# Analyzing the Behavior of Deep VQA Models

(EMNLP 2016)



Aishwarya Agrawal  Dhruv Batra  Devi Parikh

Virginia Tech

# Visual Question Answering (VQA)

What is Visual Question Answering?

# VQA Task

# VQA Task

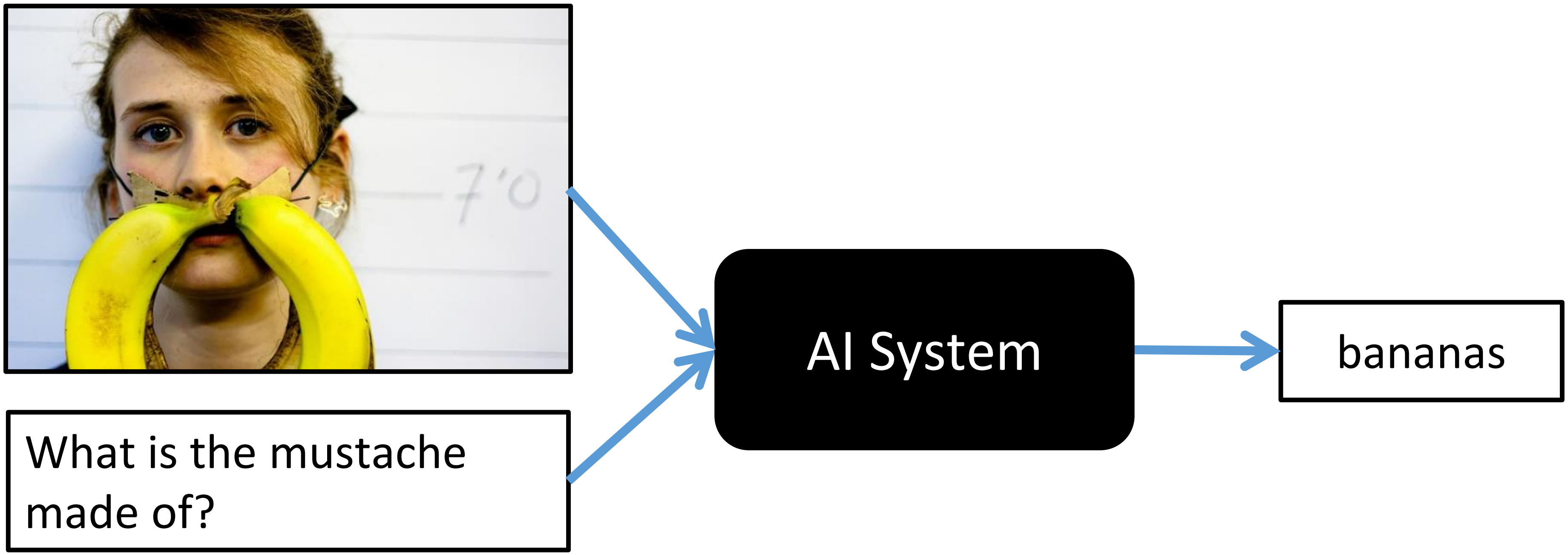

What is the mustache made of?

# VQA Task


What is the mustache made of?

AI System

# VQA Task



# Papers using VQA

**Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources**

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide
{qi.wu01,p.wang,chunhua.shen,anton.vandenhengel,anthony.dick}@adelaide.edu.au

**Simple Baseline for Visual Question Answering**

Bolei Zhou[1], Yuandong Tian[2], Sainbayar Sukhbaatar[2], Arthur Szlam[2], and Rob Fergus[2]
[1]Massachusetts Institute of Technology
[2]Facebook AI Research

**Compositional Memory for Visual Question Answering**

Aiwen Jiang[1,2] Fang Wang[2] Fatih Porikli[2] Yi Li* [2,3]
[1]Jiangxi Normal University [2]NICTA and ANU [3]Toyota Research Institute North America
[1]aiwen.jiang@nicta.com.au [2]{fang.wang, fatih.porikli}@nicta.com.au [3]yi.li@tema.toyota.com

**Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering**

Huijuan Xu, UMass Lowell, hxu1@cs.uml.edu
Kate Saenko, UMass Lowell, saenko@cs.uml.edu

**Deep Compositional Question Answering with Neural Module Networks**

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu

**Where To Look: Focus Regions for Visual Question Answering**

Kevin J. Shih, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign
{kjshih2, ss1, dhoiem}@illinois.edu

**ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering**

Kan Chen, University of Southern California, kanchen@usc.edu
Jiang Wang, Baidu Research - IDL, wangjiang03@baidu.com
Liang-Chieh Chen, UCLA, lcchen@cs.ucla.edu
Haoyuan Gao, Baidu Research - IDL, gaohaoyuan@baidu.com
Wei Xu, Baidu Research - IDL, wei.xu@baidu.com
Ram Nevatia, University of Southern California, nevatia@usc.edu

**Stacked Attention Networks for Image Question Answering**

Zichao Yang[1], Xiaodong He[2], Jianfeng Gao[2], Li Deng[2], Alex Smola[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond, WA 98052, USA
oy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

# Papers using VQA

**Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources**

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide
{qi.wu01,p.wang,chunhua.shen,anton.vandenhengel,anthony.dick}@adelaide.edu.au

**Simple Baseline for Visual Question Answering**

Bolei Zhou[1], Yuandong Tian[2], Sainbayar Sukhbaatar[2], Arthur Szlam[2], and Rob Fergus[2]
[1]Massachusetts Institute of Technology
[2]Facebook AI Research

**Compositional Memory for Visual Question Answering**

Aiwen Jiang[1,2] Fang Wang[2] Fatih Porikli[2] Yi Li* [2,3]
[1]Jiangxi Normal University [2]NICTA and ANU [3]Toyota Research Institute North America
[1]aiwen.jiang@nicta.com.au [2]{fang.wang, fatih.poril

**Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering**

Huijuan Xu Kate Saenko
UMass Lowell UMass Lowell
@cs.uml.edu saenko@cs.uml.edu

... and many more

**Deep Compositional Question Answering with Neural Module Networks**

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu

**Where To Look: Focus Regions for Visual Question Answering**

Kevin J. Shih, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign
{kjshih2, ss1, dhoiem}@illinois.edu

**ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering**

Kan Chen, University of Southern California, kanchen@usc.edu
Jiang Wang, Baidu Research - IDL, wangjiang03@baidu.com
Liang-Chieh Chen, UCLA, lcchen@cs.ucla.edu
Haoyuan Gao, Baidu Research - IDL, gaohaoyuan@baidu.com
Wei Xu, Baidu Research - IDL, wei.xu@baidu.com
Ram Nevatia, University of Southern California, nevatia@usc.edu

**Stacked Attention Networks for Image Question Answering**

Zichao Yang[1], Xiaodong He[2], Jianfeng Gao[2], Li Deng[2], Alex Smola[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond, WA 98052, USA
oy@cs.cmu.edu, {xiaohe, jfgao, deng}@microsoft.com, alex@smola.org

# Papers using VQA

## ORAL SESSION

### Image Captioning and Question Answering

Monday, June 27th, 9:00AM - 10:05AM.

These papers will also be presented at the following poster session

1 **Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data.**
Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell

2 **Generation and Comprehension of Unambiguous Object Descriptions.**
Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, Kevin Murphy

3 **Stacked Attention Networks for Image Question Answering.**
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

4 **Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction.**
Hyeonwoo Noh, Paul Hongsuck Seo, Bohyung Han

5 **Neural Module Networks.**
Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein

# Papers using VQA

## ORAL SESSION

### Image Captioning and Question Answering

Monday, June 27th, 9:00AM - 10:05AM.

These papers will also be presented at the following poster session

1 **Deep Compositional Captioning: Describing Novel Object Categories Without Paired Training Data.**
Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell

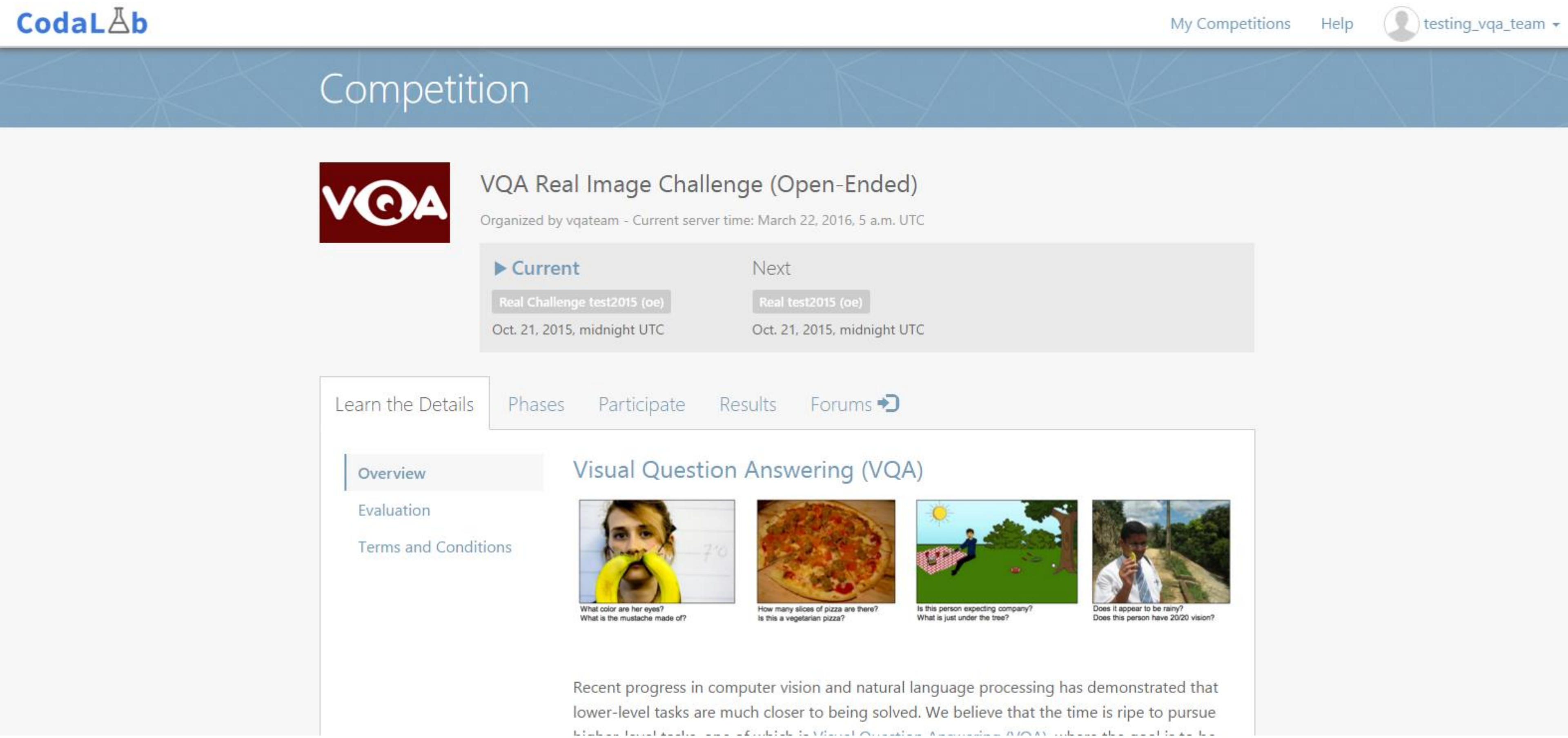
2 **Generation and Comprehension of Unambiguous Object Descriptions.**
Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, Kevin Murphy

3 **Stacked Attention Networks for Image Question Answering.**
Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola

4 **Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction.**
Hyeonwoo Noh, Paul Hongsuck Seo, Bohyung Han

5 **Neural Module Networks.**
Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein

# VQA Challenge @ CVPR16



# VQA Challenge @ CVPR16

| | By Answer Type | | | Overall |
|---|---|---|---|---|
| | Yes/No | Number | Other | |
| UC Berkeley & Sony[14] | 83.24 | 39.47 | 58 | 66.47 |
| Naver Labs[10] | 83.31 | 38.7 | 54.62 | 64.79 |
| DLAIT[5] | 83.25 | 40.07 | 52.09 | 63.68 |
| snubi-naverlabs[25] | 83.16 | 39.14 | 51.33 | 63.18 |
| POSTECH[11] | 81.67 | 38.16 | 52.79 | 63.17 |
| Brandeis[3] | 82.11 | 37.73 | 51.91 | 62.88 |
| VTComputerVison[19] | 79.95 | 38.22 | 51.95 | 62.06 |
| MIL-UT[7] | 81.98 | 37.56 | 49.75 | 61.77 |
| klab[23] | 81.53 | 39.27 | 49.61 | 61.69 |
| SHB_1026[13] | 82.07 | 36.81 | 47.77 | 60.76 |
| MMCX[8] | 80.43 | 36.82 | 48.33 | 60.36 |
| VT_CV_Jiasen[20] | 80.56 | 38.14 | 47.87 | 60.33 |
| LV-NUS[6] | 81.34 | 35.67 | 46.1 | 59.54 |
| ACVT_Adelaide[1] | 81.07 | 37.12 | 45.83 | 59.44 |
| UC Berkeley (DNMN)[15] | 80.98 | 37.48 | 45.81 | 59.44 |
| CNNAtt[4] | 81.04 | 36.44 | 45.76 | 59.33 |
| san[24] | 79.11 | 36.41 | 46.42 | 58.85 |
| UC Berkeley (NMN)[16] | 81.16 | 37.7 | 44.01 | 58.66 |
| global_vision[22] | 78.24 | 36.27 | 46.32 | 58.43 |
| vqateam-deeperLSTM_NormlizeCNN[27] | 80.56 | 36.53 | 43.73 | 58.16 |
| Mujtaba hasan[9] | 80.28 | 36.92 | 42.24 | 57.36 |
| RIT[12] | 78.82 | 35.97 | 42.13 | 56.61 |
| Bolei[2] | 76.76 | 34.98 | 42.62 | 55.89 |
| UPV_UB[18] | 78.88 | 36.33 | 40.27 | 55.77 |
| att[21] | 78.1 | 35.3 | 40.27 | 55.34 |
| vqateam-lstm_cnn[28] | 79.01 | 35.55 | 36.8 | 54.06 |
| UPC[17] | 78.05 | 35.53 | 36.7 | 53.62 |
| vqateam-nearest_neighbor[29] | 71.73 | 24.31 | 22 | 42.73 |
| vqateam-prior_per_qtype[30] | 71.17 | 35.63 | 9.32 | 37.55 |
| vqateam-all_yes[26] | 70.53 | 0.43 | 1.26 | 29.72 |

~ 30 teams

# Observations

# Observations

- Current machine performance around 60-66%

# Observations

- Current machine performance around 60-66%
- Human performance at 83%

# Observations

- Current machine performance around 60-66%
- Human performance at 83%
- How to identify where we need progress?

# Observations

- Current machine performance around 60-66%
- Human performance at 83%
- How to identify where we need progress?
- How to compare strengths and weaknesses?

# Observations

- Current machine performance around 60-66%
- Human performance at 83%
- How to identify where we need progress?
- How to compare strengths and weaknesses?
- How to develop insights into failure modes?

# Observations

- Current machine performance around 60-66%
- Human performance at 83%
- How to identify where we need progress?
- How to compare strengths and weaknesses?
- How to develop insights into failure modes?
- Need to understand the behavior of VQA models

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

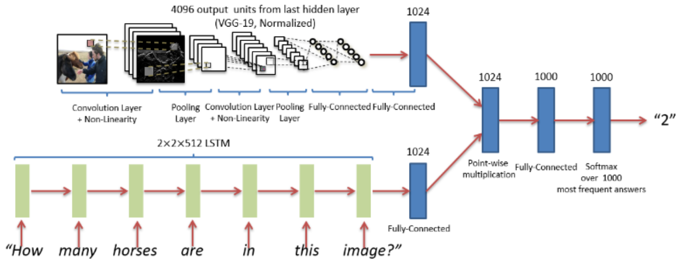
# Models

# Models

- Without attention (baseline model)

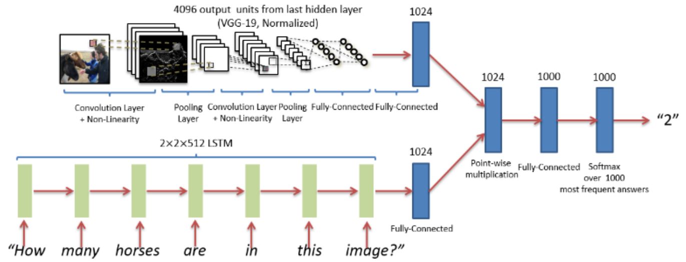
# Models

- Without attention (baseline model)
  - CNN + LSTM (Lu et al. 2015)



# Models

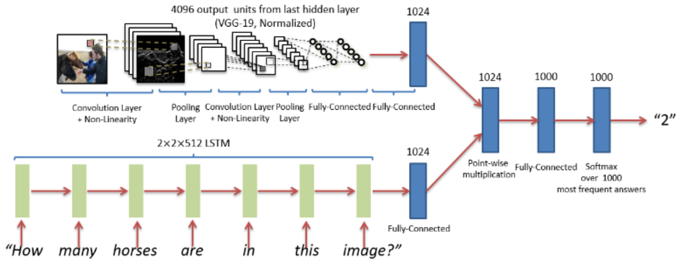
- Without attention (baseline model)
  - CNN + LSTM (Lu et al. 2015)




  - Accuracy = 54.13% (on VQA validation split)

# Models

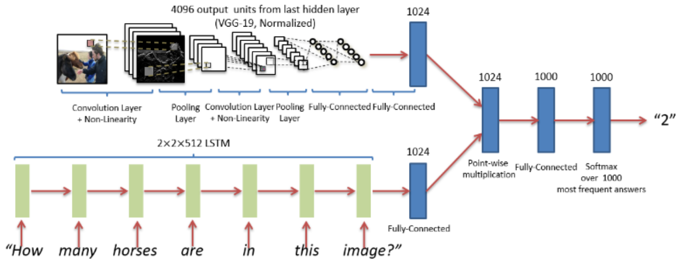
- Without attention (baseline model)
  - CNN + LSTM (Lu et al. 2015)



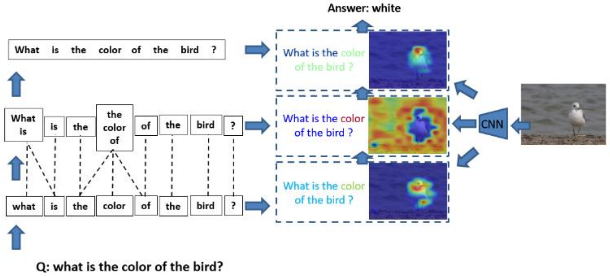

  - Accuracy = 54.13% (on VQA validation split)
- With attention

# Models

- Without attention (baseline model)
  - CNN + LSTM (Lu et al. 2015)

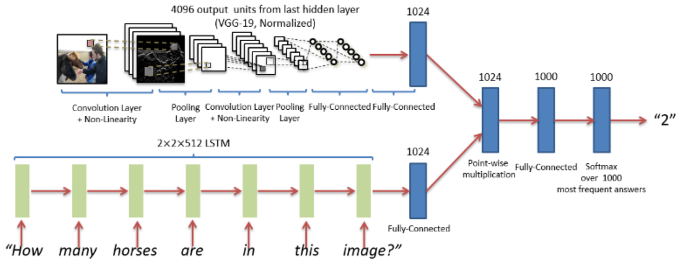



  - Accuracy = 54.13% (on VQA validation split)
- With attention
  - Hierarchical Co-attention (Lu et al. 2016)

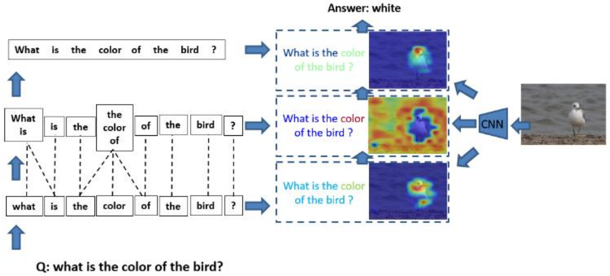


# Models

- Without attention (baseline model)
  - CNN + LSTM (Lu et al. 2015)




  - Accuracy = 54.13% (on VQA validation split)
- With attention
  - Hierarchical Co-attention (Lu et al. 2016)



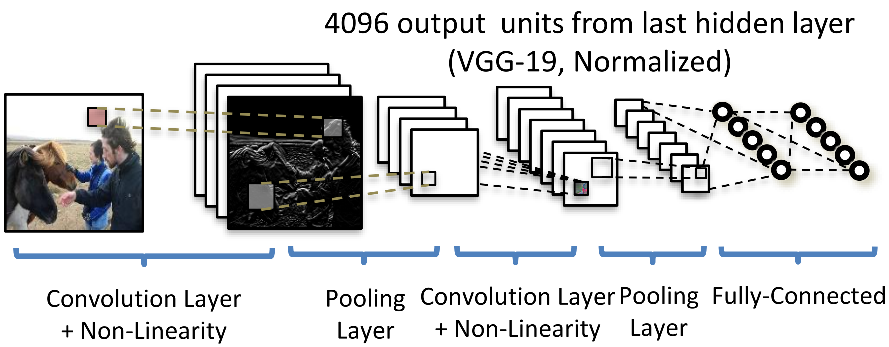

  - Accuracy = 57.02% (on VQA validation split)

# Without attention model

[Lu et al. 2015]

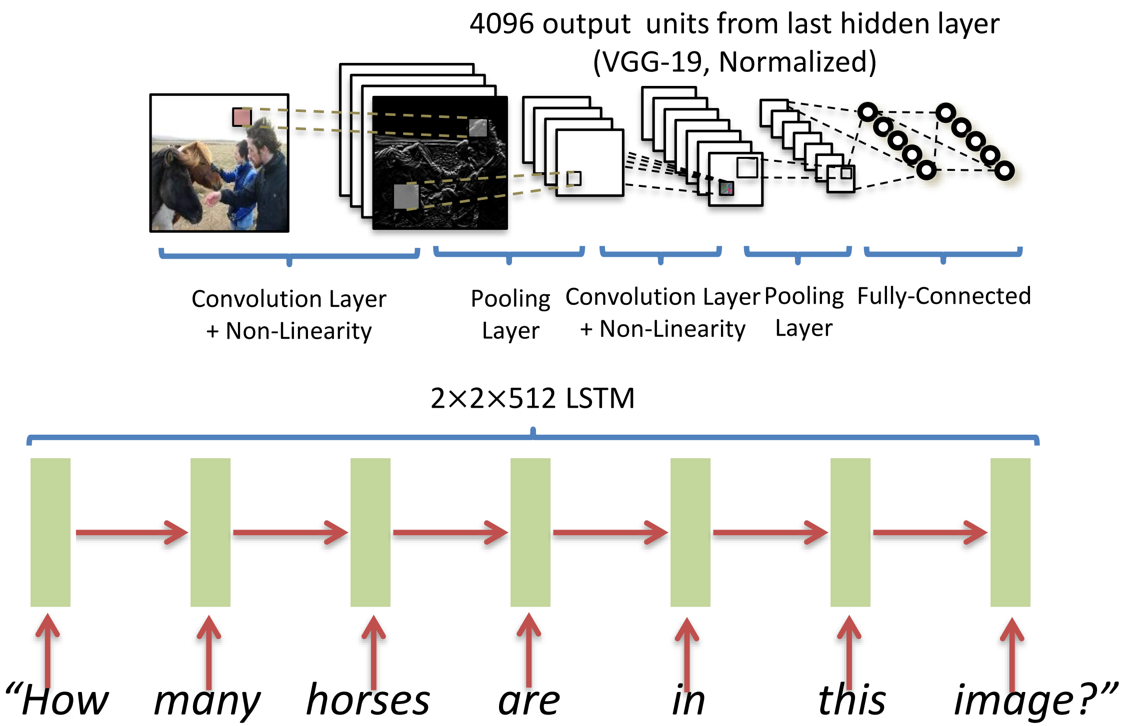
Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]



