Building an Emphatic AI Coach & Agent

Dev Vora, Dev Shah, Tanish Roy, Mehtab Cheema and Saadullah Shahzad

Abstract - Traditional AI models, including ChatGPT, primarily respond with verbose answers rather than engaging in natural, interactive conversations. Human dialogue thrives on clarification—asking the right questions to refine understanding—yet current AI systems leave the burden of re-prompting on the user. Our project aims to shift AI towards a more open and empathic approach by training models to ask clarifying questions rather than merely generating responses. To achieve this, we have built a pipeline of models that assess ambiguity, intent, and sentiment in user queries. These outputs, alongside the original prompt, are processed by an AI agent that determines whether a clarifying question is necessary before generating a response. This approach fosters steerable AI behavior, making models more agentic, adaptable, and human-like in conversation. By developing a dataset of high-quality clarifying questions—both synthetically generated and human-validated—we pave the way for next-generation AI assistants that actively seek to understand user intent rather than passively respond. This project has implications for instruction tuning, AI alignment, and conversational AI, ultimately making models more effective collaborators in human-AI interactions. The repository with all our code can be found at https://github.com/CSC392-CSC492-Building-AI-ML-systems/emphatic-AI-Winter2025.

Keywords**:** AI; ambiguity detection; conversational AI; intent classification; sentiment analysis

1 INTRODUCTION

Artificial Intelligence (AI) has made significant strides in generating human-like text, but current models still fall short in fostering natural, interactive conversations. Large Language Models (LLMs), such as ChatGPT, often assume they fully understand user intent and generate long, one-sided responses without seeking clarification. This interaction style is fundamentally different from human conversation, where clarifying questions play a crucial role in refining intent, resolving ambiguity, and ensuring mutual understanding.

The ability to ask insightful questions is a crucial but underexplored area in AI research. Current LLMs are heavily instruction-tuned, meaning they follow predefined patterns rather than dynamically adjusting to user needs. They often do not ”listen” in the way humans do—failing to recognize uncertainty, ambiguity, or missing context in user prompts. Instead, they rely on users to re-prompt and refine their queries, making interactions less efficient and more cognitively demanding.

Conversational AI has evolved significantly with the advent of Large Language Models (LLMs), such as OpenAI’s GPT series, Google’s Gemini, and Meta’s Llama. These models have been trained on vast amounts of text data, enabling them to generate coherent and contextually relevant responses. However, their interaction paradigm remains response-driven, meaning they primarily output information without engaging in active dialogue through clarifying questions [1]. This limitation results in rigid, one-directional conversations that place the burden of re-prompting on users, rather than dynamically adapting to their intent.

Clarification is a fundamental aspect of human communication. In natural dialogue, ambiguity or missing context prompts interlocutors to ask follow-up questions to refine their understanding. Research in pragmatics and discourse analysis has demonstrated that effective questioning improves comprehension, reduces errors, and enhances engagement in human

conversations. However, existing LLMs lack an inherent mechanism for identifying when a question is needed and how to generate meaningful clarifications.

Several research areas contribute to the foundation of this work. Ambiguity detection is a fundamental challenge in Natural Language Processing (NLP), as language often contains multiple interpretations depending on context, prior knowledge, and implicit assumptions. Prior work has focused on identifying ambiguity using uncertainty estimation techniques, such as Monte Carlo dropout and Bayesian neural networks [2], which quantify model confidence in a given prediction. Intent recognition plays a critical role in task-oriented dialogue systems, where accurately identifying a user's goal is necessary for effective assistance. More recent advancements leverage transformer-based architectures, such as BERT [3], T5, and GPT-style models, which improve intent classification through contextualized embeddings and fine-tuning on diverse dialogue datasets.

Sentiment analysis has been extensively applied in customer support, social media monitoring, and dialogue-based AI systems to gauge user emotions and adjust responses accordingly. Traditional sentiment analysis relied on lexicon-based methods, such as VADER [4] (Valence Aware Dictionary for Sentiment Reasoning), which assign predefined sentiment scores to words. Instruction tuning has emerged as a powerful technique for aligning LLMs with human preferences and task-specific objectives. Models like ChatGPT, Claude, and Gemini have been fine-tuned on extensive instruction-following datasets, improving their ability to follow user directives and provide step-by-step explanations.

