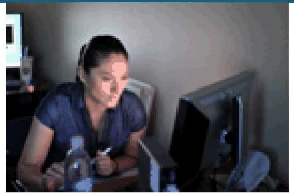
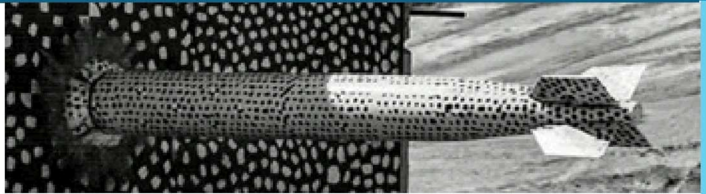
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2020-7700C

How the Presentation of Results from MLDL Models Cognitively Impacts Users





Presented By

Zoe Gastelum

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Research Team

Zoe Gastelum, PI,
Nuclear safeguards SME

Laura Matzen, Cognitive Scientist

Mallory Stites, Cognitive Scientist

Aaron Jones, Cognitive Scientist

Michael Trumbo, Cognitive Scientist

Breannan Howell, Graduate Student

Introduction

No machine learning or deep learning algorithms were hurt during the development of this research!

In our research, we examine how the errors from algorithms like MLDL impact users. To be able to experimentally manipulate our variables, we do not use any actual MLDL algorithms. However, our "MLDL results" are intended to look realistic for a deep object detection model.

We are interested in three aspects of MLDL errors:

1) Implementation of MLDL models, including visualization, transparency, confidence, and explainability;

2) Error frequency and response threshold;

3) Error type.

Motivation

Drivers:

1) In some application domains, MLDL performance now exceeds that of humans, but in other domains this remains a distant goal.

2) Newcomers to MLDL may rush into deploying models without understanding the impact on users.

Goal:

Assess cognitive impacts of MLDL errors on users to make recommendations that support the development of efficient human-and-algorithm systems.

Approach:

We introduce the research in a domain-general setting, with the intent to apply to domain-relevant datasets in Year 2.

How do errors impact user performance?

Implementation.

Error rate.

Error type.

A Deeper Look at Implementation

Implementation refers to how the MLDL model is presented to users. :

A. Visualization – the visual presentation of results of the models to users.

B. Transparency – the information provided to the user on the underlying architecture of the model, with what size and type of data it was trained, and its expected performance.

C. Confidence – the likelihood that potential targets selected by the model represent true positives.

D. Explainability – the identification of the portion of data that is most responsible for a given response.

Visualization Experimental Design

Visual search task: “find the perfect T”

Measured response time and accuracy

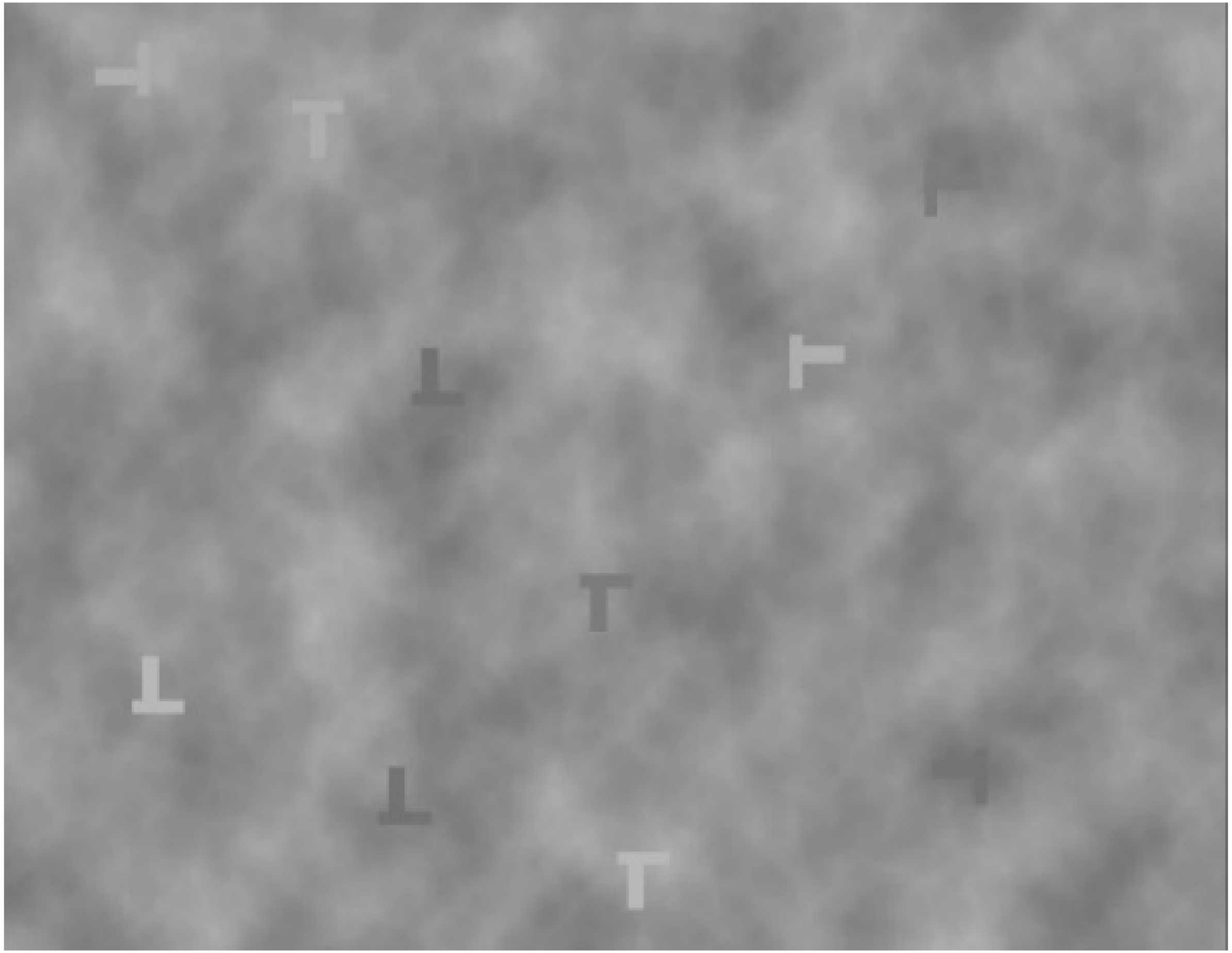
Counter-balanced six visualization types with seven MLDL response types:

