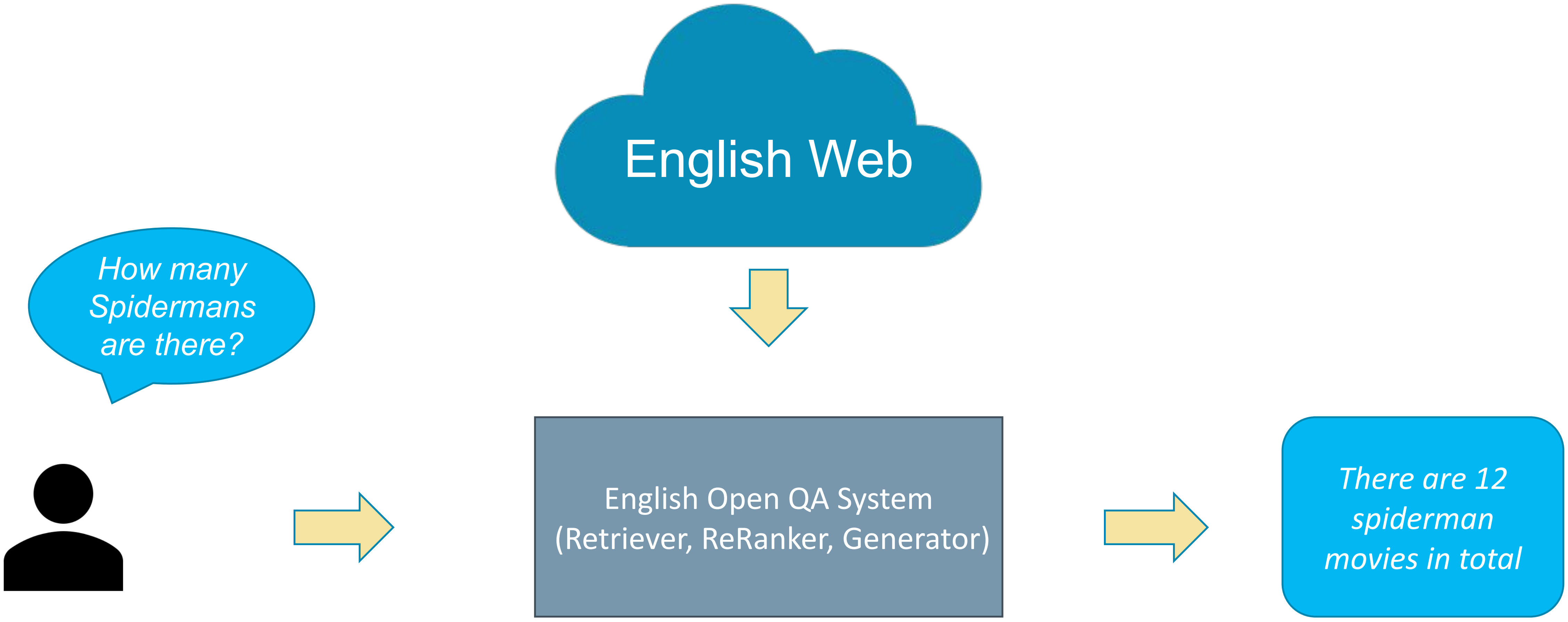
Toward a Cross-Lingual Generative Question Answering System

Benjamin Muller

Joint work done during an internship at Amazon Alexa AI
with Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind and Alessandro Moschitti

December 2021

Standard QA systems are Monolingual



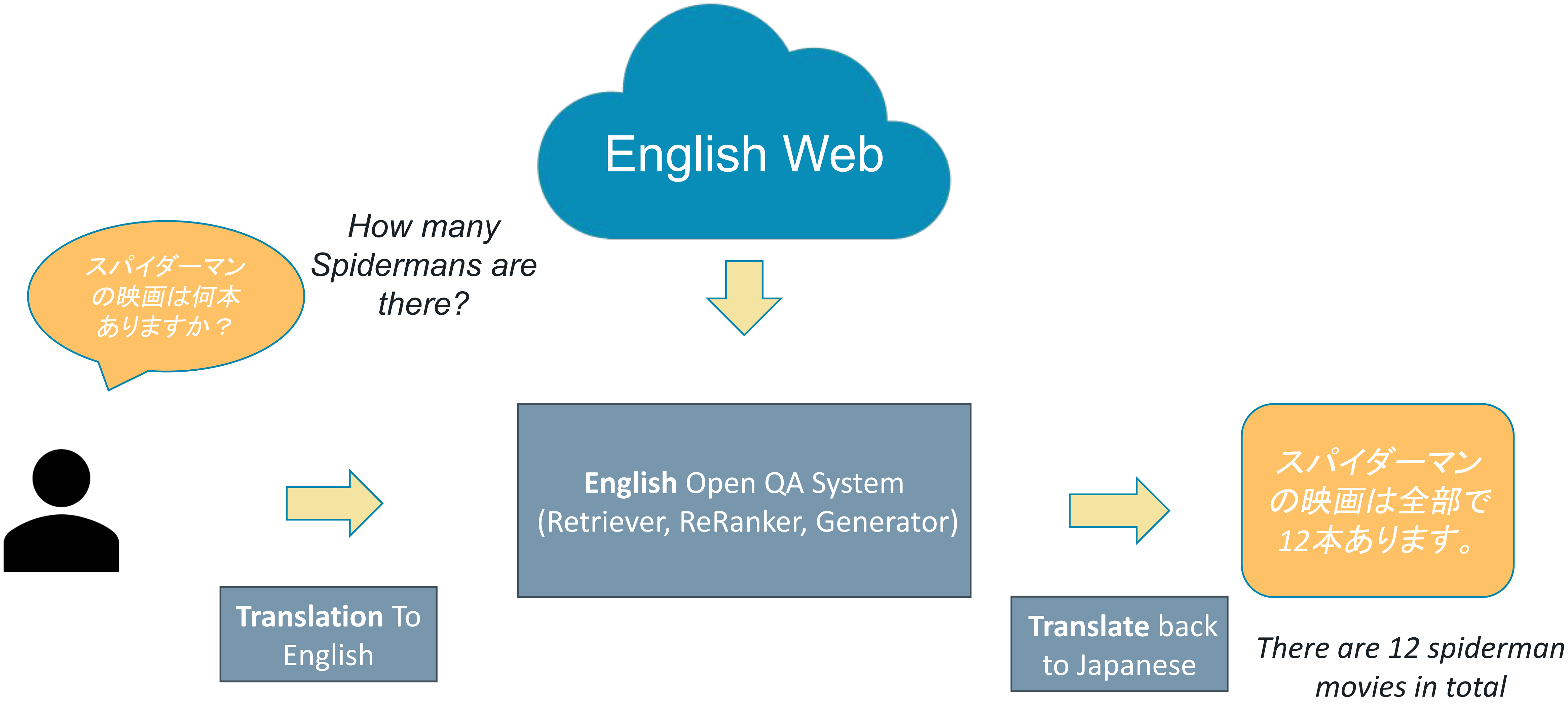
English Web

How many Spidermans are there?

English Open QA System (Retriever, ReRanker, Generator)

There are 12 spiderman movies in total

Standard QA systems are Monolingual



English Web

スパイダーマンの映画は何本ありますか？

How many Spidermans are there?

