From Heart to Words: Balancing Form and Function in the Expression of Empathy

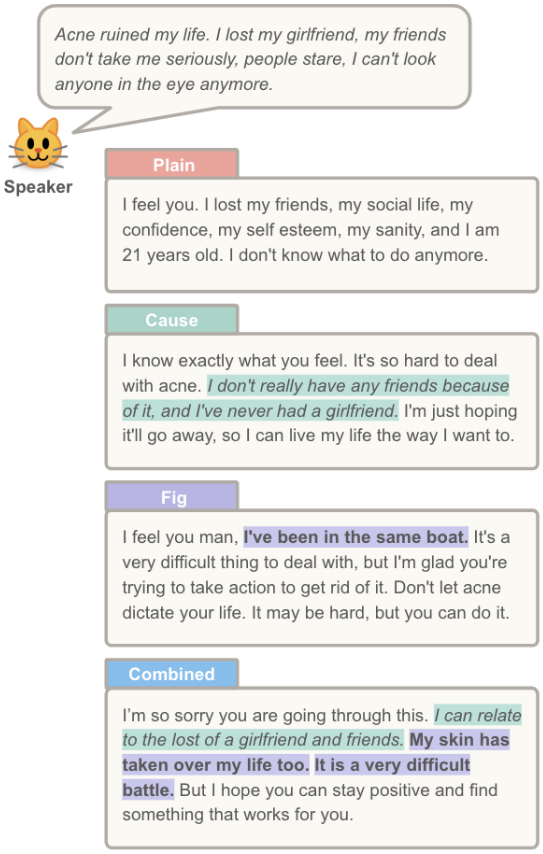
Anonymous ACL submission

Abstract

Empathy is crucial for emotionally connecting with others and providing support, a need that has grown in online communities. While empathy involves understanding others' feelings, effectively communicating that understanding is equally important. In this study, we propose a novel approach to empathetic response generation by combining figurative language with manually annotated empathy causes to address both the linguistic form and semantic context. By integrating these elements, our approach improves multiple dimensions of empathetic responses, achieving a 7.6% improvement in BLEU, a 36.7% reduction in Perplexity, and a 7.6% increase in lexical diversity (D-1 and D-2) in automated evaluations compared to models without these features. Additionally, human assessments show a 24.2% increase in empathy ratings over the same baseline. These findings highlight the synergy between figurative language and empathy causes, offering valuable insights for enhancing empathetic communication across domains.

1 Introduction

Empathy is more than simply acknowledging the emotions of others; it involves communicating that those emotions are understood in a way that makes others feel genuinely supported (Halpern, 2014). Whether through a gentle idiom like "I've got your back" or a careful choice of phrasing, people often reach beyond straightforward language to convey support, including in online communities (Barak et al., 2008; Naslund et al., 2016; Sharma et al., 2020a). There are numerous rhetorical phenomena that people leverage for this purpose. For instance, figurative language (e.g., metaphors or idioms) can enhance emotional expression by making abstract emotions more vivid and relatable (Fussell and Moss, 2014). Consider the two statements: *"I understand it's tough"* (literal) and *"I understand it feels like fighting an endless battle"* (figurative). The latter expression more vividly describes the emotional intensity of the struggle, fostering deeper resonance with the individual seeking support.



Acne ruined my life. I lost my girlfriend, my friends don't take me seriously, people stare, I can't look anyone in the eye anymore.

I feel you. I lost my friends, my social life, my confidence, my self esteem, my sanity, and I am 21 years old. I don't know what to do anymore.

I know exactly what you feel. It's so hard to deal with acne. *I don't really have any friends because of it, and I've never had a girlfriend.* I'm just hoping it'll go away, so I can live my life the way I want to.

I feel you man, **I've been in the same boat.** It's a very difficult thing to deal with, but I'm glad you're trying to take action to get rid of it. Don't let acne dictate your life. It may be hard, but you can do it.

I'm so sorry you are going through this. *I can relate to the lost of a girlfriend and friends.* **My skin has taken over my life too.** **It is a very difficult battle.** But I hope you can stay positive and find something that works for you.

Figure 1: Illustration of empathetic responses generated using different LLM fine-tuning approaches: plain, figurative language, cause annotations, and combined.

The importance of empathetic communication has motivated computational approaches to empathetic response generation. However, existing research in this area has predominantly focused on understanding the content or emotions of the speaker (the individual sharing their struggles) (Rashkin et al.; Welivita and Pu, 2023), rather than on *how* the responder (the individual providing support) can convey empathy. While some recent studies, such as the work by Welivita et al. (2023), have begun exploring communication strategies by introducing response intents (e.g., agreeing or suggesting), these approaches rely on aligning responses with speakers' emotions rather than employing linguistic tools to enhance the delivery of empathy. This highlights a gap in exploring how nuanced language—such as figurative language—can be used to effectively convey empathy, particularly when informed by the speaker's context.

