Advancing Question Generation with Joint Narrative and Difficulty Control

Bernardo Leite and Henrique Lopes Cardoso

LIACC/FEUP

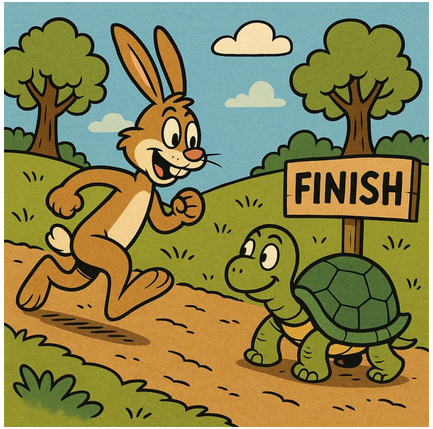
ACL 2025 VIENNA

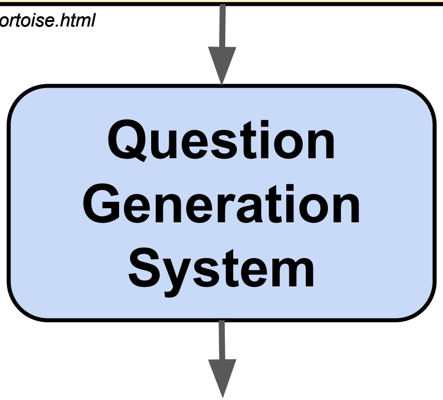
20th Workshop on Innovative Use of NLP for Building Educational Applications
July 31 and August 1, 2025, Vienna, Austria

Question Generation

“Once there were a hare and a turtle. The hare was proud of his speed and challenged the turtle to a race. Although the turtle was slow, he accepted. The hare quickly left the turtle behind but decided to rest and fell asleep. Meanwhile, the turtle kept going steadily and eventually reached the finish line first, winning the race.”

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html



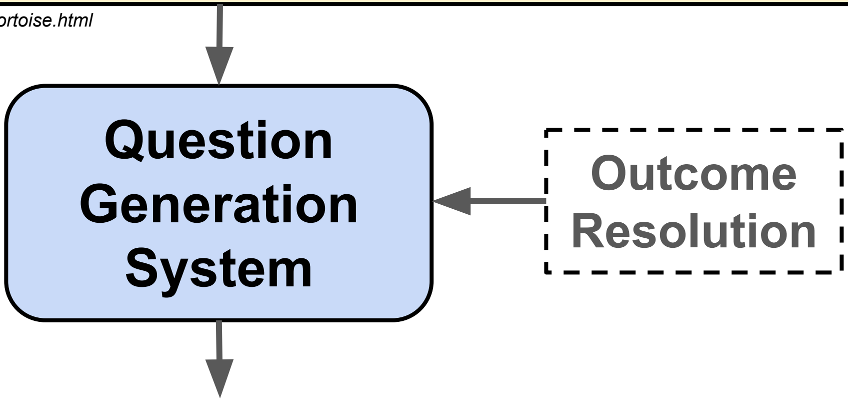


Who challenged the turtle to a race?

Question Generation with Narrative Control

“Once there were a hare and a turtle. The hare was proud of his speed and challenged the turtle to a race. Although the turtle was slow, he accepted. The hare quickly left the turtle behind but decided to rest and fell asleep. Meanwhile, the turtle kept going steadily and eventually reached the finish line first, winning the race.”

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html



Question Generation System

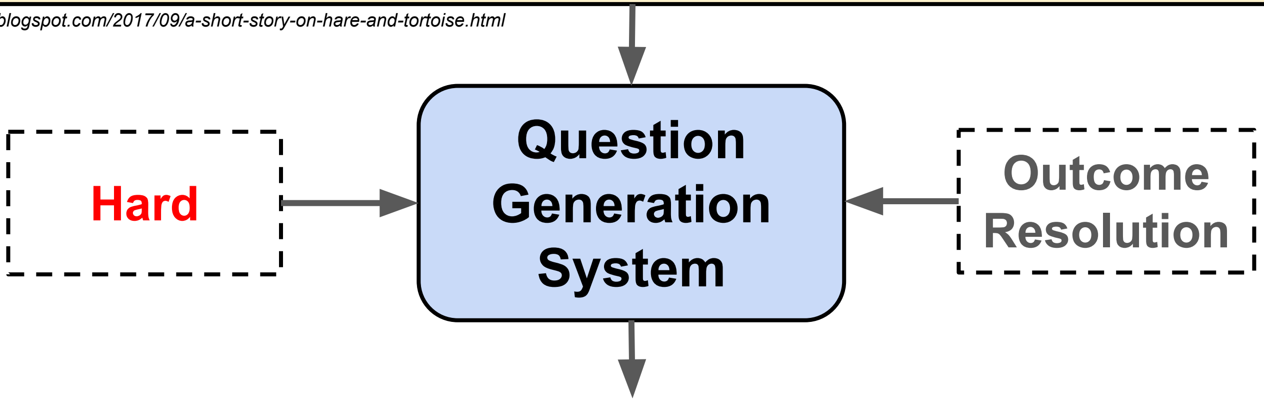
Outcome Resolution

What happened after the hare left the turtle behind?

