QuickAssist

Mid PPT

Team members:
Inbal Bolshinsky 322522673
Shani Kupiec 211521042
Almog Sasson 322758269
Nadav Margaliot 316298405

Repository

Slide 1: Project Description

Project: "QuickAssist: Intent-Aware Chatbot for Customer Support"
Intent Conditioning: *Compare response generation with and without explicit intent labels*

Task:

- **Input: Variant A (No Intent)-** Customer query, **Variant B (With Intent)-** Customer query + intent label
- **Output:** A generated response to the customer query
- **Task Details:**
  - Generate helpful and context-aware responses using an LLM
  - Compare the effect of including an intent label in the input on response quality

Data and Evaluation:

- **Dataset:** Bitext Customer Support Dataset *and* Customer-Service-for-LLM
- **Labels:** Pre-annotated intent labels provided within the dataset
- **Evaluation:**
  - **Automatic:** BERTScore measuring semantic similarity between generated and reference responses
  - **Human-like Judgement:** LLM based evaluation assessing Helpfulness, Fluency, and Appropriateness

Slide 2: Previous work

Source/Title	"Intent-Aware Dialogue Generation and Multi-Task Contrastive Learning for Multi-Turn Intent Classification," 2024	"RSVP: Customer Intent Detection via Agent Response Contrastive and Generative Pre-Training," 2023	"IntentGPT: Few-Shot Intent Discovery with Large Language Models," 2024
Approach/Model	Hidden Markov Models (HMMs) combined with Large Language Models (LLMs), Multi-task Contrastive Learning	Two-stage self-supervised model: retrieval and generative pre-training using agent responses	Few-shot in-context learning leveraging GPT-4; semantic few-shot sampling
Data	E-commerce chat logs with domain-specific intent transitions (details not specified)	Task-oriented dialogue datasets, focusing on agent utterance-response pairs (details not specified)	Minimal labeled intent datasets enhanced by embedding similarity-based sampling
Metrics	Intent classification accuracy; coherence and relevance in multi-turn conversations	Response retrieval accuracy; generation quality and relevance (exact metrics not specified clearly)	Few-shot intent classification accuracy, efficiency in intent discovery (metrics not explicitly detailed)
Results	Effective intent-driven response generation with contextually coherent conversations (precise numerical results not detailed explicitly)	Improved intent detection and response quality through contrastive and generative pre-training (numerical accuracy not explicitly stated)	Effective few-shot intent discovery with minimal labeling; improved classification performance (exact numerical results not explicitly detailed)

Slide 3: Your plan

1. Data Collection & Preparation:

- Two customer support datasets with pre-annotated intent labels are used.
- Each query is transformed into two variants:
  - Variant A: Only the customer query.
  - Variant B: Customer query + intent label as context.
- The datasets are cleaned and split into training and test sets.

2. Input Processing:

- All input texts and expected responses are preprocessed and normalized.
- Inputs are formatted into model-ready instruction–response pairs:
  - For Variant A: "Query: I need to reset my password."
  - For Variant B: "Intent: account help | Query: I need to reset my password."
- Tokenization and padding are applied to ensure consistent input lengths.

3. Model Training

- Two models are trained separately using the same architecture:
  - One trained on Variant A (query only).
  - One trained on Variant B (intent-aware inputs).
- The objective is to generate responses that are helpful, fluent, and relevant to user queries.

Slide 3: Your plan

4. Prediction (Response Generation)

- Both models generate responses for the test set queries.
- Outputs are collected along with original queries, intents, and reference responses.

