

doshisha_uzl at TRECVID 2023 MIQG Task: Medical Question Generation: Leveraging Vision- Language Summarization Models and Keyword Extraction with Flan-T5

Zihao Chen¹, Frédéric Li², Marc S. Seibel², Nele S. Brügge³, Miho Ohsaki¹
Heinz Handels^{2,3}, Marcin Grzegorzek², Kimiaki Shirahama¹

¹Doshisha University, ²Universität zu Lübeck

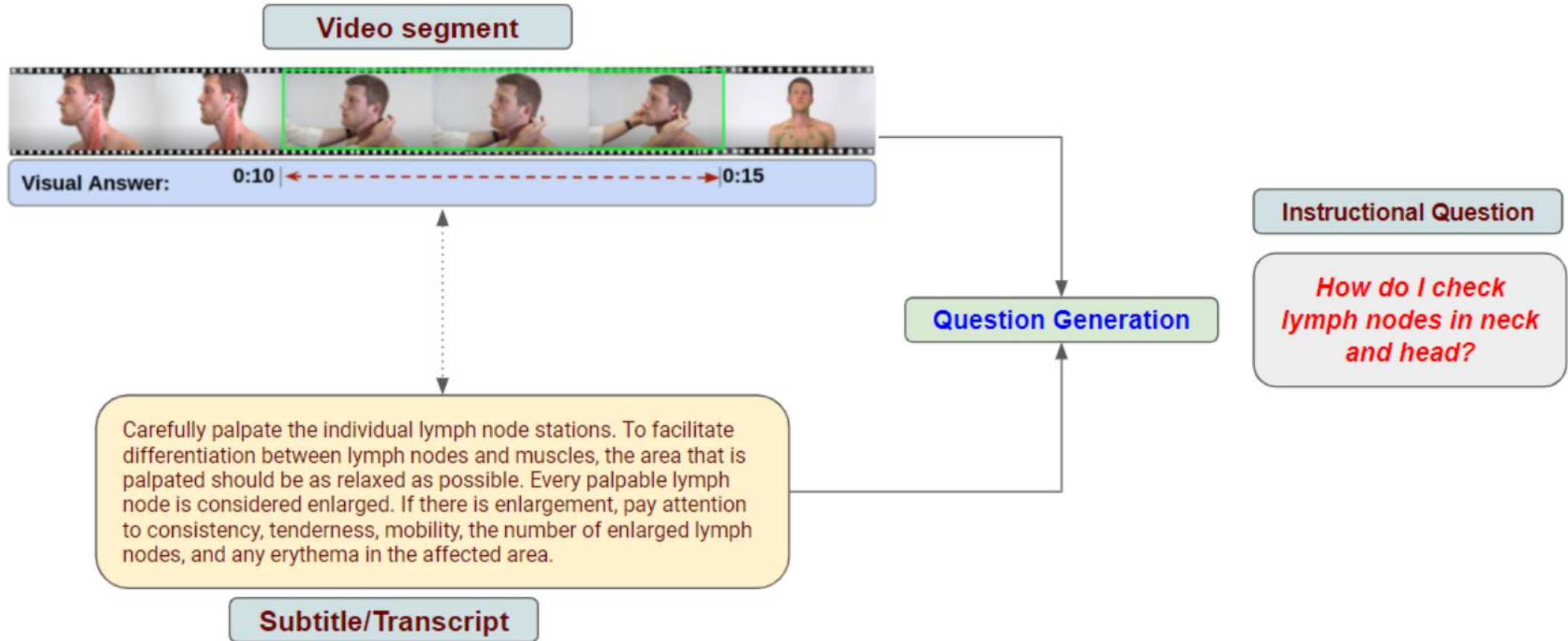
³German Research Center for Artificial Intelligence



