REIGN: Robust Training for Conversational Question Answering Models using **REI**nforced Reformulation **G**eneratio**N**

Magdalena Kaiser, Rishiraj Saha Roy and Gerhard Weikum
Max Planck Institute for Informatics, Saarbrücken, Germany

Conversational Question Answering (ConvQA)

Consider context

- Sequential, multi-turn QA
- Incomplete follow-up questions
- Challenges:
  - Implicit context
  - Ad hoc formulations

Q1: What's the 2022 LOTR TV series called?

A1: The Rings of Power (TROP)

Q2: TROP airing on?

A2: Netflix

Q3: Which actor plays Isildur in the series?

A3: Harry Sinclair

Conversational Question Answering (ConvQA)

Consider diverse formulations

- Common solution: Data augmentation

Q1: What's the 2022 LOTR TV series called?

A1: The Rings of Power (TROP)

Q2: TROP airing on?

Q21: Which **streaming service showed** TROP?

Q22: TROP available on **which network**?

Q23: On **which platform** is **the Rings of Power** airing?

Q24: **Rings of Power broadcasted** where?

Q25: Where can I **stream** the **LOTR TV series**?

Conversational Question Answering (ConvQA)

Consider diverse formulations

- Common solution: Data augmentation
- Drawbacks with classical data augmentation:
    - not model-specific
    - can be inefficient
    - challenging for ConvQA

Q1: What's the 2022 LOTR TV series called?

A1: The Rings of Power (TROP)

Q2: TROP airing on?

Q21: Which **streaming service showed** TROP?

Q22: TROP available on **which network**?

Q23: On **which platform** is **the Rings of Power** airing?

Q24: **Rings of Power broadcasted** where?

Q25: Where can I **stream** the **LOTR TV series**?

Conversational Question Answering (ConvQA)

Consider diverse formulations

- Common solution: Data augmentation
- Drawbacks with classical data augmentation:
  - not model-specific
  - can be inefficient
  - challenging for ConvQA

➡ Only select **subset of reformulations most helpful** for specific model

Q1: What's the 2022 LOTR TV series called?

A1: The Rings of Power (TROP)

~~Q2: TROP airing on?~~

Q21: Which **streaming service showed** TROP?

~~Q22: TROP available on **which network**?~~

Q23: On **which platform** is **the Rings of Power** airing?

Q24: **Rings of Power broadcasted** where?

~~Q25: Where can I **stream** the **LOTR TV series**?~~

Conversational Question Answering (ConvQA)

Consider diverse formulations

<u>Goal:</u> **Train** a more **robust** ConvQA model using a **model-specific set** of **reformulations**

Q1: What's the 2022 LOTR TV series called?

A1: The Rings of Power (TROP)

Q2: TROP airing on?

A2: Amazon Prime Video

Q3: Which actor plays Isildur in the series?

A3: Harry Sinclair

Contributions

Towards robust training and evaluation of ConvQA models

- **Taxonomy of question reformulations** for ConvQA over KGs based on string-edit distance
- RL model with **Deep Q-Network** to select helpful **reformulations guided** towards better QA performance
- About **335k** question **reformulations of test cases** in two ConvQA benchmarks
- REIGN framework with **reusable components** to judiciously augment benchmark training tailored to specific ConvQA models

