L3S

Trustworthy AI
& Digital Transformation

TIGQA: An Expert-Annotated Question-Answering Dataset in Tigrinya

Hailay Kidu Teklehaimanot, Dren Fazlija, Niloy Ganguly, Gourab K.Patro, Wolfgang Nejdl

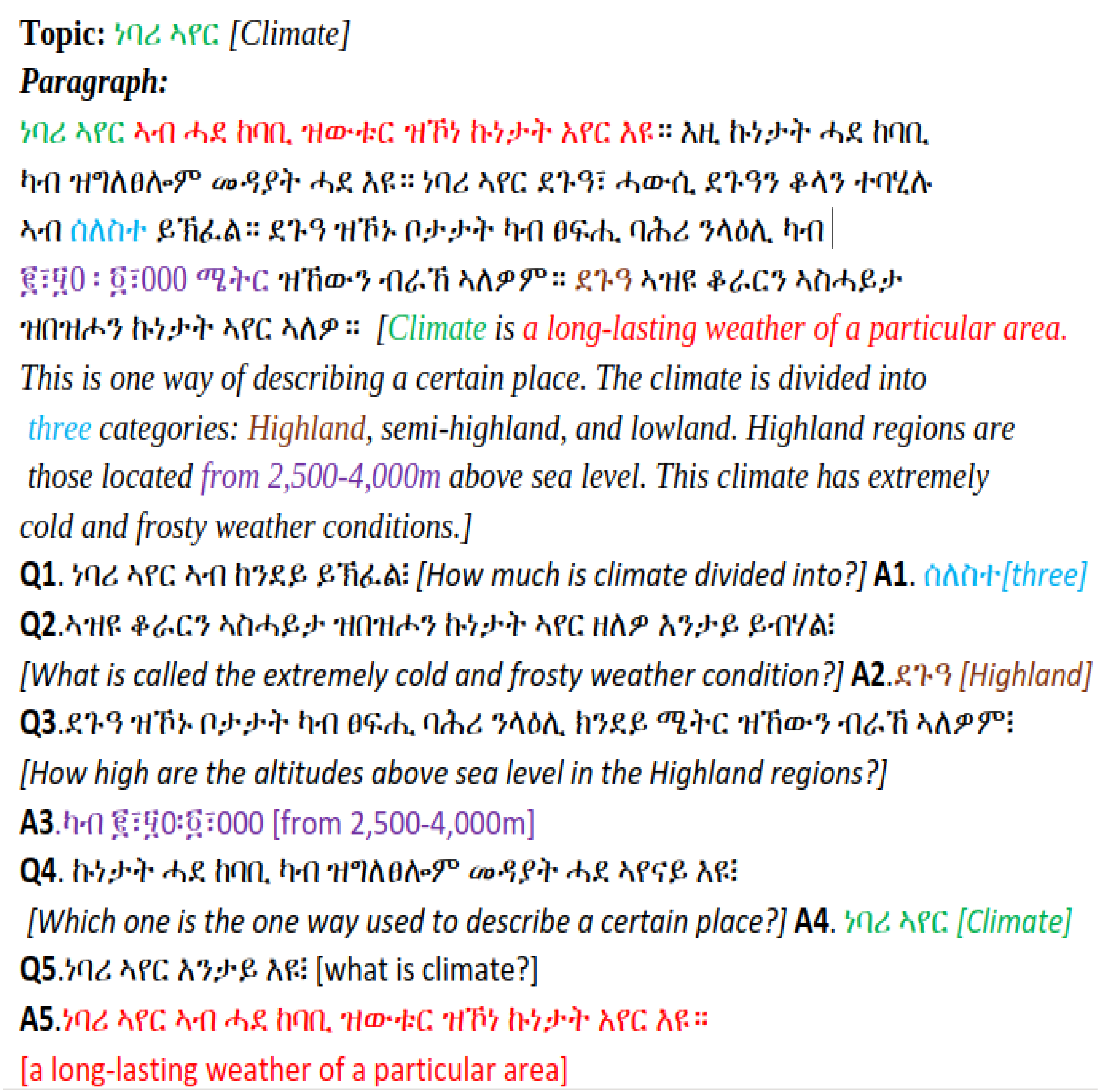
Introduction

- MRC aims to **create systems that answer questions based on their understanding of one or more given documents** (Lai et al., 2017).

- Significant advances in MRC systems focus **on high-resource languages such as English.**

- Most languages with **limited resources remain untapped** due to a lack of computational resources and **annotated datasets** (Tonja et al., 2023).

- We create the first expert annotated domain-specific dataset for Tigrinya , an east African languages spoken by more than 10 m , in Ethiopia and Eretria



Topic: ነባሪ ኣየር *[Climate]*

Paragraph:

ነባሪ ኣየር ኣብ ሓደ ከባቢ ዝውቱር ዝኾነ ኩነታት አየር እዩ። እዚ ኩነታት ሓደ ከባቢ ካብ ዝግለፅሉም መዳያት ሓደ እዩ። ነባሪ ኣየር ደጉዓ፣ ሓውሲ ደጉዓን ቆላን ተባሂሉ ኣብ ሰለስተ ይኽፈል። ደጉዓ ዝኾኑ ቦታታት ካብ ፀፍሒ ባሕሪ ንላዕሊ ካብ ፪፣፶0 ፡ ፬፣000 ሜትር ዝኸውን ብራኸ ኣለዎም። ደጉዓ ኣዝዩ ቆራርን ኣስሓይታ ዝበዝሖን ኩነታት ኣየር ኣለዎ። *[Climate is a long-lasting weather of a particular area. This is one way of describing a certain place. The climate is divided into three categories: Highland, semi-highland, and lowland. Highland regions are those located from 2,500-4,000m above sea level. This climate has extremely cold and frosty weather conditions.]*

Q1. ነባሪ ኣየር ኣብ ክንደይ ይኽፈል፧ *[How much is climate divided into?]* **A1**. ሰለስተ*[three]*

Q2.ኣዝዩ ቆራርን ኣስሓይታ ዝበዝሖን ኩነታት ኣየር ዘለዎ እንታይ ይብሃል፧ *[What is called the extremely cold and frosty weather condition?]* **A2**.ደጉዓ *[Highland]*

Q3.ደጉዓ ዝኾኑ ቦታታት ካብ ፀፍሒ ባሕሪ ንላዕሊ ክንደይ ሜትር ዝኸውን ብራኸ ኣለዎም፧ *[How high are the altitudes above sea level in the Highland regions?]*

A3.ካብ ፪፣፶0፡፬፣000 [from 2,500-4,000m]

Q4. ኩነታት ሓደ ከባቢ ካብ ዝግለፅሉም መዳያት ሓደ ኣየናይ እዩ፧ *[Which one is the one way used to describe a certain place?]* **A4**. ነባሪ ኣየር *[Climate]*

Q5.ነባሪ ኣየር እንታይ እዩ፧ [what is climate?]

A5.ነባሪ ኣየር ኣብ ሓደ ከባቢ ዝውቱር ዝኾነ ኩነታት አየር እዩ።
[a long-lasting weather of a particular area]

Figure 1: Examples of an expert annotated educational domain QA in Tigrinya (TIGQA) contexts are also a part of the dataset and answers are highlighted. The translation is in italic

Cont...