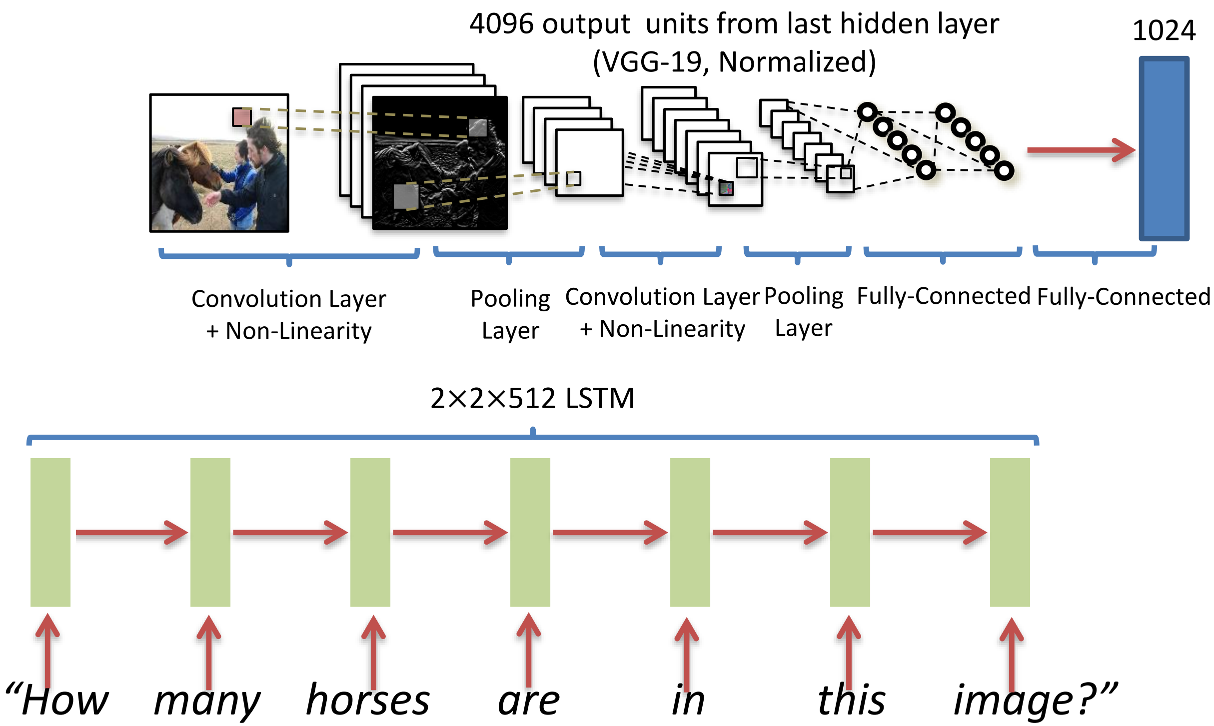

Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]



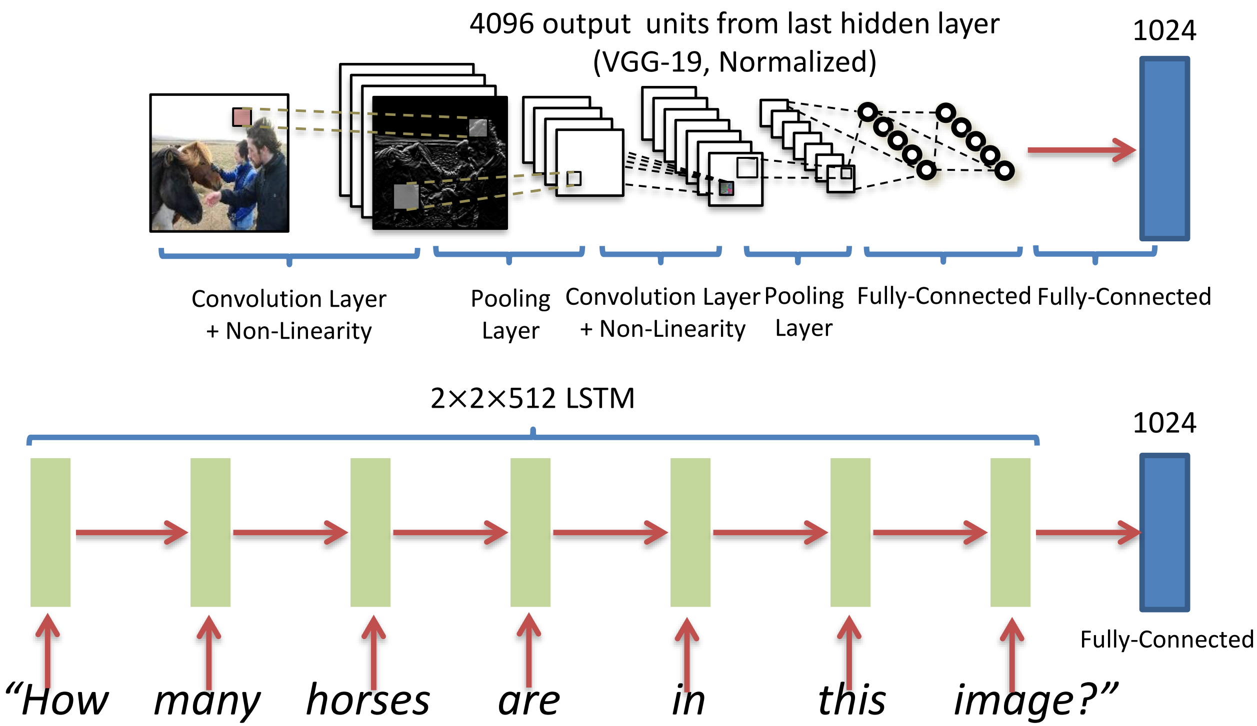

Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]



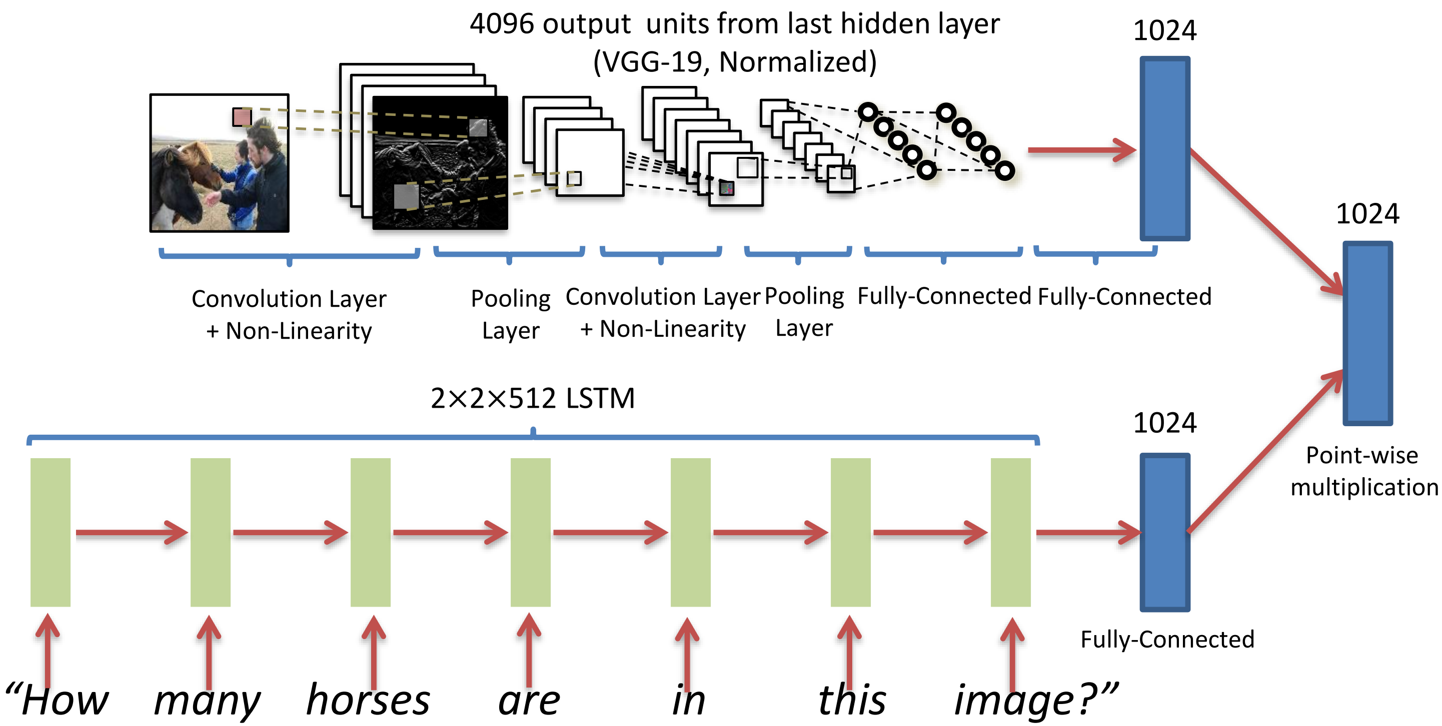

Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]

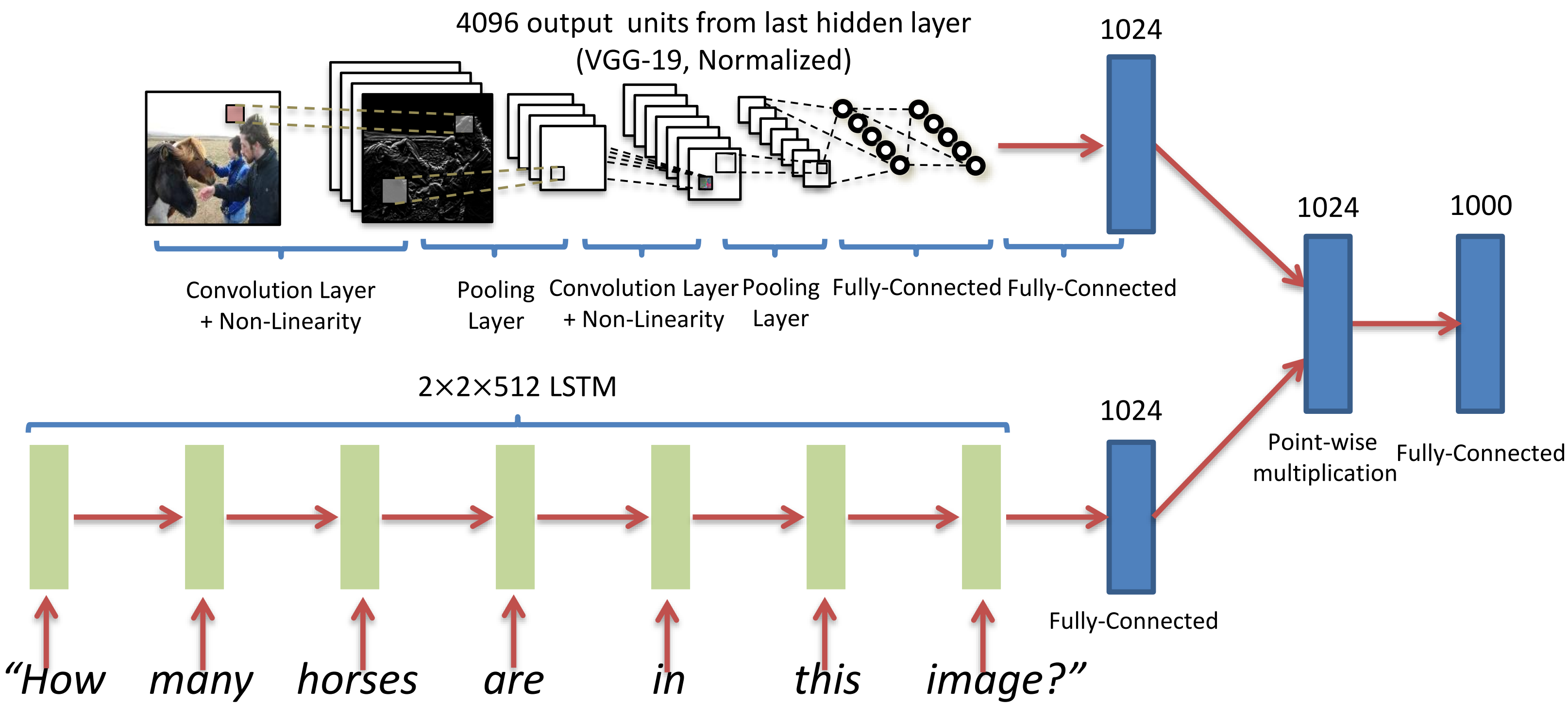



Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]

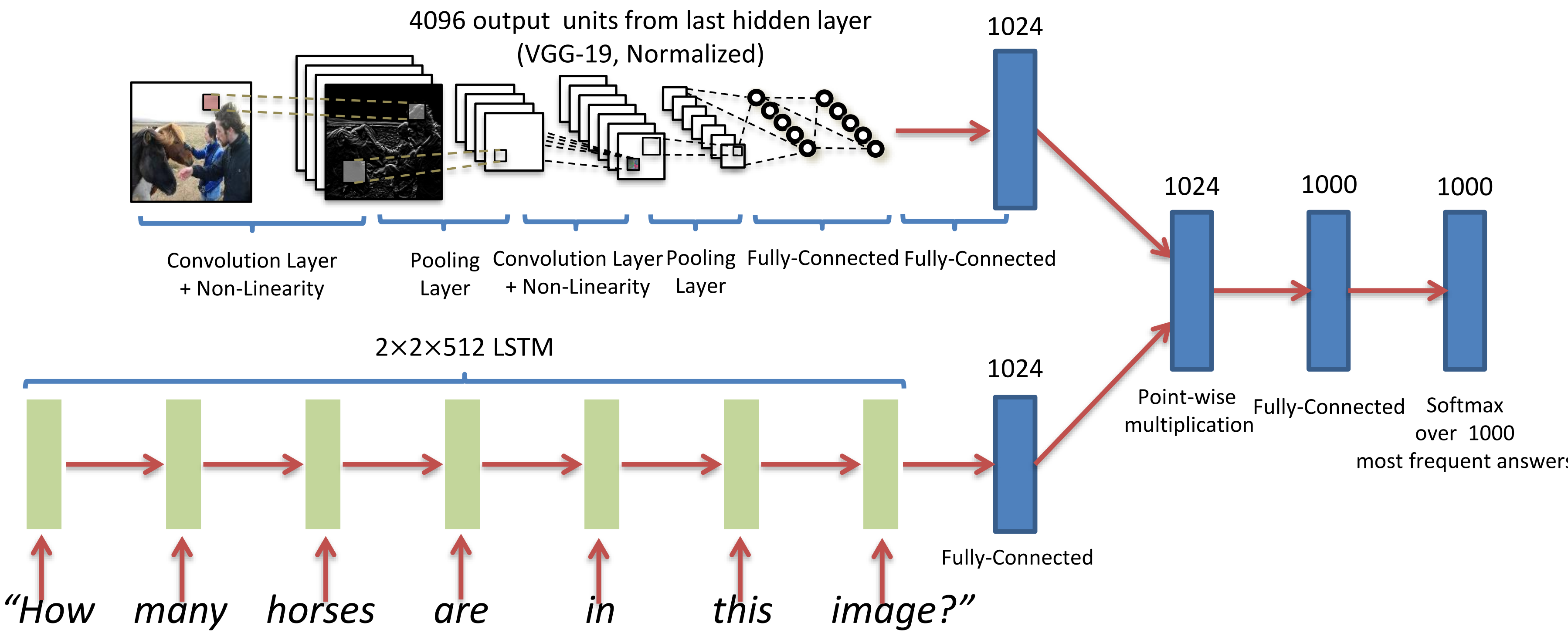


# Without attention model
## [Lu et al. 2015]



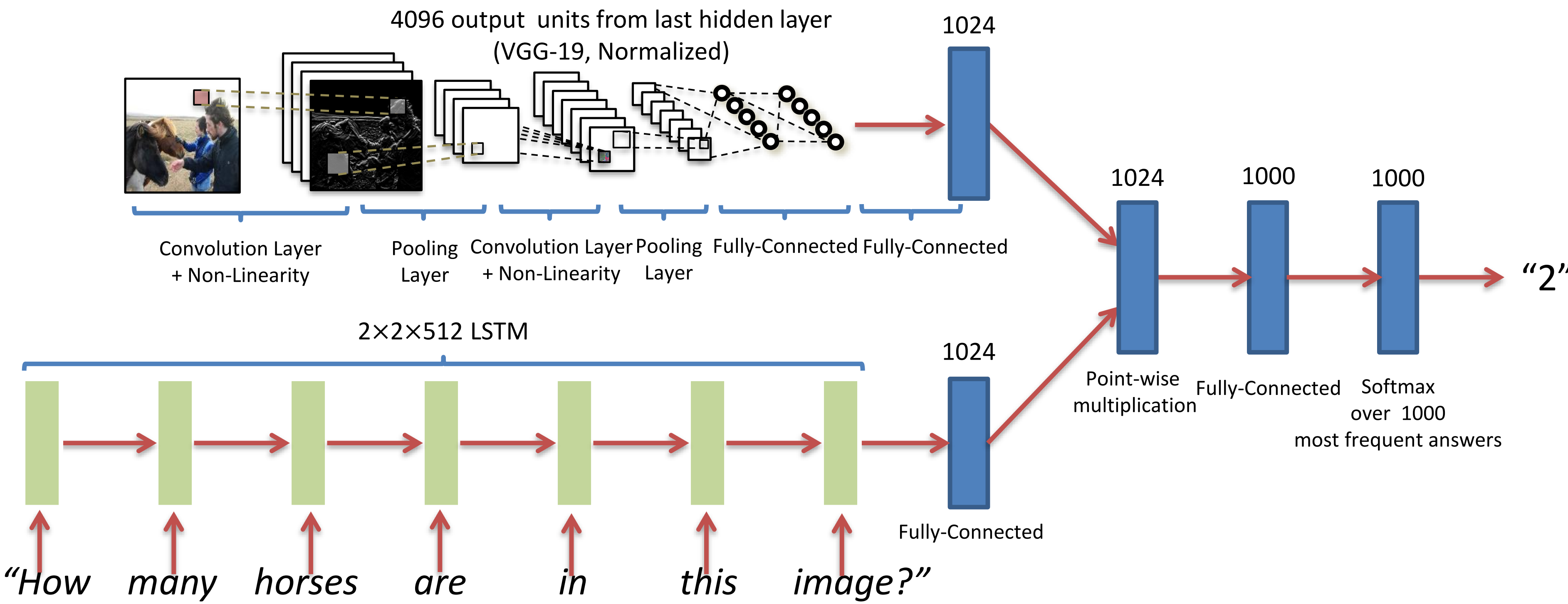

Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]




Slide credit: Devi Parikh

# Without attention model
## [Lu et al. 2015]




Slide credit: Devi Parikh

# With attention model
[Lu et al. 2016]

**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

# With attention model
## [Lu et al. 2016]

**Q: what is the color of the bird?**

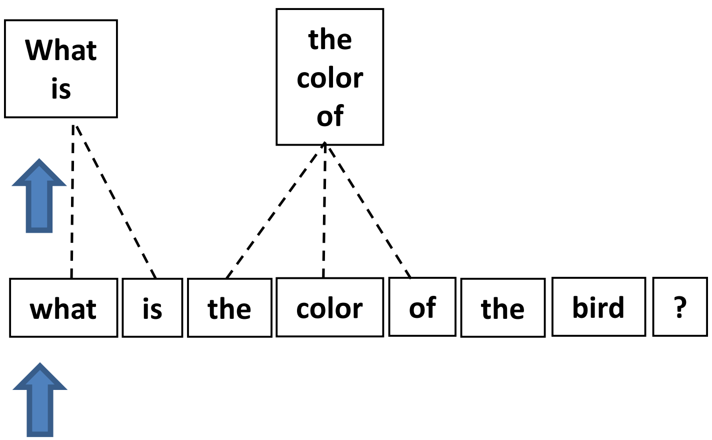
# With attention model
## [Lu et al. 2016]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **what** | **is** | **the** | **color** | **of** | **the** | **bird** | **?** |

**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

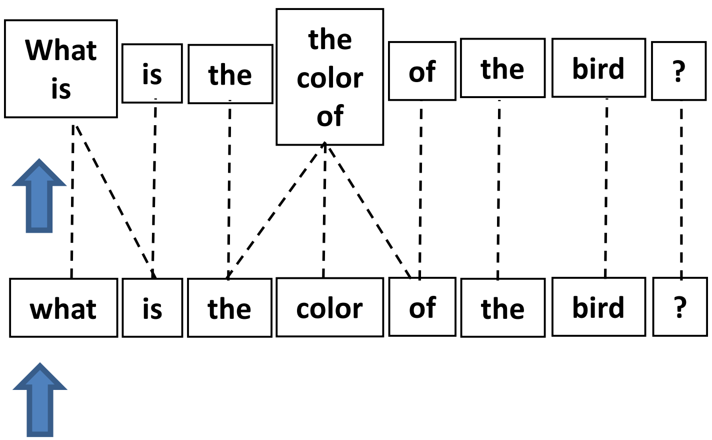
# With attention model
## [Lu et al. 2016]




**Q: what is the color of the bird?**

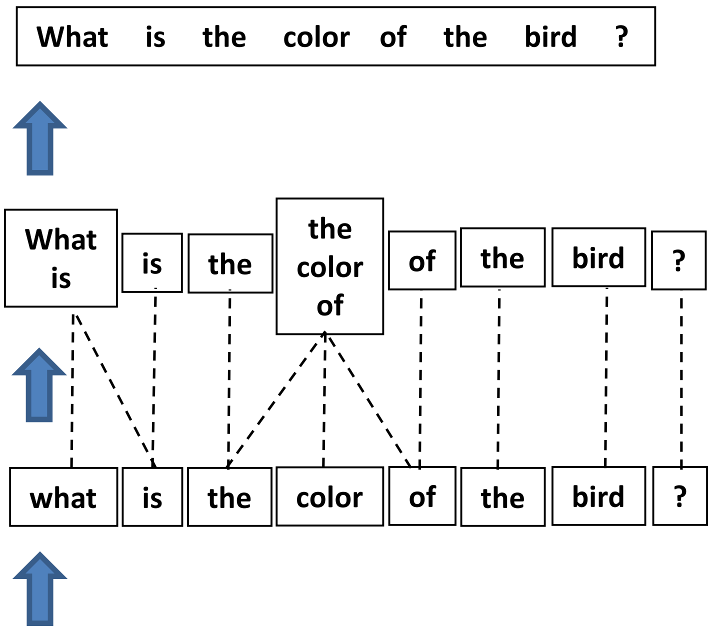
Slide credit: Adapted from Jiasen Lu

# With attention model
## [Lu et al. 2016]




**Q: what is the color of the bird?**

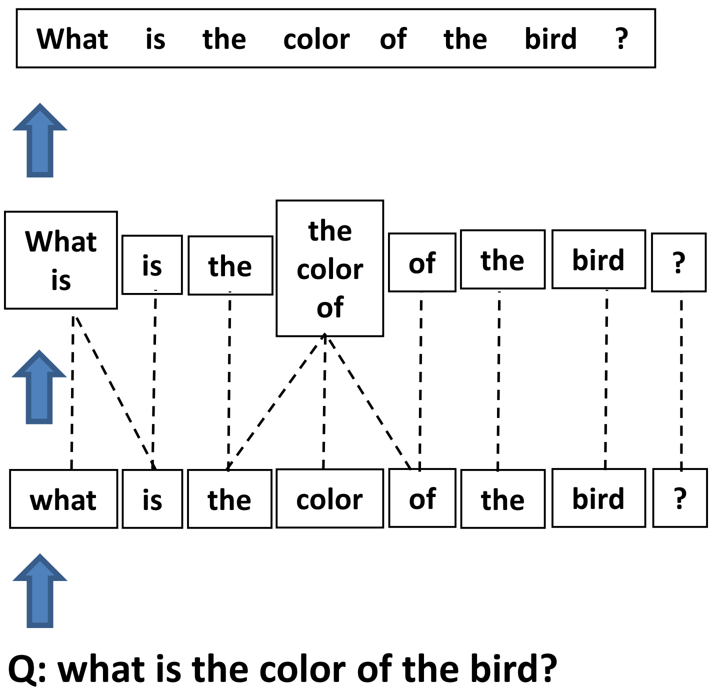
# With attention model
## [Lu et al. 2016]

