Visual Question Answering Challenge

Yash Goyal
(Georgia Tech)

Aishwarya Agrawal
(Georgia Tech)

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

VQA Task

VQA Task

What is the mustache made of?

VQA Task

AI System

What is the mustache made of?

VQA Task

What is the mustache made of?

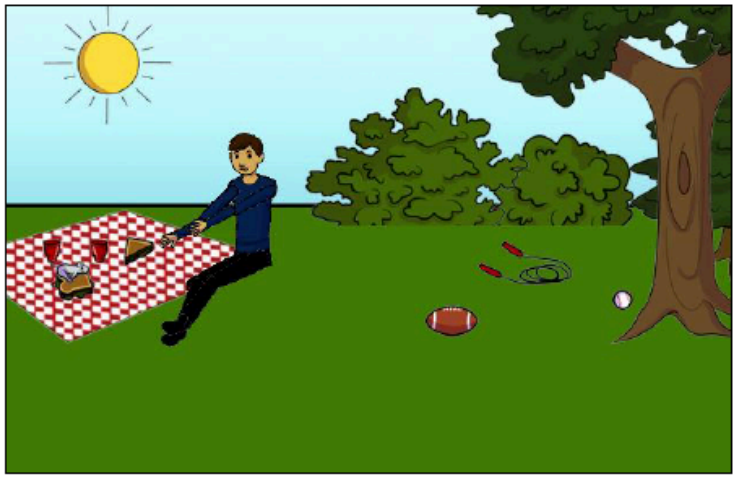
AI System

bananas

VQA v1.0 Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

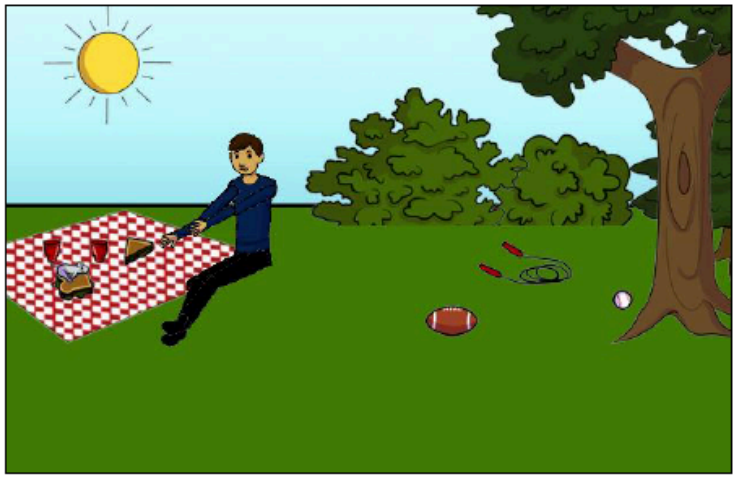
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset

About objects

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

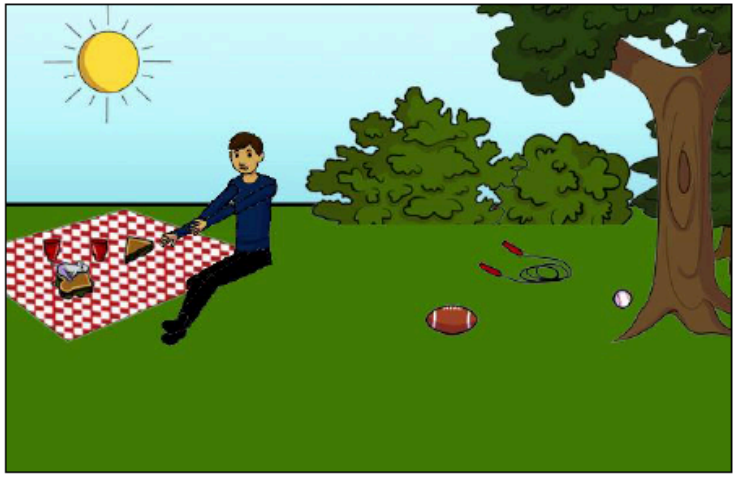
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Fine-grained recognition



Is this person expecting company?
What is just under the tree?

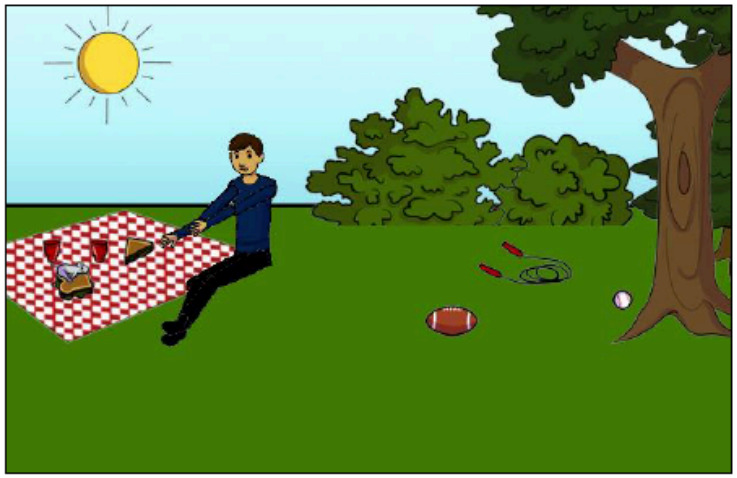
Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Counting



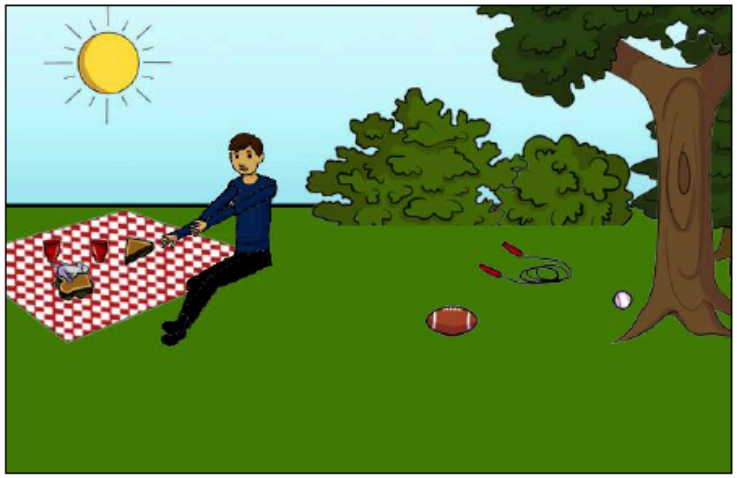
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VQA v1.0 Dataset

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

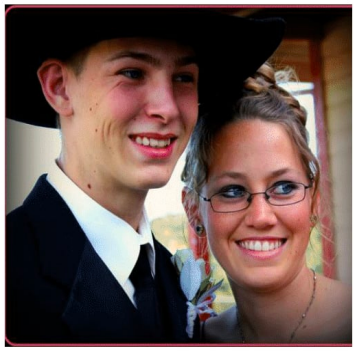
Common sense

VQA v2.0 Dataset

Who is wearing glasses?

man

woman



Where is the child sitting?

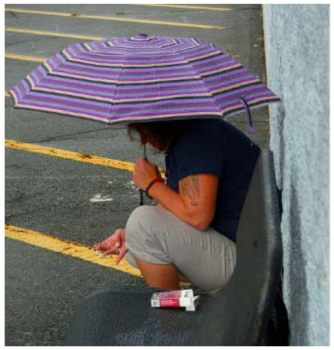
fridge

arms

Is the umbrella upside down?

yes

no



How many children are in the bed?

2

1

Who is wearing glasses?

Similar images

man

Different answers

woman

VQA v1.0

New in VQA v2.0

VQA v2.0 Dataset Stats

- >200K images
- >1.1M questions
- >11M answers

1.8 x VQA v1.0

Accuracy Metric

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

1940. COCO_train2014_000000012015

Open-Ended/Multiple-Choice/Ground-Truth

Q: WHAT OBJECT IS THIS

Ground Truth Answers:

(1) television	(6) television
(2) tv	(7) television
(3) tv	(8) tv
(4) tv	(9) tv
(5) television	(10) television

Q: How old is this TV?

Ground Truth Answers:

(1) 20 years	(6) old
(2) 35	(7) 80 s
(3) old	(8) 30 years
(4) more than thirty years old	(9) 15 years
(5) old	(10) very old

Q: Is this TV upside-down?

Ground Truth Answers:

(1) yes	(6) yes
(2) yes	(7) yes
(3) yes	(8) yes
(4) yes	(9) yes
(5) yes	(10) yes

Outline

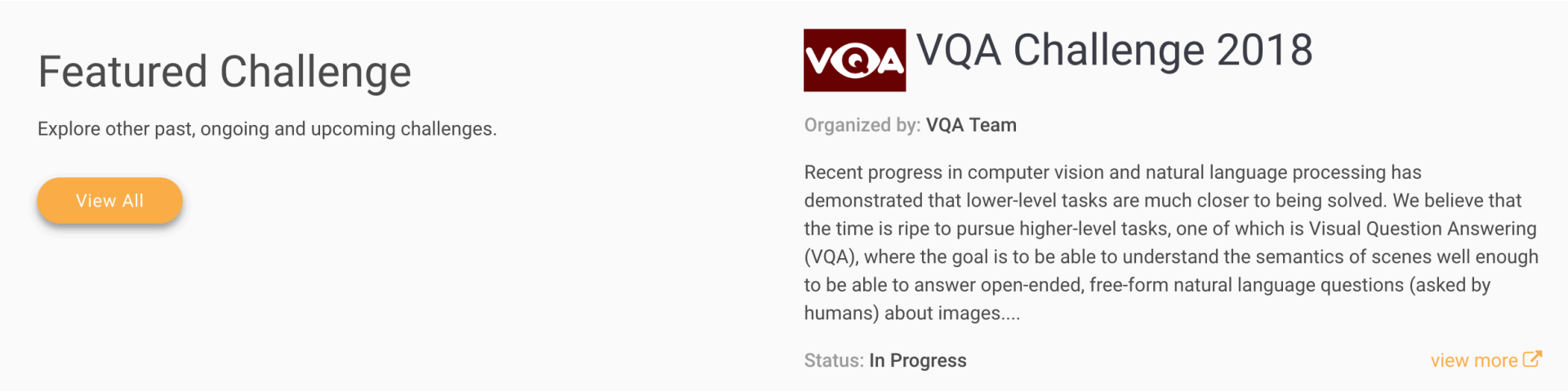
Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

VQA Challenge on https://evalai.cloudcv.org/



Dataset splits

	Images	Questions	Answers
Training	80K	443K	4.4M

Dataset size is approximate

Dataset splits

	Images	Questions	Answers
Training	80K	443K	4.4M
Validation	40K	214K	2.1M

Dataset size is approximate

Dataset splits

	Images	Questions	Answers
Training	80K	443K	4.4M
Validation	40K	214K	2.1M
Test	80K	447K	

Dataset size is approximate

Test Dataset

- 4 splits of approximately equal size
- **Test-dev (development)**
  - Debugging and Validation.
- **Test-standard (publications)**
  - Used to score entries for the Public Leaderboard.
- **Test-challenge (competitions)**
  - Used to rank challenge participants.
- **Test-reserve (check overfitting)**
  - Used to estimate overfitting. Scores on this set are never released.

Slide adapted from: MSCOCO Detection/Segmentation Challenge, ICCV 2015

Outline

Overview of Task and Dataset

Overview of Challenge

Winner Announcements

Analysis of Results

Challenge Stats

- 40 teams
- >=40 institutions*
- >=8 countries*

*Statistics based on teams that have replied

Challenge Runner-Ups

Joint Runner-Up Team 1

SNU-BI

Jin-Hwa Kim (Seoul National University)

Jaehyun Jun (Seoul National University)

Byoung-Tak Zhang (Seoul National University & Surromind Robotics)

Challenge Accuracy: **71.69**

Challenge Runner-Ups

Joint Runner-Up Team 2

HDU-UCAS-USYD

Zhou Yu (*Hangzhou Dianzi University, China*)
Jun Yu (*Hangzhou Dianzi University, China*)
Chenchao Xiang (*Hangzhou Dianzi University, China*)
Liang Wang (*Hangzhou Dianzi University, China*)
Dalu Guo (*The Unversity of Sydney, Australia*)
Qingming Huang (*University of Chinese Academy of Sciences*)
Jianping Fan (*Hangzhou Dianzi University, China*)
Dacheng Tao (*The University of Sydney, Australia*)

Challenge Accuracy**: 71.91**

Challenge Winner

FAIR-A*

Yu Jiang† (Facebook AI Research)

Vivek Natarajan† (Facebook AI Research)

Xinlei Chen† (Facebook AI Research)

Marcus Rohrbach (Facebook AI Research)

Dhruv Batra (Facebook AI Research & Georgia Tech)

Devi Parikh (Facebook AI Research & Georgia Tech)

