# BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski
Michael Collins, Kristina Toutanova

# Motivation: Inference

- Humans can infer many things from text

“The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016.”

- Pittsburgh has a sports team called the ``Penguins”
- The Sharks got second place in 2016
- The Sharks have never won the Stanley Cup

# Motivation: Testing Inference is Difficult

- Crowd-sourcing interesting examples can be challenging

“The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016.”

- The Sharks advanced to the Stanley cup in 2016
- The Sharks lost to the Pittsburgh Penguins

# Motivation: Testing Inference is Difficult

- Recognizing entailment is an artificial task
- Have to make a number of arbitrary decisions:
  - What things are important to infer?
  - How strictly should we define entailment?
  - What kinds of inferential abilities should be tested?
- Hard to interpret results

# This Work: Natural Yes/No Questions

- Yes/No questions generated without any prompting
  - No pre-specified source text or topic
  - No knowledge of the answer
  - Not required to write yes/no questions
- Paired with passages selected by independent annotators

Does Tyrion survive in Game of Thrones?

Did the US qualify for the World Cup?

# Natural Yes/No Questions

- Often require inference
- Are challenging for existing models
- Have an obvious end-task
- Real-word test of inference

# Example

**Question**: Do all neurons have the same action potential?

**Passage**: In the early development, the action potential of neurons is initially carried by calcium current. The longer opening times for the calcium channels can lead to action potentials that are considerably slower than those of mature neurons.

**Answer:** ?

# Example

**Question**: Do all neurons have the same action potential?

**Passage**: In the early development, the action potential of neurons is initially carried by calcium current. The longer opening times for the calcium channels can lead to action potentials that are considerably slower than those of mature neurons.

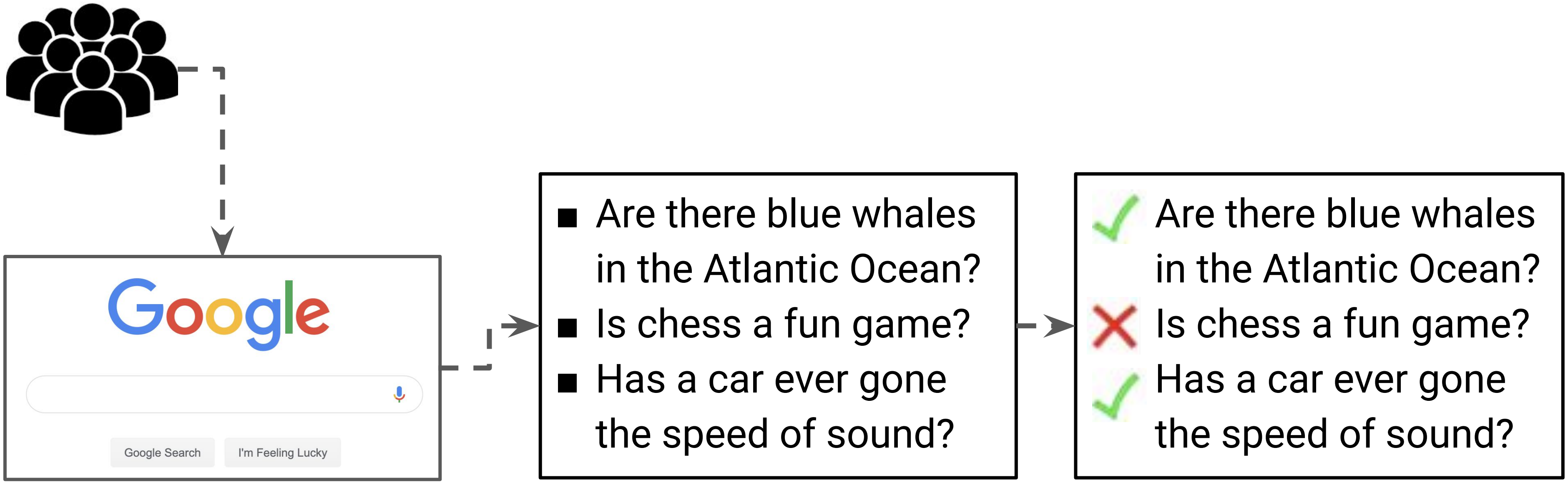
**Answer:** No

# The Rest of this Talk

- Dataset Construction
- Dataset Analysis
- Transfer Learning Baselines

# Dataset Construction

# Collecting Questions



- Are there blue whales in the Atlantic Ocean?
- Is chess a fun game?
- Has a car ever gone the speed of sound?

✓ Are there blue whales in the Atlantic Ocean?
✗ Is chess a fun game?
✓ Has a car ever gone the speed of sound?

Anonymized Queries

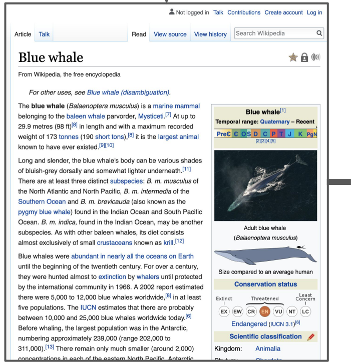
Heuristic Filtering

Manual Validation

# Collecting Passages

Are there blue whales in the Atlantic Ocean?

