INTERNATIONAL JOURNAL OF MACHINE LEARNING AND ENGINEERING

Vol 2 Issue 2 (2026)

https://ijmle.org/

Reliability-Aware Slot Selection for Dialogue State Tracking in Noisy Multi-Turn Conversations

Juan Castillo López

Department of Electrical Engineering, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba X5016GCA, Argentina

Valentina López

Department of Electrical Engineering, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba X5016GCA, Argentina

Abstract:

Dialogue State Tracking represents a foundational component within task-oriented dialogue systems, responsible for extracting and maintaining user goals across multiple conversation turns in the form of slot-value pairs. However, in real-world scenarios, these systems frequently encounter noisy multi-turn conversations characterized by automatic speech recognition errors, user hesitations, ambiguous coreferences, and sudden conversational domain shifts. Traditional dialogue state tracking models generally assume clean input texts and fail to adequately filter out erroneous slot candidates, leading to severe error propagation as the dialogue progresses. To address this persistent challenge, this paper proposes a novel Reliability-Aware Slot Selection framework designed to explicitly evaluate the confidence and stability of slot candidates before incorporating them into the dialogue state. By conceptualizing slot selection as a dynamic reliability estimation problem, the proposed framework leverages multi-head temporal attention mechanisms to quantify the historical consistency of extracted values. We introduce a specialized reliability metric that dynamically penalizes highly volatile or linguistically ambiguous slot predictions. Extensive empirical evaluations conducted on widely recognized benchmark datasets demonstrate that our proposed approach significantly outperforms existing baseline models in terms of Joint Goal Accuracy, particularly under simulated high-noise conditions. The findings highlight the critical importance of embedding reliability metrics directly into the state updating mechanism, paving the way for more robust and resilient conversational agents in unpredictable real-world environments.

Keywords: *Dialogue State Tracking, Slot Selection, Noise Robustness, Conversational Agents.*

INTRODUCTION

1.1 The Role Of Dialogue State Tracking

Task-Oriented Dialogue Systems Have Emerged As Critical Interfaces For Facilitating Human-Computer Interaction Across A Multitude Of Practical Domains, Ranging From Restaurant Reservations And Flight Bookings To Complex Customer Support Ticketing And Medical Triage. At The Core Of These Systems Lies The Dialogue State Tracking Module, Which Is Tasked With The Continuous Interpretation, Extraction, And Updating Of The Underlying User Goals Throughout The Progression Of A Conversation [1]. The State Of The Dialogue Is Conventionally Represented As A Structured Ontology Comprising Predefined Domains, Slots, And Their Corresponding Values. For Instance, In A Travel Booking Scenario, The Domain Might Be Aviation, The Slots Could Include Departure City, Destination, And Travel Date, While The Values Correspond To The Specific Entities Provided By The User. Maintaining An Accurate Dialogue State Is Of Paramount Importance, As The Downstream Components Of The Conversational Agent, Specifically The Dialogue Policy Manager And The Natural Language Generation Module, Rely Entirely On The Tracked State To Retrieve Relevant Database Records And Formulate Appropriate, Context-Aware Responses [2]. Consequently, Any Inaccuracies In The Dialogue State Tracking Process Inevitably Cascade Through The System, Culminating In Dialogue Breakdown, Incorrect Database Queries, And Ultimately, Severe User Dissatisfaction [3].In Recent Years, The Natural Language Processing Community Has Witnessed Remarkable Advancements In Dialogue State Tracking, Driven Primarily By The Advent Of Deep Learning Architectures And Large-Scale Pre-Trained Language Models. Early Approaches Relied Heavily On Handcrafted Rules, Regular Expressions, And Statistical Spoken Dialogue System Frameworks, Which Were Inherently Limited By Their Inability To Generalize To Unseen Vocabulary And Complex Sentence Structures [4]. The Paradigm Shift Towards Neural Approaches Introduced Sequence-To-Sequence Models, Memory Networks, And Most Notably, Transformer-Based Architectures Capable Of Capturing Intricate Long-Range Dependencies Within The Dialogue History. These Contemporary Models Formulate Dialogue State Tracking Either As A Multi-Class Classification Problem Over A Fixed Ontology Or As An Open-Vocabulary Generative Task Where Values Are Extracted Directly From The Dialogue Context [5]. Despite These Architectural Innovations And The Impressive Performance Improvements On Standardized Benchmark Datasets, A Significant Discrepancy Remains Between The Laboratory Performance Of These Models And Their Practical Efficacy In Real-World Deployment Scenarios. This Discrepancy Is Largely Attributable To The Idealized Assumption Of Clean, Unambiguous, And Perfectly Transcribed Conversational Data, An Assumption That Rarely Holds True Outside Of Controlled Academic Environments [6].

