QUDSELECT

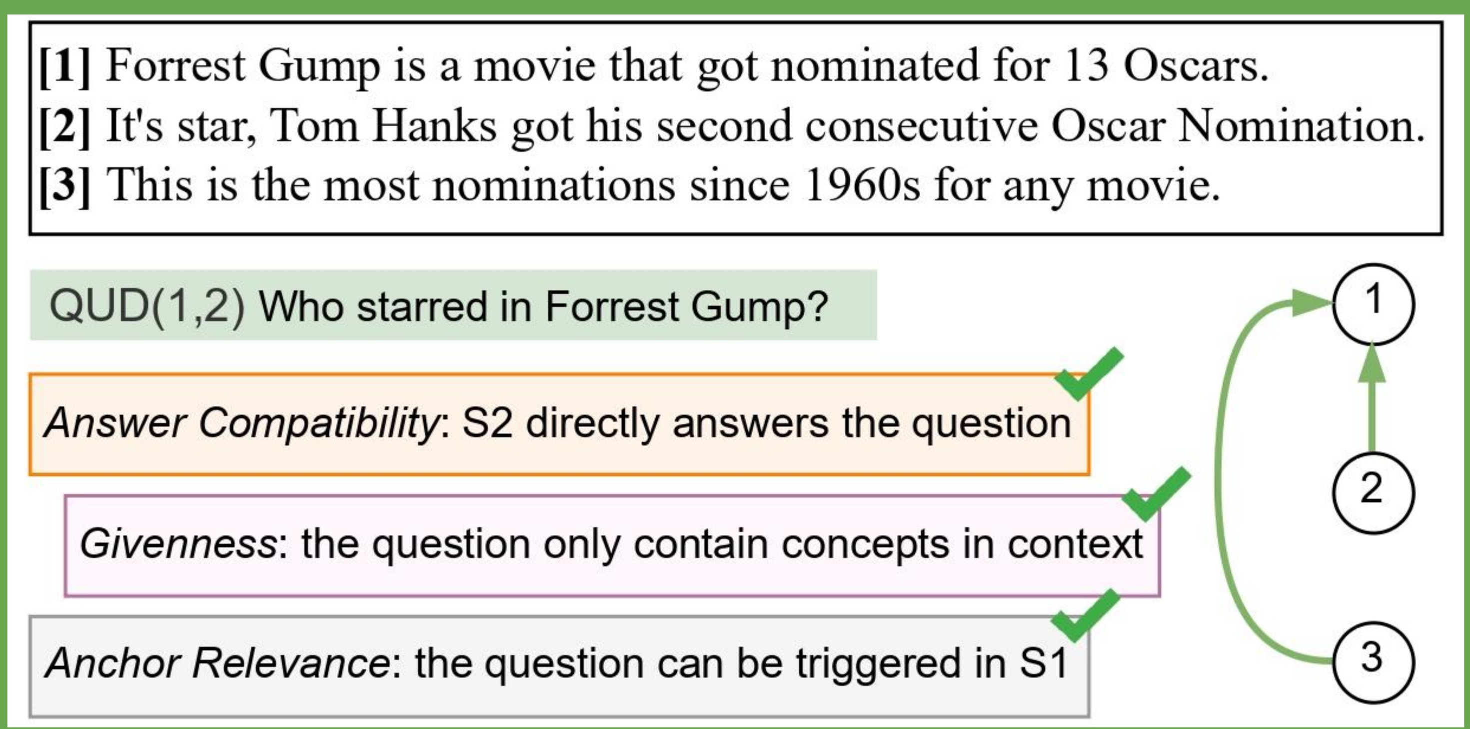
Selective Decoding for Questions Under Discussion Parsing

Ashima Suvarna*♡ Xiao Liu*◇ Tanmay Parekh♡ Kai-Wei Chang♡ Nanyun Peng♡

♡ UCLA ◇ Peking University *Equal Contribution

Read our paper!

Parse complex articles in a human-friendly manner as questions with QUDSELECT!!