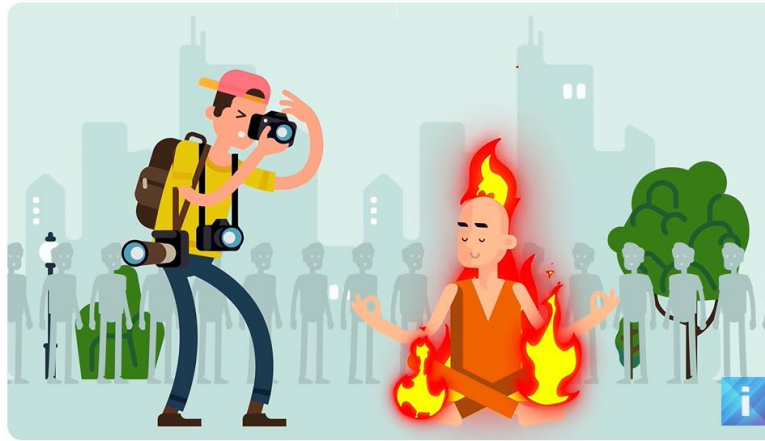
UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MEDICAL INFORMATICS



MIQG task overview



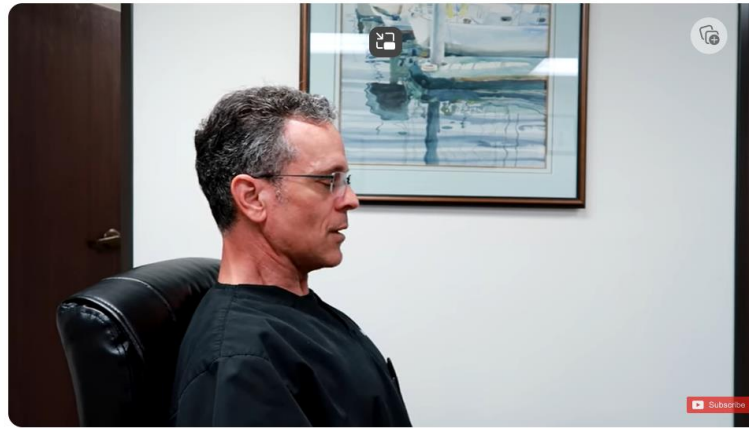
MIQG task dataset overview



How To Stop Any Pain In Minutes

animation

Video ID: 6wurtXEnBYw



Breathing Exercise: Watch what happens to my Oxygen levels when I do THIS...

No talking in the video, transcript contains meta-information.

Video ID: sFgDV68gIQY

Transcript: here we go.

Video ID: nT-P3rgI3Gg

I'm going into CPR now one two and three and four and five and six and seven and eight, seven, eight, nine and 10, 11 and 12 and 13 and 14 and 15, 16, 17, 18, 19, 20.