The REIGN pipeline

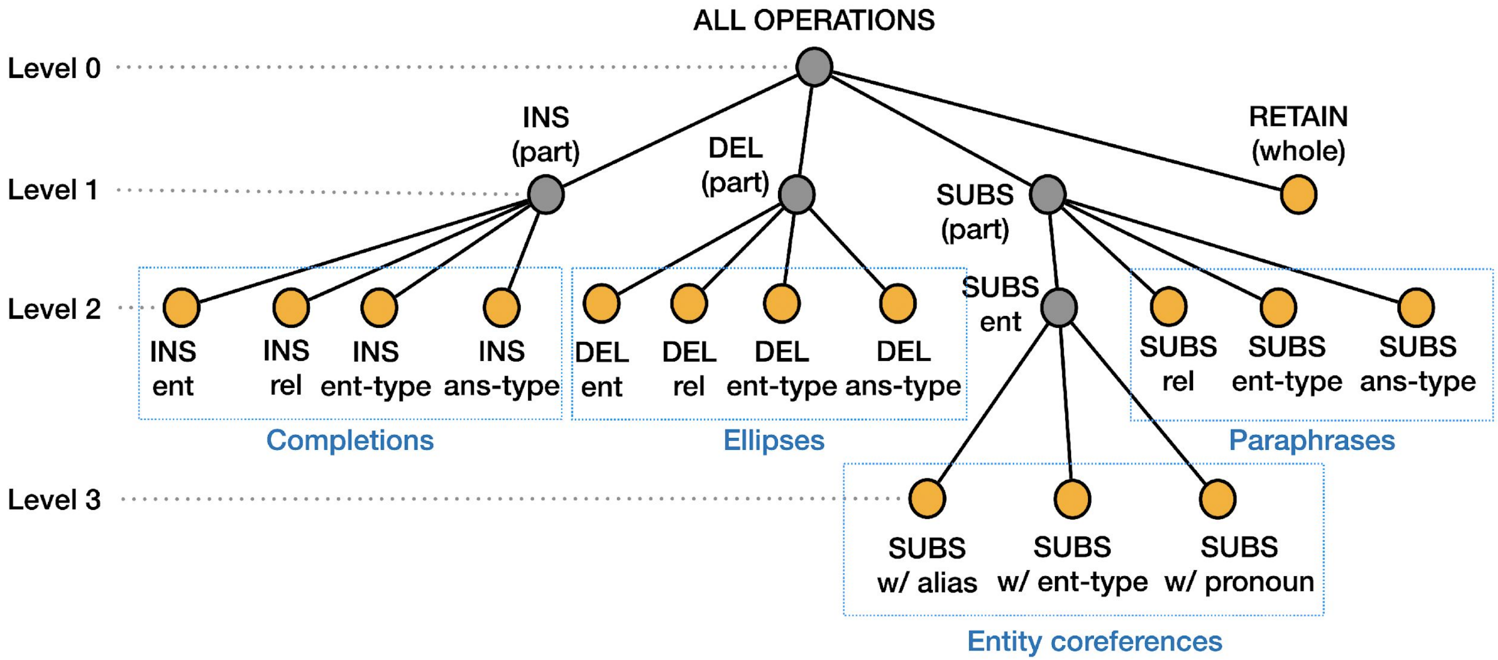
Start with (Q, A) pair from benchmark

Conversation Question 2: TROP airing on? [Gold answer: Amazon Prime Video]

The REIGN pipeline

Reformulation taxonomy

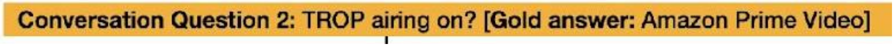
Taxonomy of ConvQA
Reformulation Categories



The REIGN pipeline

The core: Reformulation Category Selector

Taxonomy of ConvQA Reformulation Categories



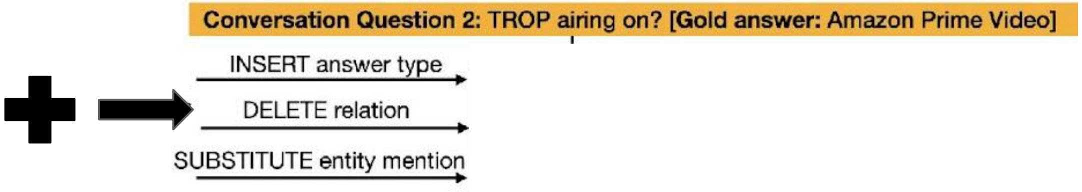
+

Reformulation Category Selector (RCS) with reinforcement learning (Deep Q-Network)

The REIGN pipeline

The core: Reformulation Category Selector

Taxonomy of ConvQA Reformulation Categories

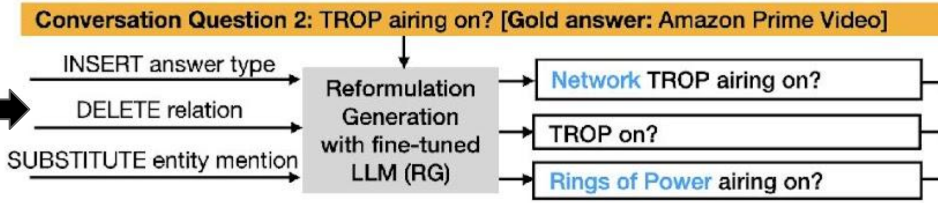


Reformulation Category Selector (RCS) with reinforcement learning (Deep Q-Network)

The REIGN pipeline

Reformulation generator creates reformulations

Taxonomy of ConvQA Reformulation Categories



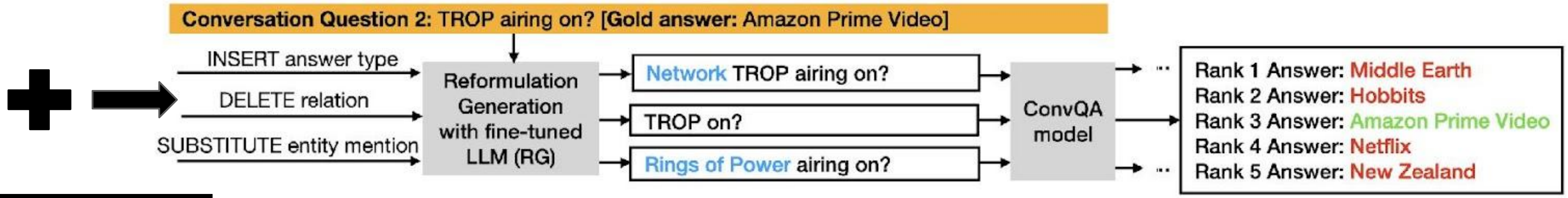
Reformulation Category Selector (RCS) with reinforcement learning (Deep Q-Network)