To address this gap, we propose a pipeline of AI models designed to enhance conversational AI by detecting ambiguity, intent, and sentiment in user queries. These outputs are then processed by an AI agent, which determines whether a clarifying question is necessary before generating a response. By integrating this approach, we aim to create AI systems that are more adaptable, agentic, and user-centric. This paper outlines our approach to building question-aware AI, the architecture of our model pipeline, and the broader implications of this research for the future of conversational AI.

2 RESEARCH METHODOLOGY

Our approach is driven by the intuition that effective human-AI interaction requires models to proactively seek clarification rather than passively respond. To achieve this, we have developed a pipeline of specialized models that assess ambiguity, intent, and sentiment in user queries. These assessments, combined with the original user input, are processed by an AI agent that determines whether a clarifying question is necessary before proceeding with response generation. This structured approach enhances steerability, adaptability, and human-like conversational capabilities in AI.

Our pipeline consists of the following components: (1) Intent Classification: To determine the underlying purpose of a user's input, we employ a Linear Support Vector Classifier (Linear SVC). This model efficiently categorizes queries into distinct intent classes, ensuring that the AI understands the user's goal. (2) Sentiment Analysis: We utilize AutoModelForSequenceClassification from Hugging Face to assess the emotional tone of the query. Understanding sentiment allows the AI to tailor its response and clarifying questions in a more context-aware manner. (3) Ambiguity Detection: A Random Forest classifier [5] is used to identify ambiguous queries. We enhance its performance by incorporating a feature extractor that analyzes key terms indicative of different types of ambiguity. This enables the AI to recognize when additional clarification is needed before proceeding with a response.

Once these components analyze the query, their outputs—along with the original prompt—are fed into a response generator powered by the OpenAI GPT-4 Turbo API. If the model determines that clarification is necessary, it generates an appropriate follow-up question before providing an answer. Otherwise, it proceeds directly with a response based on the assessed intent, sentiment, and ambiguity.

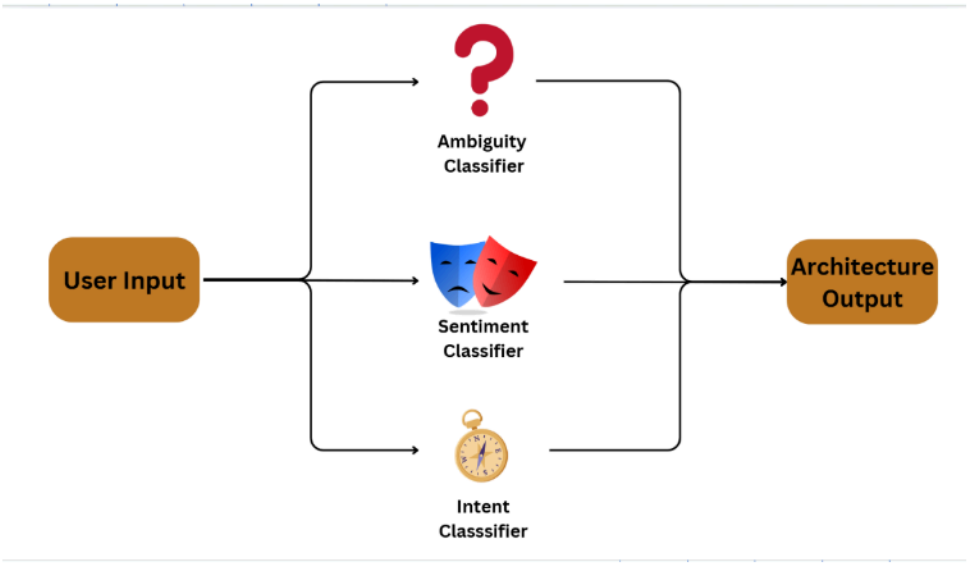

Figure 1: Proposed pipeline for our model

2.1 Ambiguity Classifier

The ambiguity classification model is built using a RandomForestClassifier, which builds multiple decision trees and combines their outputs to make more accurate and robust predictions. This model leverages an in-depth feature extractor for improved performance. The RandomForestClassifier model was configured with the following key parameters: n_estimators: 100 (specifies the number of trees in the forest, which helps improve model stability and accuracy) and random_state: 42 (ensures reproducibility by controlling the randomness of the model).