Video ID: R_GXQ77a9pU

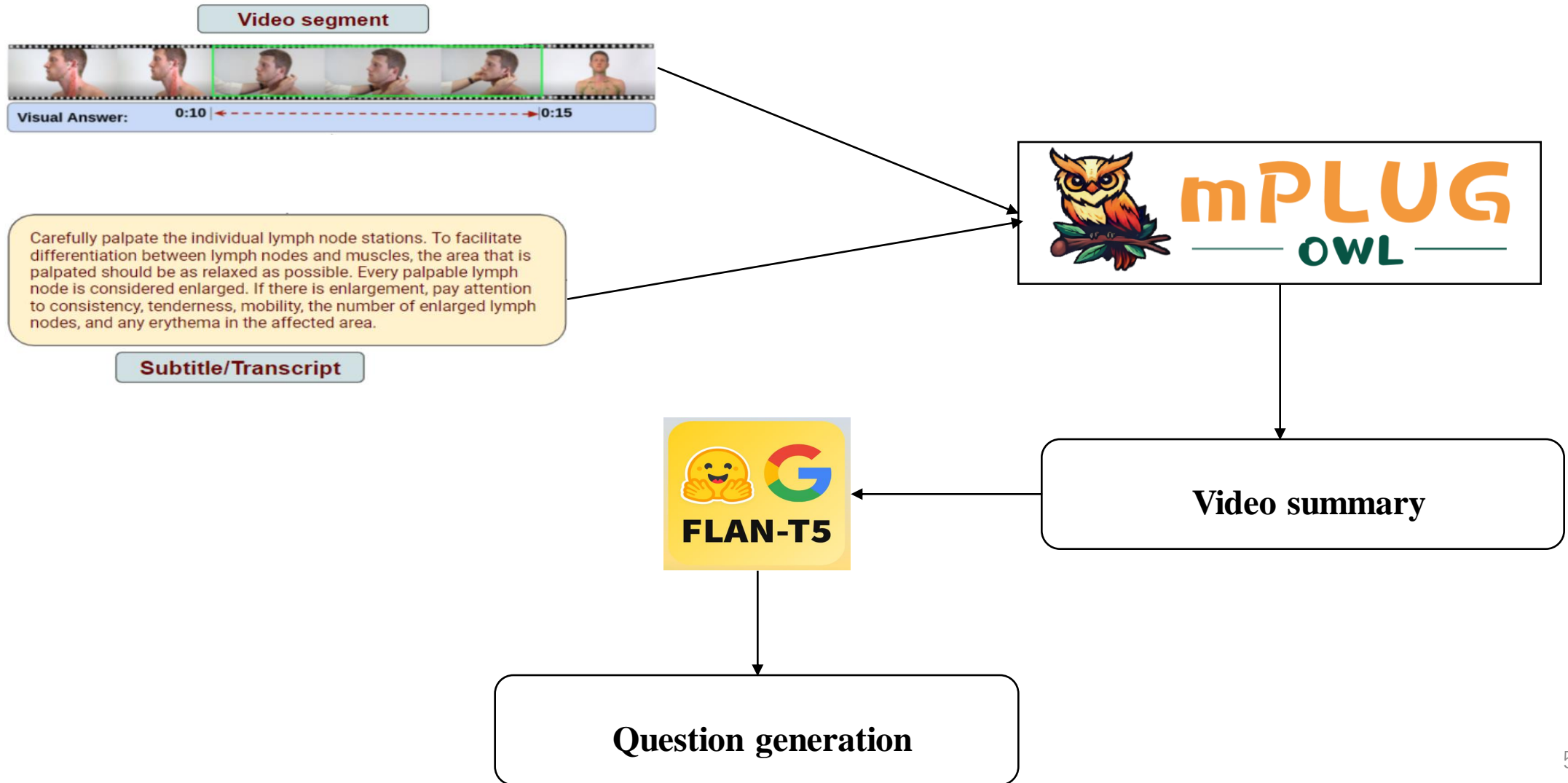
no transcript

We observe that for the provided dataset samples, either video or transcripts used separately may be insufficient on their own.

Jointly use both video and text modalities

- The most relevant content of some examples was for instance contained in the video only.
- While for others the text transcript contained information more susceptible to be useful for the generation task.

Medical Question Generation: Leveraging Vision-Language Summarization Models and Keyword Extraction with Flan-T5