MLDL Visualization
No aid
Bounding box
Heat map
Segmentation mask
Histogram
Text

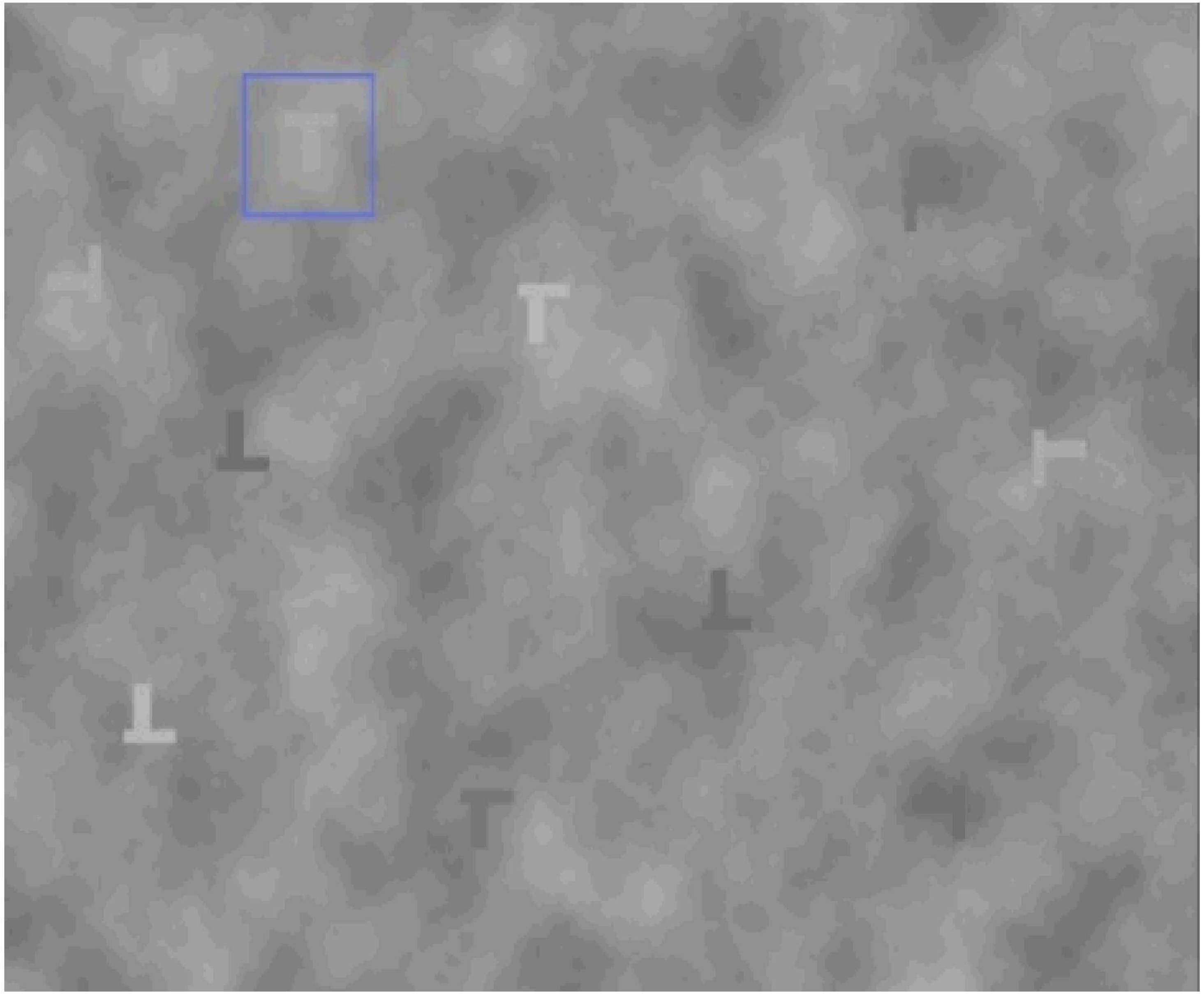
X

Response Types
True positive
True negative
False positive (two types)
False negative
False negative + false positive (two types)

