During the previous review cycle, we received constructive and insightful feedback from the reviewers and the area chair. We sincerely appreciate the recognition of our benchmark's practical value, the novelty of our intention modeling pipeline, and the multi-task design of our evaluation framework. To address the feedback and further strengthen the paper, we undertook substantial revisions that enhance the theoretical foundation, clarify methodological design, and improve empirical depth. Below we summarize the key updates:

- **Formalization of Intention Tree and Theoretical Foundation**
  We added a formal theory section (Appendix B) to rigorously define the Intention Tree, which underpins the session intention modeling framework. This includes an inductive definition of the tree structure across discrete time steps and a probabilistic decomposition of intention transitions:

  $$\mathbb{P}(P_{t+1}|\mathcal{H}_t) = \mathbb{P}(\mathcal{M}_t|\mathcal{H}_t) \cdot \mathbb{P}(P_{t+1}|\mathcal{H}_t, \mathcal{M}_t)$$

  We also introduced the decomposition of metadata into intentions, attributes, and comparisons to ground future analyses.
- **Added Intuition and Concrete Examples**
  In Appendix B.2, we now provide an "Intuition" subsection that complements the formal theory by walking through concrete intention branching examples from real user sessions. This helps illustrate how user intent shifts over time in the session.
- **Expanded Task Design Clarifications and Justifications**
  We added a new appendix (Appendix C) that explains the design rationale for each task, including:
  - How numeric scores (0–3) are mapped to choice labels.
  - Why we merged borderline categories like "Maybe Yes/No" to reduce annotator disagreement and label ambiguity.
  - Clarifications on Task 4's choice format and design motivation.
- **Detailed Annotation Analysis and Quality Control**
  We significantly expanded the analysis in Appendix by adding several new tables:
  - **Table 4**: Label distributions by task and score.
  - **Table 5**: Inter-annotator agreement breakdown (e.g., 2:1 vs. 3:0 consensus).
  - **Table 6**: Per-annotator labeling behavior, used to filter annotators with strong biases.

  We also clarified our noise mitigation strategies, such as filtering low-agreement annotations and conflict detection within session metadata.
- **Extended Fine-Tuning Analysis and Negative Results**
  Section 5.3 and Appendix F.2 now include a deeper discussion on fine-tuning. We explain cases where fine-tuning led to performance degradation and suggest potential reasons (e.g., overfitting to noisy signal, misalignment with implicit structure). This offers insight into robustness limitations for session-level modeling.
- **New Evaluation of BERT-Based Baselines**
  Appendix F.3 presents newly added discussions on the performances of RoBERTa-large and DeBERTa-v3-large. These models exhibited severe failure modes (e.g., repetitive "Yes" predictions or random token output), leading to their exclusion from the main

comparison table.

- **Few-Shot Prompting Curation Clarification**
  We added clarifications in Appendix A.4 on how few-shot exemplars were selected, ensuring balanced coverage and representativeness across task formats.
- **Presentation and Terminology Improvements**
  We corrected grammatical errors throughout the main sections and appendices, refined several ambiguous descriptions (especially in Sections 1–3), and standardized the terminology used to describe tasks and metadata components.

We believe these revisions meaningfully improve the manuscript in both clarity and technical rigor. The extended theoretical formulation, deeper error and annotation analysis, and added baselines strengthen our contribution toward understanding inter-session intention modeling in e-commerce settings.