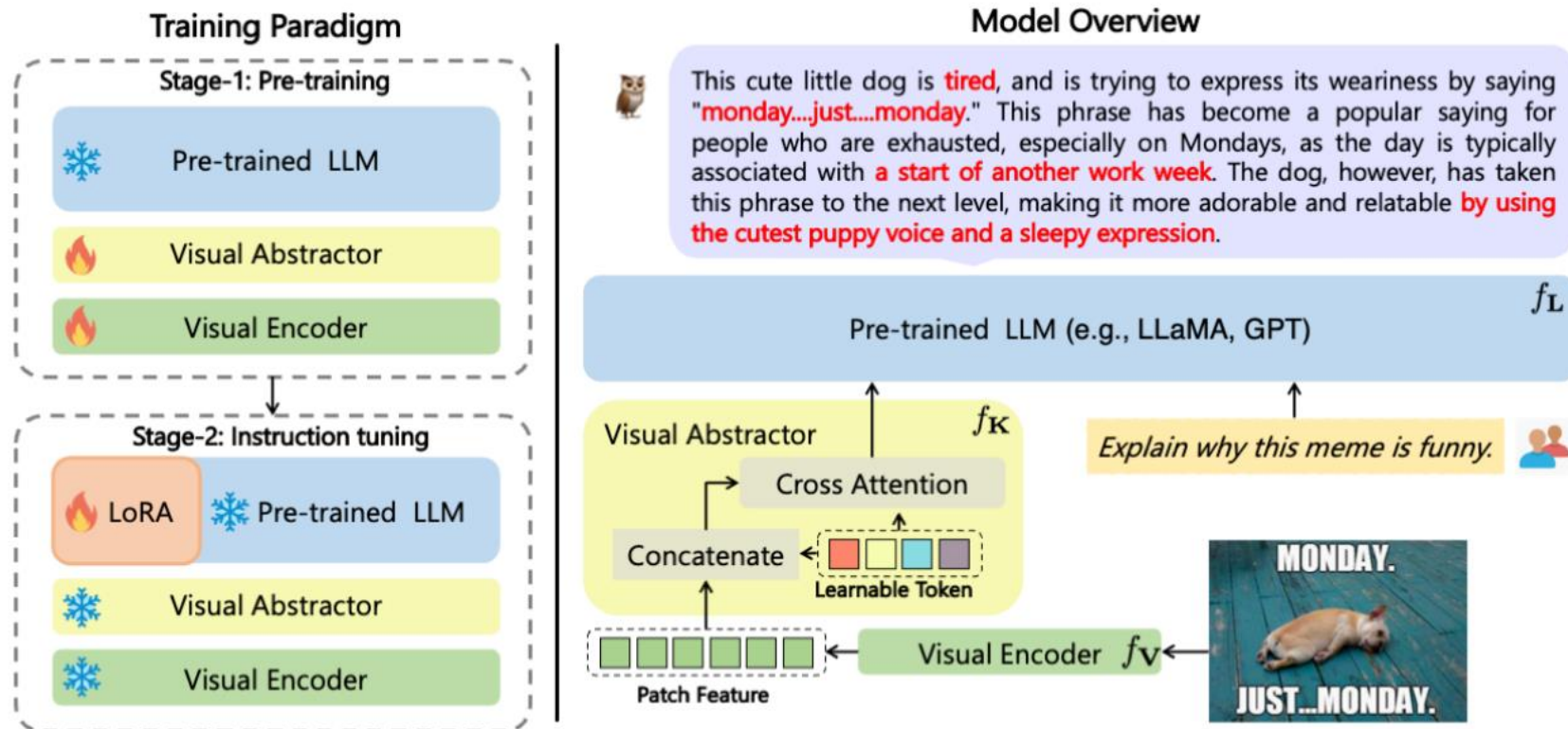
Our Main Focus

1. Extraction of mPLUG summaries
2. Question generation with Flan-T5
3. Variations of the baseline approach:
Data augmentation & Keyword extraction

mPLUG-Owl

In our work, we apply the pre-trained *mplug-owl-llama-7b-video* model.

<https://huggingface.co/MAGAr13/mplug-owl-llama-7b-video>



mPLUG-Owl

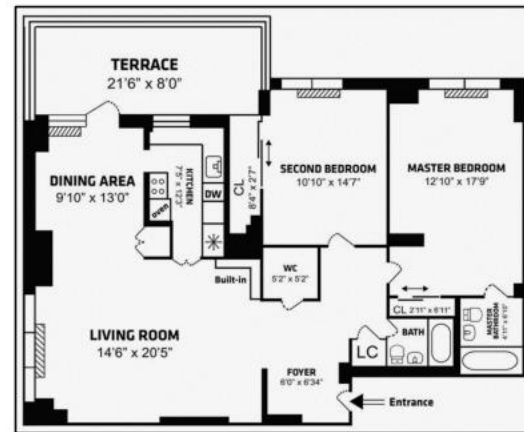
Observed abilities such as multi-image correlation and scene text understanding, vision-based document comprehension.



What is this four-panel cartoon about?



The four-panel cartoon shows an old woman walking along a city street with a cane in hand. She appears to be walking down a busy sidewalk or street corner, with many people around her. There are cars on the road, and some people are standing or walking in the background.



How many bedrooms are there in this floor plan?