Translation To English

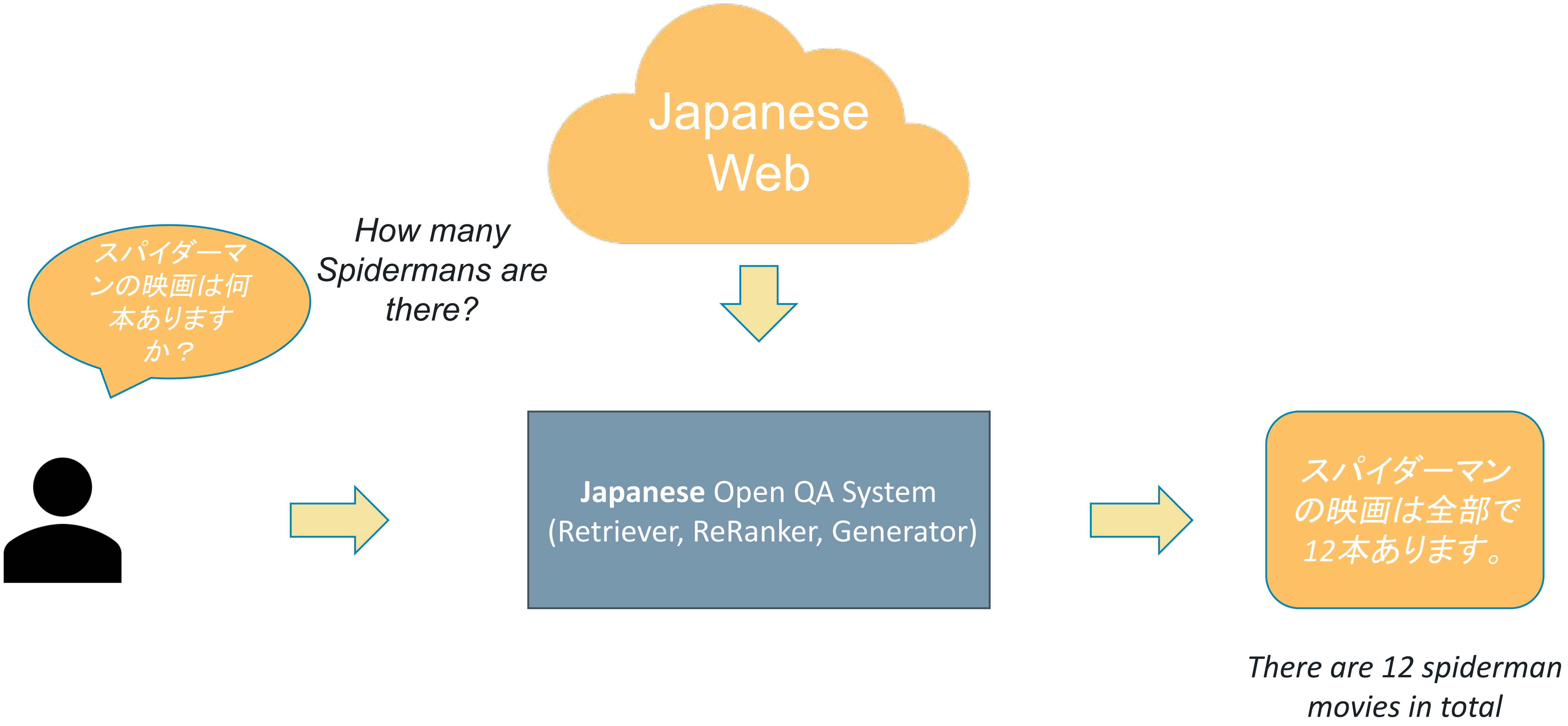
English Open QA System
(Retriever, ReRanker, Generator)

Translate back to Japanese

スパイダーマンの映画は全部で12本あります。

There are 12 spiderman movies in total

Standard QA systems are Monolingual



Japanese Web

スパイダーマンの映画は何本ありますか？

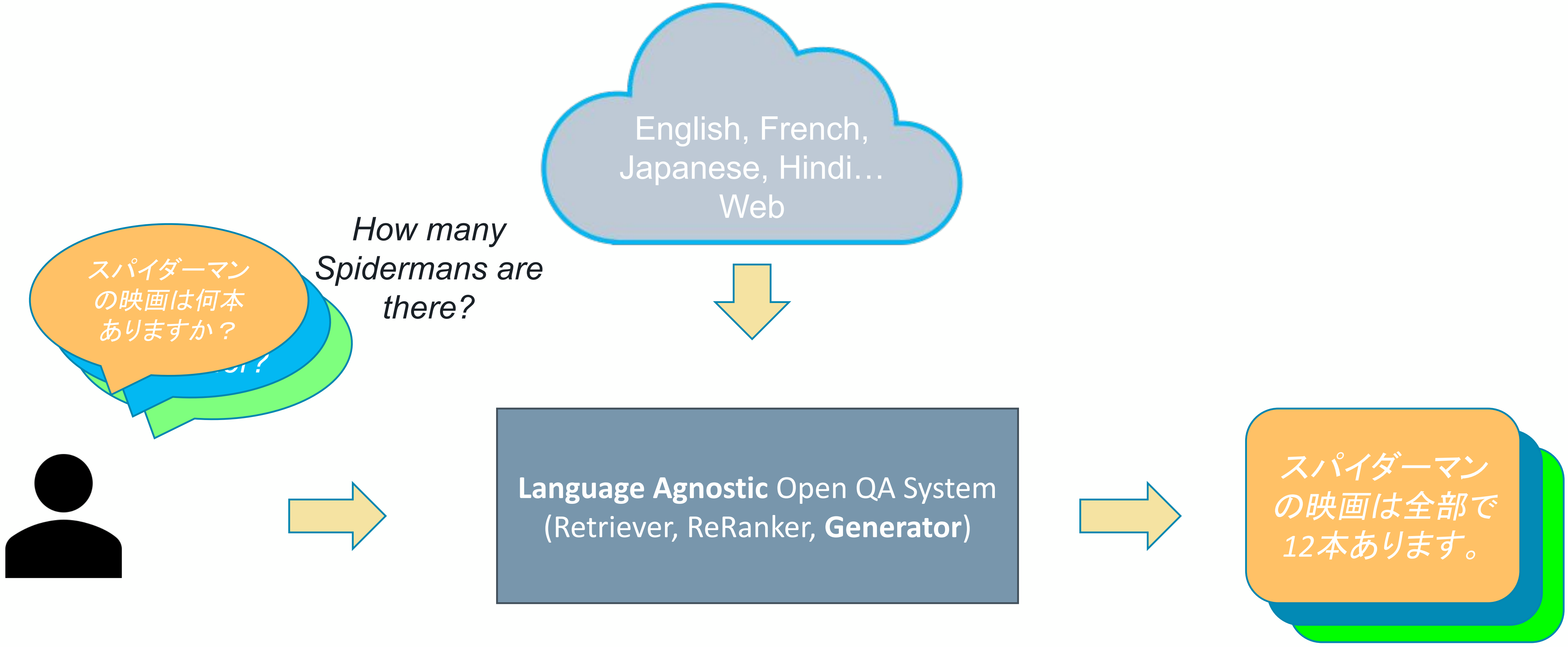
How many Spidermans are there?

Japanese Open QA System (Retriever, ReRanker, Generator)

スパイダーマンの映画は全部で12本あります。

There are 12 spiderman movies in total

Toward a Language Agnostic QA System



スパイダーマンの映画は何本ありますか？

How many Spidermans are there?

English, French, Japanese, Hindi…
Web

Language Agnostic Open QA System
(Retriever, ReRanker, **Generator**)

スパイダーマンの映画は全部で12本あります。

There are 12 spiderman movies in total

Definition and Research Question

Definition: **Language Agnostic Question Answering system**

*Answering a question in **any language** using information **from many languages***

How can we design **a Language Agnostic Question Answering** System?

- Language Agnostic Retriever / Ranker
- **Language Agnostic Generative Question Answering (GenQA): our focus**

Motivations

- GenQA delivers **better answers in English** than competing systems [Hsu et. Al 2021]
- Answering **more questions** e.g. 88% of Wikipedia items have content in less than 5 languages [Valentim et. Al 2021]
- Answering **questions more accurately**
- Scaling **the number of supported languages** without scaling cost