Challenge Accuracy: **72.41**

† equal contribution

Outline

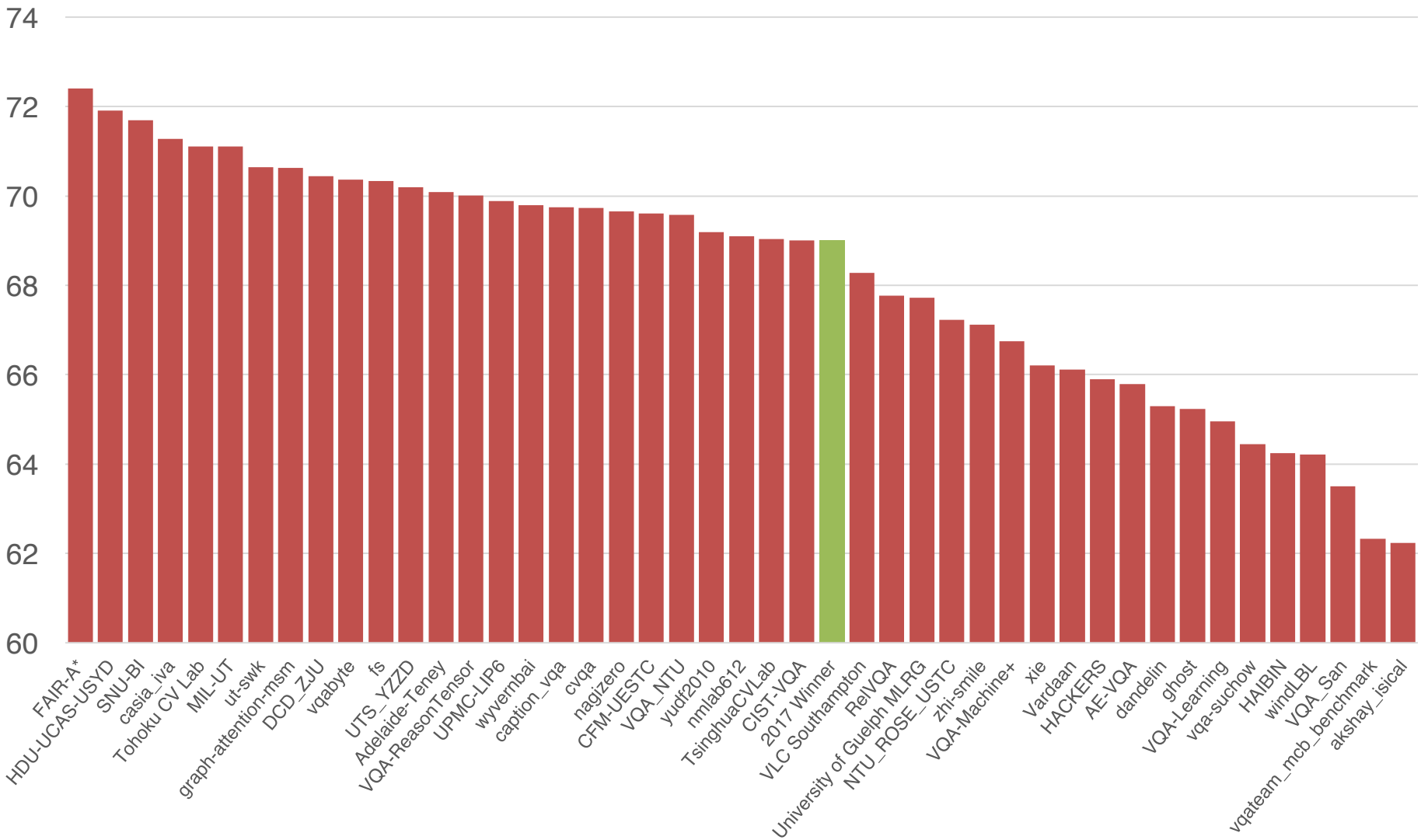
Overview of Task and Dataset

Overview of Challenge

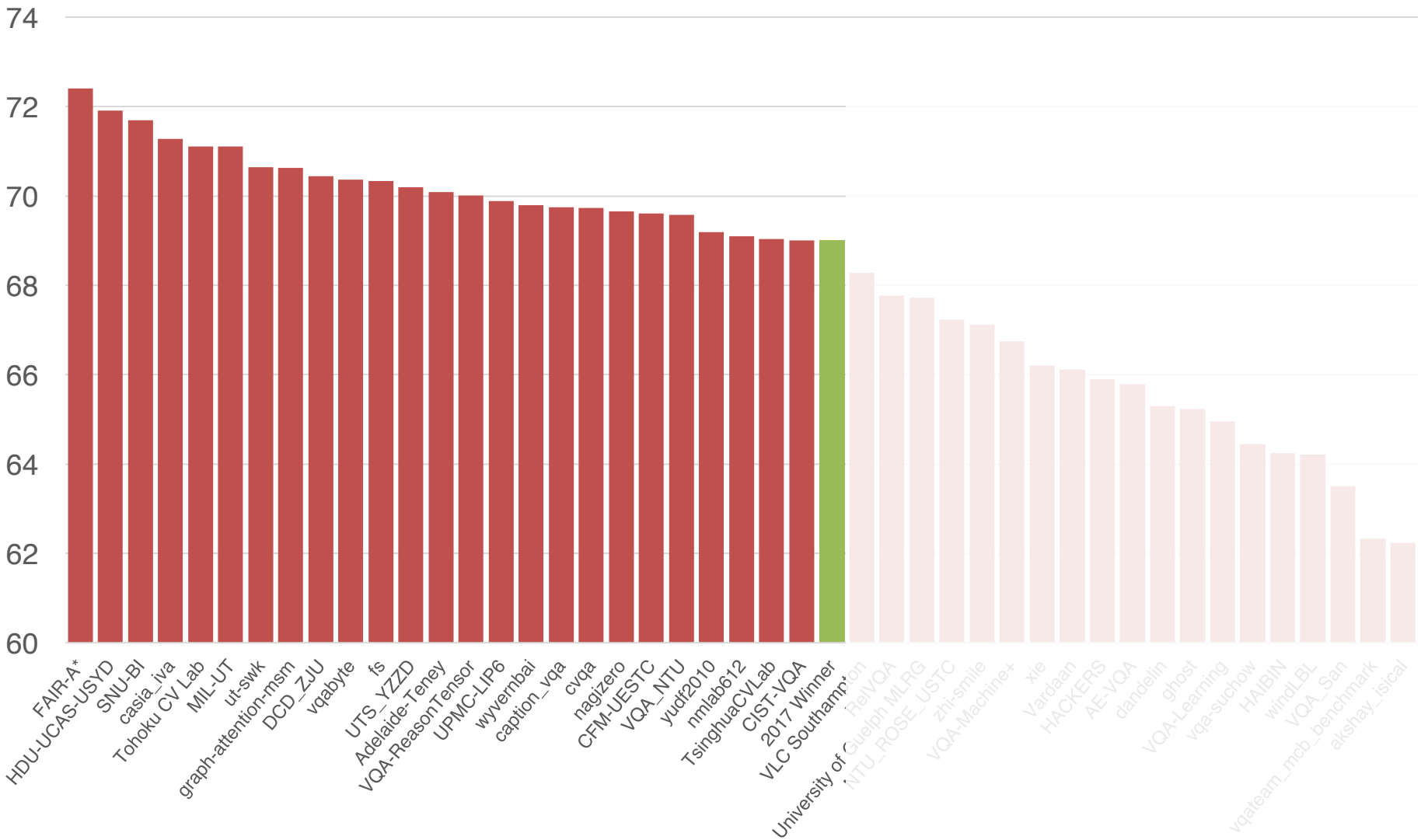
Winner Announcements

Analysis of Results

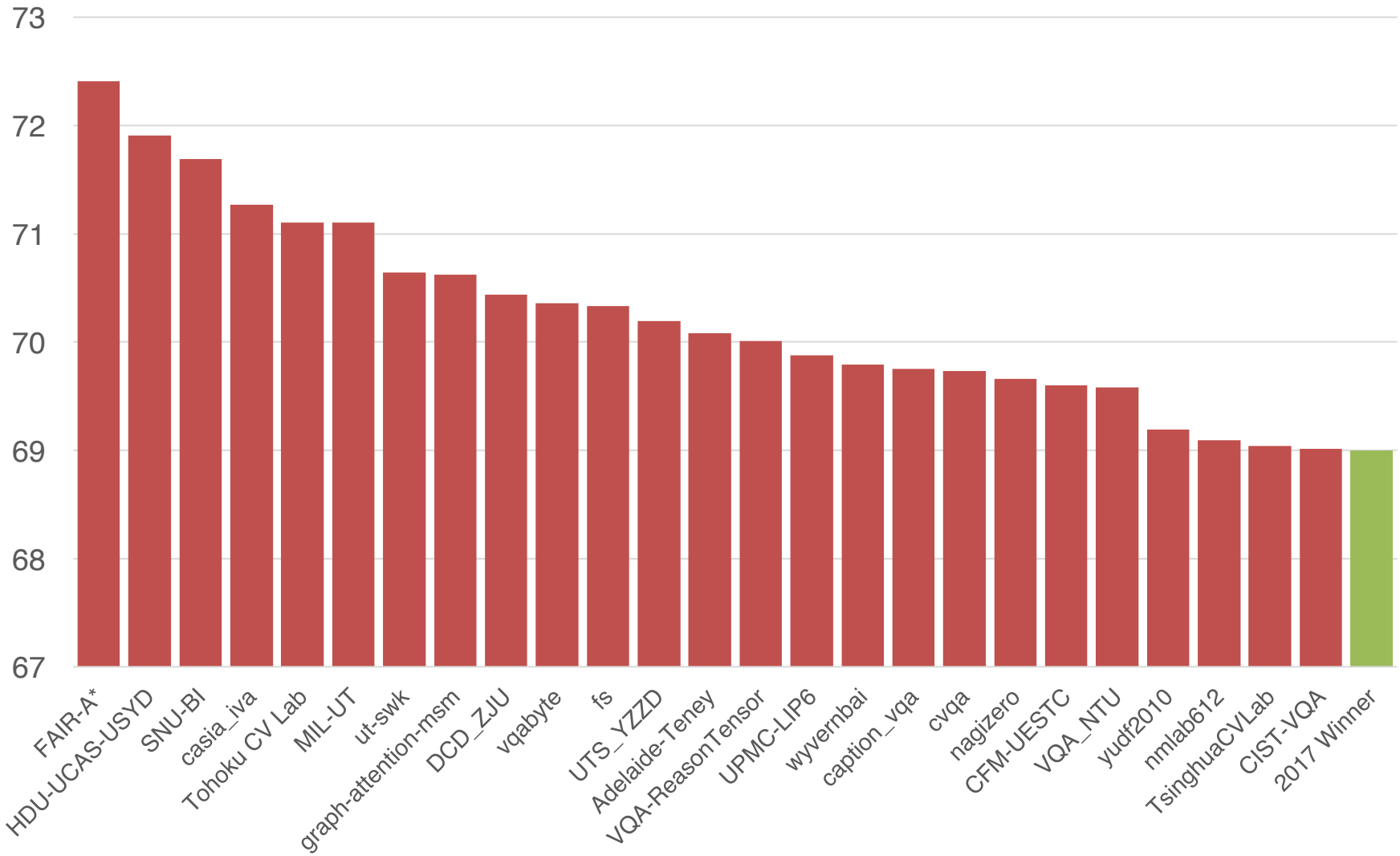
Challenge Results



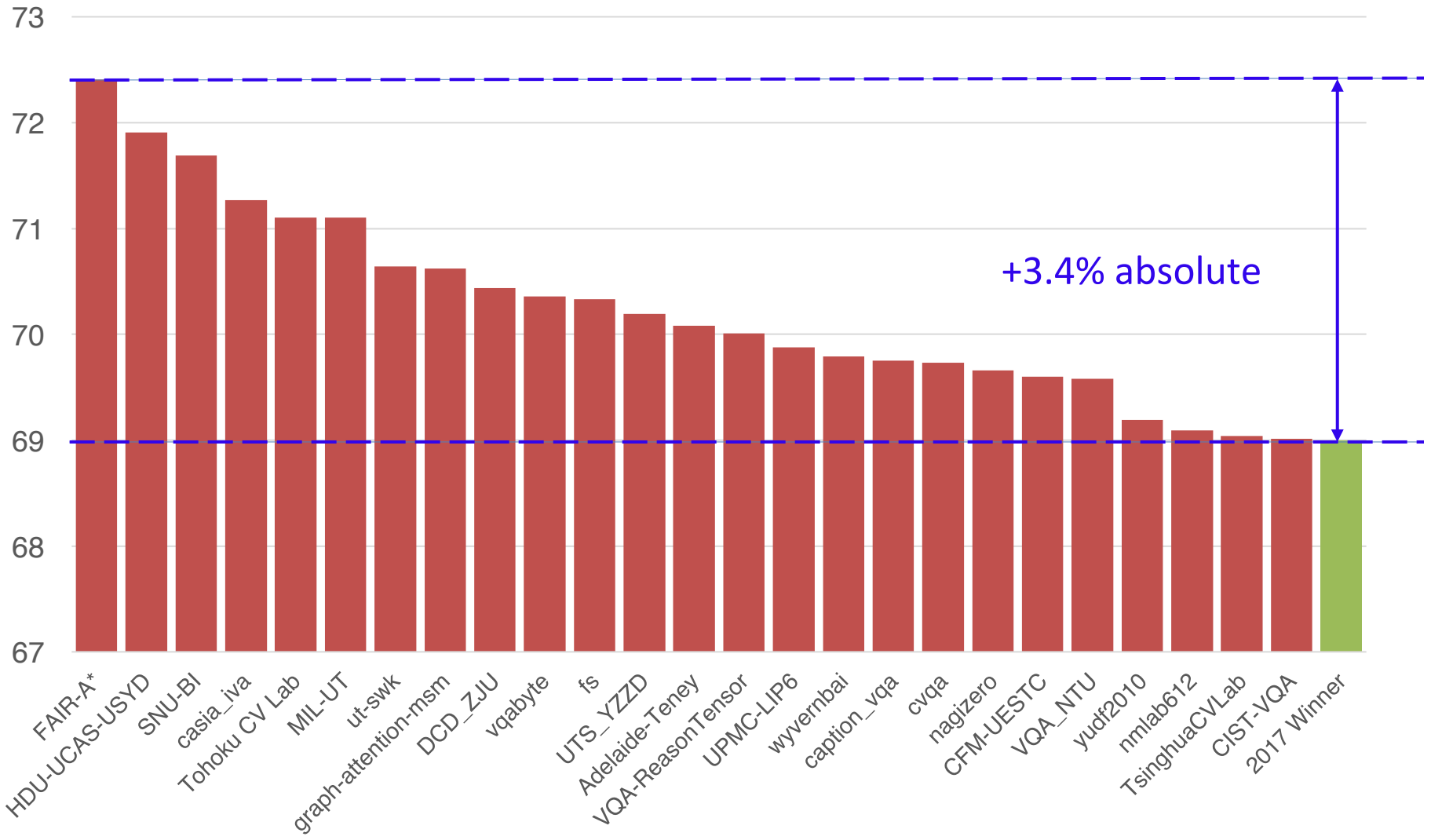
Challenge Results



Challenge Results



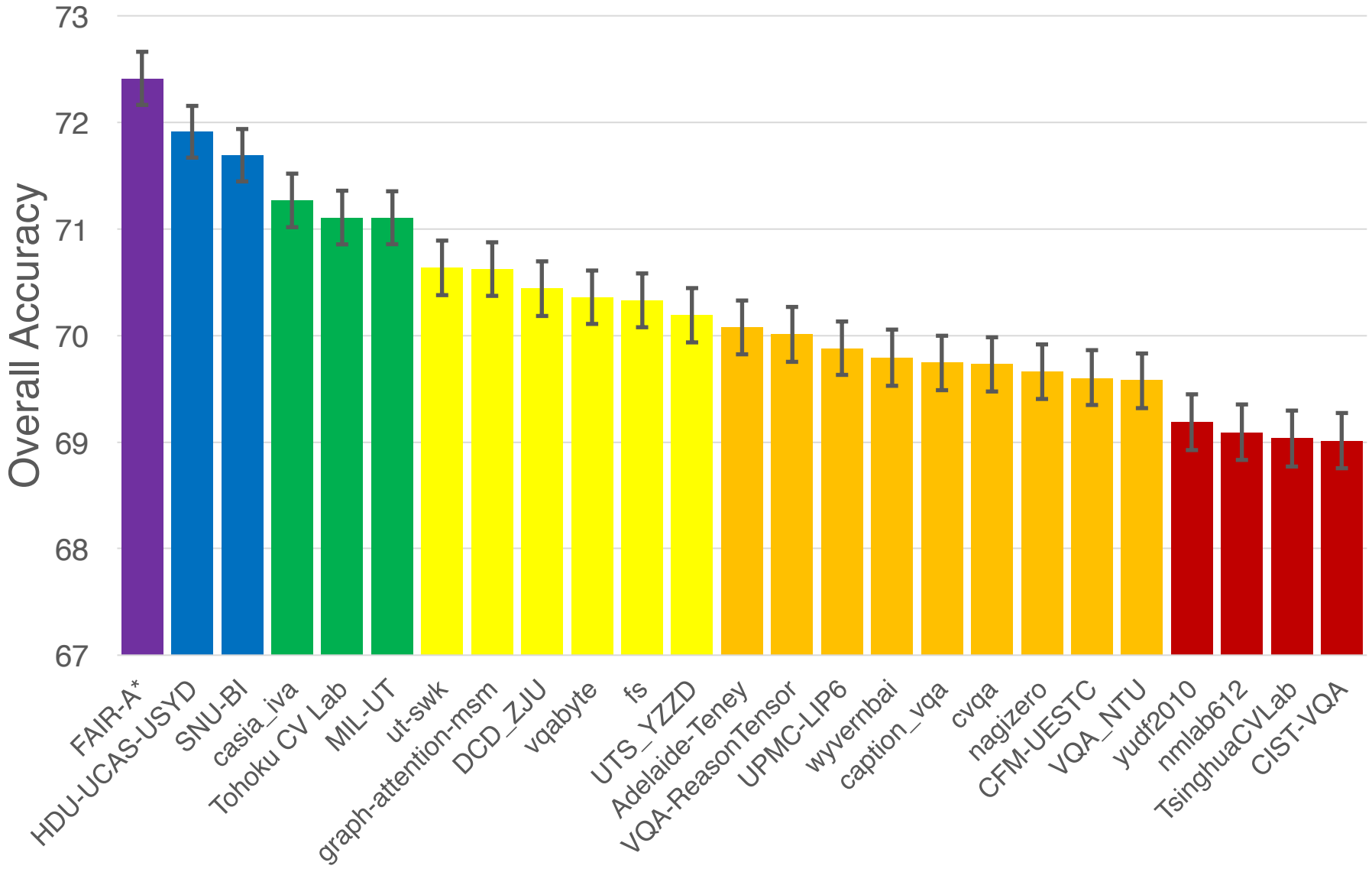
Challenge Results



Statistical Significance

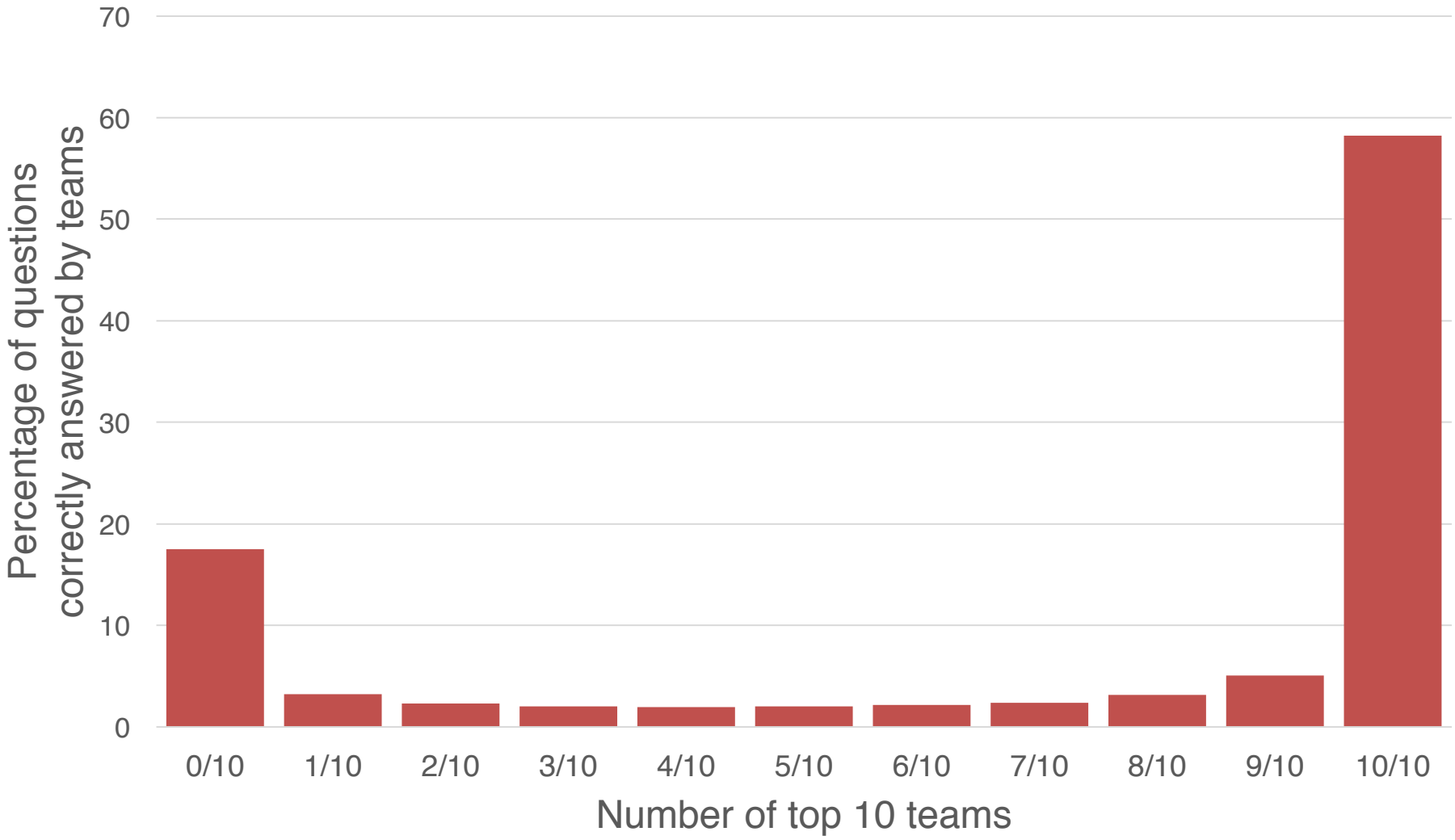
- Bootstrap samples 5000 times
- @ 95% confidence