Score: **A**

There are two bedrooms in this floor plan.

```
def Fibonacci(n):  
    if n < 0:  
        print("Incorrect input")  
    elif n == 0:  
        return 0  
    elif n == 1 or n == 2:  
        return 1  
    else:  
        return Fibonacci(n-1) + Fibonacci(n-2)  
print(Fibonacci(9))
```

Explain this code



The code snippet is a Python function named `def fibonacci(n)` that returns the `n`th element of the Fibonacci sequence. It takes an input parameter `n` that is used to define the starting index of the sequence and the number of elements to be returned, respectively. The function starts by initializing two variables, `a` and `b`, to 0. It then iterates through the loop, incrementing the value of `a` and `b` by one each time. If the current value of `a` equals `b`, it returns the result of `a` and continues to the next iteration. If the current value of `a` is greater than `b`, it returns the result of `b`, and if `a` is less than `b` it returns the result of `a` and `b`. Finally, the function returns the last value of `n` elements, which is calculated using the formula for the `n`th Fibonacci number. In this case, the input parameter `n` is set to 5, and the function returns 5.

Extraction of mPLUG summaries

```
The following is a conversation
between a curious human and AI
assistant.
Human: Here is a video and a
transcript of someone explaining the
video.
Human: <|video|>
Human: Transcript: {transcript}
Human: Summarize the instructions
provided by the video and the
transcript in less than {num_sentences}
sentences.
AI:
```

Input prompt provided to the mPLUG model



How to Get Rid of Calf Muscle Pain for Good

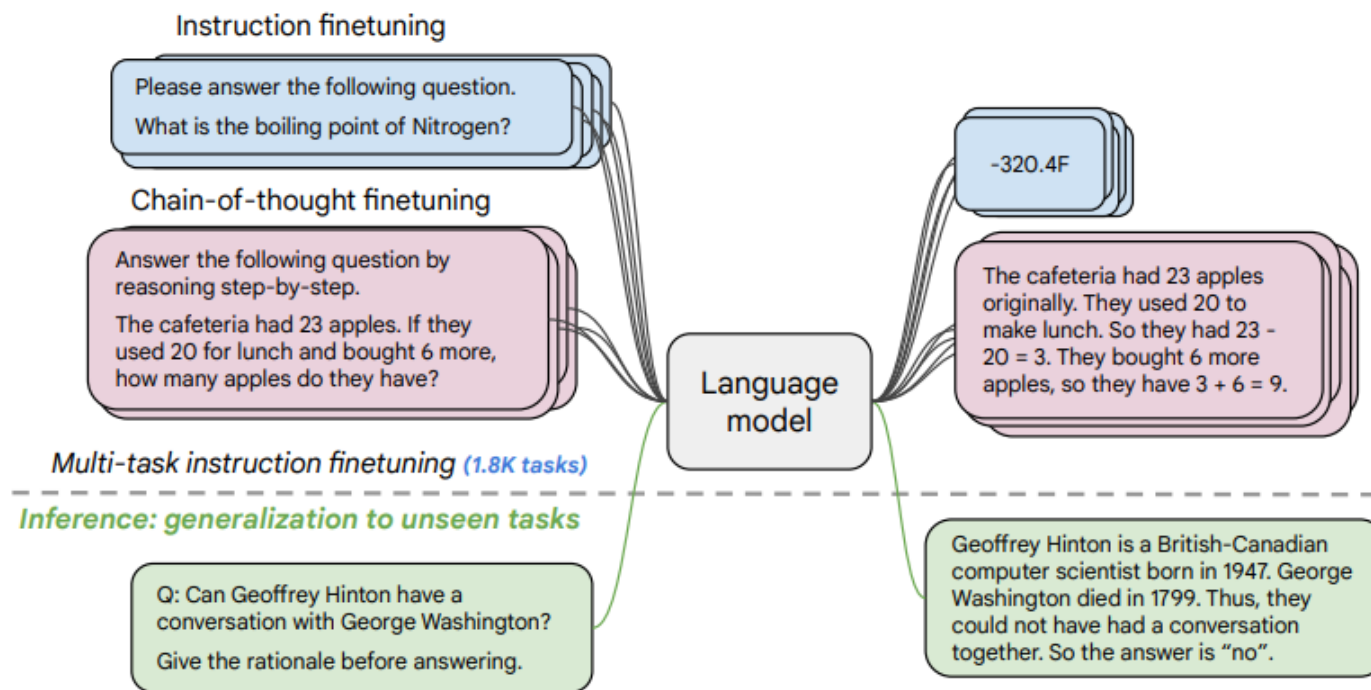
Video ID: PoSgyRG27-E
Visual answer start: 4:45
Visual answer end: 6:10

