Behavior vs Representation: A Minimal-Pair Analysis of Contextual Sensitivity in Empathetic Dialogue Models

Anonymous ACL submission

Abstract

Empathetic dialogue systems often generate supportive responses; however, it remains unclear whether these models can truly distinguish between semantically distinct emotional contexts. Current evaluation approaches mostly focus on response quality, providing limited information about contextual sensitivity and internal representation behaviors. In this study, we propose a controlled minimal-pair evaluation framework to investigate contextual sensitivity in empathy-focused language models. By creating semantically minor yet emotionally critical perturbations, we analyze both response-level adaptation and layer-wise hidden-state divergence analysis. Our findings suggest that responses that appear empathetic at a superficial level may not always reliably reflect true contextual understanding.

Keywords: empathetic dialogue, minimal-pair evaluation, contextual sensitivity, representation analysis, hidden-state analysis

1 Introduction

Empathetic dialogue systems have become a major research topic in the field of natural language processing, alongside the growing prevalence of conversation-based AI applications in psychological support, education, healthcare, and personal assistant systems (Rashkin et al., 2019). Recent large language models are capable of producing fluent and emotionally supportive responses in many conversational settings. Nevertheless, it remains uncertain whether these responses reflect genuine sensitivity to contextual-emotional differences or mainly rely on recurring supportive language patterns. Current approaches used to evaluate empathetic dialogue systems largely rely on response-level evaluation methods. Widely used evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020), mainly assess lexical or semantic similarity between generated and reference responses. Although these metrics are useful for estimating overall response quality, they provide limited insight into how models react to subtle contextual or emotional variations. To address this issue, recent studies have explored methods involving commonsense reasoning, emotional grounding, and intermediate reasoning representations for empathy-oriented response generation. Our findings suggest that response-level evaluation alone may not fully capture contextual sensitivity in empathetic dialogue models.

2 Related Work

2.1 Empathetic Dialogue Systems

Empathetic dialogue systems are designed to generate responses that are sensitive to users' emotional states. The release of the EmpatheticDialogues dataset (Rashkin et al., 2019) significantly increased research interest in this area by providing a large-scale text-based benchmark consisting of human-written conversations covering diverse emotional situations. Since then, the dataset has become one of the most widely used resources for empathetic response generation studies.

Early work on empathetic dialogue systems mainly emphasized generating fluent and coherent responses using sequence-to-sequence and transformer-based conversational models. As the field evolved, researchers began to investigate whether contextual reasoning signals could improve empathetic response quality. As a result, methods based on commonsense inference, emotional grounding, and external knowledge integration became increasingly common in em-

pathetic dialogue research. Among these approaches, COMET-based commonsense reasoning frameworks (Bosselut et al., 2019) have received considerable attention. These methods attempt to model possible emotions, intentions, and motivations associated with a speaker's utterance in order to produce more context-aware responses. Other studies have explored emotion recognition, persona modeling, and latent emotion representations to improve empathetic response quality. With the development of large language models, instruction-tuned transformer architectures have also been increasingly adopted for empathetic dialogue generation. Although these systems often produce fluent and supportive responses, relatively few studies have examined how contextual-emotional distinctions are processed within the internal representations of transformer models.

2.2 Empathy Evaluation Approaches

The evaluation of empathetic dialogue systems is still considered one of the open research problems. Current evaluation methods mostly rely on response-level evaluation approaches. Evaluation in empathetic dialogue systems is commonly performed using automatic similarity metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020). Human evaluation protocols assessing empathy, fluency, relevance, and coherence are also frequently adopted (Rashkin et al., 2019). Despite their widespread use, these evaluation strategies do not always capture whether a model is genuinely responsive to contextual-emotional differences.In particular, generic empathetic responses containing high-level supportive statements can receive high evaluation scores despite failing to adequately reflect contextual differences. As a result, it becomes difficult to determine whether a model is responding to the actual contextual conditions of the dialogue or simply producing broadly supportive responses. To investigate this issue more directly, recent studies have introduced contrastive evaluation, minimal-pair analysis, and behavioral testing frameworks (Gardner et al., 2020; Ribeiro et al., 2020). These approaches focus on how model behavior changes under small but semantically important contextual perturbations. However, studies examining the effects of contextual perturbations on transformer hidden-state representations in empathetic dialogue systems remain limited. While most existing studies focus on behavioral outputs, the internal representation dynamics of transformer models have been less explored.