In this work, we address this gap by examining how both nuanced linguistic form (focusing on figurative language) and semantic context (focusing on empathy cause) can improve empathetic response generation. We hypothesize that figurative language enriches emotional expression, while the identification and incorporation of empathetic cause helps to ensure responses are contextually aligned to the speaker's concerns. We further posit that together, these linguistic and contextual elements work in synergy to generate responses that are both emotionally engaging and contextually targeted. To investigate these hypotheses, we fine-tune large language models (LLMs) on the `AcnEmpathize` dataset (Lee and Parde, 2024), enriching the original data with both figurative language and manually annotated empathy causes (see Figure 1). We summarize our contributions as:

- We propose to incorporate figurative language into empathetic response generation, and show that it improves overall response quality across linguistic, emotional, and contextual dimensions.
- We contribute an additional layer of manual annotations for empathy cause to the AcnEmpathize dataset (Lee and Parde, 2024) and demonstrate their application in generating more contextually aligned responses.
- We integrate figurative language and empathy cause for empathetic response generation, significantly improving performance across both automated and human evaluation metrics.

2 Related Work

The importance of effectively communicating empathy has been demonstrated across multiple domains such as neuroscience, (Decety and Jackson, 2004), medicine (Riess and Kraft-Todd, 2014), and therapy (Green et al., 2005). These studies suggest how empathetic communication fosters trust, strengthens emotional connections, and improves medical and therapeutic outcomes.

Motivated by these findings, computational approaches have sought to automate empathetic response generation. A large portion of research in this area has focused on identifying and modeling the speaker's emotions. For example, Rashkin et al. introduce the widely used `EmpatheticDialogues` dataset, annotated with emotion labels to guide empathetic response generation. This dataset has inspired numerous studies, including work by Lin et al. (2020) and Majumder et al. (2020), which rely on these labels to generate responses.

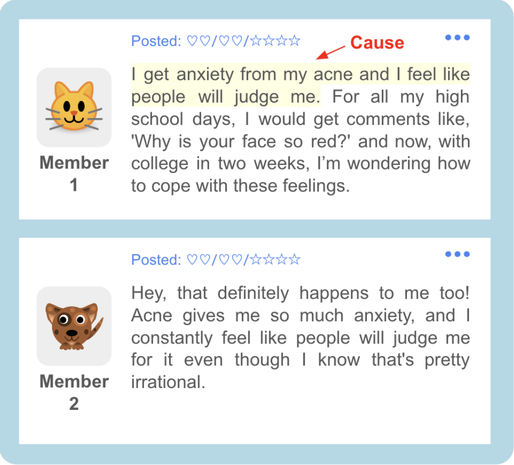
More firmly in the emotion recognition arena, researchers have explored emotion cause recognition to deepen the understanding of the speaker's context. Gao et al. (2021) and Li et al. (2021) analyze emotion causes in the `EmpatheticDialogues` dataset to identify the specific triggers behind the speaker's emotions. Similarly, Qian et al. (2023) incorporate both emotion and emotion cause recognition to train LLMs for empathetic response generation. Qian et al. (2023)'s work in particular aligns closely with our intent, although we note that empathy cause (our focus) carries subtle differences from emotion cause (See Section 3.2 for detail).

Recent works have also started considering communication strategies to guide empathetic response generation. Welivita et al. (2021) introduce a dataset combining emotion labels and response intents, such as agreeing or suggesting. In their subsequent work (Welivita et al., 2023), they demonstrate how these intents could guide the generation of emotionally supportive and empathetic responses. Similarly, Saha et al. (2022) incorporate rewriting strategies using reinforcement learning to enhance empathy in response content.

While these communication strategies contribute to shaping empathetic responses, they do not explore the linguistic tools that can further enhance these responses. Recent work on empathetic storytelling (Shen et al., 2024) shows how narrative style elements—such as tone and phrasing in the speaker's text—can influence perceived empathy. Although outside the scope of response generation, this demonstrates the potential for leveraging linguistic tools for empathetic response generation.

Figurative language is a powerful linguistic tool that enriches emotional expression (Fussell and Moss, 2014). Metaphors, in particular, make abstract emotions more relatable by enhancing the vividness and emotional impact of communication (Citron and Goldberg, 2014). Computational studies have shown that incorporating figurative language—specifically metaphors, idioms, and hyperbole—can improve predictions of both emotion (Lee et al., 2024a) and empathy (Lee et al., 2024b). Despite the clear value of figurative language to empathetic expression, as highlighted by these works, figurative language remains underexplored in empathetic response generation.

Building on the success of emotion cause annotations in improving response generation (Gao et al., 2021), we integrate *figurative language* and *empathy cause annotations* to address both the linguistic and contextual aspects of empathetic response generation. By carefully balancing these two elements, we aim to create responses that are not only emotionally engaging but also well-aligned with the speaker's concerns.



I get anxiety from my acne and I feel like people will judge me. For all my high school days, I would get comments like, 'Why is your face so red?' and now, with college in two weeks, I'm wondering how to cope with these feelings.

