On Few-Shot Prompting for Controllable Question-Answer Generation in Narrative Comprehension

Bernardo Leite, Henrique Lopes Cardoso
Faculty of Engineering - University of Porto, Portugal (FEUP)
Artificial Intelligence and Computer Science Laboratory (LIACC)
{bernardo.leite, hlc}@fe.up.pt

16th International Conference on Computer Supported Education
2-4 May, 2024, Angers, France

Agenda

- **Main Idea**
- Introduction
- Related Work
- Few-Shot Prompting for Controllable Question-Answer Generation
- Evaluation
- Error Analysis
- Conclusions & Future Directions

Controlling Question-Answer Generation (**Main Idea**)

"Once there were a hare and a turtle. The hare was proud of his speed. He asked the turtle to race..."

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

Controllable Gen. System

Character

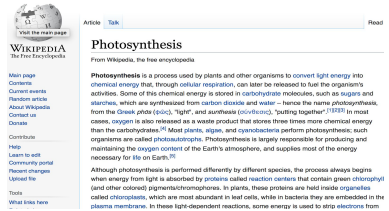
Who challenged the turtle to a race? The hare.

Introduction → Background

What is Question Generation (QG)?

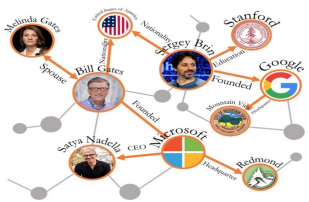
Context → question generator → Question

❑ Texts



Why do humans take O2 to produce CO2?

❑ Knowledge Graphs



Were Bill Gates and Satya Nadella once colleagues?

❑ Images

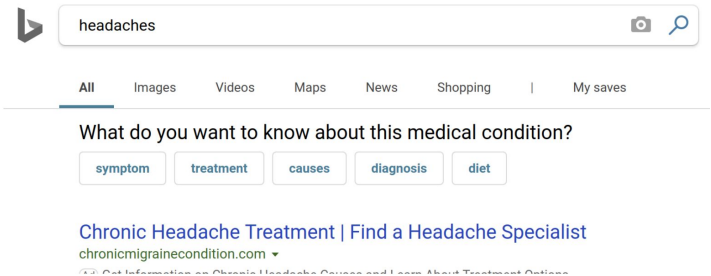
Who is having a birthday party?

based on slides: [Towards Generating Human-like Deep Questions](#) by Liangming Pan

Introduction → Background

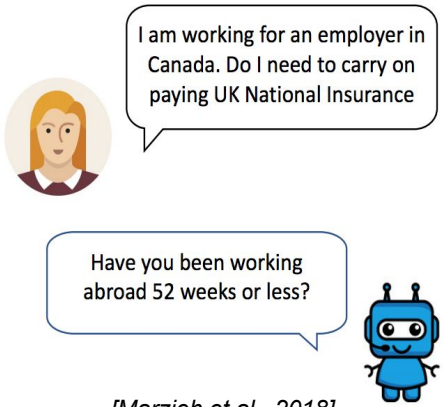
What are the main applications of Question Generation?

❑ Information Retrieval



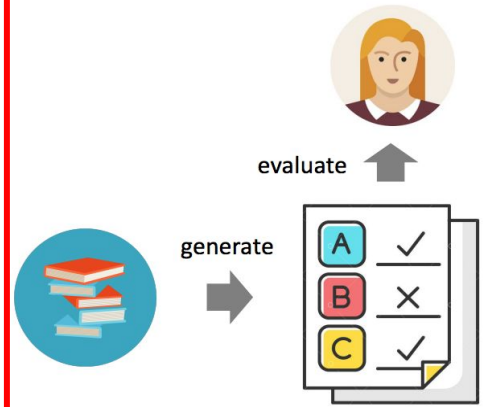
[Zamani et al., 2020]

❑ Dialogue Systems



[Marzieh et al., 2018]

