Course NLP with LLM Assignment 03

Assignment 3 Report

Omer Tarshish and Lotem Sakira

September 09, 2025

Contents

List of Tables

List of Figures

Part 0: PragmatiCQA Dataset Analysis

The PragmatiCQA paper addresses a fundamental limitation in current QA systems: they often provide only literal answers without considering cooperative dialogue. The authors argue that effective QA systems should exhibit **cooperative behavior** by anticipating follow-up questions and providing enriched responses (Qi et al., 2023).

The paper contributes in four main ways (Qi et al., 2023):

1. **Novel Dataset**: A conversational QA dataset for evaluating pragmatic reasoning.
2. **Annotation Framework**: A systematic approach for distinguishing literal versus pragmatic information spans.
3. **Evaluation Methodology**: Metrics to assess cooperative response capabilities.
4. **Empirical Analysis**: Demonstrates the limitations of current QA models in pragmatic reasoning.

Challenges Provided by the Dataset for NLP Models

In this dataset, the researchers targeted four pragmatic phenomena:

1. **Cooperative Principle**: Models must infer what additional information would be helpful beyond the literal question.
2. **Theory of Mind**: Understanding implicit user intentions and knowledge gaps.
3. **Contextual Reasoning**: Connecting information across multiple sources to provide comprehensive answers.
4. **Conversational Coherence**: Maintaining context across multi-turn interactions.

From this, we can identify four key challenges for NLP models:

- **Ambiguous Intent**: Determining what users really want to know when questions have multiple interpretations.
- **Information Selection**: Choosing relevant additional context without overwhelming the user.
- **Domain Knowledge**: Understanding complex fictional universes with intricate relationships.
- **Response Anticipation**: Predicting and addressing likely follow-up questions.

How Does the Pragmatic Answer Enrich the Literal Answer?

We examine how the pragmatic answer enriches the literal answer that a non-cooperative teacher would produce by analyzing five sample conversations from the dataset.

Example 1 – Identity Expansion

Question: "What is Batman's real name?"

Literal: "Bruce Wayne"

Pragmatic Enhancement: Adds creator information (Bob Kane, Bill Finger) and publication history (Detective Comics #27, 1939).

Value: Provides cultural and historical context that enriches understanding.

Example 2 – Explanatory Context

Question: "Does Batman have superpowers?"

Literal: "No"

Pragmatic Enhancement: Explains what Batman relies on instead (intellect, detective skills, technology, wealth).

Value: Transforms a simple negative answer into an informative explanation.

Example 3 – Comprehensive Coverage

Question: "Who are Batman's biggest enemies?"

Literal: "The Joker and Catwoman"

Pragmatic Enhancement: Mentions additional villains such as Mr. Bloom.

Value: Provides broader context about Batman's rogues gallery.

Example 4 – Contextual Background

Question: "How old was Batman when he first became Batman?"

Literal: "I don't know"

Pragmatic Enhancement: Explains the timing relative to his parents' death and his oath.

Value: Turns an unknown into a contextual explanation.

Example 5 – Comparative Similarity

Question: "Is the Batman comic similar to the movies?"

Literal: Provides only basic family background information.

Pragmatic Enhancement: Adds specific details about the tragic origin story.

Value: Provides concrete details that support the similarity claim.

This demonstrates how pragmatic QA systems function as **cooperative conversational partners** rather than mere information retrieval tools, anticipating user needs and providing contextually rich responses.

Key Insights from the Dataset Analysis

1. **Dataset Structure:**
    - The dataset contains conversations from various fandoms (Comics, TV, Movies, etc.).
    - Each response is annotated with literal and pragmatic spans.
    - There are 179 validation conversations with an average of 8.5 questions per conversation.
    - The pragmatic-to-literal span ratio is 1.19, showing rich pragmatic content.
2. **Pragmatic Phenomena (from valid examples):**
    - **Cooperative responses:** Provide more than what was explicitly asked.
    - **Context expansion:** "Bruce Wayne" → adds creator information (Bob Kane, Bill Finger).
    - **Anticipatory answers:** Joker/Catwoman → mentions additional villains.
    - **Explanatory details:** "No superpowers" → explains what Batman relies on instead.
3. **Challenges for NLP Models:**
    - **Ambiguous Intent:** Understanding what users really want to know.
    - **Multi-step Reasoning:** Connecting information from multiple sources.
    - **Domain Knowledge:** Deep understanding of complex fictional worlds.
    - **Conversation Flow:** Maintaining coherence across multiple turns.
4. **Data Quality Issues:**
    - **Significant problem:** Many spans contain "Cannot GET /wiki/..." error messages.
    - **Impact:** Approximately 82% of early examples have corrupted span data.
    - **Solution:** Dataset includes valid examples (such as Batman conversations) that demonstrate concepts.
    - **Implication:** Models must filter for valid data when training and evaluating.

