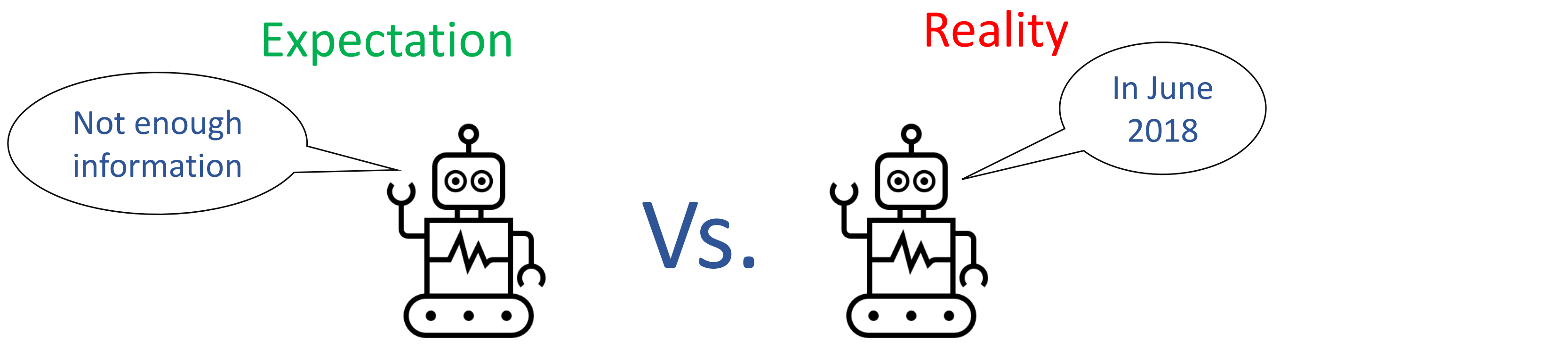
The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, Shauli Ravfogel

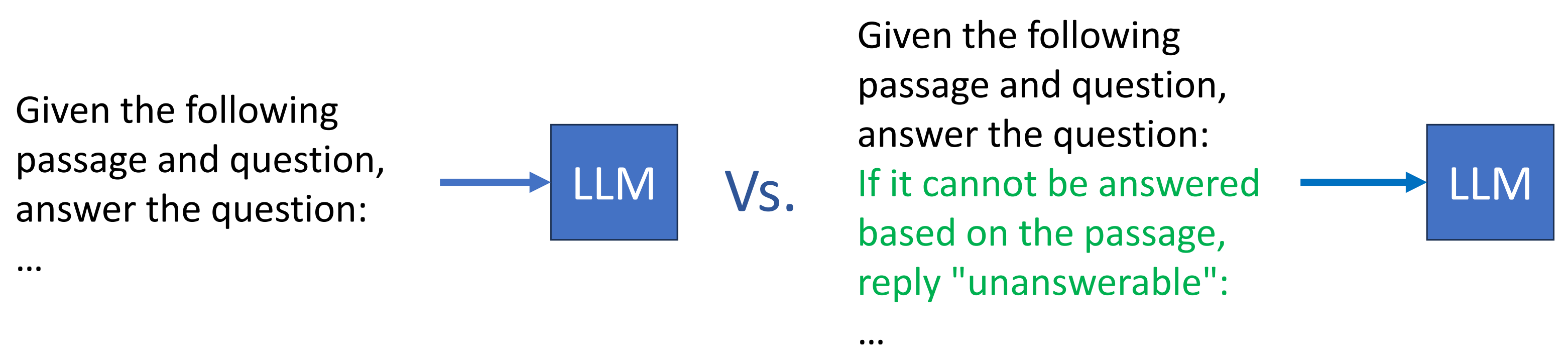
Passage:
In June 2018, Kelley told E! News that she would not be returning for season two of Dynasty.
Based on the passage, answer the Question:
when does Dynasty come back for season 2?



Do LLMs already represent (un)answerability when producing answers?

- We study (un)answerability with:
  1. Prompting
  2. (Un)answerability-aware decoding
  3. Probing
- Three models: Flan-T5-XXL (11B), Flan-UL2 (20B), OPT-IML (30B)
- Three benchmarks: SQuAD, Natural Questions (NQ), MuSiQue

Prompting


Given the following passage and question, answer the question:
…

Vs.