Joint Narrative and Difficulty Control (**This Study**)

“Once there were a hare and a turtle. The hare was proud of his speed and challenged the turtle to a race. Although the turtle was slow, he accepted. The hare quickly left the turtle behind but decided to rest and fell asleep. Meanwhile, the turtle kept going steadily and eventually reached the finish line first, winning the race.”

source: http://simpleessaywriting.blogspot.com/2017/09/a-short-story-on-hare-and-tortoise.html



Hard

Question Generation System

Outcome Resolution

What happened because the turtle kept going steadily?

Main Research Question

How effectively can we control the generation of question-answer pairs conditioned on both narrative and difficulty attributes?

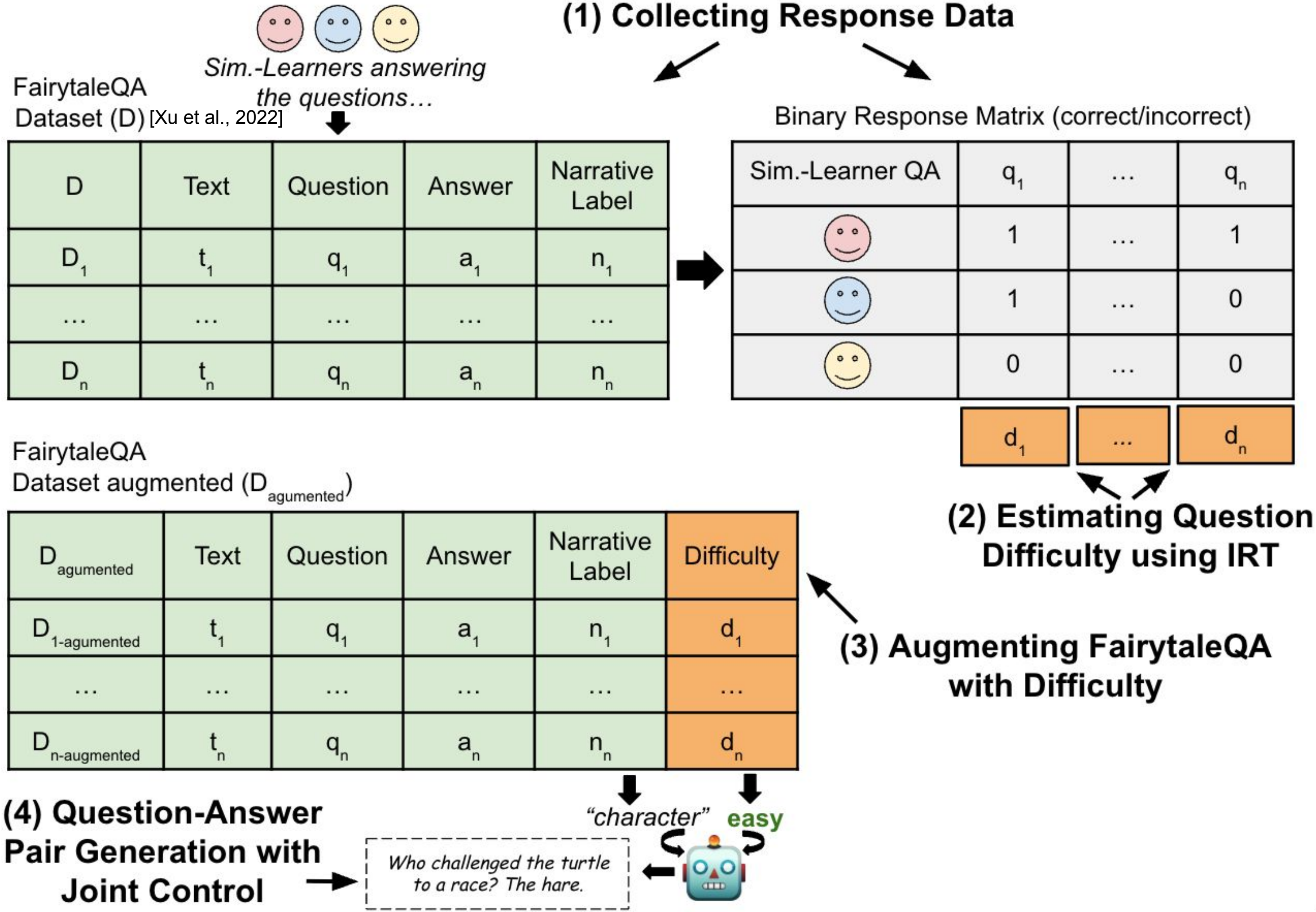
Controllable Question Generation

- **Content** Control
  - Question Reading Comprehension Skills [Ghanem et al., 2022]
  - Question Explicitness [Leite et al., 2023]
  - Question Bloom’s Taxonomy [Elkins et al., 2024] [Hwang et al., 2024]
  - Question Narrative Elements [Zhao et al., 2022] [Li and Zhang, 2024]
- **Difficulty** Control
  - QA Systems Performance [Gao et al., 2019]
  - Named Entity Popularity [Kumar et al., 2019]
  - Number of Inference Steps [Cheng et al., 2021]
  - Item Response Theory [Uto et al., 2023]