Part 1: Results Analysis

Total conversations analyzed: 179

---

1. Where the Traditional QA Model Succeeds

- Perfect Literal Answers: 3/179 (1.7%)
- Perfect Pragmatic Answers: 4/179 (2.2%)

Top Performing Topics:

- Supernanny: Literal=0.519, Pragmatic=0.391
- Alexander Hamilton: Literal=0.433, Pragmatic=0.404
- Popeye: Literal=0.422, Pragmatic=0.303
- Batman: Literal=0.409, Pragmatic=0.389
- Game of Thrones: Literal=0.397, Pragmatic=0.384

High-Scoring *Literal* Questions (F1 > 0.8): 3 cases

- Q: What year was the show released? A: 2005 (F1: 1.000)
- Q: Where did the Dinosaurs go? A: Extinction (F1: 1.000)
- Q: When was Popeye written? A: 1928 (F1: 1.000)

---

2. Where the Traditional QA Model Fails

- Zero Literal F1: 36/179 (20.1%)
- Zero Pragmatic F1: 48/179 (26.8%)
- Zero Retrieved F1: 129/179 (72.1%)

Worst Performing Topics:

- A Nightmare on Elm Street (2010): Literal=0.000, Pragmatic=0.000
- Enter the Gungeon: Literal=0.000, Pragmatic=0.000
- The Karate Kid: Literal=0.000, Pragmatic=0.500

Retrieval Failures: 44/179 (24.6%)

- A Nightmare on Elm Street (2010 film): 4 questions
- Alexander Hamilton: 17 questions
- The Wonderful Wizard of Oz (book): 3 questions
- Popeye: 20 questions

---

3. Literal vs Pragmatic Tendency

- Pragmatic > Literal: 47/179 (26.3%)
- Literal > Pragmatic: 53/179 (29.6%)
- Tied scores: 79/179 (44.1%)

Cases Where Pragmatic Context Helps:

- **Alexander Hamilton**
  Q: Who is starred as Alexander Hamilton in the musical Hamilton?
  Literal F1: 0.000 → Pragmatic F1: 1.000 (+1.000)
- **Dinosaur**
  Q: Hi. How long ago had the dinosaurs become extinct?
  Literal F1: 0.667 → Pragmatic F1: 1.000 (+0.333)
- **The Karate Kid**
  Q: When was The Karate Kid released?
  Literal F1: 0.000 → Pragmatic F1: 1.000 (+1.000)

---

4. Traditional QA Limitations Revealed

Context Length Analysis:

- Literal contexts: 77 chars (concise, targeted)
- Pragmatic contexts: 122 chars (slightly longer)
- Retrieved contexts: 62,188 chars (very long, noisy)

Answer Generation:

- Empty Literal answers: 14/179 (7.8%)
- Empty Pragmatic answers: 5/179 (2.8%)
- Empty Retrieved answers: 44/179 (24.6%)

---

5. Final Insights & Conclusions

Model Succeeds When:

- Given high-quality, targeted literal spans
- Dealing with factual, straightforward questions
- Working with well-documented topics (Batman, Game of Thrones)
- Context directly contains the answer

Model Fails When:

- Dealing with missing or corrupted source documents
- Requiring pragmatic inference or cooperative reasoning
- Working with very long, noisy retrieved contexts
- Questions need multi-step reasoning or implicit understanding

Literal vs Pragmatic Tendency:

- Model performs slightly better with literal spans (F1: 0.389)
- Pragmatic spans show close performance (F1: 0.359)
- Pragmatic improvement occurs in 26.3% of cases
- Traditional QA cannot generate truly cooperative responses
- Performance depends heavily on context quality, not reasoning ability

Assignment Goal Achieved:
This evaluation demonstrates that traditional extractive QA has clear limitations for pragmatic reasoning tasks, establishing the motivation for advanced LLM approaches in Part 2.