WIKIPEDIA



Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

## Blue whale

From Wikipedia, the free encyclopedia

*For other uses, see Blue whale (disambiguation).*

The **blue whale** (*Balaenoptera musculus*) is a marine mammal belonging to the baleen whale parvorder, Mysticeti.[7] At up to 29.9 metres (98 ft)[8] in length and with a maximum recorded weight of 173 tonnes (190 short tons),[8] it is the largest animal known to have ever existed.[9][10]

Long and slender, the blue whale's body can be various shades of bluish-grey dorsally and somewhat lighter underneath.[11] There are at least three distinct subspecies: *B. m. musculus* of the North Atlantic and North Pacific, *B. m. intermedia* of the Southern Ocean and *B. m. brevicauda* (also known as the pygmy blue whale) found in the Indian Ocean and South Pacific Ocean. *B. m. indica*, found in the Indian Ocean, may be another subspecies. As with other baleen whales, its diet consists almost exclusively of small crustaceans known as krill.[12]

Blue whales were abundant in nearly all the oceans on Earth until the beginning of the twentieth century. For over a century, they were hunted almost to extinction by whalers until protected by the international community in 1966. A 2002 report estimated there were 5,000 to 12,000 blue whales worldwide,[8] in at least five populations. The IUCN estimates that there are probably between 10,000 and 25,000 blue whales worldwide today.[6] Before whaling, the largest population was in the Antarctic, numbering approximately 239,000 (range 202,000 to 311,000).[13] There remain only much smaller (around 2,000) concentrations in each of the eastern North Pacific, Antarctic

**Blue whale**[1]

**Temporal range: Quaternary – Recent**

[2][3][4][5]

Adult blue whale

(*Balaenoptera musculus*)

Size compared to an average human

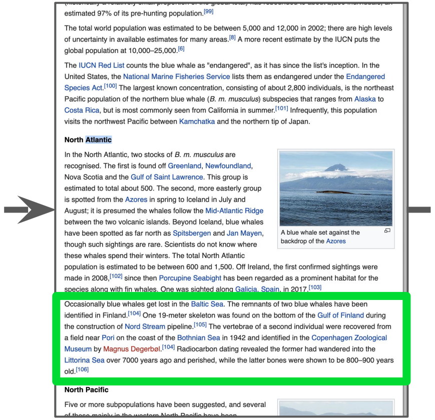
**Conservation status**

Endangered (IUCN 3.1)[6]

**Scientific classification**

| | |
|---|---|
| Kingdom: | Animalia |
| Phylum: | Chordata |

Document Selection



estimated 97% of its pre-hunting population.[99]

The total world population was estimated to be between 5,000 and 12,000 in 2002; there are high levels of uncertainty in available estimates for many areas.[8] A more recent estimate by the IUCN puts the global population at 10,000–25,000.[6]

The IUCN Red List counts the blue whale as "endangered", as it has since the list's inception. In the United States, the National Marine Fisheries Service lists them as endangered under the Endangered Species Act.[100] The largest known concentration, consisting of about 2,800 individuals, is the northeast Pacific population of the northern blue whale (*B. m. musculus*) subspecies that ranges from Alaska to Costa Rica, but is most commonly seen from California in summer.[101] Infrequently, this population visits the northwest Pacific between Kamchatka and the northern tip of Japan.

### North Atlantic

In the North Atlantic, two stocks of *B. m. musculus* are recognised. The first is found off Greenland, Newfoundland, Nova Scotia and the Gulf of Saint Lawrence. This group is estimated to total about 500. The second, more easterly group is spotted from the Azores in spring to Iceland in July and August; it is presumed the whales follow the Mid-Atlantic Ridge between the two volcanic islands. Beyond Iceland, blue whales have been spotted as far north as Spitsbergen and Jan Mayen, though such sightings are rare. Scientists do not know where these whales spend their winters. The total North Atlantic population is estimated to be between 600 and 1,500. Off Ireland, the first confirmed sightings were made in 2008,[102] since then Porcupine Seabight has been regarded as a prominent habitat for the species along with fin whales. One was sighted along Galicia, Spain, in 2017.[103]

A blue whale set against the backdrop of the Azores

Occasionally blue whales get lost in the Baltic Sea. The remnants of two blue whales have been identified in Finland.[104] One 19-meter skeleton was found on the bottom of the Gulf of Finland during the construction of Nord Stream pipeline.[105] The vertebrae of a second individual were recovered from a field near Pori on the coast of the Bothnian Sea in 1942 and identified in the Copenhagen Zoological Museum by Magnus Degerbøl.[104] Radiocarbon dating revealed the former had wandered into the Littorina Sea over 7000 years ago and perished, while the latter bones were shown to be 800–900 years old.[106]

### North Pacific

Five or more subpopulations have been suggested, and several of these mainly in the western North Pacific have been

Paragraph Selection

Occasionally blue whales get lost in the Baltic Sea. The remnants of two blue whales have been identified in Finland.[104] One 19-meter skeleton was found on the bottom of the Gulf of Finland during the construction of Nord Stream pipeline.[105] The vertebrae of a second individual were recovered from a field near Pori on the coast of the Bothnian Sea in 1942 and identified in the Copenhagen Zoological Museum by Magnus Degerbøl.[104] Radiocarbon dating revealed the former had wandered into the Littorina Sea over 7000 years ago and perished, while the latter bones were shown to be 800–900 years old.[106]