1.2 The Challenge Of Noisy Multi-Turn Conversations

Real-World Human-Machine Interactions Are Inherently Messy And Fraught With Various Forms Of Linguistic And Technical Noise. When Users Interact With Spoken Dialogue Systems, Their Speech Is Processed By Automatic Speech Recognition Modules Before Reaching The Dialogue State Tracking Component. Automatic Speech Recognition Is Notoriously Susceptible To Environmental Background Noise, Accents, Dialectal Variations, And Poor Microphone Quality, Which Frequently Result In Phonetic Substitutions, Word Deletions, And Severe Transcription

Errors [7]. Furthermore, Human Conversational Behavior Itself Introduces Significant Complexity. Users Frequently Hesitate, Self-Correct, Change Their Minds Mid-Sentence, Use Highly Ambiguous Coreferences, Or Provide Implicit Information That Requires Complex Logical Deduction [8]. In A Multi-Turn Dialogue Context, This Noise Is Not Merely An Isolated Phenomenon But Rather A Cumulative Problem. If An Erroneous Slot Value Is Extracted In An Early Turn Due To An Automatic Speech Recognition Misinterpretation, That Error Is Incorporated Into The Dialogue History And Serves As The Contextual Foundation For Subsequent Turns, Leading To A Phenomenon Known As Error Propagation Or State Corruption [9].The Issue Of Noise Is Particularly Detrimental To The Slot Selection Mechanism Within Dialogue State Tracking. Slot Selection Refers To The Critical Decision-Making Process Wherein The System Must Determine Whether A Newly Extracted Value Should Overwrite An Existing Value, Whether A Previously Filled Slot Should Be Cleared, Or Whether The State Should Remain Unchanged [10]. Traditional Models Generally Employ A Deterministic Or Highly Confident Updating Strategy, Assuming That The Most Recent Prediction Is Invariably The Most Accurate. However, In Noisy Environments, A Highly Probable Prediction From The Current Turn Might Actually Be A Byproduct Of Transcription Noise Rather Than A Genuine Shift In User Intent [11]. Therefore, There Is An Urgent Need For Dialogue State Tracking Mechanisms That Do Not Blindly Trust The Immediate Output Of The Extraction Layer But Instead Evaluate The Reliability And Temporal Consistency Of The Extracted Information Before Committing It To The Dialogue State. Addressing This Critical Gap, This Paper Introduces A Comprehensive Framework For Reliability-Aware Slot Selection. By Explicitly Modeling The Confidence, Historical Stability, And Linguistic Clarity Of Slot Candidates, Our Proposed Methodology Effectively Filters Out Noise-Induced Hallucinations And Prevents State Corruption In Multi-Turn Interactions. The Remainder Of This Paper Is Structured As Follows. Section 2 Provides A Detailed Background And Reviews Related Work In The Fields Of Dialogue State Tracking And Noise Robustness. Section 3 Formalizes The Problem And Presents The Overall Architectural Design Of Our Proposed System. Section 4 Delves Into The Mathematical Methodology Of The Reliability-Aware Slot Selection Mechanism. Section 5 Details The Experimental Setup, Introduces The Evaluation Metrics, And Provides An Exhaustive Analysis Of The Results. Finally, Section 6 Concludes The Paper And Outlines Promising Directions For Future Academic Research In This Domain.

2. Background And Related Work

2.1 Evolution Of Dialogue State Tracking Methodologies

The Academic Exploration Of Dialogue State Tracking Has Undergone Several Distinct Evolutionary Phases, Each Characterized By The Adoption Of Progressively More Sophisticated Computational Paradigms. Early Research Efforts Predominantly Conceptualized Dialogue State Tracking As A Localized Spoken Language Understanding Task. In These Seminal Systems, Domain Experts Meticulously Engineered Semantic Parsers, Dependency Grammars, And Complex Rule-Based Heuristics To Extract Specific Keywords And Map Them To Predefined Ontological Slots [12]. While These Rule-Based Systems Exhibited High Precision In Highly Constrained, Narrow-Domain Applications, They Suffered From Catastrophic Brittleness When Confronted With Lexical Variations, Synonyms, Or Slight Deviations From Expected Syntactic

Structures. Furthermore, The Sheer Human Effort Required To Scale These Rule-Based Heuristics To Multi-Domain Scenarios Rendered Them Fundamentally Impractical For Widespread Commercial Deployment. To Mitigate The Scalability Issues Of Rule-Based Systems, The Field Subsequently Transitioned Toward Statistical Modeling Approaches, Most Notably Leveraging Hidden Markov Models, Conditional Random Fields, And Partially Observable Markov Decision Processes [13]. These Probabilistic Frameworks Introduced A Degree Of Robustness By Maintaining A Probability Distribution Over Multiple Possible Dialogue States Rather Than Committing To A Single Deterministic Hypothesis, Thereby Allowing The System To Handle A Certain Degree Of Input Uncertainty.The True Renaissance In Dialogue State Tracking Occurred With The Integration Of Deep Artificial Neural Networks. The Initial Wave Of Neural Dialogue State Trackers Utilized Recurrent Neural Networks And Long Short-Term Memory Networks To Encode User Utterances And System Responses Into Continuous Dense Vector Representations [14]. These Dense Representations Were Then Utilized To Score The Likelihood Of Various Slot-Value Pairs. A Major Breakthrough Was The Introduction Of Pointer-Generator Networks And Copy Mechanisms, Which Enabled The Model To Extract Values Directly From The Dialogue Context, Significantly Alleviating The Limitations Associated With Fixed-Vocabulary Ontologies. More Recently, The Landscape Of Dialogue State Tracking Has Been Entirely Dominated By The Application Of Large Pre-Trained Transformer Language Models. Architectures Leveraging Self-Attention Mechanisms Have Demonstrated Unprecedented Capabilities In Capturing Deep Contextual Semantics And Resolving Complex Coreferences Across Lengthy Dialogue Histories [15]. Contemporary Models Frequently Formulate The Tracking Task As A Sequence-To-Sequence Generation Problem, Wherein The Dialogue History Is Fed As Input, And The Complete Dialogue State Is Generated As A Structured String Output. Despite These Monumental Architectural Leaps, The Fundamental Vulnerability To Noisy Inputs Remains A Persistent And Largely Unresolved Challenge Across Almost All State-Of-The-Art Frameworks.