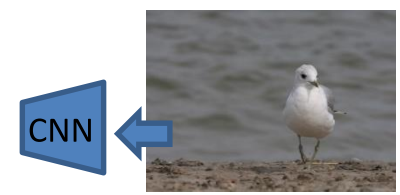



**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

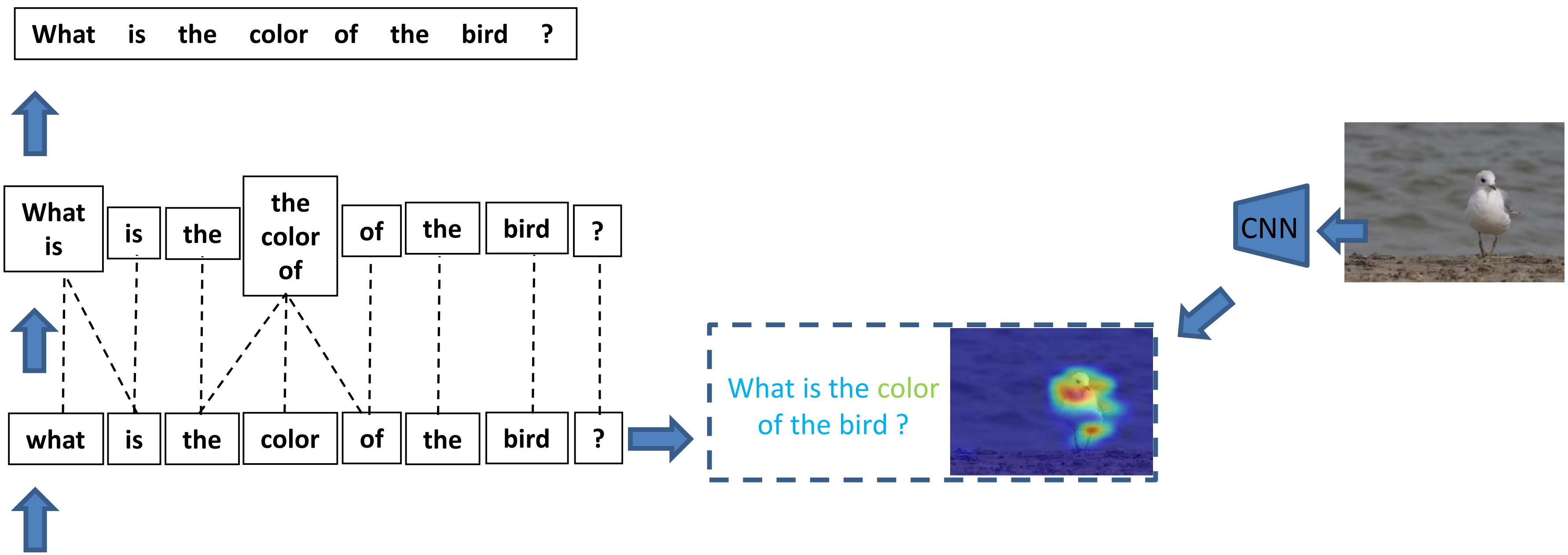
# With attention model
[Lu et al. 2016]




Slide credit: Adapted from Jiasen Lu

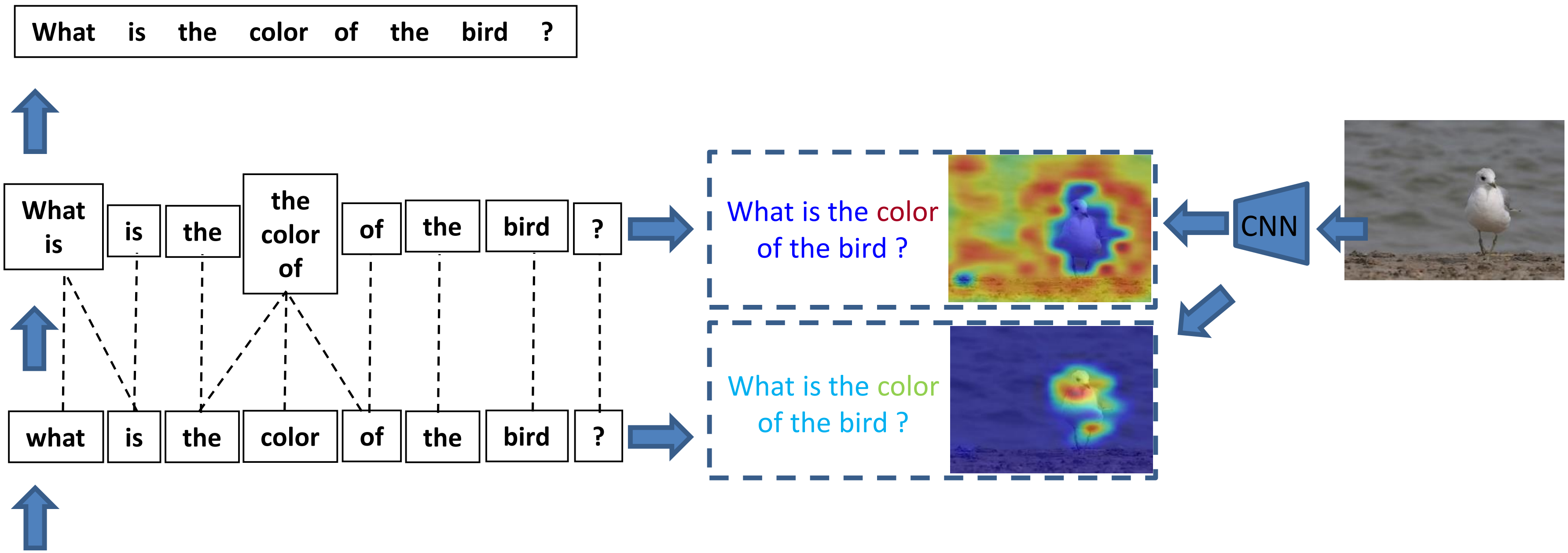
# With attention model
## [Lu et al. 2016]




**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

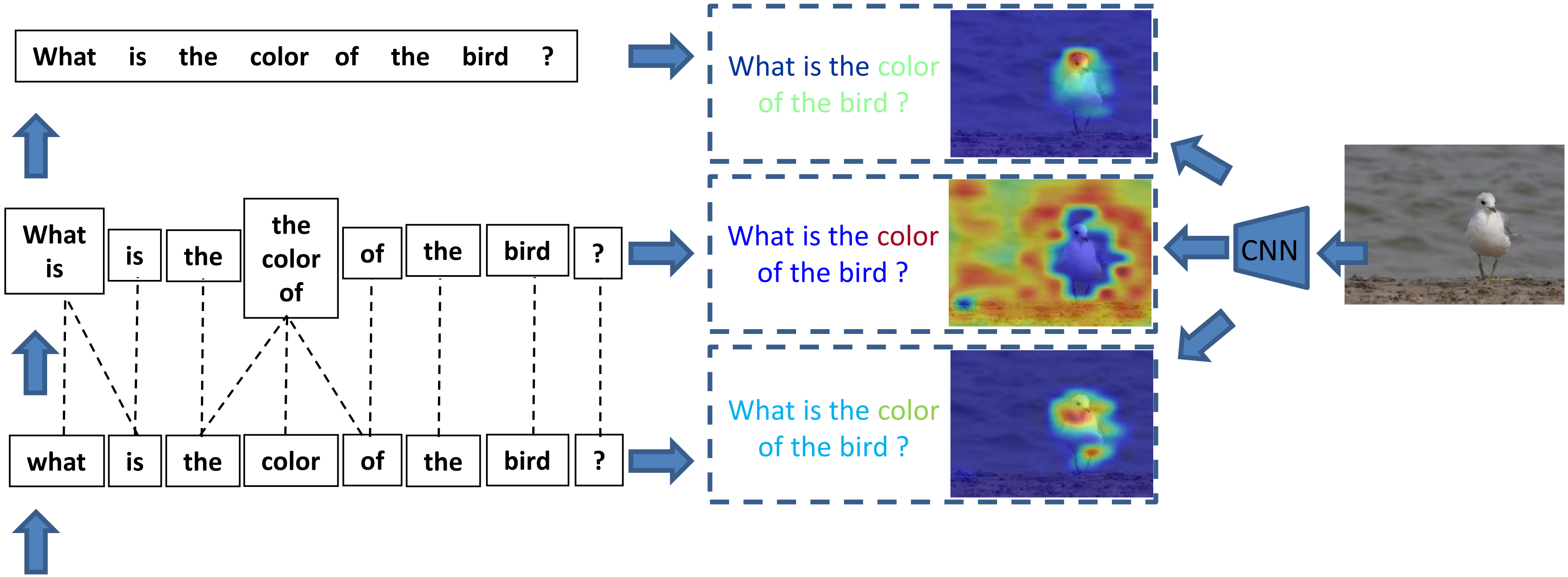
# With attention model
## [Lu et al. 2016]




**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

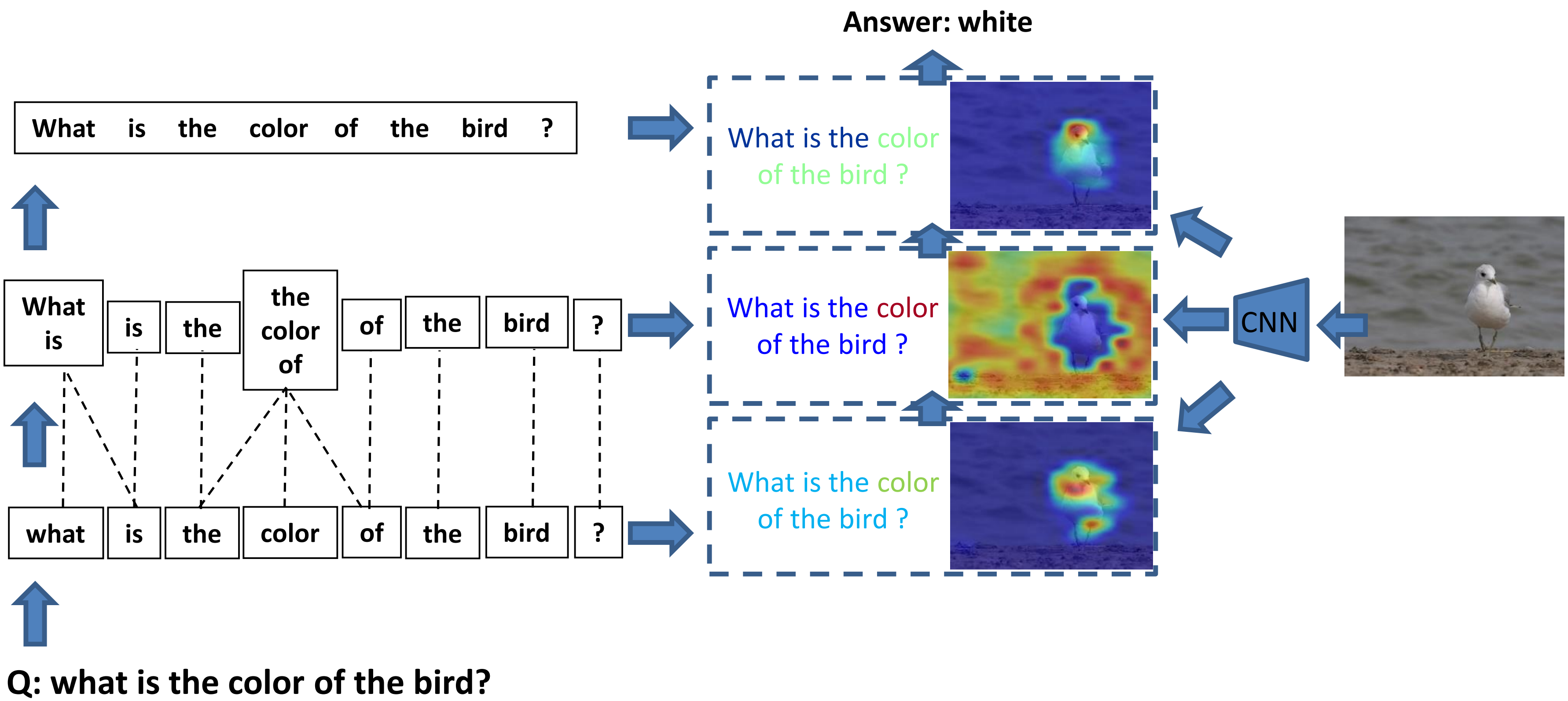
# With attention model
## [Lu et al. 2016]




**Q: what is the color of the bird?**

Slide credit: Adapted from Jiasen Lu

# With attention model
## [Lu et al. 2016]




Slide credit: Adapted from Jiasen Lu

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

# Generalization to Novel Instances

Do VQA models make mistakes because test instances are too different from training ones?

# Generalization to Novel Instances

Do VQA models make mistakes because test instances are too different from training ones?

1. Lower test accuracy $\Longrightarrow$ test QI pairs are too different from training QI pairs?

# Generalization to Novel Instances

Do VQA models make mistakes because test instances are too different from training ones?

1. Lower test accuracy ⟹ test QI pairs are too different from training QI pairs?

2. Lower test accuracy ⟹ test QI pairs are “familiar” **but** test labels are too different from training labels?

# Generalization to Novel Instances

Do VQA models make mistakes because test instances are too different from training ones?

1. Lower test accuracy ⟹ test QI pairs are too different from training QI pairs?

2. Lower test accuracy ⟹ test QI pairs are "familiar" **but** test labels are too different from training labels?

# Generalization to Novel Instances

**Experiment**

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair
2. Compute average distance between k-NN training QI pairs and test QI pair

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair
2. Compute average distance between k-NN training QI pairs and test QI pair
3. Measure correlation between average distance and test accuracy