- We investigate MRC datasets in other low-resource languages:

- Most works sourced their data from Wikipedia or translated from the SQuAD
- Such dataset creation affects the dataset quality for the following reasons:
  1. There are few Wikipedia open-source contributors in low-resource languages like Tigrinya, and their relevance and authenticity can be uncertain.
  2. Machine translations have quality issues, especially when the target language is low-resource like Tigrinya.

  Both issues warrant further research:

Contribution

- **Evaluation of machine translation (MT) models in dataset creation.**
- **Present the first expert annotated domain-specific dataset for Tigrinya.**
- **In depth_analysis such as (statistical, comparative, length, size and type).**
- **Evaluate human performance and the challenges in TIGQA.**
- **Finally, we experiment by training and fine-tuning transformer-based models**

Overall TIGQA Dataset Collection

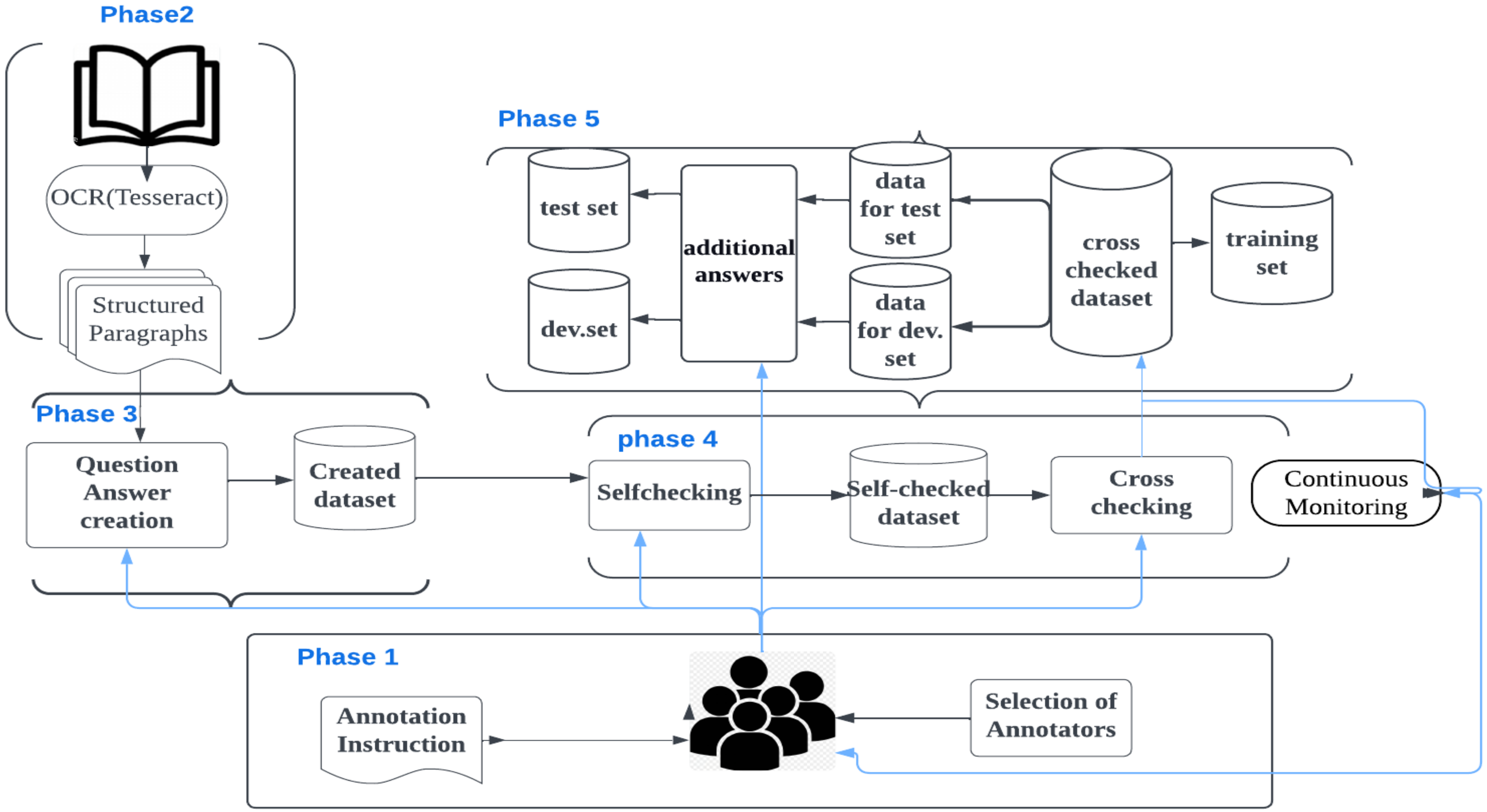
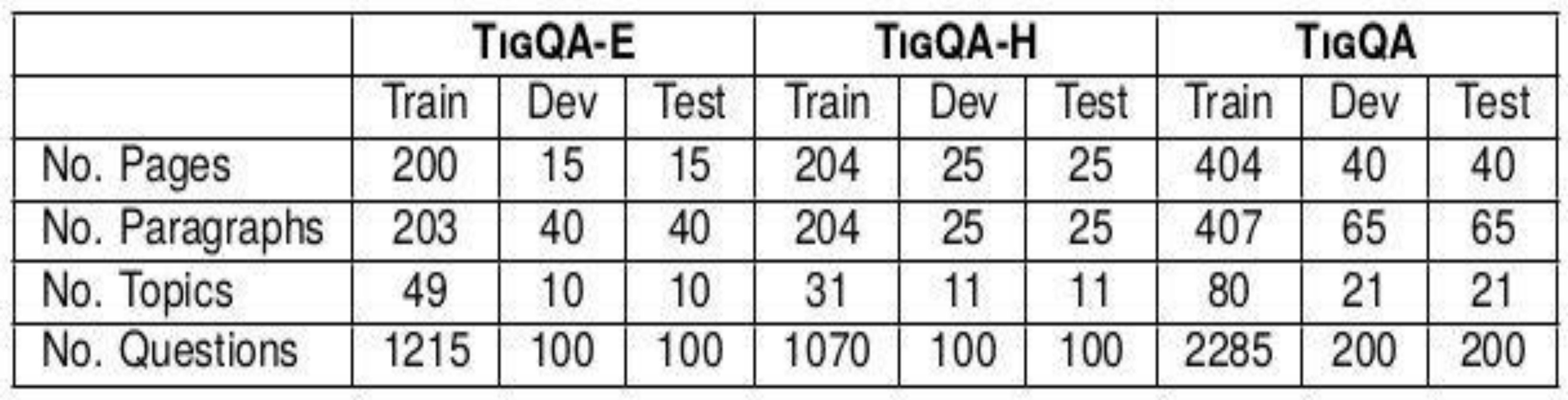


Figure 1: The overview process of creatingTIGQA dataset

Cont…

- Finally, we extracted **455** pages and **537** paragraphs from **122** diverse topics, including Climate, Social, culture, history, health, business, etc.