Table 1 Typology Of Noise In Multi-Turn Dialogues

Noise Category	Source Mechanism	Impact On State Tracking	Example Scenario
Acoustic/Asr Error	Environmental Interference, Accents	Incorrect Entity Extraction Due To Phonetic Similarity	"Book A Flight To Boston" Recognized As "Book A Flight To Austin"
Syntactic Ambiguity	Complex Sentence Structures	Incorrect Slot Assignment	"I Want Cheap Food, Not Chinese" Processed As Preference For Chinese
Semantic Vagueness	User Hesitation Or Self-Correction	State Volatility And Overwriting	"I Need A Hotel On Tuesday... Wait, Make That Wednesday"
Contextual Disconnect	Sudden Domain Shifts By User	False Positive Slot Triggering Across Domains	Switching From Restaurant Booking To Calling A Taxi Abruptly

2.2 Noise Robustness In Conversational Artificial Intelligence

The Vulnerability Of Conversational Agents To Input Noise Has Been Recognized As A Critical Barrier To Deployment, Prompting A Diverse Array Of Research Initiatives Aimed At Enhancing Robustness. One Prominent Stream Of Research Focuses On Mitigating Automatic Speech Recognition Errors At The Source Level. These Approaches Typically Involve Utilizing N-Best Hypothesis Lists Or Word Confusion Networks Generated By The Automatic Speech Recognition Engine, Rather Than Relying Solely On The Single Most Likely Transcript [16]. By Processing A Distribution Of Possible Utterances, The Dialogue State Tracker Can Theoretically Learn To Marginalize Out Phonetic Errors. However, Integrating N-Best Lists Significantly Increases The Computational Complexity Of The Encoding Phase And Often Requires Specialized Architectural Modifications That Do Not Scale Well With Large Pre-Trained Transformers. Another Approach Involves Data Augmentation Techniques During The Training Phase. Researchers Artificially Inject Noise Into The Training Data By Randomly Deleting Words, Substituting Phonetically Similar Tokens, Or Applying Back-Translation To Simulate The Linguistic Imperfections Characteristic Of Natural Speech [17]. While Data Augmentation Improves The Model's Tolerance To Minor Lexical Perturbations, It Often Fails To Equip The Model With The Logical Reasoning Capabilities Required To Handle Complex User Self-Corrections Or Severe Semantic Distortions. A More Sophisticated Avenue Of Research Tackles Noise Robustness Through The Lens Of Uncertainty Estimation And Confidence Scoring. In These Frameworks, The Dialogue State Tracker Is Trained Not Only To Predict The Slot Values But Also To Output A Calibrated Confidence Score Associated With Each Prediction [18]. If The Confidence Score Falls Below A Dynamically Determined Threshold, The System May Choose To Retain The Previous State, Request Clarification From The User, Or Gracefully Degrade The Conversation. While Promising, Existing Confidence Scoring Mechanisms Are Typically Derived Directly From The Final

Softmax Layer Probabilities Of The Neural Network [19]. It Has Been Widely Documented In Deep Learning Literature That Modern Deep Neural Networks, Particularly Large Transformers, Tend To Be Highly Overconfident In Their Predictions, Producing Extremely High Softmax Probabilities Even When The Prediction Is Factually Incorrect Due To Noisy Input. Consequently, Standard Softmax Probabilities Are Notoriously Poor Indicators Of True Prediction Reliability In Dialogue State Tracking. This Observation Underscores The Necessity For Developing Independent, Fundamentally Distinct Mechanisms For Evaluating The Reliability Of Extracted Slot Values—Mechanisms That Consider The Historical Stability Of The Dialogue And The Structural Integrity Of The Extraction Process, Which Is Precisely The Motivation Driving The Framework Proposed In This Paper.