The REIGN pipeline

Pass reformulations through ConvQA model …

System responses

Taxonomy of ConvQA Reformulation Categories

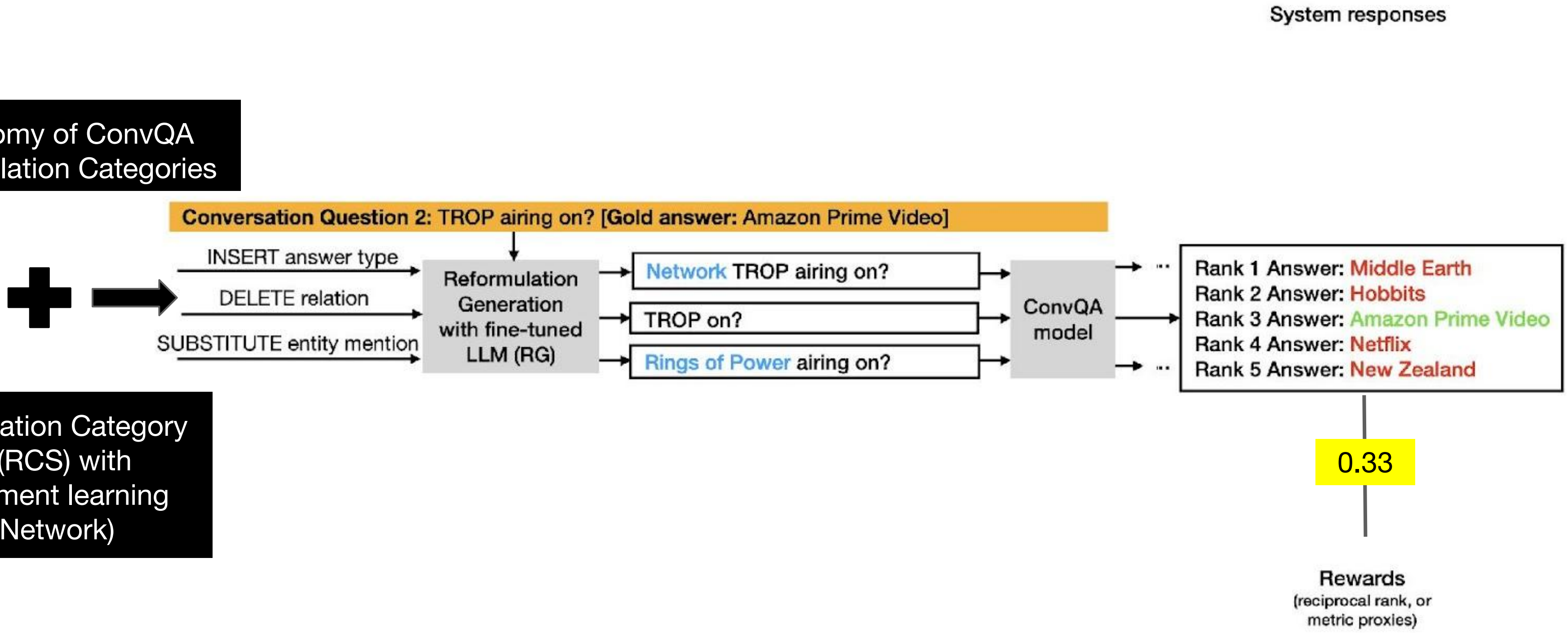


Reformulation Category Selector (RCS) with reinforcement learning (Deep Q-Network)

The REIGN pipeline

… to collect rewards …

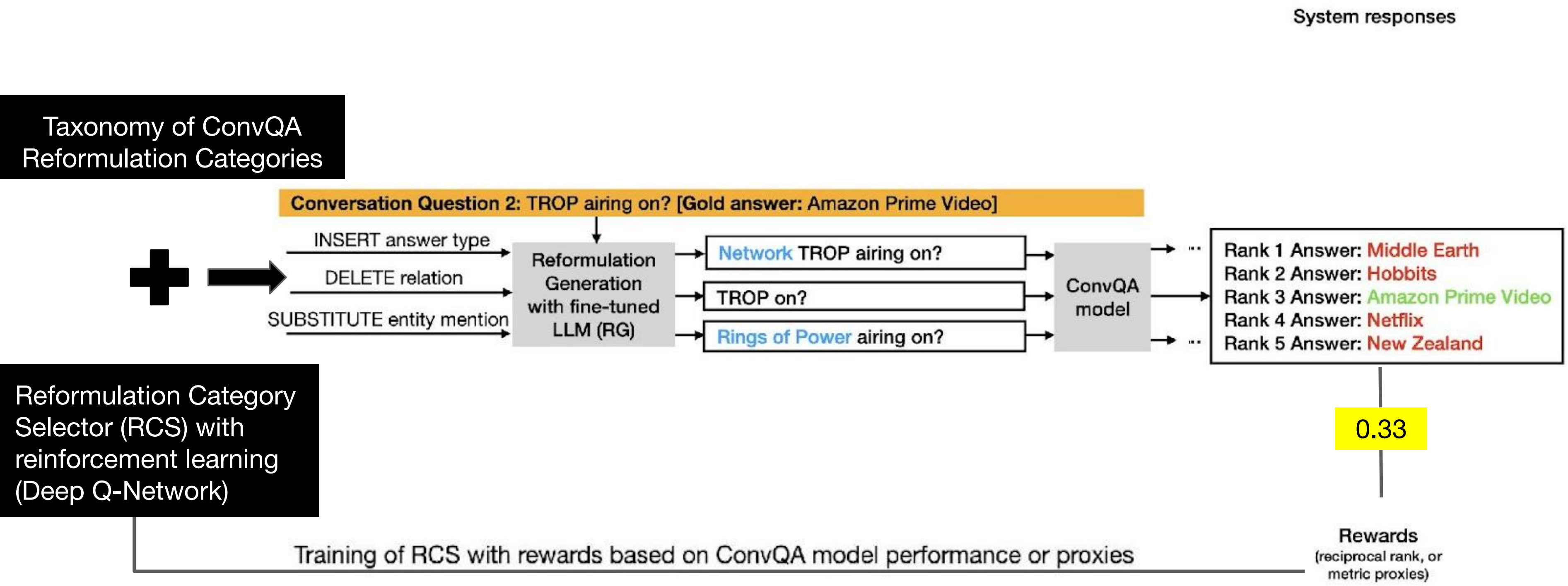
Taxonomy of ConvQA Reformulation Categories