The video and transcript explain a simple calf stretch exercise that can help alleviate aches and pain, especially in runners and those with musculoskeletal issues. It involves placing the feet out wide with weight on the toes and then performing a rocking motion, first forward and then backward, while engaging the calf muscle. The exercise is repeated fifteen times for a total of 30 seconds per side to target the muscle effectively.

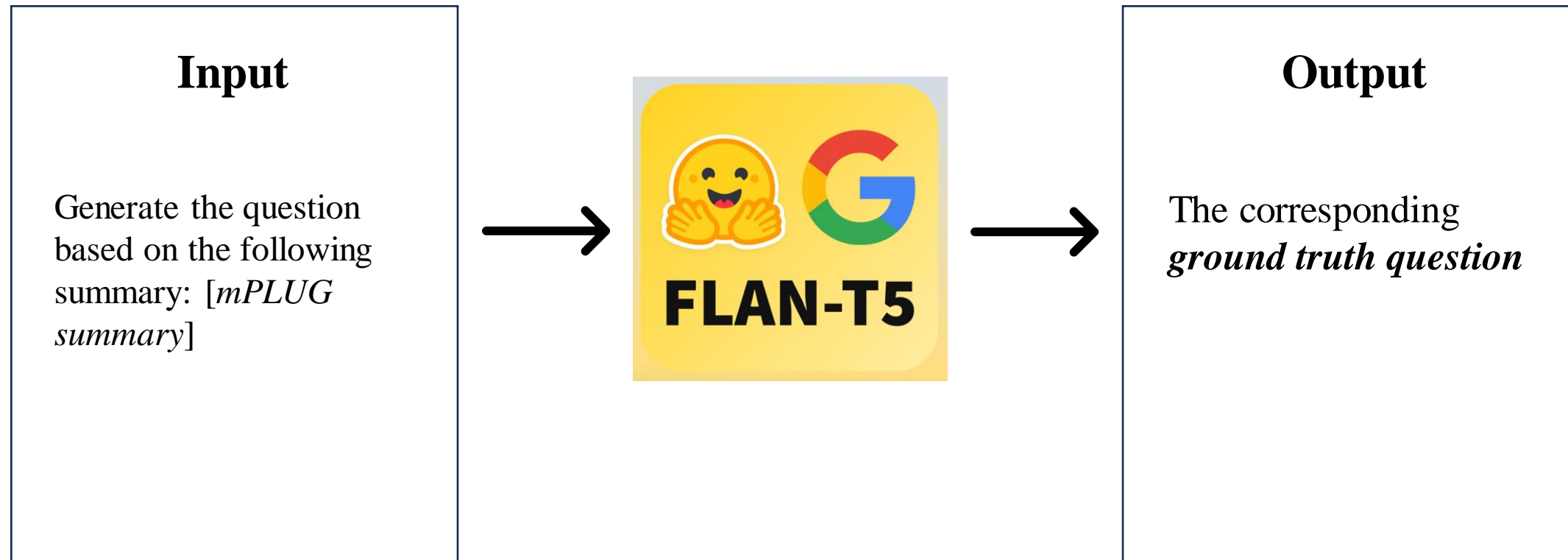
Flan-T5

In our work, we apply the pre-trained *flan-t5-base* model.

<https://huggingface.co/google/flan-t5-base>



Finetuning of Flan-T5



Generation example of Flan-T5

Model Input

['Generate the question based on the following summary: The instructions provided by the video and transcript include wrapping the bandage around the injured arm, starting at the top and applying pressure to stop the bleeding. They suggest using a shoulder rod and elevating the bandage above the heart to minimize the flow of blood.']