Yes ☑
No □

Answer Selection

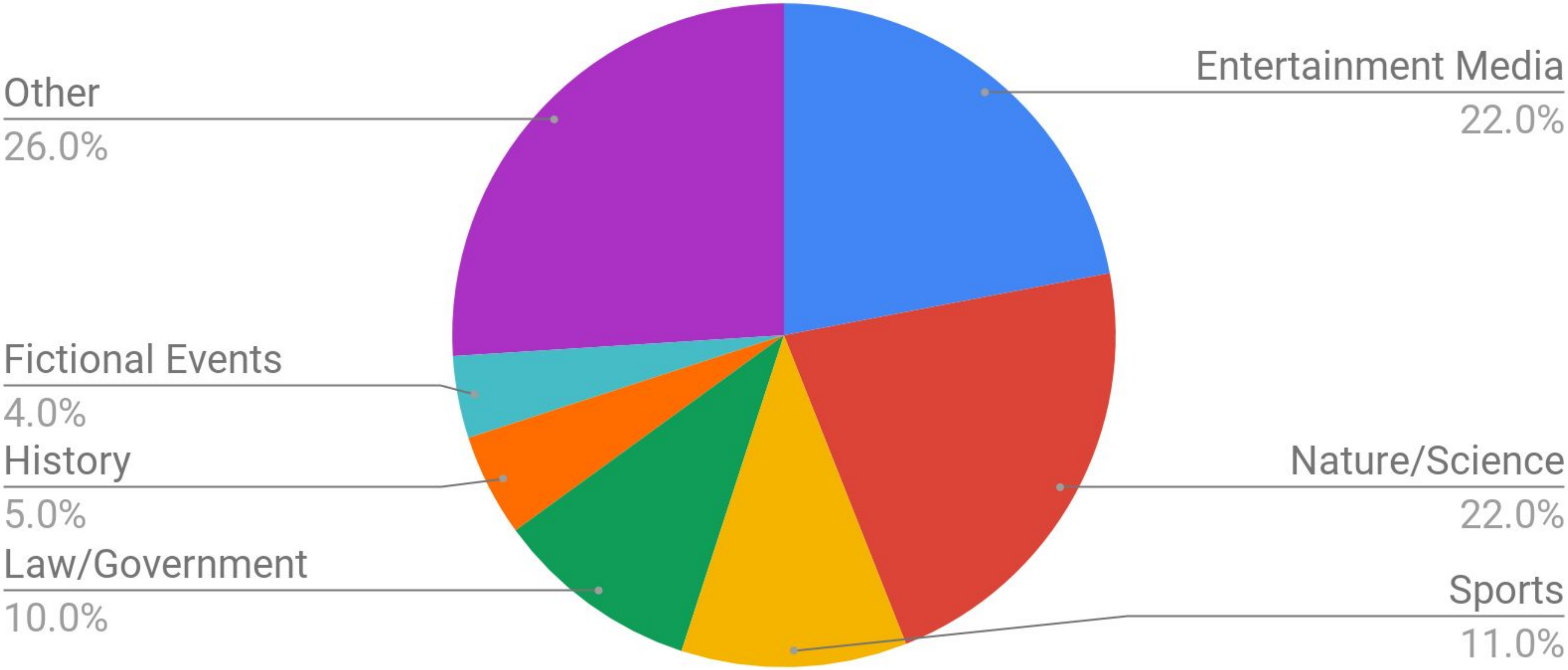
Pipeline from Natural Questions (Kwiatkowski et al., 2019)

# The Dataset

- (Question, Paragraph, Answer) triples where the answer is either “yes” or “no”
- 9.4k train questions
- 3.2k dev/test questions
- 62% “Yes” answers
- 110 average paragraph tokens
- 90% human performance

# Dataset Analysis

# Question Topics



# Paraphrasing (38.7%)

The passage explicitly asserts or refutes what is stated in the question

**Question**: Is Tim Brown in the Hall of Fame?

**Passage**: …Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame.

**Answer**: Yes

# By Example (11.8%)

The passage provides an example or counter-example to what is asserted by the question

**Question**: Has the UK been hit by a hurricane?

**Passage**:  The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands…

**Answer**: Yes

# Factual Reasoning (8.5%)

Answering the question requires using world-knowledge to connect what is stated in the passage to the question

**Question**: Was Designated Survivor filmed in the White House?

**Passage**:  The series is. . . filmed in Toronto, Ontario

**Answer**: No

# Implicit (8.5%)

The passage mentions or describes entities in the question in way that would not make sense if the answer was not yes/no

**Question**: Is static pressure the same as atmospheric pressure?

**Passage**: The aircraft designer’s objective is to ensure the pressure in the aircraft’s static pressure system is as close as possible to the atmospheric pressure…

**Answer**: No

# Missing Mention (6.6%)

We can conclude the answer is yes or no because, if this was not the case, it would have been mentioned in the passage

**Question**: Did Mickey Rourke win an Oscar for the Wrestler?

**Passage**:  In the 2008 film The Wrestler… Rourke received a 2009 Golden Globe award, a BAFTA award, and an Academy Award nomination…

**Answer**: No

# Other Inference (25.9%)

The passage states a fact that can be used to infer whether the answer is true or false, and does not fall into any of the other categories

**Question**: Is the sea snake the most venomous snake?

**Passage**:  ...the venom of the inland taipan, drop by drop, is the most toxic among all snakes

**Answer**: No

# Why are Yes/No Question Interesting?

- Rarely factoid
    - Unusual to ask “Was Obama born in 1961?”
- “No” Answers usually have to be inferred
- Easy to use non-trivial kinds of reasoning when labelling them