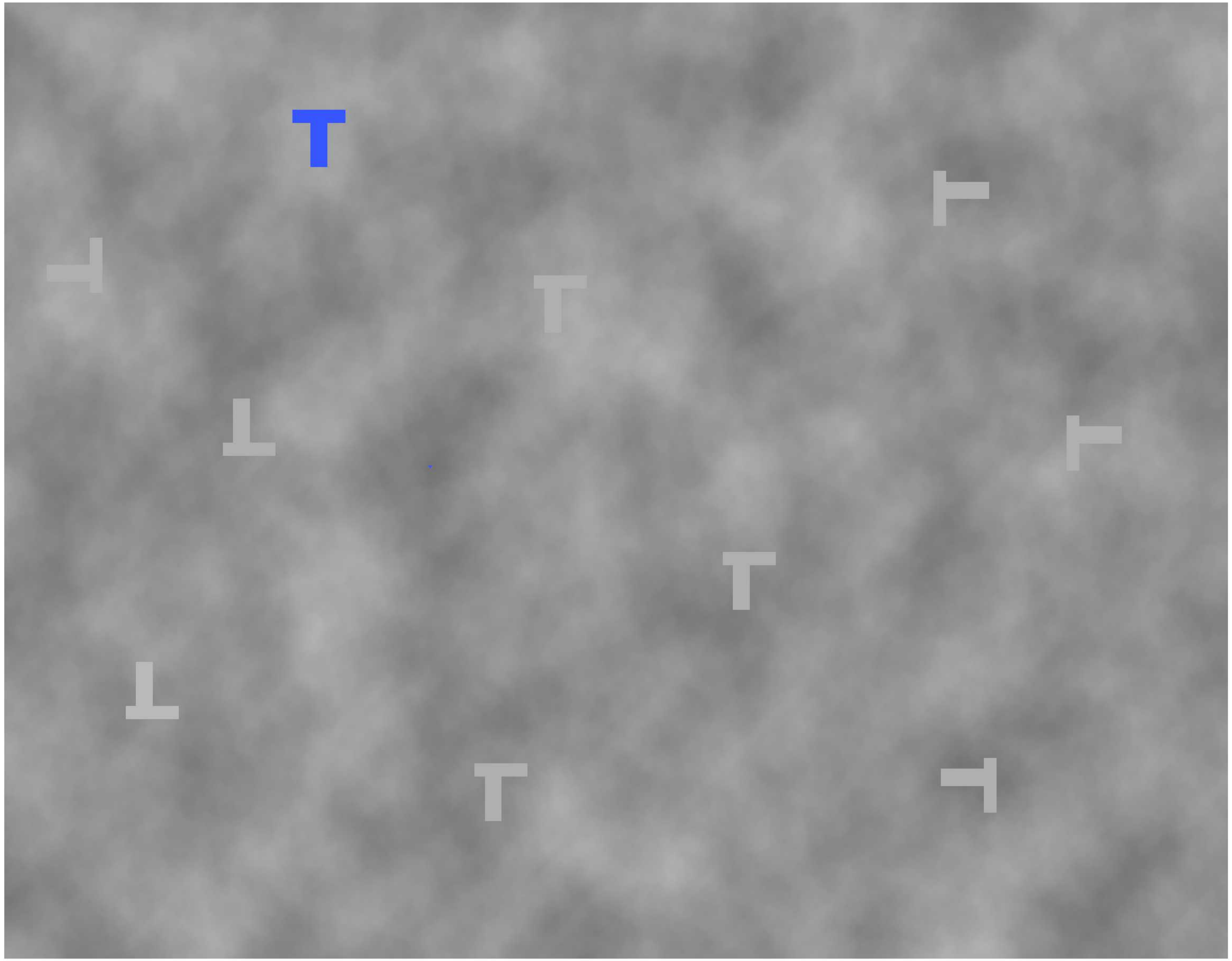
No Aid



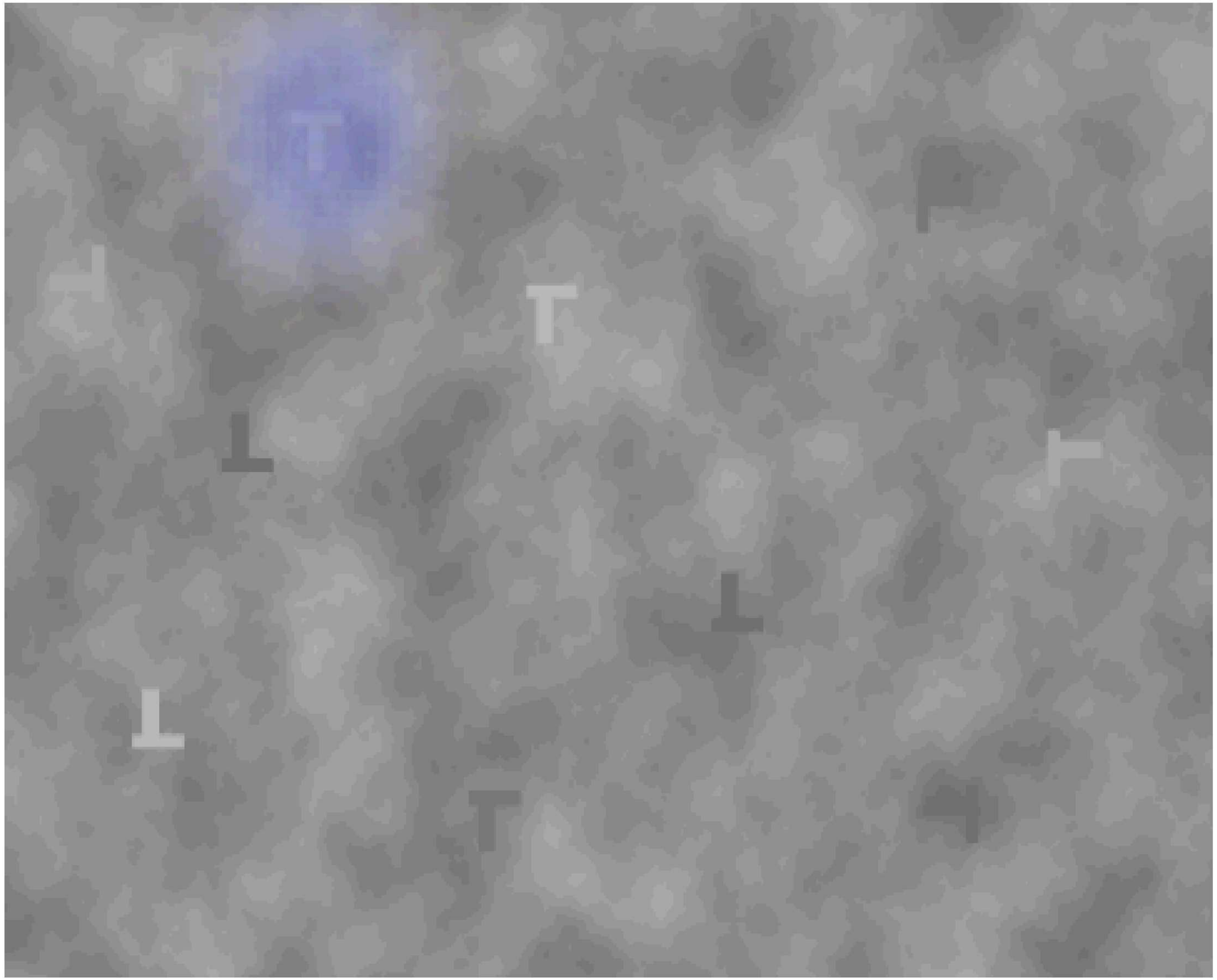
Bounding box