Implementation-Specific Limitations

Beyond the quantitative results, several limitations of the current implementation qualify how we should interpret the findings:

1. **Metrics Validity:** The SemanticF1 implementation derives precision and recall heuristically from the F1 score (via fixed multipliers). Thus, only F1 values are trustworthy for analysis.
2. **Context Handling:** Retrieved passages are truncated by characters (2000 char limit), not tokens. This can cut answers mid-span and ignores DistilBERT's max token length. Passages are concatenated naively (top-3), without windowing or stride, leading to loss of relevant information and injection of noise.

3. **Retrieval and Data Quality:** Some topics resolve to missing or corrupted sources, yielding empty contexts. In addition, the lightweight `all`-MiniLM-`L6-v2` embedder may reduce retrieval quality for nuanced or pragmatic questions.

4. **Model Choice:** The QA backbone is DistilBERT fine-tuned on SQuAD. This model is optimized for literal span extraction, not for pragmatic or cooperative reasoning, limiting task performance by design.

5. **Evaluation Scope:** As required, evaluation was restricted to the *first question* of each conversation. While faithful to the assignment, this underestimates the challenges posed by multi-turn dialogue.

6. **Analysis Artifacts:** Reported averages combine cases with valid retrieval and cases with complete retrieval failure, conflating model and retriever limitations.

These limitations explain why literal spans outperform retrieved contexts, why pragmatic improvements are modest, and why reported precision/recall should be treated with caution.

Table 1: Summary of Traditional QA Results (179 conversations)

Category	Count / Cases	Percentage	Notes / F1
Overall Success			
Perfect Literal Answers	3 / 179	1.7%	F1 = 1.0
Perfect Pragmatic Answers	4 / 179	2.2%	F1 = 1.0
High-Scoring *Literal* (F1 > 0.8)	3 cases	–	Show, Dinosaurs, Popeye
Mean F1 (Literal / Prag. / Ret.)	–	–	0.389 / 0.359 / 0.122
Failures			
Zero Literal F1	36 / 179	20.1%	–
Zero Pragmatic F1	48 / 179	26.8%	–
Zero Retrieved F1	129 / 179	72.1%	–
Retrieval Failures	44 / 179	24.6%	Missing documents
Answer Generation (Empty outputs)			
Empty Literal answers	14 / 179	7.8%	–
Empty Pragmatic answers	5 / 179	2.8%	–
Empty Retrieved answers	44 / 179	24.6%	–
Topic Examples			
Top Performers	–	–	Supernanny (0.519 / 0.391), Alexander Hamilton (0.433 / 0.404), Popeye (0.422 / 0.303)
Worst Performers	–	–	Elm Street 2010 (0.000 / 0.000), Enter the Gungeon (0.000 / 0.000), Karate Kid (0.000 / 0.500)
Literal vs Pragmatic Comparison			
Pragmatic > Literal	47 / 179	26.3%	Example: Hamilton +1.000
Literal > Pragmatic	53 / 179	29.6%	–
Tied Scores	79 / 179	44.1%	–

Part 2: Results Analysis

4.4.1 First Questions (LLM Program vs. Traditional QA)

Setup and link to prior work. Following Qi et al. (2023), we evaluate on the *first* question of each conversation (179/179 covered), holding retrieval and scoring protocols fixed. We treat *SemanticF1* (decompositional) as our summary statistic, conceptually aligned with the paper's literal/pragmatic spans ($F_1^{\text{lit}}, F_1^{\text{prag}}$). We compare our LLM program to the Part 1 extractive baseline.

Method	**Questions**	**Valid**	**Excluded (%)**	**SemanticF1**
Extractive baseline (Part 1)	179	179	0 (0%)	0.389
LLM program (first turns)	179	156	23 (12.8%)	**0.407**

Table 2: First-turn results on PragmatiCQA. LLM SemanticF1 is the corrected mean over valid questions. Original unfiltered mean: 0.355.[2]

Results on first questions. The LLM outperforms the extractive baseline on first-turn questions by an **absolute** ΔF1 of **+0.018** (0.407 vs. 0.389), i.e., a **+4.6%** *relative* gain.[3] Qualitatively, wins occur when the gold expects a short literal fact plus a concise pragmatic addendum; losses concentrate when retrieval is weak or noisy, where the LLM may over-generalize.