# Experiments

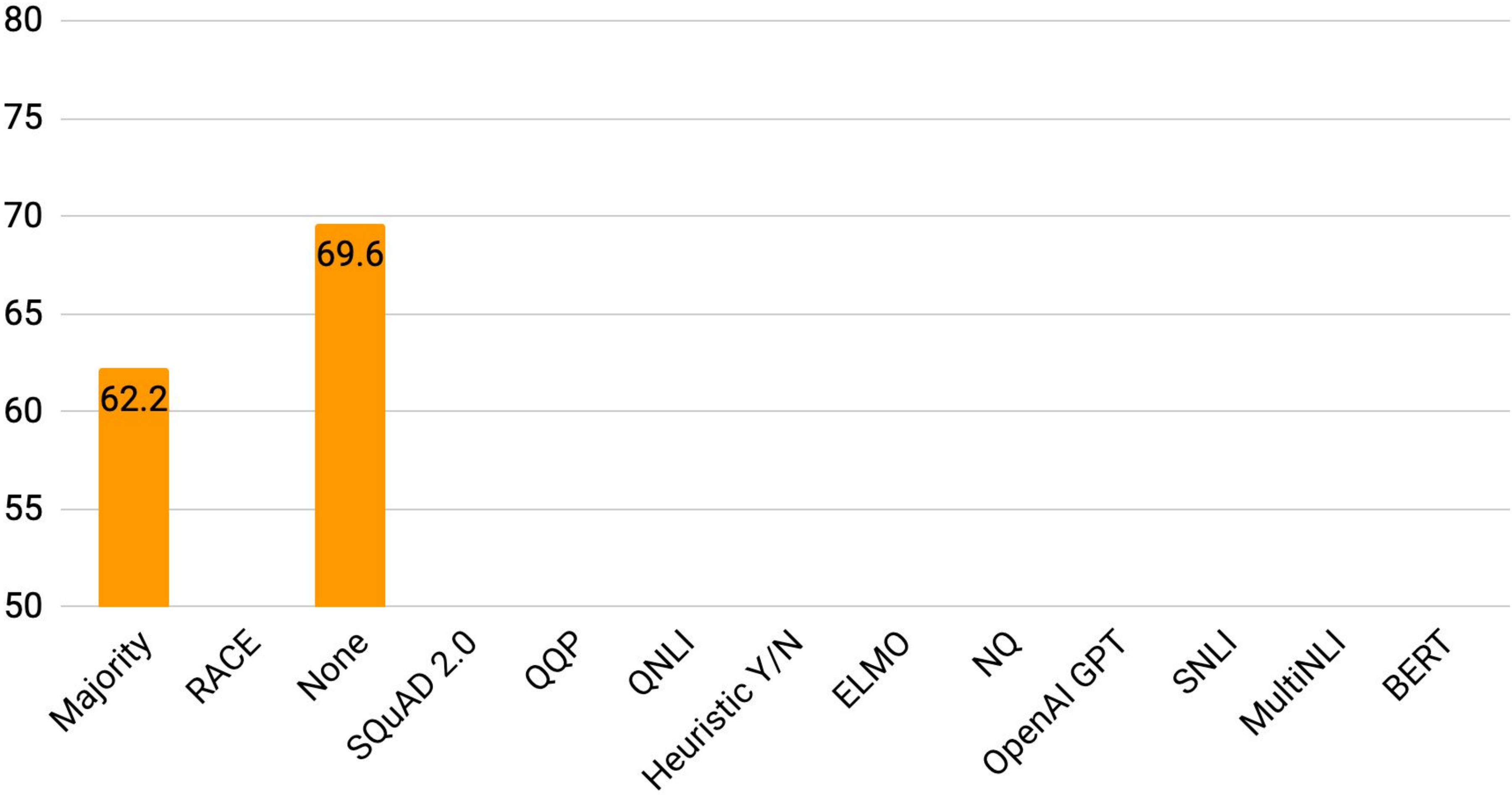
# Simple Baselines

- Majority Guess: 62.2%
- Question-Only $\text{BERT}_L$ model: 64.5%
- Passage-Only $\text{BERT}_L$ model: 66.7%
- Word-Overlap Model: 62.2%