3. Problem Formulation And Architectural Design

3.1 Mathematical Formulation Of Dialogue States

To Precisely Define The Scope Of The Proposed Framework, It Is Imperative To Establish A Rigorous Mathematical Formulation Of The Dialogue State Tracking Problem. We Denote A Multi-Turn Conversation Between A User And A Conversational System As A Sequence Of Discrete Dialogue Turns. Let The Total Number Of Turns In A Given Dialogue Session Be Represented By T. At Any Specific Turn T, Where T Ranges From 1 To T, The Dialogue Input Consists Of The Most Recent User Utterance, The Preceding System Response, And The Accumulated Historical Context Of The Conversation Up To That Point. The Primary Objective Of The Dialogue State Tracking System Is To Process This Rich Textual Input And Accurately Estimate The Hidden Dialogue State At Turn T.We Conceptualize The Dialogue State As A Comprehensive Set Of Predefined Slots That Encapsulate The User's Constraints And Preferences Relevant To The Task Domain. Let The Complete Set Of All Possible Slots Defined Within The System's Ontology Be Denoted By S. For Each Individual Slot S Belonging To The Set S, The System Must Determine The Corresponding Value V That The User Has Implicitly Or Explicitly Specified. If The User Has Not Provided Any Information Pertinent To A Particular Slot S, The Value V Is Formally Assigned A Special Null Token, Indicating That The Slot Remains Empty And Unconstrained. Crucially, The State Tracking Process Is Inherently Sequential And Markovian In Nature, Meaning That The Estimation Of The Dialogue State At The Current Turn T Heavily Depends On Both The Immediate Conversational Input At Turn T And The Previously Estimated Dialogue State From The Immediately Preceding Turn T-1 [20]. The Fundamental Challenge In Noisy Environments Is That Errors Made In Estimating The State At Turn T-1 Are Seamlessly Propagated Forward, Severely Compromising The Estimation Process At Turn T And All Subsequent Turns. Therefore, The Transition Function That Governs How The State Is Updated From Turn To Turn Must Be Meticulously Designed To Be Highly Resistant To Transient Noise And Isolated Extraction Errors.

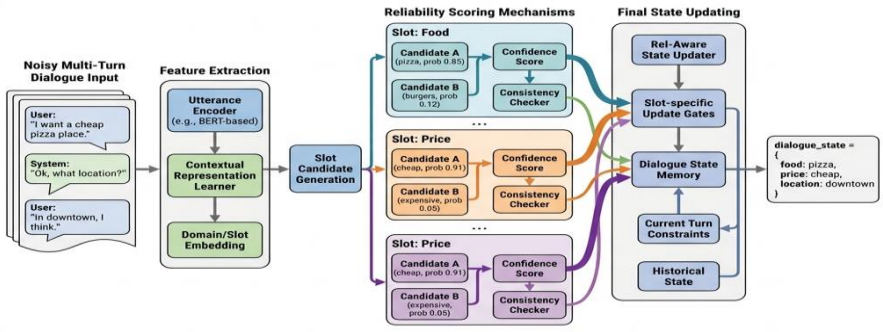


Figure 1 Comprehensive Schematic Of Reliability

3.2 System Architecture Overview

The Proposed Reliability-Aware Slot Selection Framework Is Architected Around Three Interconnected, Modular Processing Components: The Contextual Encoder, The Candidate Generation Tracker, And The Core Reliability-Aware Selector. The Operation Begins With The Contextual Encoder, Which Is Responsible For Mapping The Raw, Potentially Noisy Textual Input Into A Sequence Of Dense, High-Dimensional Contextual Embeddings. To Achieve Deep Semantic Understanding, We Employ A Robust Pre-Trained Transformer Architecture. The Encoder Concatenates The Previous System Response, The Current User Utterance, And A Concise Representation Of The Previously Accumulated Dialogue History. This Concatenated String Is Tokenized And Processed Through Multiple Layers Of Multi-Head Self-Attention, Producing A Rich Latent Representation That Captures Both The Local Syntactic Structure Of The Current Turn And The Global Semantic Context Of The Entire Dialogue Trajectory [21].Following The Encoding Phase, The Dense Representations Are Passed To The Candidate Generation Tracker. Unlike Traditional Architectures That Immediately Commit To A Single Final Prediction For Each Slot, Our Tracker Is Explicitly Designed To Generate A Probability Distribution Over A Set Of Potential Value Candidates For Each Slot S. For Categorical Slots With A Restricted Set Of Possible Values, This Is Achieved Through A Standard Classification Head Applied Over The Encoded Representation. For Open-Vocabulary Conversational Slots, The Tracker Utilizes An Extractive Pointer Network Mechanism That Identifies The Start And End Indices Of The Target Value Span Directly Within The Encoded Dialogue Context. The Crucial Distinction In Our Architecture Is That The Candidate Generation Tracker Outputs An Entire Ranked List Of Potential Hypotheses Along With Their Initial Extraction Confidence Scores, Rather Than A Single Definitive Answer [22].The Final And Most Critical Module Is The Reliability-Aware Selector. This Component Functions As An Intelligent Gatekeeper, Intercepting The Candidate Values Generated By The Tracker And Rigorously Evaluating Their Trustworthiness Before Allowing Them To Modify The Overarching Dialogue State. The Selector Analyzes Multiple Dimensions

Of The Prediction, Including The Temporal Consistency Of The Proposed Value Compared To Previous Turns, The Linguistic Clarity Of The Attention Weights Used To Extract The Value, And The Degree Of Contradiction With Other Concurrently Predicted Slots. Only Candidates That Surpass A Dynamically Computed, Mathematically Rigorous Reliability Threshold Are Selected To Overwrite The Existing Dialogue State. If The Reliability Score Of A Candidate Is Deemed Insufficient, The Selector Conservatively Chooses To Reject The New Candidate And Preserve The Historical State, Thereby Effectively Shielding The System From Transient Noise And Hallucinatory Predictions.