Statistical Significance

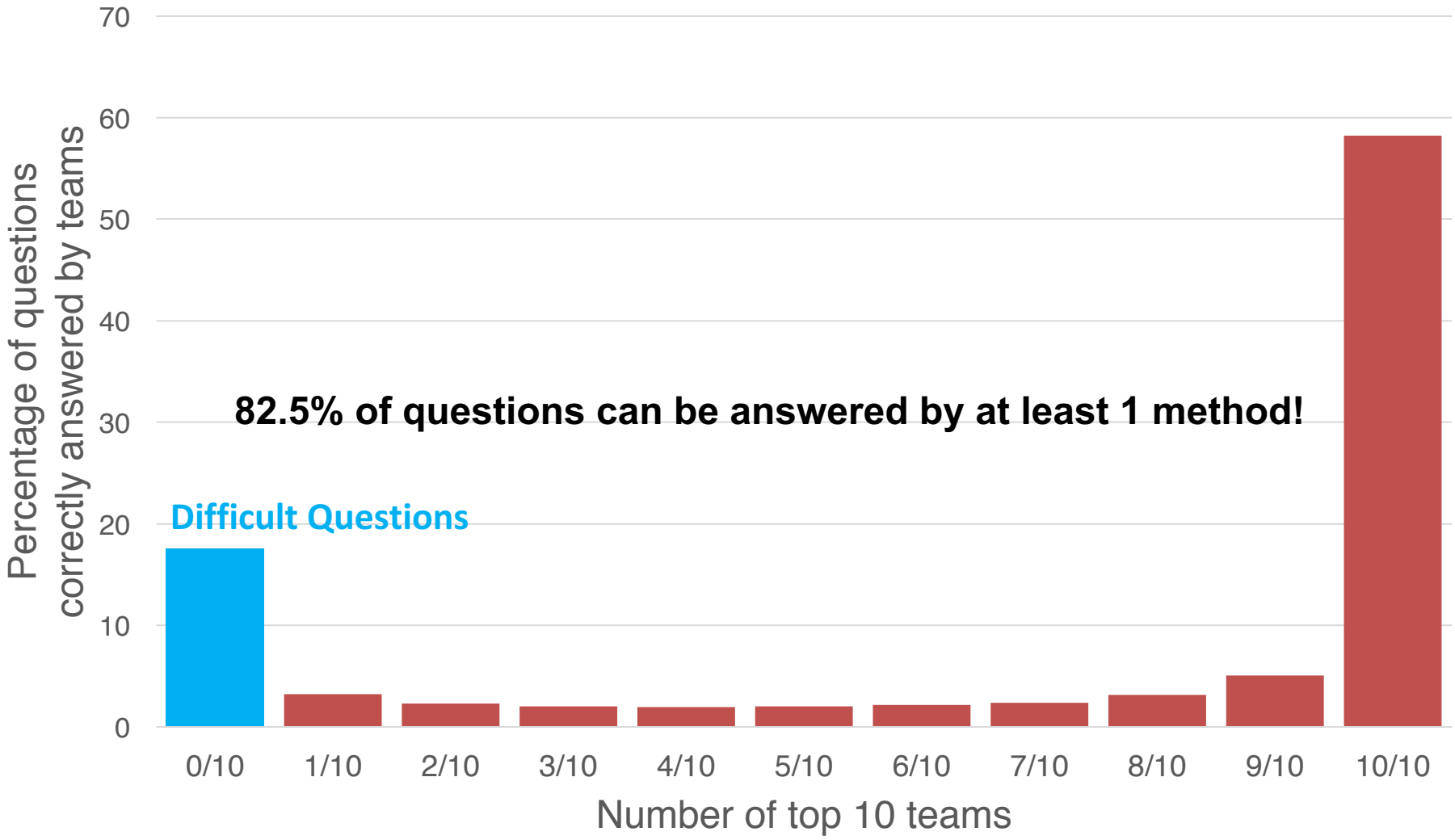


Easy vs. Difficult Questions

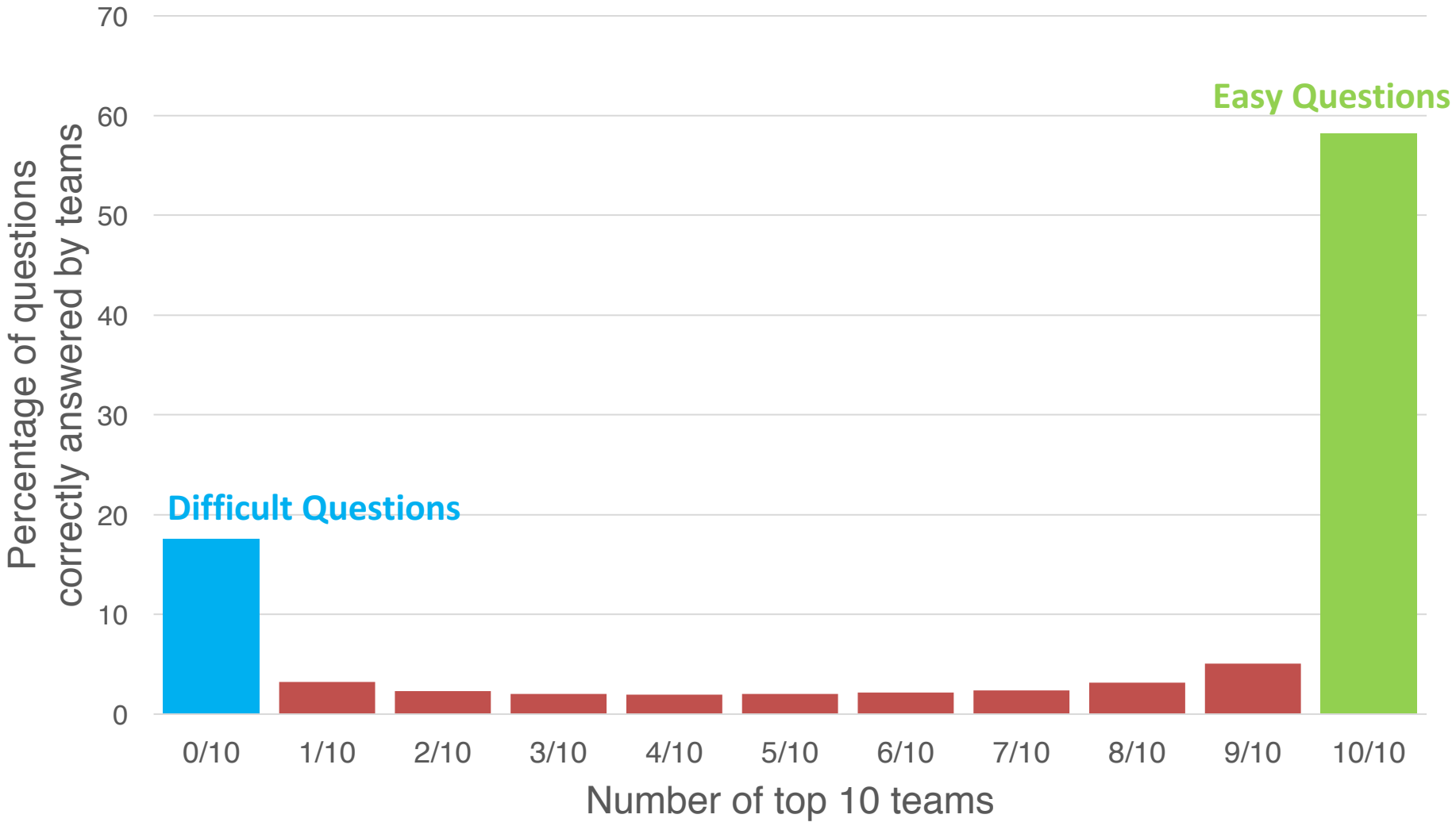
Easy vs. Difficult Questions



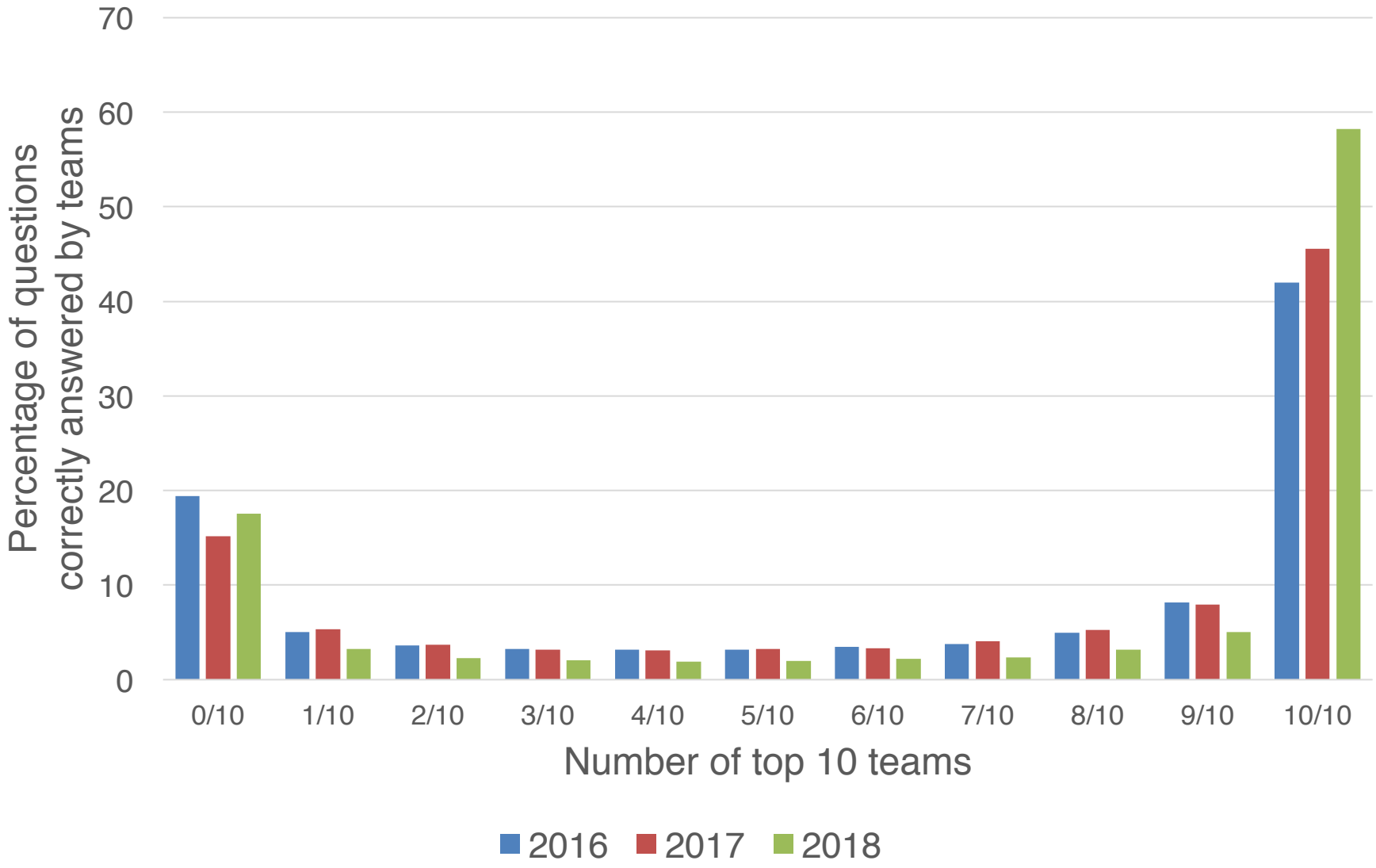
Easy vs. Difficult Questions



Easy vs. Difficult Questions

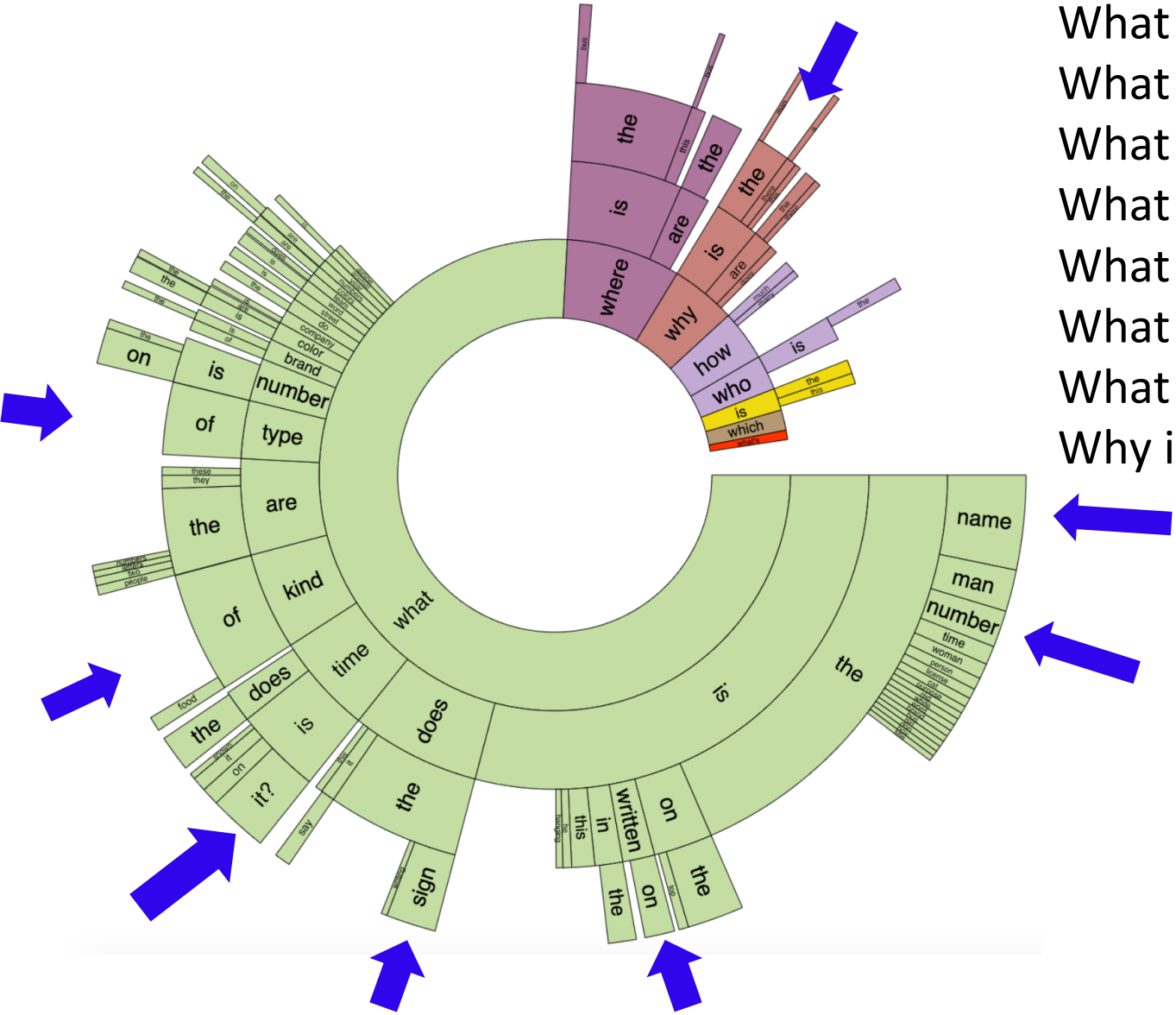


Easy vs. Difficult Questions



Difficult Questions with Rare Answers

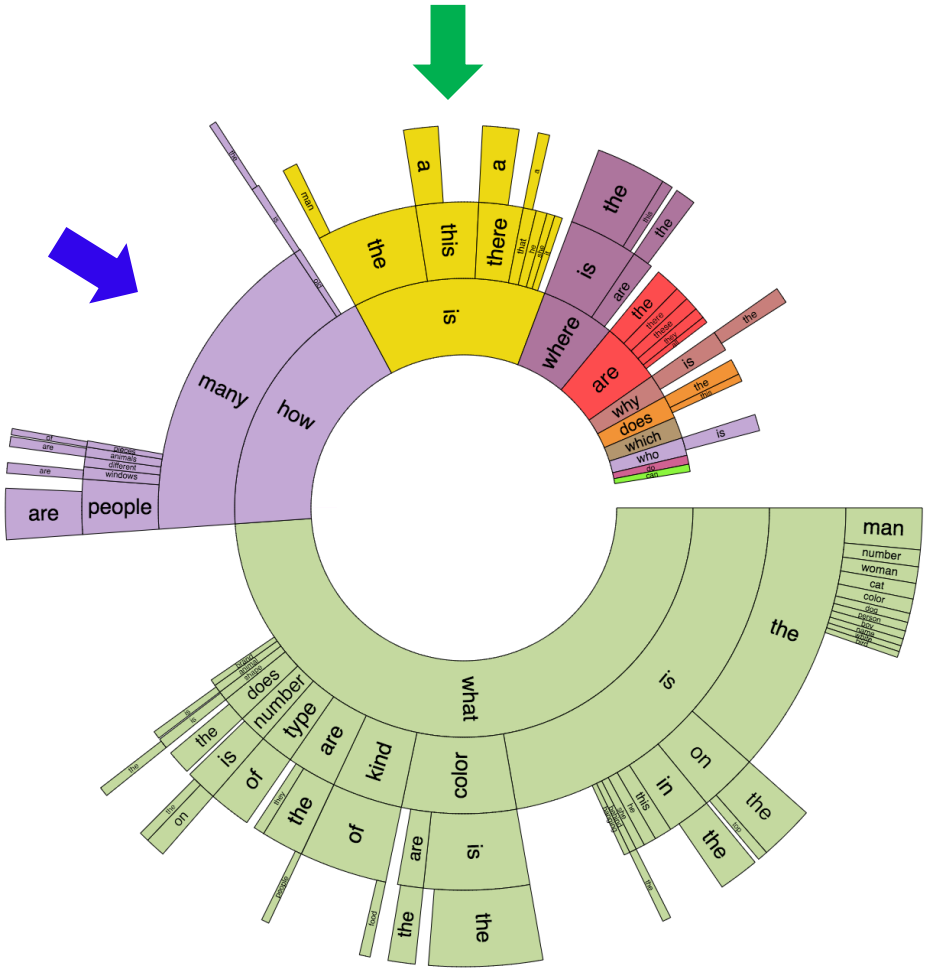
Difficult Questions with Rare Answers



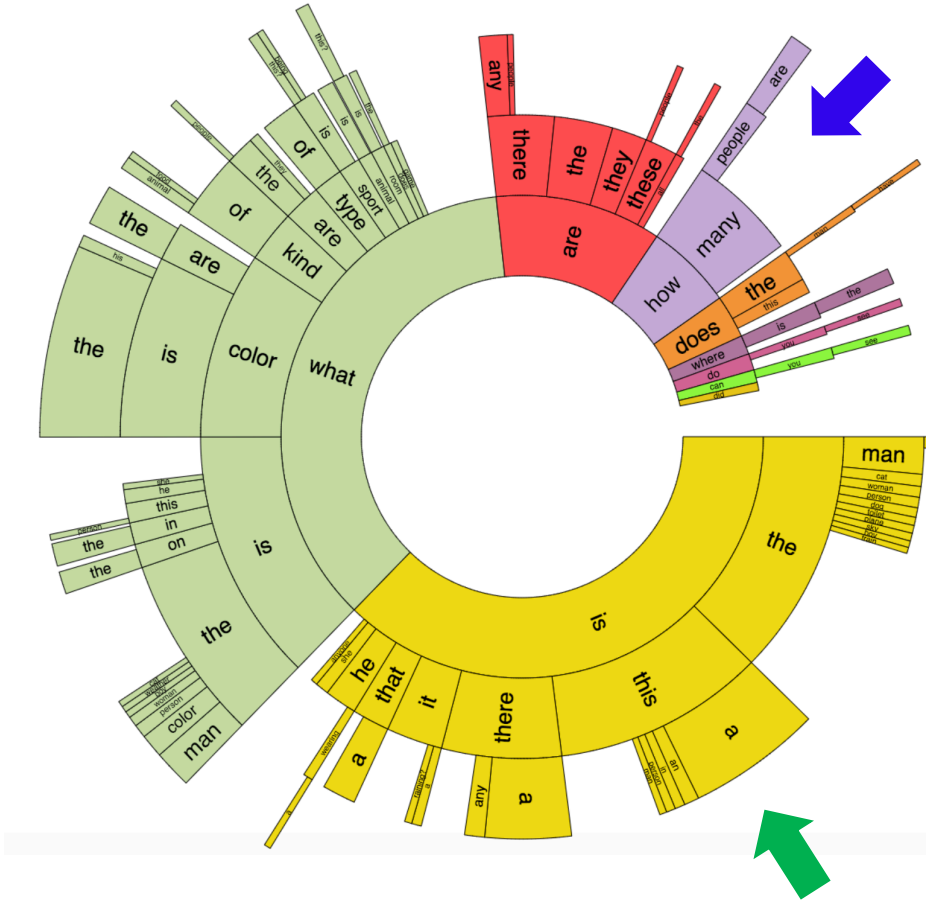
What is the name of …
What is the number on …
What is written on the …
What does the sign …
What time is it?
What kind of …
What type of …
Why is the …

Easy vs. Difficult Questions

Easy vs. Difficult Questions



Difficult Questions with Frequent Answers

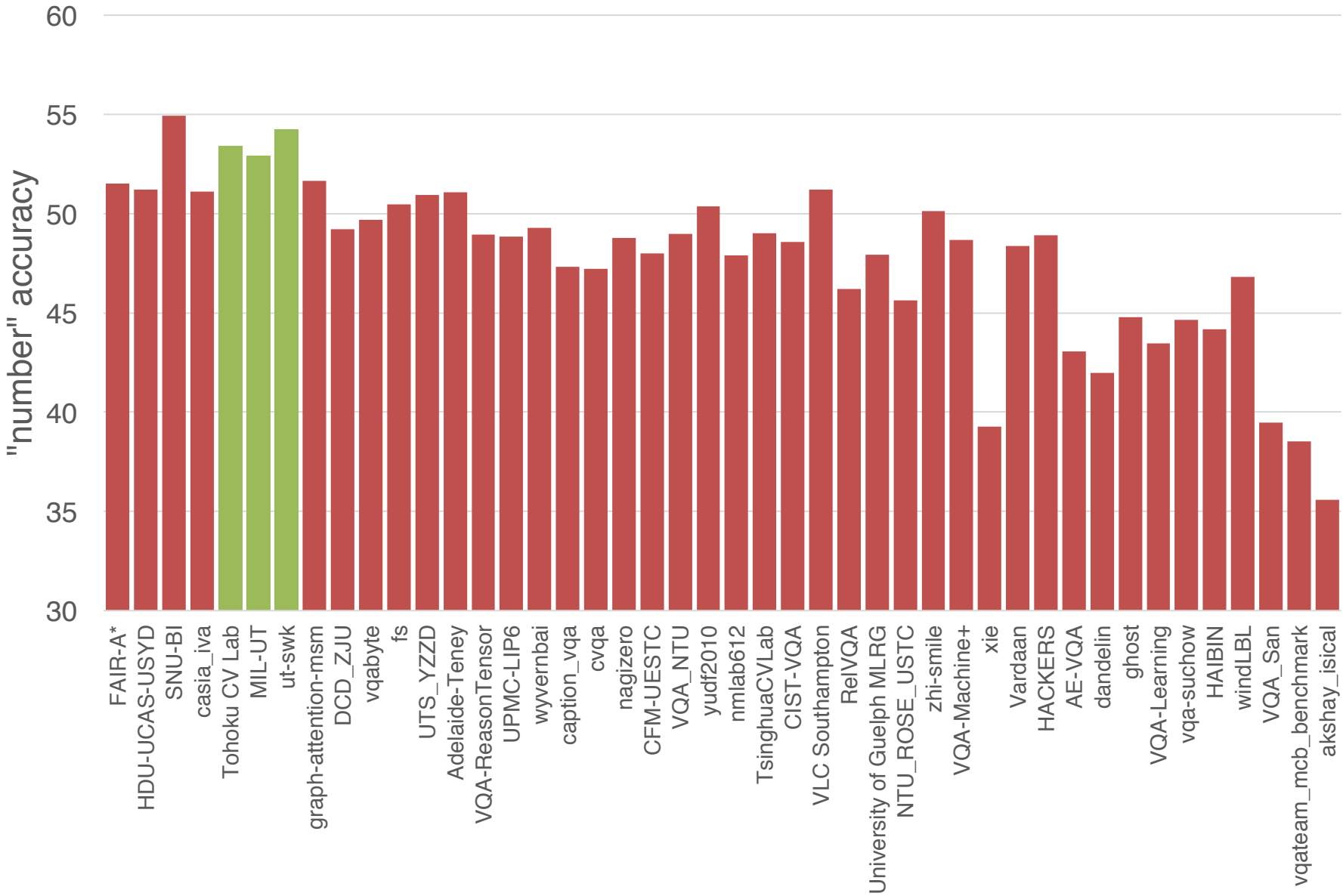


Easy Questions

Answer Type Analyses

- SNU_BI performs the best for “number” questions

Results on “number” questions



Answer Type Analyses

- SNU_BI performs the best for “number” questions
- No team statistically significantly better than the winner team for “yes/no” and “other”

Are models sensitive to subtle changes in images?

Who is wearing glasses?