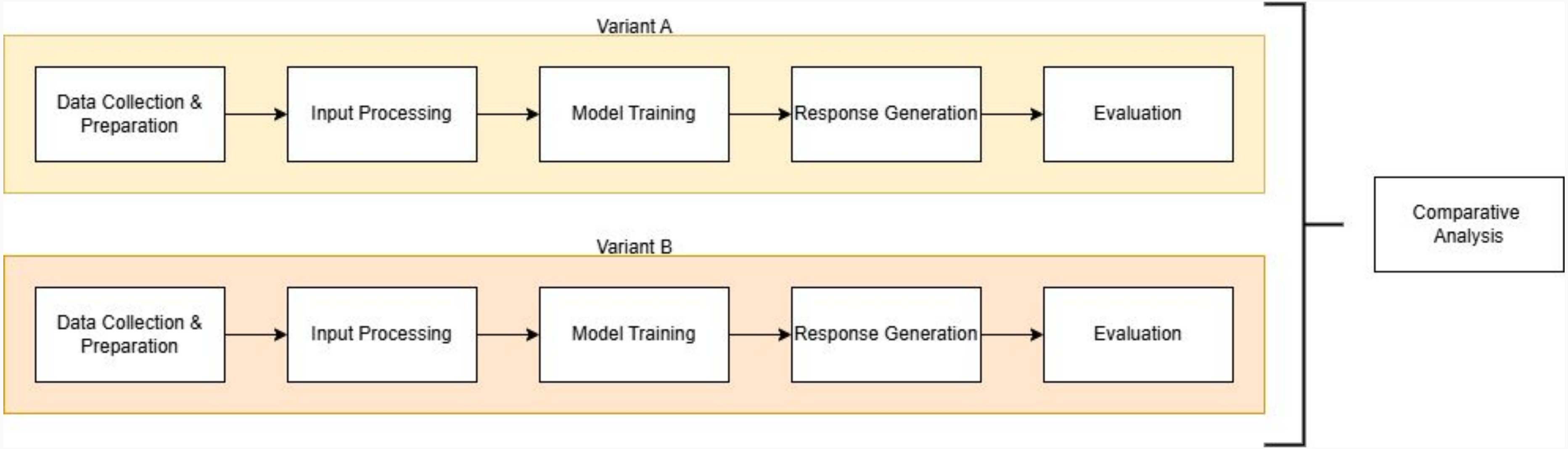
5. Evaluation

- Automatic Evaluation: Responses are compared to references using semantic similarity metrics (e.g., BERTScore).
- LLM-based Evaluation: A large language model evaluates each response on: Helpfulness, Fluency, Appropriateness

6. Comparative Analysis

- Final step compares the performance of Variant A vs. Variant B across all metrics.
- Goal: Determine whether intent conditioning improves the overall quality of model responses.