Hey, that definitely happens to me too! Acne gives me so much anxiety, and I constantly feel like people will judge me for it even though I know that's pretty irrational.

Figure 2: Example of an initial (speaker) post and an empathetic reply in the AcnEmpathize dataset. The highlighted portion in the speaker's post indicates the annotated cause that evokes empathy in the given reply.

3 AcnEmpathize Dataset

3.1 Dataset Description

We use the publicly available AcnEmpathize dataset (Lee and Parde, 2024) as the basis for our work. This dataset captures authentic emotional exchanges from acne.org, an online acne support community. It focuses on posts from the "Emotional and Psychological Effects of Acne" forum where users discuss their emotional struggles stemming from acne. The dataset features over 12K posts categorized into initial posts, written by individuals seeking support, and responses from others replying to emotional challenges. For this study, we focus on a subset of 2,492 posts in speaker-response pair format, where all responses contain empathy. These pairs correspond to 1,110 unique speaker posts, as many posts received multiple empathetic replies. This domain-specific, emotion-rich dataset forms the foundation for our work on generating and evaluating empathetic responses.

3.2 Cause Annotations

We annotated empathy cause in our subset of speaker posts to enable the generation of more targeted responses (see Figure 2 for an example of an annotated cause in a speaker-response pair), and we release these annotations publicly as an additional layer of data available for AcnEmpathize. Empathy causes refer to the specific sentences within a speaker's post that elicit an empathetic response from a responder. Unlike traditional emotion cause extraction, which identifies triggers of expressed emotions (e.g., sadness, anger) within the speaker's text (Xia and Ding, 2019; Chen et al., 2010), empathy causes highlight the textual elements that prompt an empathetic reply. These annotations help models to stay contextually relevant to the speaker's concerns.

To annotate cause sentences, we used the collaborative tool INCEpTION (Klie et al., 2018). We recruited three graduate student volunteers with formal training in natural language processing at a U.S.-based institution. Annotators were instructed to highlight cause sentence(s) in each speaker post that were most likely to prompt the corresponding empathetic reply across three rounds.

In Round 1, annotators independently labeled 10 identical conversations and participated in a discussion afterward to resolve all disagreements, resulting in eventual perfect inter-annotator agreement (IAA) using Krippendorff's alpha (Krippendorff, 1970). In Round 2, annotators labeled 90 additional identical conversations, resulting in an initial IAA of 0.70. Disagreements in this round were also resolved through discussion to reach full consensus. Pre-consensus pairwise IAA scores for the total of 100 triple-annotated samples ranged from 0.67 to 0.73, consistent with the IAA score of 0.68 reported in an existing empathy study (Sharma et al., 2020b). For the final round, the remaining conversations were divided among the annotators in a

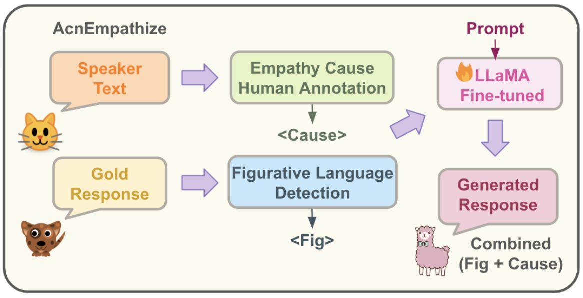


Figure 3: Overview of our pipeline for empathetic response generation. Speaker texts are manually annotated with empathy causes, while figurative language in responses is identified using a detection method. These elements are integrated during the fine-tuning of LLaMA, guided by prompts, to generate empathetic responses that are both contextually relevant and linguistically expressive.

Language Type	# Posts (%)
Idiom	1,225 (49.16%)
Metaphor	887 (35.59%)
Hyperbole	559 (22.43%)
Total Figurative	1,723 (69.14%)

Table 1: Distribution of figurative language type (idioms, metaphors, and hyperbole) in responses within the cause-annotated AcnEmpathize dataset. The counts represent the number of response posts that contain each type of figurative language. Each response may contain more than one type.

ratio of 476:476:150.[1]

3.3 Figurative Language Annotations

We did not manually annotate figurative language in AcnEmpathize; however, we automatically identified the presence of metaphors, idioms, and hyperboles. This was done using a similar technique to that proposed by Lee et al. (2024b) for empathy detection, and incorporated directly into our empathetic response generation approach. We describe this more in Section 4. This integration allows us to explore how linguistic and semantic elements jointly enhance empathetic responses.