❑ Education



[Rocha et al., 2017]

based on slides: Towards Generating Human-like Deep Questions by Liangming Pan

Introduction → Motivation

Advantages of Question Generation (QG) for Education

- Time-saving
- Resource augmentation

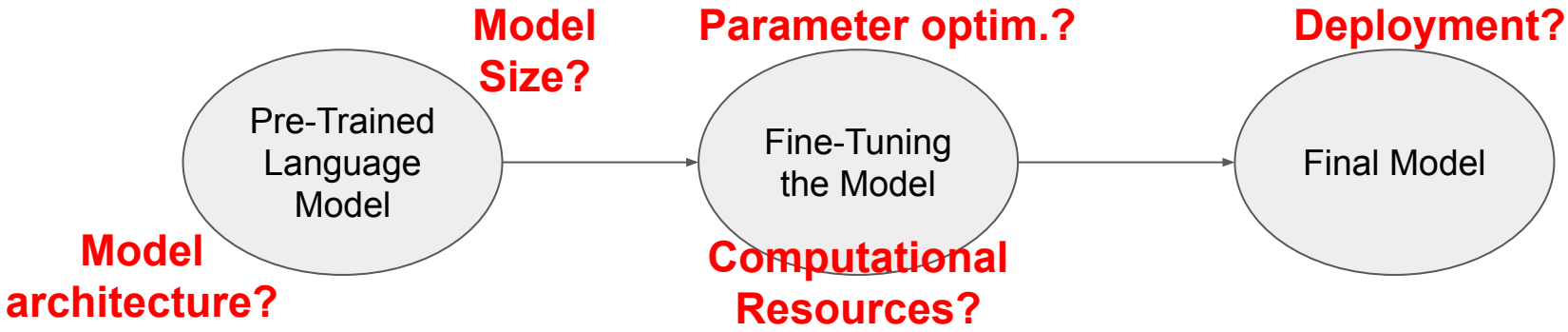
However… generated questions are generally limited in types and difficulty levels

[*Kurdi et al., 2020*] [*Wang et al., 2022*]

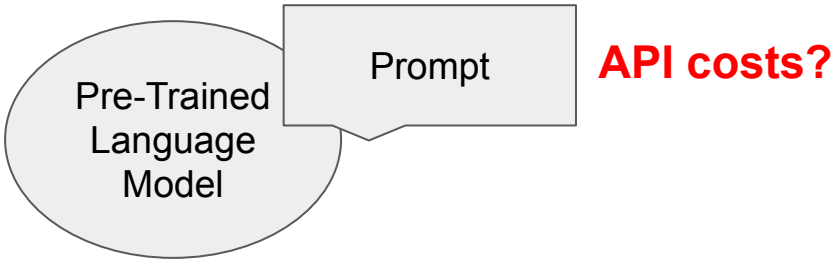
Strong desire for user control

Introduction → Motivation

Traditional Approach (Fine-Tuning)



Few-Shot Prompting



Introduction → Research Goals

Main Goal:

A feasibility analysis of a Few-Shot prompting strategy to address the Controllable Question-Answer Generation (**CQG**) task.

Main Contributions:

1. Few-Shot prompting strategy for CQG
   Generation of **both** questions and answers

2. Comparison with a reference fine-tuned model

3. Prompting and Error Analysis

Agenda

- Introduction
- **Related Work**
- Few-Shot Prompting for Controllable Question-Answer Generation
- Evaluation
- Error Analysis
- Conclusions & Future Directions

Related Work → Controllable Question-Answer Gen.

- **With Fine-Tuning**
  - T5/BART models with control labels [Ghanem et al., 2022] [Zhao et al., 2022]
- **With Few-Shot Prompting**
  - GPT-based models [Elkins et al., 2023]
- **Controllable Attributes**
  - Question Reading Comprehension Skills [Ghanem et al., 2022]
  - Question Bloom's Taxonomy [Elkins et al., 2023]
  - Question Narrative Elements [Zhao et al., 2022]
  - Question Explicitness [Leite et al., 2023]

Related Work → Controllable Question-Answer Gen.