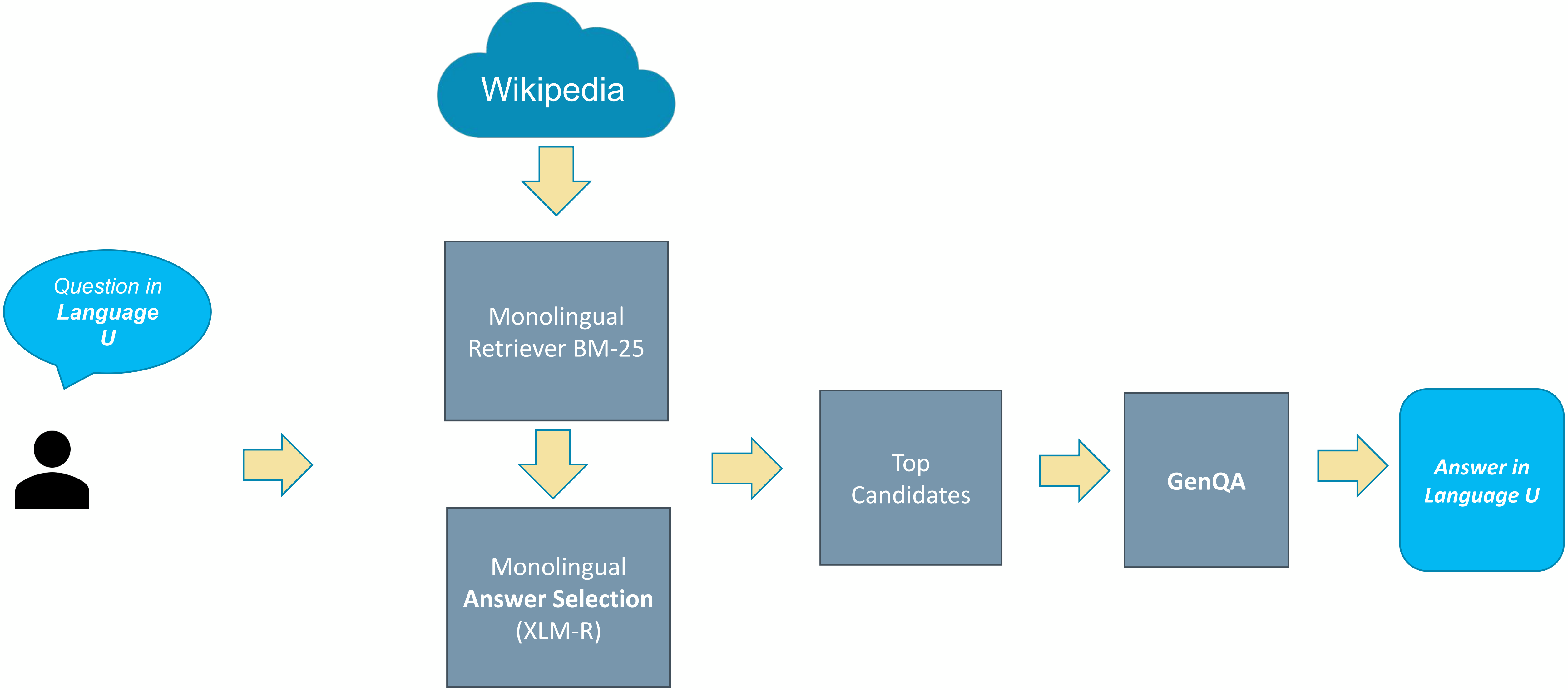
Outline

1. **The GenQA Pipeline**

2. The Gen-TyDiQA Dataset

3. Cross-Lingual GenQA in the End-To-End Setting

Monolingual Pipeline



Question in ***Language U***

Wikipedia

Monolingual Retriever BM-25

Monolingual **Answer Selection** (XLM-R)

Top Candidates

GenQA

Answer in Language U

The GenQA Task

Example of reranked answers

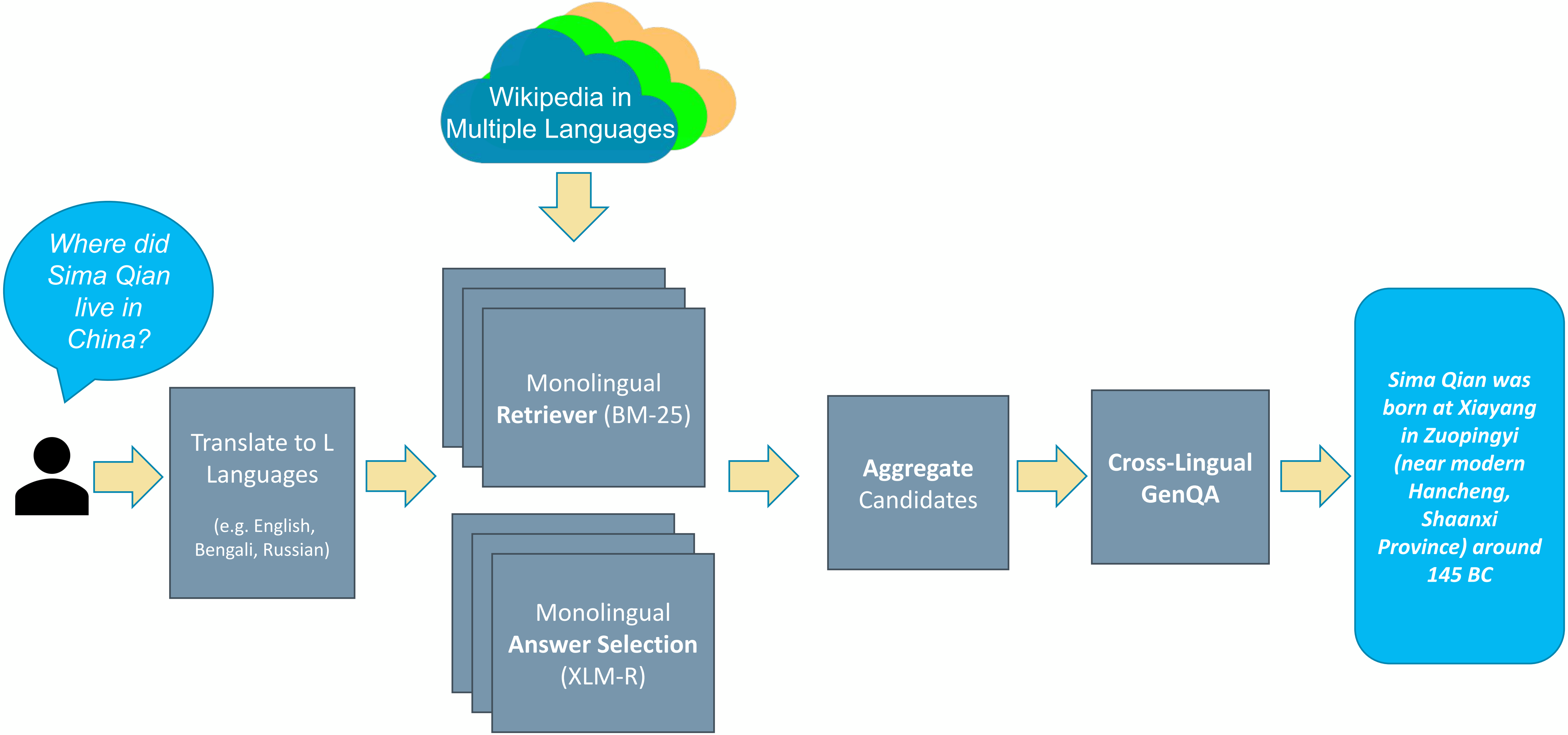
Question	how long do you need to go to school for to get a degree in music?
Top 1 candidate (incorrect)	is it possible to get a degree in music education within four years?
Top 2 candidate (correct)	in many schools, the degree in music alone averages over **4 years**.
Top 3 candidate (incorrect)	all requirements for music degrees should be completed within seven years of the date the student is admitted to the school of music.
Top 4 candidate (correct)	the average number of years it takes to complete a **bachelor's degree** in music at schools in the u.s. is 4.
Top 5 candidate (incorrect)	if you don't make money in music after 4 years then go back to school and get your masters degree.

Example of Generated Answers

The average time to get a bachelor's degree in music is four years.

[Hsu et. al 2021]

Cross-Lingual Pipeline



Where did Sima Qian live in China?

Translate to L Languages

(e.g. English, Bengali, Russian)

Wikipedia in Multiple Languages

Monolingual **Retriever** (BM-25)

Monolingual **Answer Selection** (XLM-R)

Aggregate Candidates

Cross-Lingual GenQA

Sima Qian was born at Xiayang in Zuopingyi (near modern Hancheng, Shaanxi Province) around 145 BC

Cross-Lingual GenQA: Task

Example of reranked answers

Question	how are meiosis and mitosis similar?
Candidate French	La division cellulaire et la reproduction peuvent se produire de deux façons - la mitose et la méiose. La mitose est un processus de duplication cellulaire, ou de reproduction, au cours duquel une cellule donne naissance à deux cellules filles génétiquement identiques.;
Candidate French	La méiose et la mitose sont des formes de reproduction cellulaire. Ils ont tous les deux une prophase et une anaphase.. Les deux sont des formes de division nucléaire. La méiose se produit dans les cellules génératives et la mitose dans les cellules somatiques.;
Candidate Italian	Sia la mitosi che la meiosi sono tipi di divisione cellulare che condividono molte somiglianze, ed entrambi condividono le stesse fasi fondamentali della profase, metafasi, anafasi e telofase, secondo l'Università di...
Candidate Spanish	Respuesta rápida. Tanto la mitosis como la meiosis son tipos de división celular que comparten muchas similitudes, y ambas comparten las mismas etapas básicas de profase, metafase, anafase y telofase, según la Universidad de Illinois en Chicago. Ambos comienzan con cambios en la organización del ADN en el núcleo de una célula.
Candidate German	Sowohl die Mitose als auch die Meiose sind Arten der Zellteilung, die viele Gemeinsamkeiten aufweisen, und beide teilen die gleichen grundlegenden Phasen der Prophase, Metaphase, Anaphase und Telophase, so die University of...;

Example of Generated Answers

Mitosis and meiosis are similar in their basic stages of prophase, metaphase, anaphase, and telophase.