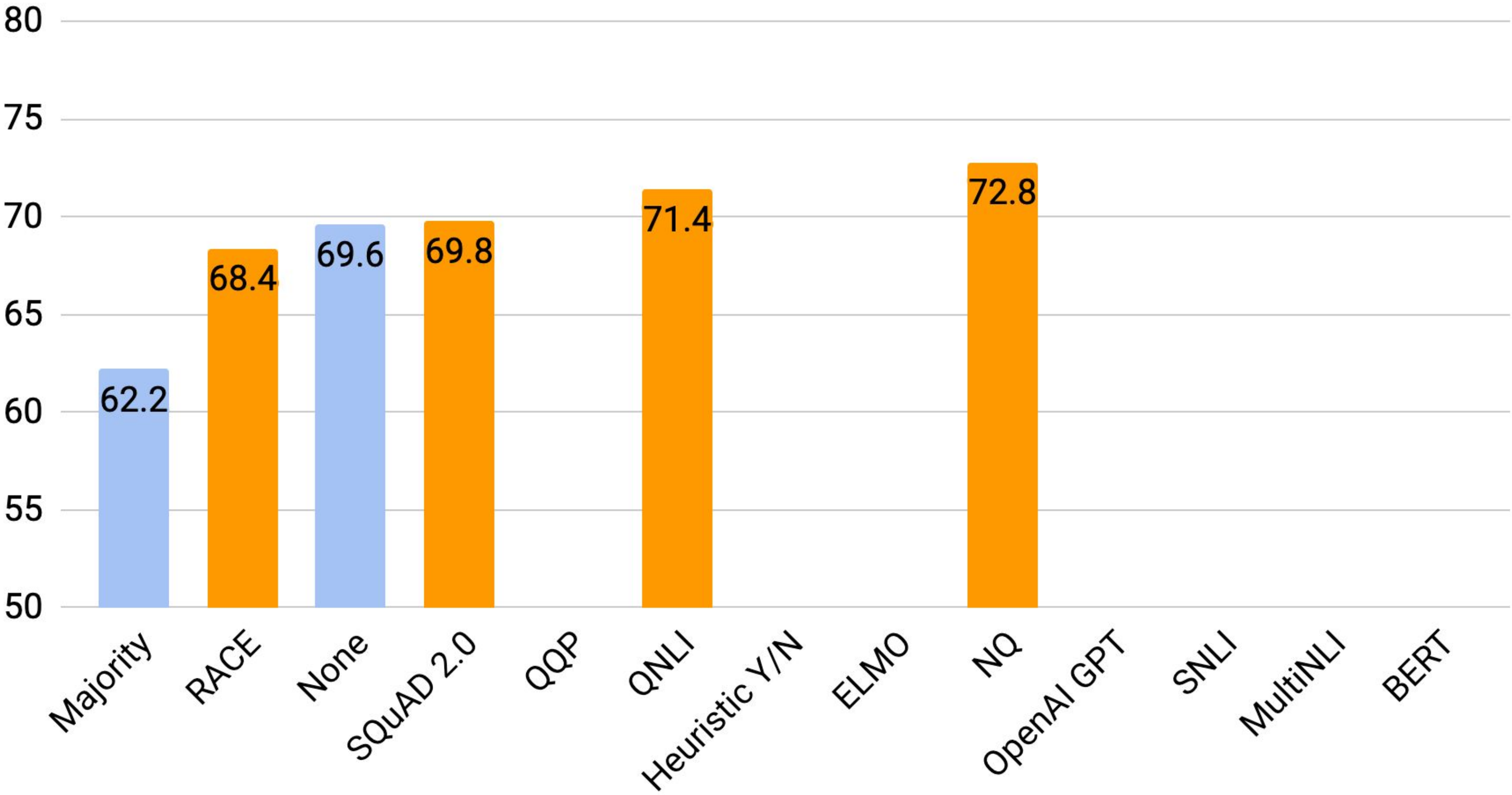
# Transfer Baselines

- Supervised transfer sources:
    - Question Answering (SQuAD, QNLI, NQ)
    - Entailment (MNLI, SNLI)
    - Paraphrasing (QQP)
    - Heuristic Y/N data (MSMarco)
- Supervised tasks are used to pre-train a standard recurrent + co-attention model (see paper for details)
- Recent unsupervised transfer methods (BERT, OpenAI GPT, ELMo)