Reformulation Category Selector (RCS) with reinforcement learning (Deep Q-Network)

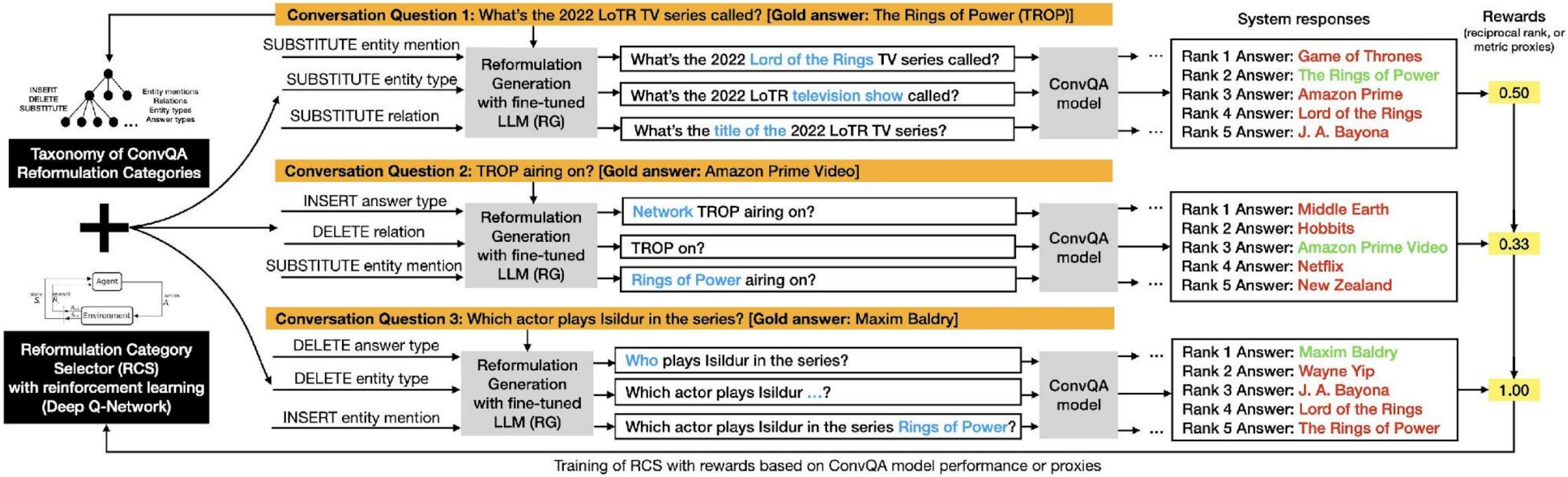
The REIGN pipeline

… to train the RCS



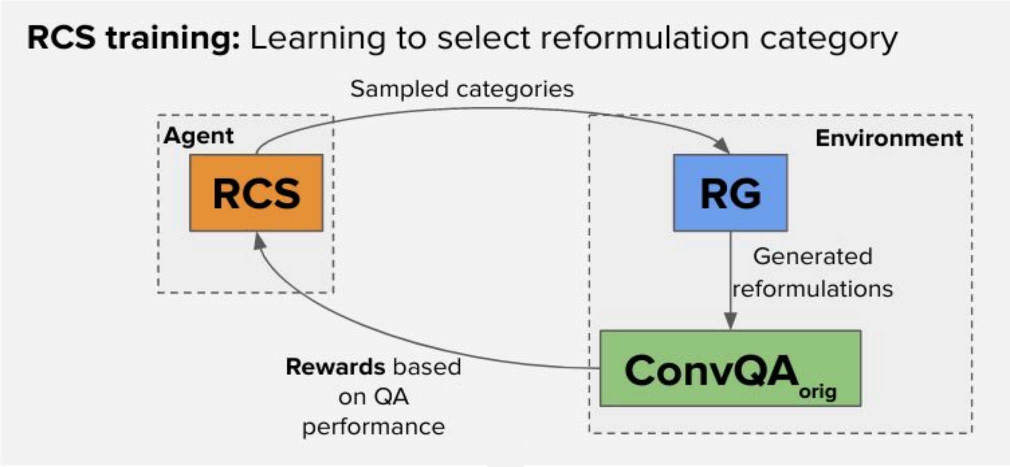
The REIGN pipeline

Repeat for all questions



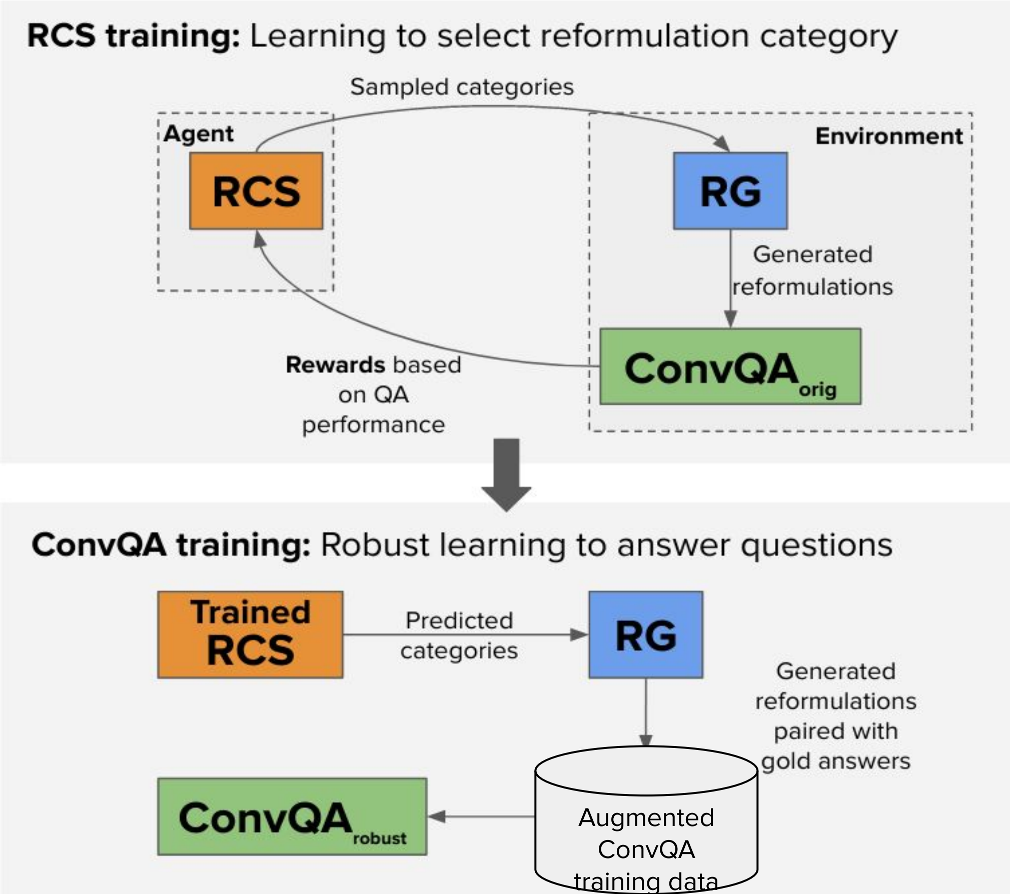
Components in REIGN

Two-step training



Components in REIGN

Two-step training



Large-scale Evaluation

Increasing robustness at inference time

- Small test sets not enough
- Reformulate questions with GPT-3.5-turbo
  - 10x with conversation history
  - 10x without conversation history
- **100k-200k questions** in total