Narrative Elements

- Character
- Setting
- Action
- Feeling
- Causal Relationship
- Outcome Resolution
- Prediction

Explicitness

- **Explicit** questions ask for answers that can be directly found in the stories
- **Implicit** questions rely on summarizing and drawing inferences from text

Data → FairytaleQA dataset

"Once there were a hare and a turtle. The hare was proud of his speed. He asked the turtle to race (...) The hare ran very fast, and the turtle was left behind. The hare thought he should take some rest…

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

Simplistic Example of dataset:

FairytaleQA
(Xu et al., 2022)

Question	Answer	Explicitness Label	Narrative Label
Who challenged the turtle to a race?	The hare.	Explicit	Character
(...)	(...)	(...)	(...)
Why did the hare decide to take some rest?	The turtle was far behind.	Implicit	Causal Relation

Agenda

- Introduction
- Related Work
- **Few-Shot Prompting for Controllable QA Gen.**
- Evaluation
- Error Analysis
- Conclusions & Future Directions

Proposed Strategy → Few-Shot Prompting for CQG

Prompt (query)

Generate questions and answers targeting the following narrative element:
causal relationship

Prompt (examples)

Text: Sarah found a lost kitten on the street and decided to take it home…
Question: Why did Sarah decide to take the kitten home?
Answer: The kitten was lost.
(...)
Text: Jack saw a friendly group of kids playing in the park, so he decided to join them…
Question: Why did Jack decide to join the group of kids playing in the park?
Answer: The group of kids was friendly.

Text: The little girl opened the door because she was curious about the room's contents…
Question:

⇩

GPT-3.5

⇩

Generated QA Pair

Why did the little girl open the door?
Answer: The little girl was curious about the room.

Procedure for Evaluating Question **Narrative Control**

Prompt (query)

Generate … targeting the following narrative element: <NAR>

Prompt (examples)

Text: {Some text…}
Question: {Some question…}
Answer: {Some answer}
(...)

Text: {Some text…}
Question:

⇩

GPT-3.5

⇩

Generated QA Pair

Why did the little girl open the door?
Answer: The little girl was curious about the room.

compare similarity

Metrics used: $ROUGE_L$-F1, BLEURT

Human-Annotated QA pair

Why was the little girl curious?
Answer: The little girl was curious because…

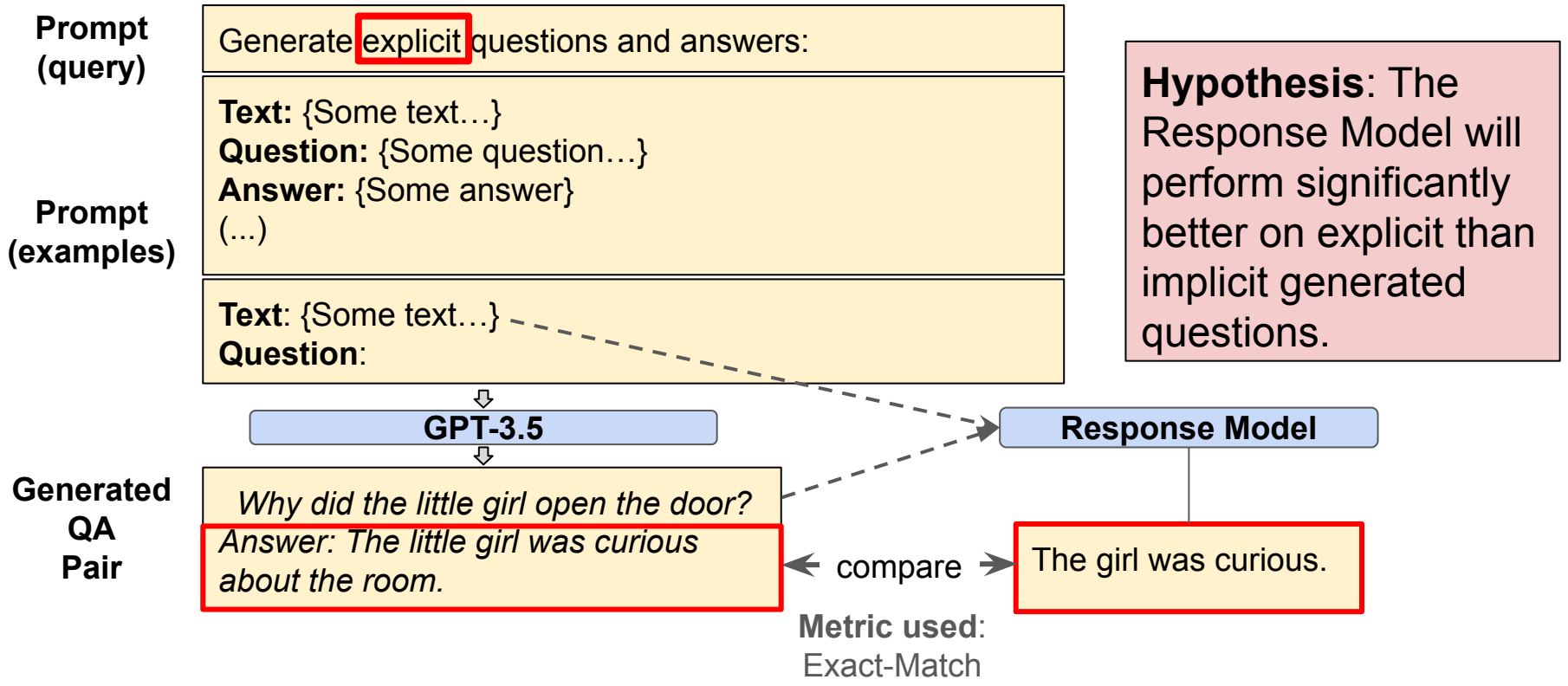
Hypothesis: Generated questions will be closer to the human-annotated questions when control attributes are incorporated.