- Beam search Output: How to wrap the bandage around the injured arm?
- Nucleus sampling Output: How to wrap a bandage around a wound?

Variations of the Baseline approach

To attempt to improve the Baseline approach, we experimented with two alternative strategies to further fine-tune the Flan-T5 model for the MIQG task.

- Training data augmentation
- Keyword extraction

Data augmentation on both mPLUG summaries and ground truth questions

This augmentation led to a total of 43,013 training samples, up from 2,710 originally. Potential duplicates were removed to avoid redundancy in the training set.

```
The following is a conversation
between a curious human and AI
assistant.
Human: Here is a video and a
transcript of someone explaining the
video.
Human: <|video|>
Human: Transcript: {transcript}
Human: Summarize the instructions
provided by the video and the
transcript in less than {num_sentences}
sentences.
AI:
```

"Summarize the instructions provided by the video and the transcript in less than n sentences" with $n \in \{1,3\}$.

Data augmentation on both mPLUG summaries and ground truth questions

We apply the *TextAugment* Python library for data augmentation on ground truth questions.

- **Word2Vec**
- WordNet
- Translate (text translated into French, then back to English), double Translate (consecutive translation loops with English/French and English/German)
- **Synonym replacement with Easy Data Augmentation (EDA)**

We found out a mistake in the runs that we submitted using augmentation, the original mPLUG summaries generated with $n = 5$ were mistakenly omitted.

Keyword extraction from the mPLUG summaries

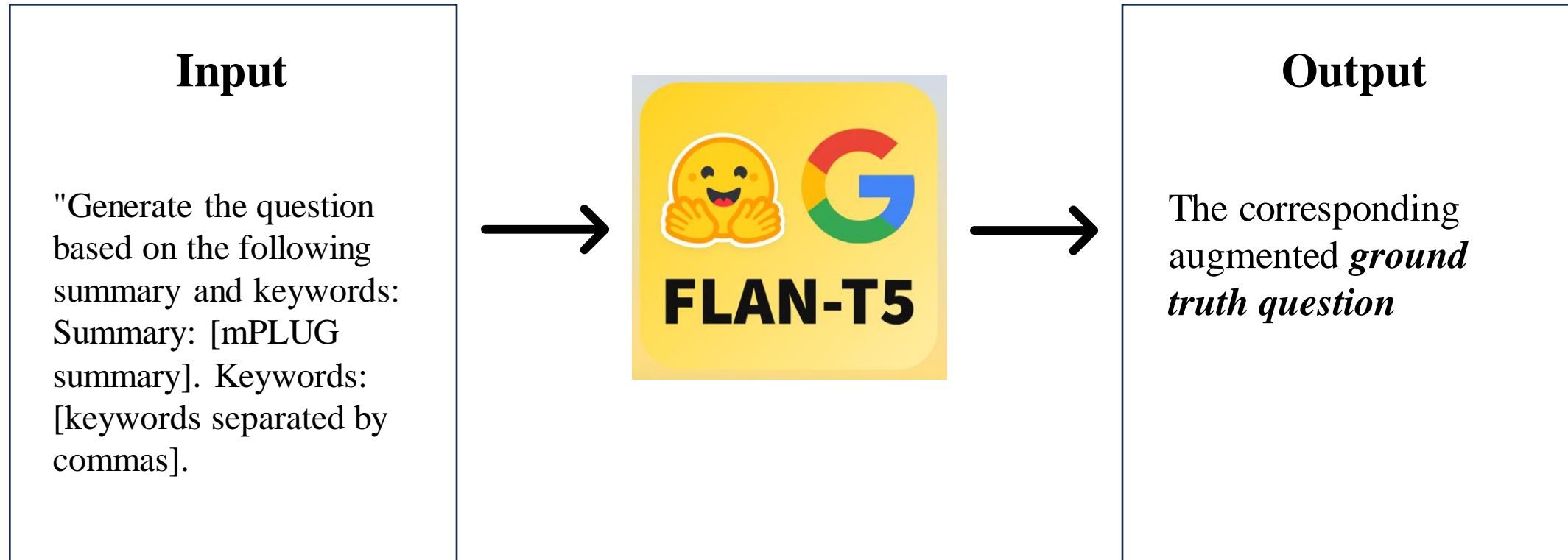
We used the *Python Keyphrase Extraction* (PKE) library that implements various statistical-, graph- or feature-based keyword extraction methods.