Architecture and Model for GenQA

Multilingual T5 [Xue et al. 2020]

Encoder-Decoder Transformer model

Pretrained on the **mC4** dataset (101 languages, 26TB, 6.3T tokens) using a **span-corruption objective**

Fine-Tuning

INPUT: **Question** \n **candidate 1** \n …\n **candidate 10**

OUTPUT: **Answer**

Experimental Setting

- **Training** on **MSMARCO** with Translated Data

Experimental Setting

- **Training** on **MSMARCO** with Translated Data
- **Evaluation on GenTyDiQA** Dataset with Gold Passages from TyDiQA and Retrieved Candidates from Multiple Languages

Experimental Setting

- **Training** on **MSMARCO** with Translated Data

- **Evaluation on GenTyDiQA** Dataset with Gold Passages from TyDiQA and Retrieved Candidates from Multiple Languages

- **Metrics**

  Automatic Evaluation with n-grams based metrics: **BLEU** and **ROUGE**

  Human Judgement Ask several turkers if a sentence is correct given a reference answer (Accuracy)

Outline

1. The GenQA Pipeline

2. **The Gen-TyDiQA Dataset**

3. Cross-Lingual GenQA in the End-To-End Setting

Extending the TyDiQA Dataset

- The ***Typologically Diverse Question Answering*** ***(TyDiQA)*** dataset is a multilingual machine reading dataset in 11 languages [Clark et. Al 2020]

(Question, 1 Wikipedia Page, Paragraph/Null, Span/Null, Yes/No/Null)

- We extend the TyDiQA dataset in two ways

1. Reformat TyDiQA to an **Answer Sentence Selection Dataset**
2. Collect Natural Answers: **Gen-TyDiQA** Dataset

The Gen-TyDiQA Dataset

For **English, Arabic, Bengali, Japanese** and **Russian**

1. **Generation Step:** Given the question and a paragraph we ask Turkers to answer the question with **well-formed and self-contained** answer
2. **Validation Step:** Turkers validate that the answer is correct (based on the information provided in the paragraph), well-formed, and self-contained

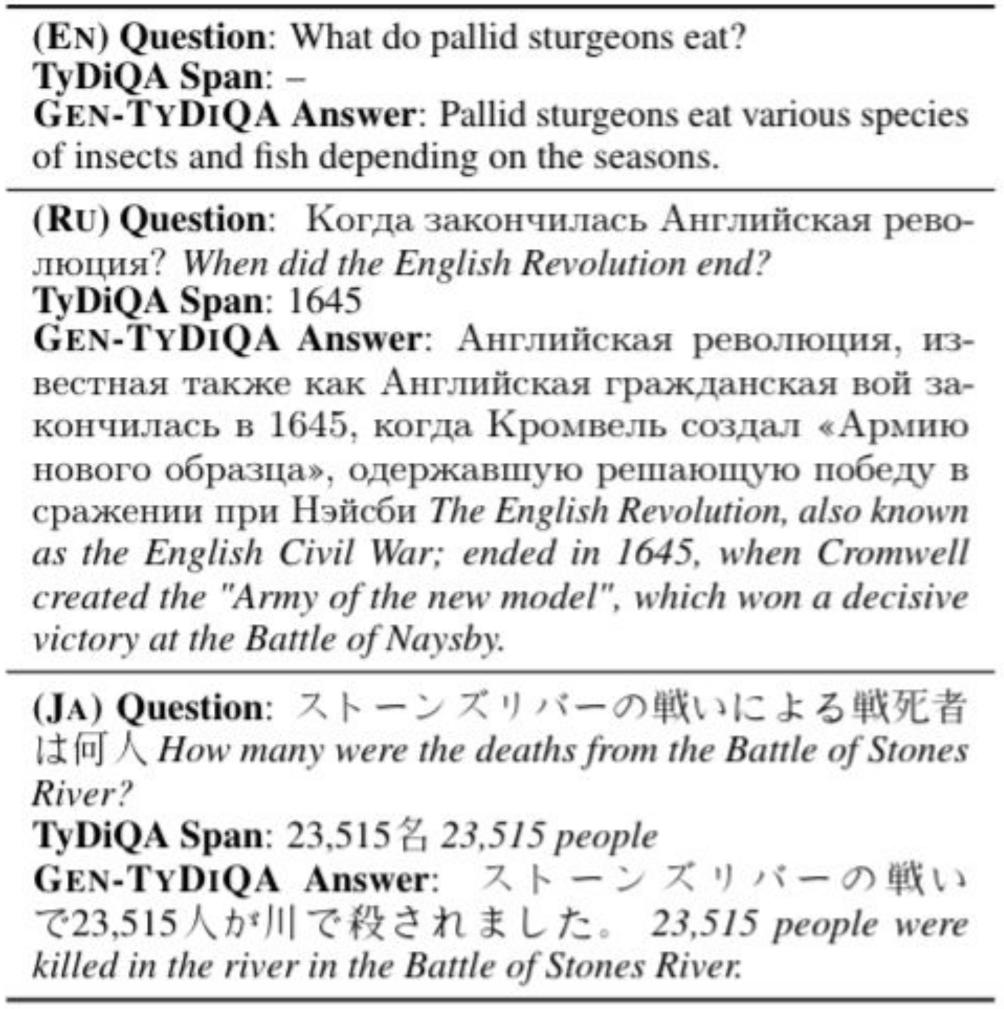
We accept answers only if:

Correct Ensured using the reference TyDiQA paragraph

Well-Formed Answers Answers should be grammatically correct, full sentences and natural sounding

Self-Contained The answer can be understood without any extra-context

Gen-TyDiQA Dataset



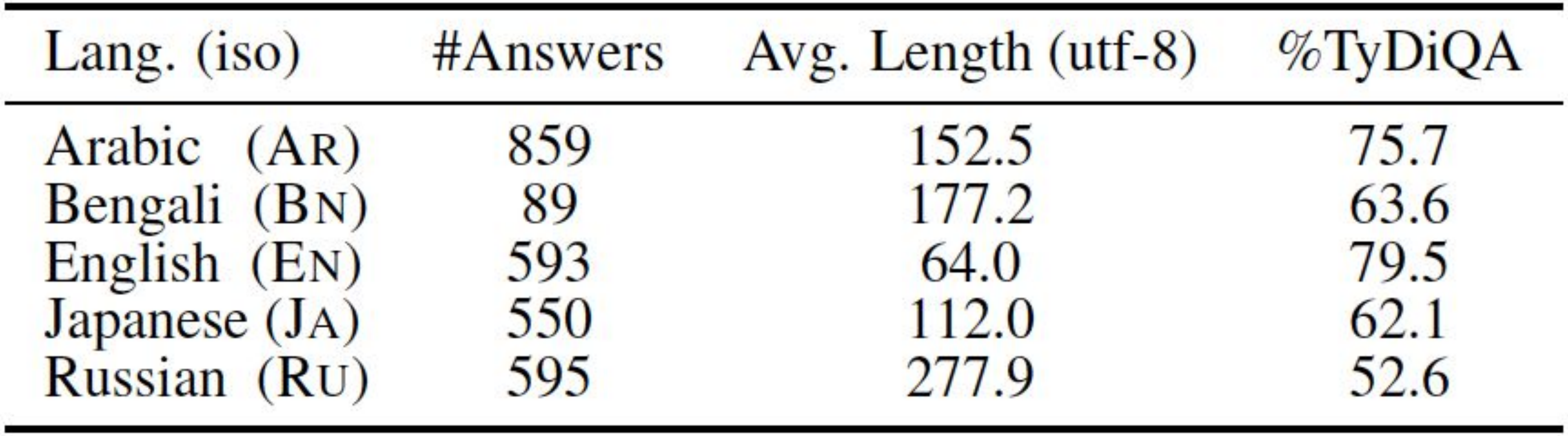
(En) Question: What do pallid sturgeons eat?
TyDiQA Span: –
Gen-TyDiQA Answer: Pallid sturgeons eat various species of insects and fish depending on the seasons.

(Ru) Question: Когда закончилась Английская революция? *When did the English Revolution end?*
TyDiQA Span: 1645
Gen-TyDiQA Answer: Английская революция, известная также как Английская гражданская вой закончилась в 1645, когда Кромвель создал «Армию нового образца», одержавшую решающую победу в сражении при Нэйсби *The English Revolution, also known as the English Civil War; ended in 1645, when Cromwell created the "Army of the new model", which won a decisive victory at the Battle of Naysby.*

(Ja) Question: ストーンズリバーの戦いによる戦死者は何人 *How many were the deaths from the Battle of Stones River?*
TyDiQA Span: 23,515名 *23,515 people*
Gen-TyDiQA Answer: ストーンズリバーの戦いで23,515人が川で殺されました。 *23,515 people were killed in the river in the Battle of Stones River.*

Table 1: Gen-TyDiQA question and answer samples.