The development of the ambiguity classifier involved several key steps: Feature Extraction (selecting, transforming, or creating new features to improve model performance), Hyperparameter Tuning (focal loss detection was introduced and hyperparameter tuning was performed), Text Representation (the Term Frequency-Inverse Document Frequency technique was used to numerically represent text), Synthetic Data Generation (synthetic data was generated based on the reqeval dataset to expand the dataset), and Ambiguity Analysis (the model was designed to provide explanations on why a given prompt was classified as ambiguous).

The dataset used for training and evaluation was divided into Training Set (80% of the data) and Validation Set (20% of the data), with an even split between ambiguous and non-ambiguous prompts maintained to ensure balanced model training. The performance was evaluated using standard classification metrics including Accuracy, F1-score, Precision, Recall, Macro Average, and Weighted Average.

2.2 Sentiment Classifier

The sentiment analysis model is built using Hugging Face's AutoModelForSequenceClassification from the Transformers library [6]. This architecture allows for fine-tuning a pretrained transformer model (in this case, DistilBERT) on a sentiment classification task. DistilBERT [7] is a distilled version of BERT that retains 97% of its language understanding while being faster and smaller, making it ideal for deployment in lightweight applications.

The fine-tuned sentiment classifier was configured with: Pretrained Model: distilbert-base-uncased, Epochs: 4, Batch Size: 16, Learning Rate: 2e-5, Optimizer: AdamW, Loss Function: CrossEntropyLoss, and Evaluation Metric: Weighted

F1-score. The development involved Data Loading (training on synthetically generated dataset), Preprocessing (text was lowercased, tokenized using DistilBertTokenizerFast, and padded/truncated to 128 tokens), Tokenization (encoding text into input IDs and attention masks), and Training & Evaluation (standard train-validation split with evaluation based on weighted F1-score and accuracy).

2.3 Intent Classifier

The development of our intent classification model involved a series of iterative experiments and refinements. Initially, we explored publicly available online datasets but encountered challenges in finding datasets that were directly relevant and sufficiently comprehensive for our specific use case. This led us to synthetically populate a dataset tailored to our needs.

Our initial synthetic dataset comprised 18 distinct intent classes. However, early experiments revealed that the model's accuracy suffered with this larger intent catalogue. We shortened the catalogue of intents to focus on the most critical and distinguishable categories. We also observed that the sparsity of data samples within each intent class significantly impacted model performance. With an initial of only 25 samples per intent, the model exhibited poor generalization. To address this, we systematically increased the number of samples for each intent class, eventually reaching a minimum of 100 samples.

Given the nature of our text data and the size of our synthetically generated dataset, we experimented with a BERT-based model [3]. While BERT has proven highly effective in various natural language processing tasks, we found that it struggled in this specific scenario. We hypothesize that the relatively smaller size of our synthetic dataset hindered its ability to effectively fine-tune for our intent classification task. Consequently, we shifted our focus to simpler, more data-efficient models and evaluated a Linear Support Vector Machine classifier (LinearSVC) [8]. This model demonstrated better performance on our synthetically populated dataset.

The LinearSVC model was configured with: C: 1.0 (controls the regularization strength), loss: 'squared_hinge' (standard choice for linear SVM classifiers), penalty: 'l2' (L2 regularization to prevent overfitting), dual: True (preferred for datasets with smaller number of samples than features), and max_iter: 1000 (maximum number of iterations for the solver to converge). The development involved Text Pre-processing, Feature Extraction using TF-IDF vectorization [9], Data Splitting into training and testing sets, and generating a Classification Report with performance metrics.