Slide 3: Your plan



Slide 4: Data exploration and baseline

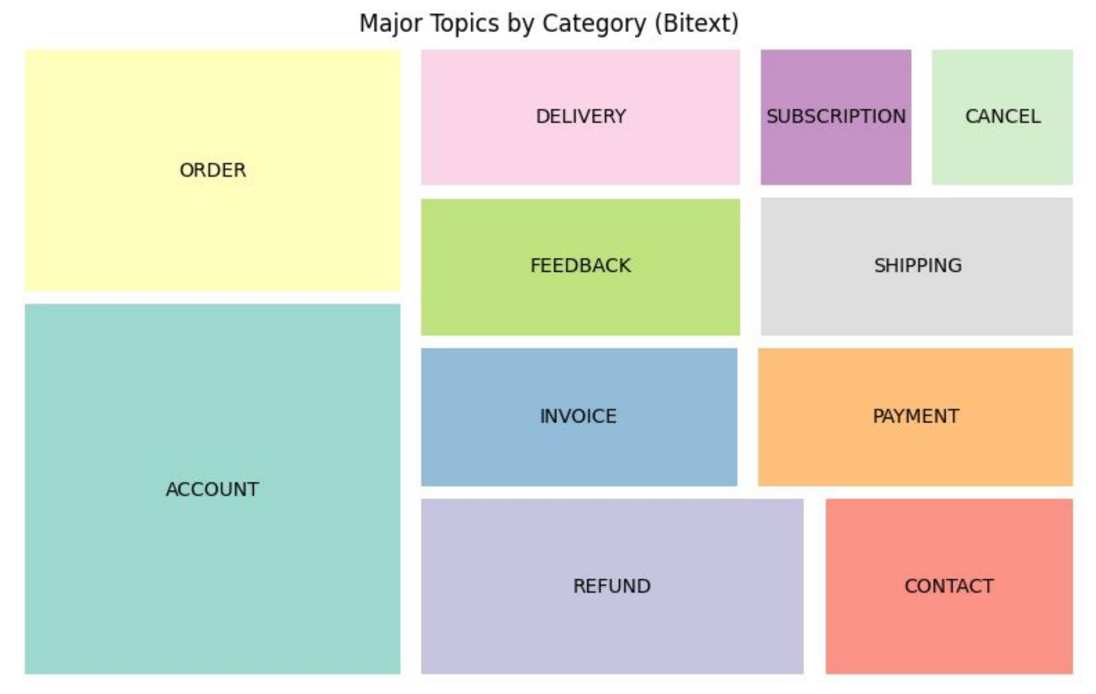
Datasets

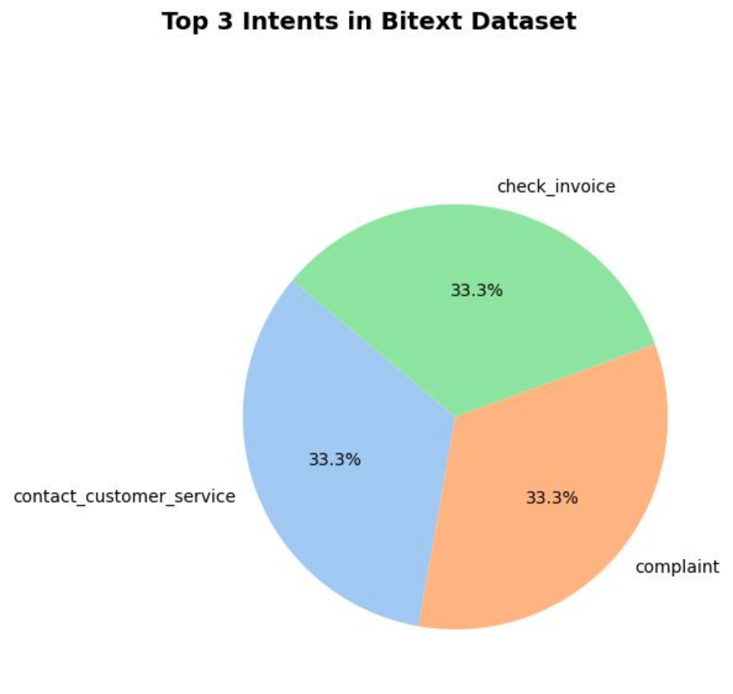
- Two public customer support datasets:
  - **Bitext:** 26,872 question-answer pairs, 27 intents, 11 categories
  - **Customer-Service-for-LLM:** 2,700 question-answer pairs, 27 intents, 11 categories
- All queries are short, average length ≈ 8.69 words [Bitext], 8.6 words [CS-for-LLM]
- Intent labels are pre-annotated, there are no missing values in both datasets.
- **Dataset split:** 80% train, 20% test, stratified by intent
- Most common intents:

➔ `contact_customer_service,complaint, check_invoice` [Bitext]

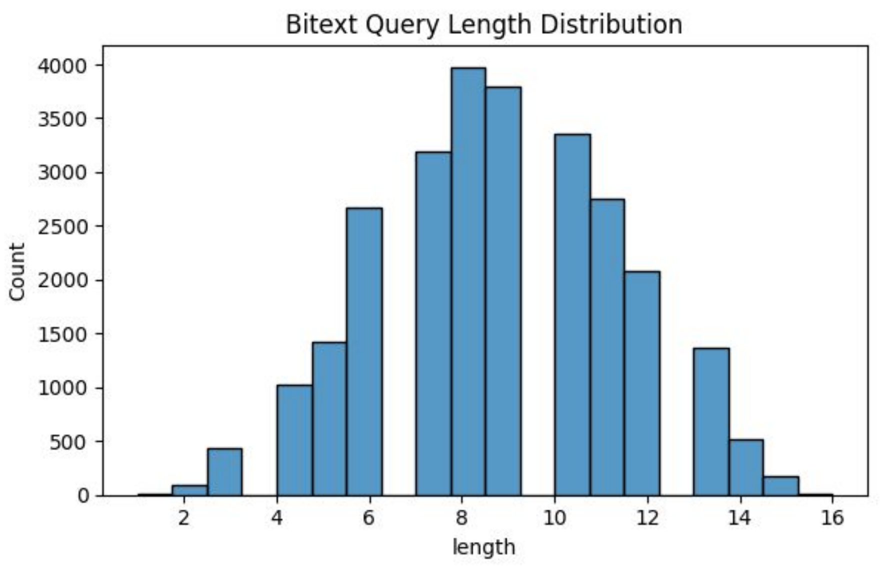
➔ `check_invoice, switch_account, edit_account` [CS-for-LLM]