Gen-TyDiQA Dataset



Lang. (iso)	#Answers	Avg. Length (utf-8)	%TyDiQA
Arabic (AR)	859	152.5	75.7
Bengali (BN)	89	177.2	63.6
English (EN)	593	64.0	79.5
Japanese (JA)	550	112.0	62.1
Russian (RU)	595	277.9	52.6

Table: Statistics on our new Gen-TyDiQA dataset

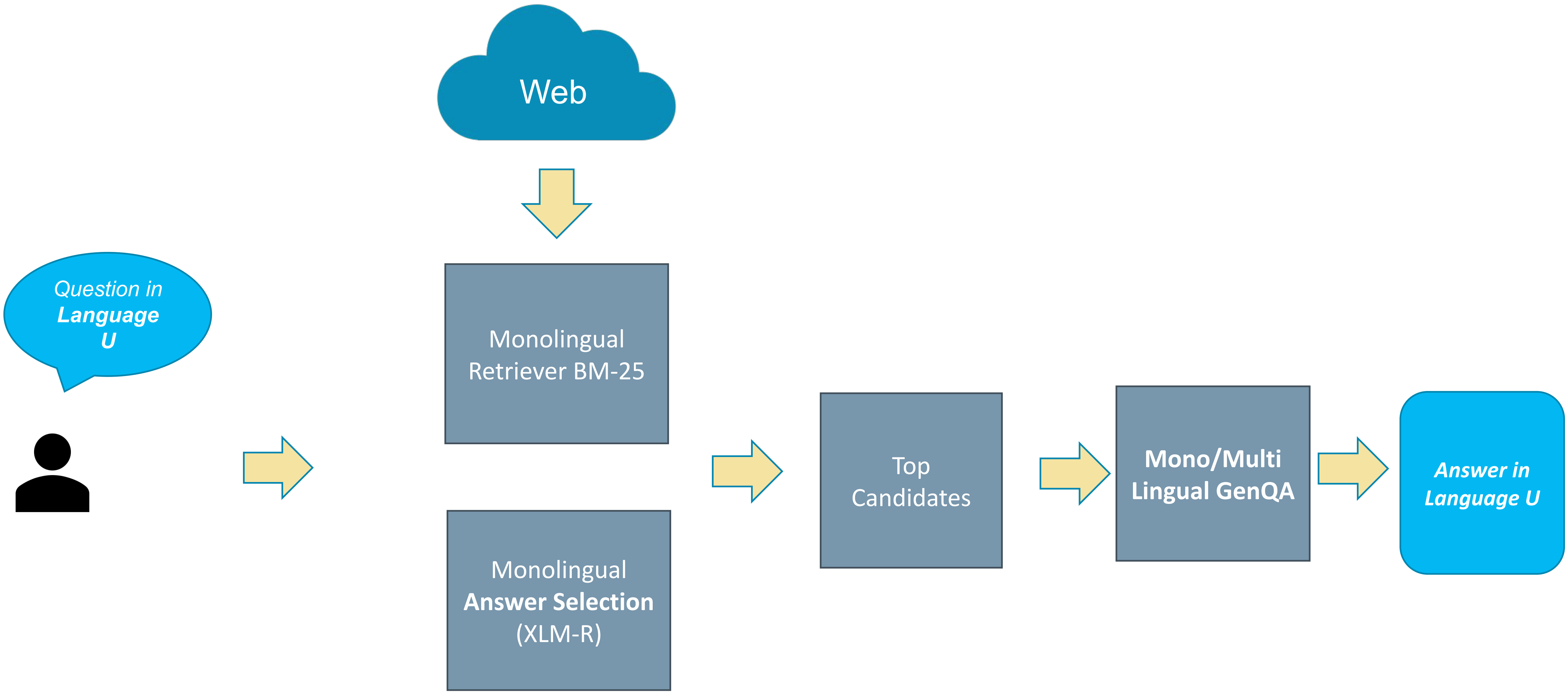
Outline

1. The GenQA Pipeline

2. The Gen-TyDiQA Dataset

3. **Cross-Lingual GenQA in the End-To-End Setting**

Monolingual Pipeline



Question in ***Language U***

Web

Monolingual Retriever BM-25

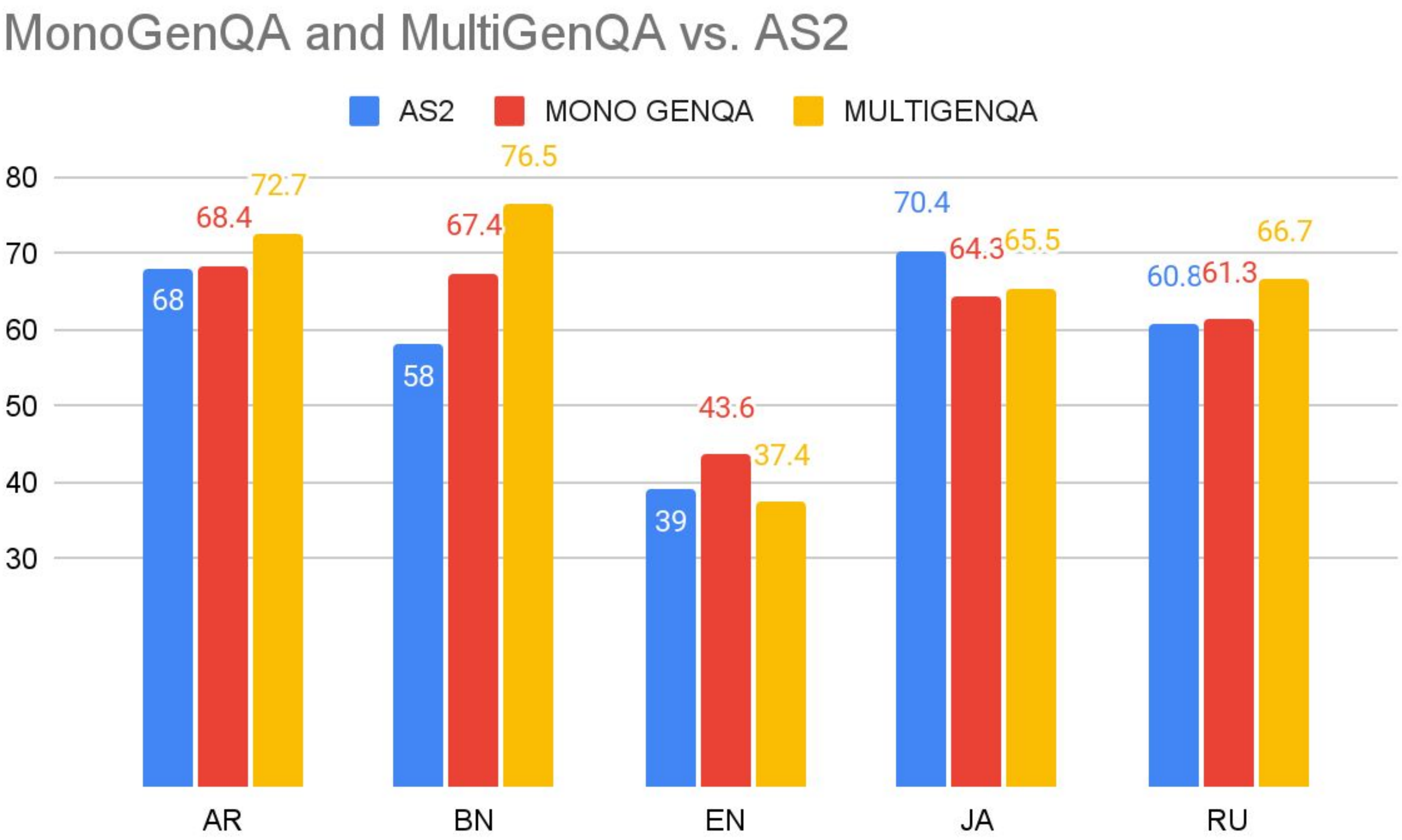
Monolingual **Answer Selection** (XLM-R)