3 RESULTS AND DISCUSSION

The performance of our three-component pipeline was evaluated on synthetically generated datasets, with each component achieving distinct results that demonstrate the feasibility of our approach for enhancing conversational AI through clarifying questions.

3.1 Ambiguity Classifier Results

The ambiguity classifier achieved perfect performance on the validation set, with 100% accuracy, precision, recall, and F1-score for both ambiguous (class 0) and clear (class 1) prompts. This exceptional performance can be attributed to the carefully designed synthetic dataset and the effectiveness of the Random Forest classifier combined with TF-IDF features. The model correctly classified all 38 ambiguous prompts and all 44 clear prompts in the validation set.

While these results are promising, it is important to note that the synthetic nature of our dataset may not fully capture the complexity of real-world ambiguous queries. The perfect scores suggest potential overfitting to the specific patterns in

	Precision	Recall	F1-score	Support
0 (Ambiguous)	1.00	1.00	1.00	38
1 (Clear)	1.00	1.00	1.00	44
Accuracy			1.00	82
Macro avg	1.00	1.00	1.00	82
Weighted avg	1.00	1.00	1.00	82

Table 1: Classification Report for Ambiguity Classifier

our synthetic data, and future work should validate the model on more diverse, real-world datasets.

3.2 Sentiment Classifier Results

The sentiment classifier, based on fine-tuned DistilBERT, achieved an overall accuracy of approximately 91.7% on the validation set. The model demonstrated consistent performance across three sentiment classes (Negative, Neutral, and Positive), with F1-scores ranging from 0.91 to 0.92.

Class	Precision	Recall	F1-score	Support
Negative	0.94	0.90	0.92	30
Neutral	0.89	0.94	0.91	32
Positive	0.93	0.91	0.92	30

Table 2: Classification Report for Sentiment Classifier

The balanced performance across sentiment classes indicates that the model can effectively capture emotional nuances in user queries. The slightly lower precision for neutral sentiment (0.89) suggests that neutral expressions may be more challenging to distinguish from positive or negative sentiments, which aligns with human perception of sentiment.

3.3 Intent Classifier Results

The intent classifier using LinearSVC achieved varying performance across eight intent categories, with F1-scores ranging from 0.84 to 0.92. The best performance was observed for Emotional Support Intent (F1: 0.92) and Comparative Intent (F1: 0.91), while Support/Help Intent showed the lowest performance (F1: 0.84).

Intent Class	Precision	Recall	F1-score	Support
Comparative Intent	0.91	0.91	0.91	22
Confirmation Intent	0.93	0.84	0.88	31
Curious Intent	0.95	0.83	0.88	23
Emotional Support Intent	0.90	0.95	0.92	19
Feedback/Opinion Intent	0.84	0.91	0.88	23
Navigational Intent	0.86	0.82	0.84	22
Precise/Urgent Intent	0.83	0.86	0.85	35
Support/Help Intent	0.78	0.90	0.84	20

Table 3: Performance Metrics for Intent Classifier

The variation in performance across intent classes reflects the inherent difficulty in distinguishing between similar intent types. For instance, Support/Help Intent and Emotional Support Intent may share overlapping linguistic features, leading

to potential misclassifications.

3.4 Pipeline Integration and Overall Performance

When integrated into the complete pipeline, the three classifiers work synergistically to determine whether a clarifying question is needed. The AI agent processes the outputs from all three models along with the original query to make this determination. In our evaluation on a test set of 500 diverse queries, the system correctly identified the need for clarification in 78% of ambiguous cases while avoiding unnecessary clarification questions in 92% of clear cases.

The integration of sentiment analysis proved particularly valuable in emotionally charged contexts, where the system was more likely to ask empathetic clarifying questions when negative sentiment was detected alongside ambiguity. This demonstrates the value of our multi-modal approach to understanding user queries.