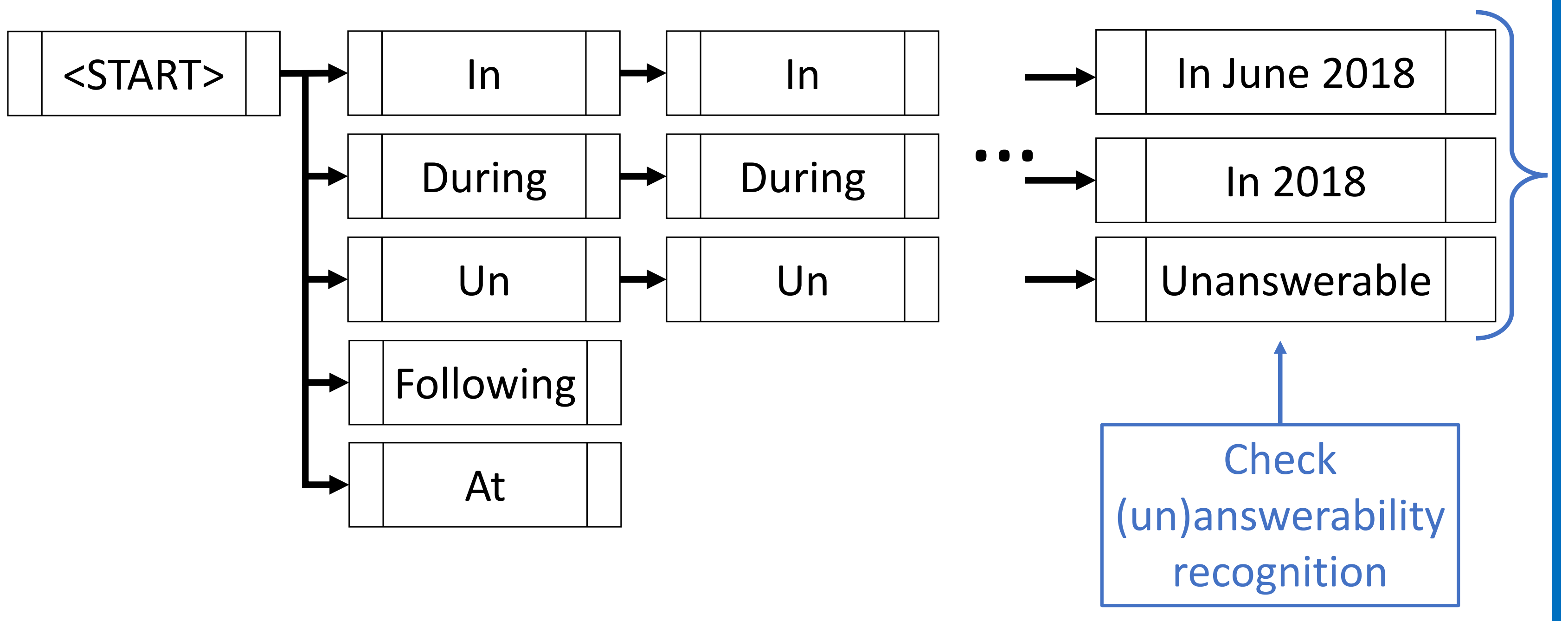
Given the following passage and question, answer the question:
If it cannot be answered based on the passage, reply "unanswerable":
…

Results

		SQuAD	NQ	MuSiQue
Flan-T5$_{xxl}$	Regular	37.3	5.7	8.2
(11B)	+hint	**91.5**	**85.6**	**74.7**
Flan-UL2	Regular	46.3	13.8	6.8
(20B)	+hint	**92.3**	**83.9**	**67.3**
OPT-IML	Regular	43.9	21.8	17.1
(30B)	+hint	**85.4**	**85.3**	**54.1**