Relation Between Question Difficulty and Learner Ability

Multi-attribute Control

Overall Methodology for Joint Control



(1) Collecting Response Data

FairytaleQA Dataset (D) [Xu et al., 2022]

Sim.-Learners answering the questions…

D	Text	Question	Answer	Narrative Label
D_1	t_1	q_1	a_1	n_1
…	…	…	…	…
D_n	t_n	q_n	a_n	n_n

Binary Response Matrix (correct/incorrect)

Sim.-Learner QA	q_1	…	q_n
☺	1	…	1
☺	1	…	0
☺	0	…	0

d_1	...	d_n

(2) Estimating Question Difficulty using IRT

FairytaleQA Dataset augmented ($D_{agumented}$)

$D_{agumented}$	Text	Question	Answer	Narrative Label	Difficulty
$D_{1\text{-}agumented}$	t_1	q_1	a_1	n_1	d_1
…	…	…	…	…	…
$D_{n\text{-}augmented}$	t_n	q_n	a_n	n_n	d_n

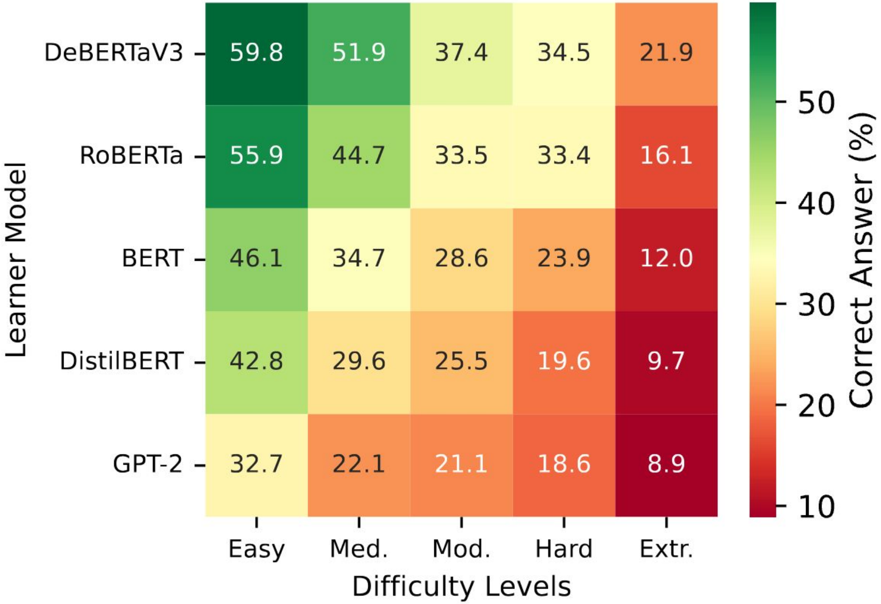
(3) Augmenting FairytaleQA with Difficulty

(4) Question-Answer Pair Generation with Joint Control

"character" easy

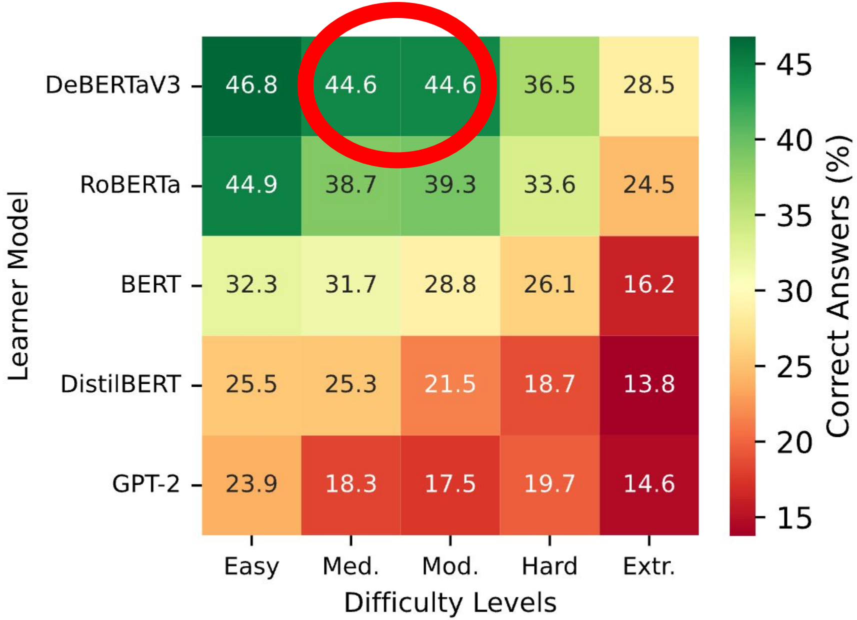
Who challenged the turtle to a race? The hare.

