**Problem & Background**
- **Background:** Question Answering Models can be inconsistent.
  - "A bird has feathers,"
  - "A sparrow is a bird,"
  - "A sparrow does not have feathers."
- **BeliefBank:** Existing Framework [1]
  - Save questions and answers to external memory
  - Manually create constraints from ConceptNet
  - Run MAX-SAT solver to filter current and past answers
  - Feed back relevant constraints to QA model [2]
- **Our Goal:** Eliminate manual creation of constraints

**Methods & Experiments**
- **Dataset:** 12k silver facts about animals and plants from ConceptNet [1]. Split into 20% validation, 80% test.
- **Hyper-parameters tuned on validation set.** batch size, sentence similarity threshold, default flipper confidence, flip threshold, entailment threshold, scoring mechanism.
- **Baseline:** Unturbored questions out of QA model.
- **Metrics:**
  - Consistency = \( 1 - \frac{\text{# violated constraints}}{\text{# applicable constraints}} \)
  - F1-score = \( 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \)

**Analyses**

**Results**

**Conclusions**
- Greedy local constraint solving is comparable to global maxSAT constraint solving in this context.
- NLI-based constraints are not sufficient to significantly improve QA accuracy and consistency.
- NLI models don’t have sufficient entailment detection.

---

**Figure 1. Our Approach.**

**Figure 2. Accuracy Scores and Consistency Scores.**

**Figure 3. NLI Constraint Entailment experiment.**

---