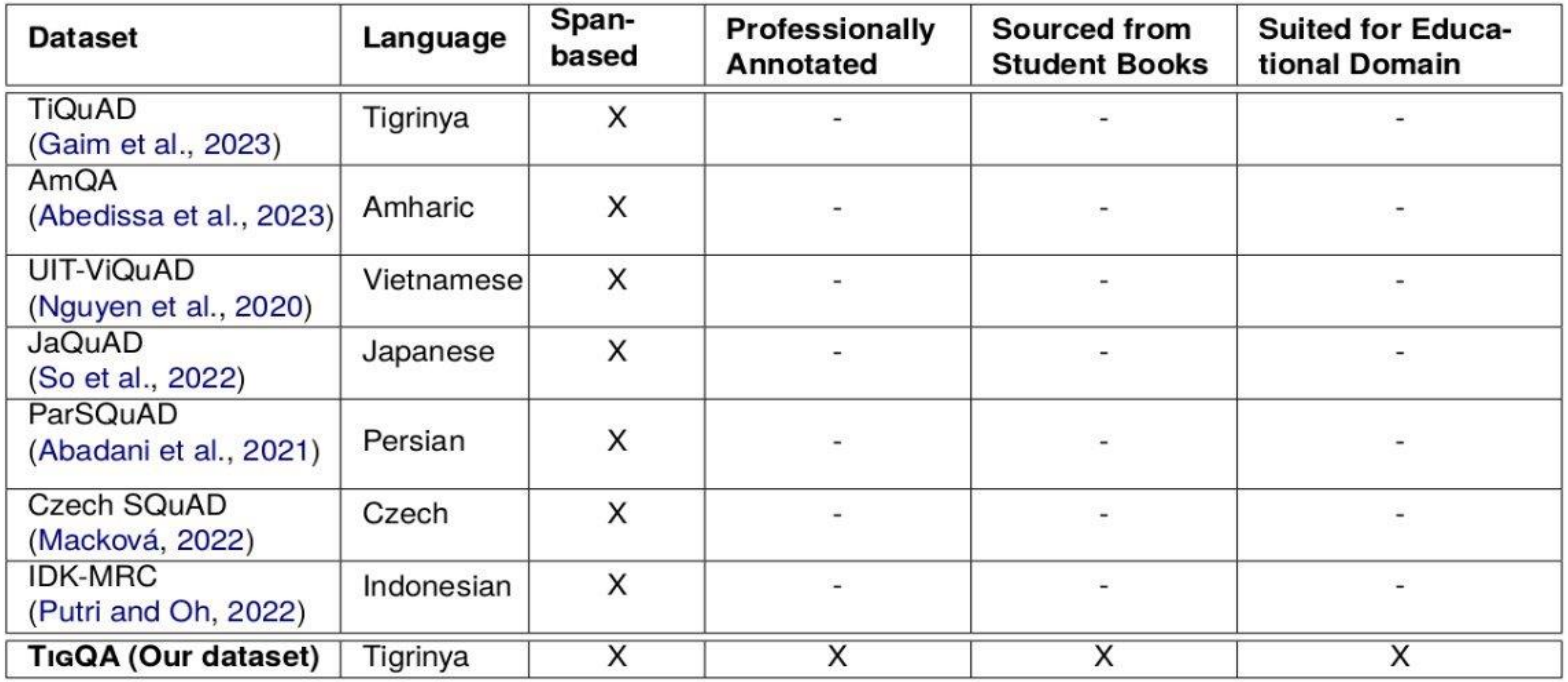
- We further module into TIGQA-E and TIGQA-H



	TIGQA-E			TIGQA-H			TIGQA		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
No. Pages	200	15	15	204	25	25	404	40	40
No. Paragraphs	203	40	40	204	25	25	407	65	65
No. Topics	49	10	10	31	11	11	80	21	21
No. Questions	1215	100	100	1070	100	100	2285	200	200

Table 1: TIGQA dataset statistics

Cont…



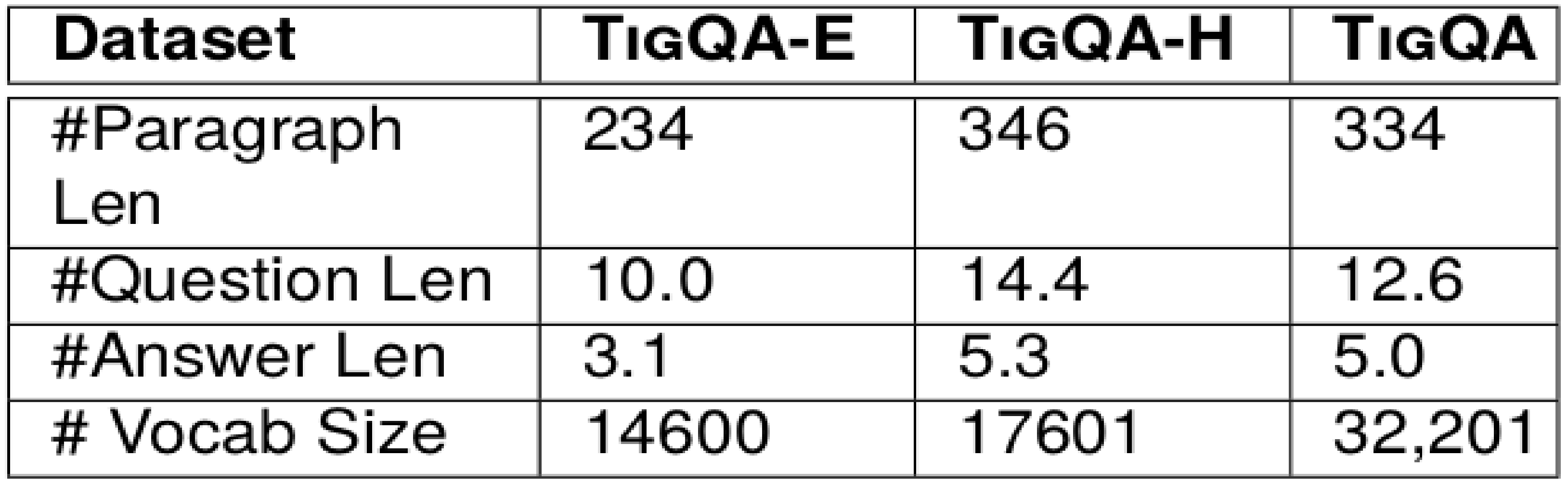
Dataset	Language	Span-based	Professionally Annotated	Sourced from Student Books	Suited for Educational Domain
TiQuAD (Gaim et al., 2023)	Tigrinya	X	-	-	-
AmQA (Abedissa et al., 2023)	Amharic	X	-	-	-
UIT-ViQuAD (Nguyen et al., 2020)	Vietnamese	X	-	-	-
JaQuAD (So et al., 2022)	Japanese	X	-	-	-
ParSQuAD (Abadani et al., 2021)	Persian	X	-	-	-
Czech SQuAD (Macková, 2022)	Czech	X	-	-	-
IDK-MRC (Putri and Oh, 2022)	Indonesian	X	-	-	-
TIGQA (Our dataset)	Tigrinya	X	X	X	X

Table 6: Comparison of TIGQA with existing low-resource MR/QA datasets. Our dataset is unique because it is fully annotated by experts, which is suited for educational domains and contextually and culturally relevant to the local use cases; others use Wikipedia and news articles as sources and employ crowd workers.

This signifies that our dataset is exceptional and represents the first instance of subject matter experts' annotation in the low-resource language Tigrinya.