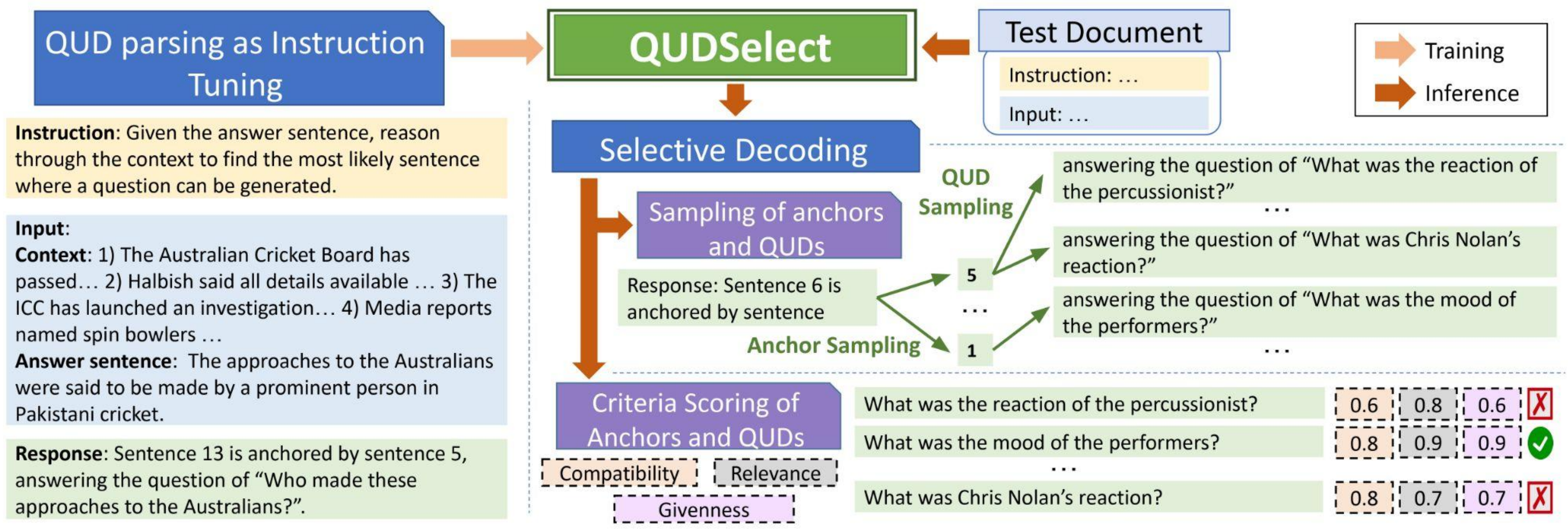


Joint Framework to **parse QUD structures** by integrating key **theoretical criteria:**

Answer Compatibility | **Givenness** | **Anchor Relevance**

We view QUD parsing as **instruction tuning task** and **selectively decode** candidate questions and anchors.

QUDSELECT Framework



- We first **instruction tune a joint QUD parser.** Given the answer sentence, models are instructed to predict the anchor and question together.
- Then we propose **selective decoding** where we apply the three key **principles of QUD as our criteria** to assess the quality of generated ⟨anchor sentence, question⟩ pairs and select the best.
- We implement **reference-free** and **training-free** scorers for each of them, namely, answer compatibility, givenness and anchor relevance.

Experimental Setup

DATASET : DCQA (22k questions, 606 news articles)

BASELINES: LLaMA2-7B, Mistral-7B, Pipeline, GPT-4

EVALUATION CRITERIA : Satisfaction of Theoretical Criteria of QUD - answer compatibility, givenness, anchor relevance