Similar images

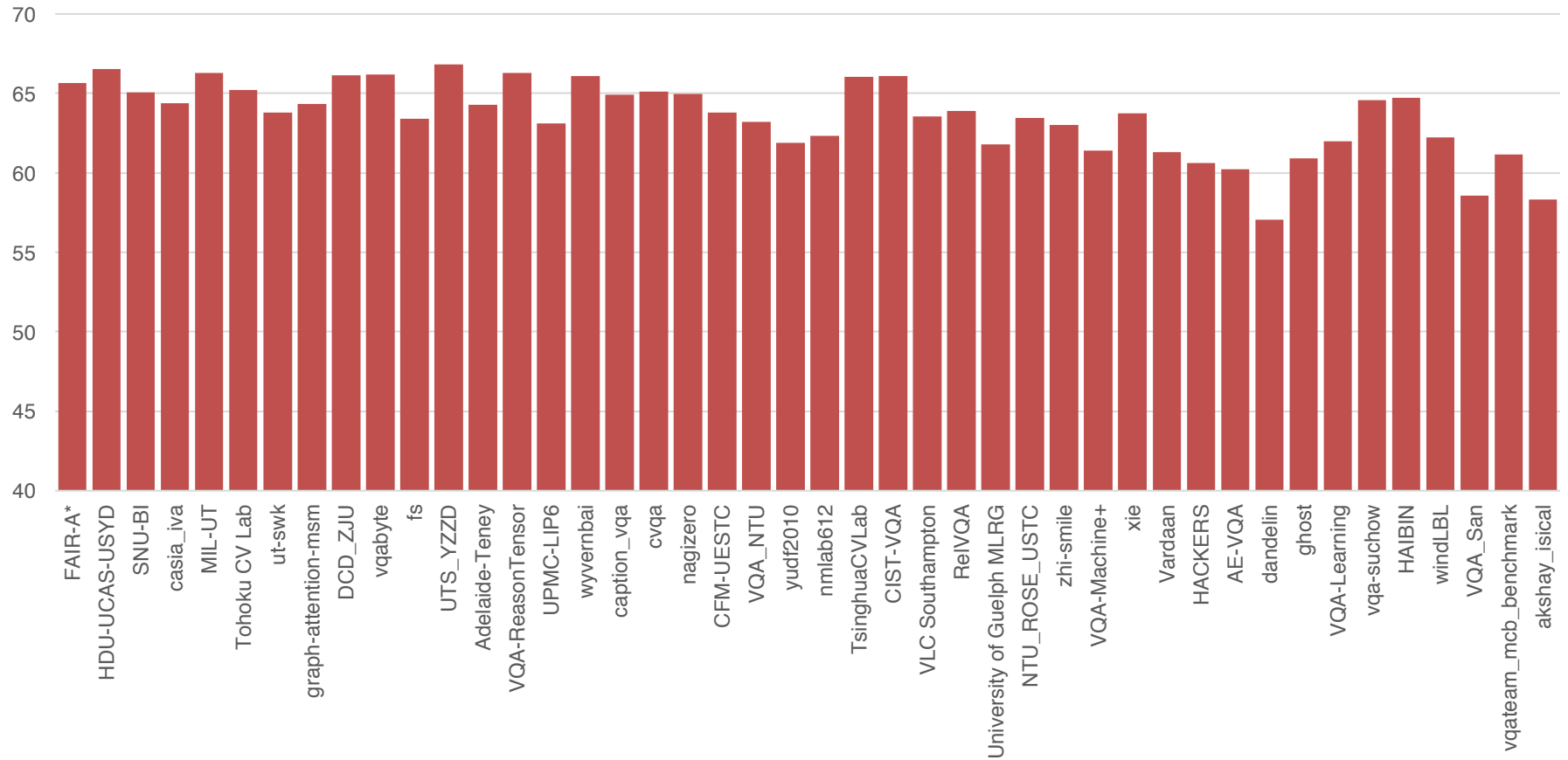
man ⟷ **woman**

Different answers

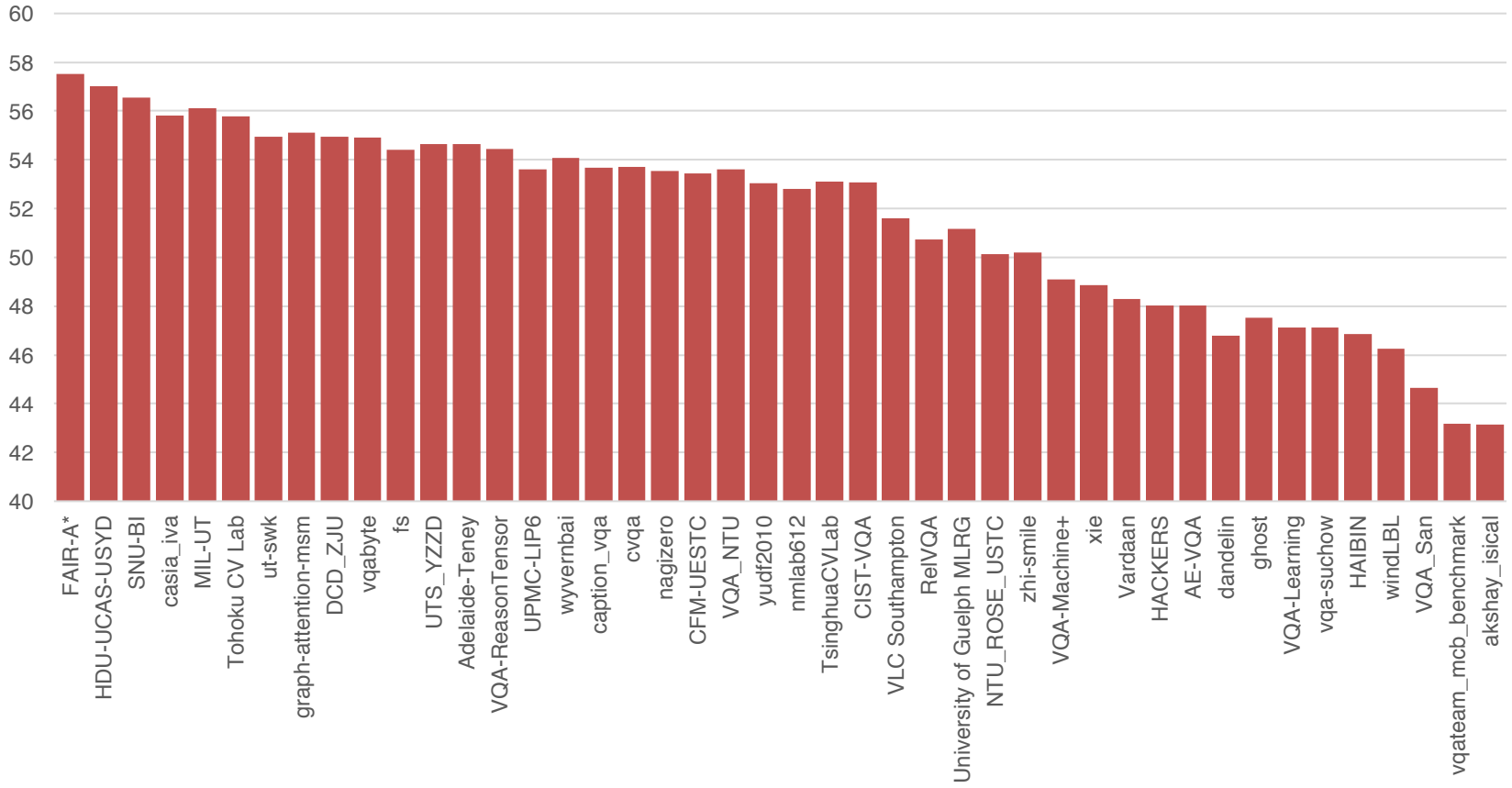
Are models sensitive to subtle changes in images?

- Are predictions different for complementary images?
- Are predictions accurate for complementary images?

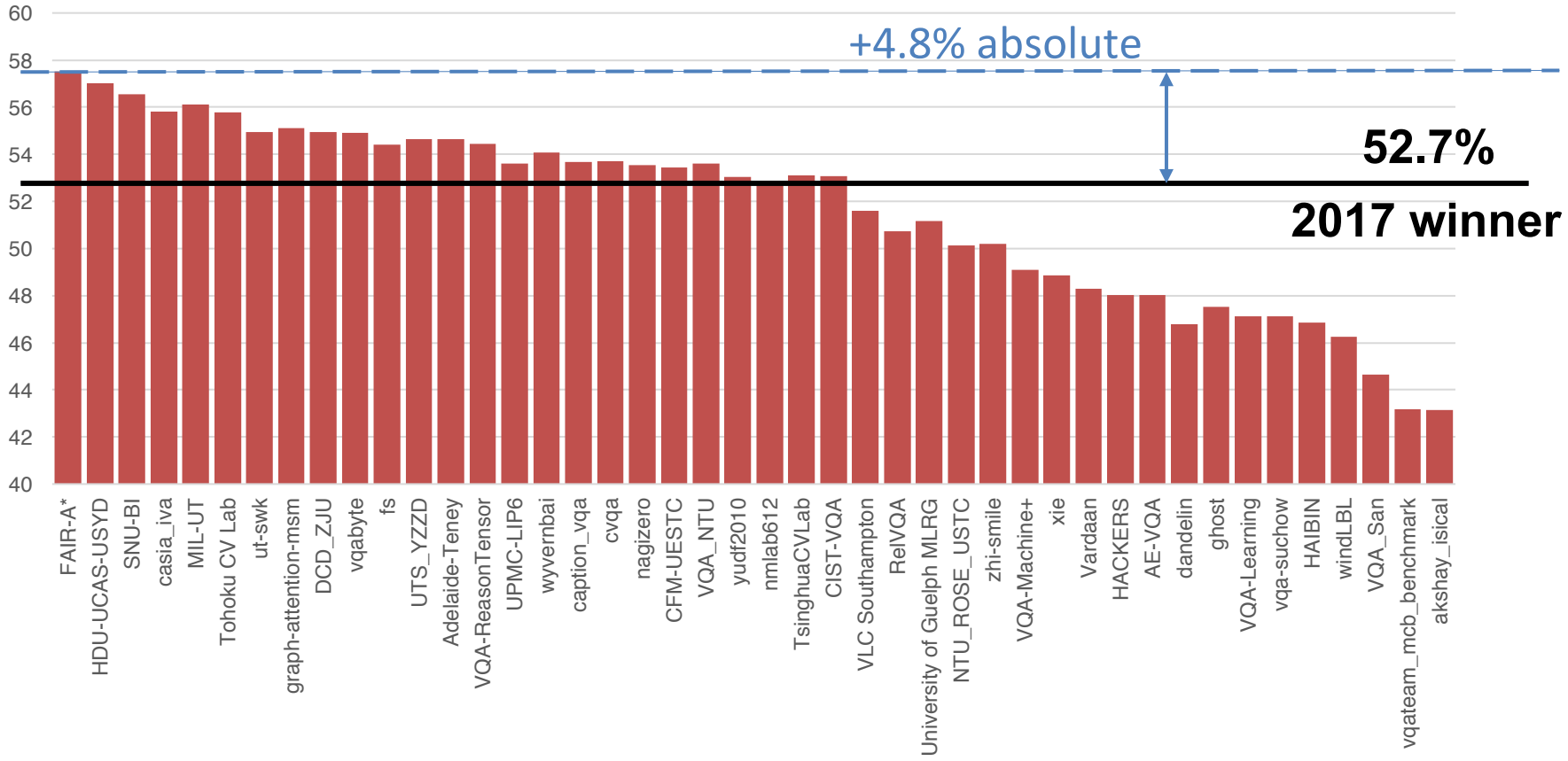
Are predictions **different** for complementary images?



Are predictions **accurate** for complementary images?



Are predictions accurate for complementary images?



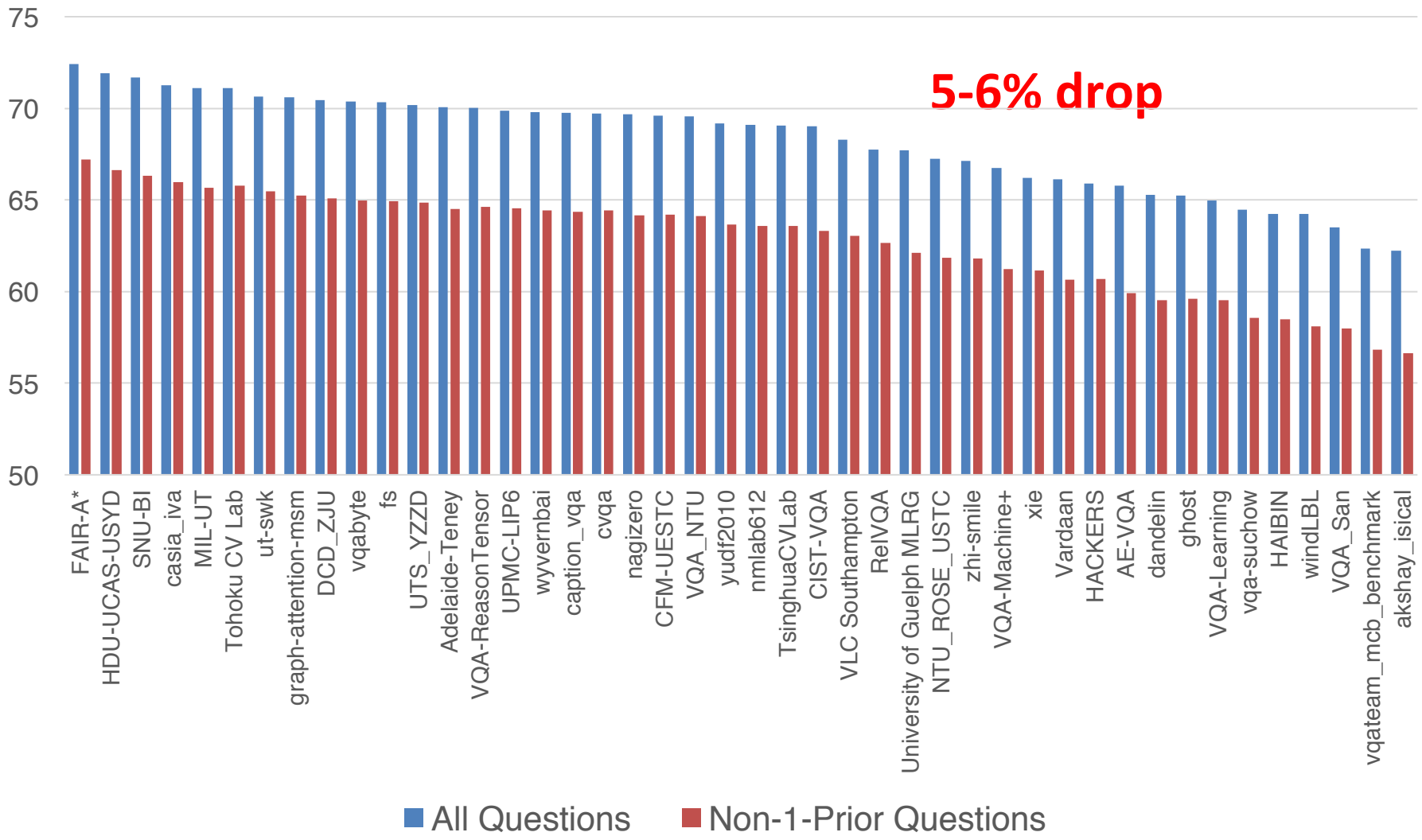
Are models driven by priors?

Only consider those questions whose answers are not popular (given the question type) in training

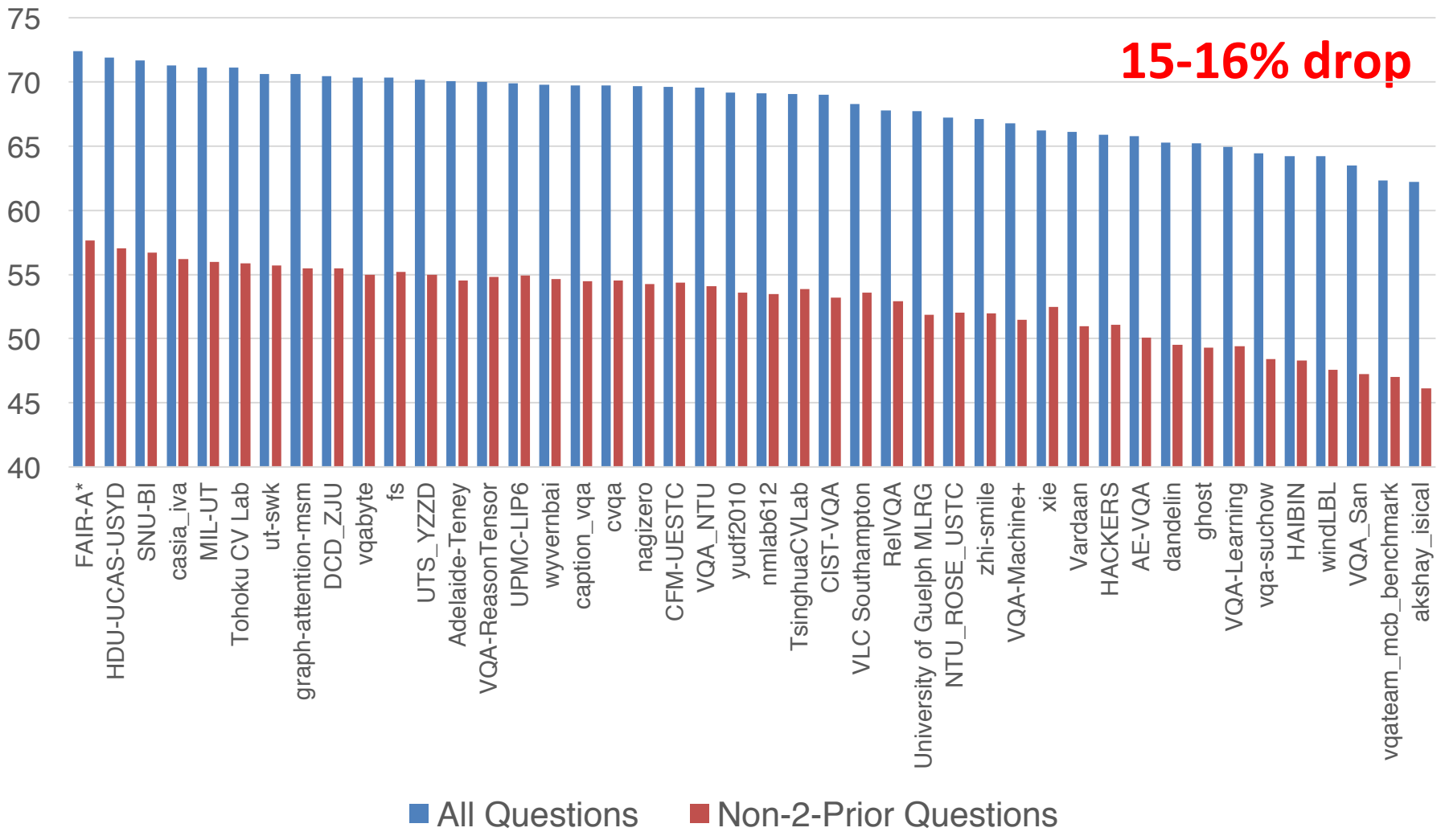
- 1-Prior: Test answers are not the top-1 most common in training
- 2-Prior: Test answer are not the top-2 most common in training

Agrawal et al., CVPR 2018

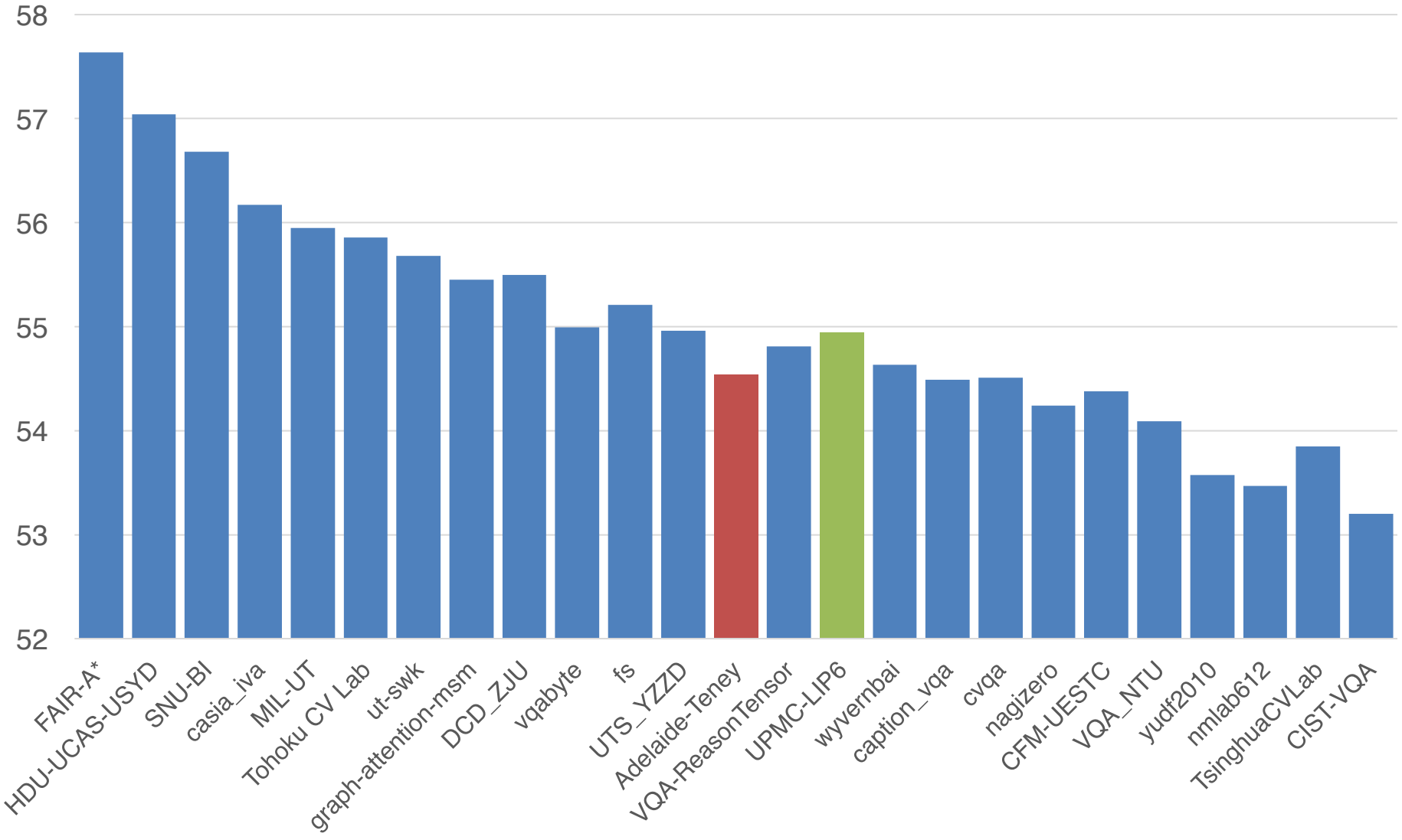
Are models driven by priors?