# Generalization to Novel Instances

**K-NN Space**

# Generalization to Novel Instances

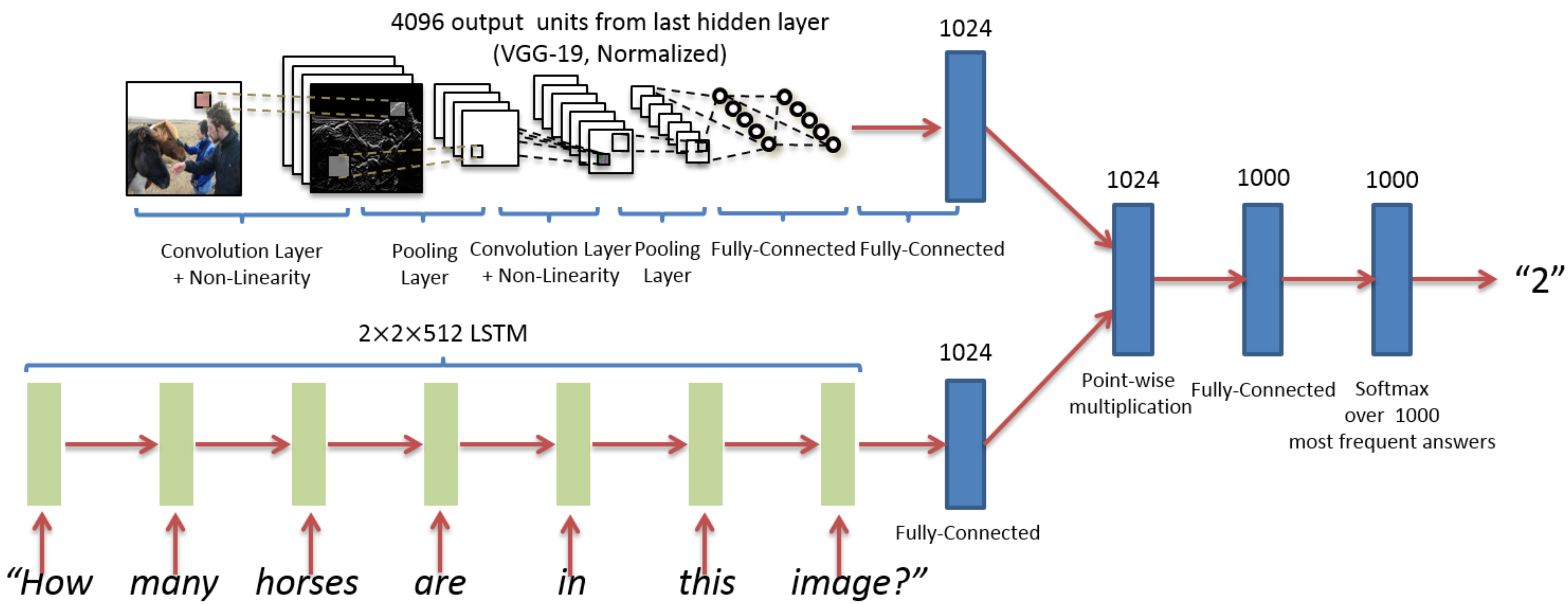
**K-NN Space**

Combined Q+I embedding

# Generalization to Novel Instances

**K-NN Space**

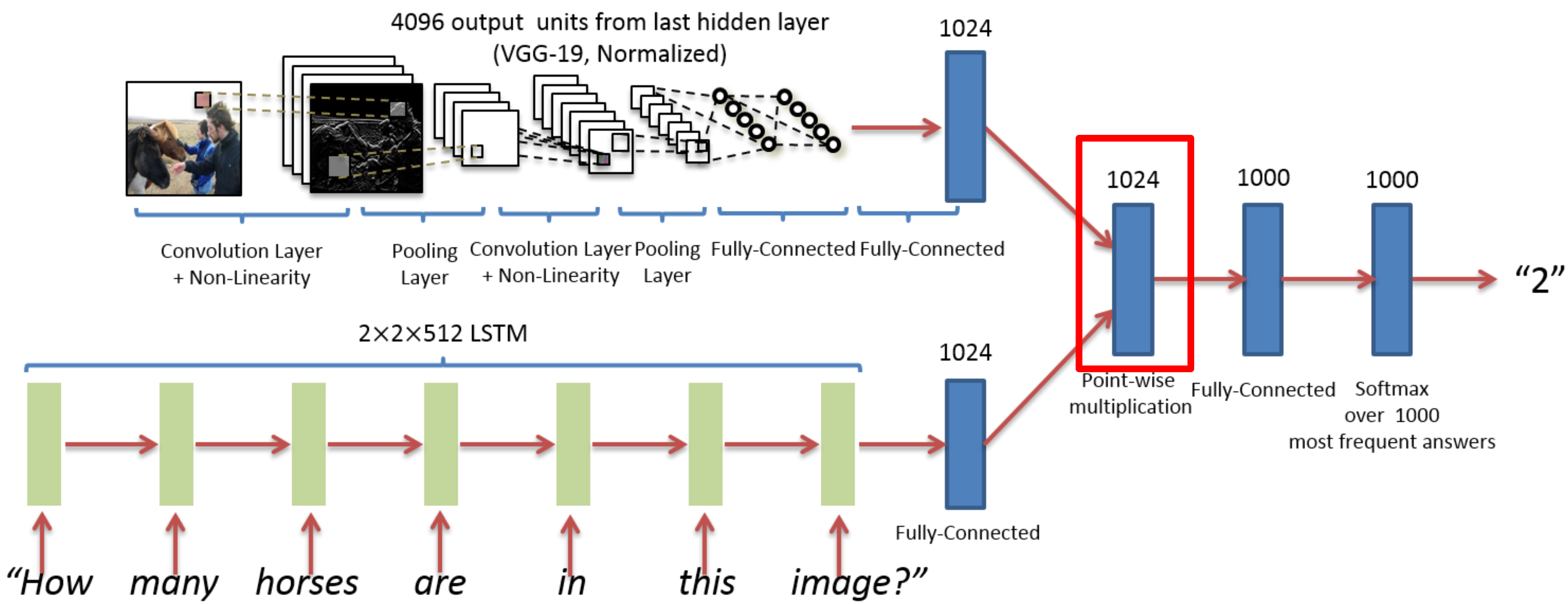
Combined Q+I embedding



# Generalization to Novel Instances

**K-NN Space**

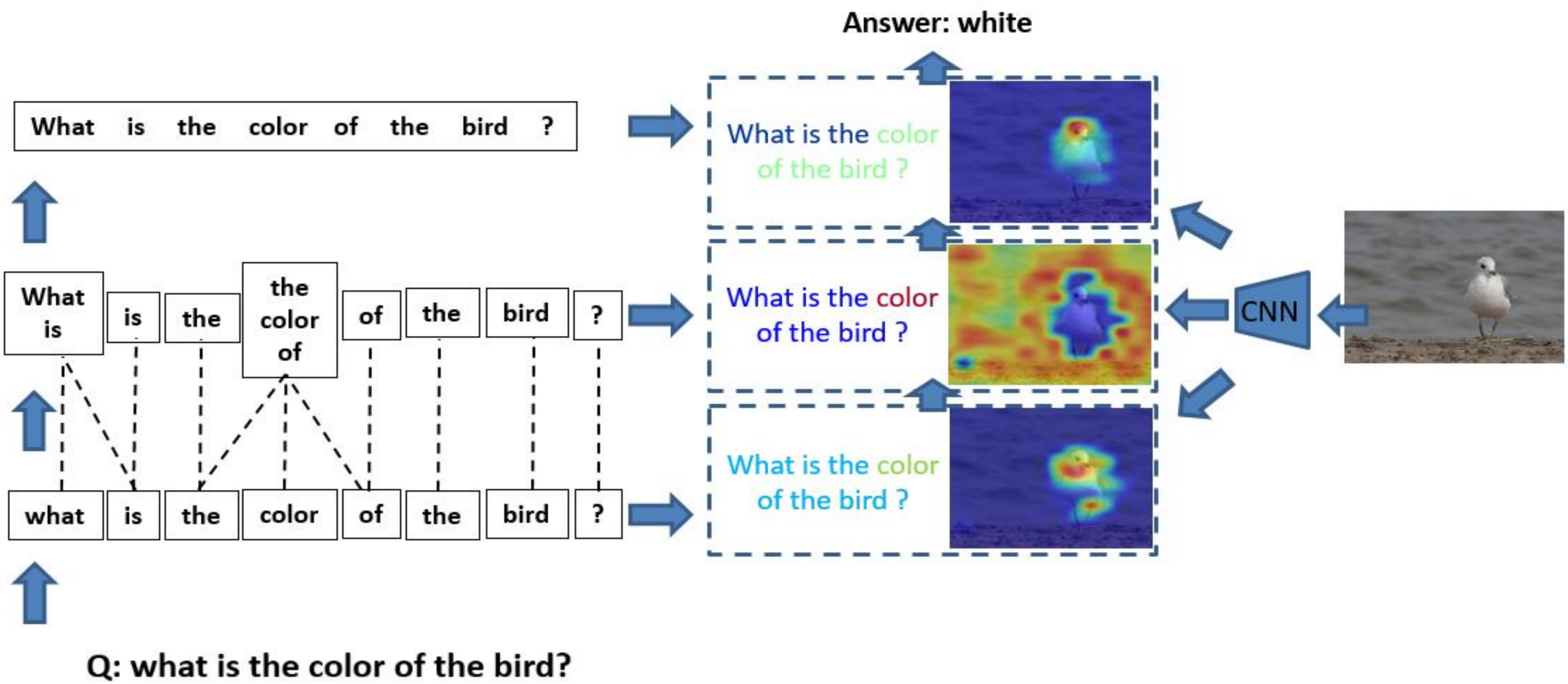
Combined Q+I embedding



# Generalization to Novel Instances

**K-NN Space**

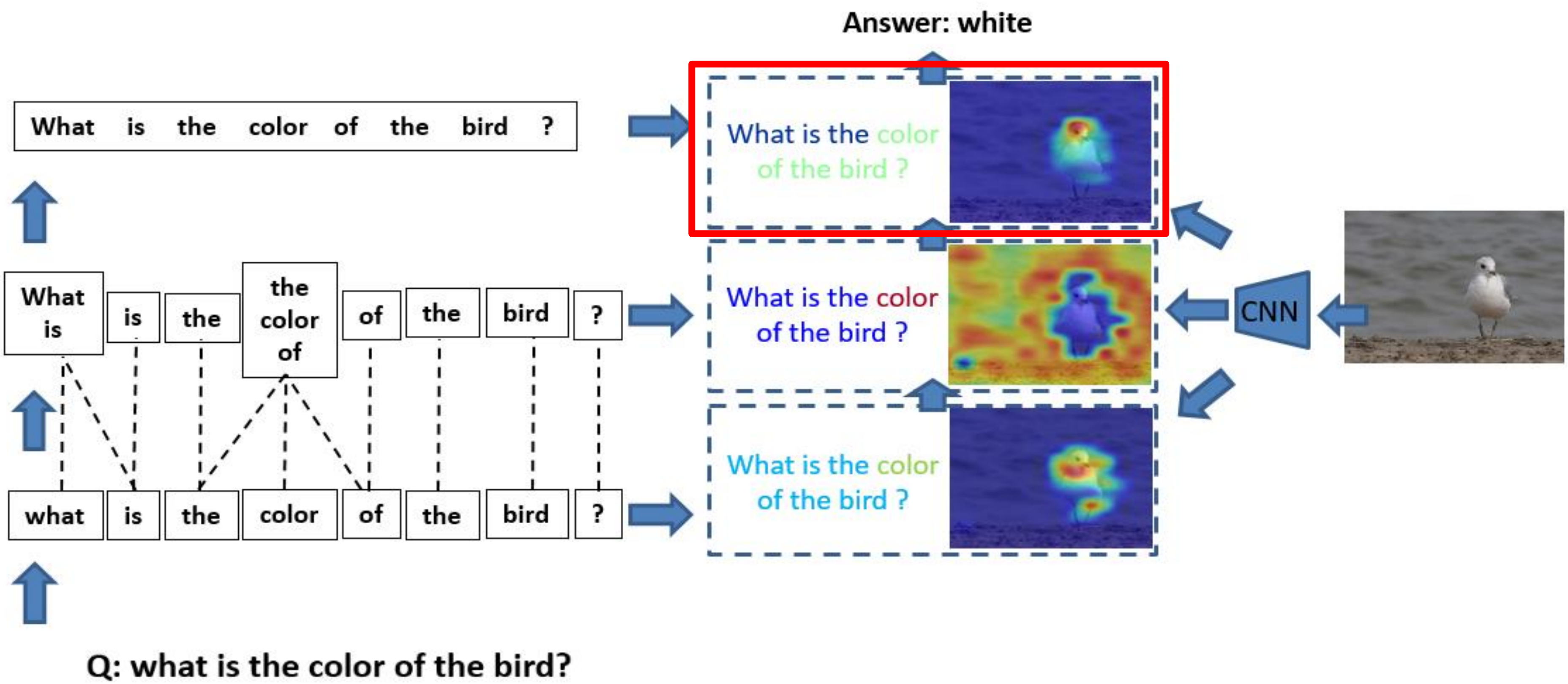
Combined Q+I embedding



# Generalization to Novel Instances

**K-NN Space**

Combined Q+I embedding



# Generalization to Novel Instances

**Results**

# Generalization to Novel Instances

**Results**

Significant negative correlation

# Generalization to Novel Instances

**Results**

Significant negative correlation

| | Without Attention | With Attention |
|---|---|---|
| Correlation | -0.41 (@ k=50) | -0.42 (@ k=15) |

# Generalization to Novel Instances

**Results**

Significant negative correlation

| | Without Attention | With Attention |
|---|---|---|
| Correlation | -0.41 (@ k=50) | -0.42 (@ k=15) |

*VQA models are not very good at generalizing to novel test QI pairs*

# Generalization to Novel Instances

**Results**

Significant negative correlation

| | Without Attention | With Attention |
|---|---|---|
| Correlation | -0.41 (@ k=50) | -0.42 (@ k=15) |

*VQA models are not very good at generalizing to novel test QI pairs*

⇩

*VQA models are "myopic"*

# Generalization to Novel Instances

**Results**

# Generalization to Novel Instances

**Results**

- Significant percentage of mistakes can be successfully predicted

# Generalization to Novel Instances

**Results**

- Significant percentage of mistakes can be successfully predicted

| | Without Attention | With Attention |
|---|---|---|
| % of mistakes that can be successfully predicted | 67.5% | 66.7% |

# Generalization to Novel Instances

**Results**

- Significant percentage of mistakes can be successfully predicted

| | Without Attention | With Attention |
|---|---|---|
| % of mistakes that can be successfully predicted | 67.5% | 66.7% |

- The analysis provides a way for models to predict their own oncoming failures

# Generalization to Novel Instances

**Results**

- Significant percentage of mistakes can be successfully predicted

| | Without Attention | With Attention |
|---|---|---|
| % of mistakes that can be successfully predicted | 67.5% | 66.7% |

- The analysis provides a way for models to predict their own oncoming failures → human-like models

# Test Sample

Q: What type of reception is being attended?

## Test Sample

Q: What type of reception is being attended?

Predicted Ans: cake

## Test Sample

Q: What type of reception is being attended?

GT Ans: wedding

Predicted Ans: cake

## Test Sample

## Nearest Neighbor Training Samples

Q: What type of reception is being attended?

GT Ans: wedding

Predicted Ans: cake

## Test Sample

Q: What type of reception is being attended?

GT Ans: wedding

Predicted Ans: cake

## Nearest Neighbor Training Samples

Q: What type of exercise equipment is shown?

GT Ans: bike

**Test Sample** | **Nearest Neighbor Training Samples**

Q: What type of reception is being attended?

GT Ans: wedding

Predicted Ans: cake

Q: What type of exercise equipment is shown?

GT Ans: bike

Q: What type of dessert is this man having?

GT Ans: cake

## Test Sample

Q: What type of reception is being attended?

GT Ans: wedding

Predicted Ans: cake

## Nearest Neighbor Training Samples

Q: What type of exercise equipment is shown?

GT Ans: bike

Q: What type of dessert is this man having?

GT Ans: cake

Q: What dessert is on the table?

GT Ans: cake

# Generalization to Novel Instances

Do VQA models make mistakes because test instances are too different from training ones?

1. Lower test accuracy $\Longrightarrow$ test QI pairs are too different from training QI pairs?

2. Lower test accuracy $\Longrightarrow$ test QI pairs are "familiar" **but** test labels are too different from training labels?

# Generalization to Novel Instances

**Experiment**

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair
2. Compute average distance (in Word2Vec space) between GT answers of k-NN training QI pairs and GT answer of test QI pair

# Generalization to Novel Instances

**Experiment**

1. Find k-NN training QI pairs, for each test QI pair
2. Compute average distance (in Word2Vec space) between GT answers of k-NN training QI pairs and GT answer of test QI pair
3. Measure correlation between average distance and test accuracy

# Generalization to Novel Instances

**Results**

# Generalization to Novel Instances

**Results**

Significant negative correlation

# Generalization to Novel Instances

**Results**

Significant negative correlation

| | Without Attention | With Attention |
|---|---|---|
| Correlation | -0.62 (@ k=50) | -0.62 (@ k=15) |

# Generalization to Novel Instances

**Results**

Significant negative correlation

| | Without Attention | With Attention |
|---|---|---|
| Correlation | -0.62 (@ k=50) | -0.62 (@ k=15) |

*VQA models tend to regurgitate answers seen during training*

## Test Sample

Q: What color are the safety cones?

# Test Sample

Q: What color are the safety cones?

Predicted Ans: orange

## Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

**Test Sample**

**Nearest Neighbor Training Samples**

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

## Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

## Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

## Test Sample

## Nearest Neighbor Training Samples

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange

## Test Sample

Q: What color are the safety cones?

GT Ans: green

Predicted Ans: orange

## Nearest Neighbor Training Samples

Q: What color are the cones?

GT Ans: orange

Q: What color is the cone?

GT Ans: orange


Q: What color are the cones?

GT Ans: orange

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

# Listening to the Entire Question

Q: How many horses are on the beach?

Predicted Ans: 2

# Listening to the Entire Question

Q: How

Predicted Ans?

Q: How many horses are on the beach?

Predicted Ans: 2

# Listening to the Entire Question

Q: How Predicted Ans?

Q: How many Predicted Ans?

Q: How many horses are on the beach?

Predicted Ans: 2

# Listening to the Entire Question

Q: How many horses are on the beach?

Predicted Ans: 2

Q: How Predicted Ans?

Q: How many Predicted Ans?

Q: How many horses

Q: How many horses are

Q: How many horses are on

Q: How many horses are on the

Q: How many horses are on the beach

Q: How many horses are on the beach?

# Listening to the Entire Question

Q: How many horses are on the beach?

Predicted Ans: 2

Q: How Predicted Ans?

Q: How many Predicted Ans?

Q: How many horses

Q: How many horses are

Q: How many horses are on

Q: How many horses are on the

Predicted Ans?

Q: How many horses are on the beach

Q: How many horses are on the beach?

# Listening to the Entire Question

**Experiment**

# Listening to the Entire Question

**Experiment**

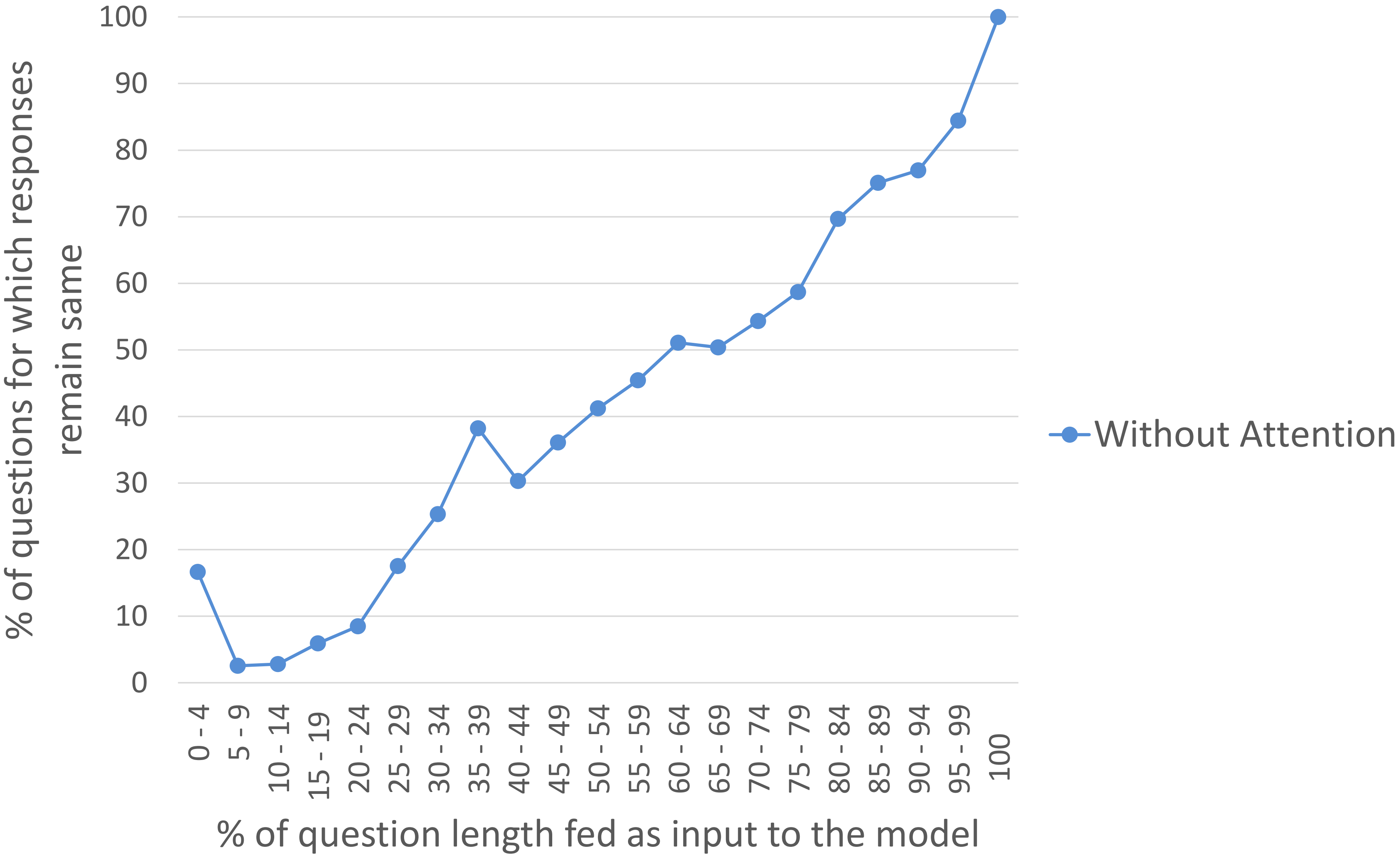
1. Test the model with partial questions of increasing lengths

# Listening to the Entire Question

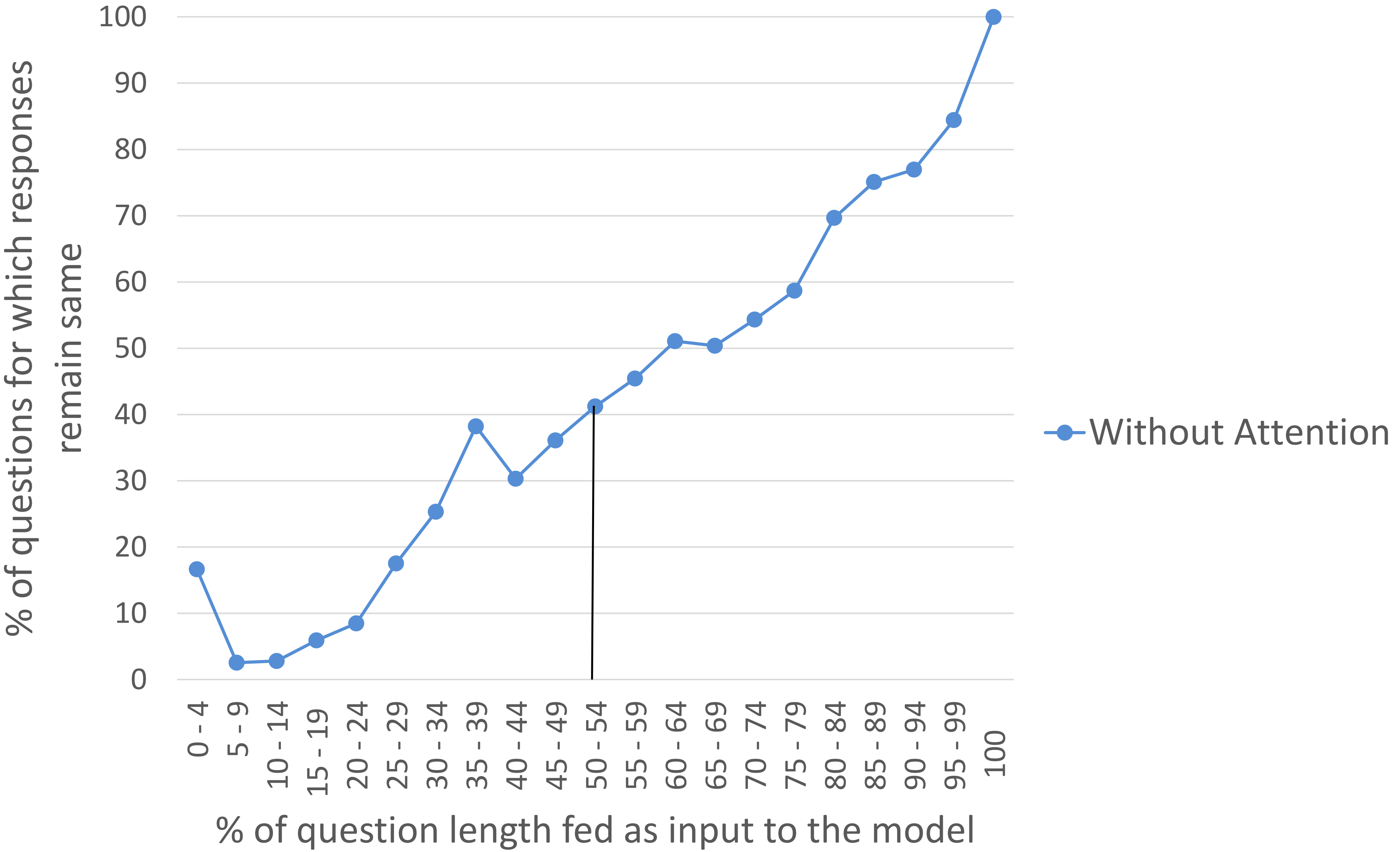
**Experiment**

1. Test the model with partial questions of increasing lengths
2. Compute percentage of questions for which partial question responses are same as full question responses