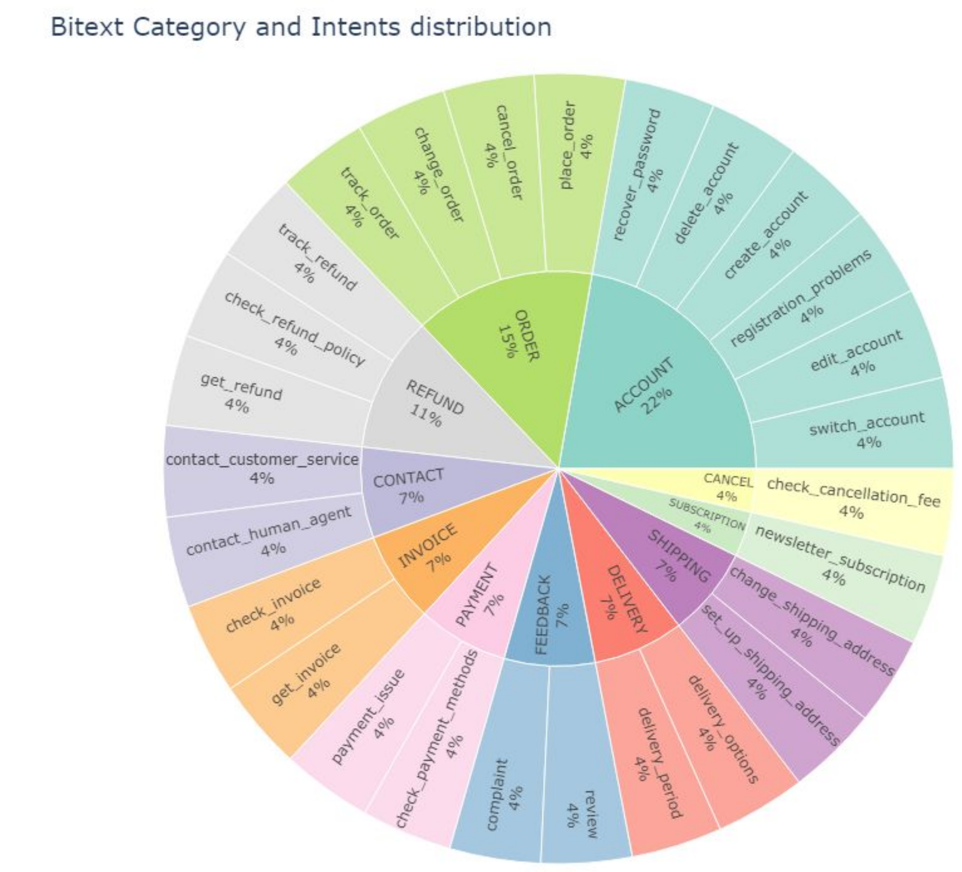
Slide 4: Data exploration and baseline





Slide 4: Data exploration and baseline





Slide 4: Data exploration and baseline

Baseline Details:

- Model: google/flan-t5-small
- Input Format:
  - Variant A (No Intent): Only the customer query is provided as input.
- Dataset: Bitext Customer Support Dataset
  - Subset Used: 20% of the dataset
- Evaluation Set Size: 500 customer queries

Evaluation Metrics:

- Automatic Evaluation:
  - BERTScore (F1): 0.8761 — capturing semantic similarity between generated and reference responses
- LLM-Based Evaluation: Helpfulness: 3.20 / 5, Fluency: 4.02 / 5, Appropriateness: 3.46 / 5.

Slide 5: Conclusions

Insights -> Recommendations:

- **Insight:** The model produces coherent replies, but sometimes lacks task specificity.
  Recommendation: Augment the input with explicit intent labels (e.g., "Intent: Cancel Order | Query: ...") to guide generation toward more targeted responses (include the intent label that is already in the dataset)

- **Insight:** No comparison was made between intent-conditioned and non-intent inputs.
  Recommendation: Train and evaluate a second variant with intent labels added to the prompt to quantify their impact on fluency and helpfulness.(As discussed in slide 2 using the variant A and variant B models).

- **Insight:** Evaluation was limited to a single instruction format.
  Recommendation: Expand analysis to include both variants (with and without intent) under identical evaluation metrics (BERTScore + LLM judgment).

Slide 5: Conclusions

Insights -> Recommendations:

- **Insight:** The dataset includes pre-annotated intent categories, but they may be too granular, redundant, or semantically overlapping.
  Recommendation: Use sentence embeddings of the intent labels themselves (e.g., "cancel membership", "terminate account") to detect semantic similarity between them.
  Then apply clustering (e.g., KMeans, HDBSCAN, or BERTopic) to group similar intents into higher-level clusters — reducing fragmentation and making model predictions more robust.

- **Insight:** Performance evaluation is mostly average across all samples.
  Recommendation: Perform per-intent or per-cluster performance analysis to identify which intents benefit most from conditioning and where the model struggles.