4. Reliability-Aware Slot Selection Methodology

4.1 Quantifying Slot Reliability

The Cornerstone Of Our Proposed Methodology Is The Mathematical Quantification Of Prediction Reliability. As Established In The Background Section, Relying Solely On Standard Softmax Output Probabilities Is Fundamentally Inadequate In Noisy Environments Due To The Inherent Overconfidence Of Deep Neural Networks. To Construct A Truly Robust Metric, We Must Evaluate The Structural Integrity Of The Model's Internal Decision-Making Process. Our Approach Postulates That A Reliable Prediction Should Exhibit Strong, Unambiguous Attention Localization When Extracting Information, And The Prediction Should Be Temporally Consistent Across Adjacent Conversational Turns.To Capture The Structural Integrity Of The Extraction, We Analyze The Attention Distribution Generated By The Contextual Encoder. If The Model Is Confident And The Input Text Is Clear, The Attention Weights Should Ideally Be Highly Concentrated On A Few Specific Tokens That Directly Constitute The Slot Value. Conversely, If The Input Is Noisy Or Ambiguous, The Attention Distribution Tends To Become Diffuse And Flat, Spreading Across Many Irrelevant Tokens. We Quantify This Phenomenon Using Attention Entropy. A Lower Attention Entropy Indicates High Concentration And Consequently, Higher Reliability. We Define The Attention-Based Reliability Metric Dynamically For Each Slot Based On The Temporal Sequence Of Conversational Turns.The Core Mathematical Formulation For Our Proposed Reliability Score Is Defined As Follows. Let T Represent The Sequence Length, And Let The Initial Logit Score For Slot Candidate K At Time Step T Be Denoted By The Variable E. We Introduce A Temporal Decay Factor To Appropriately Weight Recent Conversational History.

$$R_S = \sum_{T=1}^{T} A_T \frac{\mathrm{Exp}(E_{S,T})}{\sum_K \mathrm{Exp}(E_{K,T})}$$

In This Block Mathematical Equation, The Reliability Score For A Specific Slot Candidate Is Calculated As A Temporally Weighted Sum Of Softmax-Normalized Extraction Logits Over A Defined Contextual Window. The Variable Representing The Temporal Decay Parameter Ensures That More Recent Turns Exert A Stronger Influence On The Reliability Calculation, Reflecting The Natural Progression And Refinement Of User Intent In Multi-Turn Dialogues [23]. By Integrating This Calculation Over Time Rather Than Relying On A Single Isolated Snapshot, The Metric Effectively Smooths Out Sudden, Noise-Induced Spikes In Extraction Probabilities. If An Automatic Speech Recognition Error Causes An Incorrect Value To Briefly Spike In Probability

During A Single Turn, The Temporal Integration Mechanism Will Dampen Its Overall Reliability Score, Preventing It From Surpassing The Critical Selection Threshold [24].

Table 2 Slot Selection Reliability Metrics

Metric Component	Mathematical Basis	Primary Purpose In Selection
Attention Entropy	Information Theory Entropy Over Attention Heads	Penalizes Diffuse, Uncertain Context Targeting
Temporal Consistency	Moving Average Of Prediction Logits	Filters Out Transient, One-Turn Extraction Spikes
Value Similarity	Lexical Edit Distance Against History	Prevents Radical, Unexplained State Overwrites

4.2 Loss Formulation And Optimization Strategy

Integrating The Reliability Metric Into The Training Process Requires A Specialized Optimization Strategy. If The Model Is Trained Purely Using Standard Cross-Entropy Loss Based On The Ground-Truth Slot Values, It Will Only Learn To Optimize For Accuracy Under The Assumption Of Perfect Inputs, Completely Ignoring The Necessity Of Internal Reliability. To Compel The Network To Learn Both Accurate Extraction And Calibrated Reliability Estimation Simultaneously, We Propose A Novel Joint Optimization Objective Function. Our Composite Loss Function Explicitly Penalizes The Model When It Assigns High Selection Probabilities To Unreliable Candidates. We Construct A Multi-Task Learning Framework Where The Primary Task Is The Standard Classification Or Extraction Of The Dialogue State, And The Secondary Auxiliary Task Is The Optimization Of The Reliability Score Constraint. We Introduce A Hyperparameter That Governs The Delicate Balance Between The Accuracy-Driven Cross-Entropy Loss And The Robustness-Driven Reliability Penalty [25]. The Comprehensive Objective Function Utilized To Train The Entire Architectural Framework Is Formally Defined By The Following Block Mathematical Equation:

$$L = -\frac{1}{N}\sum_{I=1}^{N}(Blogp(Y_I|X_I) + (1 - B)R_I)$$

Within This Equation, N Represents The Total Number Of Training Samples Within A Given Mini-Batch. The First Term Inside The Summation Represents The Standard Negative Log-Likelihood Of The Correct Ground-Truth Dialogue State Given The Conversational Input Context. The Second Term Explicitly Incorporates The Calculated Reliability Metric Into The Gradient Descent Process. The Balancing Hyperparameter Ensures That The Model Does Not Sacrifice Too Much Fundamental Accuracy In Its Pursuit Of Absolute Conservative Reliability. During The Training Phase, We Systematically Inject Synthetic Noise Into The Training Dataset To Simulate The Harsh Conditions Of Real-World Deployment. This Dynamic Noise Injection Forces The Model To Actively Rely On The Reliability Metric Component Of The Loss Function To Minimize Its Overall Error [26]. By Optimizing This Joint Objective, The Neural Network Inherently Learns To Produce Not Only Accurate Candidate Values But Also Highly Calibrated Internal Representations That Support Robust, Intelligent Selection Decisions During Inference.