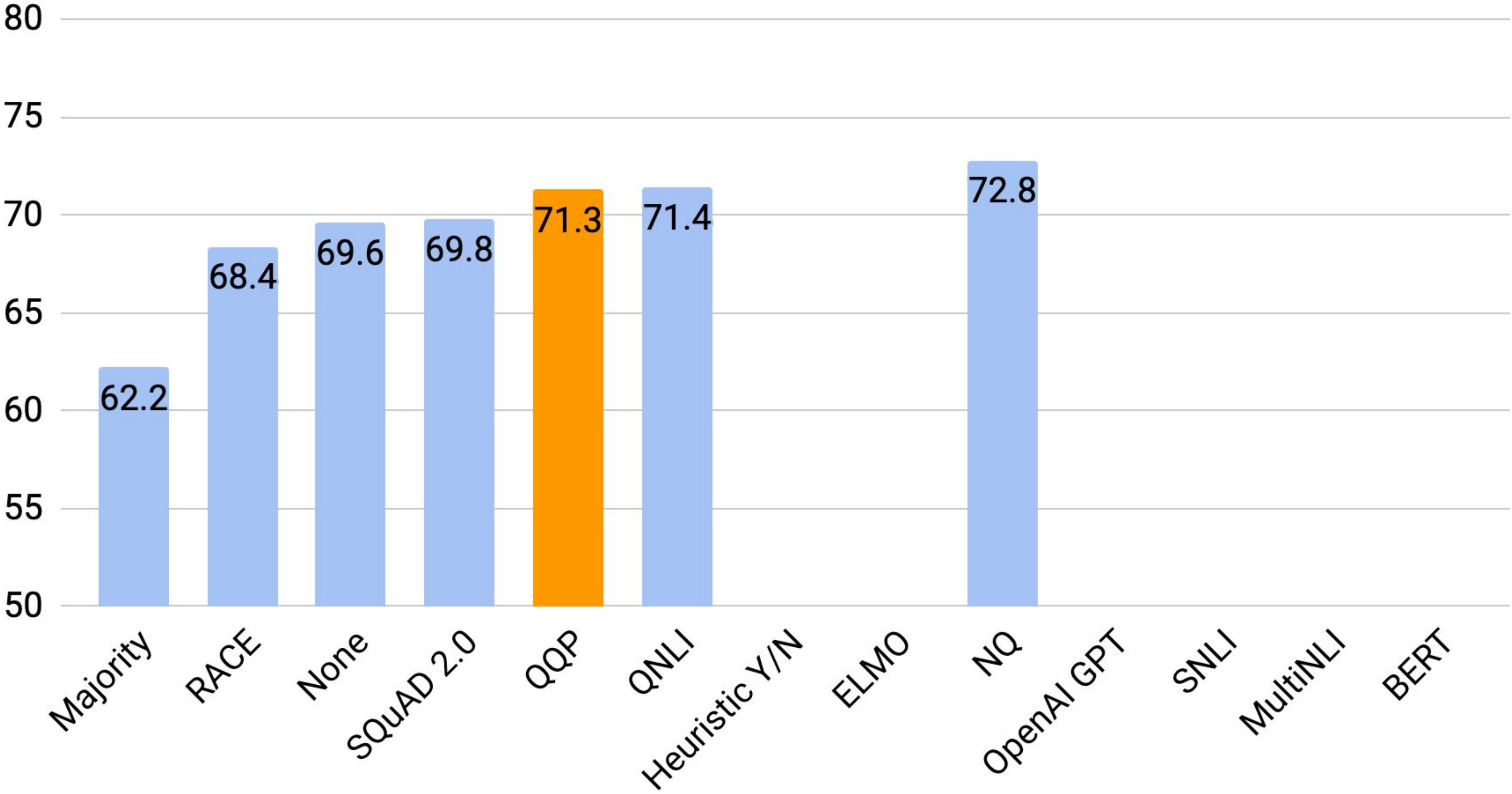
# No Transfer


80
75
70
65
60
55
50
62.2
69.6
Majority
RACE
None
SQuAD 2.0
QQP
QNLI
Heuristic Y/N
ELMO
NQ
OpenAI GPT
SNLI
MultiNLI
BERT