Results for Difficulty Control



Learner Model	Easy	Med.	Mod.	Hard	Extr.
DeBERTaV3	59.8	51.9	37.4	34.5	21.9
RoBERTa	55.9	44.7	33.5	33.4	16.1
BERT	46.1	34.7	28.6	23.9	12.0
DistilBERT	42.8	29.6	25.5	19.6	9.7
GPT-2	32.7	22.1	21.1	18.6	8.9

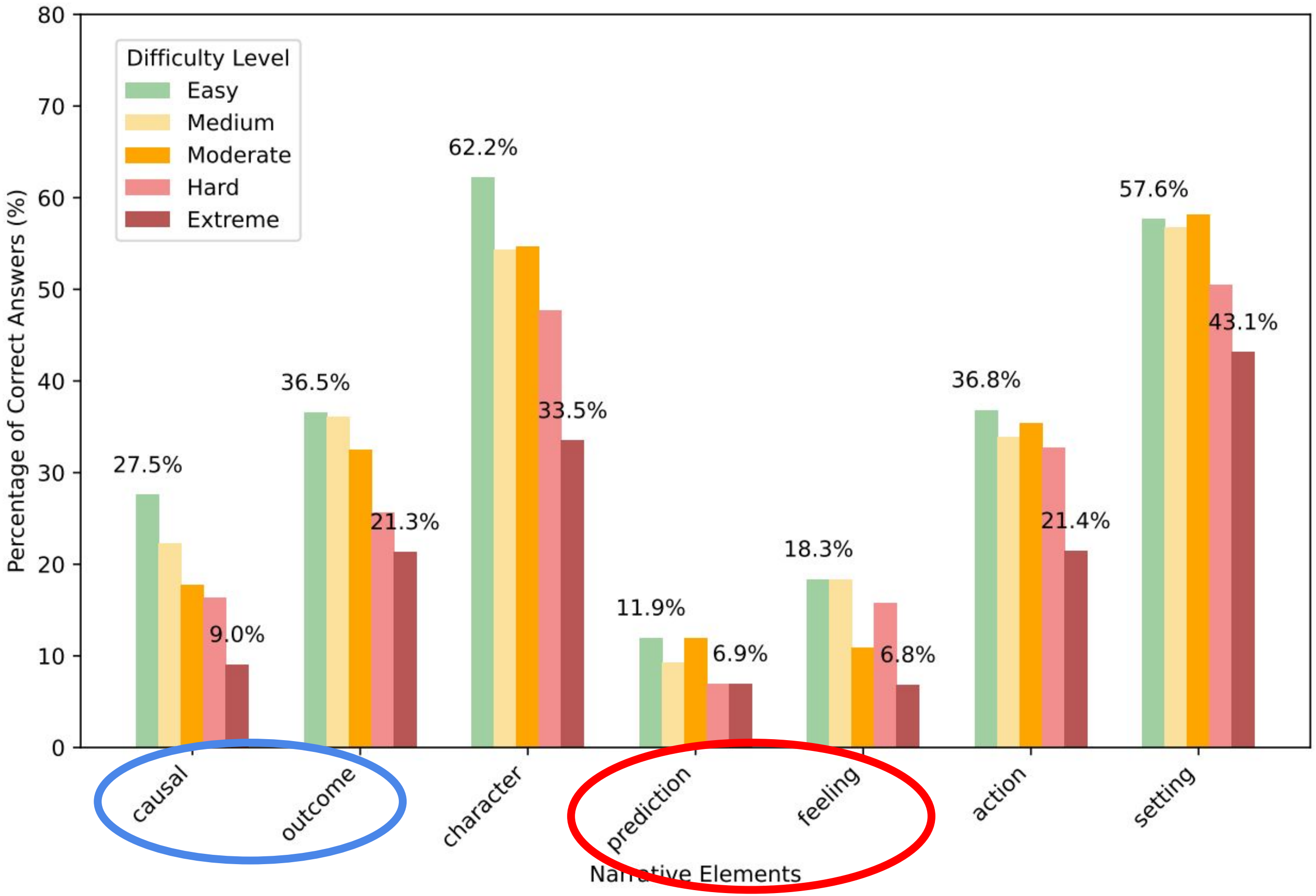
Dif + Text → QA



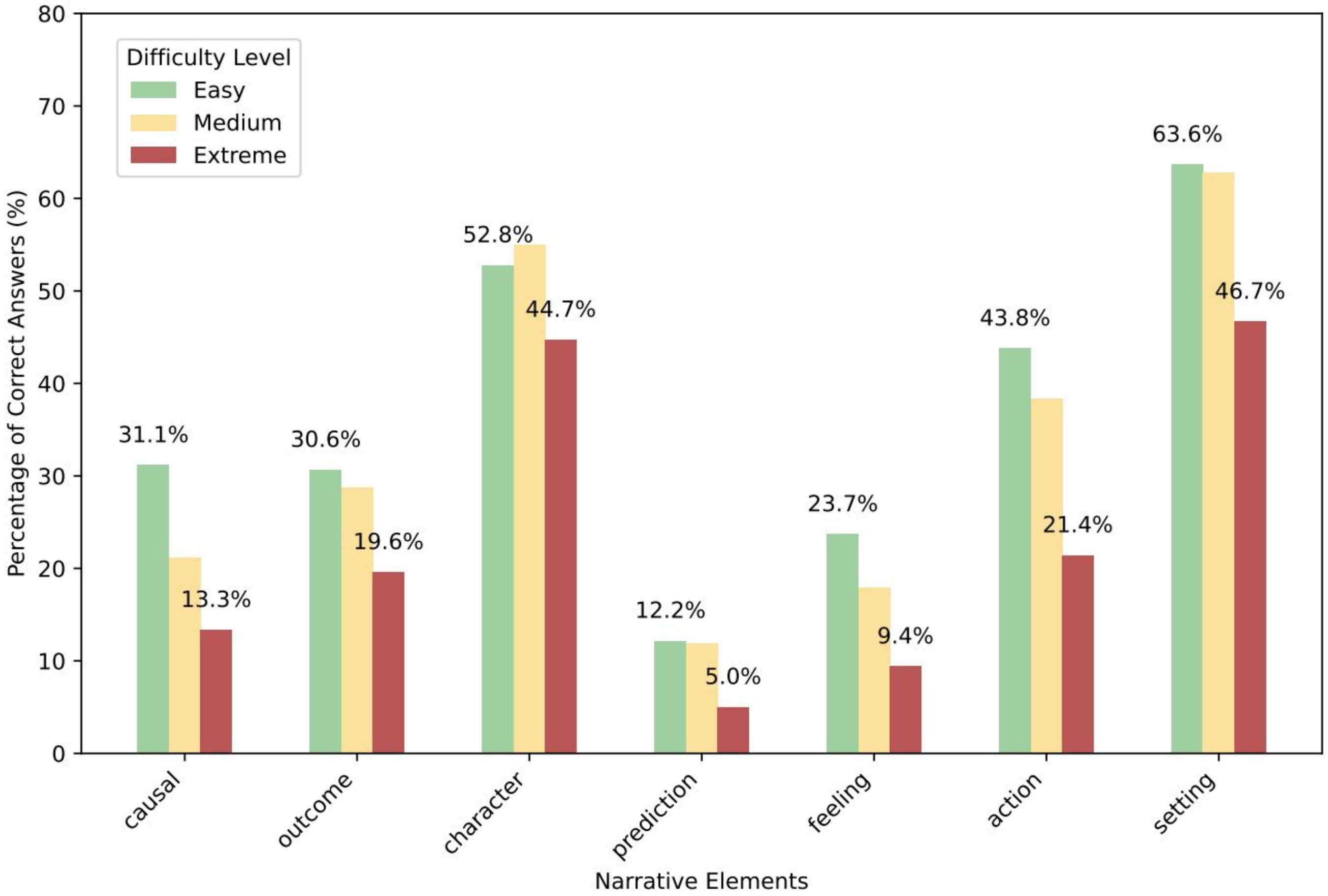
Learner Model	Easy	Med.	Mod.	Hard	Extr.
DeBERTaV3	46.8	44.6	44.6	36.5	28.5
RoBERTa	44.9	38.7	39.3	33.6	24.5
BERT	32.3	31.7	28.8	26.1	16.2
DistilBERT	25.5	25.3	21.5	18.7	13.8
GPT-2	23.9	18.3	17.5	19.7	14.6

Nar + Dif + Text → QA

Results for Joint Narrative and Difficulty Control



Results for Joint Narrative and Difficulty Control



Additional Findings

- Harder Questions are More Diverse
- Error Analysis
    - Hallucinated Content (14%)
    - Nonsensical QA pairs (10%)
- Repeated QA Pairs using Beam Search

Key Takeaways

- Joint Control for Question Generation *can* be feasible
- More Datasets, More Attributes, More Control
- Controlled Question Generation with *Real* Students

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO DEI DEPARTAMENTO DE ENGENHARIA INFORMÁTICA LIACC fct Fundação para a Ciência e a Tecnologia REPÚBLICA PORTUGUESA