Human Evaluation

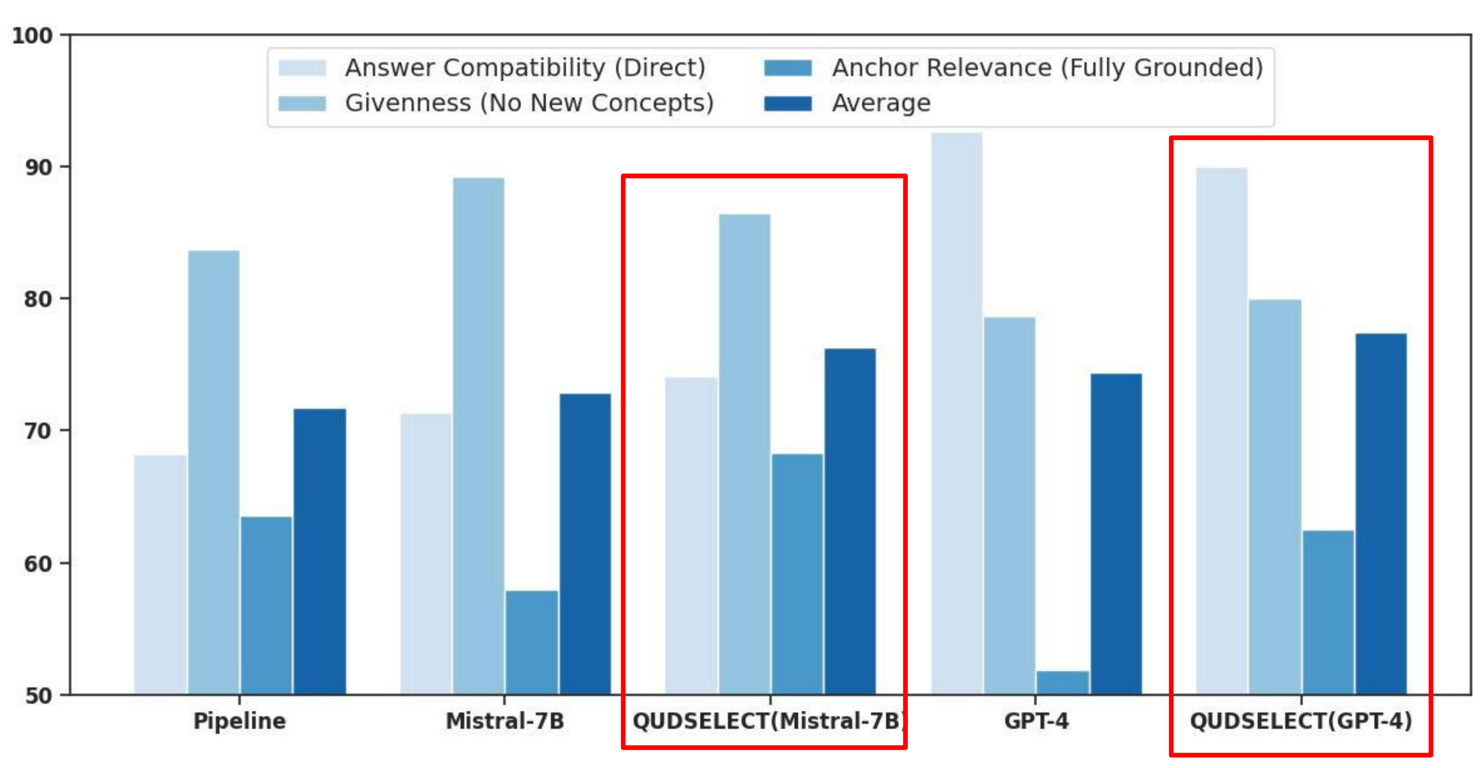
Model	Answer Compatibility Dir. (↑)	Unfocus.	No Ans.(↓)	Givenness No New (↑)	Ans. leak. (↓)	Hall. (↓)	Anchor Relevance Fully G. (↑)	Partial. G.	No G. (↓)	Avg. (↑)
HUMAN EVALUATION										
Pipeline	52.5	15.0	32.5	53.8	28.7	17.5	50.0	32.5	17.5	52.1
Mistral-7B	67.0	15.4	17.6	60.3	23.6	16.1	58.6	29.0	12.4	62.0
+ QUDSELECT	**67.1**	20.0	**12.9**	**77.6**	**20.0**	**2.4**	**68.2**	24.7	**7.1**	**71.0**

QUDSELECT outperforms baselines by **~9%** on human evaluation. Human annotators find that QUDSELECT leads to **directly answerable questions**, **fully grounds** in context and **satisfies givenness**.