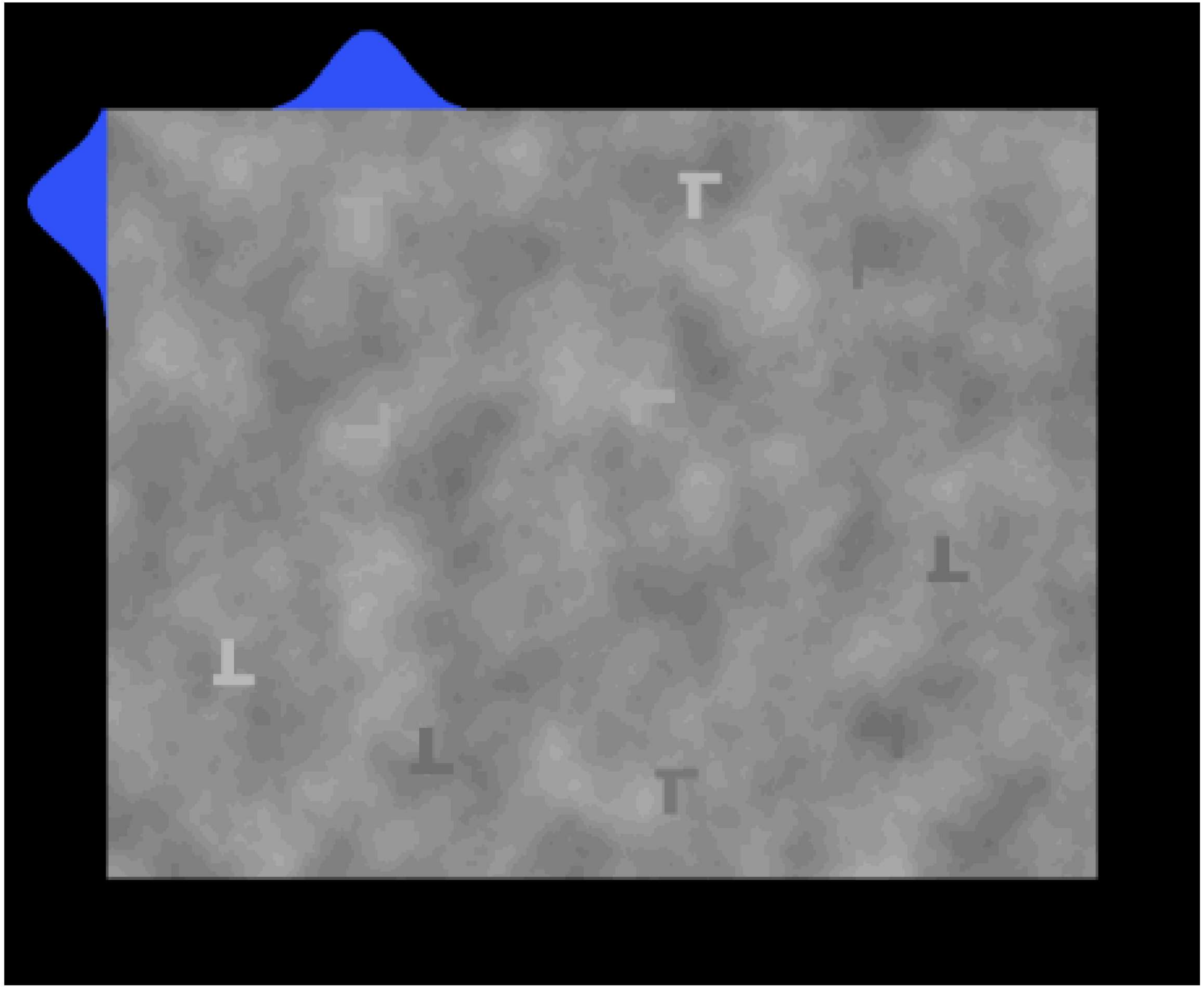
Mask



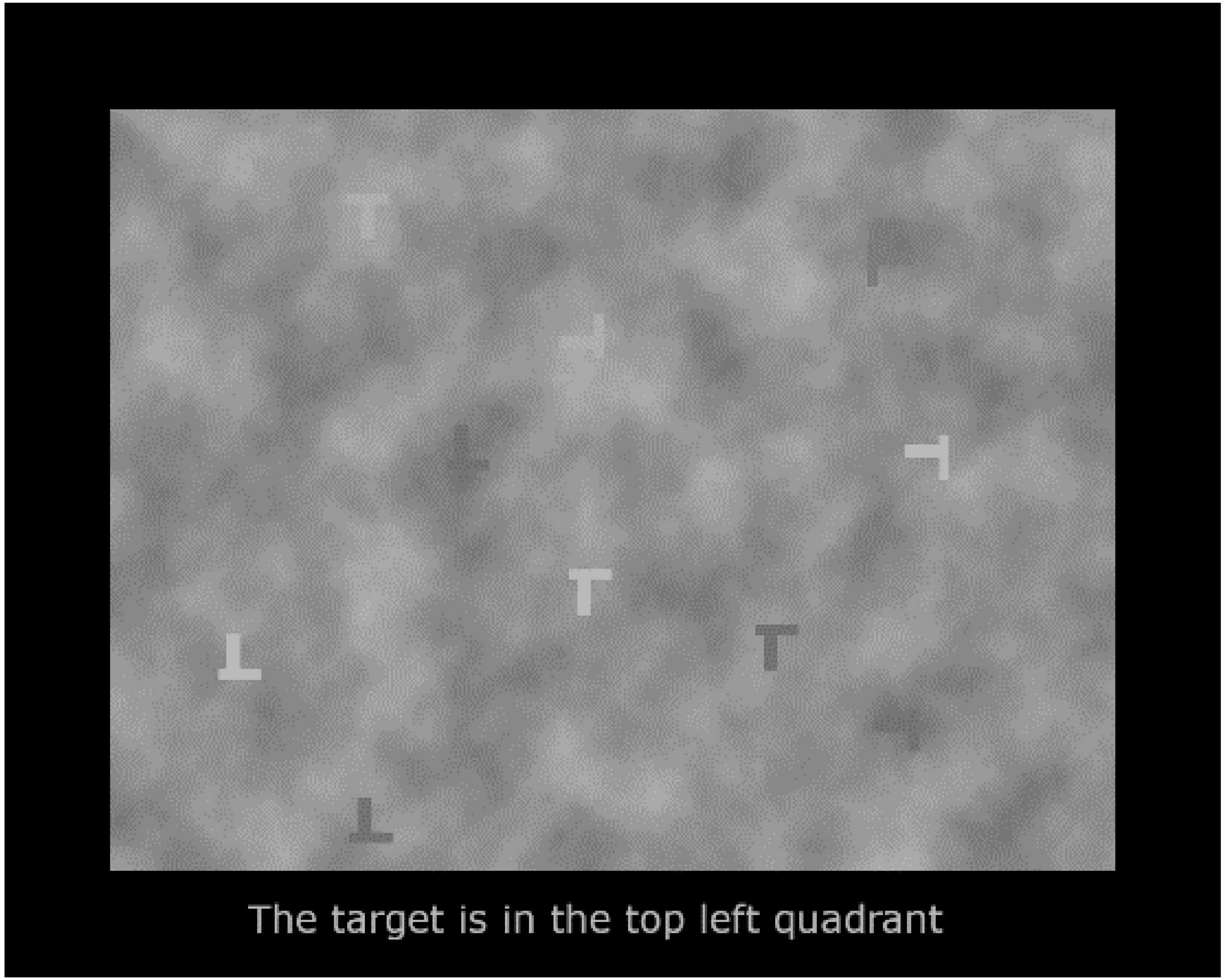
Heat map



Histogram