4 Empathetic Response Generation

Using the new layer of cause annotations for the AcnEmpathize dataset, we sought to generate empathetic responses that meaningfully address the struggles expressed in speaker posts through both emotional engagement and contextual alignment. Prior work (Lee et al., 2024b) has shown that incorporating figurative language—specifically idioms, metaphors, and hyperbole—meaningfully enhances empathy detection. Motivated by this finding, we detect these figurative expressions in response texts using a prompt-based approach to incorporate them into empathy generation. We study their use both independently and in concert with empathy cause annotations when generating emapthetic responses. An overview of our full pipeline, incorporating both figurative language detection and empathy cause annotations, is presented in Figure 3. Although gold empathy cause labels are used in this study to demonstrate proof of concept and assess the contribution of high-quality cause labels to empathetic response generation, a promising future direction (and one facilitated by the new cause labels) involves the automated detection of empathy cause alongside automated figurative language detection.

4.1 Figurative Language Detection

We detect idioms, metaphors, and hyperbole in the response texts of the dataset using the method proposed by Lai et al. (2023), which leverages a multitask framework built on mT5 (Xue, 2020). This method identifies figurative language through template-based prompt learning. Specifically, we use the detection prompt:

```
Which figure of speech does this
text contain? (A) Literal (B)
[Task] | Text: [Text]
```

In the prompt, `[Task]` corresponds to one of the figurative language types: idiom, metaphor, or hyperbole. Each sentence in the response text is iteratively assessed for each figurative language type, and the results are recorded as binary indicators.

As shown in Table 1, approximately 69% of empathetic replies contain figurative language, with idioms being the most common (49.16%), followed by metaphors (35.59%) and hyperbole (22.43%). These statistics highlight the frequent use of figurative language in empathetic responses within the dataset, further justifying its incorporation alongside cause annotations to enrich responses both linguistically and semantically.

4.2 Experiments

We explore different approaches for generating empathetic responses by fine-tuning the LLaMA-3-8B

[1] One annotator had an unavoidable and unexpected schedule constraint.

model (Touvron et al., 2023) using our gold cause annotations and silver (automatically detected) figurative language annotations. Before fine-tuning, we evaluated its out-of-the-box performance using zero-shot empathetic response generation. This served as a baseline for comparison with the fine-tuning strategies we explore: a plain approach, a cause-based approach, a figurative language-focused approach, and a combined approach that integrates both cause annotations and figurative language. Unlike the zero-shot baseline, the fine-tuned models are trained to learn response patterns and expressions present in the dataset. Each approach employs different prompts tailored to the specific objectives of the fine-tuning strategy.

Plain. In plain fine-tuning, the model is trained without cause or figurative language annotations. Thus, the model learns to generate empathetic responses based on the natural patterns and style of replies present in the dataset. During generation, the following prompt is used:

```
Given the input text, generate an
empathetic response.
```

Cause. In this approach, we fine-tune the model using cause annotations in the speaker posts, as described in Section 3.2. We specify gold-standard causes by wrapping them in `<cause>` tags. The prompt correspondingly acknowledges these tags:

```
Given the input text with
<cause> tags, generate a
targeted empathetic response
that acknowledges the specific
concerns expressed.
```

Figurative Language (Fig). In this approach, we fine-tune the model using figurative language identified in the response texts, based on the detection method outlined in Section 4.1. We classify sentences containing idioms, metaphors, or hyperbole, and mark the classified sentences with `<idiom>`, `<metaphor>`, and `<hyperbole>` tags. These tags expose the model to examples of how figurative expressions enhance the emotional depth of empathetic replies. For generation, we design the prompt to flexibly include figurative language where appropriate, without explicitly specifying particular expressions:

```
Given the input text, generate
an empathetic response that
uses figurative language where
appropriate, specifically
<idiom>, <metaphor>, or
<hyperbole>.
```

Combined (Fig + Cause). Our final approach integrates both figurative language and cause annotations to enrich responses both linguistically and semantically. Speaker texts are tagged with `<cause>` tags around labeled triggers of empathy, while response texts are tagged with `<idiom>`, `<metaphor>`, and `<hyperbole>` during training. During generation, we use the following prompt:

```
Given the input text, generate
an empathetic response that
very strongly emphasizes the
use of figurative language,
specifically <idiom>, <metaphor>,
and <hyperbole>, optimizing
<idiom> and <metaphor> to
maximize emotional support,
while addressing the concerns
indicated by the <cause> tags.
```

The emphasis on idioms and metaphors in the prompt is motivated by a prior study on empathy detection using the same dataset (Lee et al., 2024b), which shows their statistically significant association with empathy labels. While hyperbole remains a useful linguistic device, idioms and metaphors are prioritized to maximize the emotional supportiveness of the responses, with cause annotations incorporated to maintain contextual relevance.

Training Details

All experiments utilized a 4-bit quantized version of the LLaMA-3-8B model, implemented with the FastLanguageModel framework[2] to optimize memory usage and computational efficiency. For the fine-tuning approaches, the model was trained on both speaker posts and responses to generate empathetic outputs. The model was fine-tuned using a batch size of 1, a learning rate of 5e-5, and the AdamW optimizer in 8-bit mode. Training was conducted for three epochs using three NVIDIA 2080 TI GPUs with FP16 or BF16 support, utilizing the PEFT framework with a LoRA (Hu et al., 2021) configuration (rank = 16, alpha = 16, dropout = 0).