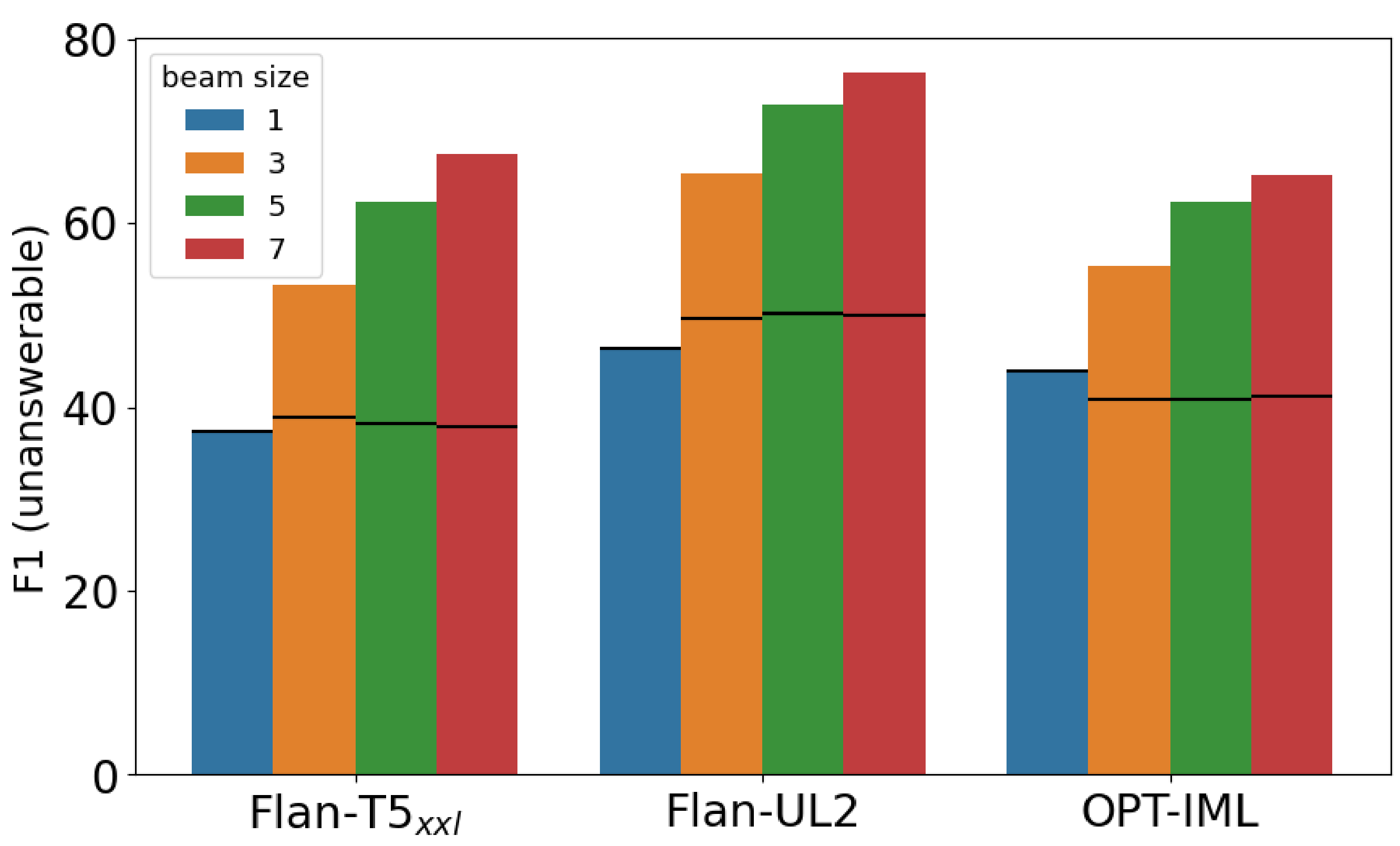
(Un)answerability-aware Decoding

Passage:
In June 2018, Kelley told E! News that she would not be returning for season two of Dynasty.
Based on the passage, answer the Question:
when does Dynasty come back for season 2?