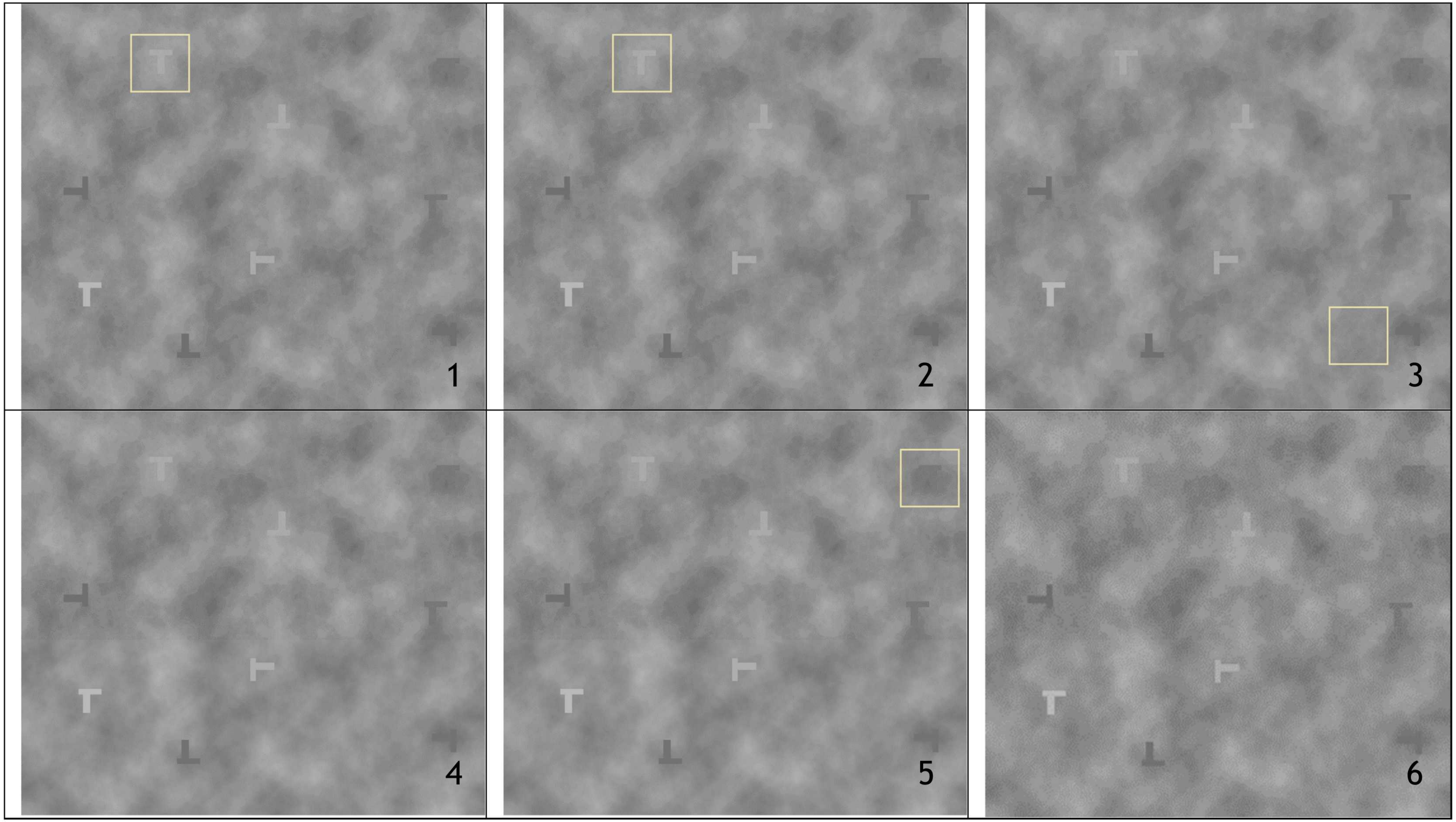


Text



The target is in the top left quadrant

A review of response types…



Results Up Front