Advancing Question Generation with Joint Narrative and Difficulty Control

Bernardo Leite and Henrique Lopes Cardoso

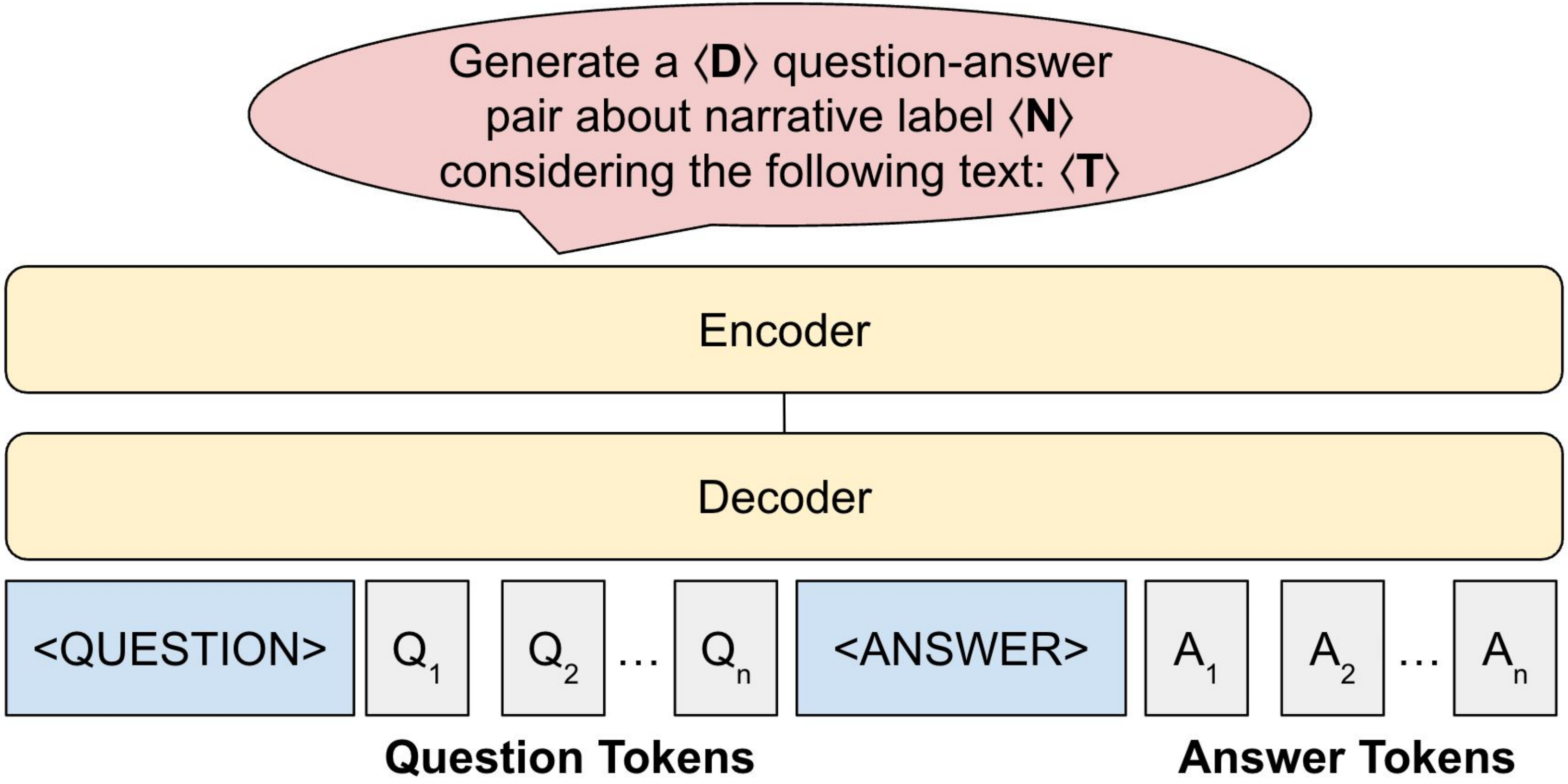
LIACC/FEUP

ACL 2025 VIENNA

20th Workshop on Innovative Use of NLP for Building Educational Applications
July 31 and August 1, 2025, Vienna, Austria

Supplementary Information

Question-Answer Pair Generation Model



Generate a ⟨**D**⟩ question-answer pair about narrative label ⟨**N**⟩ considering the following text: ⟨**T**⟩

