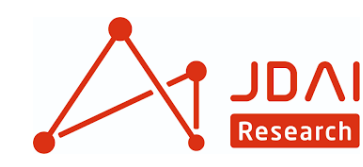# Answering Open-domain Questions of Varying Reasoning Steps from Text

Peng Qi[1,3]*, Haejun Lee[2]*, Oghenetegiri "TG" Sido[3], Christopher D. Manning[3]

*=Equal contribution   1 JD AI Research   2 Samsung Research   3 Stanford University

tl;dr: We present a new open-domain QA dataset, 🍺 **BeerQA**, which evaluates models in a more realistic setting of non-predetermined hops of reasoning. We also present a multi-task Transformer model, **IRRR**, that iterates between retrieval, reading comprehension, and reranking to answer questions that require arbitrary steps of reasoning, which achieves 👍 performance on prior benchmarks and BeerQA.

Background & Main Contributions

Current open-domain QA research is dichotomized into **single-hop** and **multi-hop** datasets and models.

	Single-hop	Multi-hop
Datasets	SQuAD Open TriviaQA Natural Questions ...	HotpotQA QAngaroo
Models	DrQA Dense Passage Retrieval Sparse Transformer Matching ...	GoldEn Retriever Question Decomposition Multi-hop DPR ...

But whether a question is **single-hop** or **multi-hop** is often difficult to determine just by looking at the question itself, because it is highly **context-dependent**!

Facebook, Inc. was founded in Cambridge, *Massachusetts*.	**Question:** In which U.S. state was Facebook founded?	Facebook, Inc. was founded in a Harvard dorm room. Harvard University is located in *Massachusetts*.
Matthew Perry is the youngest among the six main cast members of Friends.	**Question:** Which main cast member of the TV show *Friends* is the youngest?	*Friends*' main cast include Jennifer Aniston, Matthew Perry, ... Jennifer Aniston (born Feb 11, 1969) ... *Matthew Perry* (born Aug 19, 1969) ...

Our main contributions: a dataset and a model for more realistic, varying-hop open-domain QA

🍺 **BeerQA**
- SQuAD + HotpotQA "time-traveled" to 2020
- 500+ test questions requiring 3+ Wiki pages to answer

IRRR: Iterative Retriever, Reader, and Reranker
- Open-domain QA with varying hops in a single multi-task Transformer model
- Competitive performance + explainable reasoning

🍺 BeerQA

We combine SQuAD, HotpotQA, and 500+ newly annotated 3+ hop test questions on a 2020 version of the English Wikipedia to construct a new dataset.

SQuAD
- 100k questions
- On Wikipedia passages from **2016**
- Passages taken **across entire pages**, from **a small number of popular pages**
- Each question requires **one** Wikipedia page to answer

HotpotQA
- 113k questions
- On Wikipedia passages from **2017**
- Passages taken from **introductory paragraphs** in **a large number of pages**
- Each question requires **two** Wikipedia pages to answer

➡️ Find every paragraph in a 2020 version of the English Wikipedia
(sometimes the text has been significantly paraphrased or the world has changed too much that this is not always possible!)