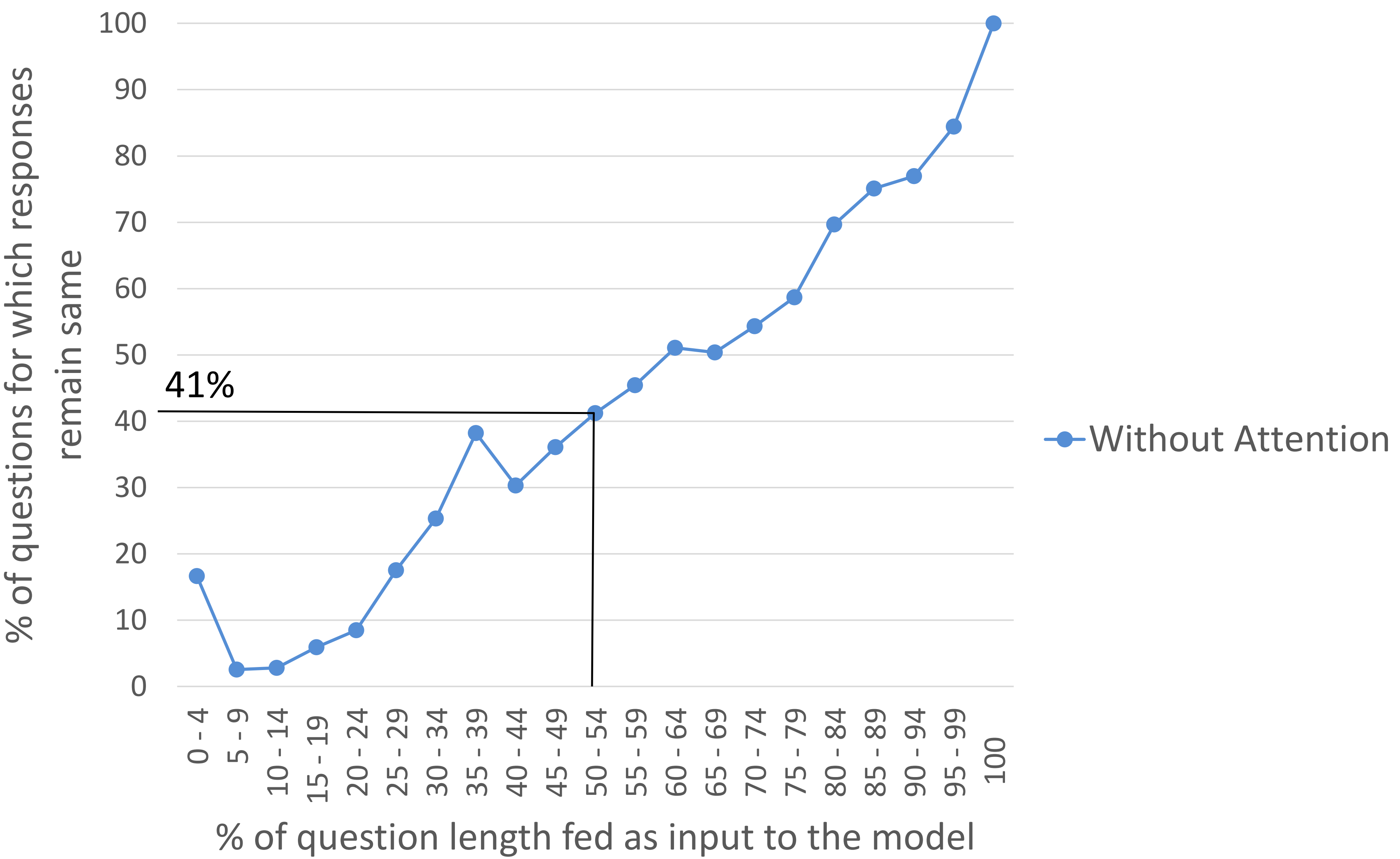
# Listening to the Entire Question



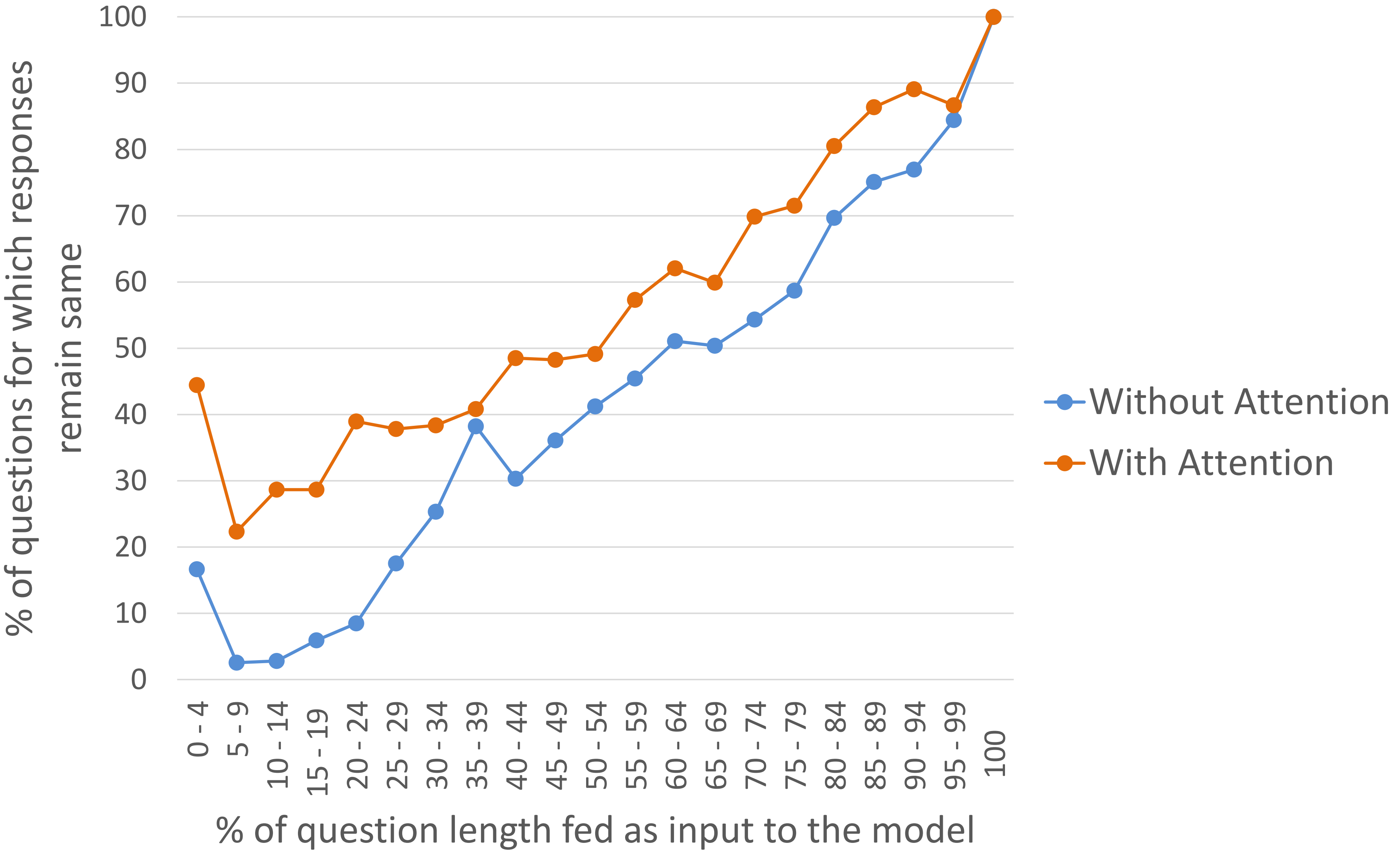
# Listening to the Entire Question



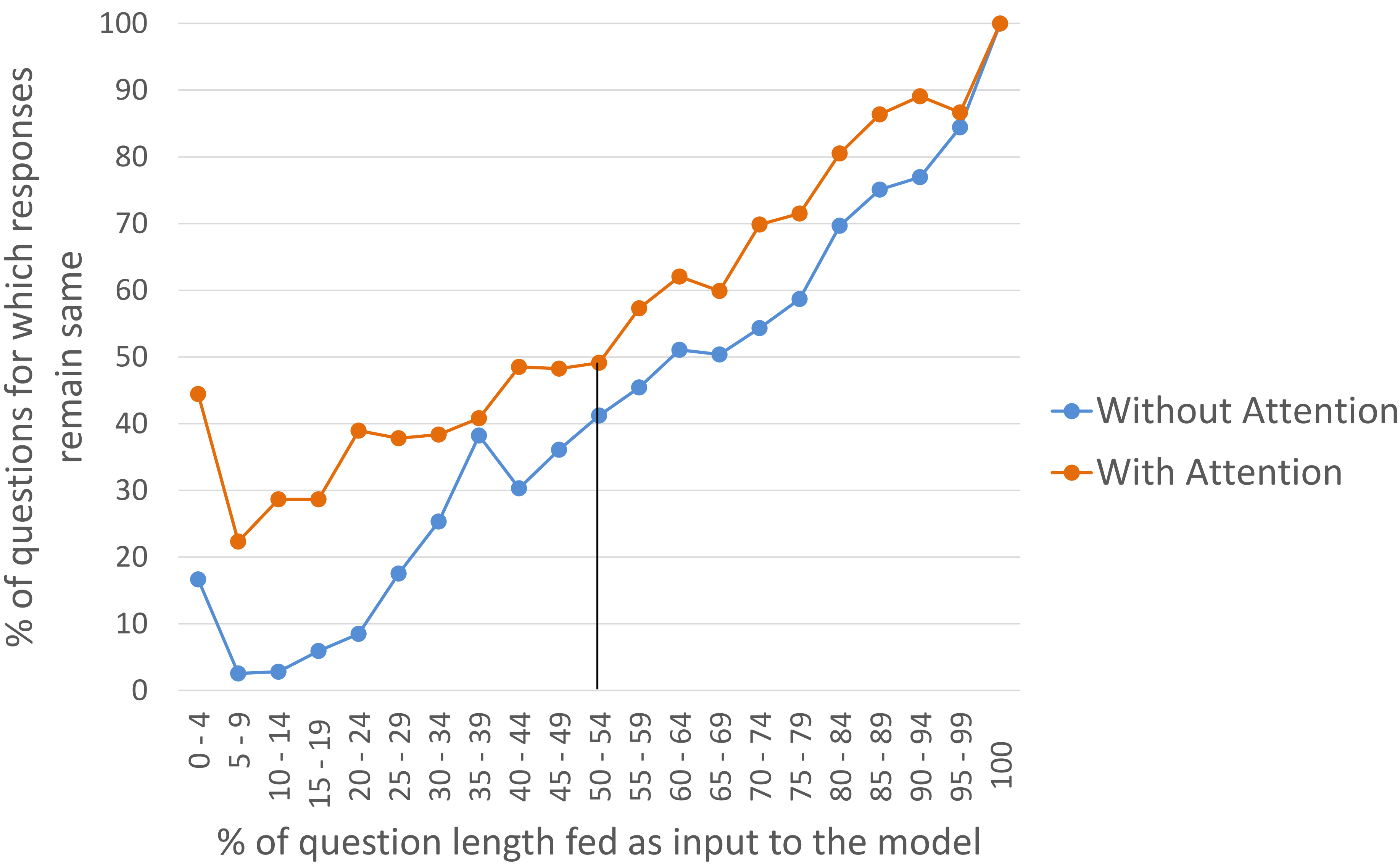
# Listening to the Entire Question



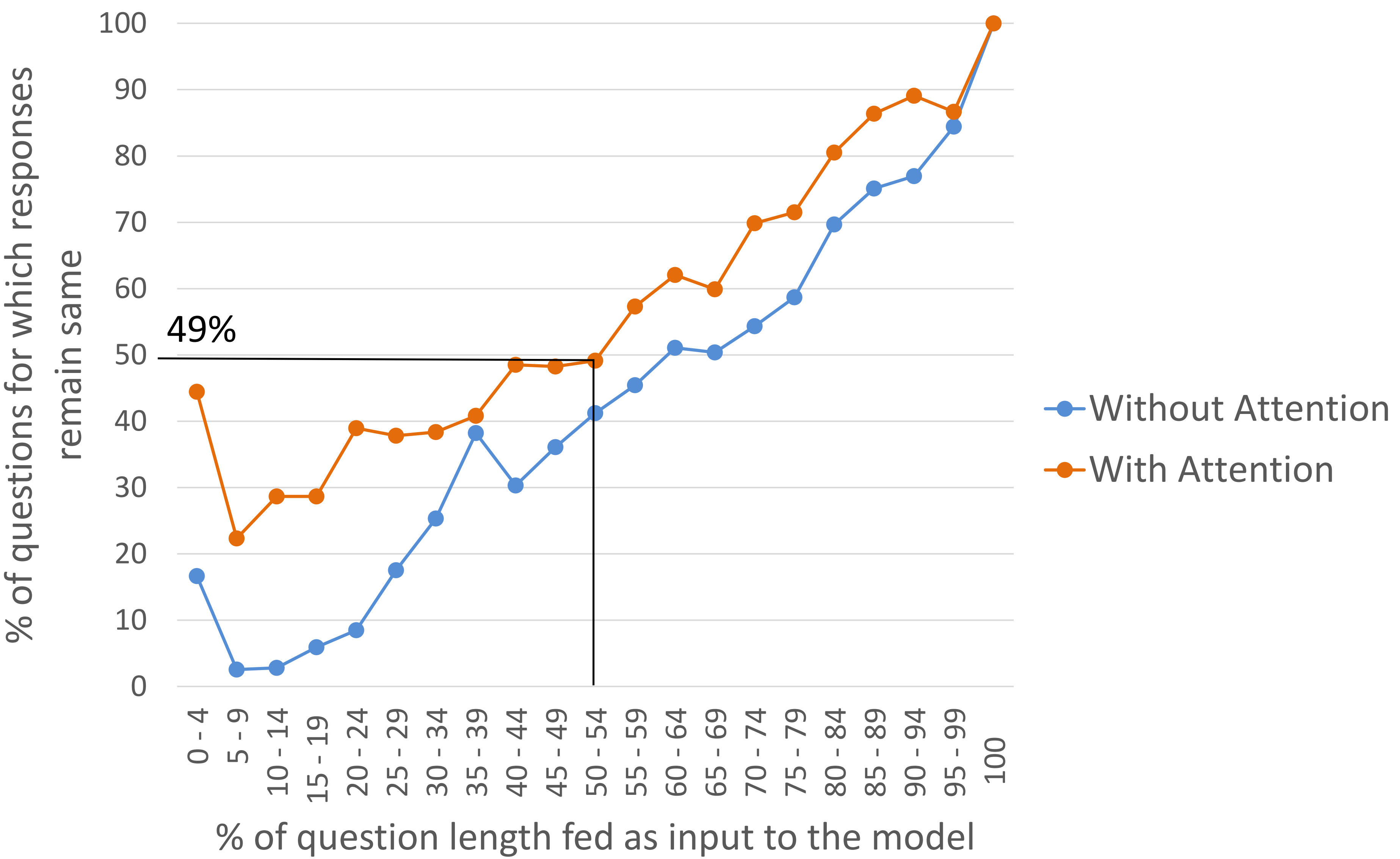
# Listening to the Entire Question



# Listening to the Entire Question



# Listening to the Entire Question



# Listening to the Entire Question

**Result**

VQA models converge on predicted answer after half the question for significant % of questions

# Listening to the Entire Question

**Result**

VQA models converge on predicted answer after half the question for significant % of questions

| | Without Attention | With Attention |
|---|---|---|
| % of questions | 41% | 49% |

# Listening to the Entire Question

**Result**

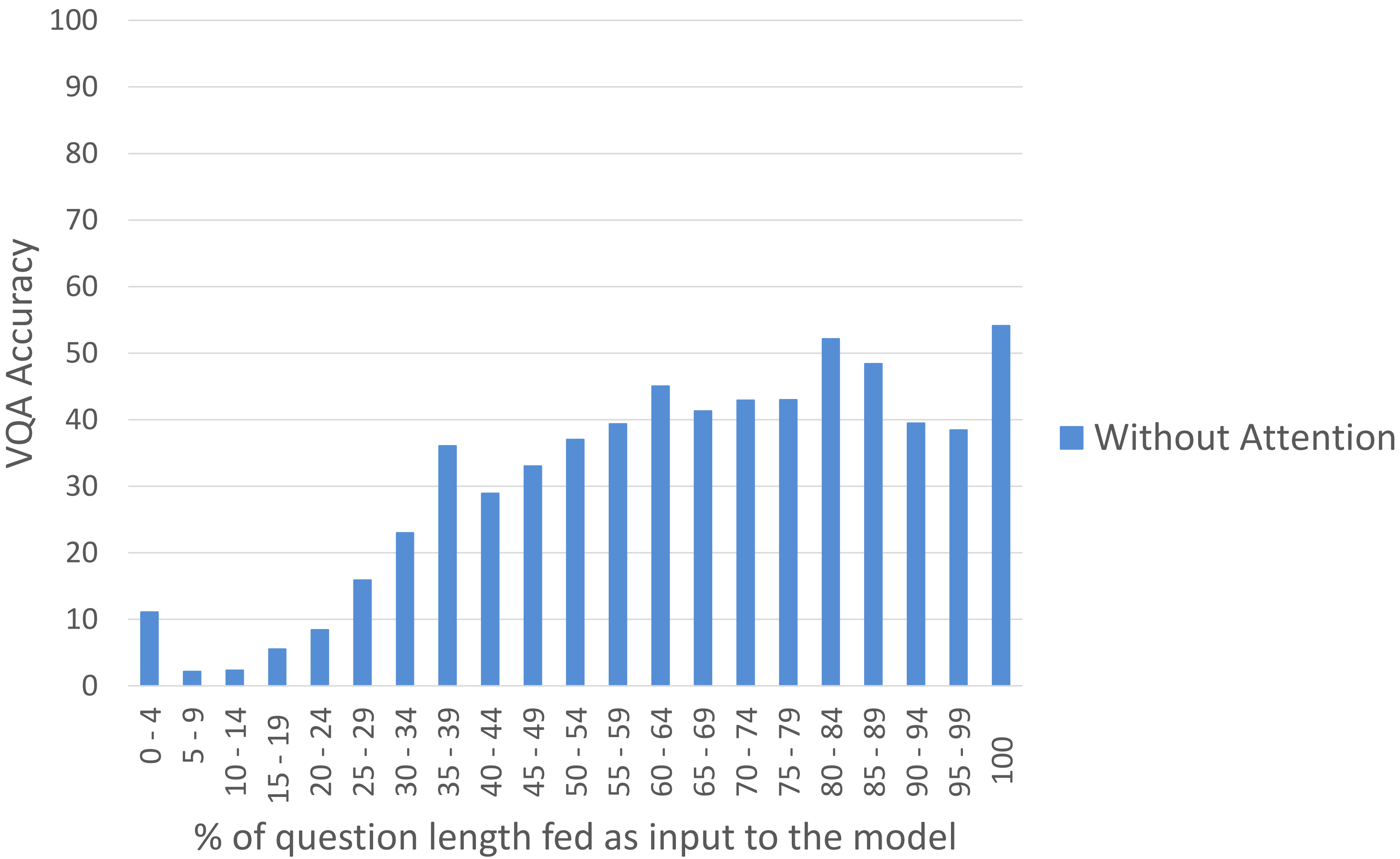
VQA models converge on predicted answer after half the question for significant % of questions

| | Without Attention | With Attention |
|---|---|---|
| % of questions | 41% | 49% |