Analyses

1. Length, Size and Vocabulary analyze



Dataset	TigQA-E	TigQA-H	TigQA
#Paragraph Len	234	346	334
#Question Len	10.0	14.4	12.6
#Answer Len	3.1	5.3	5.0
# Vocab Size	14600	17601	32,201

Table 2: Statistics of TigQA where Len denotes length and Vocab denotes Vocabulary TigQA

2. Question Type Analyses

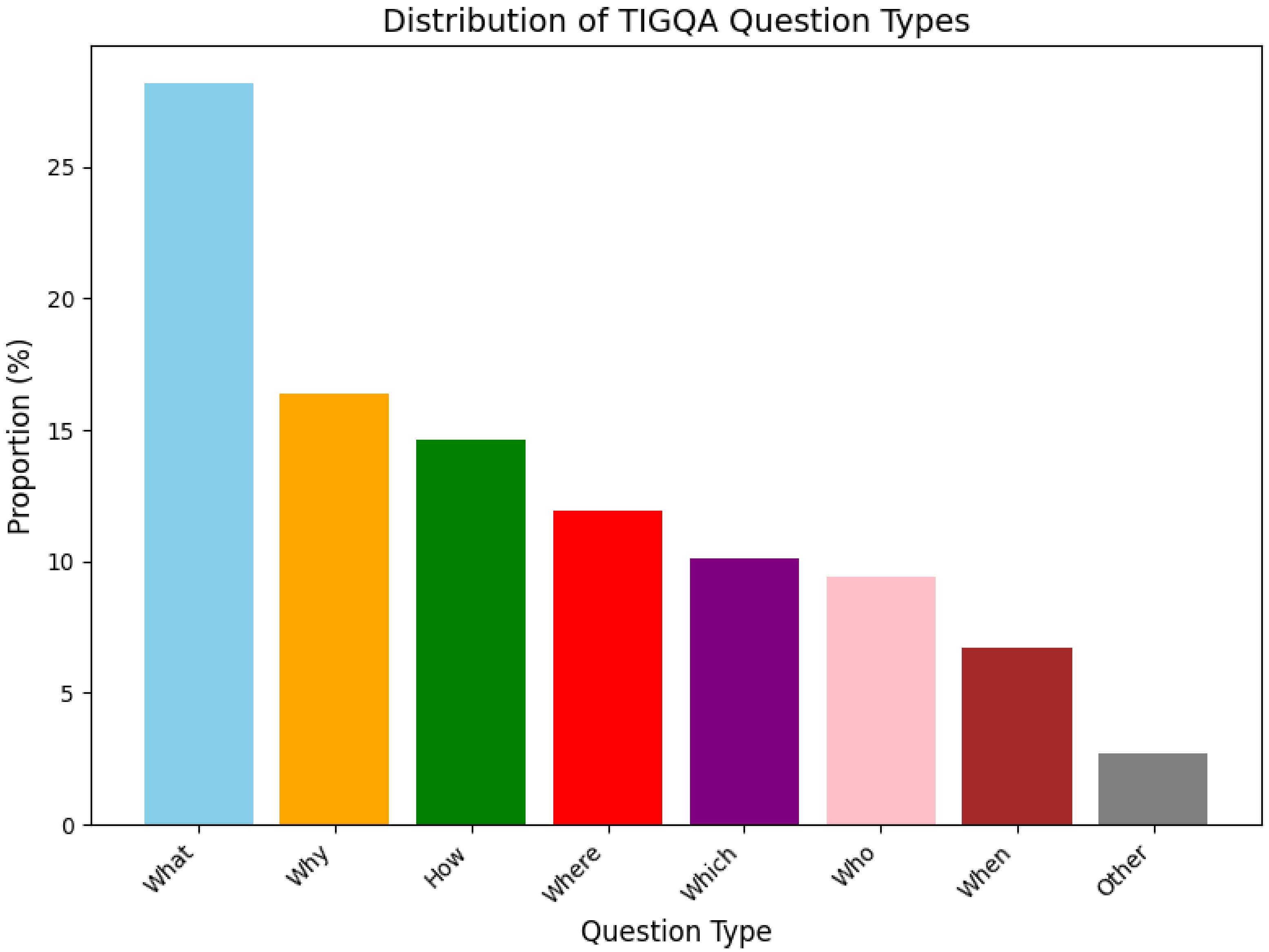
3. Reasoning Types Analyses

Vocabulary size is a higher portion in TIGQA-H than in TIGQA-E -> this indicates that TIGQA-H requires **more reasoning-type questions** based on difficulty at each grade level.

2. Question Type Analyses

What	እንታይ	28.2%	ኣብ ሓደ ከባቢ ዙውቱር ዝኾነ ከባቢ ኣየር እንታይ እዩ፧(what do we call the long-lasting weather of a particular area?)
Why	ንምንታይ፤ ስለምንታይ	16.4%	ተስፋይ ንምንታይ ትምህርቲ ከቋርፅ ደልዩ፧[Why did Tesfay want to quit school?]
How	ከመይ ፤ከንደይ	14.6%	ሕምም ዓሶ ካብ ሰብ ናብ ሰብ ኸመይ ይመሓላለፍ፧ [How is tuberculosis transmitted from person to person?]
Where	ኣበይ ፤ናበይ፤ካበይ	11.9%	ናይ ዓለምና ኦሎምፒክ ፳፳ ኣበይ ተሳሊጡ፧[Where was the 2020 World Olympics held?]
Which	ኣየናይ ፤ኣየነይቲ	10.1%	ኣየናይ እንስሳ እዩ ነቲ ፓርክ ፍሉይ ድምቀት ዝህቦ፧[Which animal gives the park a special brightness]
Who	መን	9.4%	ናይቲ ትምህርቲ ቤት ርእሰ መምህር መን ይብሃሉ ፧[Who is the principal of the school ?]
When	መዓዝ፤መኣዝ	6.7%	እቲ ፈተና መዓስ እዩ ዝጅምር፧[When does the test start?]
other	ጥቀስ	2.7%	መንግስቲ ኣብዚ ሰሙን ካብ ዝገበሮም ስምምዕነት ዝተወሰኑ ጥቀስ፧ [Name some of the agreements the government made this week?]

Figure 2: Question type distribution in TIGQA dataset: grouped by interrogative words. The highlighted color implies the interrogative words in Tigrinya.



Note that: the expected answer types are beyond proper noun entities.