Experimental Setup

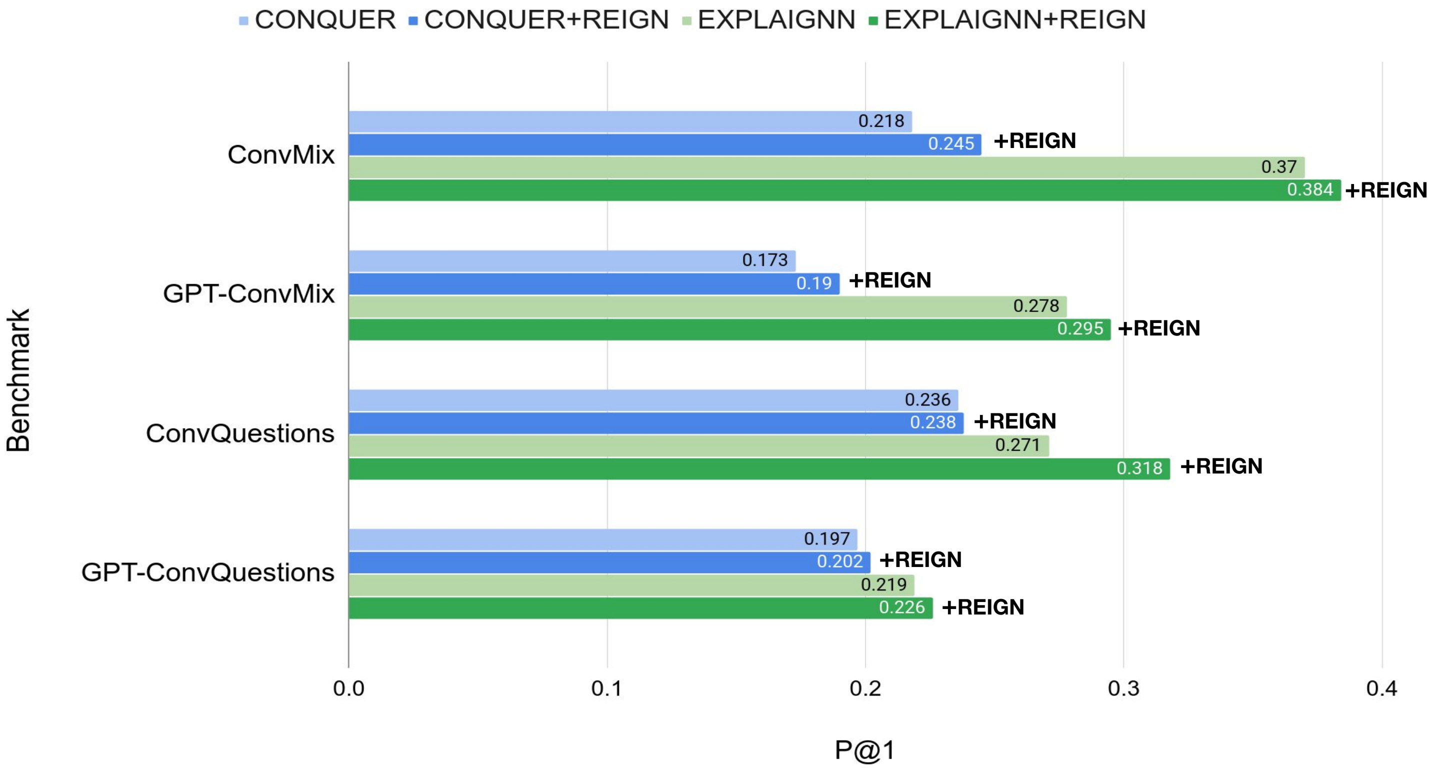
REIGN coupled with ConvQA models

- REIGN applied to **two ConvQA models**: *CONQUER, EXPLAIGNN;*
- REIGN applied on **two benchmarks**: *ConvQuestions, ConvMix*
- Results on original testsets and 20x larger GPT-augmented testsets (indicated with GPT-ConvMix / GPT-ConvQuestions)

Results

Improves performance of underlying ConvQA model

Models coupled with REIGN are able to answer more questions correctly

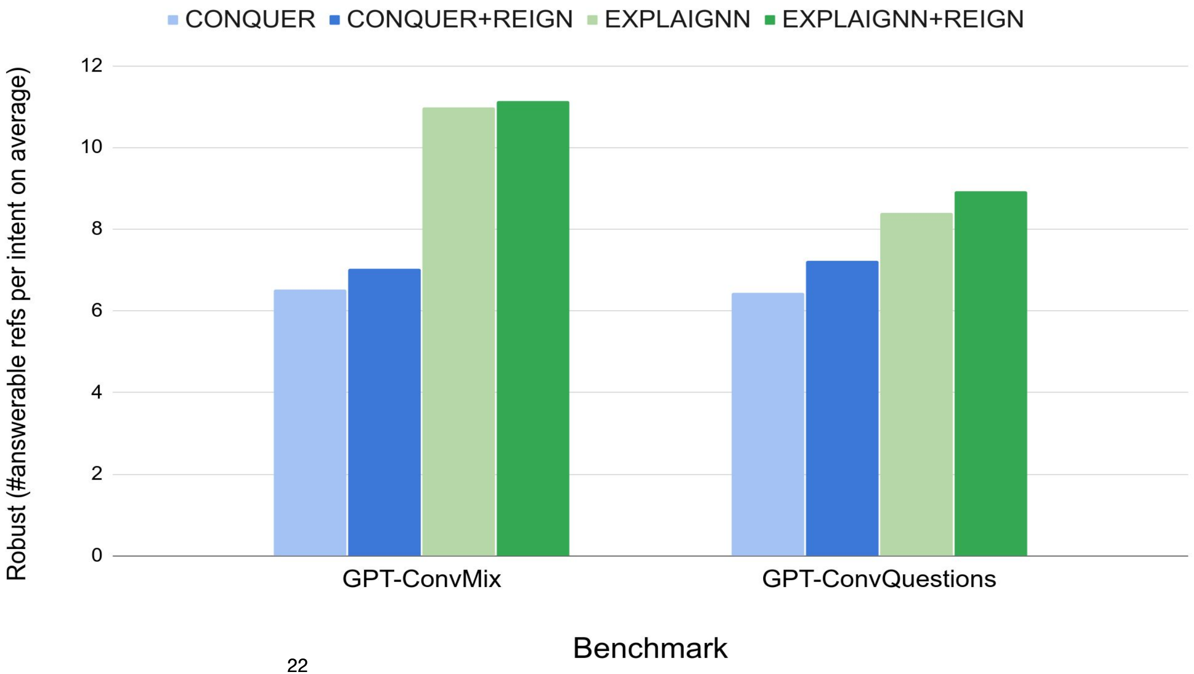


Results

Improves robustness to different surface forms

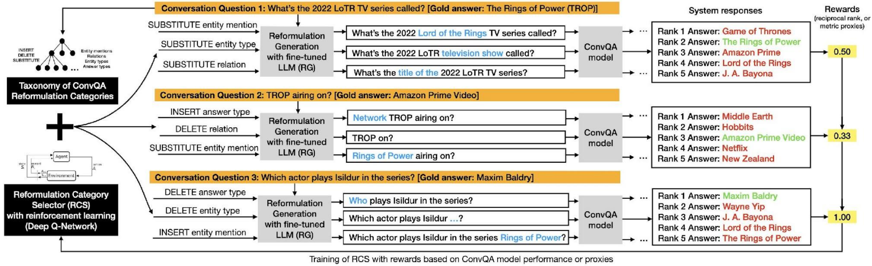
New metric *Robust*: average of #answerable reformulations per original test question (0-21)

Models coupled with REIGN are able to answer more reformulations per question intent correctly



REIGN: Wrap-up

Takeaways



- **Improved** ConvQA models by **training** with **reformulations**
- Reformulations **generated at scale** in **systematic way** by **reformulation taxonomy**
- More **robust** and **efficient training** by selecting **set** of most **helpful reformulations** for underlying model
- **Enlarged test set** generated with LLM for model **stress-testing**

reign.mpi-inf.mpg.de

Thank you!