It was chosen to use the Topical Page Rank approach with a number of extracted keywords set to three.

[The video and transcript explain a simple calf stretch exercise that can help alleviate aches and pain, especially in runners and those with musculoskeletal issues. It involves placing the feet out wide with weight on the toes and then performing a rocking motion, first forward and then backward, while engaging the calf muscle. The exercise is repeated fifteen times for a total of 30 seconds per side to target the muscle effectively.]

['simple calf stretch exercise', 'musculoskeletal issues', 'calf muscle']

Finetuning of Flan-T5 using augmented data



Our Submitted Runs Results

Approach	BLEU	BLEU-4	ROUGE-2	ROUGE-L	BERTScore
run-1	0.15828	0.05153	0.27752	0.47825	0.91092
run-2	0.14352	0.04546	0.24478	0.44685	0.90523
run-3	0.14593	0.03875	0.27339	0.47439	0.91099
run-4	0.13289	0.03404	0.24421	0.45659	0.90780
run-5	0.09300	0.01627	0.20023	0.40716	0.90248
Overall min	0	0	0.12262	0.26083	0.85332
Overall mean	0.10041	0.02867	0.22098	0.41484	0.89717
Overall max	0.15828	0.05153	0.27752	0.47924	0.91099

Run	Search strategy	Augmentation	Keywords
run-1	beam	no	yes
run-2	beam	yes	no
run-3	beam	no	no
run-4	beam	yes	yes
run-5	nucleus	no	no

Our run-1 achieves top 1 BLEU, BLEU-4, ROUGE-2 scores, and top 2 ROUGE-L and BertScore for the MIQG task of TRECVID 2023.

Additionally, our run-3 achieves top 1 BertScore overall.

Discussion

- Dataset

In some cases, videos lacked **descriptive transcripts** or contained only instructions through on-screen text or actions. On the other hand, the dataset encompasses a diverse range of video types, including **animation-style videos**.

- Limitations

Sometimes simple augmentation techniques may lead to a degradation of performances.

Discussion

- Methods

The individual impact of the different data augmentations, hyperparameters, finetuning strategies, and prompts.

Keyword extraction is not the most reliable right now.

Despite the different summaries being close to each other, the most relevant keywords extracted were obtained for $n = 5$ in some cases.

```
mPLUG prompt: "Summarize the
instructions provided by the video
and the transcript in less than n
sentences."
```

```
Output for sample 1 of the training
set:
```

```
n = 5: "In the video, a physical
therapist explains the Epley maneuver
[...]. Both involve a series of head
movements [...]."
```

```
Keywords: "head movements", "epley
maneuver", "movement"
```

```
n=3: "The instructions for performing
the Epley maneuver to treat vertigo
involve a patient turning their head
[...]. After 30 seconds, they roll
over onto their left side, tilt their
head down towards their left shoulder,
and maintain this position for 30
seconds, [...]."
```

```
Keywords: "left side", "head",
"seconds"
```

```
n = 1: "The video and transcript
describe two Epley exercises for
treating benign paroxysmal positional
vertigo (BPPV). The first exercise
involves [...], and the second exercise
requires them to [...]."
```

```
Keywords: "second exercise", "first
exercise", "epley exercises"
```

Conclusion and Future work

- Proposed approach

mPLUG model (summarize the contents of the video clip using both video and text)

Flan-T5 model (generate questions using the mPLUG summaries as input)

Either text augmentation and/or keyword extraction were tested as well.

- Results

Best results with keywords and without augmentation. Our run-1 gets top 1 or 2 in all metrics.

- Future work

1. testing augmentation on the questions and summaries separately
2. devising a more systematic strategy for hyper-parameter selection
3. exploring learning-based keyword extraction methods and different prompt strategies

Thank you!!!