Emotional Intelligence Alignment and Conversational Coherence in Dialogue Systems on ConvEval

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the comparative impact of emotional intelligence alignment techniques on conversational coherence metrics in dialogue systems evaluated on the ConvEval benchmark. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluate On-the-job Learning Dialogue Systems and a Case Study for Natural Language Understanding. Research question: What is the comparative impact of emotional intelligence alignment techniques on conversational coherence metrics in dialogue systems evaluated on the ConvEval benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The simulated user follows a specific scenario involving recipe requests and system responses.	×	0.03
The first step of on-the-job learning, detecting system misunderstanding, is trivial with the simple user simulation.	×	0.12
The parameters for the initial training are the same except for the epoch size, which is set to the length of trainLEARN	×	0.00
The user simulation has a decision diagram describing the user's and the system's behavior, found in the appendix.	×	0.03
The F1-score evolution during simulation is shown in Figure 2 for testINITIAL, testLEARN, testUNKNOWN, and testWEIGHTED.	×	0.00
The datasets used for training, simulating, and evaluating the on-the-job learning dialogue system include trainINITIAL,	×	0.12
The data for the concept detection task was generated using patterns and mentions because no initial data was available.	×	0.04
The patterns were written manually by two people from the research domain, and the mentions were scraped from the intern	×	0.01
The sets of patterns and mentions are randomly split into disjoint sets: INITIAL, LEARN, and UNKNOWN.	×	0.00
The simulation dataset is built from INITIAL and LEARN with a mention and patterns distribution following the assumption	×	0.03
The development dataset follows the same assumption as the simulation dataset.	×	0.06
The F1-scores for different models on various test sets are provided in Table (p8).	×	0.03
The modelSIMU achieves the highest F1-scores across all test sets, including testINITIAL, testLEARN, testUNKNOWN, testWE	×	0.01

References

- `http://arxiv.org/abs/1908.11706v1`

- http://arxiv.org/abs/2304.00180v1
- http://arxiv.org/abs/2102.13589v1