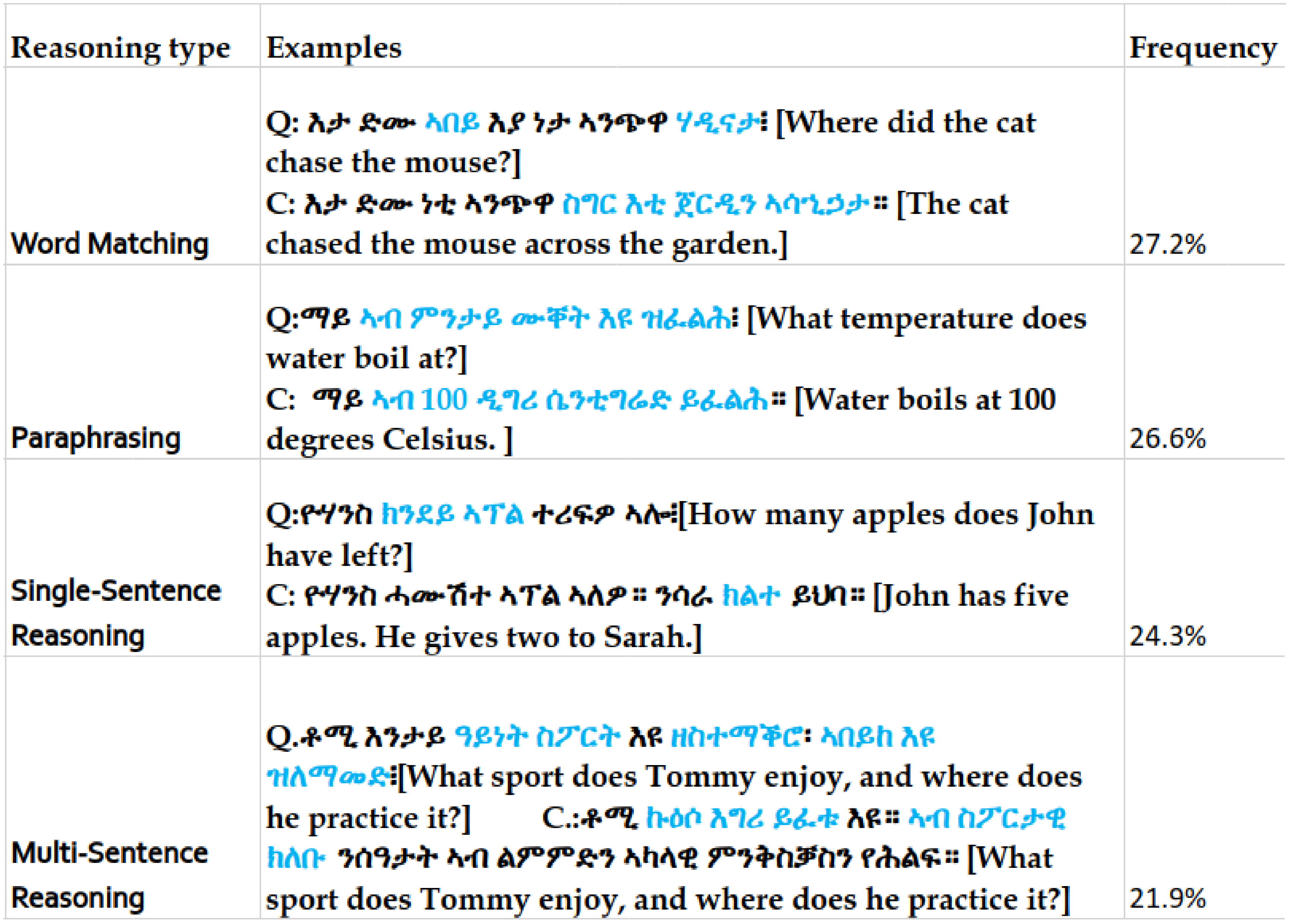
Cont...

- The dominant question types: **What** (እንታይ) and **Why** (ንምንታይ፤ ስለምንታይ)
  - Collectively comprising **59.2%** of all questions

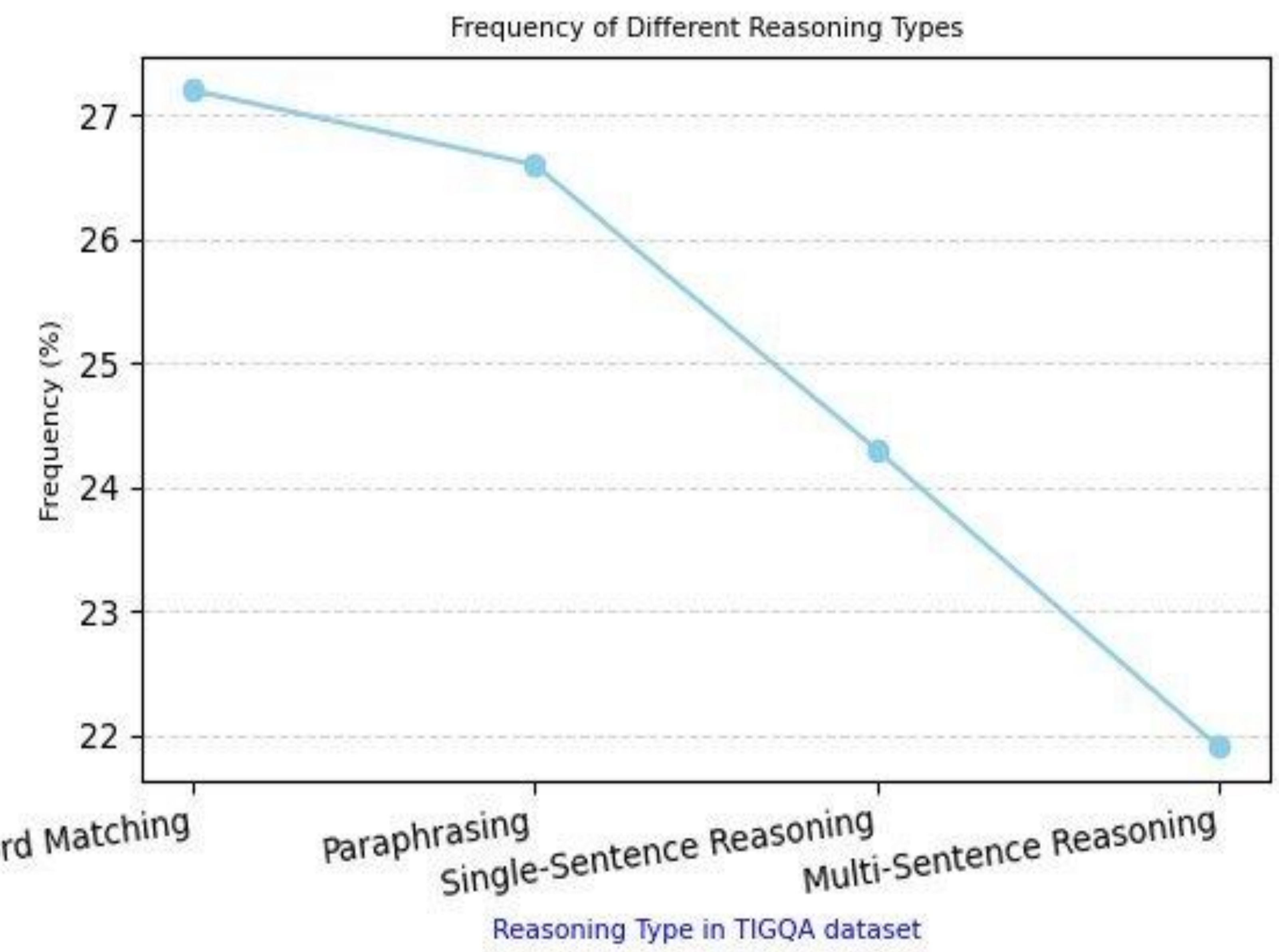
- Answering the **predominant** questions **what** and **why** requires a **profound** comprehension of the **rhetorical** structure and nuanced descriptions.

- Responses to such questions typically involve **entire clauses**, rather than mere phrases embedded within a context.