Are models driven by priors?



Are models driven by priors?



Improvement from 2017 challenge

- 1-Prior: Best performance improved by 3.8%
- 2-Prior: Best performance improved by 3.3%

Are models compositional?

Only consider those questions which are compositionally novel:

- QA pair is not seen in training
- Constituting concepts seen in training

Agrawal et al., Arxiv 2018

Are models compositional?

Training

Q: What color is the plate?

A: Green

Q: What color are stop lights?

A: Red

Testing

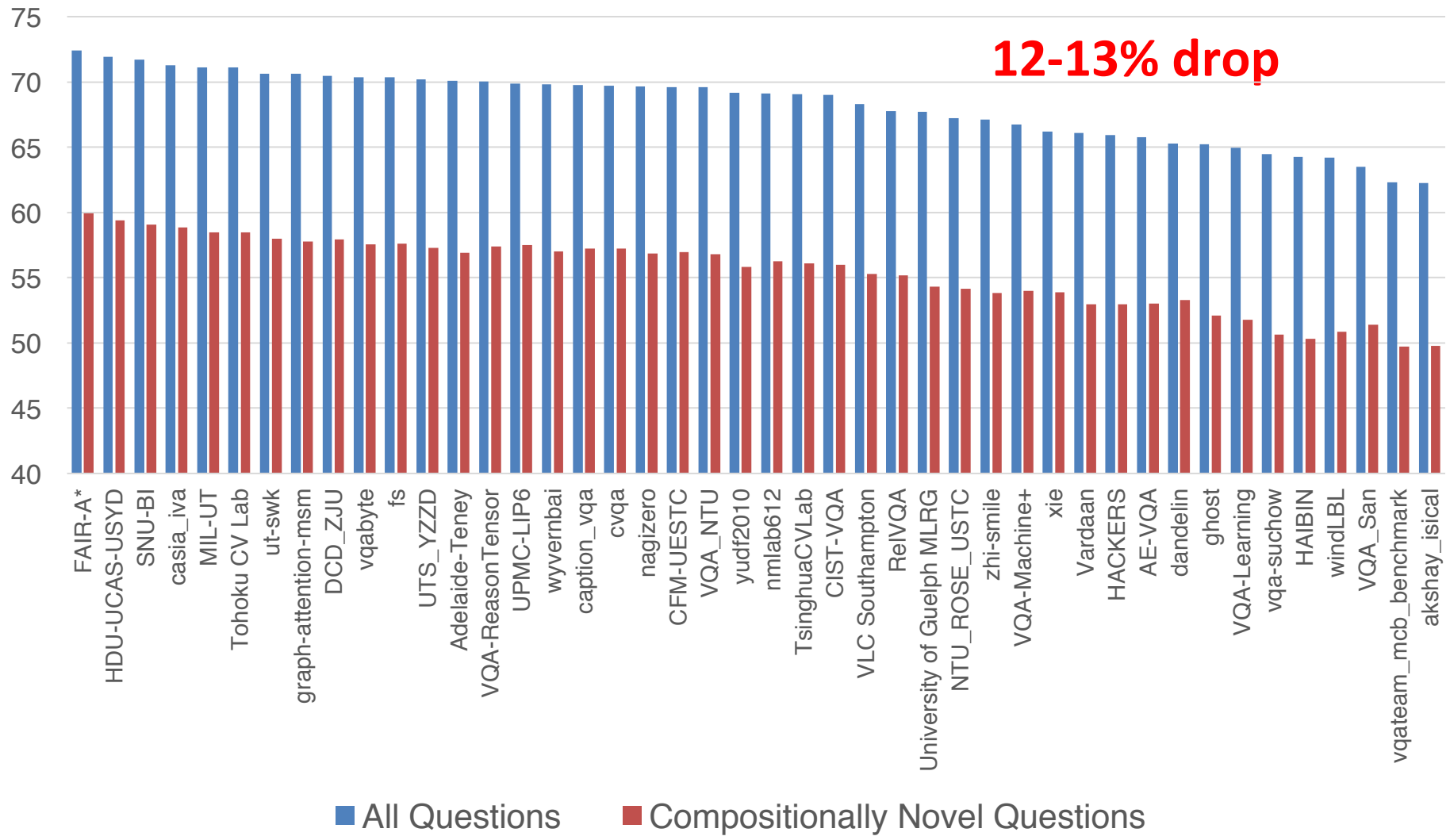
Q: What color is the stop light?

A: Green

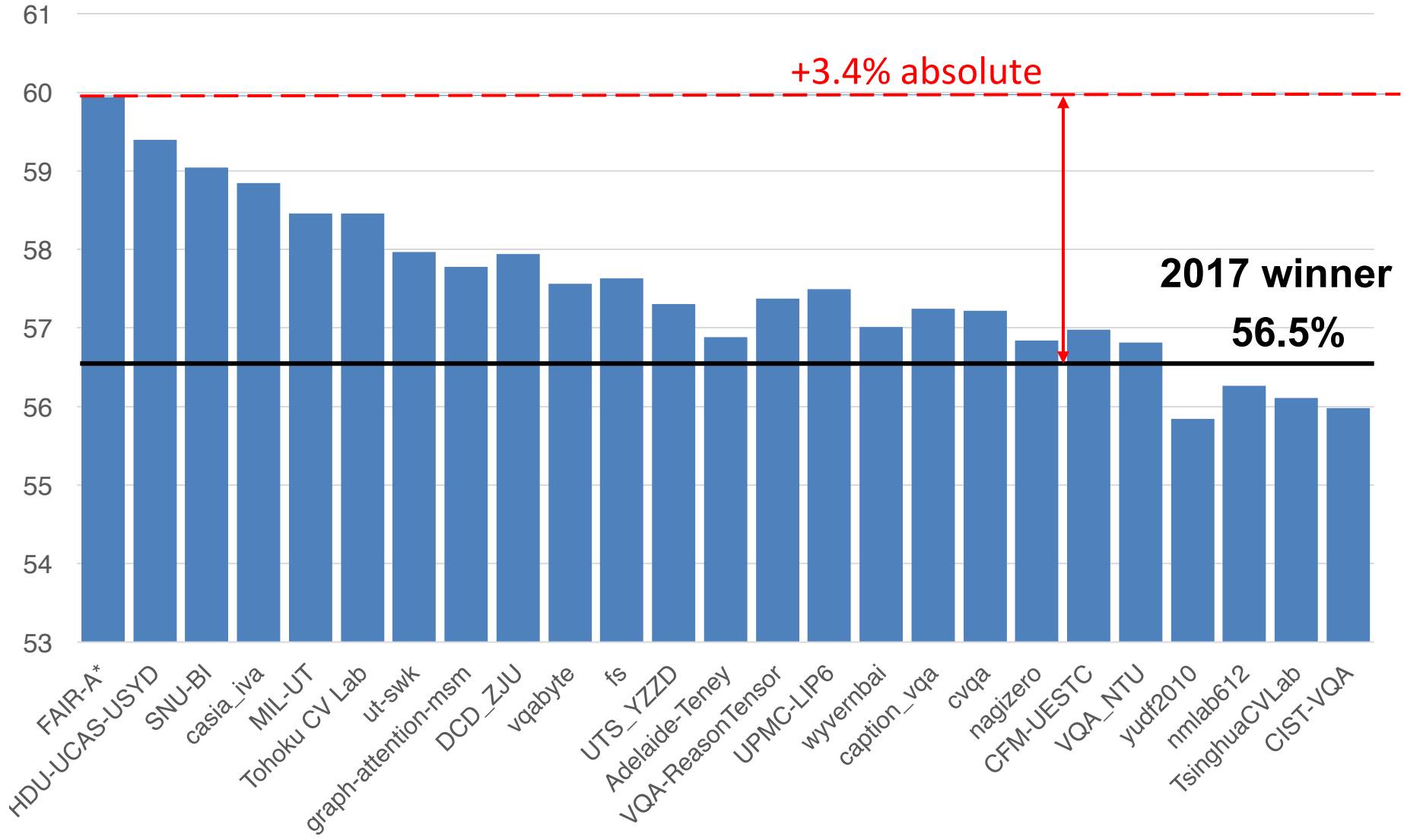
Q: What is the color of the plate?

A: Red

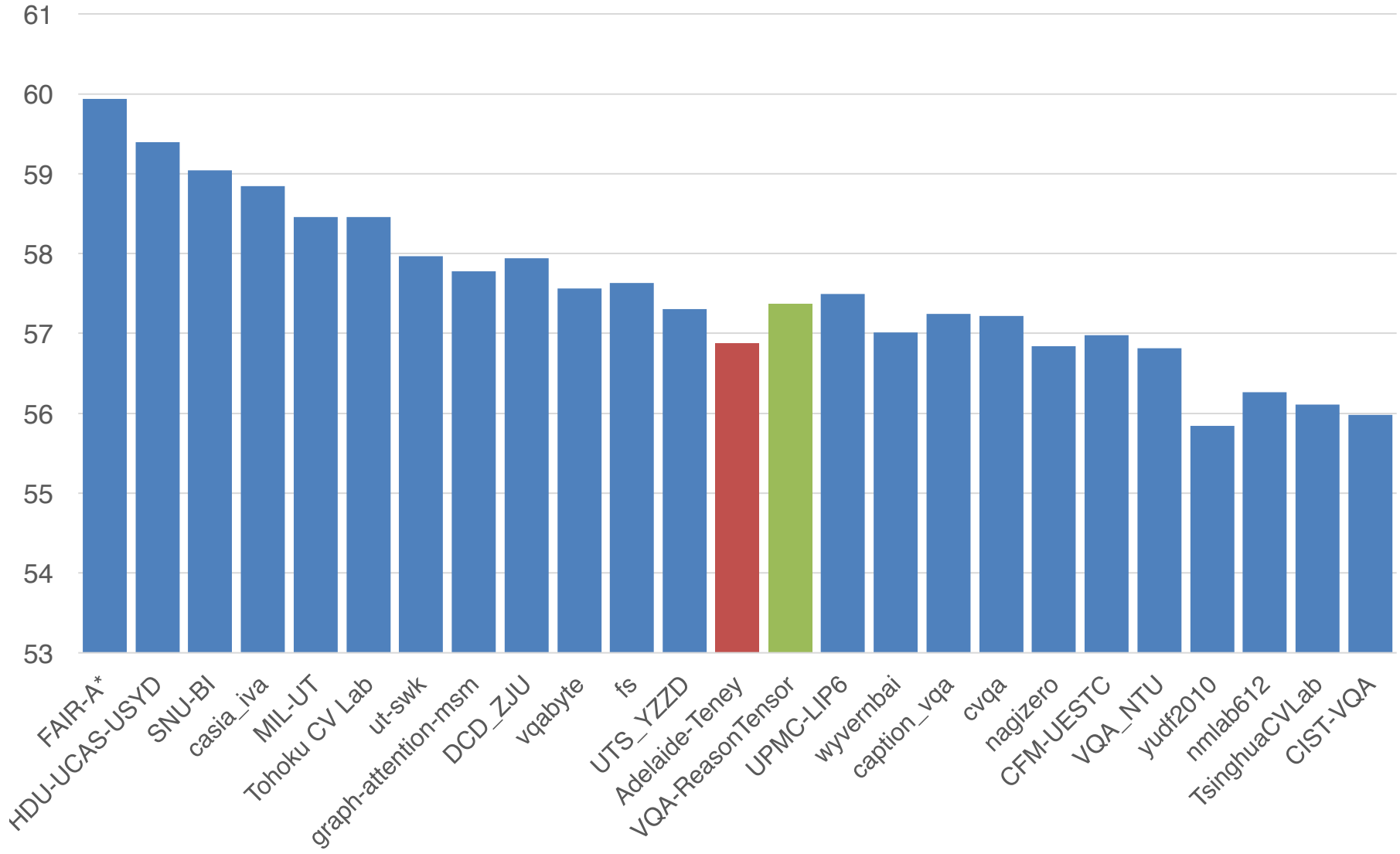
Are models compositional?



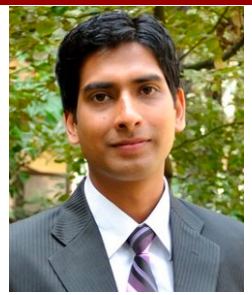
Are models compositional?



Are models compositional?