Results for Question **Narrative Control**

	Narrative Control		ROUGE$_L$-F1	BLEURT
Fine-Tuning (reference)	No		0.335	0.394
	Yes	*improves*	0.429	0.438
Few-Shot Prompting	No		0.339	0.397
	Yes	*improves*	0.409	0.445

Improved results when **Narrative Control is Enabled**.

Procedure for Evaluating Question **Explicitness**

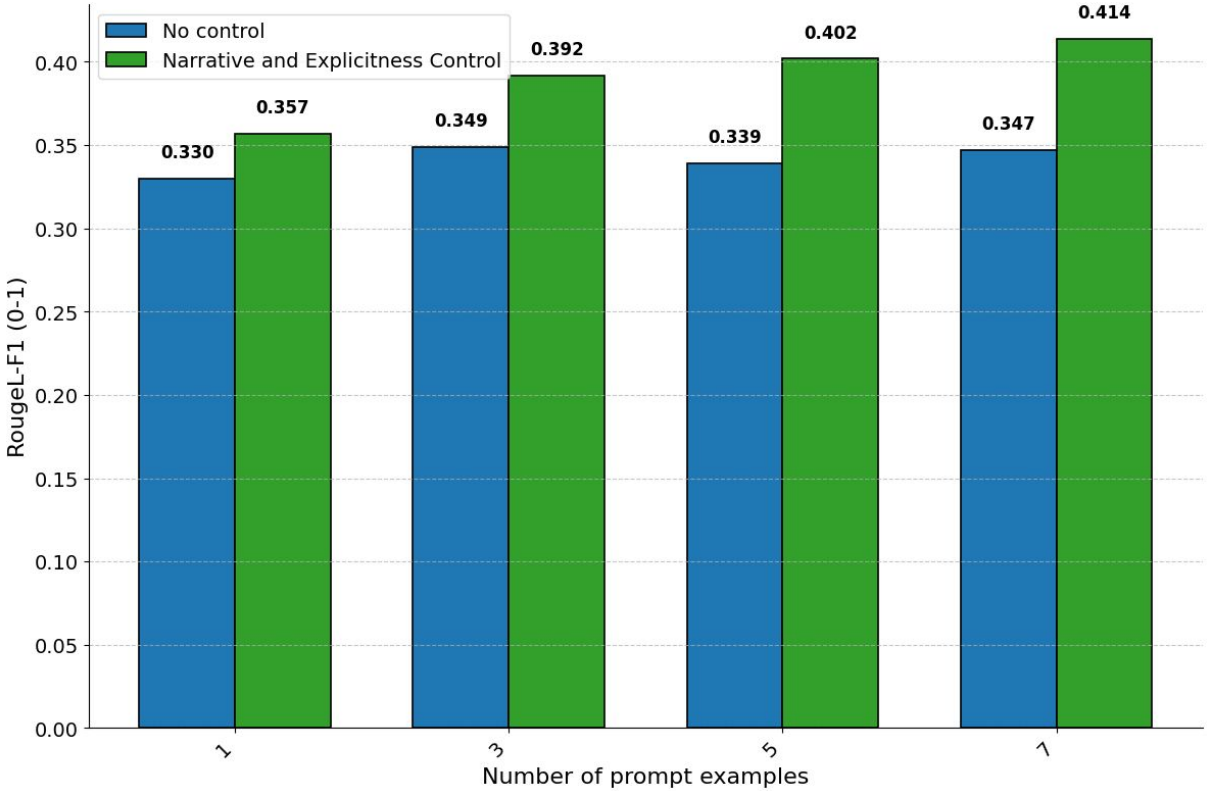


Results for Question **Explicitness Control**

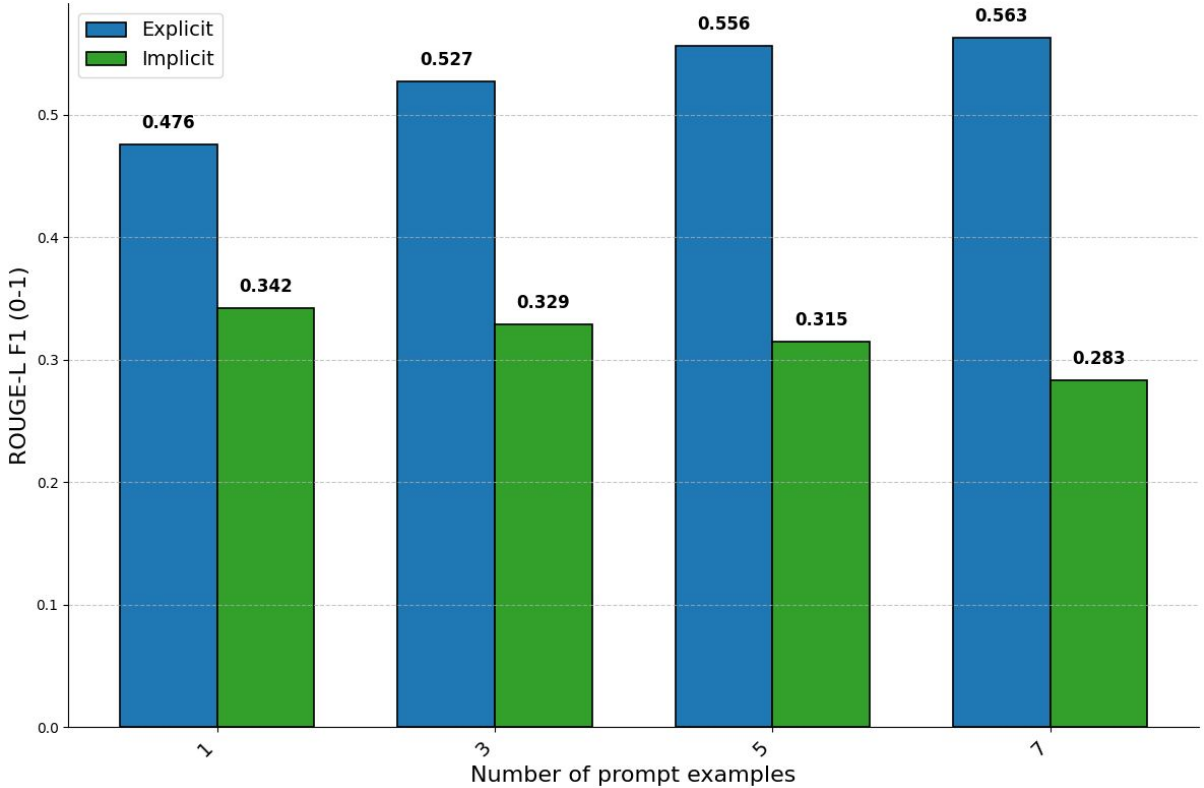
	$\mathbf{ROUGE_L}$-F1		
	Overall	Explicit	Implicit
Fine-Tuning (reference)	0.517	0.580	0.352 ↓ *decreases*
Few-Shot Prompting	0.754	0.785	0.673 ↓ *decreases*