3. Reasoning Types Analyses



Reasoning type	Examples	Frequency
Word Matching	Q: እታ ድሙ ኣበይ እያ ነታ ኣንጭዋ ሃዲናታ፧ [Where did the cat chase the mouse?] C: እታ ድሙ ነቲ ኣንጭዋ ስግር እቲ ጀርዲን ኣሳጒጻታ። [The cat chased the mouse across the garden.]	27.2%
Paraphrasing	Q:ማይ ኣብ ምንታይ ሙቐት እዩ ዝፈልሕ፧ [What temperature does water boil at?] C: ማይ ኣብ 100 ዲግሪ ሴንቲግሬድ ይፈልሕ። [Water boils at 100 degrees Celsius.]	26.6%
Single-Sentence Reasoning	Q:ዮሃንስ ክንደይ ኣፕል ተሪፍዎ ኣሎ፧[How many apples does John have left?] C: ዮሃንስ ሓሙሽተ ኣፕል ኣለዎ። ንሳራ ክልተ ይህብ። [John has five apples. He gives two to Sarah.]	24.3%
Multi-Sentence Reasoning	Q.ቶሚ እንታይ ዓይነት ስፖርት እዩ ዘስተማቕሮ፡ ኣበይከ እዩ ዝለማመድ፧[What sport does Tommy enjoy, and where does he practice it?] C.:ቶሚ ኩዕሶ እግሪ ይፈቱ እዩ። ኣብ ስፖርታዊ ክለቡ ንሰዓታት ኣብ ልምምድን ኣካላዊ ምንቅስቓስን የሕልፍ። [What sport does Tommy enjoy, and where does he practice it?]	21.9%



word matching is the most accessible type and is the most significant subset of our datasets (27.2%).

MT Error analyses

- We evaluate **auto** and **manual** translations then **categorize the errors.**

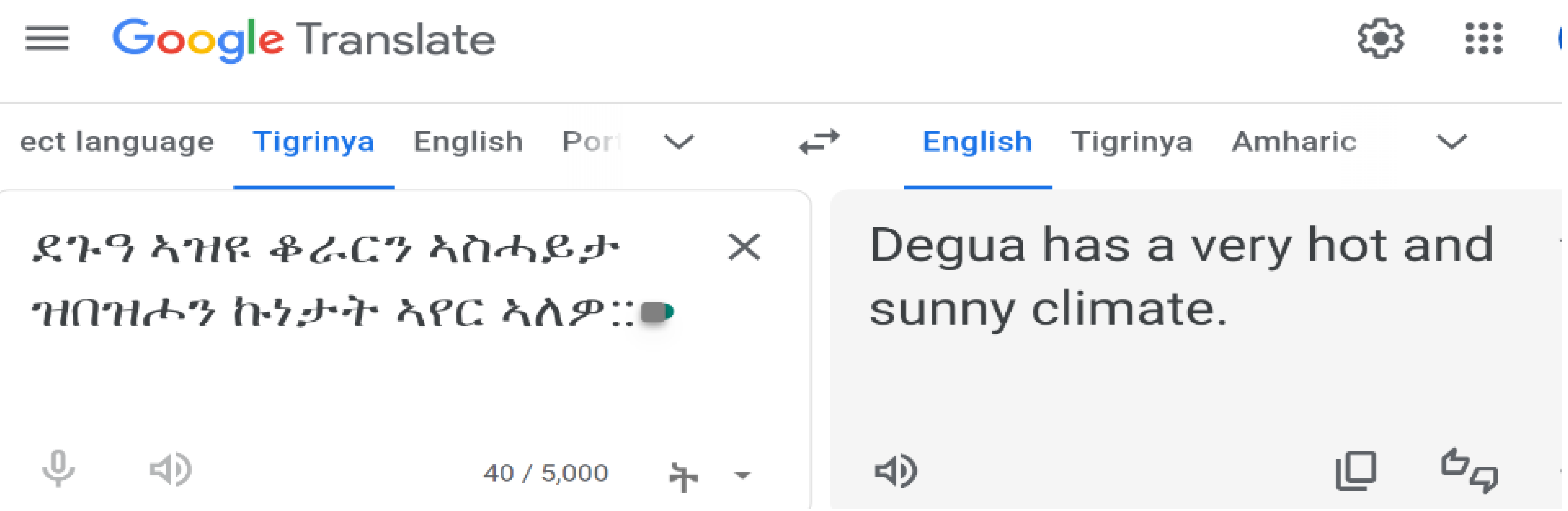
Sample of **150 triplets extracted** from SQuAD and **50 Tigrinya triplets** manually created following the SQuAD format from the student textbook.

- **<u>Goal:</u>** translate these pairs using three publicly available MT systems

Errors classified: **Mistranslation, Omission, Untranslated**

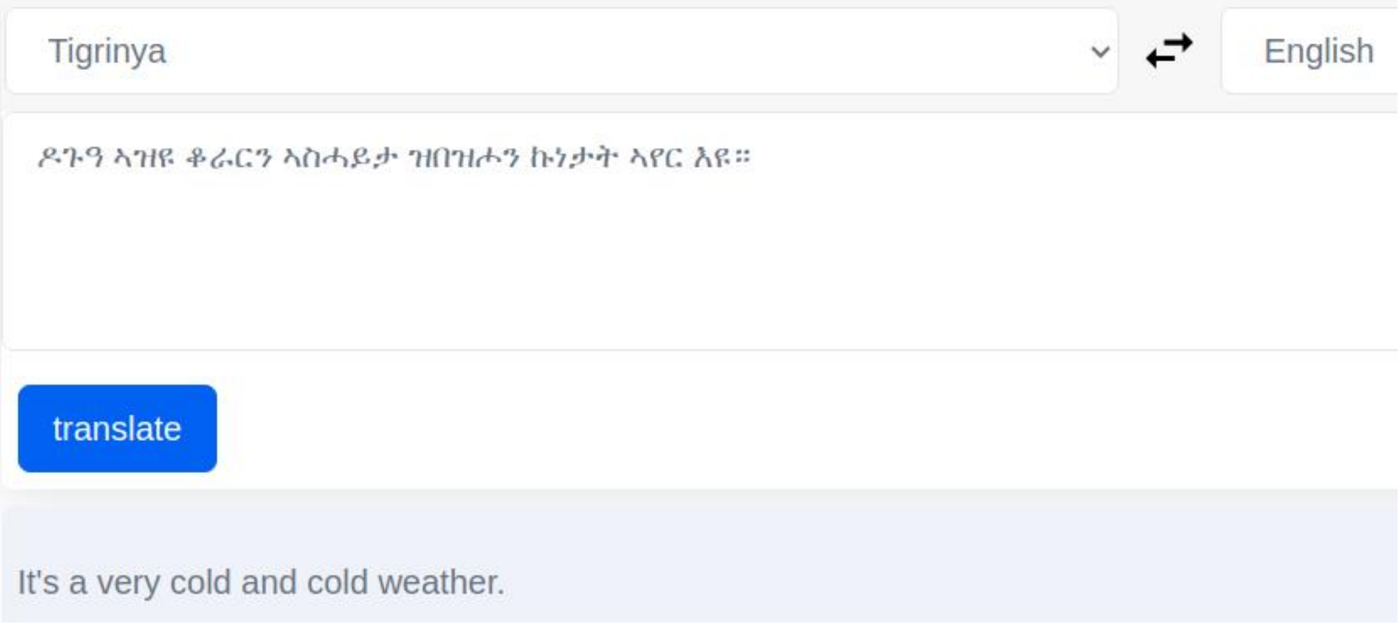
Results: Tigrinya to English

- Human Translation: **Highland has extremely cold and frosty weather conditions.**