Average answer recall

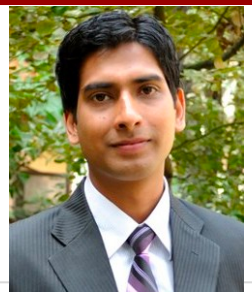


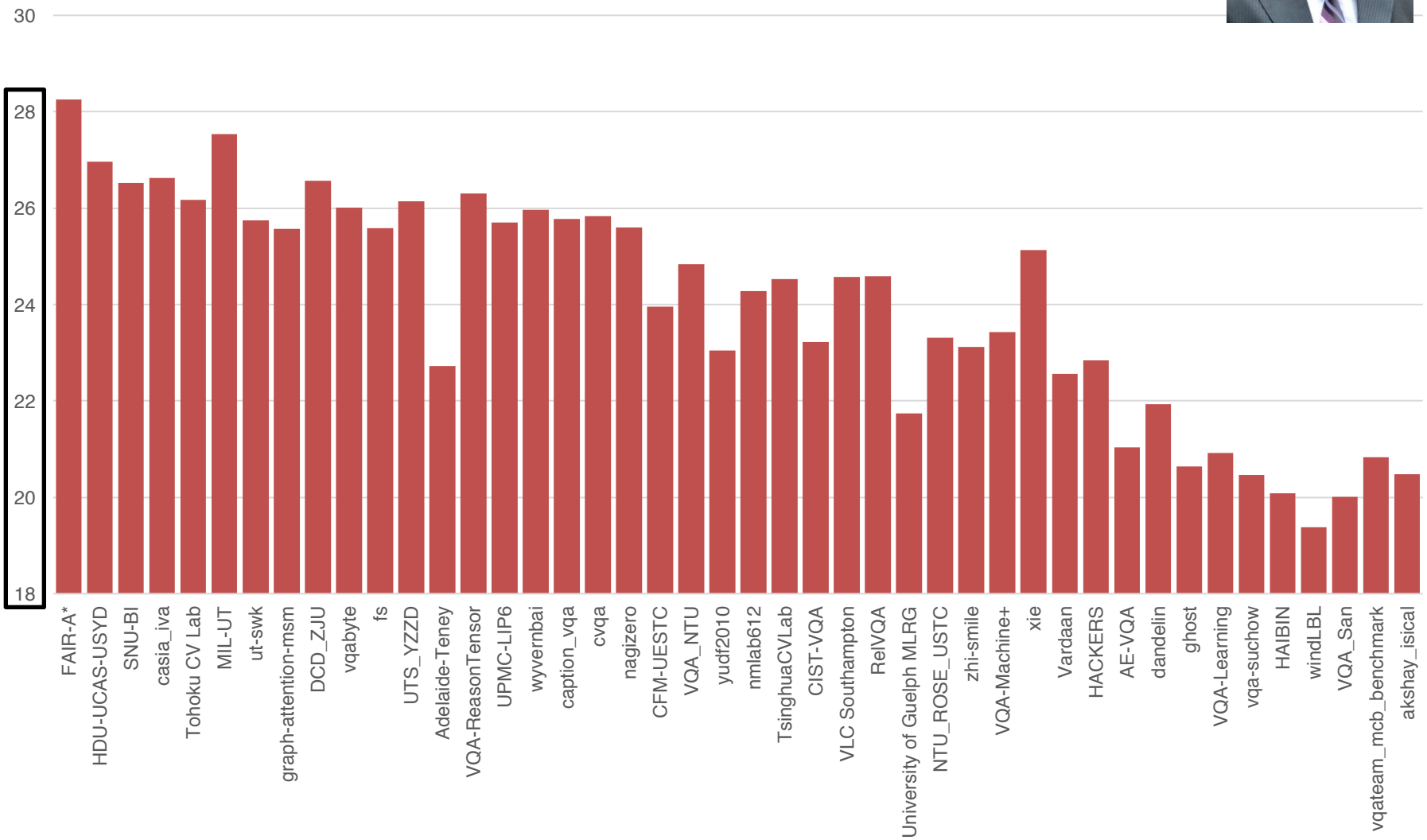
- New accuracy metric proposed in Kafle and Kannan, ICCV 17
  - Also known as “Normalized accuracy”

- Method:
  - Computes accuracy for each unique answer
  - Take the mean over all unique answers

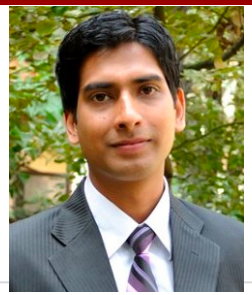
- Rewards models which perform well for rare answers