Response Model performs **better on explicit** than implicit generated questions (confirms the hypothesis).

Varying the Nr. of Prompt Examples for **Narrative** Control



Varying the Nr. of Prompt Examples for **Explicitness** Control



Eval. → Linguistic Quality of Gen. Questions and Answers

		Generated Questions			Generated Answers		
	Control	**PPL ↓**	**Dist-3 ↑**	**Gram. ↓**	**PPL ↓**	**Dist-3 ↑**	**Gram. ↓**
Fine-Tuning (reference)	No	197.192	0.776	0.013	303.331	0.668	0.033
	Explicitness	175.717	0.789	**0.005**	336.649	0.662	0.028
	Narrative	168.303	0.782	0.018	343.050	0.597	**0.020**
	Narrative + Explicitness	183.665	0.789	0.013	352.672	0.560	0.025
Few-Shot Prompting	No	166.160	0.787	**0.005**	248.966	0.725	0.038
	Explicitness	**143.270**	0.791	0.013	240.593	**0.734**	0.036
	Narrative	155.761	**0.797**	0.008	**224.536**	0.679	**0.020**
	Narrative + Explicitness	153.056	0.790	0.010	260.307	0.671	0.033

Agenda

- Introduction
- Related Work
- Few-Shot Prompting for Controllable Question-Answer Generation
- Evaluation
- **Error Analysis**
- Conclusions & Future Directions

Error Analysis

Analysed 105 generated QA pairs, via few-shot prompting (GPT 3.5) with narrative control.

We identify 3 types of issues:

1. Narrative Misalignment
2. Generic Questions
3. Lengthy Answers

Issue 1: Narrative Misalignment (found 19 in 105)

Example:

"There was once a fisherman who was called Salmon (...). He had a wife called Maie; (...) In winter they dwelt in a little cottage by the shore (...). The cottage on the rock was even smaller than the other; it had a wooden bolt instead of an iron lock to the door, a stone hearth, a flagstaff, and a weather-cock on the roof…"

Target Narrative: CHARACTER

Question: What did the cottage on the rock have?

Answer: A wooden bolt instead of an iron lock to the door, a stone hearth, a flagstaff, and a weather-cock on the roof.

The question is not related to any character.

Issue 2: QA Pair Ambiguity (found 1 in 105)

Example:

> *"I am going to tell you a story about a poor young widow woman, who lived in a house called Kittlerumpit, though whereabouts in Scotland the house of Kittlerumpit stood nobody knows. Some folk think that it stood in the neighbourhood of the Debateable Land…"*

Target Narrative: SETTING

Question: Whereabouts in Scotland was Kittlerumpit located?
Answer: Nobody knows.

Too vague.

Issues 2 and 3: Generic Questions & Long Answers (found 5 in 105)

"Ahti", said they, "is a mighty king who lives in his dominion of Ahtola, and has a rock at the bottom of the sea, and possesses besides a treasury of good things. He rules over all fish and animals of the deep; he has the finest cows and the swiftest horses that ever chewed grass at the bottom of the ocean."...

Target Narrative: CHARACTER

Question: Who is ahti? **Generic.**

Very long answer.

Answer: Ahti is a mighty king who lives in his dominion of Ahtola, and has a rock at the bottom of the sea, and possesses besides a treasury of good things. He rules over all fish and animals of the deep; he has the finest cows and the swiftest horses that ever chewed grass at the bottom of the ocean.