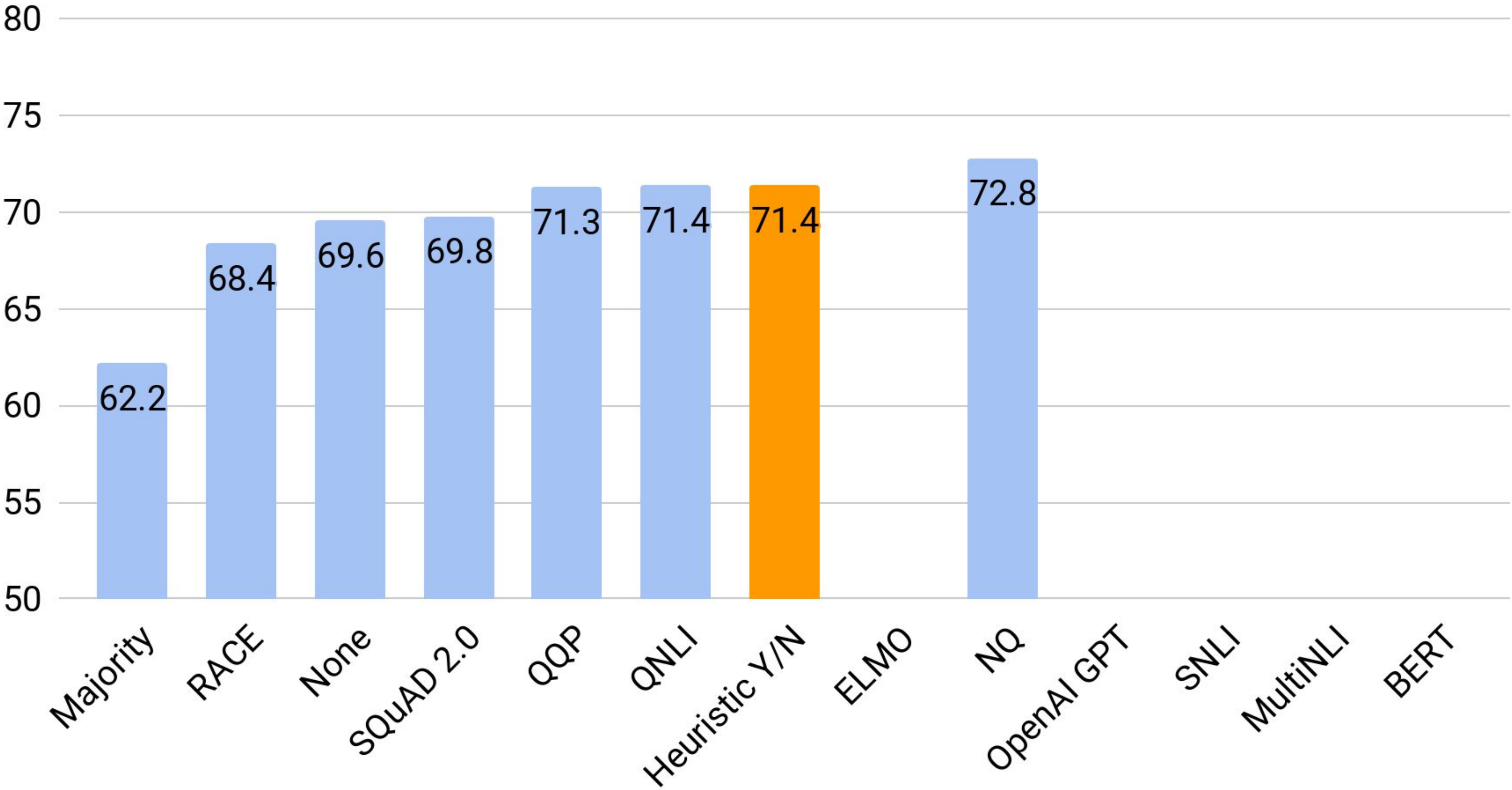
# Question Answering



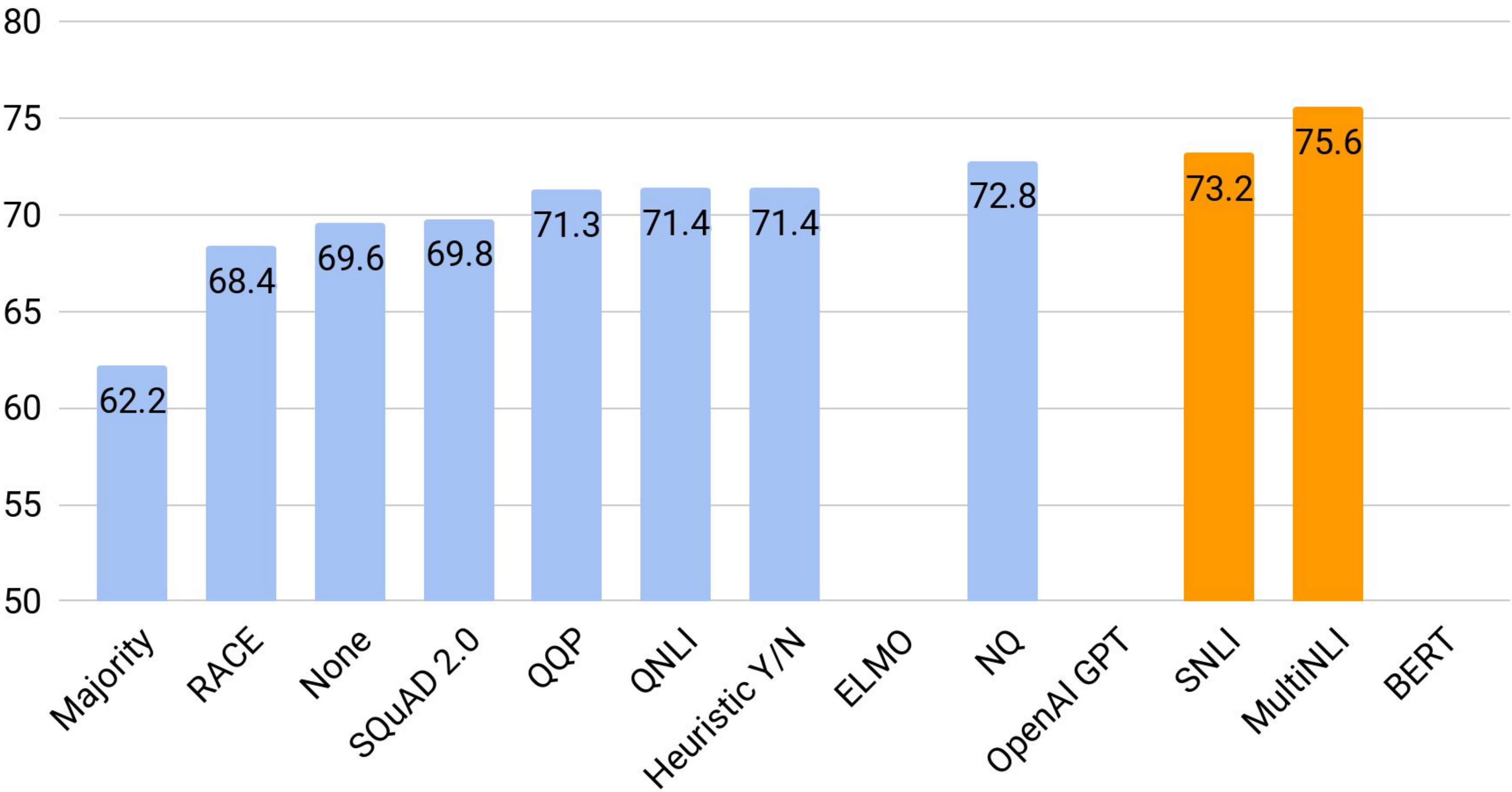
# Paraphrasing



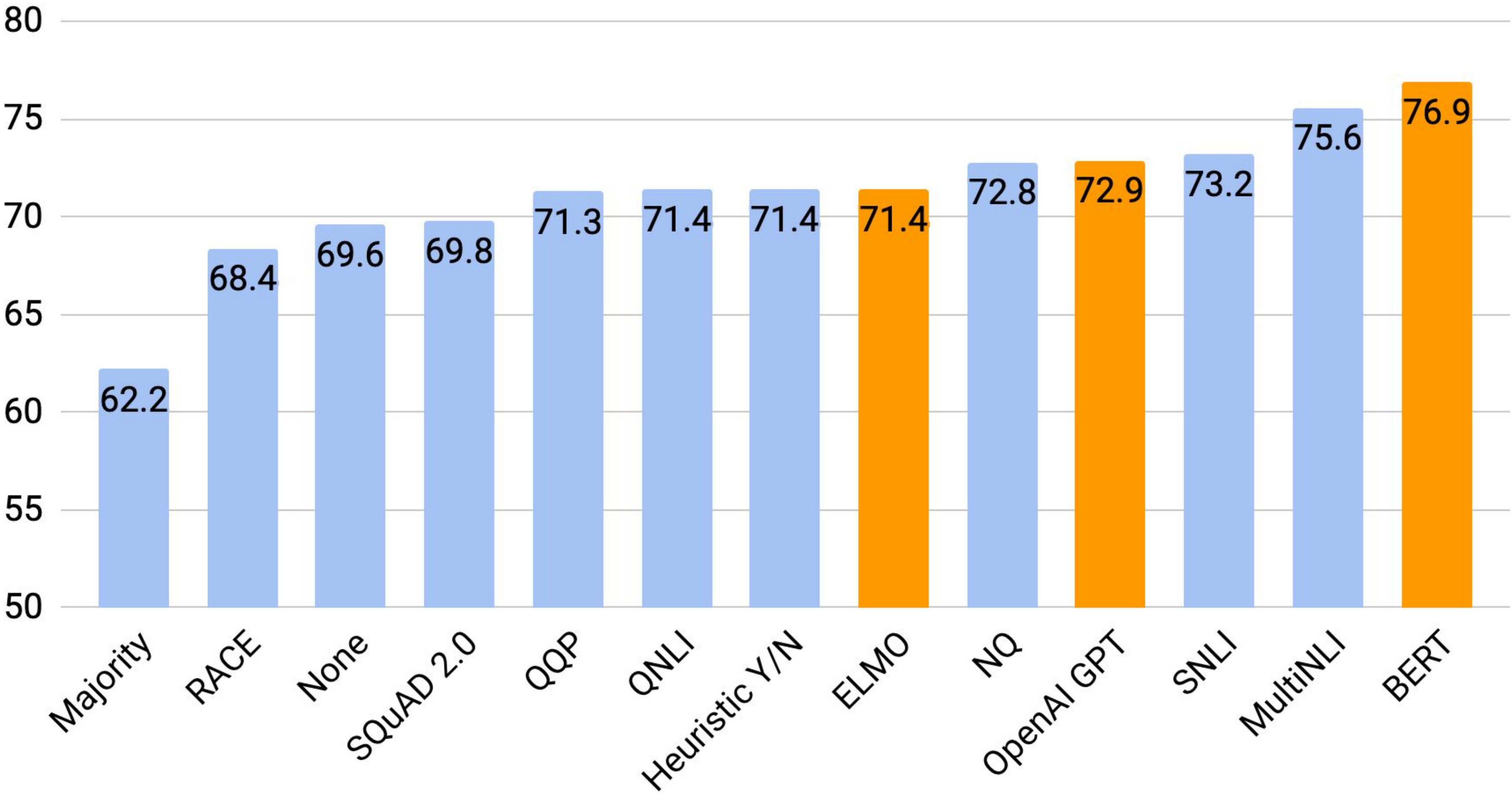
# Heuristic Y/N



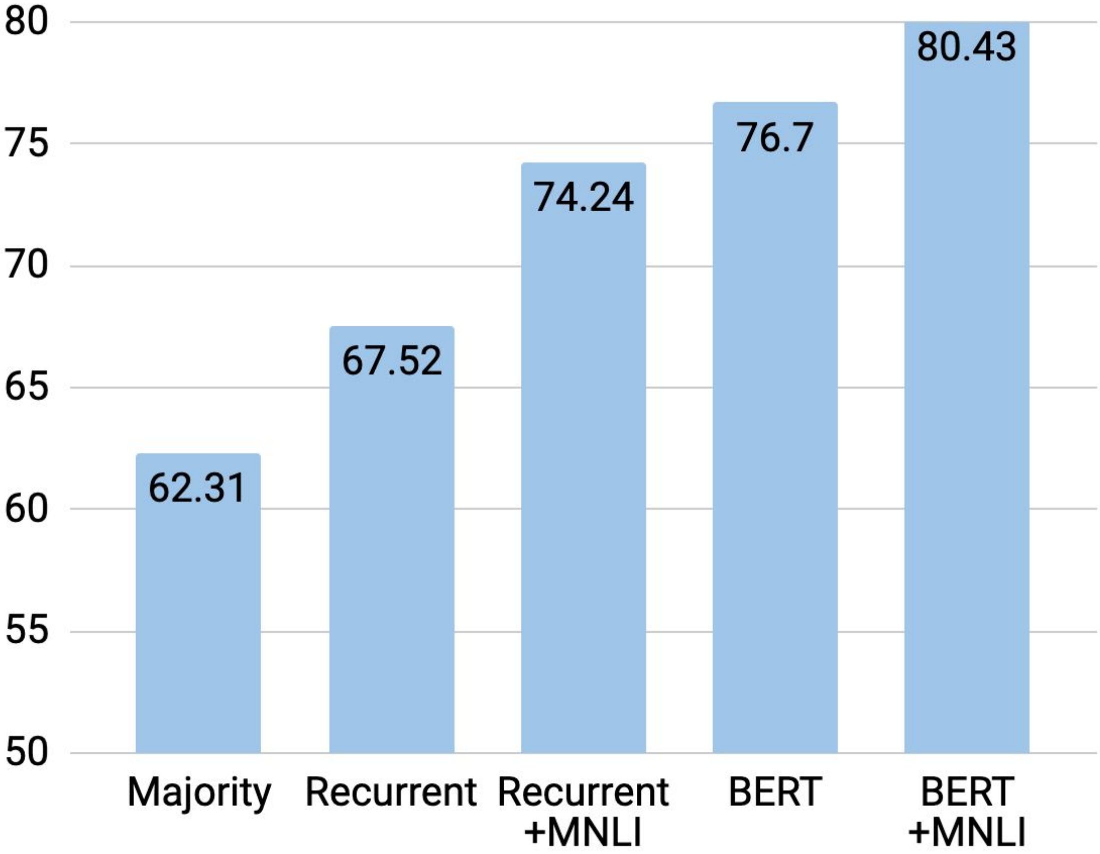
## Entailment


80
75
70
65
60
55
50
62.2
68.4
69.6
69.8
71.3
71.4
71.4
72.8
73.2
75.6
Majority
RACE
None
SQuAD 2.0
QQP
QNLI
Heuristic Y/N
ELMO
NQ
OpenAI GPT
SNLI
MultiNLI
BERT

# Unsupervised



# Test Set Results



# Thank You

Data: goo.gl/boolq

Will become part of the SuperGLUE benchmark (Wang et al., 2019)

- super.gluebenchmark.com