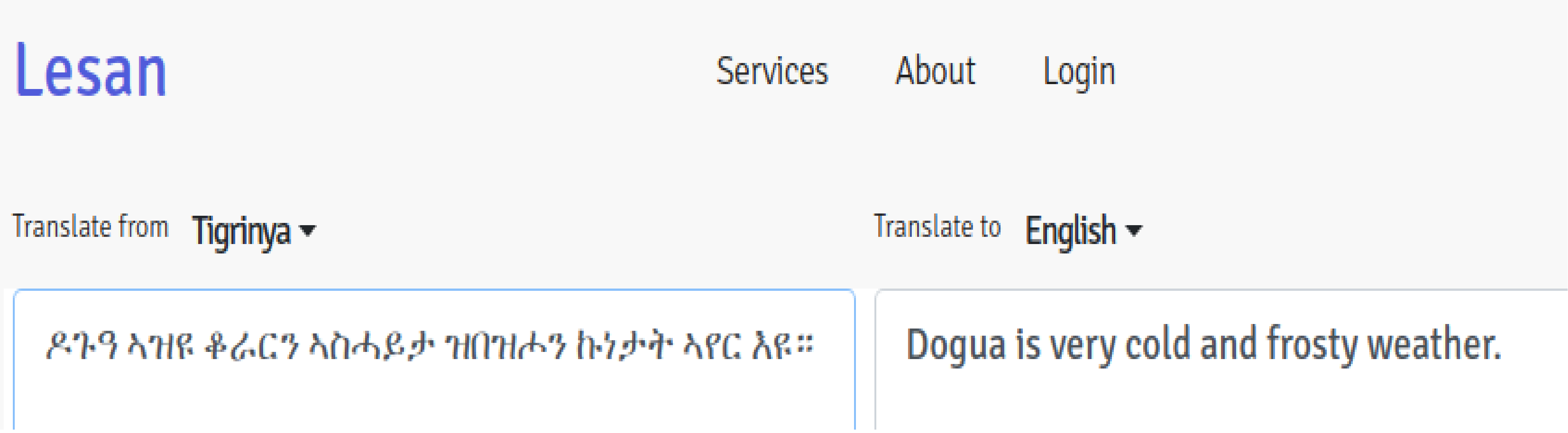
ደጉዓ ኣዝዩ ቆራርን ኣስሓይታ ዝበዝሖን ኩነታት ኣየር ኣለዎ::

Degua has a very hot and sunny climate.



ዶጉዓ ኣዝዩ ቆራርን ኣስሓይታ ዝበዝሖን ኩነታት ኣየር እዩ።

It's a very cold and cold weather.



ዶጉዓ ኣዝዩ ቆራርን ኣስሓይታ ዝበዝሖን ኩነታት ኣየር እዩ።

Dogua is very cold and frosty weather.

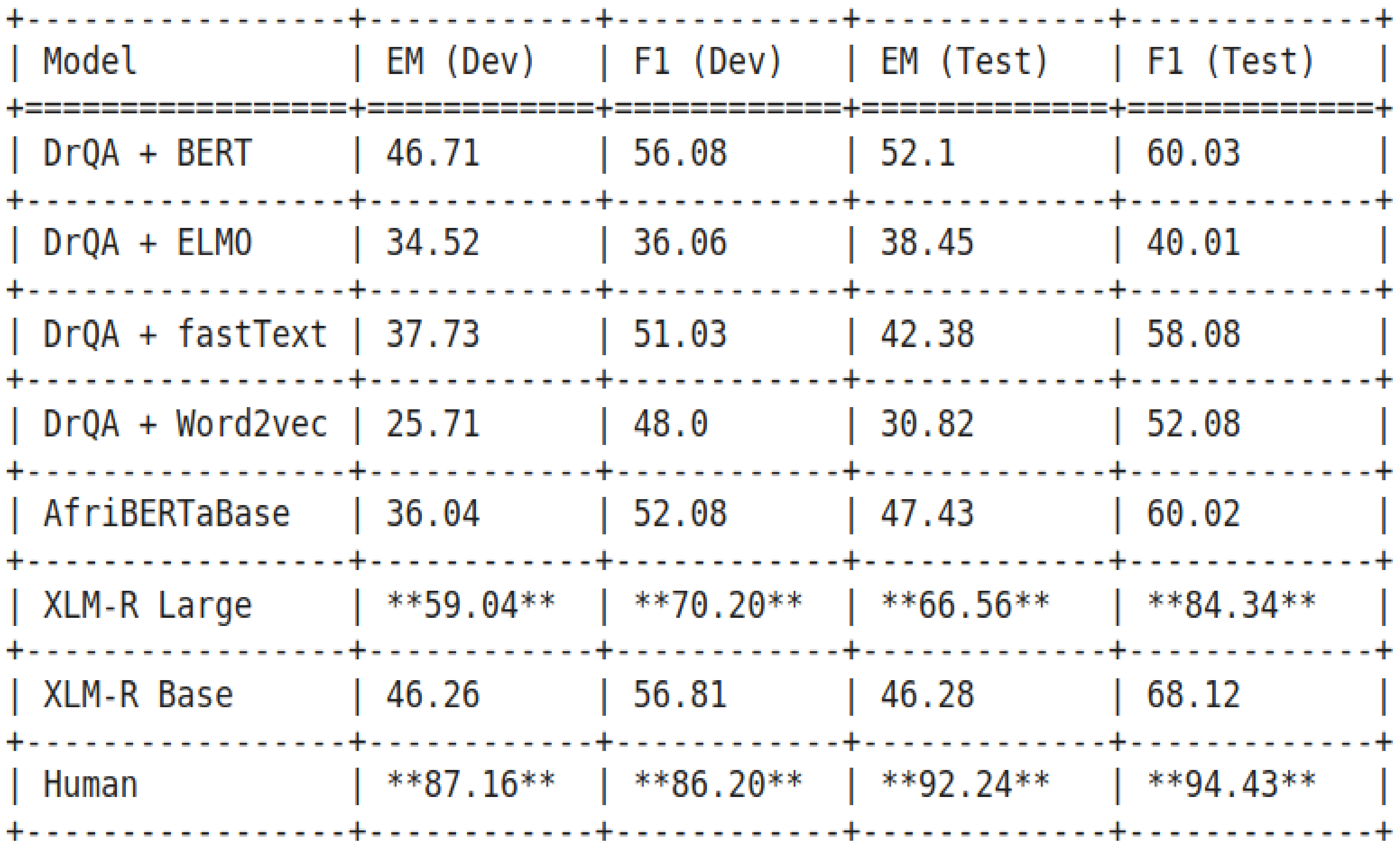
Experimental Settings and Evaluation

- Two evaluation metrics are used
    1. **Exact string match (EM)**
    2. **F1 score**,