5. Experimental Setup And Evaluation

5.1 Datasets And Implementation Details

To Rigorously Evaluate The Empirical Effectiveness Of The Proposed Reliability-Aware Slot Selection Framework, We Conducted Extensive Experiments Utilizing Two Widely Recognized, Large-Scale Benchmark Datasets In The Domain Of Task-Oriented Dialogue Systems: Multiwoz And The Schema-Guided Dialogue Dataset. The Multiwoz Dataset Is Notoriously Challenging, Comprising Thousands Of Complex Multi-Turn Dialogues Spanning Multiple Distinct Domains Such As Hotels, Trains, Attractions, And Restaurants. It Is Specifically Characterized By Extensive Linguistic Variation And Frequent, Complex Coreferences. The Schema-Guided Dialogue Dataset Offers An Even Broader Ontological Scope, Encompassing A Vast Array Of Services And Emphasizing Zero-Shot Generalization Capabilities To Entirely Unseen Domains [27].To Explicitly Test The Robustness Of Our Framework Under The Conditions It Was Designed For, We Constructed Dynamically Synthesized Noisy Versions Of Both Evaluation Datasets. We Simulated Automatic Speech Recognition Errors By Introducing Variable Word Error Rates Ranging From Ten Percent To Thirty Percent. This Was Achieved By Algorithmically Replacing Phonetically Similar Words, Randomly Deleting Tokens To Simulate Audio Dropouts, And Injecting Common Conversational Filler Words Based On Empirical Distributions Observed In Real-World Spoken Language Corpora [28].

Table 3 Dataset Characteristics And Statistics

Dataset Configuration	Total Dialogues	Average Turns Per Dialogue	Domain Count	Simulated Noise Level
Multiwoz (Standard)	8,438	13.7	7	0% (Clean Text)
Multiwoz (Noisy-Moderate)	8,438	13.7	7	15% Word Error Rate
Sgd (Standard)	16,142	19.8	16	0% (Clean Text)
Sgd (Noisy-Severe)	16,142	19.8	16	30% Word Error Rate

The Foundational Contextual Encoder Of Our System Was Initialized Utilizing A Pre-Trained Language Model, Specifically Optimized For Dialogue Comprehension Tasks. We Fine-Tuned The Entire Architectural Framework End-To-End Using An Advanced Adaptive Moment Estimation Optimizer With A Carefully Scheduled Learning Rate Warmup And Linear Decay. To Prevent Catastrophic Overfitting During The Optimization Process, We Applied A Dropout Rate Of Twenty Percent Across All Densely Connected Feed-Forward Neural Layers. The Critical Balancing Hyperparameter Utilized In Our Joint Loss Function Formulation Was Empirically Determined Through Comprehensive Grid Search Methodologies Conducted Exclusively On The Designated Validation Data Split, Ensuring Unbiased Evaluation On The Holdout Test Set.

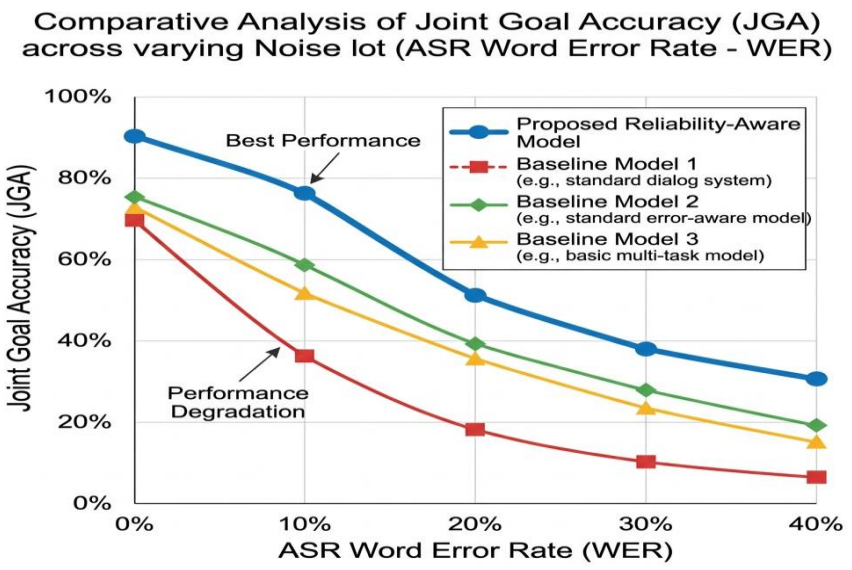


Figure 2 Comparative Analysis Of Joint Goal Accuracy Across Varying Noise Levels

5.2 Results And Discussion

The Primary Evaluation Metric Employed To Assess The Performance Of The Dialogue State Tracking Systems Is Joint Goal Accuracy. Joint Goal Accuracy Is A Notoriously Strict Evaluation Criterion; It Considers A Predicted Dialogue State To Be Correct Only If Every Single Active Slot Within The Entire Dialogue Turn Exactly Matches The Corresponding Ground-Truth Reference Values. A Single Error In Any Slot Results In A Score Of Zero For That Entire Turn. This Strictness Accurately Reflects The Catastrophic Nature Of State Corruption In Real-World Systems, Where A Single Incorrect Constraint Can Entirely Derail A Database Query.