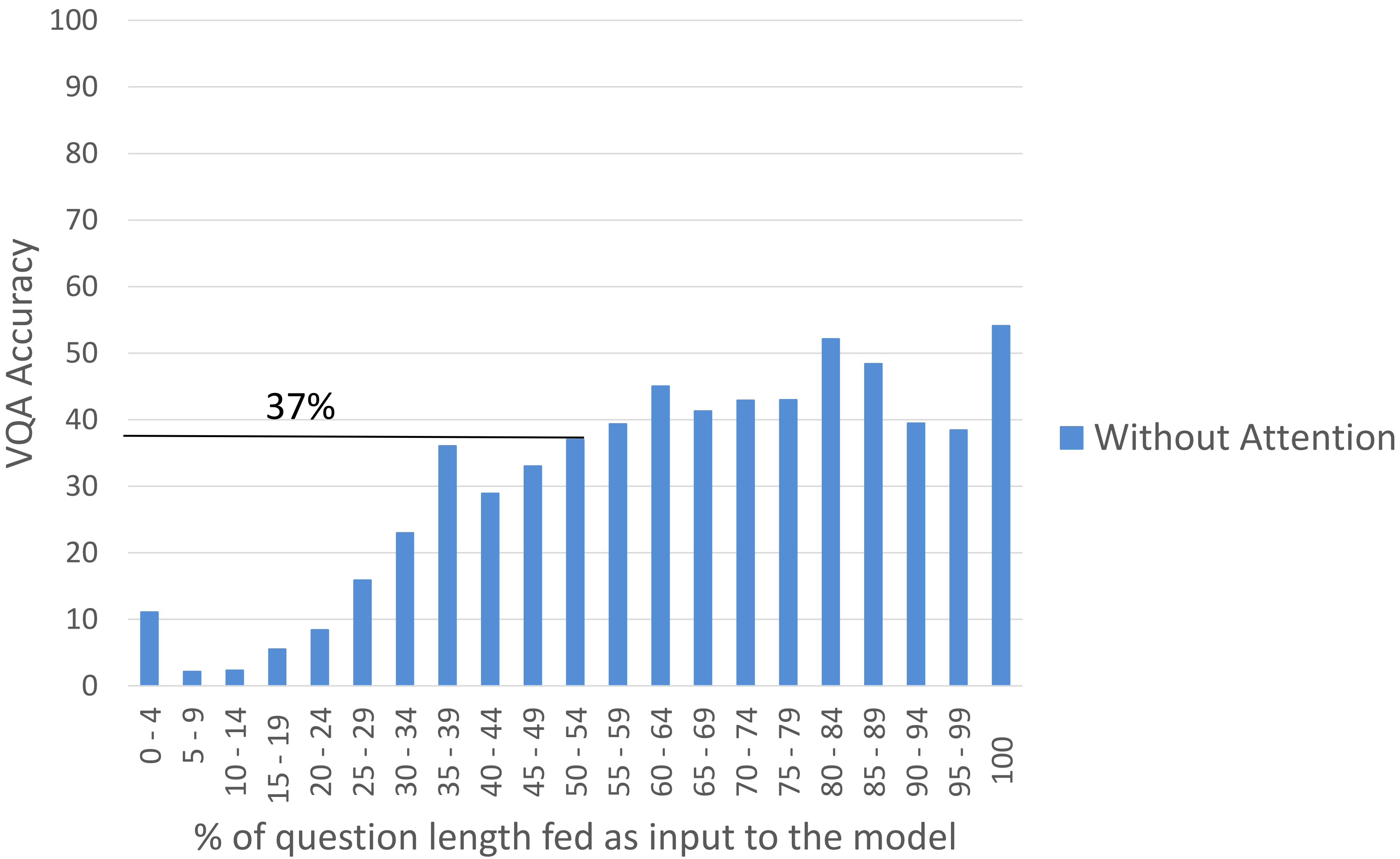
⇩

*VQA models often "jump to conclusions"*

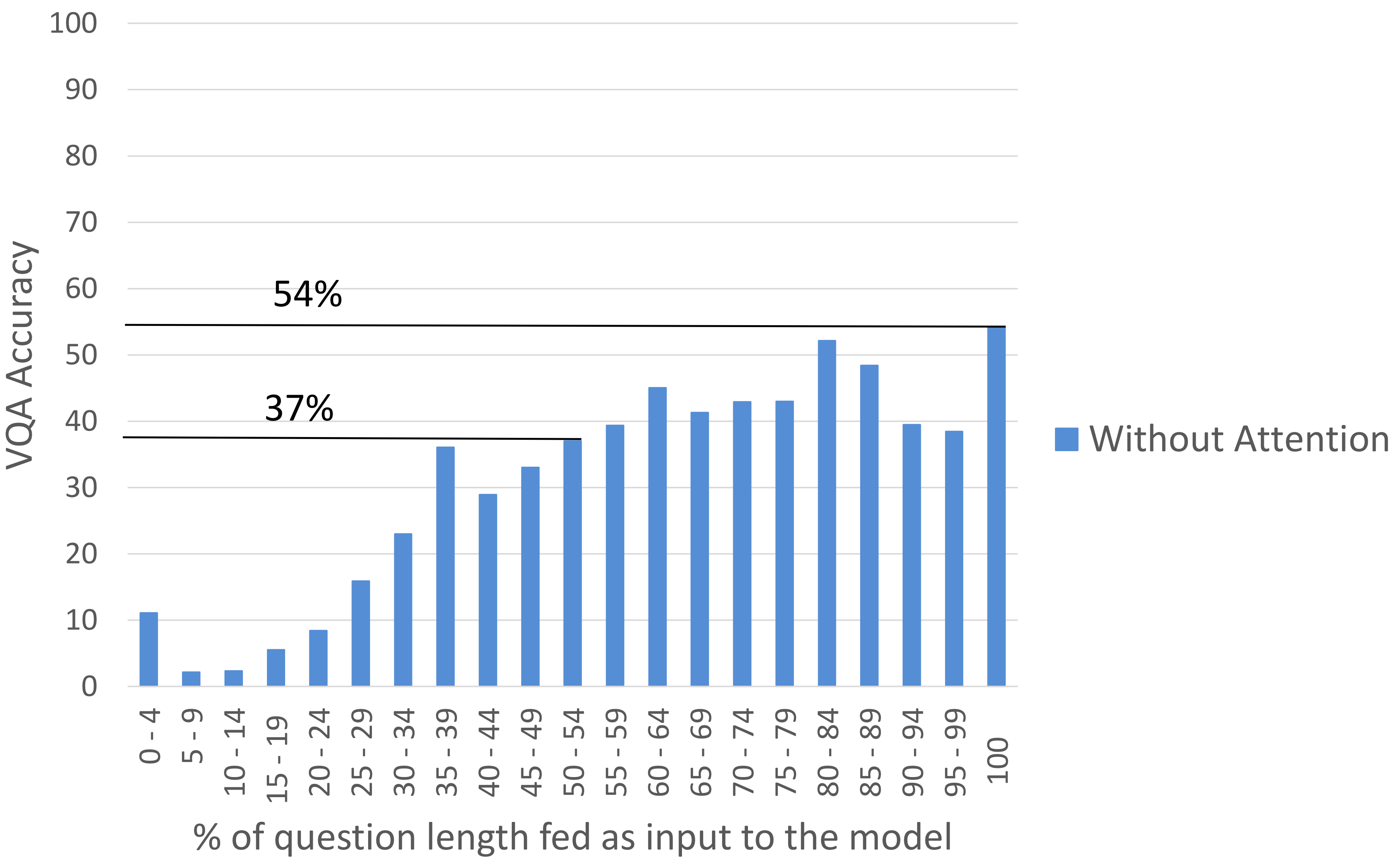
# Listening to the Entire Question



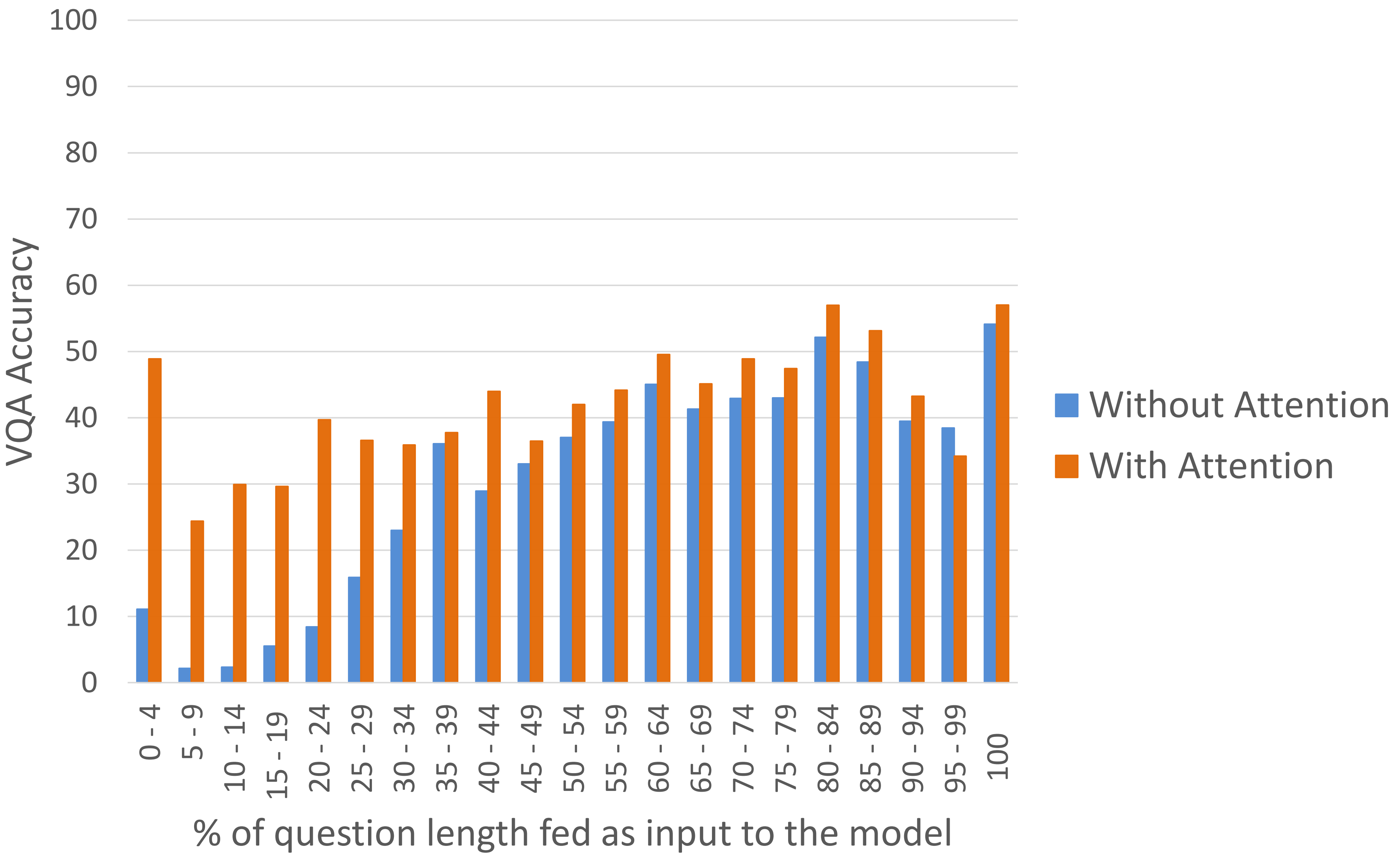
# Listening to the Entire Question



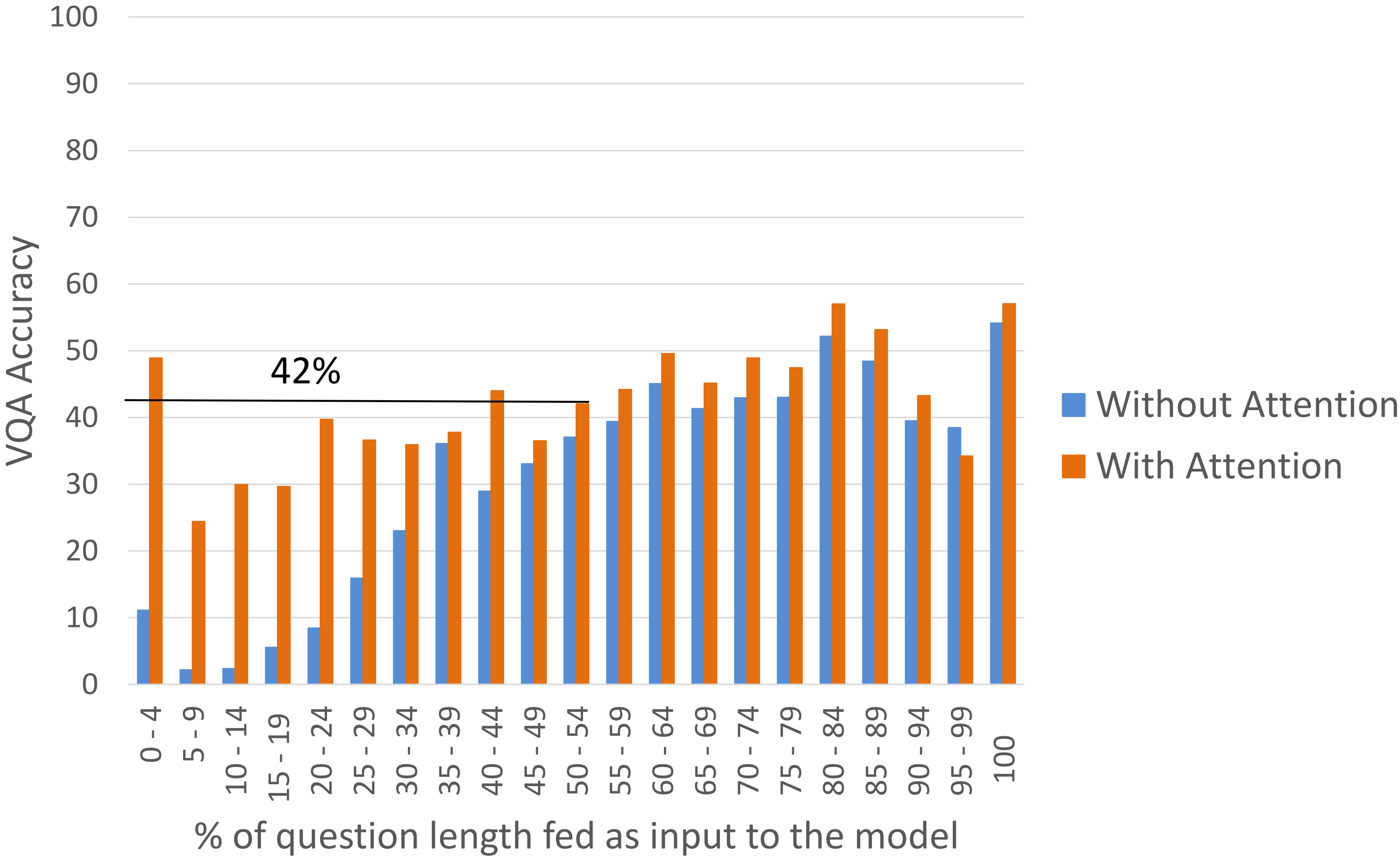
# Listening to the Entire Question



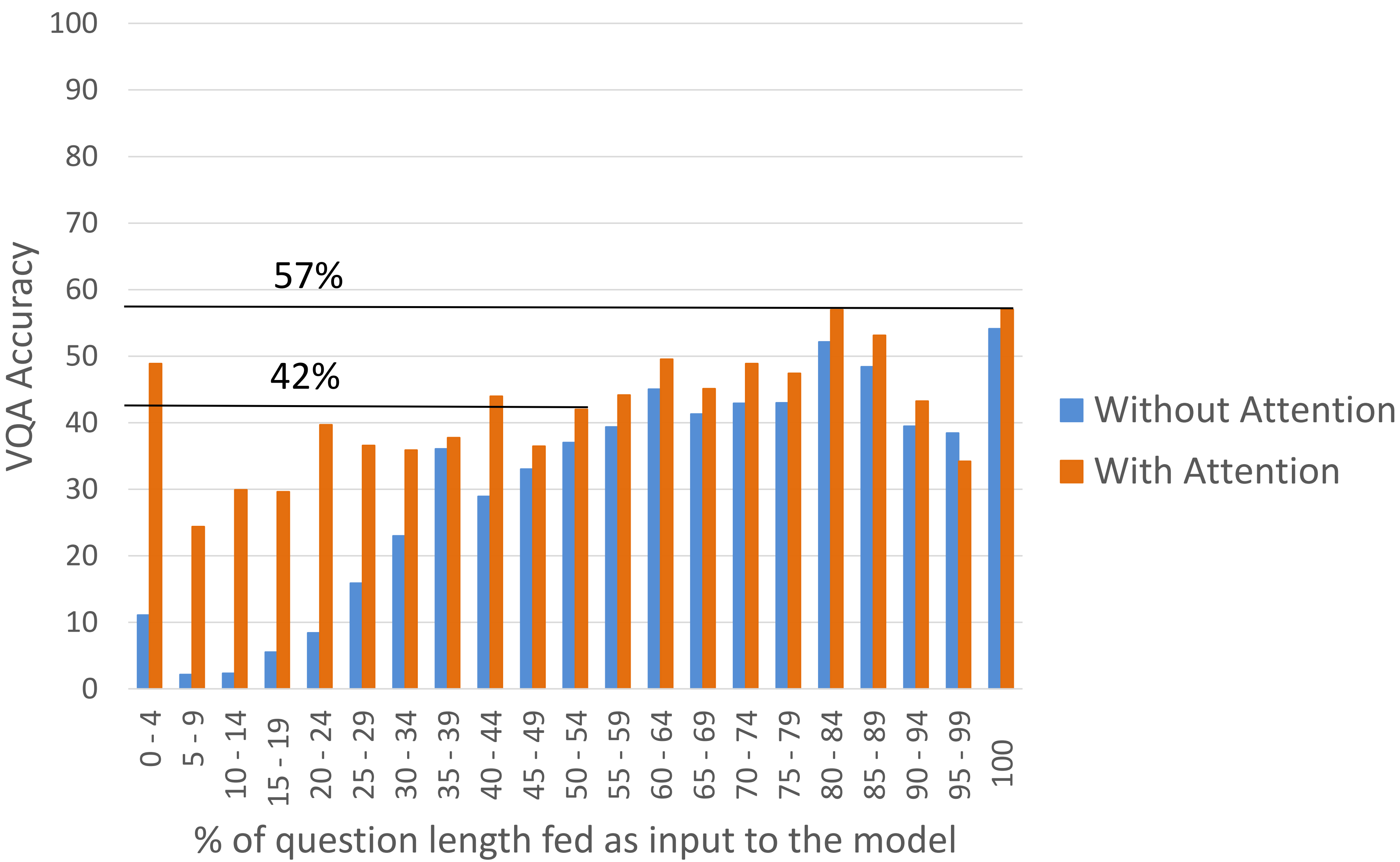
# Listening to the Entire Question



# Listening to the Entire Question



# Listening to the Entire Question



# Correct Response


**Q:** Are **A:** military

**Q:** Are they **A:** yes

**Q:** Are they playing **A:** yes

**Q:** Are they playing a **A:** yes

**Q:** Are they playing a game? **A:** yes

GT Ans: yes

# Incorrect Response

**Q:** How **A:** no

**Q:** How many **A:** 2

**Q:** How many horses **A:** 2

**Q:** How many horses are **A:** 2

**Q:** How many horses are on **A:** 2

**Q:** How many horses are on the **A:** 2

**Q:** How many horses are on the beach? **A:** 2

GT Ans: 6

# Incorrect Response

**Q:** Is **A:** kitchen

**Q:** Is the **A:** outside

**Q:** Is the bench **A:** no

**Q:** Is the bench made **A:** no

**Q:** Is the bench made of **A:** no

**Q:** Is the bench made of metal? **A:** no

GT Ans: yes

# Incorrect Response

**Q:** What **A:** umbrella

**Q:** What season **A:** summer

**Q:** What season of **A:** summer

**Q:** What season of year **A:** summer

**Q:** What season of year was **A:** summer

**Q:** What season of year was this **A:** summer

**Q:** What season of year was this photo **A:** summer

**Q:** What season of year was this photo taken **A:** summer

**Q:** What season of year was this photo taken in? **A:** summer

GT Ans: spring

# Outline

Do VQA models
generalize to novel instances?

Do VQA models
'listen' to the entire question?

Do VQA models
really 'look' at the image?

# Looking at the Image

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

Q: How many zebras?

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

Q: How many zebras?

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

Q: How many zebras?

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

Q: How many zebras?

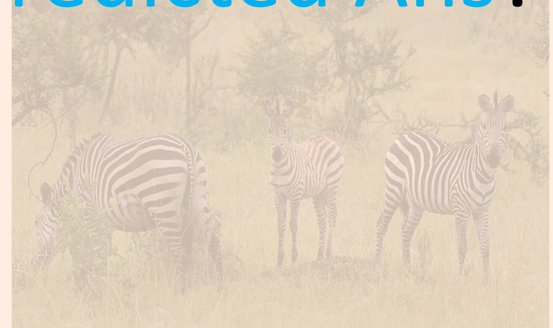
# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

Q: How many zebras?

Predicted Ans?



# Looking at the Image

**Experiment**

# Looking at the Image

**Experiment**

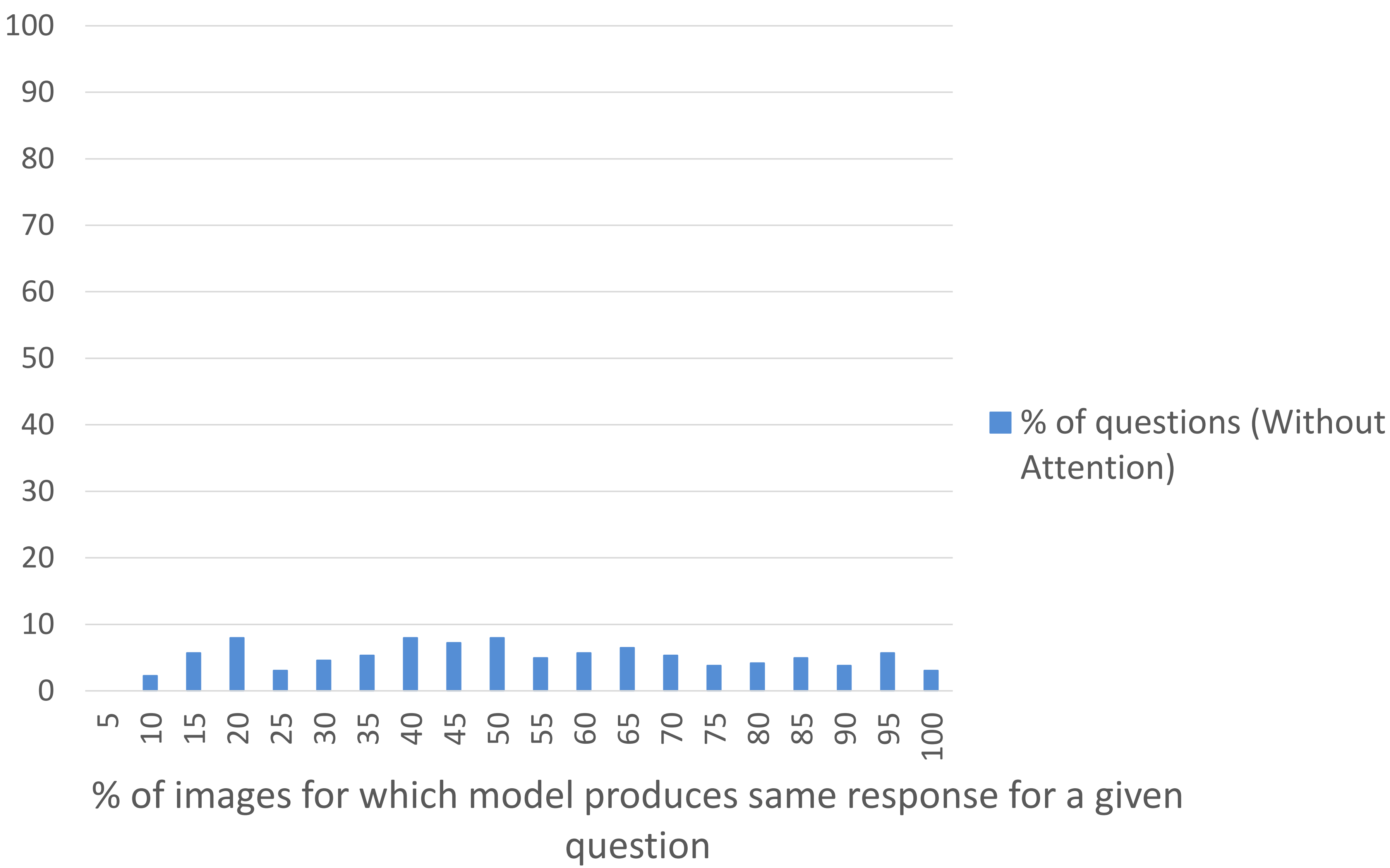
1. Compute the % of times (say X), the response does not change across images for a given question

# Looking at the Image

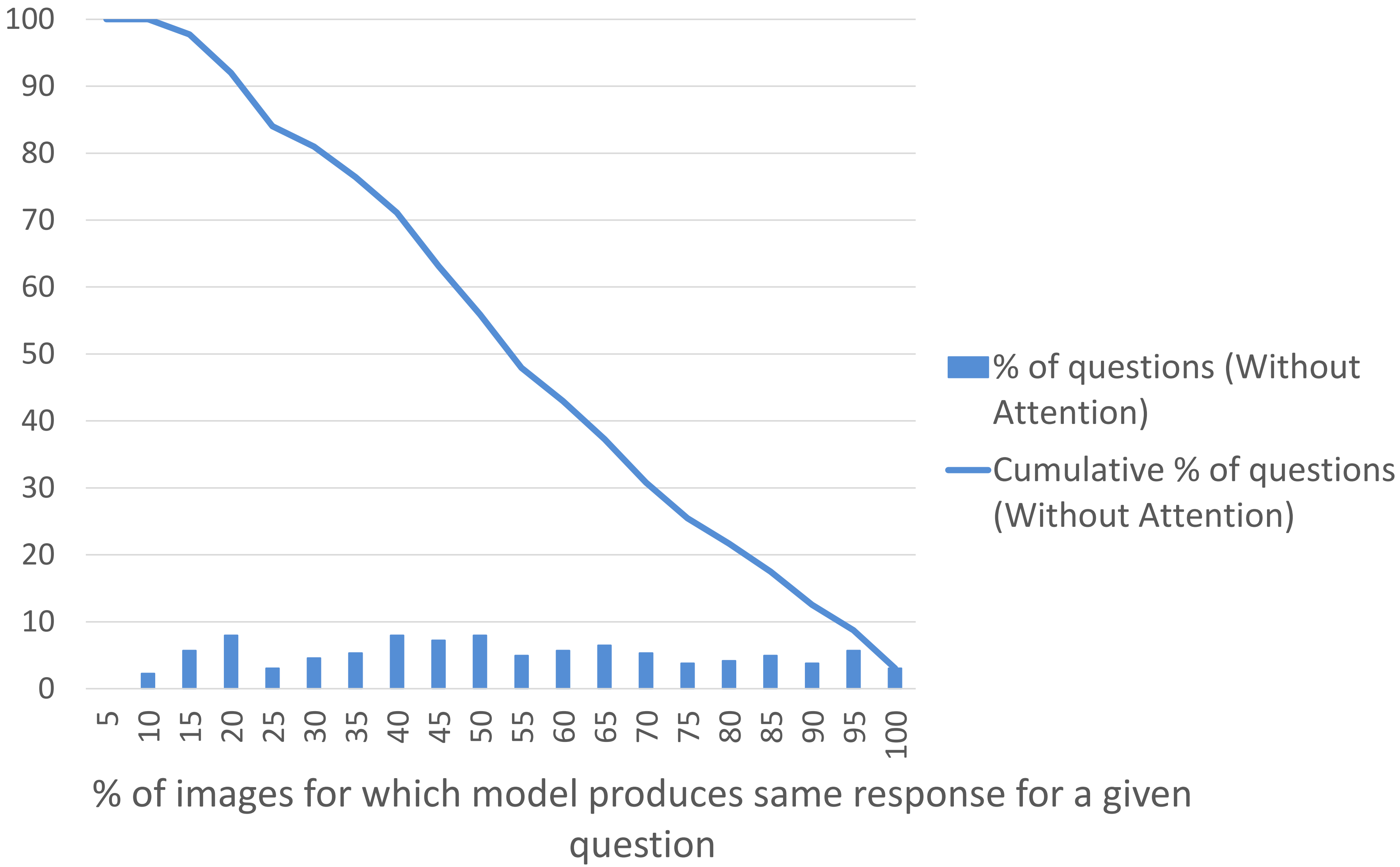
**Experiment**

1. Compute the % of times (say X), the response does not change across images for a given question
2. Plot histogram of X across questions