Check (un)answerability recognition

Results

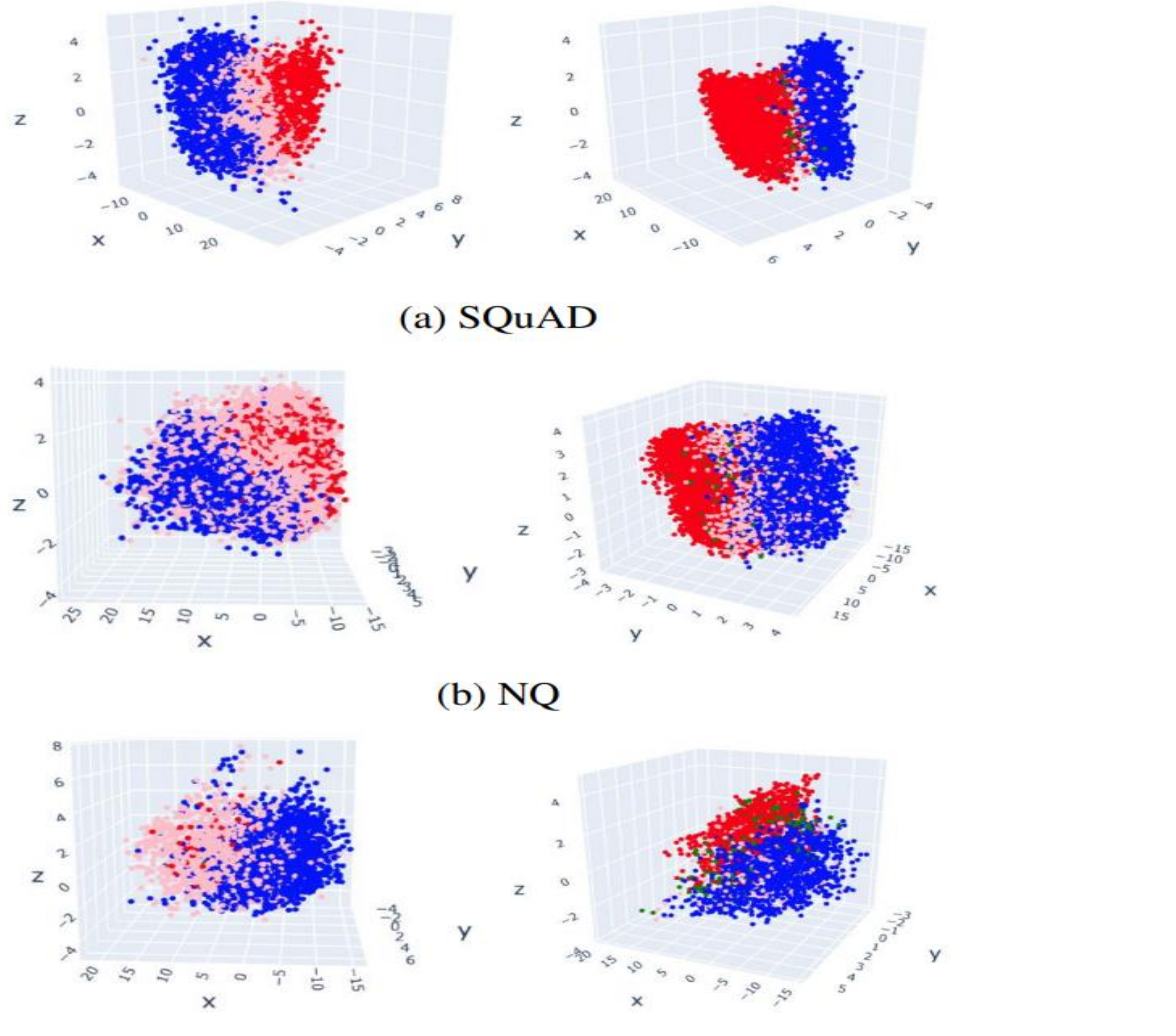


Probing

- Train linear classifier on model's embedding space.

Model		SQuAD		NQ		MuSiQue	
		1st layer	last layer	1st layer	last layer	1st layer	last layer
Flan-T5$_{xxl}$	regular	40.1	89.9	23.0	86.1	47.4	77.5
(11B)	+hint	40.0	89.4	26.6	86.2	38.6	77.3
Flan-UL2	regular	39.4	90.4	42.2	87.3	15.1	78.3
(20B)	+hint	39.6	89.9	41.5	87.9	41.6	78.3
OPT-IML	regular	48.4	82.8	40.8	85.5	45.6	75.5
(30B)	+hint	48.4	83.9	45.3	86.2	45.6	84.9

- PCA projection of embedding space onto 3-D plane.