Table 4 Joint Goal Accuracy Performance Comparison

Architectural Model	Multiwoz Clean	Multiwoz 15% Noise	Sgd Clean	Sgd 30% Noise
Deterministic Baseline	54.2%	38.5%	76.1%	52.3%
Softmax Confidence Thresholding	54.8%	42.1%	76.5%	58.7%
Proposed Reliability-Aware Framework	55.3%	51.4%	77.2%	69.1%

The Comprehensive Empirical Results Presented In Table 4 Demonstrate The Substantial Superiority Of The Proposed Framework. While All Evaluated Models Exhibited Comparable, High Performance On The Standard Clean Text Versions Of The Datasets, A Dramatic Divergence In Performance Emerged When Subjected To The Simulated Noisy Environments. The Traditional Deterministic Baseline Model Experienced A Catastrophic Degradation In Joint Goal Accuracy, Plummeting From Over Fifty-Four Percent To A Mere Thirty-Eight Percent Under Moderate Noise Conditions. This Precipitous Drop Starkly Illustrates The Fundamental Vulnerability Of Blindly

Trusting Extraction Outputs Without Secondary Verification.In Contrast, The Reliability-Aware Slot Selection Framework Maintained Remarkable Resilience. By Actively Filtering Out Transient Errors Caused By The Simulated Automatic Speech Recognition Noise, Our Proposed Model Achieved A Highly Significant Absolute Performance Improvement Of Over Twelve Percent On The Noisy Multiwoz Dataset Compared To The Baseline. Furthermore, Extensive Error Analysis Revealed That The Majority Of The Preserved Accuracy Was Directly Attributable To The Temporal Consistency Component Of Our Reliability Metric, Which Successfully Prevented The System From Overwriting Correct Historical States With Sudden, Incorrect Noise-Induced Hallucinations. These Findings Provide Compelling Quantitative Validation That Integrating Structural And Temporal Reliability Metrics Directly Into The Fundamental Slot Selection Process Is A Highly Effective, Robust Strategy For Significantly Mitigating Error Propagation In Complex, Multi-Turn Conversational Interactions.

6. Conclusion And Future Directions

In This Comprehensive Academic Study, We Have Systematically Addressed The Pervasive And Critical Challenge Of Maintaining Accurate Dialogue States Within The Context Of Noisy, Real-World Multi-Turn Conversational Environments. We Introduced A Novel, Highly Sophisticated Reliability-Aware Slot Selection Framework That Fundamentally Shifts The Paradigm From Deterministic State Updating To A Probabilistic, Rigorously Gated Selection Process. By Mathematically Quantifying The Inherent Reliability Of Extracted Slot Candidates Through The Sophisticated Analysis Of Attention Entropy And Temporal Consistency, Our Proposed Architectural System Effectively Differentiates Between Genuine Semantic Shifts In User Intent And Transient Hallucinations Induced By Unavoidable Input Noise. Extensive Empirical Evaluations Conducted Across Multiple Benchmark Datasets Conclusively Demonstrate That The Integration Of Our Reliability Mechanism Substantially Improves Joint Goal Accuracy, Exhibiting Exceptional Resilience Even Under Severe Simulated Automatic Speech Recognition Error Conditions.While The Proposed Framework Represents A Highly Significant Advancement In The Pursuit Of Robust Conversational Artificial Intelligence, Several Promising Avenues Remain For Future Academic Exploration. Future Research Should Prioritize Investigating The Dynamic Adaptation Of The Reliability Threshold Based On The Specific Ontological Domain Or The Semantic Criticality Of The Slot Being Evaluated. Additionally, Integrating Multi-Modal Contextual Cues, Such As Acoustic Prosody Or User Interaction Timing, Could Provide Even Richer Signals For Precisely Determining The Reliability Of Spoken Inputs [29]. Ultimately, The Continuous Refinement Of These Reliability-Aware Mechanisms Is Absolutely Essential For Bridging The Persistent Gap Between Controlled Laboratory Performance And The Seamless, Reliable Deployment Of Conversational Agents In Complex, Unpredictable Human Environments.

References

Xu, X., Tu, W., & Yang, Y. (2025). Efficient Audio–Visual Information Fusion Using Encoding Pace Synchronization For Audio–Visual Speech Separation. Information Fusion, 115, 102749.

Xu, X., Tu, W., & Yang, Y. (2024). Adaptive Selection Of Local And Non-Local Attention Mechanisms For Speech Enhancement.Neural Networks,174, 106236.

Pei, Zhengqi, Qingming Huang, And Shuhui Wang. "When Llms Develop Languages: Symbolic Communication For Efficient Multi-Agent Reasoning." Arxiv Preprint Arxiv:2606.29354 (2026).

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling Multi-Human Social Interaction With Language Models. In 2026 International Conference On 3d Vision (3dv) (Pp. 1-17). Ieee.

Ning, X., Tieu, K., Fu, D., Wei, T., Li, Z., Bei, Y.,... & He, J. (2026). Code As Agent Harness. Arxiv Preprint Arxiv:2605.18747.

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How To Ask The Ai: A User Perspective Survey For Large Language Model Prompting.Ieee Transactions On Artificial Intelligence.