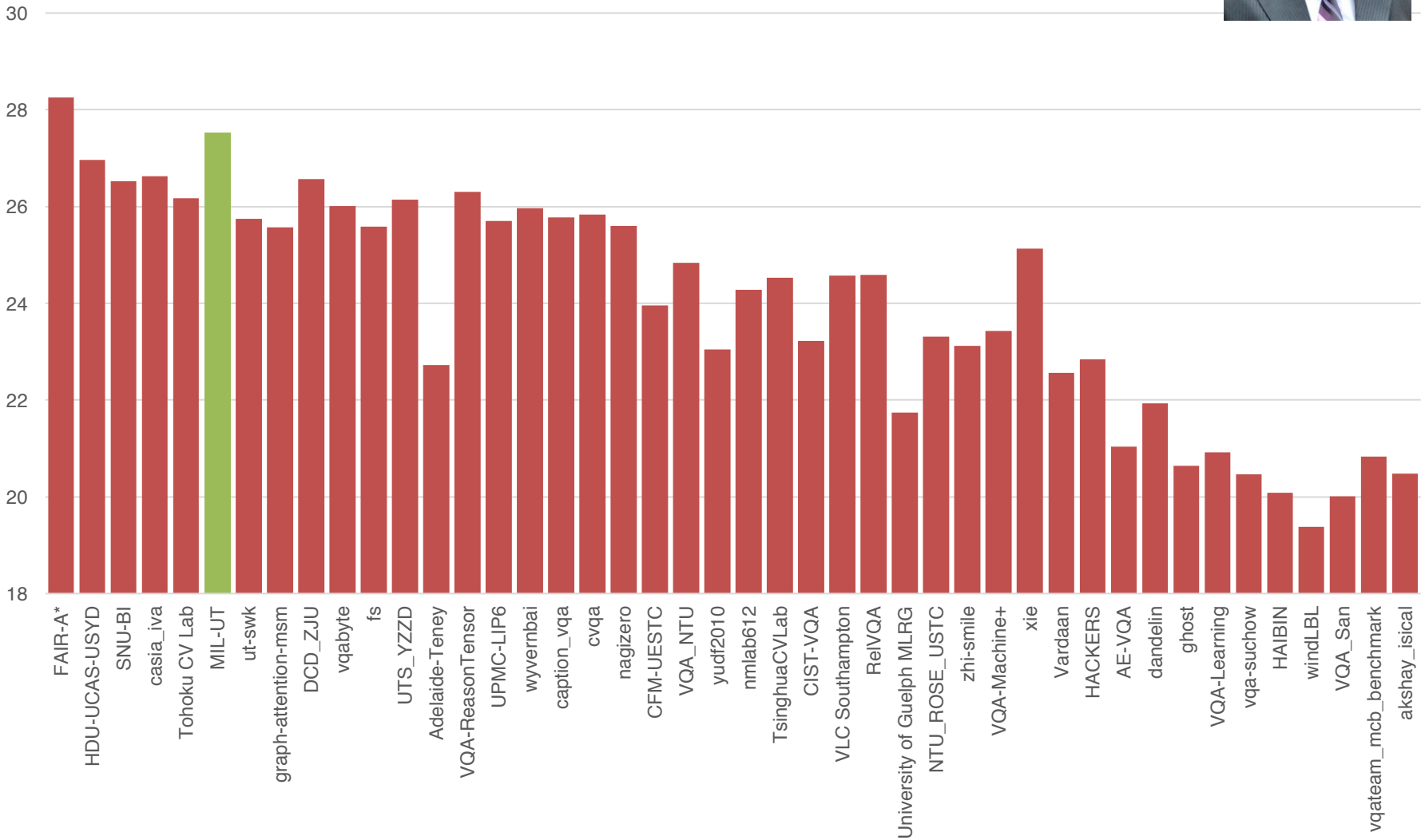
Average answer recall



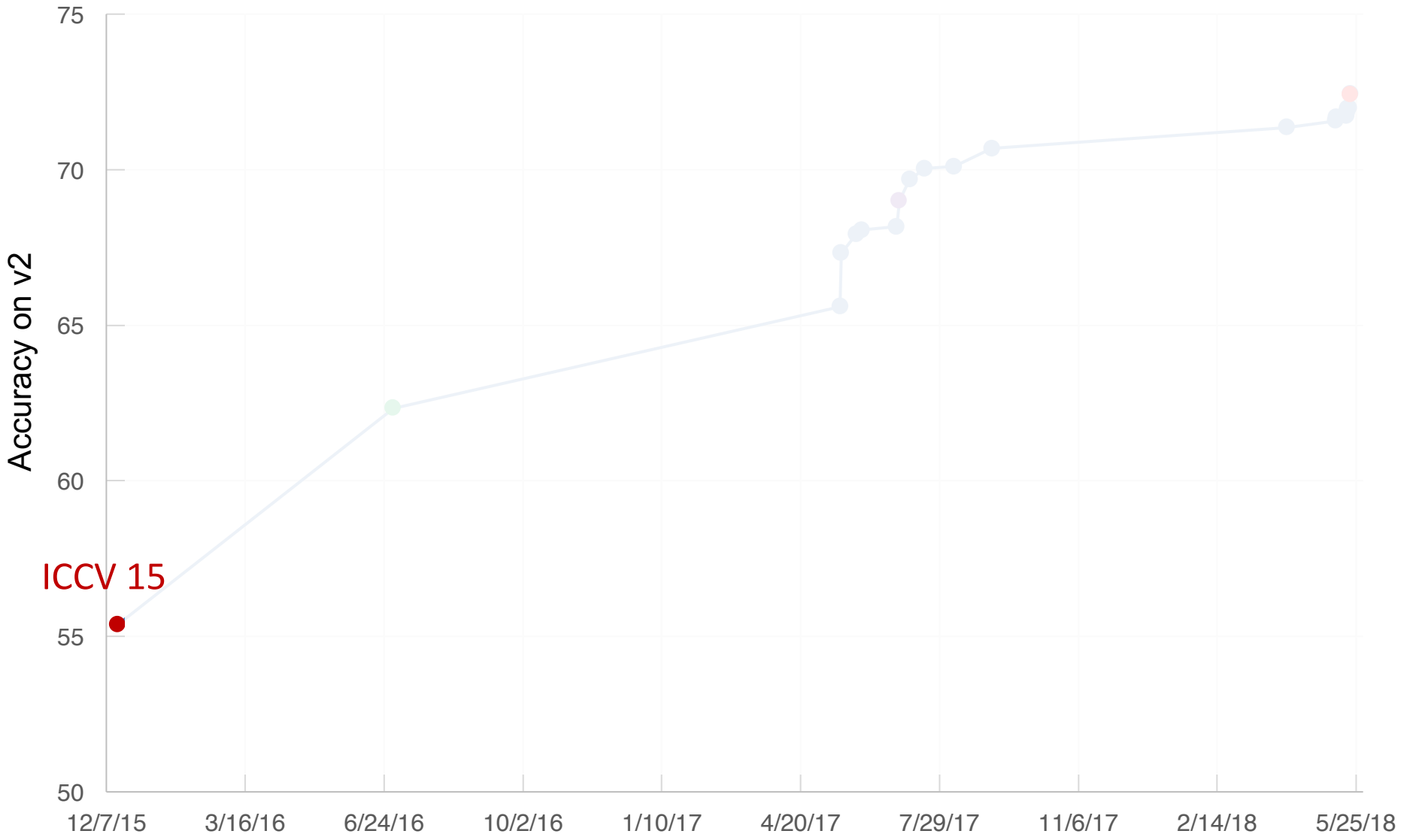


Average answer recall

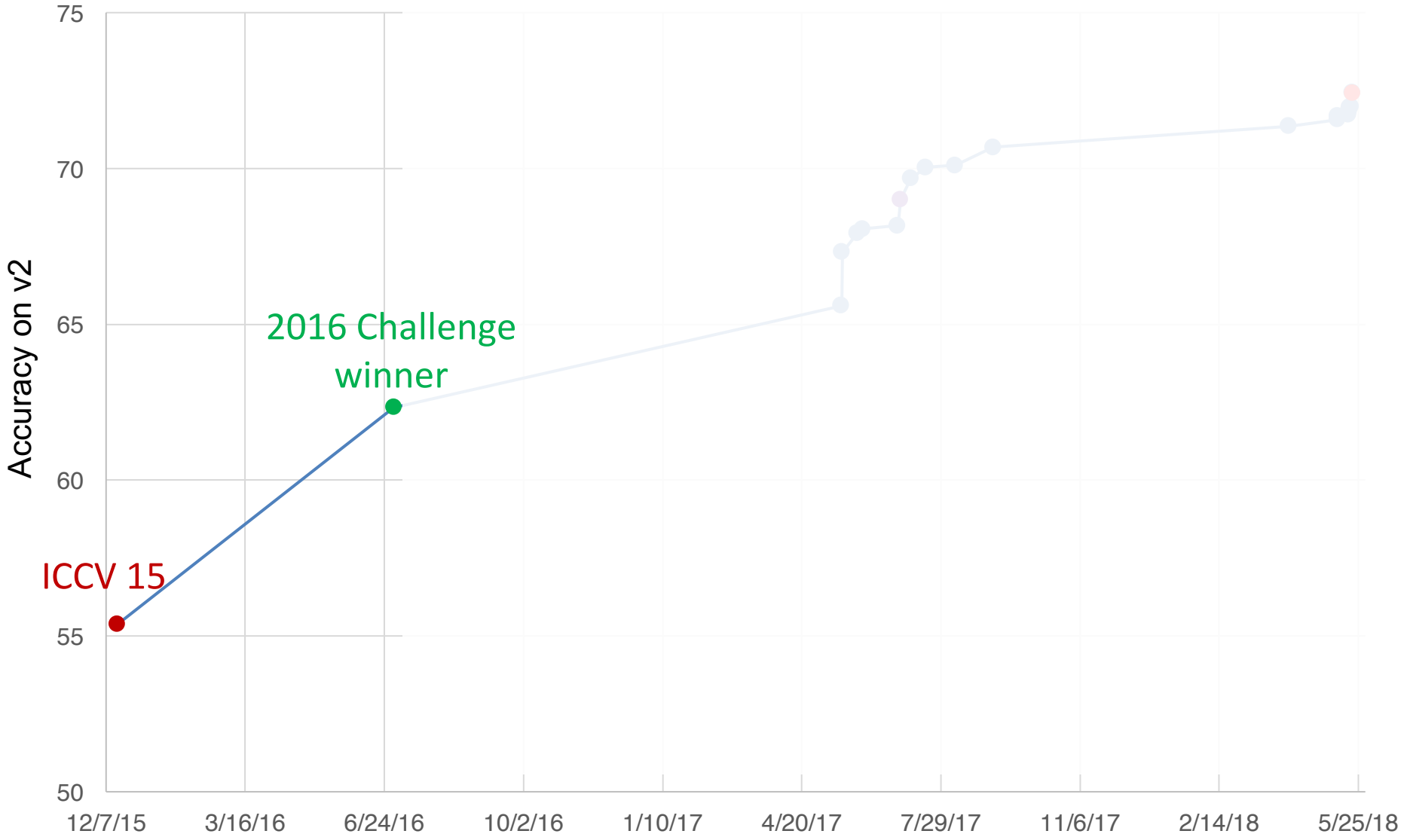




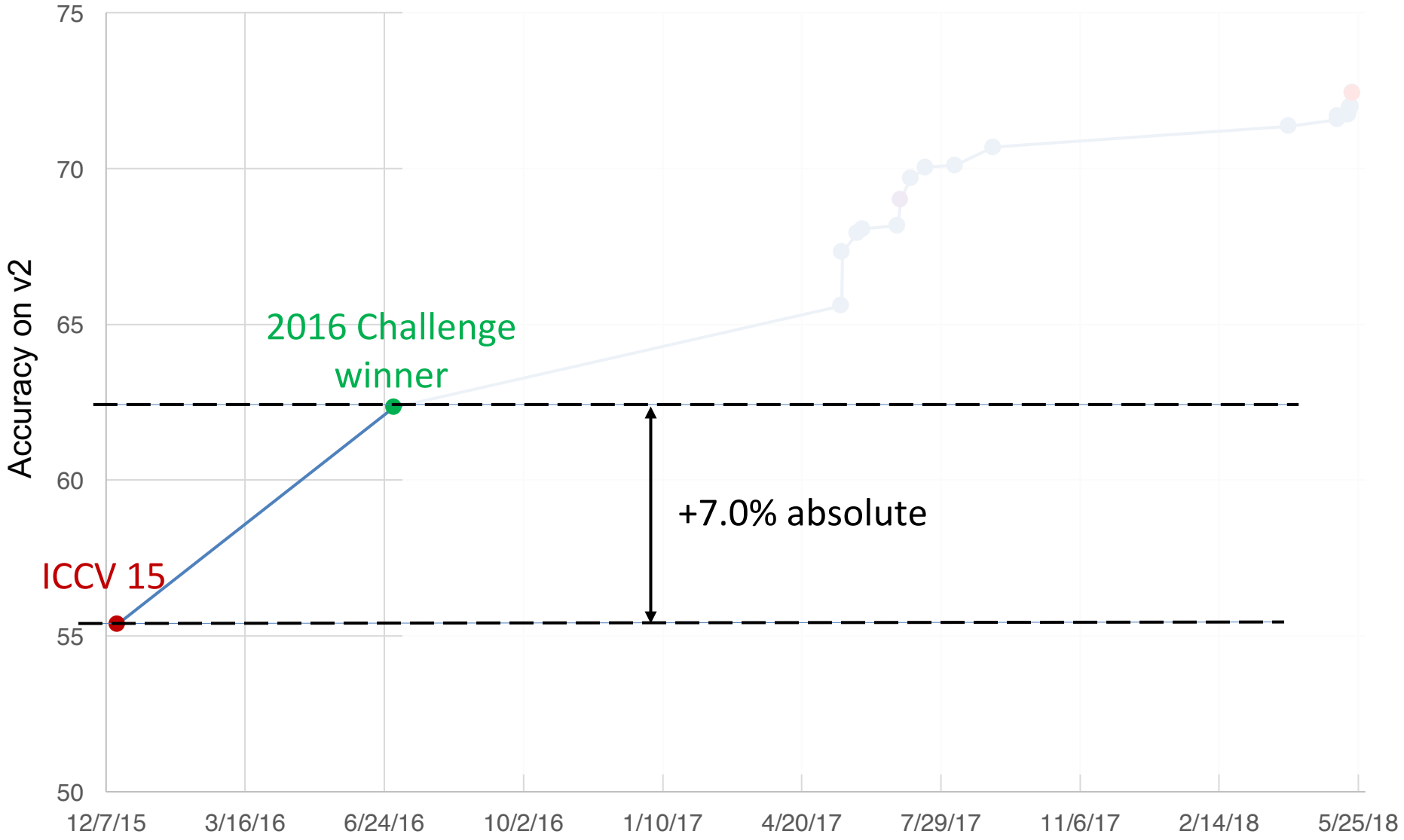
Progress in VQA



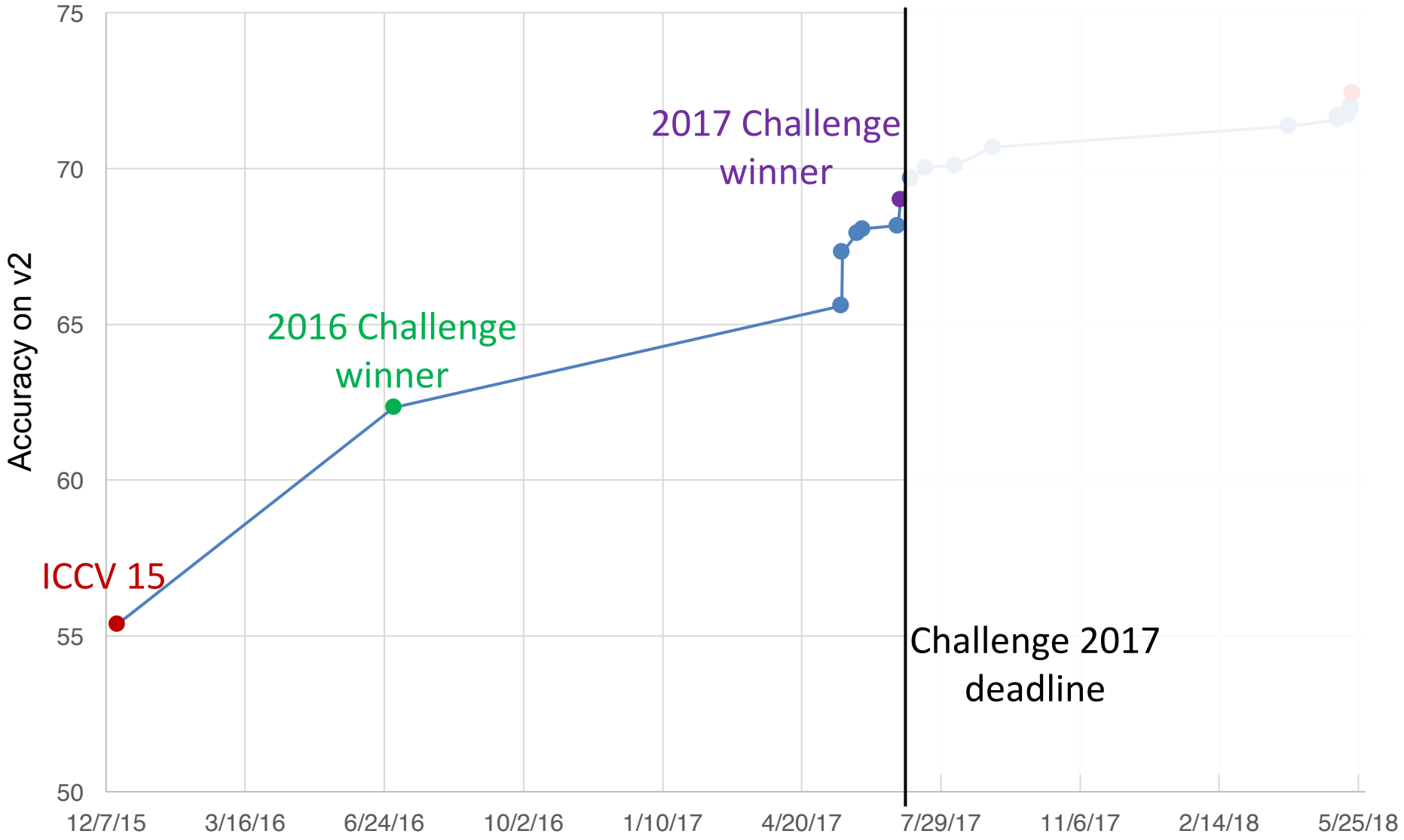
Progress in VQA



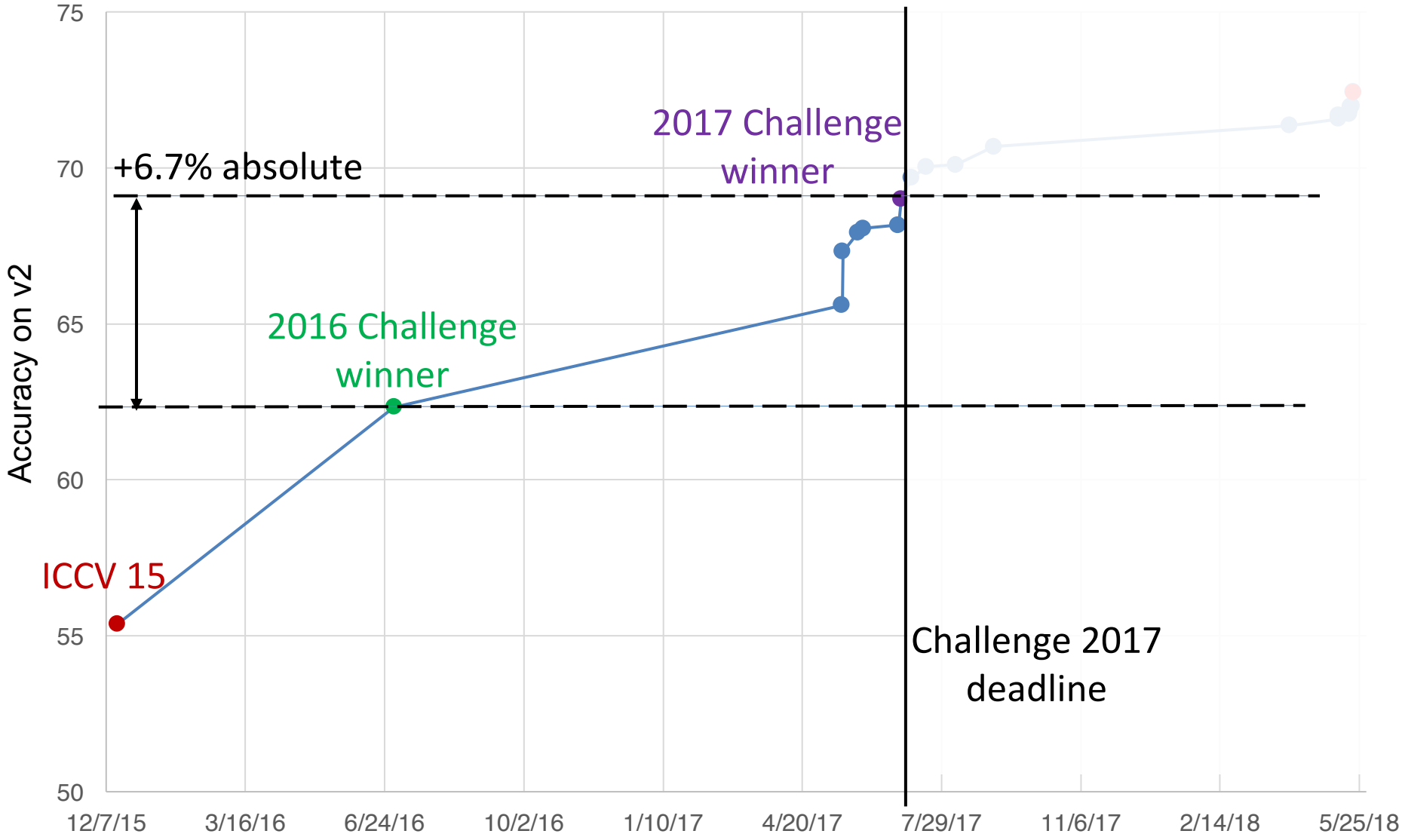
Progress in VQA



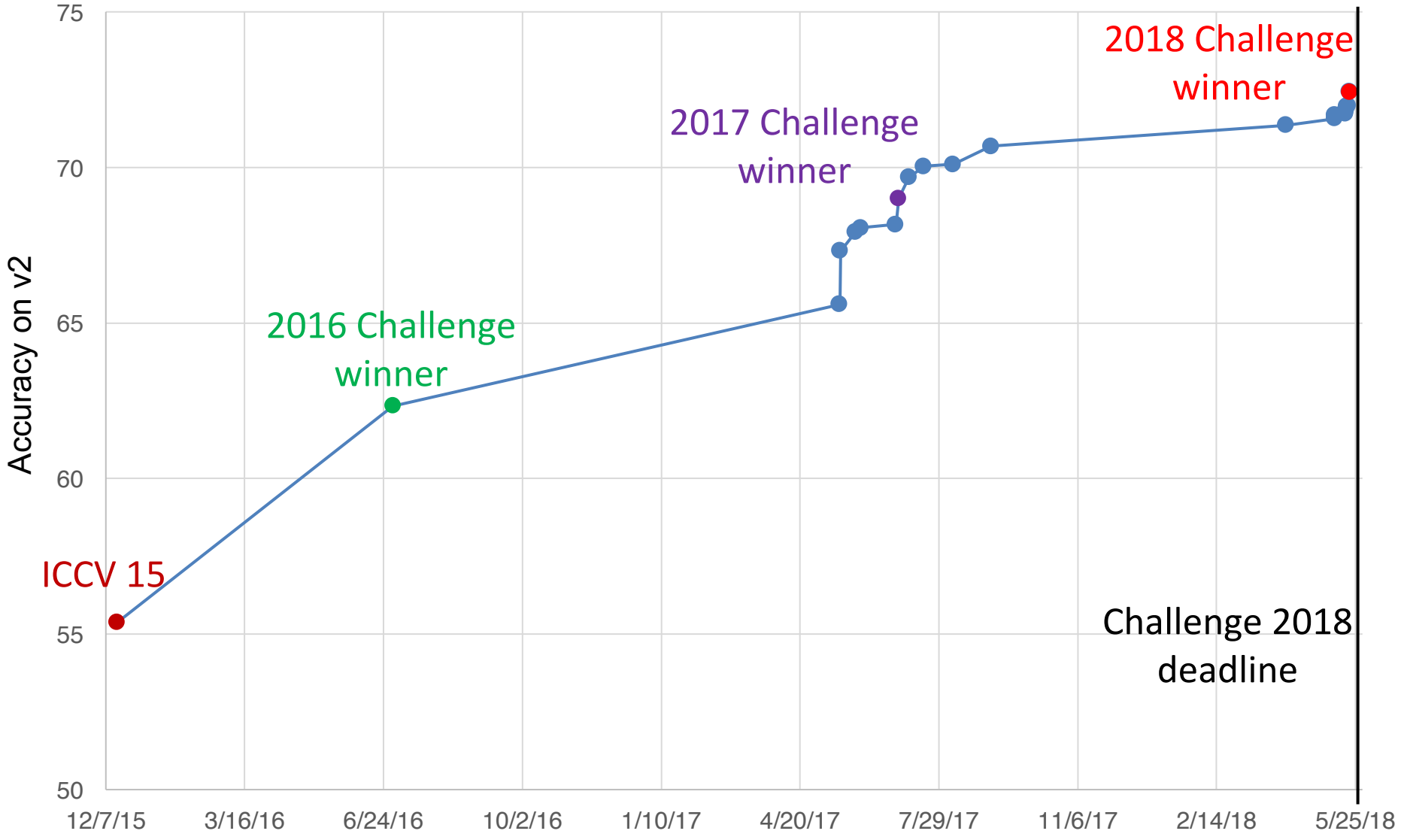
Progress in VQA