Backup slides

REIGN

Detailed results: Main results, domain-wise, turn-wise

Benchmark →	ConvMix [14] Test			GPT-ConvMix Test				ConvQuestions [12] Test			GPT-ConvQuestions Test			
Method ↓	**P@1**	**MRR**	**Hit@5**	**P@1**	**MRR**	**Hit@5**	**Robust**	**P@1**	**MRR**	**Hit@5**	**P@1**	**MRR**	**Hit@5**	**Robust**
Conquer [35]	0.218	0.272	0.337	0.173	0.224	0.287	6.531	0.236	0.287	0.360	0.197	0.245	0.304	6.447
Conquer [35] + Reign	**0.245***	**0.292***	**0.346***	**0.190***	**0.236***	**0.289***	**7.035***	**0.238**	**0.290***	**0.371***	**0.202***	**0.252***	**0.310***	**7.224***
ExplaiGNN [15]	0.370	0.438	0.526	0.278	0.346	0.433	10.983	0.271	0.355	0.466	0.219	0.290	0.382	8.400
ExplaiGNN [15] + Reign	**0.384***	**0.446***	**0.531**	**0.295***	**0.361***	**0.448***	**11.130***	**0.318***	**0.411***	**0.529***	**0.226***	**0.302***	**0.402***	**8.925***

Table 5: Main results comparing Reign-enhanced ConvQA models with their standalone versions. GPT-augmented test sets are 20x original sizes. Reign is applied zero-shot on ConvQuestions. The higher value per column per QA model is in bold.

Method ↓ / Domain →	**Books**	**Movies**	**Music**	**TV series**	**Soccer**
Conquer [35]	0.227	0.175	0.159	0.141	0.163
Conquer [35] + Reign	**0.239***	**0.200***	**0.167***	**0.160***	**0.184***
ExplaiGNN [15]	0.298	**0.287***	0.265	0.274	0.265
ExplaiGNN [15] + Reign	**0.333***	0.283	**0.301***	**0.281***	**0.275***

Table 6: Domain-wise P@1 results on GPT-ConvMix testset.

Method ↓ / Turn →	**1**	**2**	**3**	**4**	**5**	**6-10**
Conquer [35]	0.205	0.193	0.177	0.184	0.160	0.133
Conquer [35] + Reign	**0.210***	**0.214***	**0.194***	**0.204***	**0.184***	**0.147***
ExplaiGNN [15]	0.333	0.297	0.286	0.292	0.277	0.205
ExplaiGNN [15] + Reign	**0.350***	**0.318***	**0.311***	**0.305***	**0.291***	**0.216***

Table 7: Turn-wise P@1 results on GPT-ConvMix testset.

REIGN

Detailed results: Category predictions, design choices

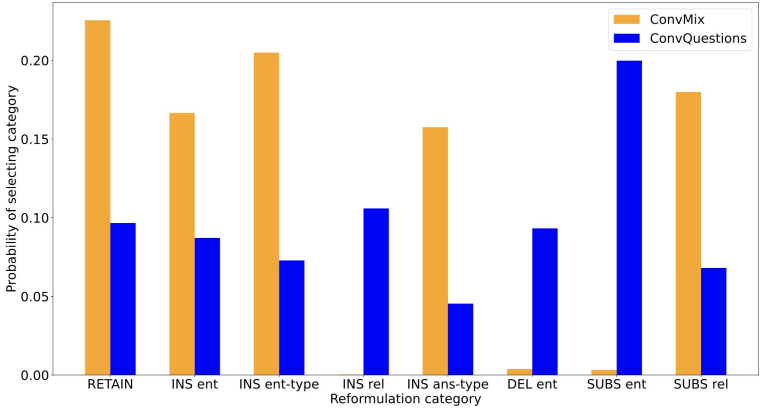


Figure 4: Common category predictions by the RCS DQN.

Row	Configuration	P@1	MRR	Hit@5	#Data
1	RCS (DQN, top-5) + RG (BART) [Full]	**0.190**	**0.236**	0.289	43.6k
2	RCS (DQN, top-3) + RG (BART)	0.184	0.231	0.288	30.5k
3	RCS (DQN, top-1) + RG (BART)	0.178	0.228	0.288	15.9k
4	No RCS (All cats) + RG (BART)	0.188	0.234	0.292	126k
5	No RCS (Random cats) + RG (BART)	0.182	0.232	**0.293**	42k
6	No RCS (Sample cats) + RG (BART)	0.185	0.231	0.287	41.9k
7	No RCS (INS part) + RG (BART)	0.183	0.230	0.288	42k
8	No RCS (DEL part) + RG (BART)	0.172	0.218	0.273	42k
9	No RCS (SUBS part) + RG (BART)	0.183	0.228	0.282	58.8k
10	No RCS + No RG (Question completion)	0.175	0.224	0.284	15.1k
11	No RCS + No RG (Question rewriting)	0.180	0.230	0.291	15.1k

Table 8: Large-scale effects of design choices in REIGN (with CONQUER on GPT-CONVMIX, all differences systematic).

REIGN

Detailed results: GPT test sets, prompts

Benchmark	Train	Dev	Test	GPT-Test
ConvMix [14]	8.4k (1680)	2.8k (560)	4.8k (760)	100.8k (760)
ConvQuestions [12]	33.6k (6720)	11.2k (2240)	11.2k (2240)	235.2k (2240)

Table 2: Benchmark sizes as #questions (#conversations). Reformulations are also counted as individual questions to be answered. Questions for the GPT-Test sets subsume the original test questions.