2.3 Representation Analysis and Probing Studies

Studies on representation analysis and probing, which examine the internal representation behavior of Transformer-based language models, have increased significantly in recent years. These studies have shown that Transformer hidden-state representations encode syntactic, semantic, and contextual information across layers (Hewitt and Manning, 2019; Rogers et al., 2020). In particular, probing methods are widely used to investigate in which layers specific linguistic or semantic information is represented (Belinkov, 2022). Similarly, hidden-state similarity analyses and representational geometry studies provide important tools for understanding how Transformer models distinguish between different inputs. Although representation-level analysis approaches have been widely used in machine translation, natural language inference, and reasoning-based NLP tasks, they have been studied only to a limited extent in the context of empathetic dialogue systems. This study aims to address this gap in the existing literature by jointly examining the effects of contextual perturbations on both behavioral outputs and transformer hidden-state representations in empathetic dialogue systems. In particular, analyzing behavior–representation mismatch scenarios enables a more detailed examination of the true contextual sensitivity capacity of empathetic dialogue models.

3 Methodology

3.1 Benchmark Construction and Minimal-Pair Perturbation

In this study, the EmpatheticDialogues (ED) dataset was used as the primary source dataset (Rashkin et al., 2019). However, the dataset was not used directly as a conventional benchmark for empathetic response generation. Instead, a controlled minimal-pair perturbation benchmark was constructed to analyze contextual sensitivity in empathetic dialogue models.

The main objective of the study is to investigate whether empathetic dialogue systems genuinely adapt to contextual-emotional distinctions or mainly rely on recurring supportive response patterns. For this reason, benchmark construction focused specifically on semantically small yet emotionally critical contextual changes.

During the benchmark creation process, examples with strong contextual polarity were selected manually and semi-automatically. In particular, the following perturbation categories were considered:

- success/failure
- acceptance/rejection
- support/betrayal
- loss/recovery
- worsening/improvement
- moral reversal

These categories correspond to situations in which relatively small lexical modifications produce substantial emotional or social meaning shifts. This setup makes it possible to examine whether the model mainly relies on superficial lexical cues or demonstrates sensitivity to contextual distinctions.

A hard-case benchmark consisting of 100 minimal-pair examples was created for the experiments. The purpose of the benchmark is not large-scale performance evaluation, but rather controlled behavioral and representational analysis under contextual perturbations. Therefore, priority was given to diagnostically informative examples instead of benchmark scale.

Each benchmark instance contains:

- original dialogue context
- perturbed dialogue context
- generated empathetic responses
- perturbation metadata

The perturbation strategy aims to maximize contextual-emotional transformation while preserving most of the lexical structure of the dialogue. Only a small number of emotionally critical expressions were modified in each example.

Original Context	**Perturbed Context**	**Perturbation Type**
I failed my exam.	I passed my exam.	success/failure
Someone made fun of me.	Someone stood up for me.	support/betrayal
I never cheated on my partner.	I cheated on my partner.	moral reversal

Table 1: Example minimal-pair perturbations used in the benchmark.

For example:

> Original: "I'm looking forward to my wedding."
> Perturbed: "I'm dreading my wedding."

Although the lexical structure remains largely preserved, the emotional meaning of the context changes substantially (Table 1).

The following principles were preserved during perturbation construction:

1. lexical preservation
2. semantic contrast
3. emotional polarity shift
4. contextual consistency

This perturbation strategy is inspired by controlled semantic perturbation approaches used in contrastive evaluation and minimal-pair analysis studies (Gardner et al., 2020; Ribeiro et al., 2020).

3.2 Model and Inference Setup

The experiments were conducted using the `theprint/Empathetic-Llama-3.2-3B-Instruct` model, an open-source instruction-tuned transformer model designed for empathetic dialogue generation. The model was selected primarily because its open-source architecture enables direct extraction of hidden-state representations.

Since hidden-state access is generally unavailable in proprietary GPT-style systems, representation-level analyses cannot be performed easily in such models. For this reason, an open-source transformer architecture was preferred in this study.

A total of 100 hard minimal-pair benchmark examples were evaluated. For each example, the following outputs were collected:

- generated responses for the original context
- generated responses for the perturbed context
- layer-wise hidden-state representations

For behavioral similarity analysis, sentence embeddings were obtained using a pretrained `sentence-transformers` model. Hidden-state analyses were performed using residual hidden representations extracted from all transformer layers.