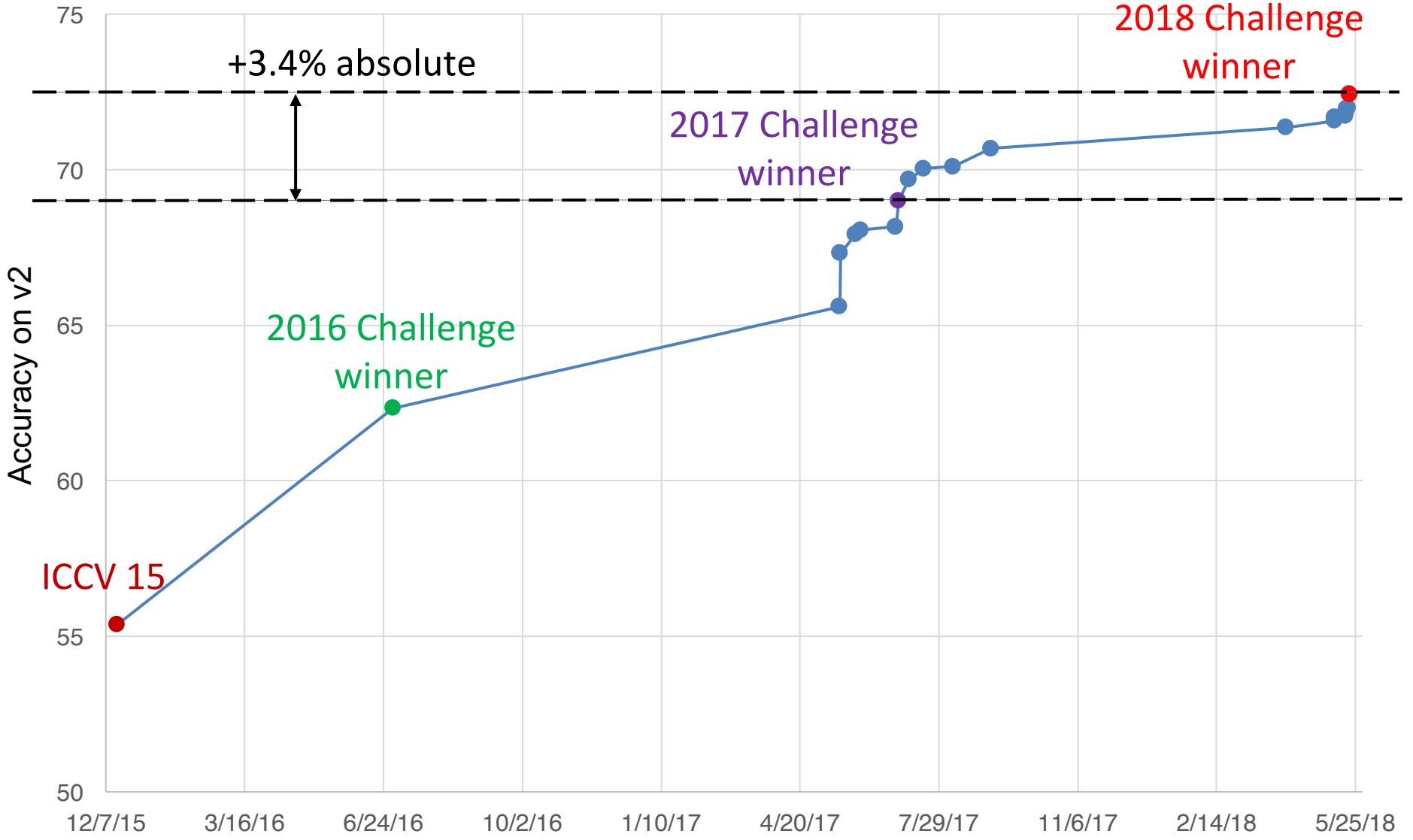
Progress in VQA



Progress in VQA



Progress in VQA



Visual Dialog Challenge 2018

VisDial v1.0

- ~130k images (COCO)
- 10-round dialog / image
- ~1.3 million QA pairs
- Evaluation
  - Automatic metrics
  - Human annotations

- Deadline: **mid-August, 2018**
- Results: **September 8th, 2018** at ECCV 2018

visualdialog.org/challenge/2018

Thanks!

Questions?