➡️ Filter out paragraphs that don't contain the answer

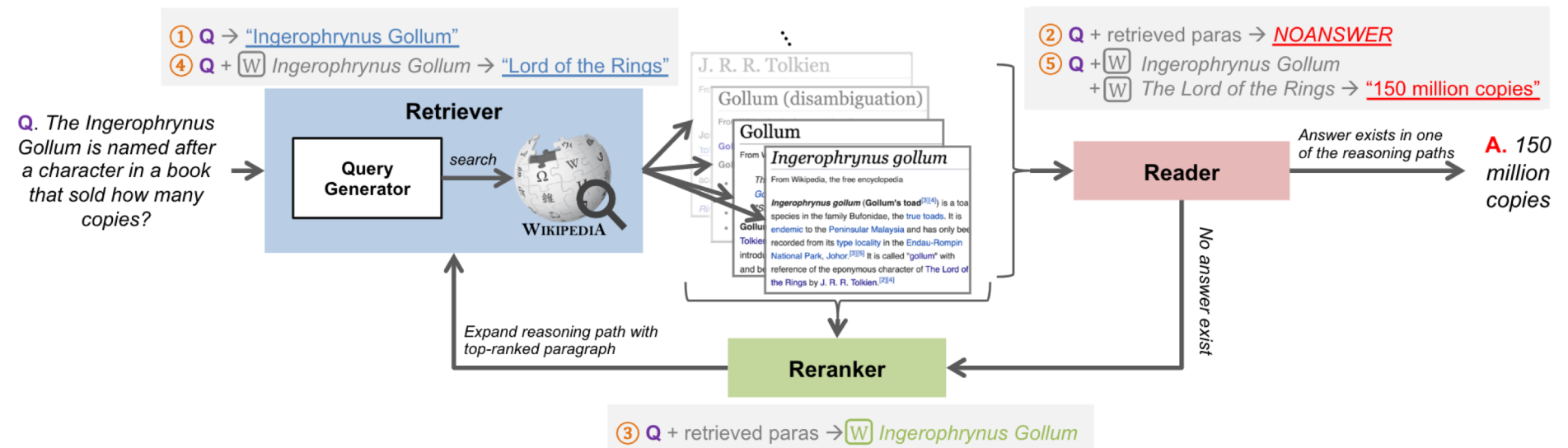| **3+ hop challenges** 530 test-only questions	**SQuAD-2020** ~76k questions	**HotpotQA-2020** ~87k questions

🍺 **BeerQA:** ~163k questions of varying reasoning steps

IRRR: Iterative Retriever, Reader, and Reranker

Retriever(Question, retrieved docs) → Natural language queries
- Prior work restrict queries to be substrings of "question + concatenated retrieved docs"
- We expand to **arbitrary subsequences** and propose an algorithm to find good ones for retriever training in $O(n)$ time
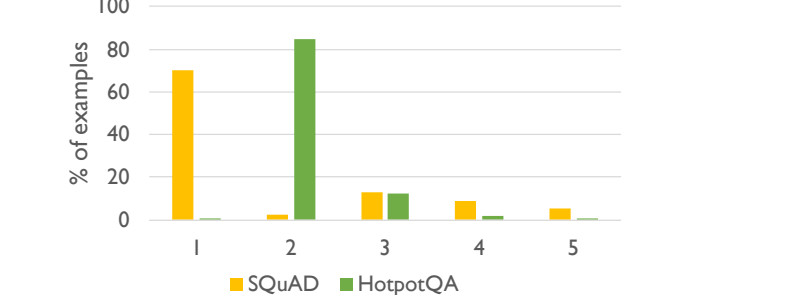

Recall of "Paragraph #2" (%)