[2] We implemented it using the unsloth GitHub repository https://github.com/unslothai/unsloth.

5 Evaluation

We compare conditions using both automated metrics and human evaluation to provide a well-rounded assessment of the generated responses, encompassing linguistic quality, lexical alignment, and empathetic support. We describe both evaluation frameworks below.

5.1 Automatic Evaluation

We assess different aspects of response quality using perplexity (PPL), BLEU, and Distinct-1 (D-1) and Distinct-2 (D-2) scores. PPL is measured using the Hugging Face transformers library (Wolf et al., 2020) to evaluate the likelihood of gold responses under the model's probability distribution, indicating fluency and coherence. BLEU evaluates lexical overlap between generated and gold responses (Papineni et al., 2002). We compute it using sentence_bleu from the NLTK library (Bird et al., 2009), averaging across multiple gold responses, with smoothing function applied to handle sparsity (Chen and Cherry, 2014). Finally, we calculate the ratio of unique unigrams (D-1) and bigrams (D-2) to the total number of tokens (Li et al., 2015) (also using the NLTK library) to measure the lexical diversity of generated responses.

5.2 Human Evaluation

Additionally, we conducted a human evaluation to provide a more holistic assessment of the generated responses. The evaluation was performed by three graduate student volunteers with formal training in NLP at a U.S.-based institution.[3] The annotators rated the responses based on the following criteria inspired by Rashkin et al.:

- **Empathy**: The response's ability to demonstrate understanding of the speaker's feelings.
- **Relevance**: The extent to which the response is appropriate and on-topic.
- **Fluency**: The ease of understanding and linguistic clarity.

We randomly sampled 111 sets of four generated responses, each corresponding to a single speaker post.[4] To minimize potential bias, the responses generated from each of the four approaches were shuffled before being presented to annotators. Annotators were also asked to select the response they consider **Most Supportive**, beyond the individual scores for Empathy, Relevance, and Fluency. These counts were weighted based on majority agreement (i.e., responses selected by at least two annotators) for evaluation. Complete evaluation guidelines, including definitions and examples for each criterion, are provided in Appendix A.

Approach	PPL (↓)	BLEU (↑)	D-1 (↑)	D-2 (↑)
Zero-shot	14.944	0.058	0.587	0.847
Plain	14.182	0.764	0.515	0.751
Cause	13.990	0.772	0.523	0.755
Fig	9.100	0.775	0.561	**0.814**
Combined	**8.980**	**0.822**	**0.569**	**0.814**

Table 2: Performance of the zero-shot baseline and fine-tuned approaches on automated metrics for empathetic response generation. *Combined* shows the best overall performance. While *Zero-shot* achieves the highest D-1 and D-2 scores, it is excluded from further evaluation due to its extremely low BLEU score (0.058). (↑) means higher is better, (↓) means lower is better.

Approach	E	R	F	Most Supportive (%)
Plain	3.565	3.631	3.207	3.19%
Cause	3.889	4.024	3.799	11.70%
Fig	4.195	4.021	4.132	32.98%
Combined	**4.426**	**4.135**	**4.189**	**52.13%**

Table 3: Performance on the fine-tuned approaches in human evaluation for empathetic response generation. The values represent the average scores for Empathy (E), Relevance (R), and Fluency (F) across all evaluated samples, along with the Most Supportive (%) column reflecting the percentage of responses selected as Most Supportive by the majority of annotators. *Combined* achieves the best overall performance across all metrics.

6 Results

6.1 Automatic Evaluation

Table 2 summarizes the results of evaluating the Zero-shot baseline and fine-tuned approaches using the automated metrics. *Combined (Fig + Cause)* achieves the best overall performance, balancing highest BLEU (0.822) and lowest Perplexity (8.980) with competitive D-1 (0.569) and D-2 (0.814) scores. This suggests that balancing figurative language and cause annotations yields responses that are not only linguistically diverse but also coherent and contextually aligned.

After *Combined*, *Fig* achieves the best

[3] Two annotators are not involved in the cause annotation process, while one annotator participated in both tasks.

[4] These 111 sets represent 10% of the 1,110 unique speaker posts used for generation.

overall performance, reducing PPL by 35.8% (14.182→9.100), increasing D-1 by 8.9% (0.515→0.561), D-2 by 8.4% (0.751→0.814), and BLEU by 1.4% (0.764→0.775) compared to *Plain*. While both *Fig* and *Cause* enhance response quality, *Fig* has a more pronounced impact on lexical diversity (D-1, D-2) and overall coherence (PPL), whereas *Cause* demonstrates modest improvements in semantic alignment, with a 1.1% increase in BLEU (0.764→0.772).

In contrast, the *Zero-shot* baseline, despite achieving the highest diversity scores (D-1: 0.587, D-2: 0.847), shows extremely poor performance in other key metrics. Its BLEU score is especially low at 0.058 and it has the highest PPL (14.944), reflecting poor fluency and alignment with gold responses. Due to these limitations, we excluded *Zero-shot* from further human evaluation and analysis to ensure a more meaningful comparison across fine-tuned approaches.