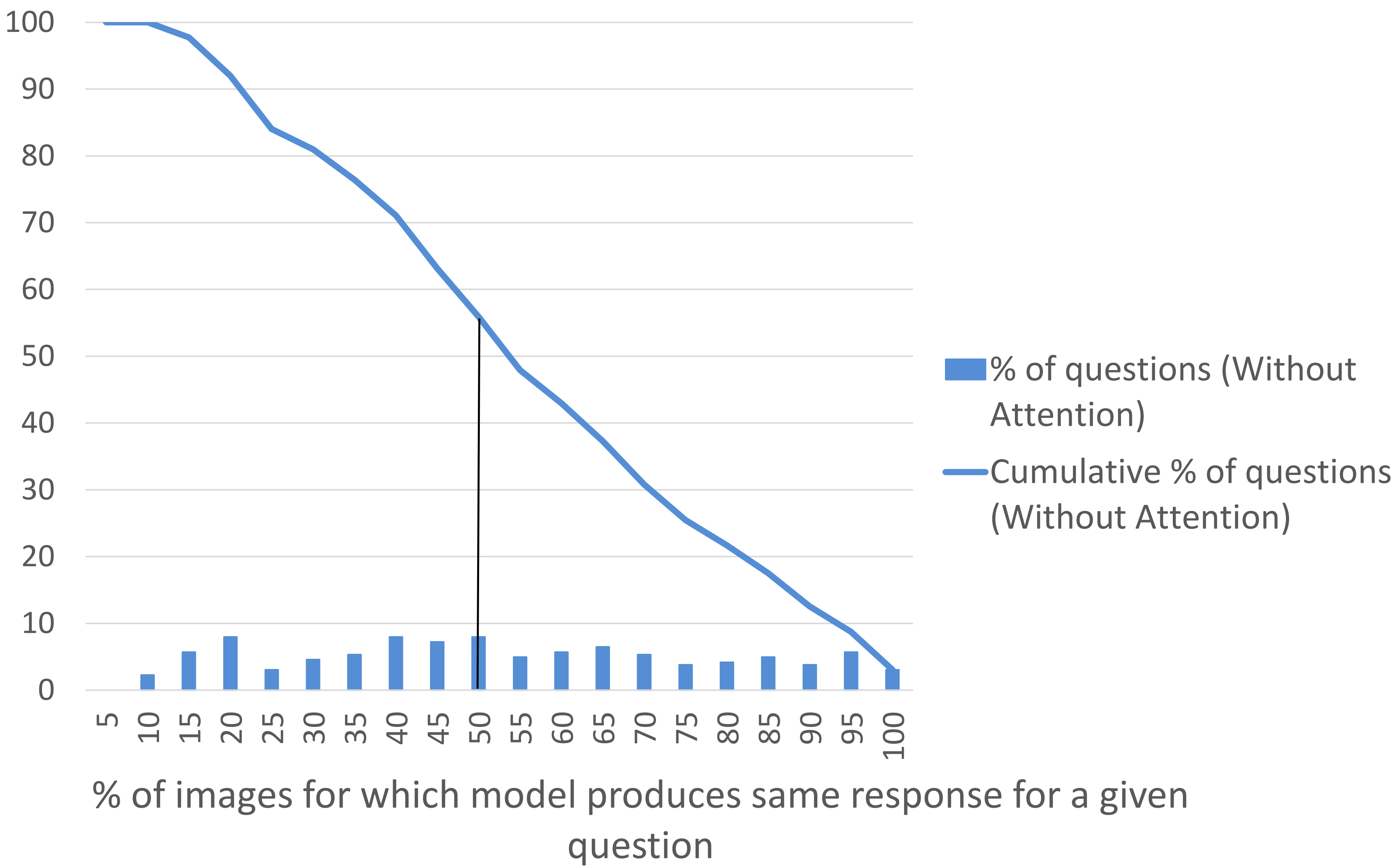
# Looking at the Image



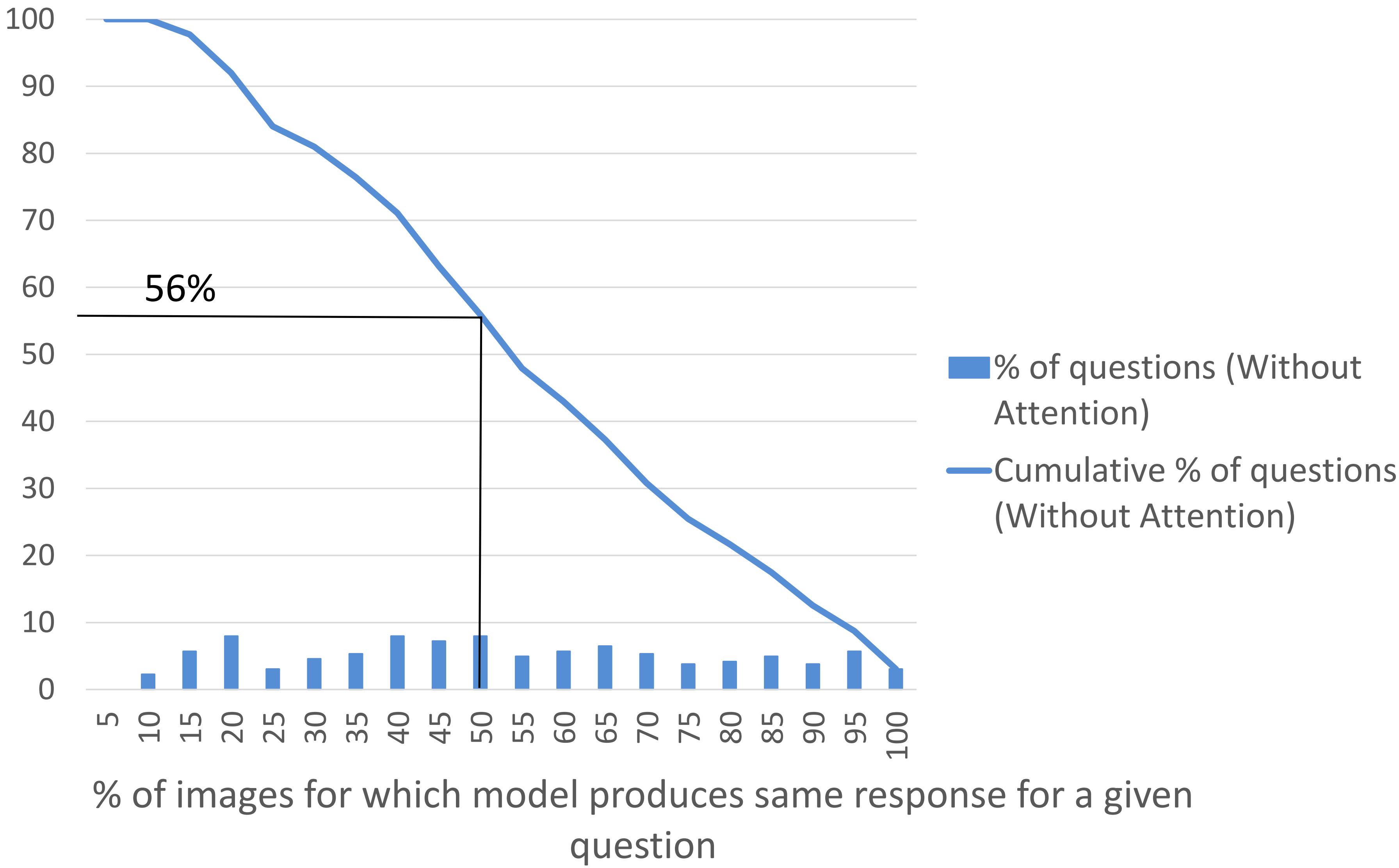
# Looking at the Image



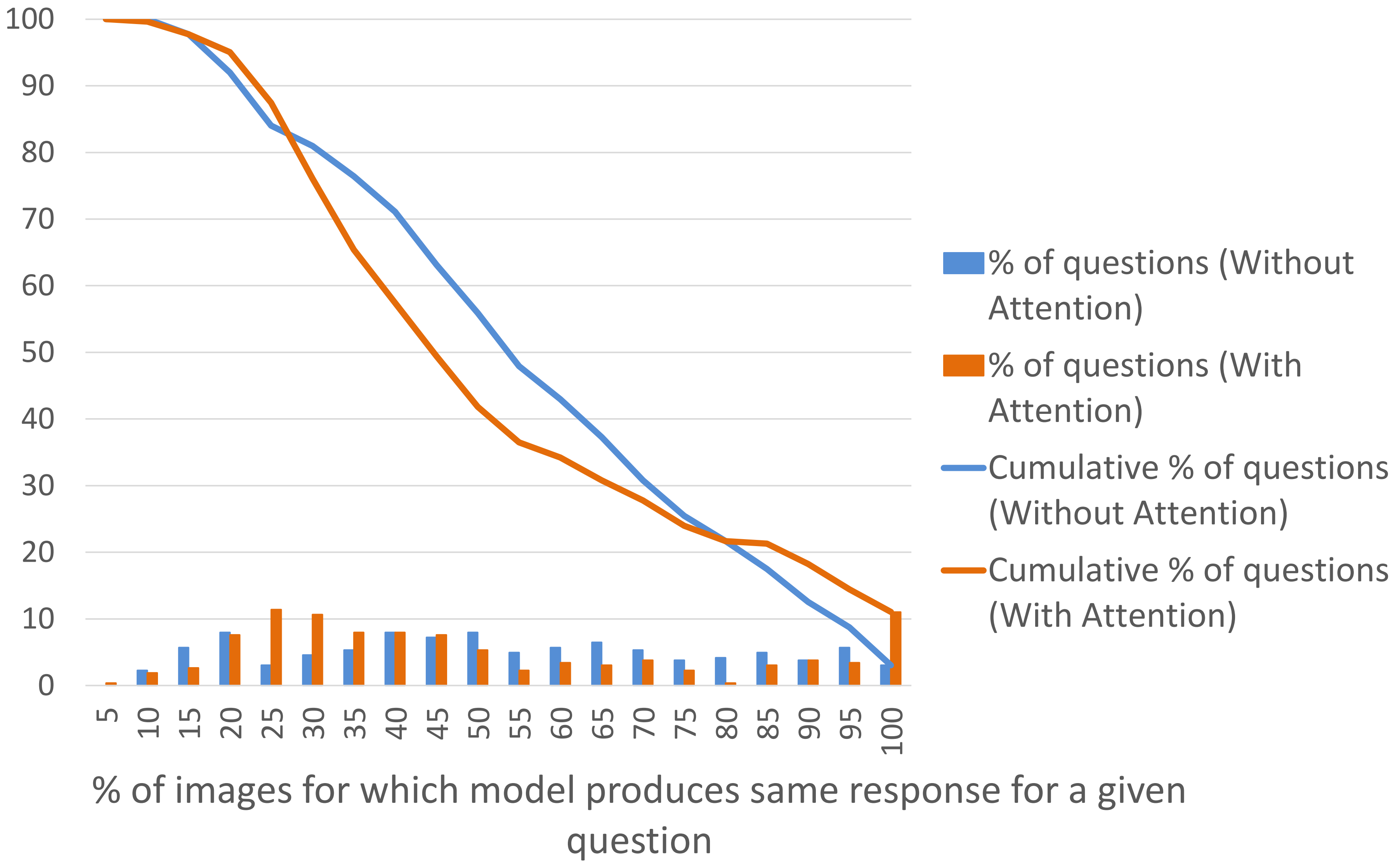
# Looking at the Image



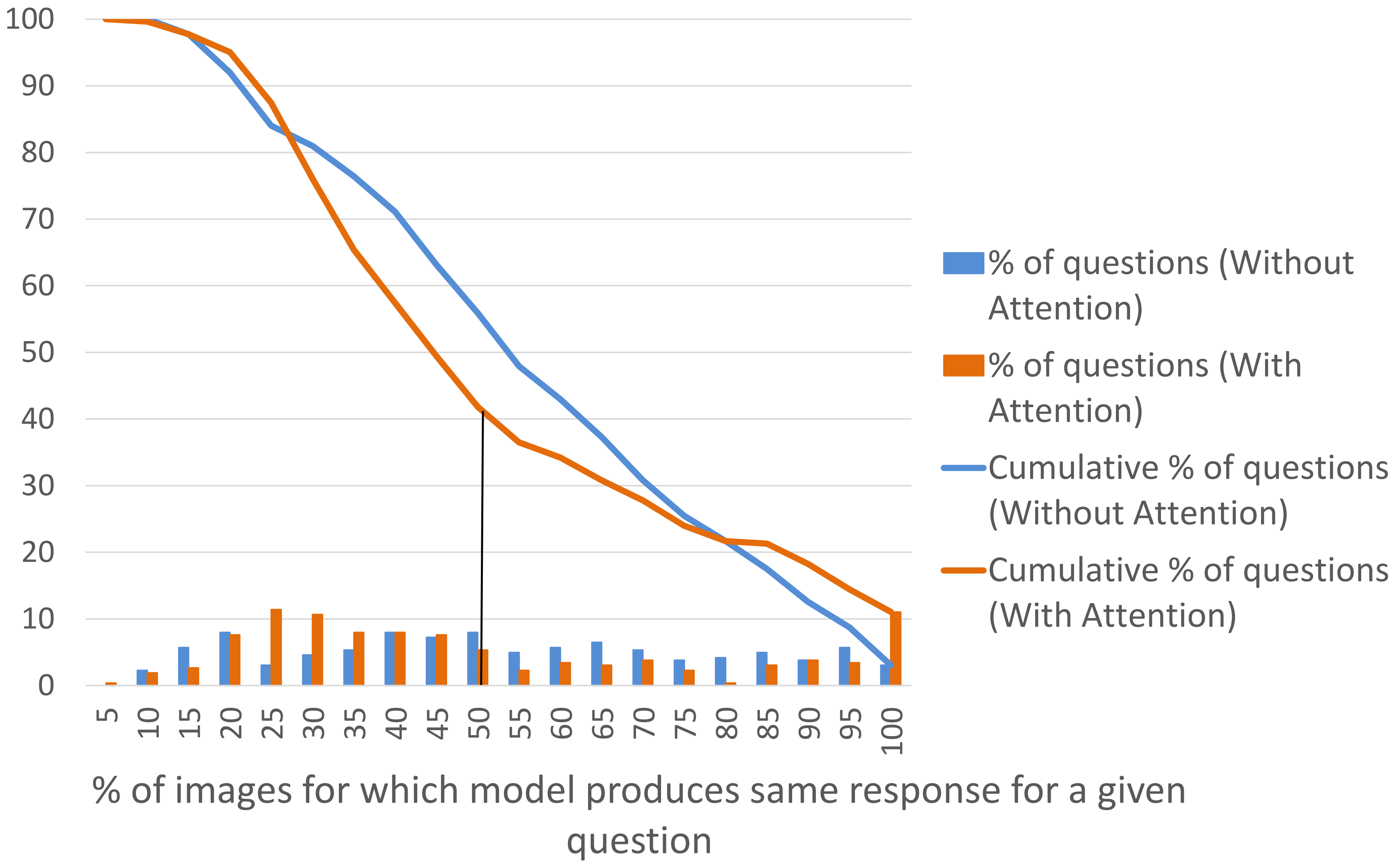
# Looking at the Image



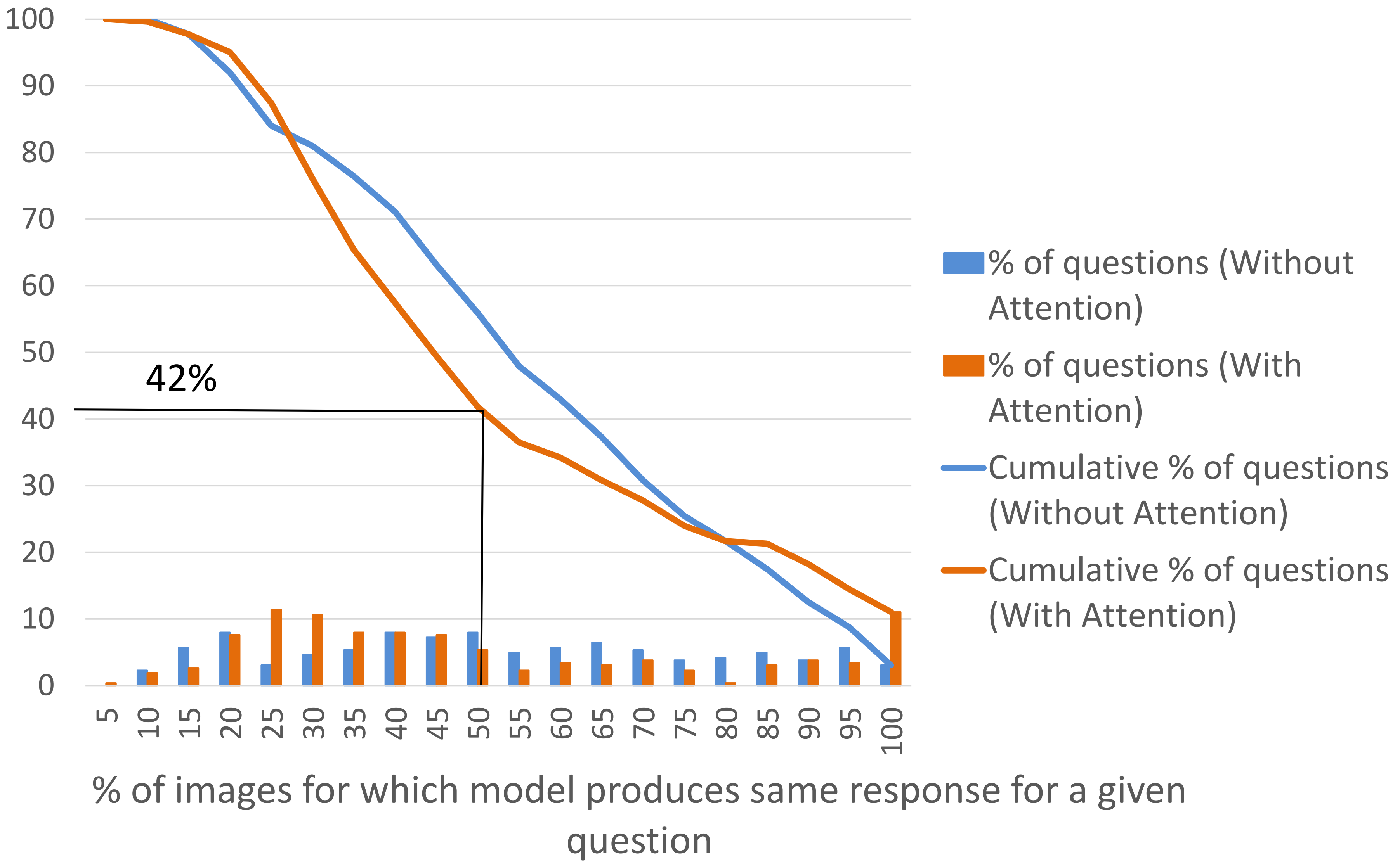
# Looking at the Image



# Looking at the Image



# Looking at the Image



# Looking at the Image

**Results**

1. VQA models do not change answers across images for significant % of questions

# Looking at the Image

**Results**

1. VQA models do not change answers across images for significant % of questions

| | Without Attention | With Attention |
|---|---|---|
| % of questions | 56% | 42% |

# Looking at the Image

**Results**

1. VQA models do not change answers across images for significant % of questions

| | Without Attention | With Attention |
|---|---|---|
| % of questions | 56% | 42% |

⇩

*VQA models are "stubborn"*

# Looking at the Image

**Results**

1. VQA models do not change answers across images for significant % of questions

| | Without Attention | With Attention |
|---|---|---|
| % of questions | 56% | 42% |

⇩

*VQA models are "stubborn"*

*Attention based models are less "stubborn" than non-attention based models*

# Looking at the Image

# Looking at the Image

Q: What does the red sign say?

# Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

# Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

**Correct Response**

# Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

**Correct Response**


**Incorrect Responses**

# Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

**Correct Response**


**Incorrect Responses**


# Looking at the Image

Q: What does the red sign say?

Predicted Ans: stop

## Correct Response


## Incorrect Responses



# Looking at the Image

# Looking at the Image

Q: How many zebras?

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

**Correct Response**

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

**Correct Response**

**Incorrect Responses**

# Looking at the Image

Q: How many zebras?

Predicted Ans: 2

**Correct Response**

**Incorrect Responses**

# Looking at the Image

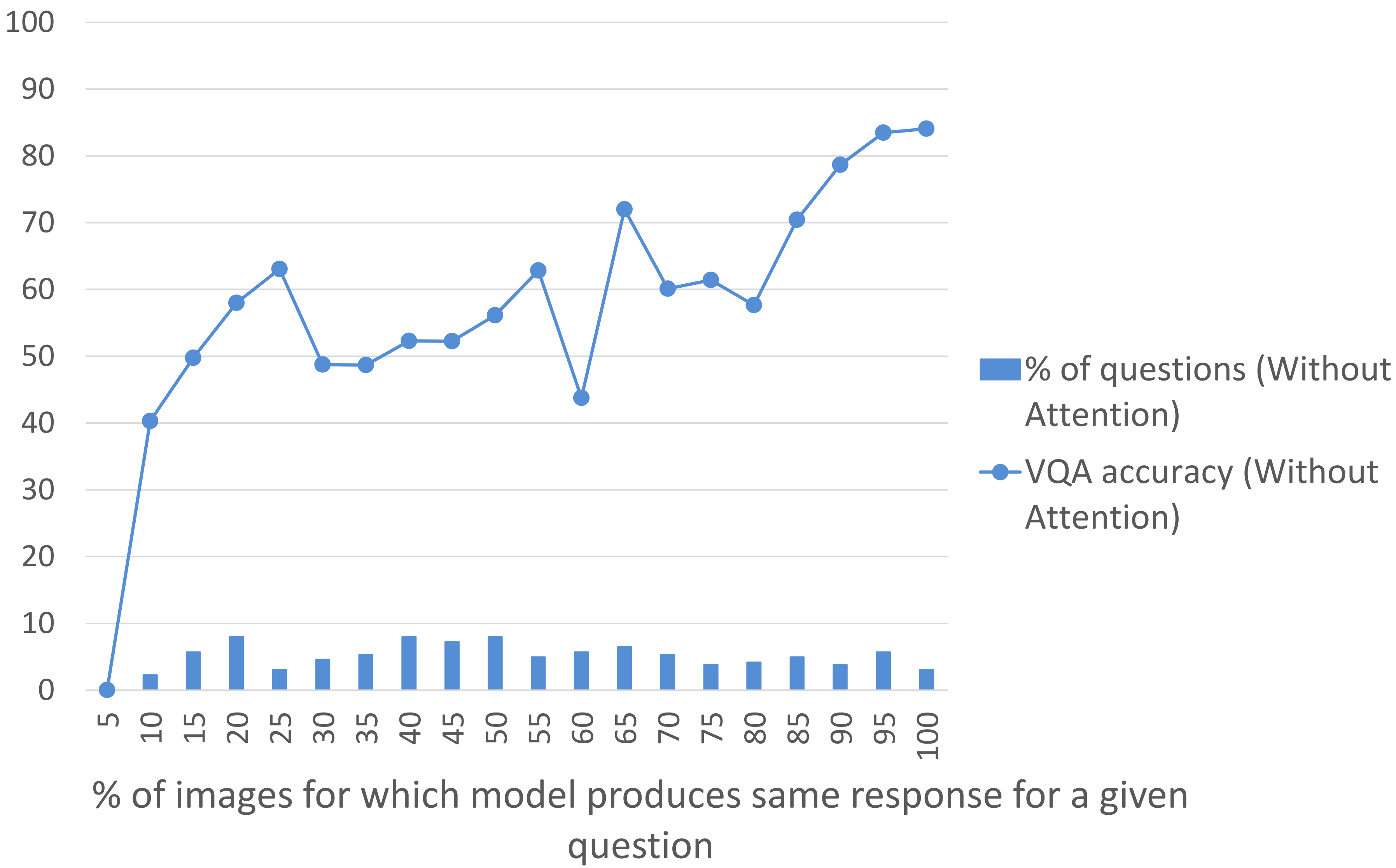
Q: How many zebras?