Participants did better on identifying stimuli without targets.

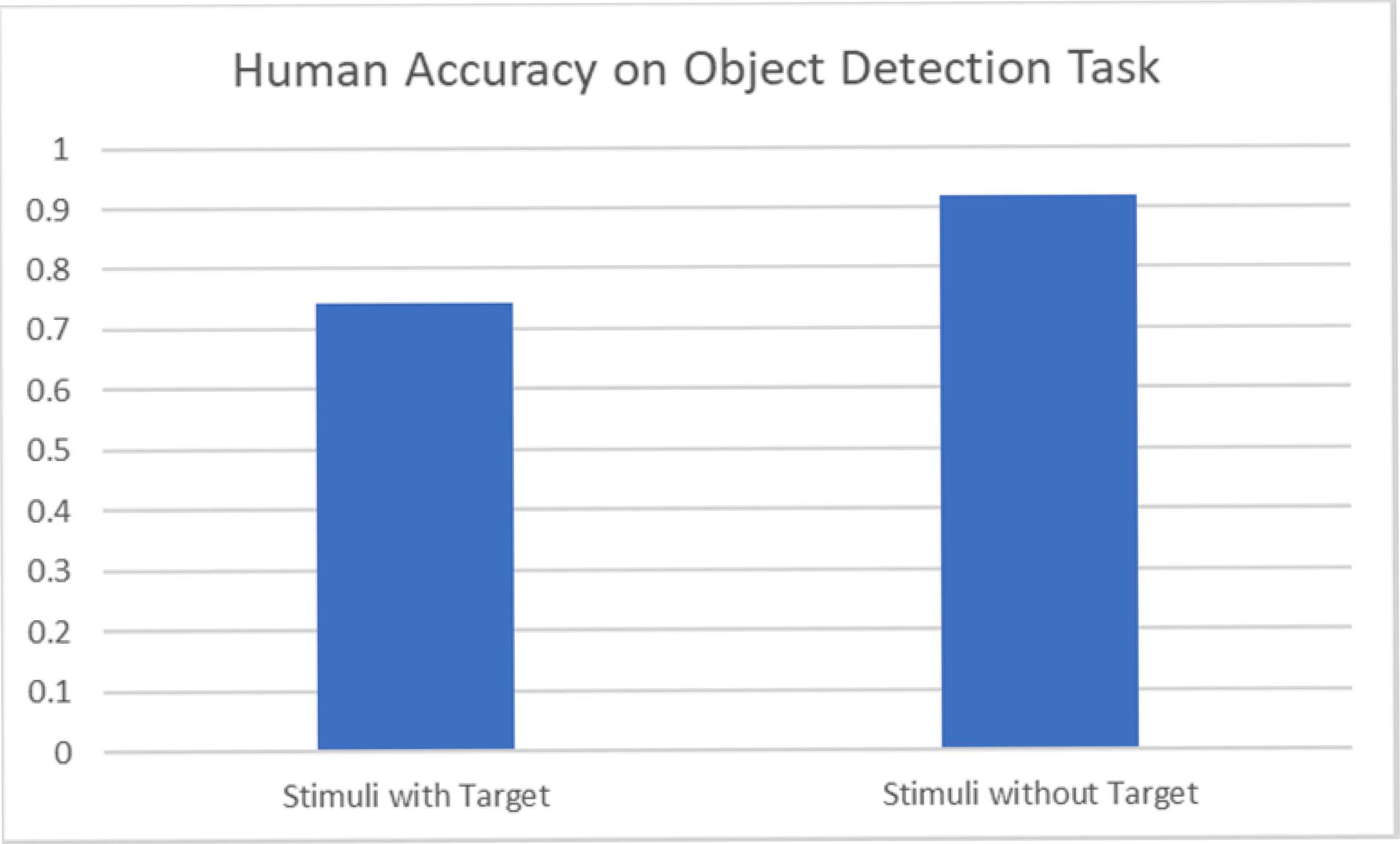
Participants performed the same or better with MLDL model visualizations than without, despite poor model performance.