QUDs generated by Different Models

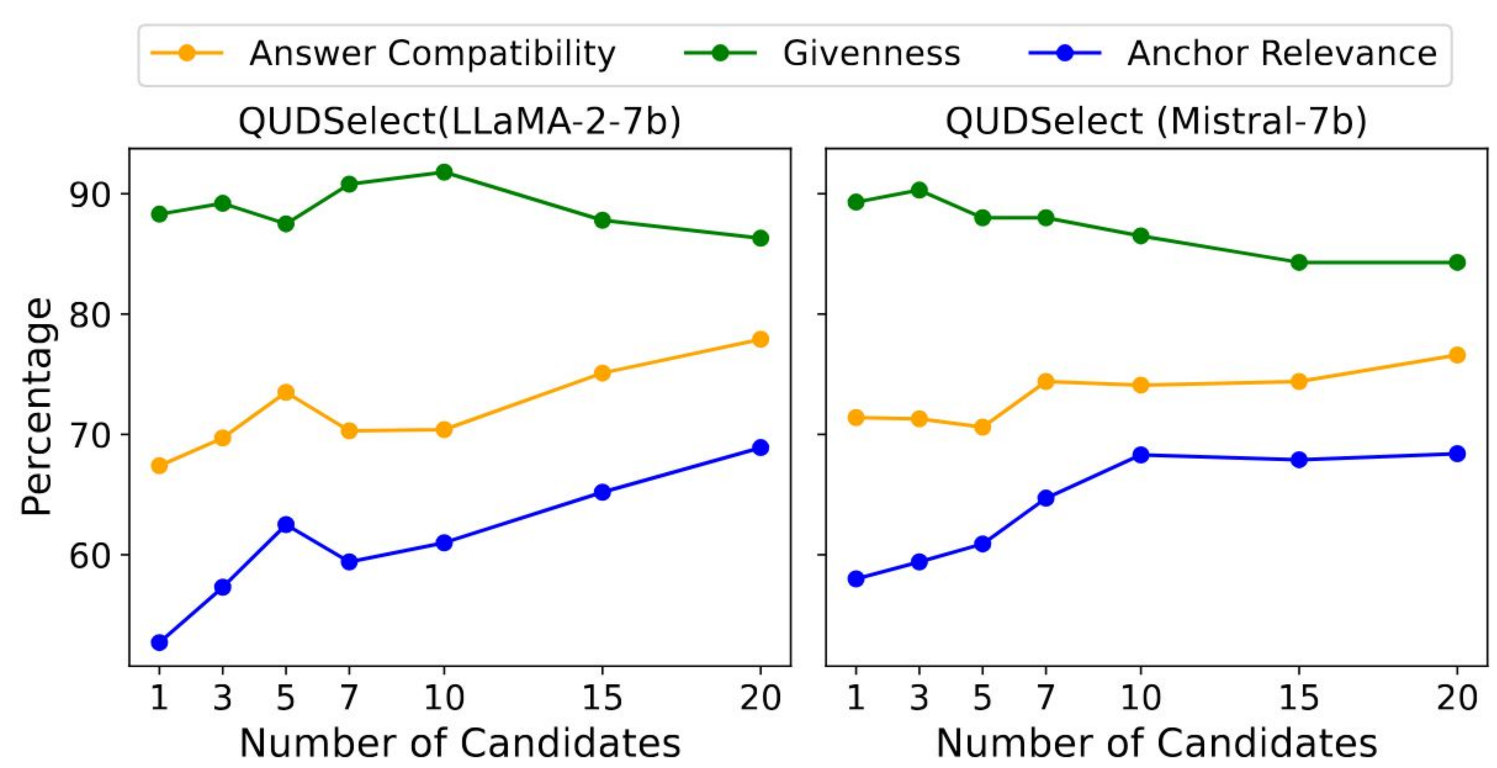
Pipeline (Ko et al. (2023))	
Answer: s_3 Anchor: s_2 QUD: "What does Glenn think is the future outlook on nuclear materials?"	✗Non answer ✗Answer leakage ✓Partially grounded
Answer: s_4 Anchor: s_2 QUD: "Who is the Sen. Glenn from?"	✗Nonsensical question
LLaMA2	
Answer: s_4 Anchor: s_3 QUD: "What is deadly contra-band?"	✗Non answer ✓No new concepts ✗Partially grounded
Answer: s_3 Anchor: s_1 QUD: "Why is it difficult to trace nuclear material?""	✗Non answer ✓No new concepts ✓Fully grounded
QUDSELECT (LLaMA2)	
Answer: s_4 Anchor: s_2 QUD: "Who requested the re-port?"	✓Direct answer ✓No new concepts ✓Fully grounded
Answer: s_3 Anchor: s_1 QUD: "What is the reason for the inability to trace nuclear material?"	✓Indirect Answer ✓No new concepts ✗Partially grounded
QUDSELECT (Mistral)	
Answer: s_3 Anchor: s_1 QUD: "Why is it important that U.S. exports of nuclear material cannot be adequately traced from country to country?"	✓Direct answer ✓No new concepts ✓Fully grounded
Answer: s_4 Anchor: s_2 QUD: "Who commissioned the report?"	✓Direct answer ✓No new concepts ✓Fully grounded

Automatic Evaluation



- Train **supervised classifiers** for each evaluation criterion on expert annotated **data from QUDEVAL.**
- Our automatic evaluators achieve average macro **F1 of 0.47** across all evaluation criteria.

Number of Candidates vs QUDSELECT



We see an **upward trend as k grows** for **Answer Compatibility** and **Anchor Relevance** while Givenness is sacrificed by a small margin for better overall performance.