Predicted Ans: 2

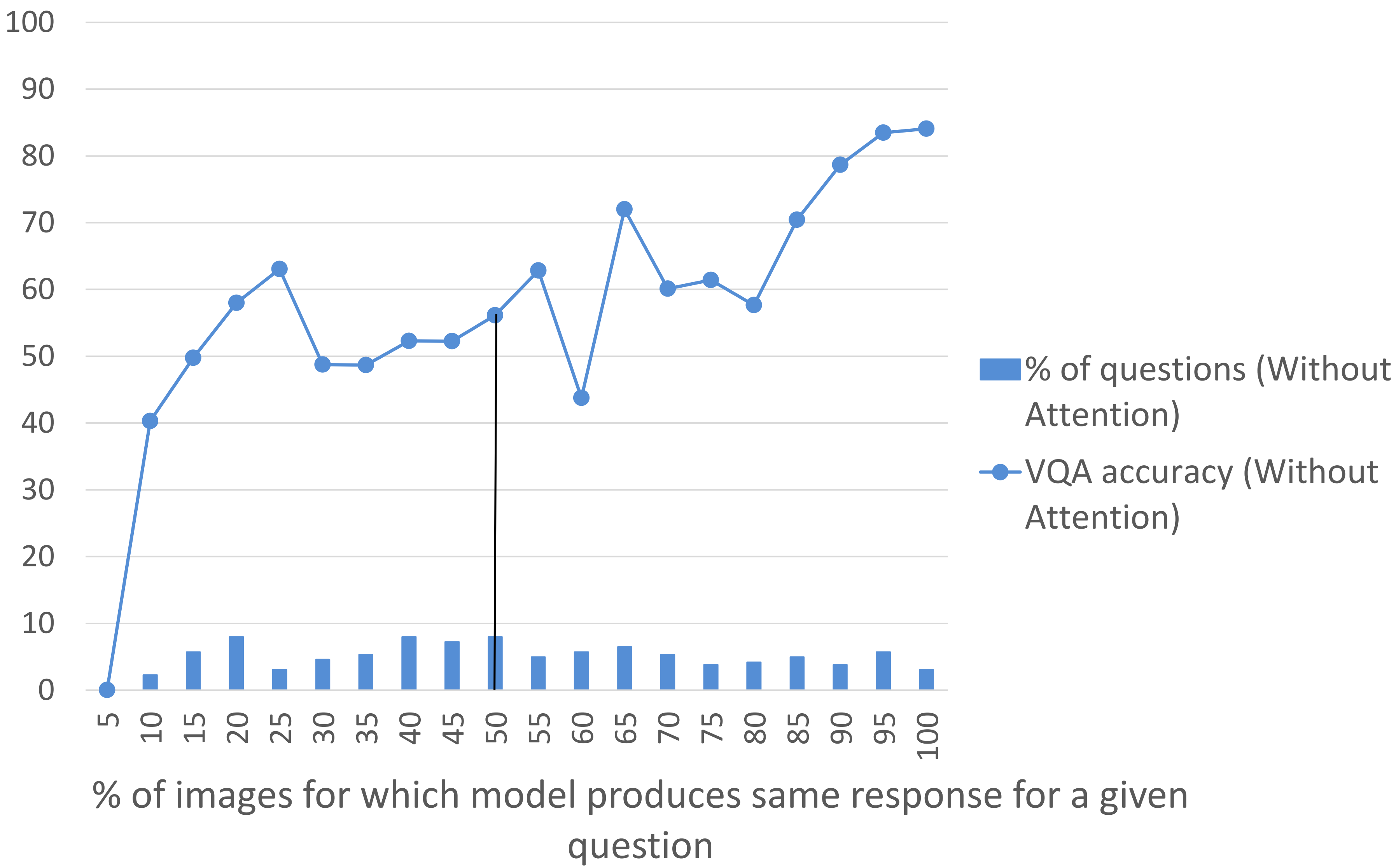
**Correct Response**

**Incorrect Responses**

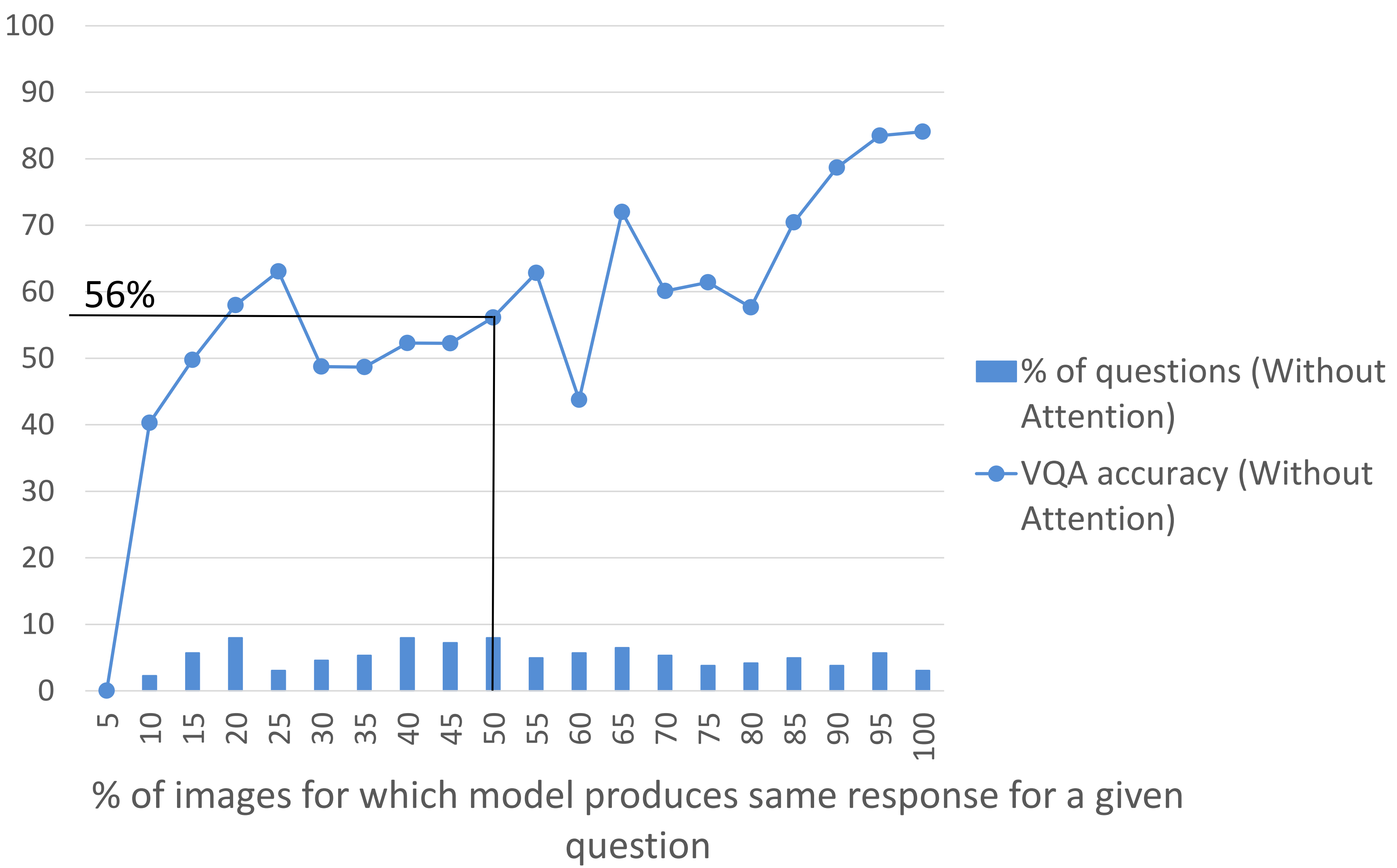
# Looking at the Image



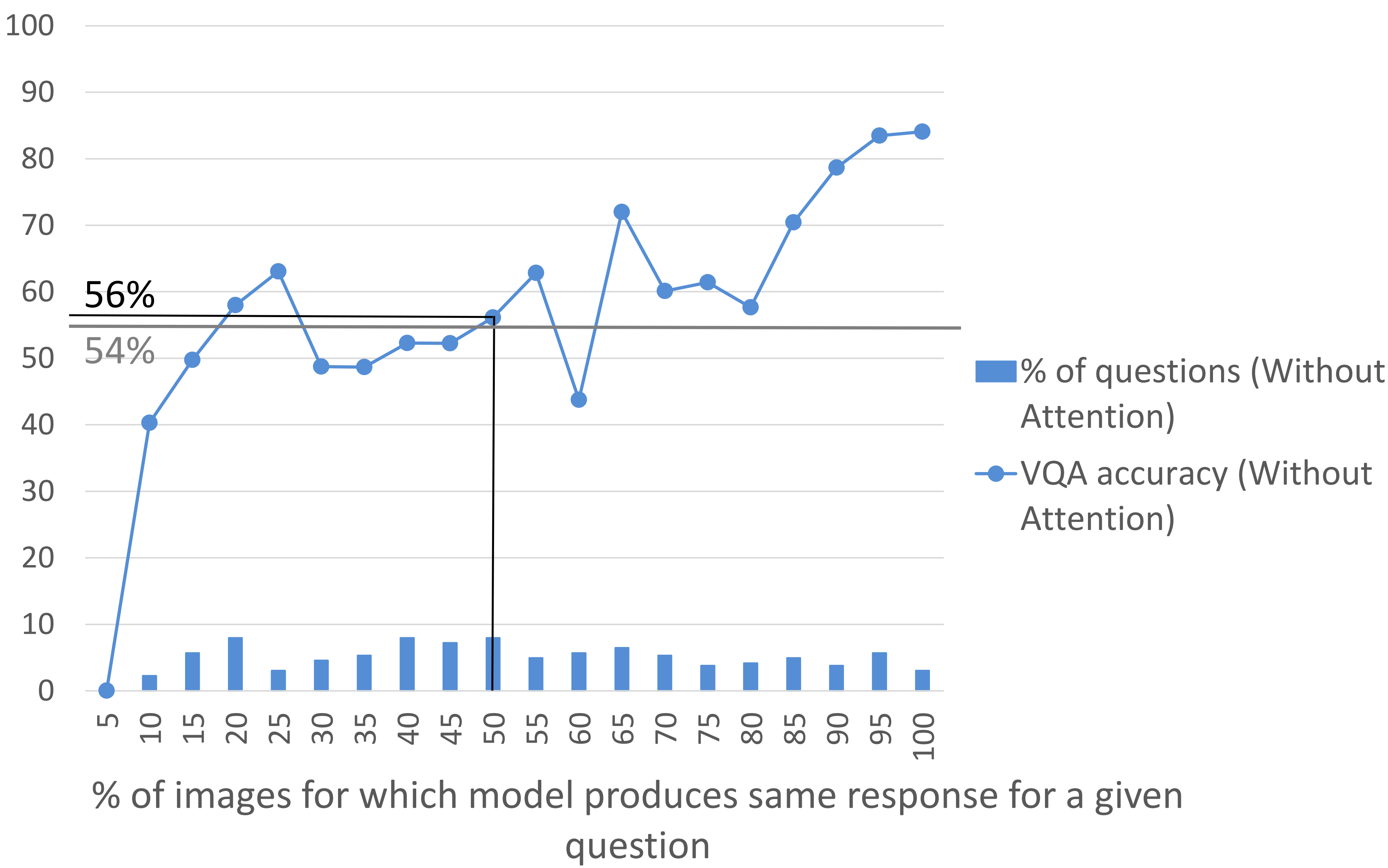
# Looking at the Image



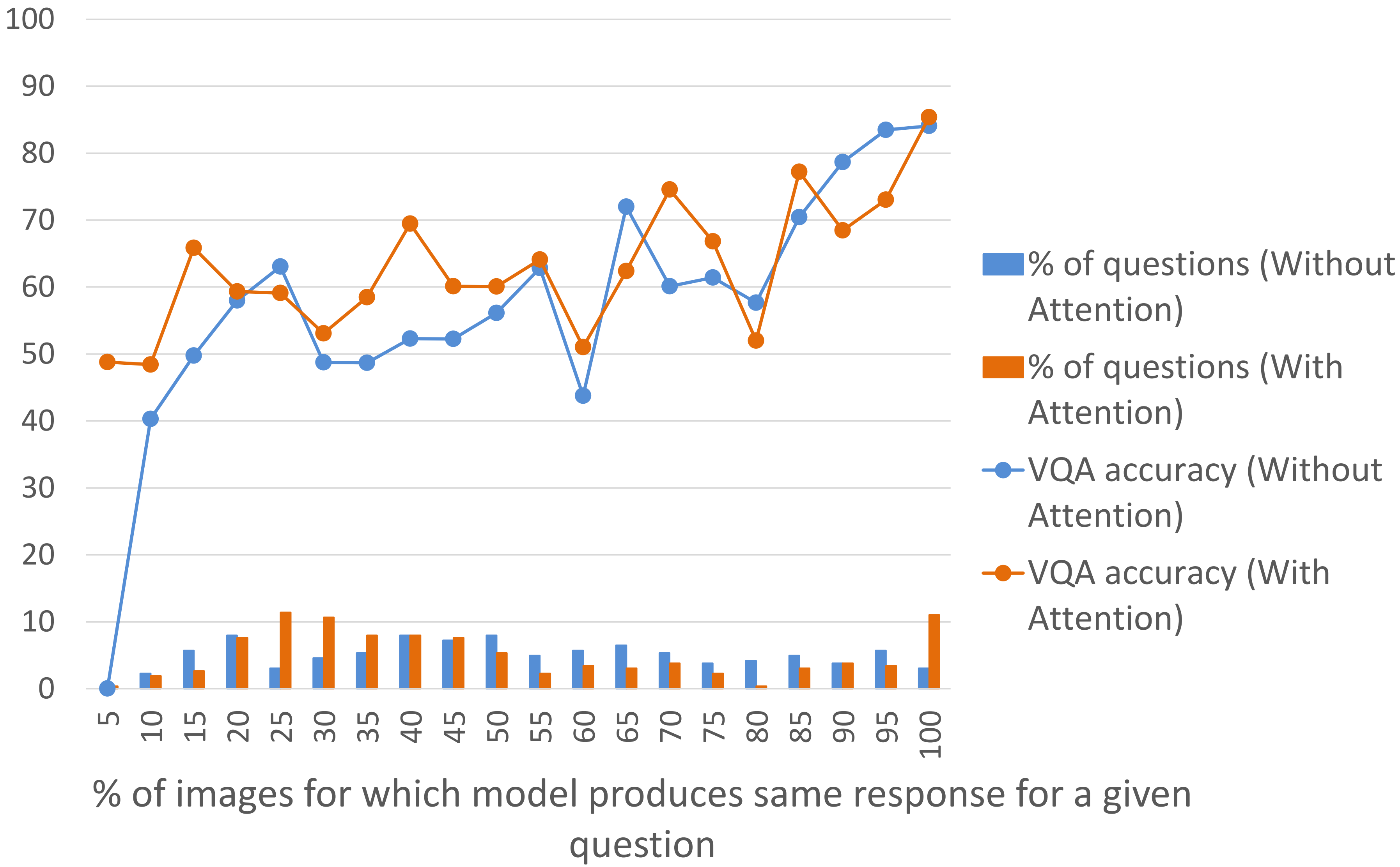
# Looking at the Image



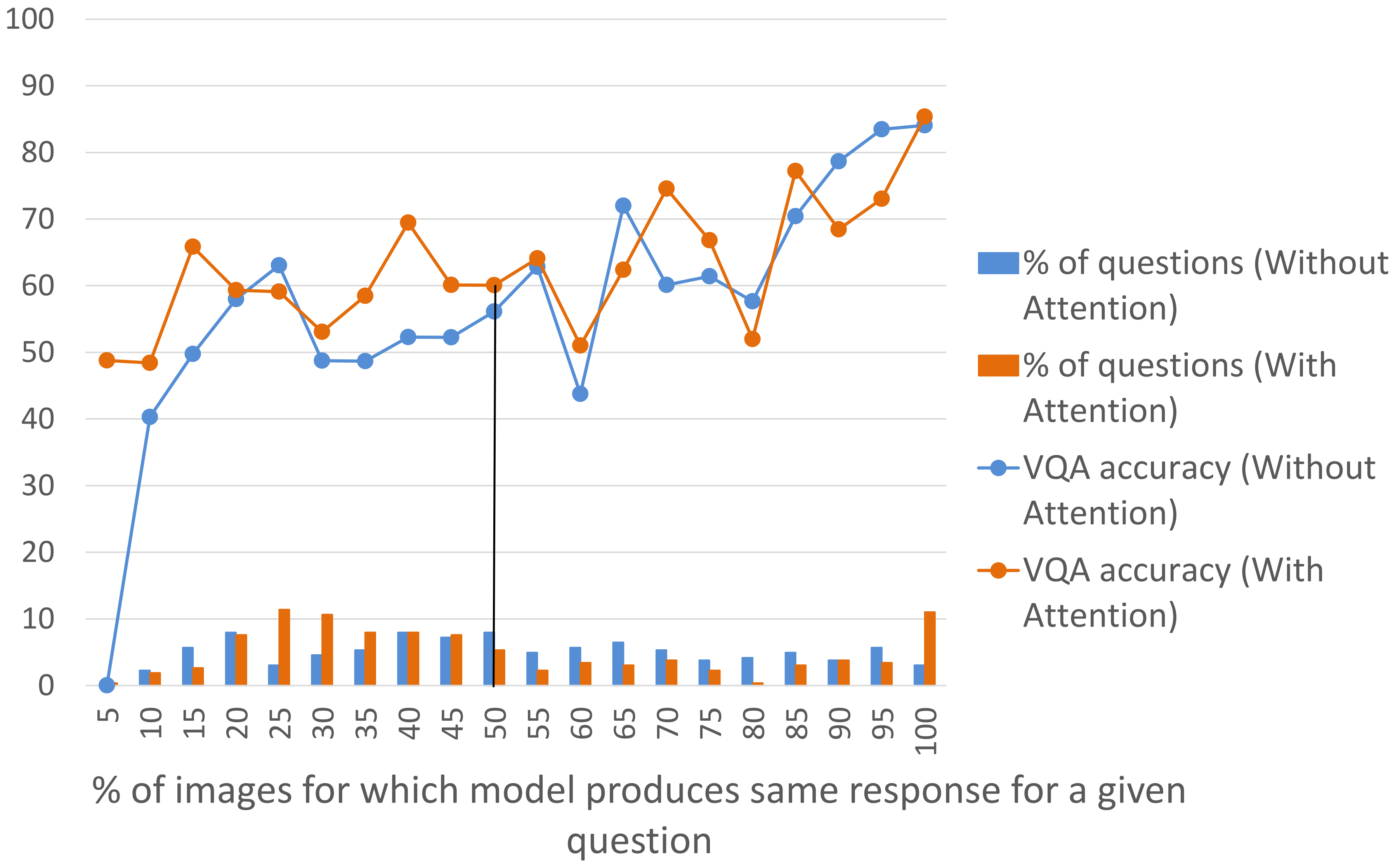
# Looking at the Image



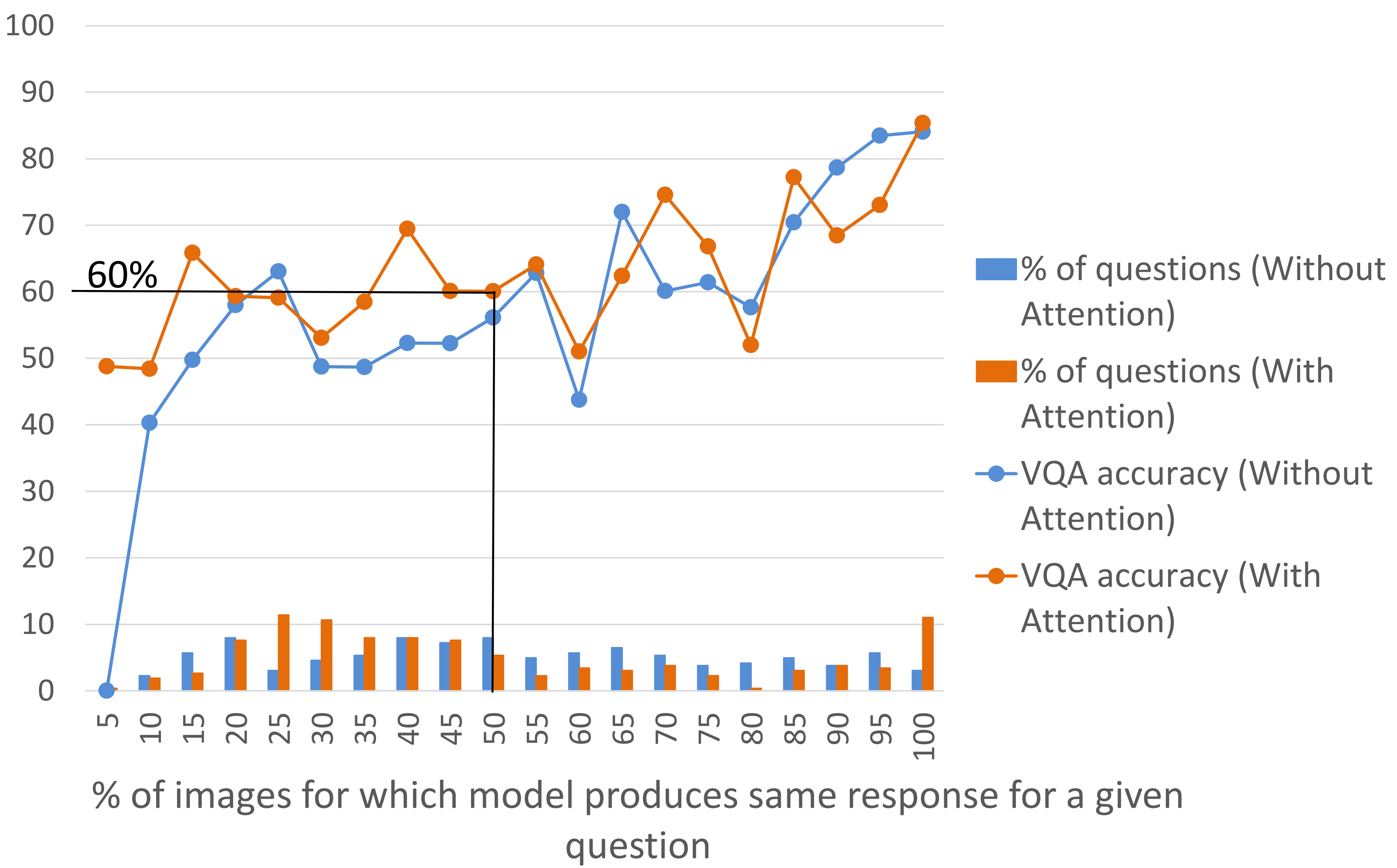
# Looking at the Image



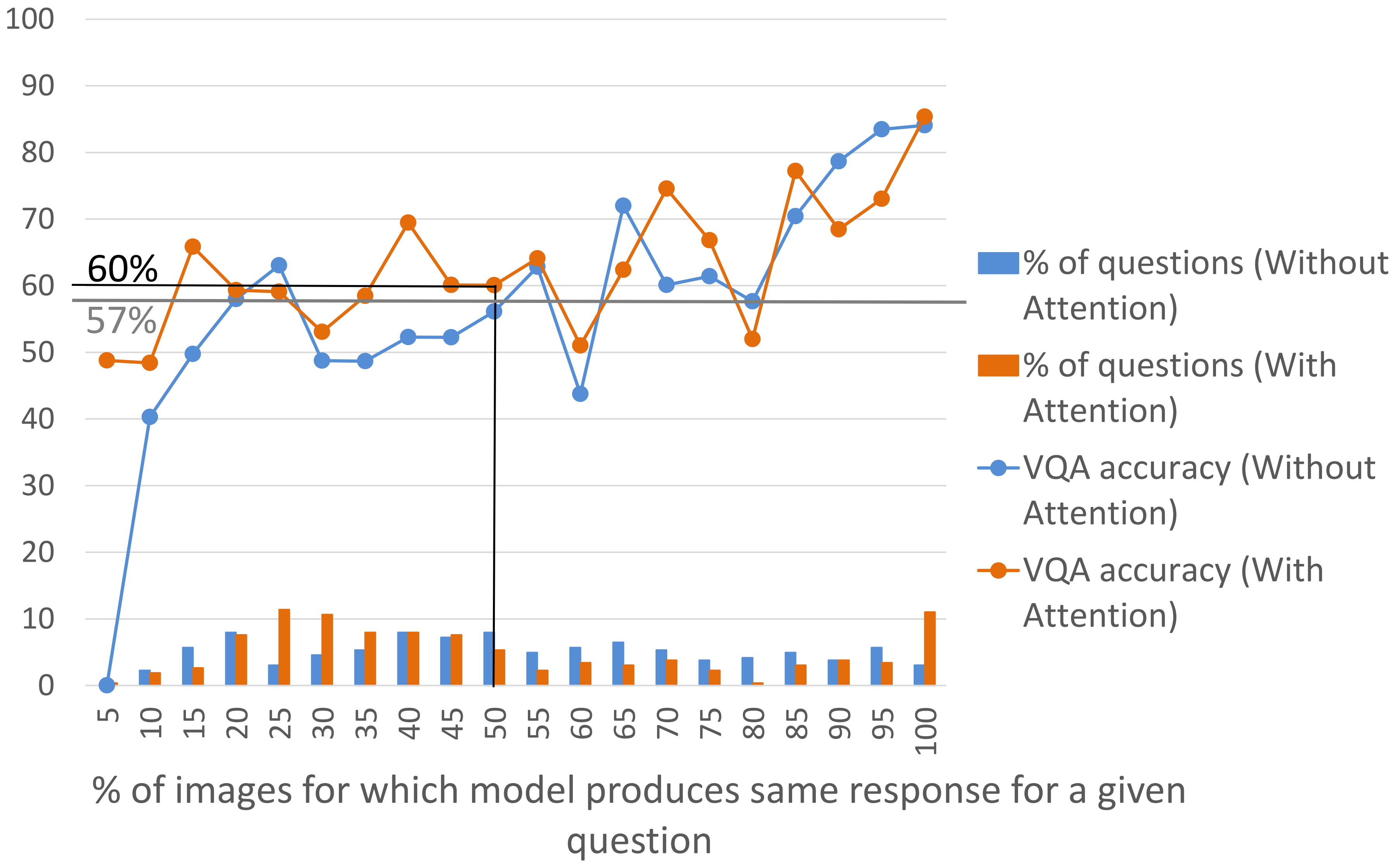
# Looking at the Image



# Looking at the Image



# Looking at the Image



# Looking at the Image

Q: What covers the ground?

Predicted Ans: snow

**All Correct Responses**

# Looking at the Image

**Observations**

# Looking at the Image

**Observations**

1. Producing same responses across images seems to be statistically favorable

# Looking at the Image

**Observations**

1. Producing same responses across images seems to be statistically favorable
2. Label biases in the dataset

# Conclusion

# Conclusion

- Novel techniques for characterizing the behavior of deep VQA models

# Conclusion

- Novel techniques for characterizing the behavior of deep VQA models
- Today's VQA models –

# Conclusion

- Novel techniques for characterizing the behavior of deep VQA models
- Today's VQA models –
  - are "myopic"

# Conclusion

- Novel techniques for characterizing the behavior of deep VQA models
- Today's VQA models –
  - are "myopic"
  - often "jump to conclusions"

# Conclusion

- Novel techniques for characterizing the behavior of deep VQA models
- Today’s VQA models –
  - are “myopic”
  - often “jump to conclusions”
  - are “stubborn”

# To be noted

# To be noted

- Correct behavior depending on dataset?

# To be noted

- Correct behavior depending on dataset?
- Good to know the current behavior

# To be noted

- Correct behavior depending on dataset?
- Good to know the current behavior
- Is the behavior desired?

# To be noted

- Correct behavior depending on dataset?
- Good to know the current behavior
- Is the behavior desired?
- Anthropomorphic adjectives purely pedagogical

Thanks!

Questions?