Greedy decoding was used during inference in order to minimize random variability introduced by stochastic sampling strategies.

3.3 Response Sensitivity Metric

Behavioral adaptation was measured by calculating the semantic difference between responses generated for original and perturbed contexts.

Sentence embedding-based semantic similarity methods are widely used in NLP research (Reimers and Gurevych, 2019). In this study, the concept of *response sensitivity* was used as an operational metric representing the degree of behavioral adaptation to contextual perturbations.

Cosine similarity was first computed between the response embeddings of original and perturbed contexts. Response sensitivity was then defined as:

$$ResponseSensitivity = 1 - \cos(r_o, r_p) \quad (1)$$

where:

- r_o: original response embedding
- r_p: perturbed response embedding

Higher response sensitivity values indicate stronger behavioral adaptation to contextual perturbations, whereas lower values suggest that the model continues to produce similar generic supportive responses.

3.4 Hidden-State Divergence Analysis

The second major analysis focused on transformer hidden-state representations. The objective is to examine how contextual perturbations influence the model's internal representation space.

During inference, hidden-state representations were extracted from all transformer layers. For each benchmark example, cosine similarity was calculated between the hidden-state representations of original and perturbed contexts at every layer.

Previous probing and representation analysis studies have shown that transformer hidden states capture syntactic, semantic, and contextual information across layers (Hewitt and Manning, 2019; Rogers et al., 2020; Belinkov, 2022). Cosine similarity-based representation comparison methods are also widely used for examining representational proximity in transformer architectures.

In this study, representational divergence between original and perturbed contexts was operationalized using the following hidden-distance metric:

$$HiddenDistance_l = 1 - \cos(h_o^l, h_p^l) \quad (2)$$

where:

- h_o^l: hidden-state representation of the original context at layer l
- h_p^l: hidden-state representation of the perturbed context at layer l

This analysis makes it possible to track how semantically minor but emotionally important perturbations propagate across transformer layers.

3.5 Behavior–Representation Analysis Framework

To examine the relationship between behavioral adaptation and representational divergence, benchmark examples were evaluated within a two-dimensional analysis framework.

In this framework:

- response sensitivity
- hidden-state divergence

Category	Meaning
True contextual adaptation	Both hidden-state divergence and behavioral adaptation are high
Generic empathy collapse	Both representational and behavioral changes remain limited
Surface-level response shift	Behavioral adaptation exists despite limited hidden divergence
Internal shift without adaptation	Hidden-state divergence is high despite limited behavioral adaptation

Table 2: Behavioral and representational adaptation categories.

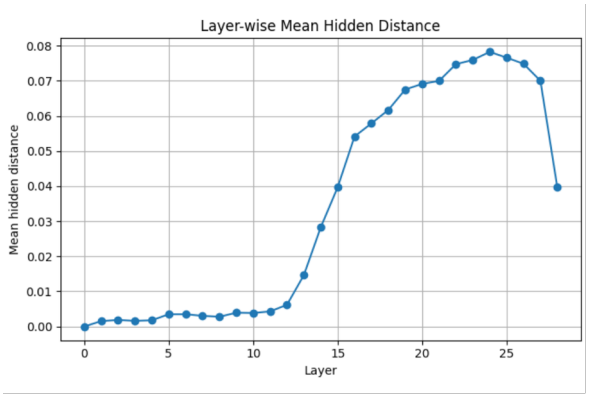

Figure 1: Layer-wise mean hidden-state divergence across transformer layers.

were analyzed jointly in order to identify different adaptation behaviors.

Examples were grouped into four primary categories(Table 2):

This framework enables systematic analysis of whether contextual perturbations influence the model only behaviorally or also at the representational level. In particular, examining mismatch patterns between behavioral adaptation and hidden-state divergence provides a more detailed perspective on contextual sensitivity in empathetic dialogue systems.

4 Results and Discussion

4.1 Layer-wise Sensitivity

Layer-wise hidden-state analyses indicate that contextual perturbations generate representational divergence at different stages of the transformer architecture. In the early layers, similarity values between original and perturbed contexts remained relatively high. More substantial divergence emerged in the middle and later layers.

Average hidden-state divergence values began to increase around layer 14 and reached their highest levels particularly between layers 22 and 24. These findings indicate that contextual perturbations are processed progressively across transformer layers (Figure 1).