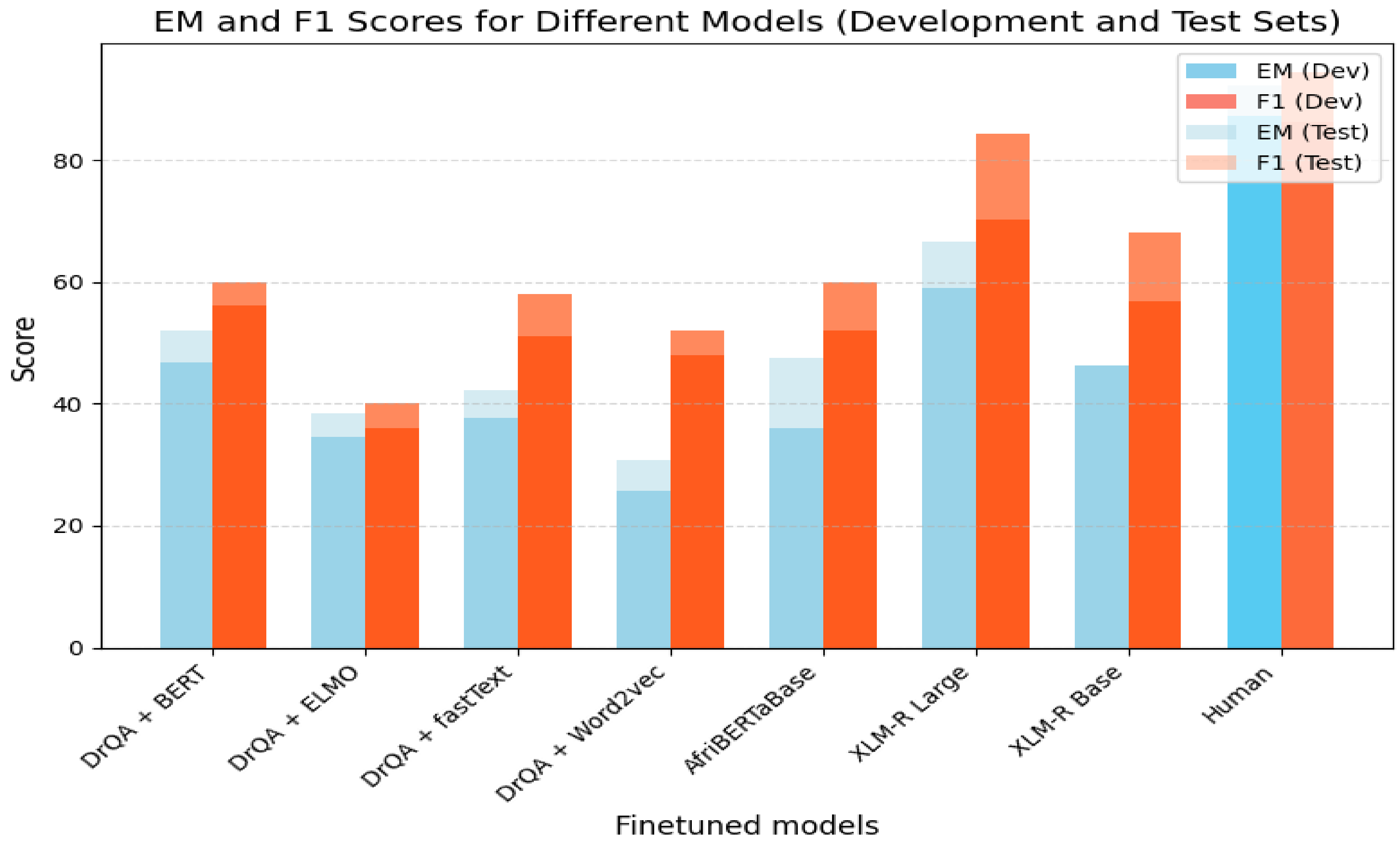
Baseline Models:

- **AfriBERTaBase** (Ogueji et al., 2021a): M**ultilingual pretraining language model**
- **DrQA** (Chen et al., 2017): A **neural network-based QA** model
- **XLM-R** (Conneau et al., 2020): A state-of-the-art **cross-lingual** model

Results



```
+-----------------+-------------+-------------+--------------+--------------+
| Model           | EM (Dev)    | F1 (Dev)    | EM (Test)    | F1 (Test)    |
+=================+=============+=============+==============+==============+
| DrQA + BERT     | 46.71       | 56.08       | 52.1         | 60.03        |
+-----------------+-------------+-------------+--------------+--------------+
| DrQA + ELMO     | 34.52       | 36.06       | 38.45        | 40.01        |
+-----------------+-------------+-------------+--------------+--------------+
| DrQA + fastText | 37.73       | 51.03       | 42.38        | 58.08        |
+-----------------+-------------+-------------+--------------+--------------+
| DrQA + Word2vec | 25.71       | 48.0        | 30.82        | 52.08        |
+-----------------+-------------+-------------+--------------+--------------+
| AfriBERTaBase   | 36.04       | 52.08       | 47.43        | 60.02        |
+-----------------+-------------+-------------+--------------+--------------+
| XLM-R Large     | **59.04**   | **70.20**   | **66.56**    | **84.34**    |
+-----------------+-------------+-------------+--------------+--------------+
| XLM-R Base      | 46.26       | 56.81       | 46.28        | 68.12        |
+-----------------+-------------+-------------+--------------+--------------+
| Human           | **87.16**   | **86.20**   | **92.24**    | **94.43**    |
+-----------------+-------------+-------------+--------------+--------------+
```



Result of Human and model performances on both Dev and Test sets of TIGQA dataset

Discussion

- The comparative performance of our models is compared against human performance in both dev. and test sets.

- Among the **neural network-based QA** models, **DrQA with BERT** embeddings consistently achieves the highest performance of EM 52.10% and F1 score 60.03% in test set, followed by **fast text** embeddings, while **ELMO** and **Word2Vec** embeddings show comparatively lower performance across both sets.

- Transformer-based models, particularly **XLM-R Large** and **XLM-R Base**, outperformed other configurations on both sets, demonstrating the effectiveness of cross-lingual pretraining.

- Human performance significantly **surpassed** all model-based approaches, highlighting the complexity of QA tasks in Tigrinya.

- Our experment shows the potential of **transformer-based models like XLM-R** for Tigrinya QA, with XLM-R Large achieving the highest **EM score of 66.56% and 84.34%** on the test set.

- Fine-tuning multilingual models remains a promising approach for low-resource languages, but further research is needed to address specific linguistic challenges and improve performance.

Cont...

- Regarding Exact Match (EM) and F1-core scores, XLM-R Large demonstrates significantly superior performance to the other models. Specifically, the model achieves an F1 score of **84.34%** on the test set. However, its Exact Match accuracy is **66.56%,** considerably lower than its F1 score.

- This suggests that while the model identifies relevant answers, it struggles to **precisely match human responses**, indicating need for improvement in aligning its outputs more closely with human performance benchmarks.

Summary & Limitations

- We present TIGQA, the first domain specific expert annotated dataset for low resource language Tigrinya
- Evaluate the quality of Machine Translation for dataset creation in low resource settings
- Estimate human performance on the dataset
- We demonstrate that annotated dataset significantly impacts the Machine reading comprehension and evaluation process

Limitations:

- TIGQA does not have adversarial questions to assess a model's ability to abstain
- The dataset can be further augmented to include adversarial examples and increase its size in the future.
- The dataset can be found at: https://github.com/hailaykidu/TigQA-Dataset

Contact

Hailay Kidu Teklehaymanot

teklehaymanot@l3s.de

www.l3s.de

Official GitHub Repository

@l3s_luh   @l3s-research-center   @forschungszentruml3s   @l3s-research-center