Encoder

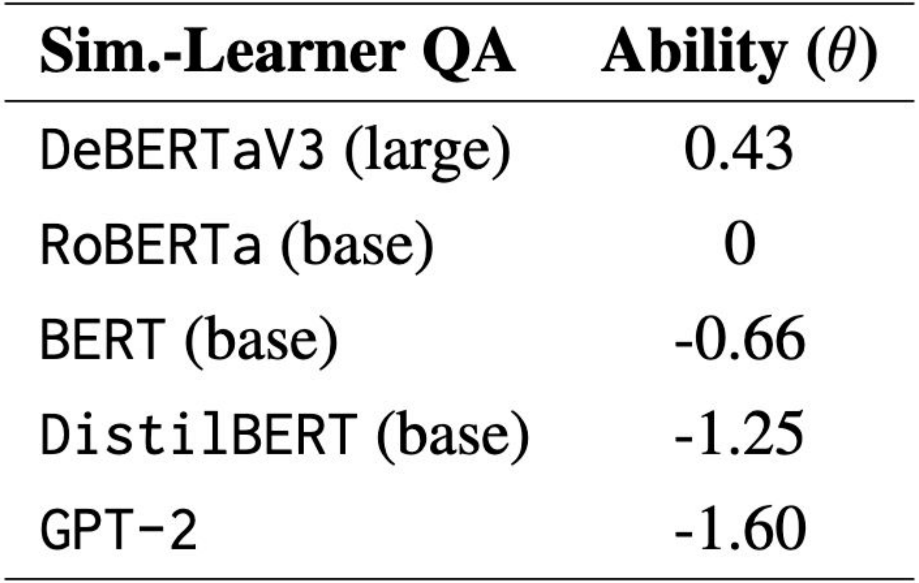
Decoder

<QUESTION> Q_1 Q_2 … Q_n <ANSWER> A_1 A_2 … A_n

Question Tokens

Answer Tokens

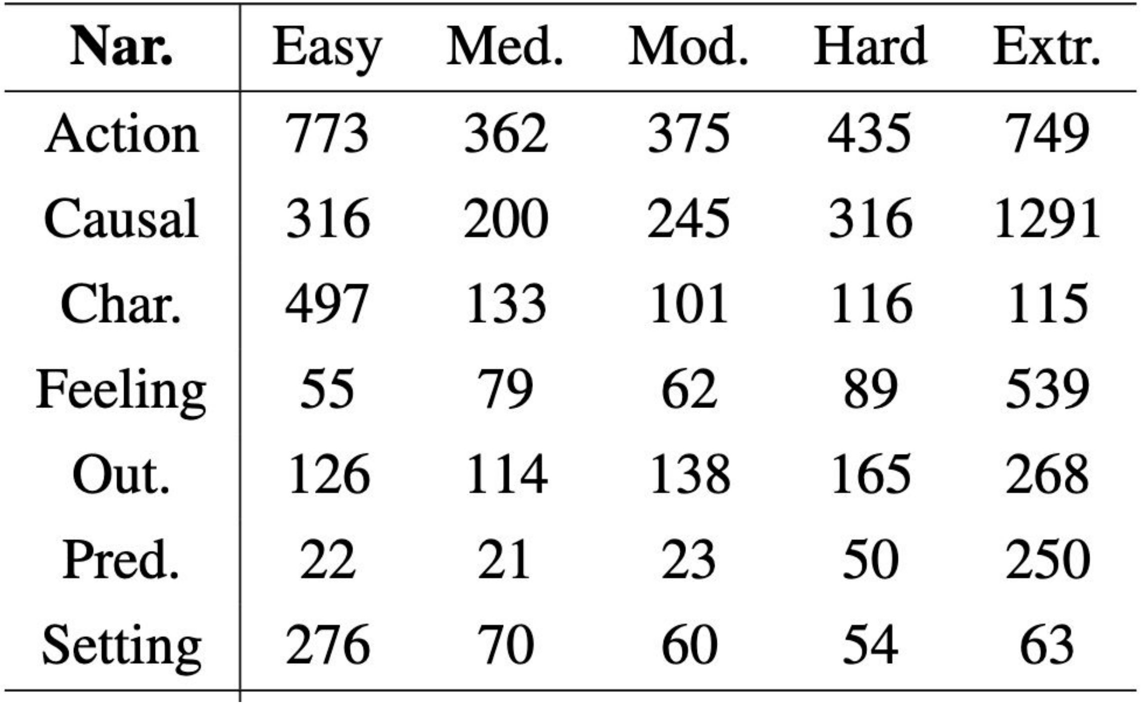
Simulated-Learners Estimated Ability



Sim.-Learner QA	**Ability (θ)**
DeBERTaV3 (large)	0.43
RoBERTa (base)	0
BERT (base)	-0.66
DistilBERT (base)	-1.25
GPT-2	-1.60

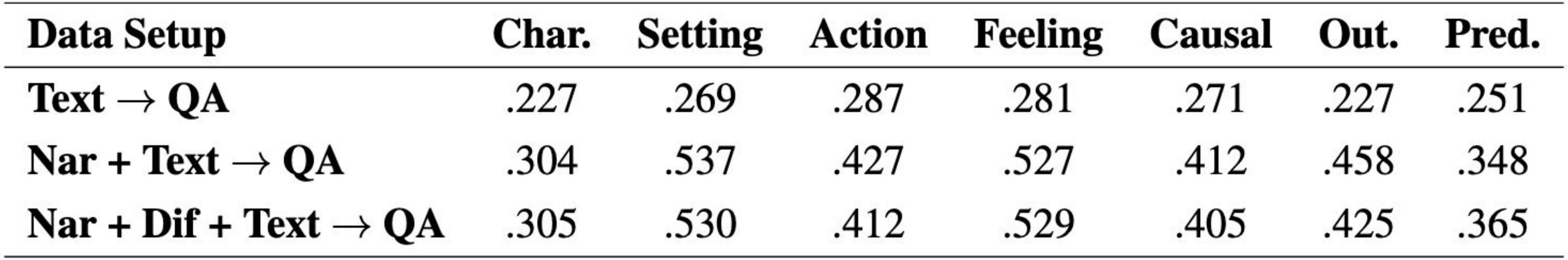
This study: *Advancing Question Generation with Joint Narrative and Difficulty Control*