— GoldEn Retriever   — IRRR
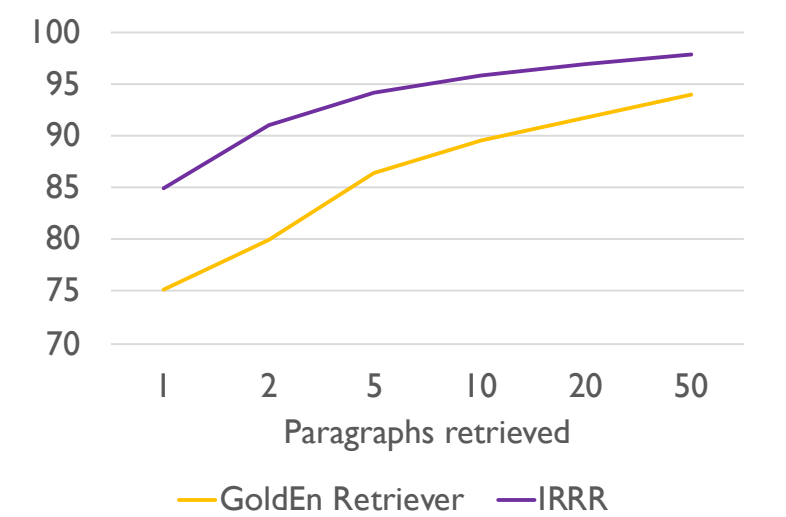
Reader(Question, retrieved docs) → Span answer/yes/no/NOANSWER
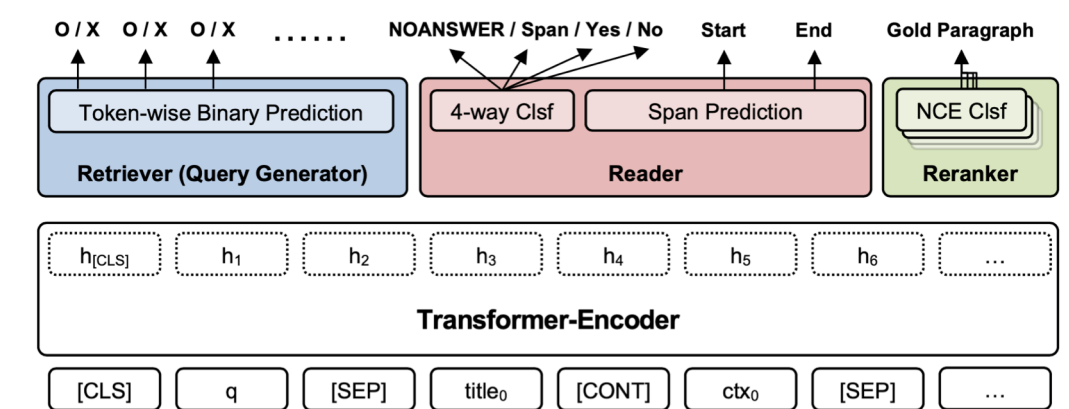- **Dynamically stops** the iterative process when an answer can be predicted