Conclusions

- Evaluation indicates Narrative & Explicitness control, however…
  error analysis revealed instances with failed control

- Best controllability when prompt examples increase

- Few-Shot strategy can outperform the reference model in certain scenarios

When is it worth to use the Few-Shot strategy for controlling QA generation?

  - Data availability is limited
  - One favors for a “plug-and-play” AI-assisted approach

Future Directions 🔭

- Prompt **Design**
- Verify QA Pair **Alignment**
- Controllability **+ Adaptivity**

On Few-Shot Prompting for Controllable Question-Answer Generation in Narrative Comprehension

Bernardo Leite, Henrique Lopes Cardoso
Faculty of Engineering - U. Porto, Portugal (FEUP)
Artificial Intelligence and Computer Science Laboratory (LIACC)
{bernardo.leite, hlc}@fe.up.pt

16th International Conference on Computer Supported Education
2-4 May, 2024, Angers, France

Appendix

Experimental Setup → Data Preparation

From the original dataset, we have prepared 4 different prompting setups:

- **No Control (baseline)**:
  "Generate questions and answers from text:"
- **Narrative Control**:
  "Generate questions and answers targeting the following narrative element: <NAR>."
- **Explicitness Control**:
  "Generate <EX> questions and answers:"
- **Narrative + Explicitness Control**:
  "Generate <EX> questions and answers targeting the following narrative element: <NAR>:"

Response Model Results for Question **Explicitness Control**

Response Model is fine-tuned

	$\mathbf{ROUGE_L}$-F1		
	Overall	Explicit	Implicit
Fine-Tuning (reference)	0.661	0.716	0.513
Few-Shot Prompting	0.481	0.531	0.351

Response Model is GPT-3.5

	$\mathbf{ROUGE_L}$-F1		
	Overall	Explicit	Implicit
Fine-Tuning (reference)	0.517	0.580	0.352
Few-Shot Prompting	0.754	0.785	0.673

Results for All Controllable Settings

	Data Setups	ROUGEL-F1 ↑	BLEU-4 ↑	BLEURT ↑
Reference Model	section → question + answer	0.335	0.137	0.394
	ex + section → question + answer	0.333	0.138	0.398
	nar + section → question + answer	0.429	**0.201**	0.438
	nar + ex + section → question + answer	**0.442**	0.198	0.442
Few-Shot Prompting	section → question + answer	0.339	0.108	0.397
	ex + section → question + answer	0.358	0.123	0.411
	nar + section → question + answer	0.409	0.168	**0.445**
	nar + ex + section → question + answer	0.402	0.177	0.441

Results for All Controllable Settings

		ROUGEL-F1 ↑			EXACT-MATCH ↑		
	Data Setups	Overall	Explicit	Implicit	Overall	Explicit	Implicit
Reference Model	ex + section → question + answer	**0.661**	**0.716**	**0.513**	0.371	0.413	**0.259**
	nar + ex + section → question + answer	0.628	0.681	0.487	**0.383**	**0.434**	0.250
Few-Shot Prompting	ex + section → question + answer	0.481	0.531	0.351	0.119	0.143	0.056
	nar + ex + section → question + answer	0.490	0.556	0.315	0.155	0.185	0.074

Results for All Controllable Settings (Per Nar. Element)

		Narrative Elements						
	Data Setups	Chara.	Setting	Action	Feeling	Causal	Out.	Pred.
Reference Model	section → question + answer	0.320	0.279	0.372	0.300	0.381	0.273	0.240
	nar + section → question + answer	**0.360**	0.550	0.461	0.517	0.409	0.374	0.379
	nar + ex + section → question + answer	0.350	**0.615**	0.461	**0.568**	**0.419**	**0.447**	**0.450**
Few-Shot Prompting	section → question + answer	0.254	0.307	0.449	0.305	0.303	0.324	0.300
	nar + section → question + answer	0.277	0.380	0.496	0.532	0.377	0.387	0.335
	nar + ex + section → question + answer	0.296	0.365	**0.498**	0.516	0.367	0.337	0.327