Comparison to the paper's baseline. The paper's text-to-text FiD (BART-large + DPR) is principled for multi-document generation but remains challenged by faithfulness, disambiguation, and pragmatic recovery. Against that backdrop, our LLM's *first-turn* gains derive from surfacing small, salient pragmatic nuggets despite retrieval noise. (Note: our "3.4×" figure is a directional comparison under non-identical setups/metrics and should be interpreted qualitatively.)

Caveats. (i) Only *first-turn* questions are considered here; multi-turn evaluation (Section 4.4.2) can favor extractive spans under overlap-based metrics. (ii) Our pipeline truncates by characters and concatenates passages naively, limiting headroom for both systems.

4.4.2 All Turns (LLM Program)

Setup. We evaluate the LLM program on *all* turns in each conversation. Of 1,526 total questions in the split, 1,496 were processed; 1,291 were valid after filtering.

[2] Precision/recall reported by our tool are heuristic; F1 is the reliable figure. The LLM's corrected average (0.407) excludes 23 invalid cases (e.g., corrupted/missing contexts); the raw mean over all 179 is 0.355. Timestamp: 2025-09-09_19:35:28.

[3] Relative gain computed as $(0.407 - 0.389)/0.389$.

Method	Questions	Valid	Excluded (%)	SemanticF1
LLM program (all turns)	1,496	1,291	205 (13.7%)	**0.410**

Table 3: Multi-turn results on PragmatiCQA. SemanticF1 is the corrected mean over valid questions; the original unfiltered mean was 0.354. Timestamp: 2025-09-09_23:18:39.

Results on later questions. A split by position shows essentially flat means: first-turn (valid n=152) F_1=0.410; later-turns (valid n=1,139) F_1=0.410. Performance varies by depth (e.g., turn 6: 0.436, n=76; turn 8: 0.501, n=51) but does not increase monotonically with more history.

Comparison to Part 1. Relative to the Part 1 extractive baseline on first turns (0.389), the multi-turn LLM average (0.410) reflects a **+5.5%** relative improvement.[4]

Where history helps. Gains typically occur when dialog history clarifies coreference (entities, pronouns), disambiguates underspecified requests, or highlights a short literal fact plus a small pragmatic addendum. Failures remain tied to noisy/overlong retrieval and occasional over-generalization.

[4]Computed as (0.410−0.389)/0.389. Baseline was not re-run in multi-turn mode; comparison is provided for context only.

Discussion Questions

Comparison of Models

Traditional extractive QA excels when the gold answer appears verbatim in short, targeted contexts; it is brittle under long/noisy concatenations and does not provide cooperative enrichment.

The LLM program integrates literal facts with pragmatic addenda and tolerates imperfect phrasing, aligning better with free-form gold answers. Quantitatively, it improves over the Part 1 baseline on first turns (0.407 vs. 0.389, Δ=+0.018, +4.6% relative) and sustains a similar level on all turns (0.410 corrected mean over 1,291 valid questions). Weaknesses include occasional hallucinations and verbosity that can depress span-based F1 despite pragmatically useful content. (Qi et al., 2023).

First vs. later questions

Empirically, we observe *negligible* average differences between first and later turns in our setup: first-turn F_1=0.410 (valid n=152) vs. later-turns F_1=0.410 (valid n=1,139). There are pockets where more history helps (e.g., turn 8 mean 0.501, n=51), but improvements are not monotonic with depth. This suggests that history helps in specific discourse patterns (coreference, disambiguation), while retrieval quality remains the main bottleneck. (Qi et al., 2023).

Theory of Mind (ToM)

Our LLM displays *functional* ToM-like behavior—anticipating likely follow-ups and adding relevant context—consistent with the dataset's cooperative design. However, behavior under retrieval failure (confident additions without evidence) indicates sophisticated pattern matching rather than explicit belief modeling. In line with Qi et al. (2023), high F_1^{prag} is a useful but imperfect proxy for "true" pragmatic competence; grounding and retrieval fidelity are decisive. (Qi et al., 2023).

References

Peng Qi, Nina Du, Christopher D. Manning, and Jing Huang. Pragmaticqa: Pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6175–6191, 2023.