Interestingly, hidden-state similarity increased again to some extent in the final layers. This observation may indicate partial representational reconvergence in layers closer to response generation.

This pattern appeared more clearly in perturbation categories involving acceptance/rejection and loss/recovery contrasts.

4.2 Correlation Analysis

To investigate the relationship between behavioral adaptation and representational divergence, correlations between response sensitivity and layer-wise hidden-distance values were analyzed.

The hidden-distance value from layer 24 was selected as the reference layer representing contextual divergence behavior in later transformer stages. The analysis revealed a moderate positive correlation between response sensitivity and hidden-state divergence:

$$r(ResponseSensitivity, HiddenDistance_{24}) \approx 0.40 \quad (3)$$

These findings indicate a partial association between representational divergence and behavioral adaptation. In general, examples with stronger hidden-state divergence also tended to exhibit higher behavioral adaptation.

However, the moderate correlation value suggests that representational divergence and behavioral adaptation are not perfectly aligned. In several cases, the model generated noticeable representational changes while still producing highly similar supportive responses (Figure 2).

The following example illustrates the partial mismatch between representational divergence and behavioral adaptation.

Although the model appears to partially distinguish contextual perturbations at the representational level in this example, behavioral adaptation remains limited. This observation suggests that internal contextual representations and observable behavioral adaptation may not depend entirely on the same mech-

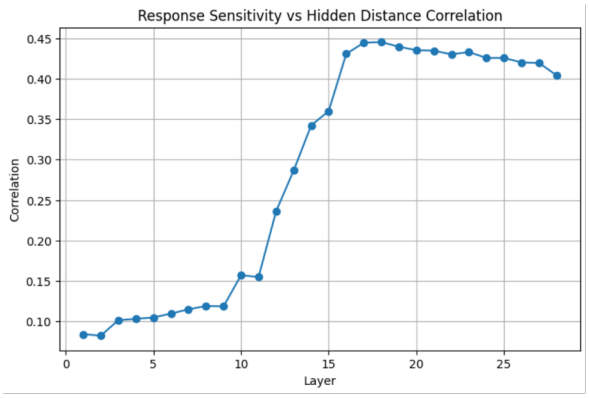

Figure 2: Correlation between response sensitivity and hidden-state divergence across transformer layers.

Metric	Value
Response sensitivity	0.478
Layer-24 hidden distance	0.201

Table 3: Example of moderate behavioral adaptation despite representational divergence.

anisms within the transformer architecture.

4.3 Failure Taxonomy

Benchmark examples were analyzed according to the adaptation categories defined earlier in the methodology section (Table 2).

To examine the relationship between behavioral adaptation and representational divergence more systematically, examples were evaluated within a two-dimensional analysis framework based on:

- response sensitivity
- hidden-state divergence

The resulting distribution reveals several distinct adaptation patterns. Approximately one-third of the benchmark examples exhibited both strong representational divergence and strong behavioral adaptation. Conversely, a comparable number of examples showed only limited changes at both levels (Figure 3).

In particular, examples categorized as *generic empathy collapse* showed minimal representational and behavioral change despite contextual perturbations. By contrast, examples classified as *internal shift without adaptation* demonstrated relatively strong hidden-state divergence while behavioral responses remained broadly similar.

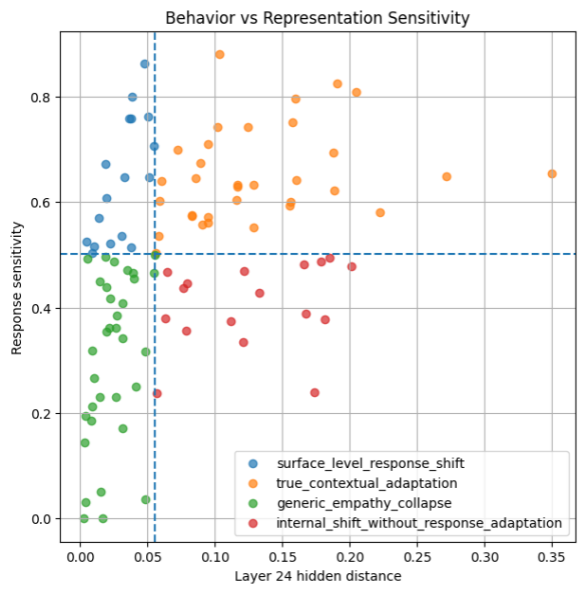

Figure 3: Distribution of benchmark examples across behavioral and representational adaptation patterns.