Reformulate the 'Question' 10 times in a short, informal way. Assume third person singular if not obvious from the question.
'History': {Conversation history}
'Question': {Question}
'Reformulation':

[Books] History: *How many Pulitzer Prizes has John Updike won? 2.*
Question: *Which was the first book to win him the award?*
Ref 1: *What book earned John Updike his first Pulitzer Prize?*
Ref 2: *What was the author's first book to win a Pulitzer?*
Ref 3: *Title of John Updike's first Pulitzer Prize-winning book?*

[Movies] History: *Which year did the Hobbit An unexpected journey released? 2012.*
Question: *What is the book based on?*
Ref 1: *What's the book about?*
Ref 2: *What's the book's topic?*
Ref 3: *What's the book's subject?*

[Music] History: *Which singer sang the number Single Ladies? Beyonce. What is the year of its release? 2008. Who is her spouse? Jay-Z . What is his date of birth? 4 December 1969.*
Question: *Was Kanye West a composer of the song?*
Ref 1: *Did Kanye West contribute to the lyrics of the song?*
Ref 2: *Did Kanye West perform the song with Beyonce?*
Ref 3: *Was Kanye West featured in the song?*

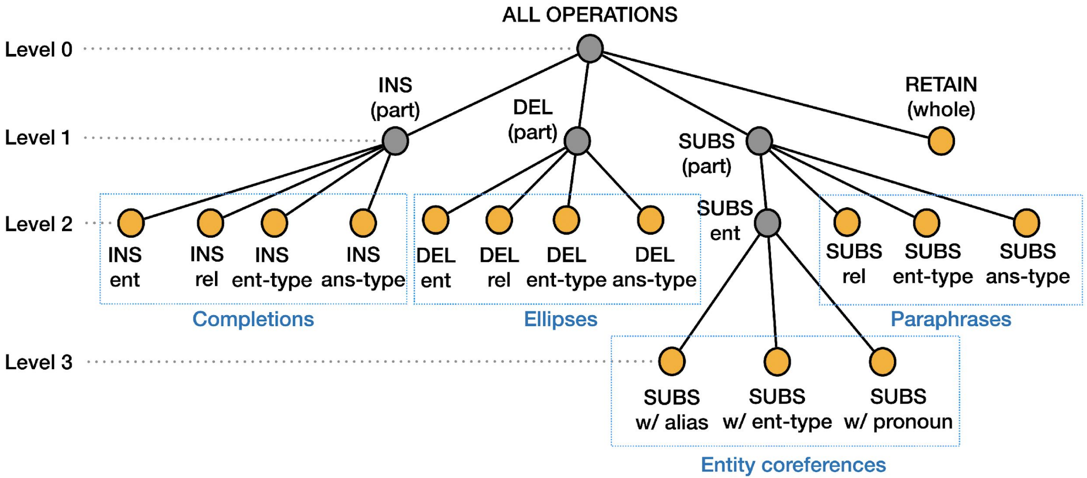
[TV series] History: *What is the release year of the TV series See? 2019.*
Question: *created by?*
Ref 1: *Who's responsible for it?*
Ref 2: *Who's the mastermind?*
Ref 3: *Who's the author?*

[Soccer] History: *Pele scored how many goals in international play? 77. Has he scored the most goals? No.*
Question: *Did Messi beat his goal total?*
Ref 1: *Did Messi surpass Pele's international goal record?*
Ref 2: *Has Messi scored more international goals than Pele?*
Ref 3: *Did Messi break Pele's goal-scoring record?*

Table 3: Examples of GPT reformulations for test sets.

REIGN

Detailed results: REIGN reformulations



[Books] History: *Which book won the 2017 Pulitzer Prize for Fiction? The Underground Railroad. subject of the book? Slavery in the United States. publisher of the novel? Doubleday.*
Question: *author of the fiction?*

Ref 1: *creator of the fiction?* [SUBS rel]
Ref 2: *Which individual is author of the fiction?* [INS ans-type]
Ref 3: *author of the fiction The Underground Railroad?* [INS ent]

[Movies] History: *Who was the director of The Lord of the Rings? Peter Jackson.* **Question:** *Who played Frodo Baggins?*

Ref 1: *Who Frodo Baggins?* [DEL rel]
Ref 2: *Who portrayed Frodo Baggins ?* [SUBS rel]
Ref 3: *Who played Frodo Baggins in it?* [SUBS pronoun]

[Music] History: -
Question: *Formation year of the band U2?*

Ref 1: *Formation year of the rock band U2?* [SUBS ent-type]
Ref 2: *Which year is Formation year of the band U2?* [INS ans-type]
Ref 3: *Formation year of U2?* [DEL ent-type]

[TV series] History: *Who played as Marty in Ozark series? Jason Bateman. and Wendy Byrde? Laura Linney. who is the director of the series? Jason Bateman. How many episodes are in the series? 30.*
Question: *production company of the series?*

Ref 1: *production company of the series television series?* [INS ent-type] (noisy)
Ref 2: *production company of the series Ozark?* [INS ent]
Ref 3: *production house of the series?* [SUBS rel]

[Soccer] History: *What is the full name of footballer Neymar? Neymar da Silva Santos Junior. Birthplace of Neymar? Brazil . When was he born? 5 February 1992.*
Question: *Which club does he play now?*

Ref 1: *Which club does he play now association football player?* [INS ent-type]
Ref 2: *Which club does he play now Neymar?* [INS ent]
Ref 3: *Which Football team does he play now?* [SUBS ans-type]

Table 4: Examples of REIGN-generated reformulations along with respective reformulation categories, used for training.