6.2 Human Evaluation

The results of human evaluation on the fine-tuned approaches are summarized in Table 3. Similar to the performance on automated metrics, *Combined (Fig + Cause)* demonstrates the best overall performance across all criteria, achieving the highest average scores for Empathy (4.426), Relevance (4.135), and Fluency (4.189). *Fig* follows closely, with strong scores for Empathy (4.195) and Fluency (4.132). It also performs comparably to *Cause* in Relevance (4.021 vs. 4.024), indicating that figurative language alone can align responses well with the speaker's context. *Cause* excels in Relevance (4.024), confirming that it effectively addresses the content of the speaker's text, with less pronounced scores for Empathy (3.889) and Fluency (3.799) compared to *Fig* and *Combined*. In contrast, *Plain* lags behind across all three metrics, with the lowest scores for Empathy (3.565), Relevance (3.631), and Fluency (3.207).

For the Most Supportive metric, which reflects perceived supportiveness (as described in Section 5), 84.7% (94 out of 111) of evaluated samples reach majority agreement, with at least two annotators selecting the same response. Among these, the responses generated by *Combined (Fig + Cause)* are selected the most frequently as being the Most Supportive, taking up 52.13% (49 responses out of 94) of evaluated samples. *Fig* follows, with 32.98% (31 responses out of 94), while *Cause* and *Plain* are selected less frequently, with 11.7% (11 responses) and 3.19% (3 responses) of evaluated samples, respectively.

Approach	tone_pos	prosocial	cogproc	adj
Gold	3.495	0.966	15.940	6.910
Plain	2.686	0.892	15.147	6.249
Cause	2.478	0.814	15.347	6.181
Fig	3.560	1.414	**16.304**	6.718
Combined	**3.603**	**1.425**	16.136	**6.927**

Table 4: LIWC analysis of *Gold* and generated responses across different approaches. The selected features include tone_pos, reflecting positive tone; prosocial, capturing supportive language; cogproc, representing cognitive engagement and contextual reasoning; and adj, measuring linguistic richness through descriptive adjectives.

7 Analysis of Generated Responses

In this section, we analyze the generated responses to gain deeper insights into various dimensions of empathetic expression.

Psycholinguistic Insights

By using LIWC (Tausczik and Pennebaker, 2010) psycholinguistic features, we examine emotional, social, cognitive, and linguistic aspects related to empathy in generated and gold responses (*Gold*). We use the LIWC 2022 edition[5] to extract and select four psycholinguistic features from each response:

- **tone_pos**: Encompasses words related to positive emotions (Tausczik and Pennebaker, 2010). Their presence can contribute to creating uplifting and supportive responses.
- **prosocial**: Captures social supportiveness, reflecting language that signals a willingness to help or show care (Pennebaker et al., 2015).
- **cogproc**: Indicates cognitive engagement, such as reasoning and understanding. Ensures that the response is relevant and thoughtful.
- **adj**: Measures the use of descriptive adjectives, capturing the vividness and expressiveness of the responses.

Table 4 provides results for these selected features across the generated and gold responses. *Combined (Fig + Cause)* demonstrates the most well-rounded performance, surpassing both *Gold*

[5] https://www.liwc.app/

and other generated methods in most metrics (tone_pos: 3.603, prosocial: 1.425, adj: 6.927), except in cogproc, where it ranks second (16.136 vs. *Fig*: 16.304). Overall, it effectively balances positive tone, social supportiveness, cognitive engagement, and linguistic richness in the generated responses.

Fig also excels, achieving the highest cognitive engagement score in cogproc (16.304 vs. Gold: 15.940). This demonstrates how figurative language can enhance reasoning and thoughtful engagement beyond emotion expression. It also significantly boosts the score for prosocial (1.414 vs. *Gold*: 0.966) which makes it particularly effective in shaping socially supportive responses. While *Fig* does not surpass *Gold* in adj (6.718 vs. *Gold*: 6.910), it remains competitive in its role to leverage descriptive adjectives in empathetic text.

Cause shows nuanced results, with a slight improvement in cogproc (15.347, an increase from *Plain*: 15.147) but a lower tone_pos score (2.478 vs. *Plain*: 2.686). This suggests that while *Cause* enhances reasoning, it may benefit from complementary strategies to elevate positive tone and social supportiveness. These findings highlight the synergy between *Cause* and *Fig*, as evidenced by *Combined*, which effectively balances their strengths to enhance empathetic responses.

What Makes a Response Supportive?

We extend our analysis beyond all generated responses to focus on the Most Supportive responses, identified by majority agreement during human evaluation (see Section 5). These responses were compared against others to explore the role of Empathy, Relevance, and Fluency in determining perceived supportiveness.