4.4 Behavioral vs. Representational Mismatch

One of the central findings of the study is the partial mismatch observed between behavioral responses and internal transformer representations.

In several examples, the model appeared to distinguish contextual perturbations internally while continuing to generate generic supportive responses behaviorally. This observation suggests that contextual representation and response generation processes may not be fully aligned within the transformer architecture.

This pattern was particularly noticeable in acceptance/rejection and loss/recovery perturbations, where hidden-state divergence increased substantially in later layers despite limited behavioral adaptation.

The following example illustrates the distinction between behavioral adaptation and representational divergence.

Conversely, some examples exhibited behavioral adaptation despite relatively limited hidden-state divergence. This suggests that certain response changes may result from more superficial response reformulation processes rather than deeper contextual differentiation.

Taken together, these findings indicate that behavioral empathy and internal contextual representations do not exhibit a simple one-

Component	Content
Original context	"Someone was making fun of me."
Perturbed context	"Someone stood up for me."
Generated response	"I'm sorry that happened to you..."
Observation	The model partially distinguishes the contextual change while maintaining a broadly supportive response pattern.

Table 4: Example of behavioral–representational mismatch.

to-one correspondence. A model may internally distinguish contextual differences without consistently translating these distinctions into adaptive empathetic responses.

These observations also highlight an important limitation of response-level empathy evaluation approaches. When only generated responses are evaluated, models may appear more context-sensitive than their internal representational behavior actually suggests.

Overall, the findings demonstrate that representation-level analysis can provide important complementary insights for evaluating contextual sensitivity in empathetic dialogue systems.

5 Conclusion and Limitations

This study presents an analytical framework that simultaneously examines the sensitivity of empathetic dialogue models to contextual perturbations at both the behavioral and representational levels. Using a controlled minimal-pair benchmark created from the EmpatheticDialogues dataset, the study analyzed both the generated responses and the hidden-state representations of an empathy-focused open-source transformer model. The findings reveal that behavioral adaptation and internal representation dynamics in empathetic dialogue models are not always fully aligned. In particular, layer-wise hidden-state analyses indicate that contextual-emotional distinctions are distinctly encoded in the middle and late layers of the transformer architecture. However, only a moderate correlation was observed between behavioral adaptation and representational divergence. One of the most striking findings of the study is that, in some cases, while the model appears to distinguish contextual perturbations at the representational level, it continues to produce generic empathetic responses behaviorally. This suggests that internal contextual encoding and observable behavioral adaptation processes may not rely entirely on the same mechanisms. In addition, the proposed behavior–representation analysis framework has enabled the systematic analysis of different contextual adaptation patterns in empathetic dialogue models. In particular, patterns such as generic empathy collapse and internal representational shifts without strong behavioral adaptation indicate that response-level evaluation approaches may not always accurately reflect the models' true contextual understanding capabilities. These findings suggest that representation-level analysis approaches can provide important complementary insights in the evaluation of empathetic dialogue systems. In particular, examining transformer hidden-state dynamics offers an additional perspective for understanding whether models can truly distinguish contextual perturbations. The study has some limitations. First, the analyses were conducted on a single open-source empathy-tuned model. Whether similar representational behavior patterns persist across different model families or larger instruction-tuned models constitutes an important area of research for future studies. Additionally, the benchmark was conducted on a limited-size hard-case subset created for controlled analysis. Therefore, the study aims to provide diagnostic contextual sensitivity analysis rather than assessing large-scale benchmark performance. However, thanks to the controlled minimal-pair structure, the effects of semantically critical contextual perturbations on transformer representations could be examined more clearly. Another limitation of the study is that the measurement of representational divergence was performed using a cosine similarity-based approach. The use of different representation similarity methods constitutes an important area of research for future studies. Additionally, the analyses were limited to the hidden-state divergence level. More detailed interpretability methods, such as attention dynamics, neuron-level specialization, or probing-based representation analysis, could provide deeper insights in the future.

Overall, the study indicates that empathetic dialogue models should be evaluated not only based on the responses they generate but also through their internal transformer representations. In particular, the findings regarding behavior–representation mismatch suggest that current empathy evaluation paradigms may not fully reflect the models' actual contextual understanding capabilities.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*.

Matt Gardner, Yoav Artzi, Victoria Basmov, and 1 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of EMNLP*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

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 ACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of ACL*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.