Data Distribution



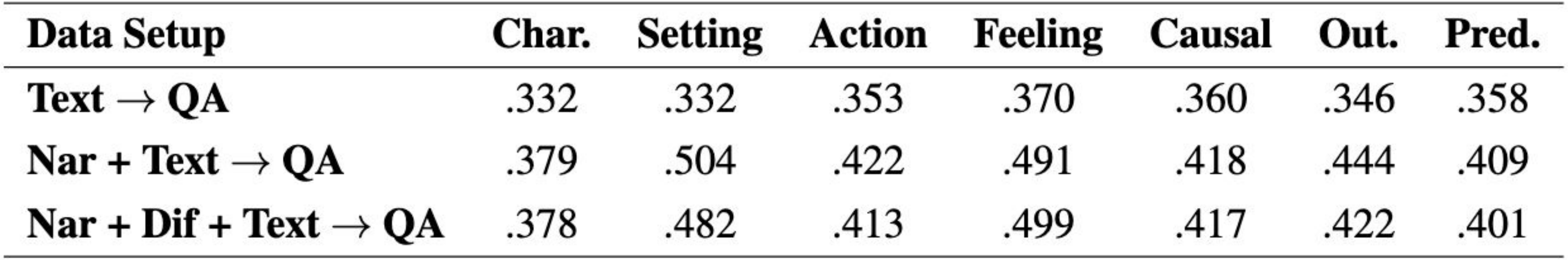
Nar.	Easy	Med.	Mod.	Hard	Extr.
Action	773	362	375	435	749
Causal	316	200	245	316	1291
Char.	497	133	101	116	115
Feeling	55	79	62	89	539
Out.	126	114	138	165	268
Pred.	22	21	23	50	250
Setting	276	70	60	54	63

Results for Narrative Control



Data Setup	Char.	Setting	Action	Feeling	Causal	Out.	Pred.
Text → QA	.227	.269	.287	.281	.271	.227	.251
Nar + Text → QA	.304	.537	.427	.527	.412	.458	.348
Nar + Dif + Text → QA	.305	.530	.412	.529	.405	.425	.365

Lexical similarity (ROUGE$_L$-F1) between generated and ground-truth questions

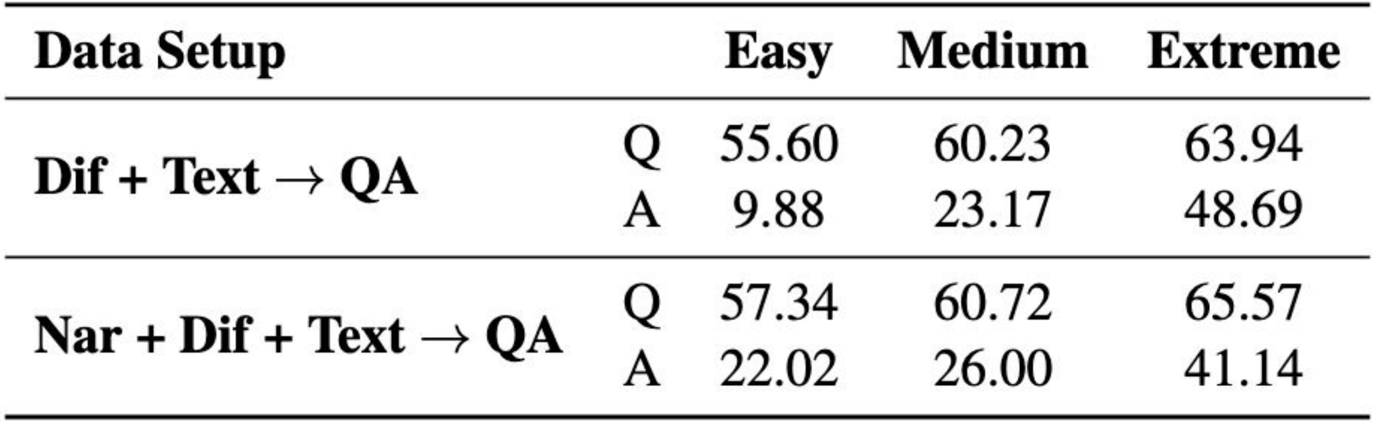


Data Setup	Char.	Setting	Action	Feeling	Causal	Out.	Pred.
Text → QA	.332	.332	.353	.370	.360	.346	.358
Nar + Text → QA	.379	.504	.422	.491	.418	.444	.409
Nar + Dif + Text → QA	.378	.482	.413	.499	.417	.422	.401

Semantic similarity (BLEURT) between generated and ground-truth questions

Linguistic Features Influenced By Control

The degree of **lexical novelty** between the generated question-answer pairs and the source text plays a key role.



Data Setup		Easy	Medium	Extreme
Dif + Text → QA	Q	55.60	60.23	63.94
	A	9.88	23.17	48.69
Nar + Dif + Text → QA	Q	57.34	60.72	65.57
	A	22.02	26.00	41.14

PINC (Chen and Dolan, 2011) values