Response Type	Empathy	Relevance	Fluency
Most Supportive	**4.67**	**4.37**	**4.40**
Other Responses	3.78	3.80	3.64

Table 5: Average scores for Empathy, Relevance, and Fluency in Most Supportive and Other Responses. All differences were tested using a paired t-test and found statistically significant ($p < 0.001$).

Our analysis reveals that a balance among Empathy, Relevance, and Fluency is critical for perceived supportiveness (See Table 5). While Empathy scores were consistently high for Most Supportive responses (average: 4.67), high empathy alone was insufficient. When we observed responses that received a perfect empathy score (5) but weren't selected as being the Most Supportive, 64.71% (22 out of 34) of such responses had lower fluency scores (average: 3.18 vs. Most Supportive: 4.40). Similarly, 35.29% (12 out of 34) of responses with perfect empathy scores that were not selected as Most Supportive had lower relevance (average: 3.78 vs. Most Supportive: 4.37). While low relevance may have some influence, it appears to be a less critical breaking factor than fluency. This is supported by our effect size analysis using Cohen's d (fluency: 1.24, relevance: 0.85), aligning with research that frames supportiveness as a multidimensional construct requiring high empathy, contextual alignment, and linguistic clarity (Cutrona, 1990; Halpern, 2001; Burleson, 2003).

8 Conclusion

In this study, we introduced a novel approach to empathetic response generation by integrating figurative language and manually annotated empathy causes. Our approach significantly improves the quality of responses across emotional, contextual, and linguistic dimensions. This improvement is reflected in automated metrics, with BLEU improving by 7.6% (0.764 → 0.822), PPL reducing by 36.7% (14.182 → 8.980), and lexical diversity increasing by an average of 7.6% (D-1: 0.515 → 0.569, D-2: 0.751 → 0.814) from *Plain* fine-tuning. Human evaluation further confirmed these results, with *Combined (Fig + Cause)* achieving the highest average ratings for Empathy (4.426, +24.2%), Relevance (4.135, +13.9%), and Fluency (4.189, +30.6%) out of 5. These findings, supported by our psycholinguistic analysis, underscore the synergy between figurative language and empathy causes. By effectively balancing linguistic expressiveness with semantic alignment, this work advances empathetic response generation, moving beyond conventional approaches focused solely on understanding speaker's emotions.

9 Limitations

Our study is limited in several aspects. Human evaluation inherently involves subjectivity, which could introduce variability in assessing empathy, relevance, and fluency. Efforts were made to mitigate this by carefully crafting an evaluation guideline and shuffling responses; however, subjectivity remains a potential limitation. Additionally, the findings are based on the AcnEmpathize dataset, which focuses on an acne support community. The results may not necessarily generalize to other contexts. In follow-up work, researchers are encouraged to test and adapt these strategies to diverse domains that require support.

10 Ethical Considerations

This study utilizes the AcnEmpathize dataset, which is based on publicly available and anonymized data, ensuring compliance with ethical standards for research involving online communities. The dataset does not include any personal identifying information, and all annotation tasks were conducted by volunteers who were informed about the research goals and methods. The dataset and annotations are intended solely for research purposes, with the aim of advancing empathetic communication through computational methods.

Acknowledgments

Parts of the writing in this paper were paraphrased with input from ChatGPT.

References

Azy Barak, Meyran Boniel-Nissim, and John Suler. 2008. Fostering empowerment in online support groups. *Computers in human behavior*, 24(5):1867–1883.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brant R Burleson. 2003. The experience and effects of emotional support: What the study of cultural and gender differences can tell us about close relationships, emotion, and interpersonal communication. *Personal relationships*, 10(1):1–23.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the ninth workshop on statistical machine translation*, pages 362–367.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595.

CE Cutrona. 1990. Type of social support and specific stress: Toward a theory of optimal matching. *Social support: An interactional view/Wiley*.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

David Green et al. 2005. *Troubled talk: Metaphorical negotiation in problem discourse*, volume 15. Walter de Gruyter.

Jodi Halpern. 2014. From idealized clinical empathy to empathic communication in medical care. *Medicine, Health Care and Philosophy*, 17:301–311.

Jordi Halpern. 2001. From detached concern to empathy: Humanizing medical practice.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations*, pages 5–9.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. *arXiv preprint arXiv:2306.00121*.

Gyeongeun Lee and Natalie Parde. 2024. Acnempathize: A dataset for understanding empathy in dermatology conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 143–153.

Gyeongeun Lee, Zhu Wang, Sathya N Ravi, and Natalie Parde. 2024a. Empatheticfig at wassa 2024 empathy and personality shared task: Predicting empathy and emotion in conversations with figurative language. In *Proceedings of the 14th Workshop on Computational*

Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 441–447.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024b. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 519–529.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2041–2045.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13622–13623.

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. *arXiv preprint arXiv:2010.01454*.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

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

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

H Rashkin, EM Smith, M Li, and YL Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. arxiv 2018. *arXiv preprint arXiv:1811.00207*.