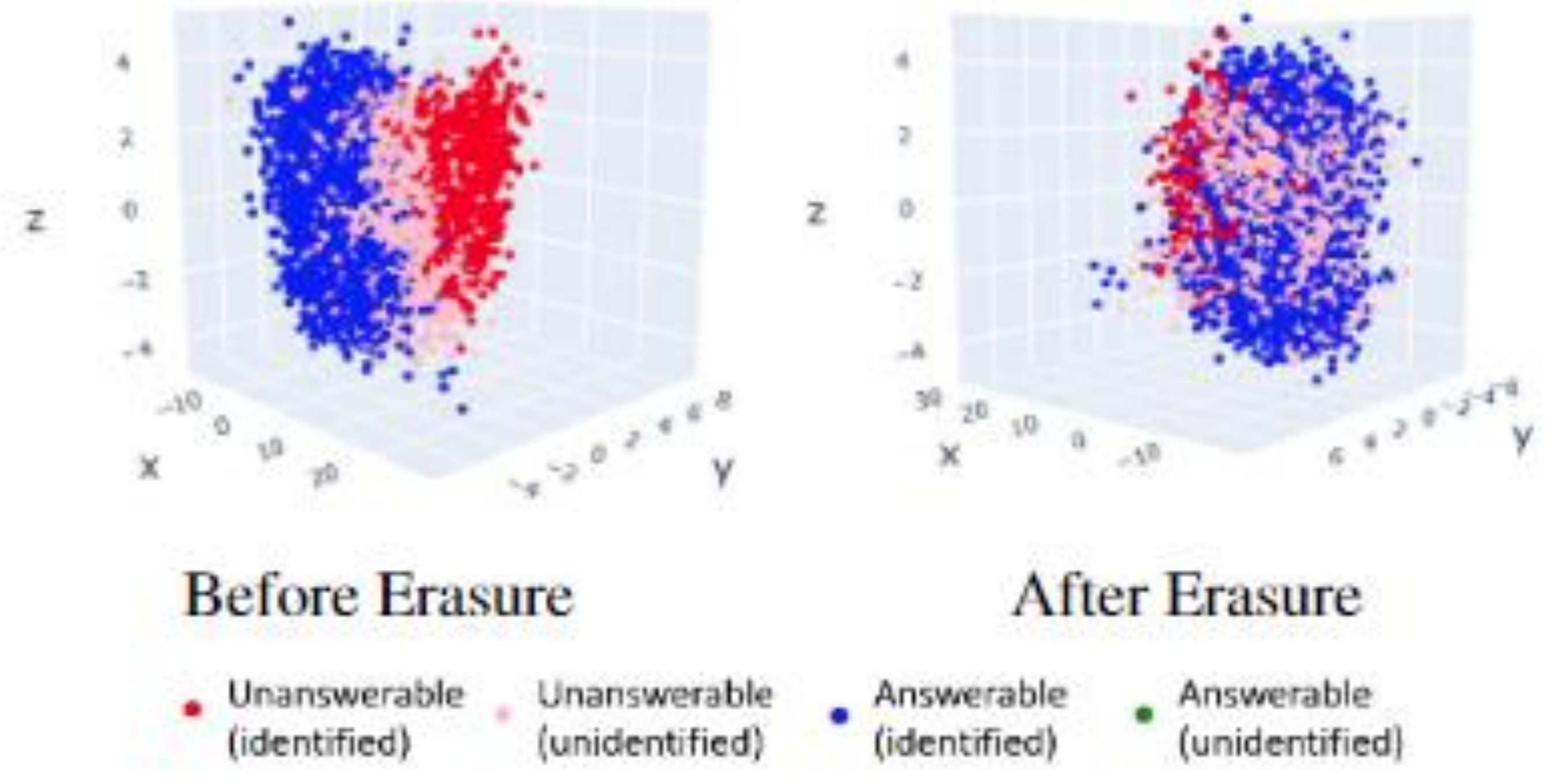
(a) SQuAD

(b) NQ

(c) MuSiQue

🟥	Unanswerable (identified)	🟦	Answerable (identified)
●	Unanswerable (unidentified)	🟩	Answerable (unidentified)

- Erase answerability subspace



		All		Answerable	
k-beam Type		EM	F1	EM	F1
Regular	w\o erasure	60.2	63.8	87.1	94.1
	with erasure	50.2	55.1	82.0	91.8
Relaxed	w\o erasure	60.9	64.4	87.1	94.1
	with erasure	50.8	55.7	82.0	91.8