Top Candidates

Mono/Multi Lingual GenQA

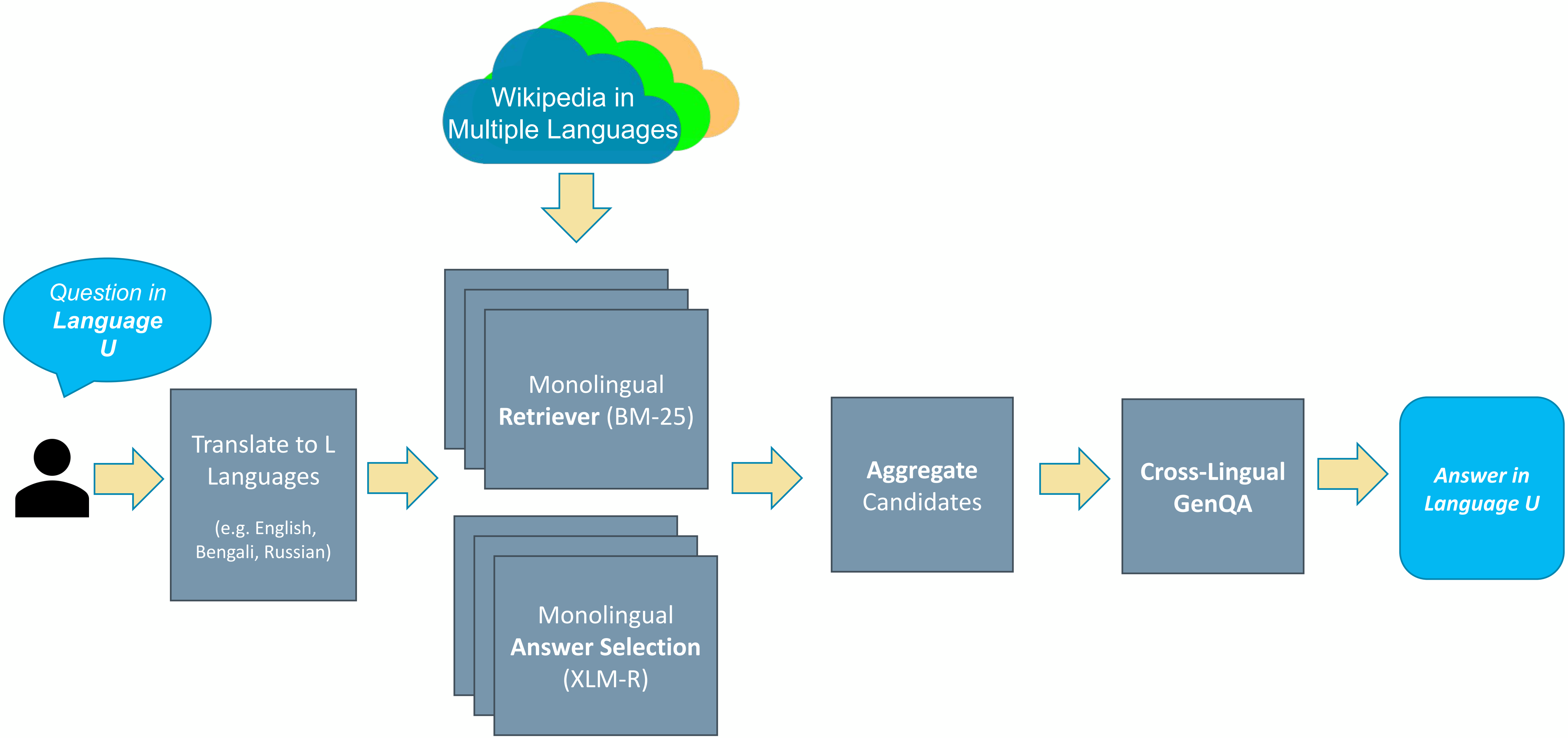
Answer in Language U

Results



- Monolingual GenQA and Multilingual GenQA outperforms the AS2 approach in all cases except JA
- Multilingual GenQA outperforms Monolingual GenQA for all languages except English (multilingual modelling)

Cross-Lingual Pipeline



Wikipedia in Multiple Languages

Question in ***Language U***

Translate to L Languages

(e.g. English, Bengali, Russian)

Monolingual **Retriever** (BM-25)

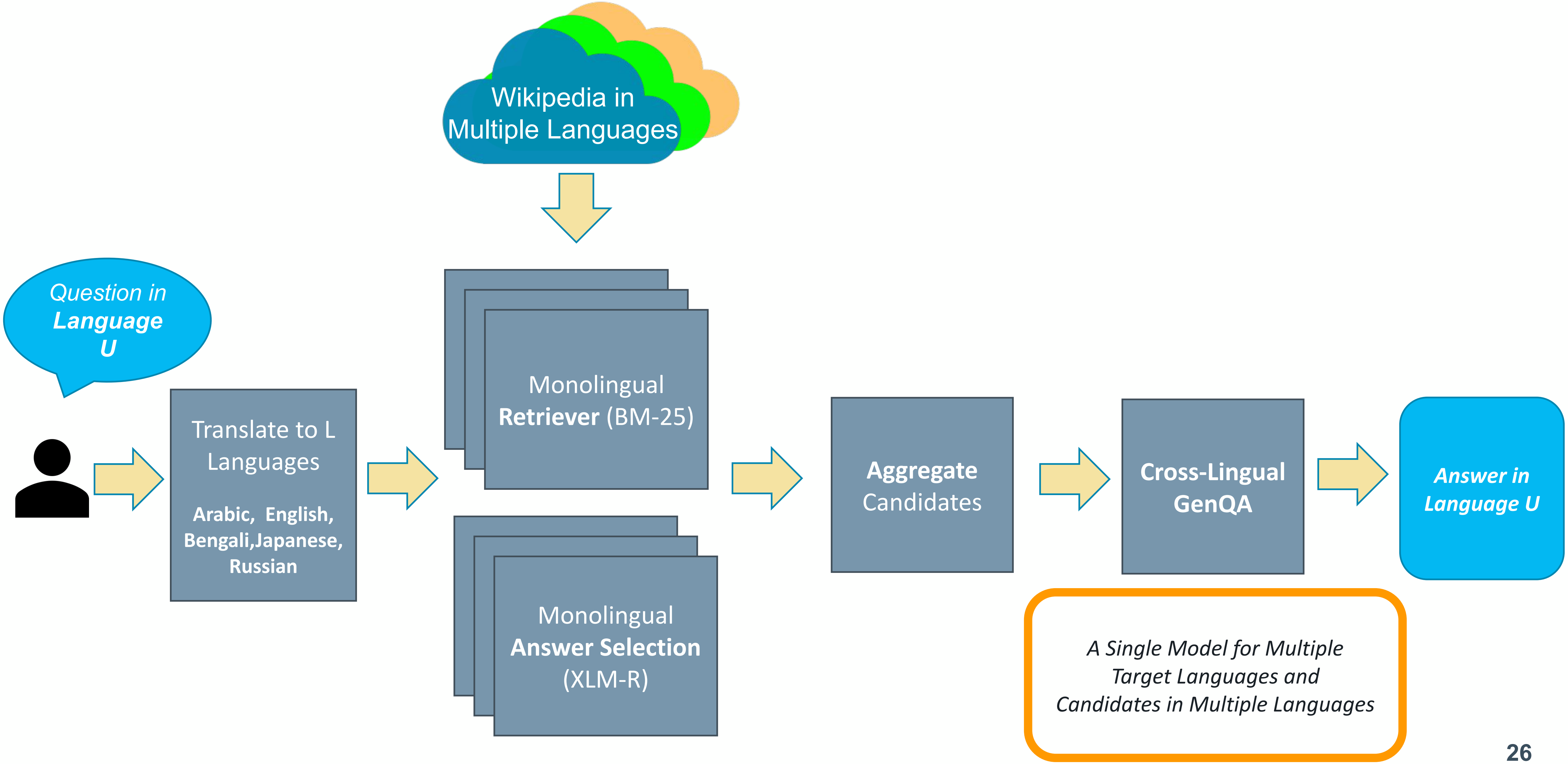
Monolingual **Answer Selection** (XLM-R)

Aggregate Candidates

Cross-Lingual GenQA

Answer in Language U

Cross-Lingual Pipeline



Wikipedia in Multiple Languages

Question in ***Language U***

Translate to L Languages

Arabic, English, Bengali,Japanese, Russian

Monolingual **Retriever** (BM-25)

Monolingual **Answer Selection** (XLM-R)