Helen Riess and Gordon Kraft-Todd. 2014. Empathy: a tool to enhance nonverbal communication between clinicians and their patients. *Academic Medicine*, 89(8):1108–1112.

Tulika Saha, Vaibhav Gakhreja, Anindya Sundar Das, Souhitya Chakraborty, and Sriparna Saha. 2022. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2650–2656.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *arXiv preprint arXiv:2405.17633*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Anuradha Welivita and Pearl Pu. 2023. Use of a taxonomy of empathetic response intents to control and interpret empathy in neural chatbots. *arXiv preprint arXiv:2305.10096*.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Anuradha Welivita, Chun-Hung Yeh, and Pearl Pu. 2023. Empathetic response generation for distress support. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, 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*, pages 38–45.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Appendix

A.1 Human Evaluation Guideline

Human annotators were provided with a common file containing 111 samples, each consisting of a speaker text and four generated responses (see Table 6). Each annotator was given an individual evaluation file (see Table 7) to record scores for the generated responses with the following instruction:

- **Instruction**: For each entry, read the "Speaker text" and four responses. Rate empathy (E), relevance (R), and fluency (F) for each response on a scale of 1-5 (1: not at all, 3: somewhat, 5: very likely), using the format ERF. Finally, choose the response that feels most supportive.

They were also provided with the definitions and examples of each measure, as detailed below.

A.1.1 Empathy (E)

- **Definition**: Does the response show understanding of the speaker's feelings?

  *Note: Empathy doesn't necessarily involve having the exact same experience or simply agreeing to the speaker. If a response includes any of the following empathy communication mechanisms (adapted from Sharma et al. (2020b)), you can assign at least a 3.

  - **Emotional Reactions**: Does the response express or allude to warmth, compassion, concern, or similar feelings of the responder towards the seeker? (e.g., *Everything will be fine*; *I feel really sad for you.*)
  - **Interpretations**: Does the response communicate an understanding of the seeker's experiences and feelings? In what manner? (e.g., *I understand how you feel*; *This must be terrifying*; *I also have anxiety attacks at times which makes me really terrified.*)
  - **Explorations**: Does the response make an attempt to explore the seeker's experiences and feelings? (e.g., *What happened?*; *Are you feeling alone right now?*)

- **Example**:

  Speaker text: I have acne and worry that my boyfriend will think it's gross.

  Responses:

  - *Just get over it.* (1)
  - *A lot of people worry about their acne around others.* (3)
  - *I completely understand feeling self-conscious about acne, especially around people who matter to you. I've felt that way too.* (5)

A.1.2 Relevance (R)

- **Definition**: Is the response appropriate to the conversation? Is it on-topic?

- **Example**:

  Speaker text: I have acne and worry that my boyfriend will think it's gross.

  Responses:

  - *I hope to get hired soon.* (1)
  - *A lot of people feel self-conscious about their skin.* (3)
  - *It's understandable to feel self-conscious about acne around someone you care about, like your boyfriend.* (5)

A.1.3 Fluency (F)

- **Definition**: Is the response easy to understand? Does it flow smoothly?

- **Example**:

  Speaker text: I hate acne.

  Responses:

  - *I acne understand your concerns about.* (1)
  - *Acne is annoying. It is tiring. It is bad.* (3)
  - *I understand your frustration with acne. It's tough to deal with every day, and it can be tiring.* (5)

A.1.4 Most Supportive

- **Definition**: Imagine you are the person who shared the concerns in the "Speaker text" column. Which of the four responses ("Response 1", "Response 2", "Response 3", "Response 4") would make you feel the most supported?

Index #	Speaker text	Response 1	Response 2	Response 3	Response 4
1	At the end of June I'm moving in with my current boyfriend. It makes me sad, however, that the ONLY thing that is bothering me about moving in with him is my acne.	I totally understand how you feel. My boyfriend has been so supportive, but I think it's just natural for me to feel embarrassed about my skin around him.	I feel the same way. My boyfriend is always telling me how beautiful I am and that he doesn't care about my acne, but I still feel like he deserves better.	I know exactly how you feel. I've been living with my boyfriend for 3 years now and was really self-conscious about my skin. It's tough, but I've learned that you can't let it get to you.	I can relate to this. I've been seeing my boyfriend for 4 months now and he has seen me at my worst. It's a tough situation. I'm sure you'll figure it out.
⋮	⋮	⋮	⋮	⋮	⋮

Table 6: Example common file entries provided to annotators, showing the speaker text and four generated responses used for evaluation.

Index #	Speaker text	Response 1	Response 2	Response 3	Response 4	Most Supportive
1	At the end of June I'm moving in with my current boyfriend. It makes me sad, however, that the ONLY thing that is bothering me about moving in with him is my acne.	ERF	ERF	ERF	ERF	Choose from 1-4
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7: Example evaluation entries provided to annotators for scoring Empathy (E), Relevance (R), Fluency (F), and selecting the Most Supportive response from the four responses.