## Performance Report: Comprehensive Evaluation with Insights and Recommendations

### 1. Test Set Evaluation Summary

The final, best-performing pipeline was evaluated on the unseen 10% test set (100 samples) to ensure performance generalization.

| Metric | Value |
| --- | --- |
| **Overall Test Accuracy** | **0.9000 (90.00%)** |
| **Macro Avg. F1-Score** | 0.8964 |
| **Weighted Avg. F1-Score** | 0.8964 |

The model demonstrates strong performance, successfully classifying 90 out of 100 test utterances correctly.

### 2. Detailed Performance by Intent

| Intent | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| **email\_send** | **1.0000** | **1.0000** | **1.0000** | 20 |
| **calendar\_schedule** | 0.9524 | **1.0000** | 0.9756 | 20 |
| **knowledge\_query** | 0.9000 | 0.9000 | 0.9000 | 20 |
| **general\_chat** | 0.7917 | 0.9500 | 0.8636 | 20 |
| **web\_search** | 0.8667 | **0.6500** | 0.7429 | 20 |

### 3. Key Insights and Analysis

#### **Strengths**

* **High Performance on Core Tasks:** The model achieved perfect F1-scores for **email\_send** and near-perfect scores for **calendar\_schedule**. These intents likely contain highly distinct and specific keywords (e.g., "send email," "schedule meeting," "calendar") that the TF-IDF feature set effectively captures.
* **Reliable Classification:** An overall accuracy of 90% is highly competitive for a classical machine learning model like Logistic Regression combined with TF-IDF features.

#### **Weaknesses**

* **Recall for web\_search:** This intent had the lowest Recall at 0.6500, meaning 35% of actual web\_search requests were misclassified as another intent. This suggests potential semantic overlap between web searches and other intents, possibly knowledge\_query or general\_chat, where the core request (e.g., "What is...") can be ambiguous.
* **Precision for general\_chat:** This intent had the lowest Precision (0.7917), indicating that when the model *predicted* general\_chat, it was wrong about 21% of the time. This category often includes broad, informal language, making it challenging to differentiate from the background noise of other intent-specific phrases.

### 4. Recommendations for Improvement

1. **Augment the web\_search Intent Data:** The poor Recall for web\_search suggests the model hasn't learned the full variety of ways users express this intent. Increase the number and diversity of utterances labeled as web\_search in the training data, especially those that may overlap with knowledge\_query.
2. **Explore Contextual Embeddings:** The use of unigrams only, as determined by the tuning, limits the model's ability to capture meaning from multi-word phrases (bigrams and trigrams were discarded). Switching to deep learning approaches (e.g., BERT, Word2Vec) could capture richer, contextual meaning, which would be highly beneficial for distinguishing between ambiguous intents like general\_chat and web\_search.