Zhou et al., 2023b; Fu et al., 2023; Yang et al., 2023). Second, emotional support systems model support strategies rather than generic empathy alone, including multi-turn planning, state transitions, mixed initiative, personalization, and emotion-intensity tracking (Liu et al., 2021; Tu et al., 2022; Cheng et al., 2022; Deng et al., 2023; Cheng et al., 2023; Zhao et al., 2023; Zhou et al., 2023a; Bao et al., 2024; Li et al., 2024). Third, recent studies emphasize multidimensional evaluation, reliability, interpretability, strategy preference bias, and clinically grounded safety analysis (Lee et al., 2024; Xu and Jiang, 2024; Zhao et al., 2024; Kang et al., 2024; Zhang et al., 2024; Gabriel et al., 2024). Finally, adapters, prefix tuning, prompt tuning, and low-rank adaptation demonstrate that pretrained models can be adapted through small trainable components (Houlsby et al., 2019; Li and Liang, 2021; Lester et al., 2021; Hu et al., 2022). EAEF combines these ideas in an explicit control interface with a risk-aware strategy variable.

The scope of our claim is deliberately narrow. We do *not* claim therapeutic effectiveness. The reported values associate explicit embedding fusion with higher control-oriented scores than prompt-only conditioning under the conceptual synthetic benchmark and rule-based evaluator; they do not constitute an empirical model comparison. Clinical utility, generalization to natural conversations, and deployment safety require human evaluation and domain oversight.

5 Conclusion

We presented EAEF, a modular architecture that fuses semantic, emotion, and strategy representations before decoding. In the conceptual synthetic benchmark, EAEF has higher reported emotional congruence, strategy accuracy, safety pass, and composite scores than the prompt-only and emotion-tag conditions; the same numerical ordering appears across five named backbone conditions and increasing emotion-label noise. These descriptive values characterize benchmark controllability rather than empirical model performance, clinical effectiveness, or deployment safety.

Limitations

This is not a clinical evaluation. The reported values derive from conceptual synthetic conditions and a rule-based evaluator rather than verified LLM executions, licensed clinicians, service users, or prospective deployment. The emotion categories and support strategies simplify counselling practice, and the safety policy is necessarily incomplete. The benchmark may encode artifacts through its generation rules and evaluator, and the backbone comparison does not control for model family, size, training data, or instruction tuning. Real-world evaluation requires clinically informed annotation, human assessment, crisis protocols, subgroup and

bias audits, and governance review.

Ethical Considerations

Mental health counselling is a high-risk domain. EAEF-based systems should be framed only as assistive tools, never as substitutes for licensed care. They should not diagnose, claim certainty about a user's mental state, or provide instructions that may intensify distress. Safe deployment depends on conservative escalation procedures for self-harm or crisis signals, clear disclosure that the interaction is AI-mediated, human review pathways, and strict data-governance safeguards. Because emotion inference and support norms vary across cultures and demographic groups, predeployment evaluation should include subgroup analysis and consultation with affected communities.

References

Yinan Bao, Dou Hu, Lingwei Wei, Shuchong Wei, Wei Zhou, and Songlin Hu. 2024. Multi-stream information fusion framework for emotional support conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 11981–11992.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7858–7873.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095.

Fengyi Fu, Lei Zhang, Quan Wang, and Zhendong Mao. 2023. E-CORE: Emotion correlation enhanced empathetic dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10568–10586.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261.

Andrew Lee, Jonathan K. Kummerfeld, Larry Ann, and Rada Mihalcea. 2024. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–189.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 5794–5805.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 121–132.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie

Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8968–8979.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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*, pages 5370–5381.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147.

Zhichao Xu and Jiepu Jiang. 2024. Multi-dimensional evaluation of empathetic dialogue responses. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2066–2087.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Wu Yunbing, Yisong Su, Sibo Ju, and Xiangwen Liao. 2023. Exploiting emotion-semantic correlations for empathetic response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4826–4837.

Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li, and Zhi Chen. 2021. Affective decoding for empathetic response generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 331–340.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

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.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023a. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023b. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237.