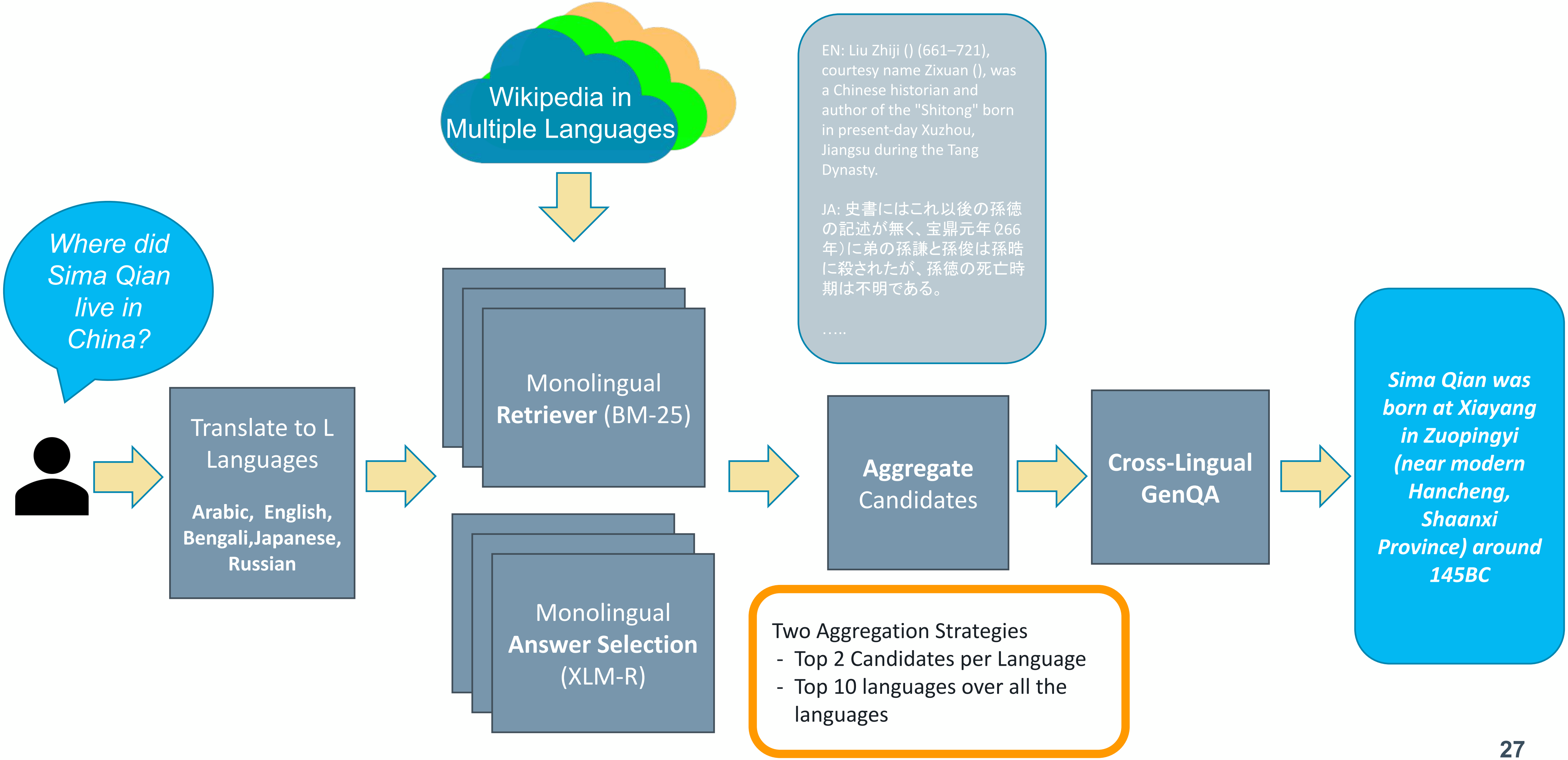
Aggregate Candidates

Cross-Lingual GenQA

Answer in Language U

A Single Model for Multiple Target Languages and Candidates in Multiple Languages

Cross-Lingual Pipeline



Where did Sima Qian live in China?

Translate to L Languages

Arabic, English, Bengali,Japanese, Russian

Wikipedia in Multiple Languages

Monolingual **Retriever** (BM-25)

Monolingual **Answer Selection** (XLM-R)

EN: Liu Zhiji () (661–721), courtesy name Zixuan (), was a Chinese historian and author of the "Shitong" born in present-day Xuzhou, Jiangsu during the Tang Dynasty.

JA: 史書にはこれ以後の孫徳の記述が無く、宝鼎元年（266年）に弟の孫謙と孫俊は孫晧に殺されたが、孫徳の死亡時期は不明である。

…..

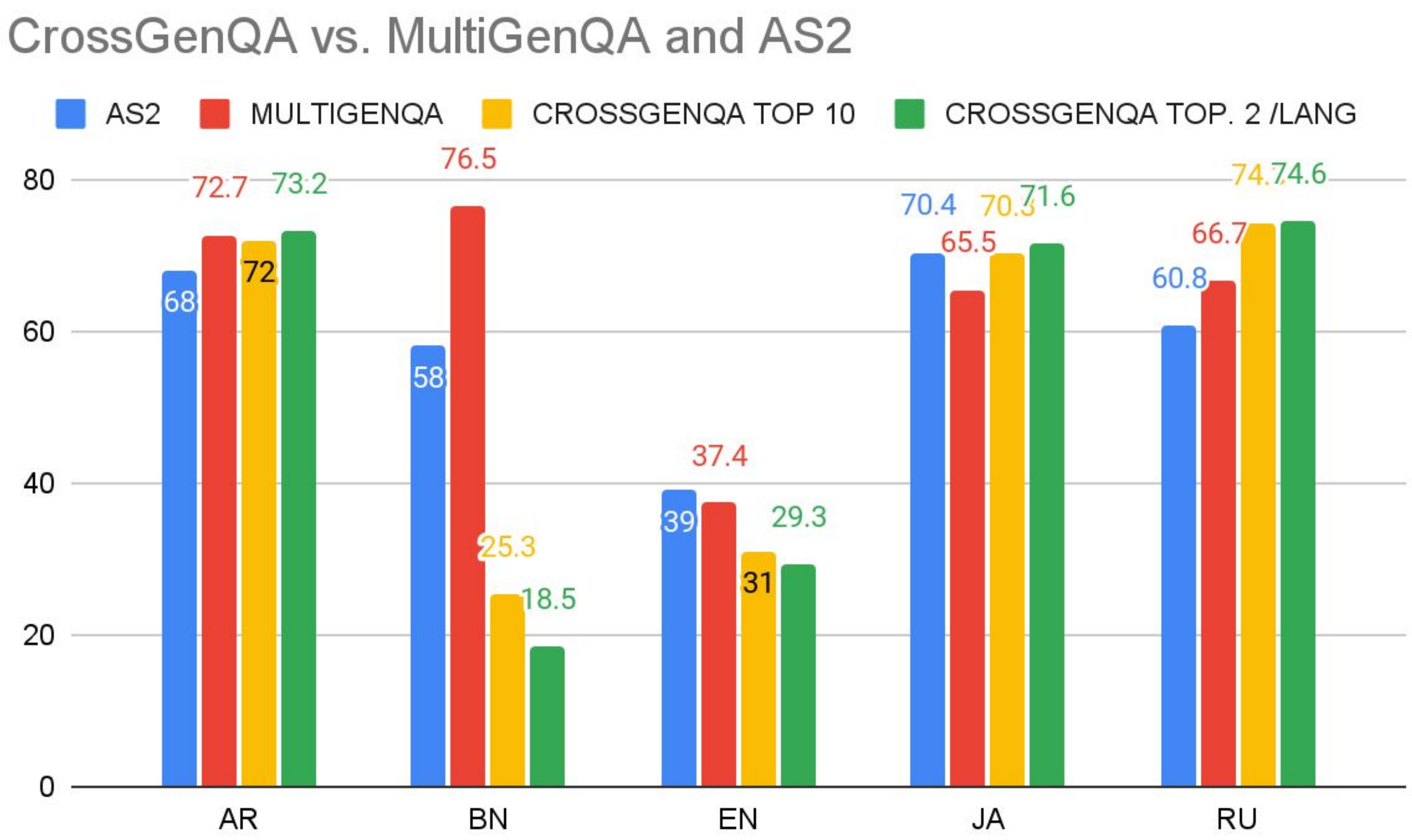
Aggregate Candidates

Cross-Lingual GenQA

Sima Qian was born at Xiayang in Zuopingyi (near modern Hancheng, Shaanxi Province) around 145BC

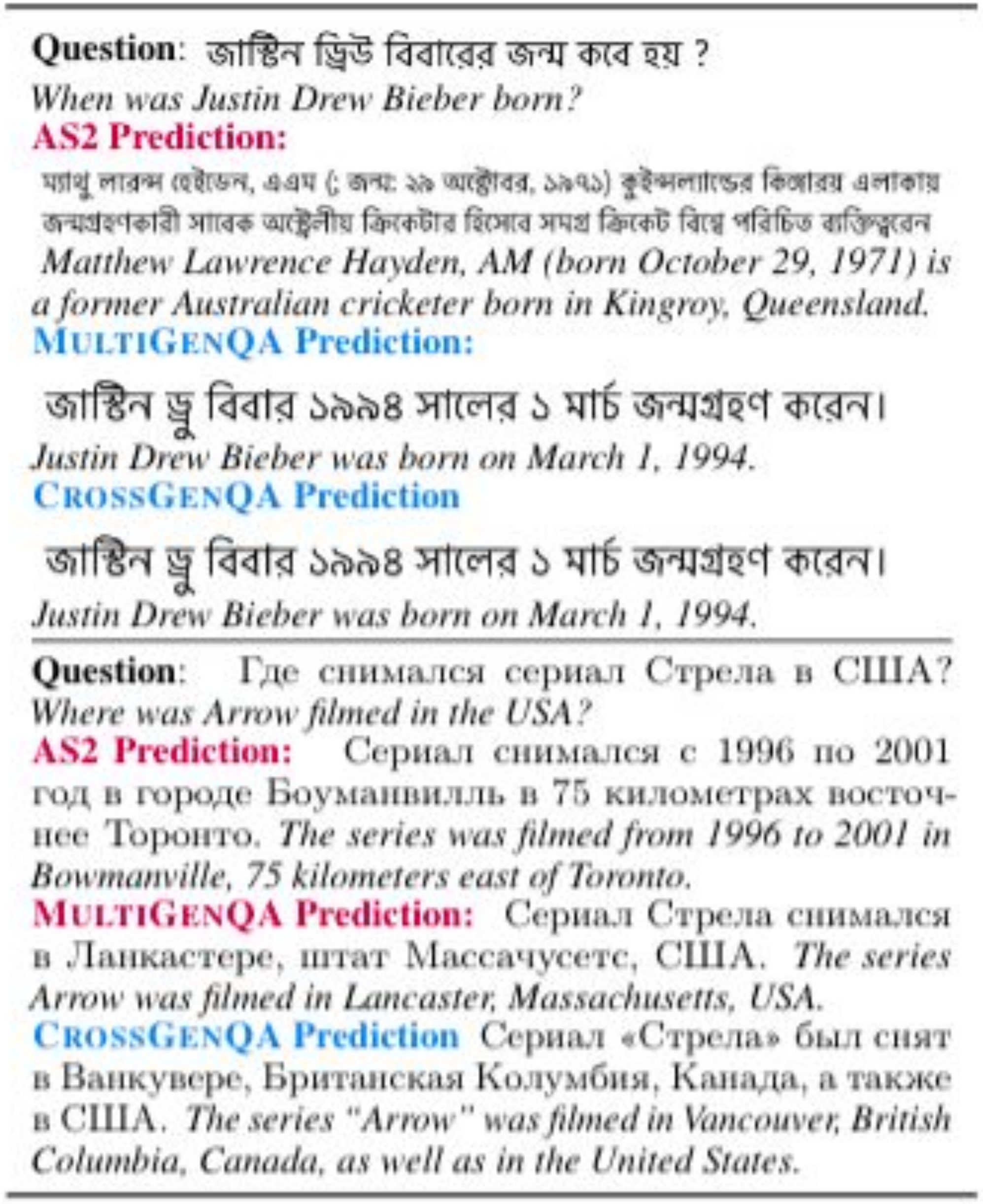
Two Aggregation Strategies
- Top 2 Candidates per Language
- Top 10 languages over all the languages