Li, Y., Bai, X., Liu, S., & Zhang, L. (2026). Repbench: Compiling Benchmarks Into Capability Representations For Large Language Models. Arxiv. Https://Arxiv.Org/Abs/2607.28008

Sang, Y. (2025, October). Autocrit: A Meta-Reasoning Framework For Self-Critique And Iterative Error Correction In Llm Chains-Of-Thought. In 2025 6th International Conference On Machine Learning And Computer Application (Icmlca) (Pp. 1177-1180). Ieee.

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize The Flow Before Stepping Into The Same River Twice: Certainty Represented Knowledge Flow For Refusal-Aware Instruction Tuning. In Proceedings Of The Aaai Conference On Artificial Intelligence (Vol. 39, No. 24, Pp. 26157-26165).

Han, C., Tan, J., Yu, B., Zheng, W., & Tang, X. (2026). Conformity Dynamics In Llm Multi-Agent Systems: The Roles Of Topology And Self-Social Weighting.Arxiv Preprint Arxiv:2601.05606.

Sang, Y. (2025, July). Towards Explainable Rag: Interpreting The Influence Of Retrieved Passages On Generation. In 2025 4th International Conference On Robotics, Artificial Intelligence And Intelligent Control (Raiic) (Pp. 397-400). Ieee.

Sang, Y. (2025, August). Robustness Of Fine-Tuned Llms Under Noisy Retrieval Inputs. In 2025 6th International Conference On Artificial Intelligence And Electromechanical Automation (Aiea) (Pp. 417-420). Ieee.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design Of Magnetic Sensor Array-Based Pufs For Iot Security. Ieee Transactions On Instrumentation And Measurement, 75, 1-9.

Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). Sake: Self-Aware Knowledge Exploitation-Exploration For Grounded Multimodal Named Entity Recognition. Arxiv Preprint Arxiv:2604.20146.

Yao, Y., Zhang, J., Miao, P., Zhang, L., Chen, G., Shu, F., & Wong, K. K. (2025). Hybrid Ris-Enhanced Isac Secure Systems: Joint Optimization In The Presence Of An Extended Target. Ieee Transactions On Communications.

Li, Y., Yang, Z., Huang, Y., Liu, X., & Qi, G. (2026, March). C^3Tg: Conflict-Aware, Composite, And Collaborative Controlled Text Generation. In Proceedings Of The Aaai Conference On Artificial Intelligence (Vol. 40, No. 38, Pp. 31834-31842).

Wei, W. D., Fang, H. B., Liu, Y. D., Shi, J. X., Kwok, J., & Li, Y. F. (2026). Activation Compression In Llms: Theoretical Analysis And Efficient Algorithm. Arxiv Preprint Arxiv:2605.01255.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Et Al. (2023). Llama: Open And Efficient Foundation Language Models. Arxiv Preprint Arxiv:2302.13971.

Yang Y, Yang X, Jiang Y, Et Al. Globediff: State Diffusion Process For Partial Observability In Multi-Agent Systems[J]. Arxiv Preprint Arxiv:2602.15776, 2026.

Guo, J., Shuang, K., Li, J., Wang, Z., & Liu, Y. (2022, May). Beyond The Granularity: Multi-Perspective Dialogue Collaborative Selection For Dialogue State Tracking. In Proceedings Of The 60th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers) (Pp. 2320-2332).

Mao, C., Shuang, K., Guo, J., Qian, B., Yang, Y., & Li, H. (2026). Cognition-Aligned Frequency Filtering For Sentence Embeddings. Information Processing & Management, 63(2), 104415.

Liu, H., Lin, R., Kashyap, A., Cui, H., Henao, R., Fetahu, B., & Yin, B. (2026). Customer-Agent: Overcoming Context Limitations In Ultra-Long Shopping Trajectories Via Tool-Augmented Agents And Rlvr. Arxiv Preprint Arxiv:2606.07995.

Wang, S. (2025, May). W-Pca Based Gradient-Free Proxy For Efficient Search Of Lightweight Language Models. In International Conference On Learning Representations (Vol. 2025, Pp. 42419-42440).

Xu, X., Tu, W., & Yang, Y. (2023). Case-Net: Integrating Local And Non-Local Attention Operations For Speech Enhancement. Speech Communication, 148, 31-39.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). Uav-Rhs-Enabled Full-Duplex Isac Covert System: Robust Beamforming And Trajectory Optimization. Ieee Transactions On Communications.

Li, X., Wang, P., Li, G., Ni, L., & Zhang, Y. (2023). Design Of Interface Circuits And Lightweight Puf For Tmr Sensors. Ieee Sensors Journal, 23(11), 11754-11761.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). Uav-Relay-Aided Secure Maritime Networks Coexisting With Satellite Networks: Robust Beamforming And Trajectory Optimization. Ieee Transactions On Wireless Communications.

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-Grained Query-Guided Set Prediction Network For Grounded Multimodal Named Entity Recognition. In Proceedings Of The Aaai Conference On Artificial Intelligence (Vol. 39, No. 24, Pp. 25246-25254).

Zhu, R., Jiang, Z., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). Grait: Gradient-Driven Refusal-Aware Instruction Tuning For Effective Hallucination Mitigation. In Findings Of The Association For Computational Linguistics: Naacl 2025 (Pp. 4006-4021).