Reasoning steps / example

■ SQuAD   ■ HotpotQA

Reranker(Question, prev retrieved docs, newly retrieved docs) → Best newly retrieved doc to append to the reasoning path
- **Contrastive loss** between the gold paragraph and top retrieved candidates
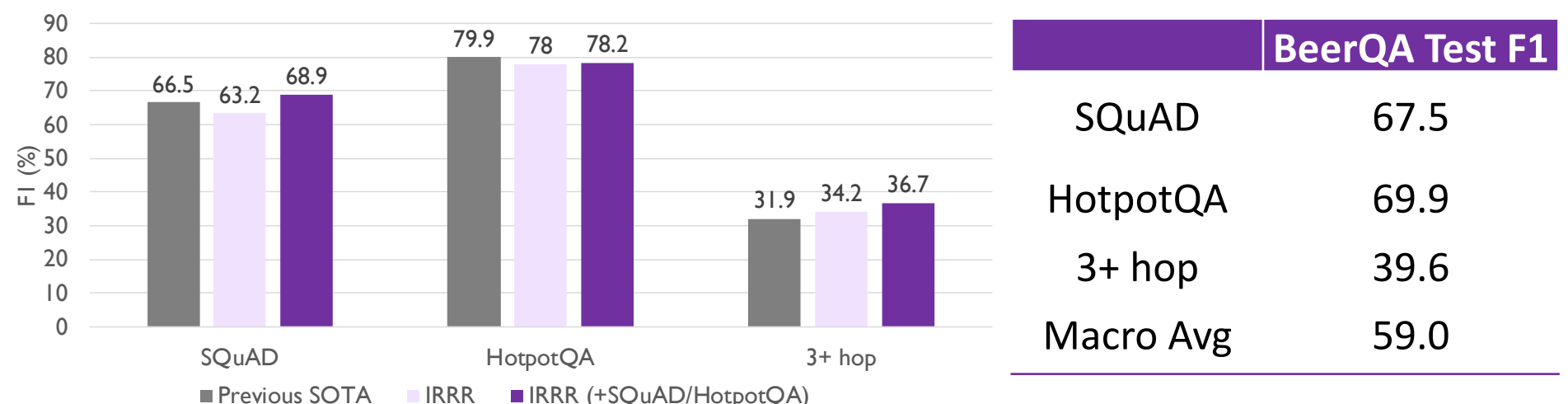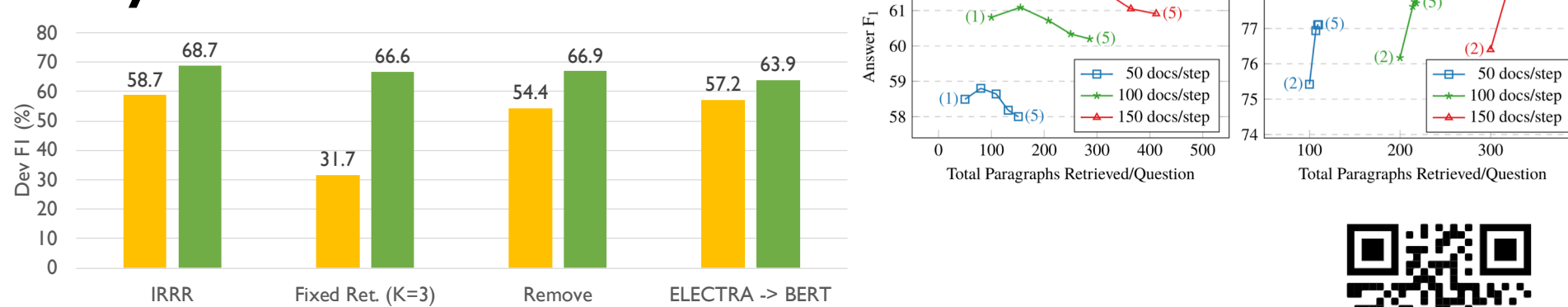- **Augmented reasoning paths** during training

Experiments

Setup

	Train		Test		
Benchmark	SQuAD'16		SQuAD'16		
	HotpotQA'17		HotpotQA'17	3+ hop	
Benchmark+	SQuAD'16	HotpotQA'17	SQuAD'16		
	SQuAD'16	HotpotQA'17	HotpotQA'17	3+ hop	
New benchmark	SQuAD'20	HotpotQA'20	🍺 BeerQA		

Benchmark Results

■ Previous SOTA   ■ IRRR   ■ IRRR (+SQuAD/HotpotQA)

	BeerQA Test F1
SQuAD	67.5
HotpotQA	69.9
3+ hop	39.6
Macro Avg	59.0

Analysis

Question	The **Ingerophrynus gollum** is **named** after a **character** in a **book** that **sold** how many **copies**?
Step 1 *(Non-Gold)*	Ingerophrynus is a genus of true toads with 12 species. ... In 2007 a new species, "Ingerophrynus gollum", was added to this genus. This species is named after the character Gollum created by **J. R. R. Tolkien.**
Query	Ingerophrynus gollum book sold copies **J. R. R. Tolkien**
Step 2 *(Gold)*	Ingerophrynus gollum (Gollum's toad) is a species of **true** toad. ... It is called "gollum" with reference of the eponymous character of The **Lord of the Rings** by J. R. R. Tolkien.
Query	Ingerophrynus gollum character book sold copies J. R. R. Tolkien **true Lord of the Rings**
Step 3 *(Gold)*	The Lord of the Rings is an epic high fantasy novel written by English author and scholar J. R. R. Tolkien. ... is one of the best-selling novels ever written, with **150 million copies** sold.
Answer/GT	**150 million copies**

🍺 BeerQA

IRRR