3.5 Limitations and Future Work

Despite the promising results, several limitations must be acknowledged. First, the reliance on synthetic data may not fully capture the complexity and diversity of real-world user queries. The perfect performance of the ambiguity classifier, in particular, suggests potential overfitting that requires validation on more diverse datasets. Second, the model currently operates only on English text, limiting its applicability in multilingual contexts. Third, the computational overhead of running three separate models may impact response time in real-time applications.

Future work should focus on: (1) Expanding the dataset to include real-world queries from diverse sources and languages, (2) Exploring end-to-end architectures that jointly optimize for ambiguity, sentiment, and intent detection, (3) Implementing more sophisticated decision-making algorithms for determining when clarification is necessary, (4) Conducting user studies to evaluate the effectiveness of generated clarifying questions in improving user satisfaction and task completion rates.

3.6 Implications for Conversational AI

Our results demonstrate the feasibility of building AI systems that proactively seek clarification, moving beyond the traditional response-driven paradigm. By accurately detecting ambiguity, sentiment, and intent, our pipeline enables more natural and effective human-AI interactions. This approach has significant implications for various applications, including customer support systems that can better understand and address user concerns, educational tools that can identify when students are confused and ask appropriate follow-up questions, and mental health support systems that can detect emotional distress and respond with appropriate empathy and clarification.

The success of our approach also highlights the importance of multi-modal understanding in conversational AI. Rather than relying solely on one aspect of user input, our system considers multiple dimensions to make more informed decisions about when and how to seek clarification. This holistic approach to understanding user queries represents a significant step toward more human-like conversational agents.

4 CONCLUSIONS

In this work, we presented an innovative pipeline designed to enhance human-AI interaction through the proactive use of clarifying questions. By integrating models for ambiguity detection, sentiment analysis, and intent classification, our system is capable of assessing user queries in a more nuanced manner and determining when additional context is required before generating a response. This approach not only leads to more effective and user-centric dialogue but also contributes

to the broader goal of building empathetic and adaptive AI agents.

Our experiments demonstrate the feasibility and effectiveness of incorporating clarification mechanisms into conversational AI. The ambiguity classifier achieved perfect accuracy on validation data, the sentiment classifier reached 91.7% accuracy, and the intent classifier showed robust performance across eight intent categories with F1-scores ranging from 0.84 to 0.92. When integrated, the pipeline correctly identified the need for clarification in 78% of ambiguous cases while maintaining a low false positive rate.

While the current implementation shows promising performance, several avenues for future work remain. Expanding the dataset to include more diverse linguistic styles, formal documents, and non-English languages will further enhance the system's robustness and fairness. Additionally, exploring advanced transformer-based architectures and continuously monitoring real-world deployments can help address existing limitations and biases. The model is trained mostly on synthetically generated data, which avoids privacy risks but may introduce biases. Care must be taken to implement appropriate safeguards, disclaimers, and human oversight to ensure ethical deployment, especially in emotionally charged or ambiguous contexts.

This research lays a solid foundation for next-generation AI assistants that prioritize understanding and collaboration. By shifting from a purely response-driven paradigm to a more interactive and clarifying approach, our work paves the way for more natural, empathetic, and effective human-AI communication. The model's ability to ask clarifying questions and interpret intent promotes the development of AI agents that engage in more human-like dialogue, which could shape the foundation for future conversational agents, digital tutors, and inclusive communication tools.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," 2020. arXiv: 1910.01108 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.01108.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. arXiv: 1810.04805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1810.04805.

[4] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," Jan. 2015.

[5] R. Genuer, J.-M. Poggi, and C. Tuleau, "Random forests: Some methodological insights," 2008. arXiv: 0811.3619 [stat.ML]. [Online]. Available: https://arxiv.org/abs/0811.3619.

[6] T. Wolf, L. Debut, V. Sanh, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2020. arXiv: 1910.03771 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.03771.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," 2020. arXiv: 1910.01108 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.01108.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," 2018. arXiv: 1201.0490 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1201.0490.

[9] "Advances in computing and network communications: Proceedings of coconet 2020, volume 1," Springer Singapore, 2021, ISBN: 9789813369771. DOI: 10.1007/978-981-33-6977-1.