Results



- Cross-GenQA outperforms the baselines for Arabic, Japanese and Russian
- Why is it failing for English and Bengali?

Predictions Samples



Question: জাস্টিন ড্রিউ বিবারের জন্ম কবে হয় ?
When was Justin Drew Bieber born?
AS2 Prediction:
ম্যাথু লারেন্স হেইডেন, এএম (; জন্ম: ২৯ অক্টোবর, ১৯৭১) কুইন্সল্যান্ডের কিঙ্গারয় এলাকায় জন্মগ্রহণকারী সাবেক অস্ট্রেলীয় ক্রিকেটার হিসেবে সমগ্র ক্রিকেট বিশ্বে পরিচিত ব্যক্তিত্বরেন
Matthew Lawrence Hayden, AM (born October 29, 1971) is a former Australian cricketer born in Kingroy, Queensland.
MULTIGENQA Prediction:
জাস্টিন ড্রু বিবার ১৯৯৪ সালের ১ মার্চ জন্মগ্রহণ করেন।
Justin Drew Bieber was born on March 1, 1994.
CROSSGENQA Prediction
জাস্টিন ড্রু বিবার ১৯৯৪ সালের ১ মার্চ জন্মগ্রহণ করেন।
Justin Drew Bieber was born on March 1, 1994.

Question: Где снимался сериал Стрела в США? *Where was Arrow filmed in the USA?*
AS2 Prediction: Сериал снимался с 1996 по 2001 год в городе Боуманвилль в 75 километрах восточнее Торонто. *The series was filmed from 1996 to 2001 in Bowmanville, 75 kilometers east of Toronto.*
MULTIGENQA Prediction: Сериал Стрела снимался в Ланкастере, штат Массачусетс, США. *The series Arrow was filmed in Lancaster, Massachusetts, USA.*
CROSSGENQA Prediction Сериал «Стрела» был снят в Ванкувере, Британская Колумбия, Канада, а также в США. *The series "Arrow" was filmed in Vancouver, British Columbia, Canada, as well as in the United States.*

Impact of Cultural Gap on CrossGenQA

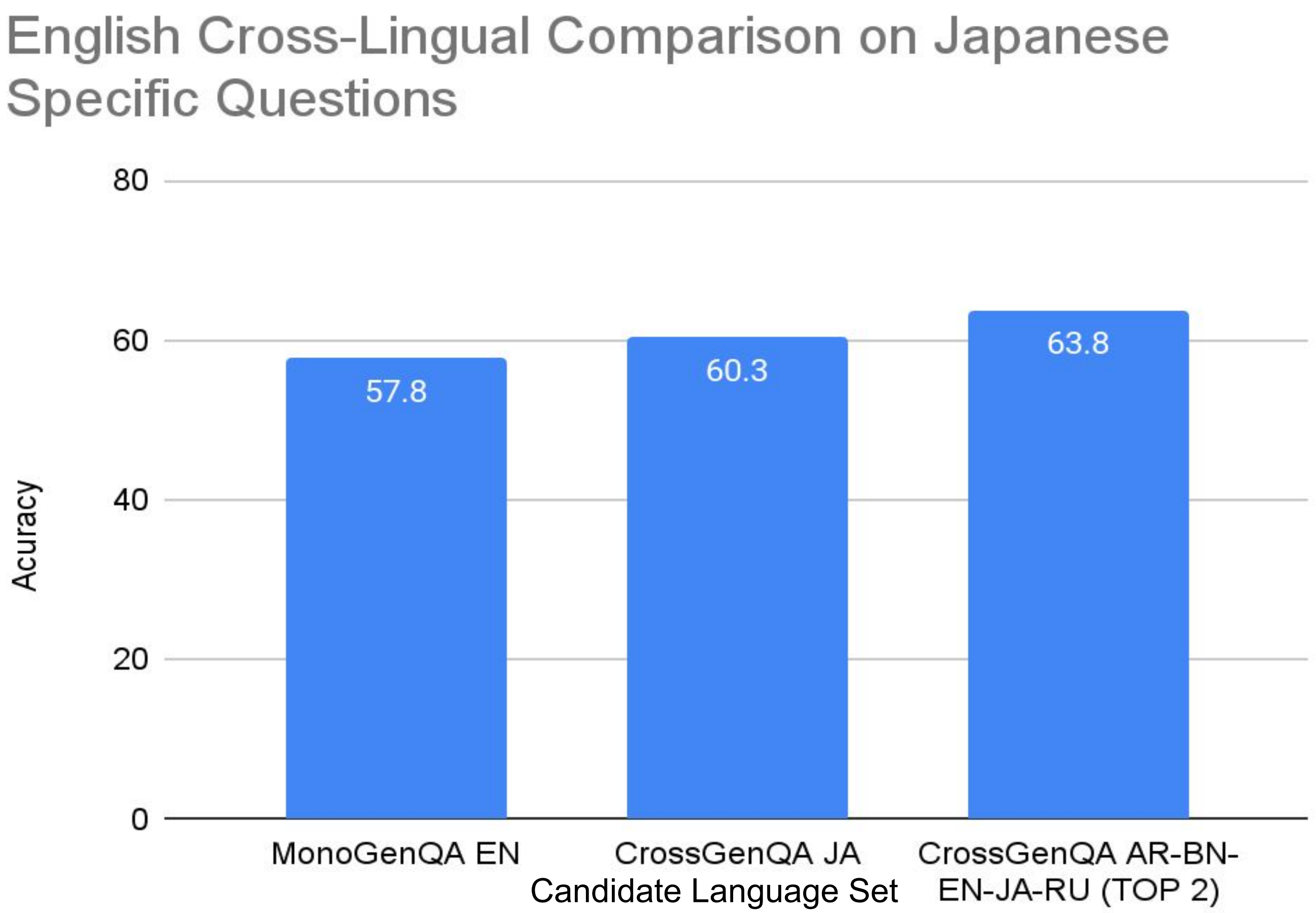
Why is CrossGenQA underperforming on English?

Hypothesis English Questions could be more English-specific than other languages

Experiment We run the same experiment on questions specific to the Japanese culture in English

Example
What was the first head of the Communist Party of Japan?
When was Prince Mochihito born?

Results



➔ **Cross-Lingual GenQA approaches outperform Monolingual GenQA in English** for Japanese culture-specific Questions

Conclusion and Takeaways

- **We release** a new dataset for evaluating GenQA Models on **Arabic, Bengali, Russian, Japanese and English**
- **Open-Retrieval Generative QA** can be made **Multilingual** without collecting training data for all the languages
- **It can provide answers** to questions **that have information in languages different from the user language**
- **Our approach** can be used to **open-up culturally centric open QA models**

Thank you!

Bibliography

- XOR QA: Cross-lingual Open-Retrieval Question Answering [Asai et. al 2020]
- Answer Generation for Retrieval-based Question Answering Systems [Hsu et. al 2021]
- Multilingual Answer Sentence Reranking via Automatically Translated Data [Vu and Moschitti 2021]
- mT5: A massively multilingual pre-trained text-to-text transformer [Xue et. al. 2020]
- TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages [Clark et. al. 2020]