Visualizations that were on the stimulus provided more benefit to accuracy and RT than those on the side.

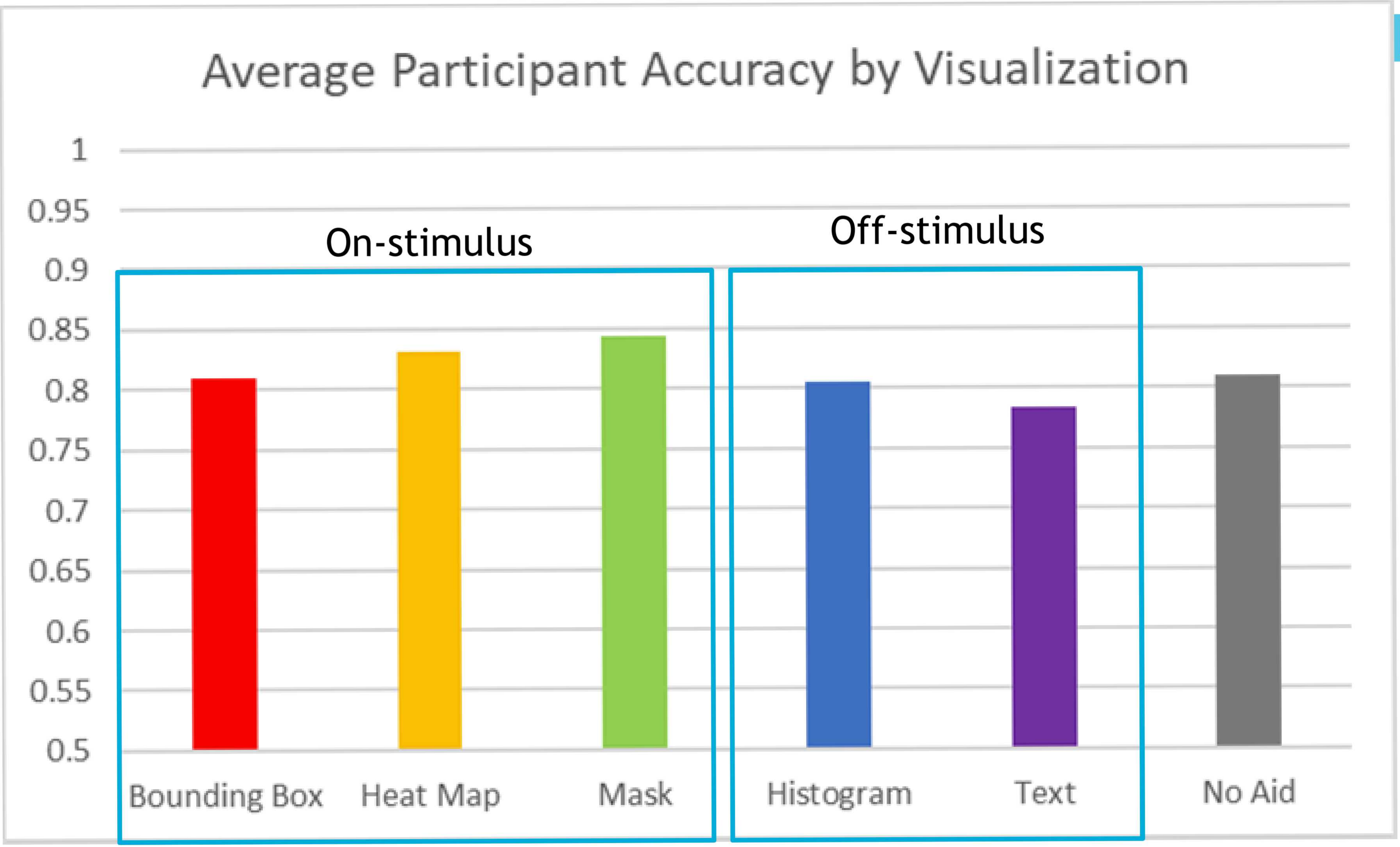
The exception to value added from MLDL visualization was the FN + FP error condition.

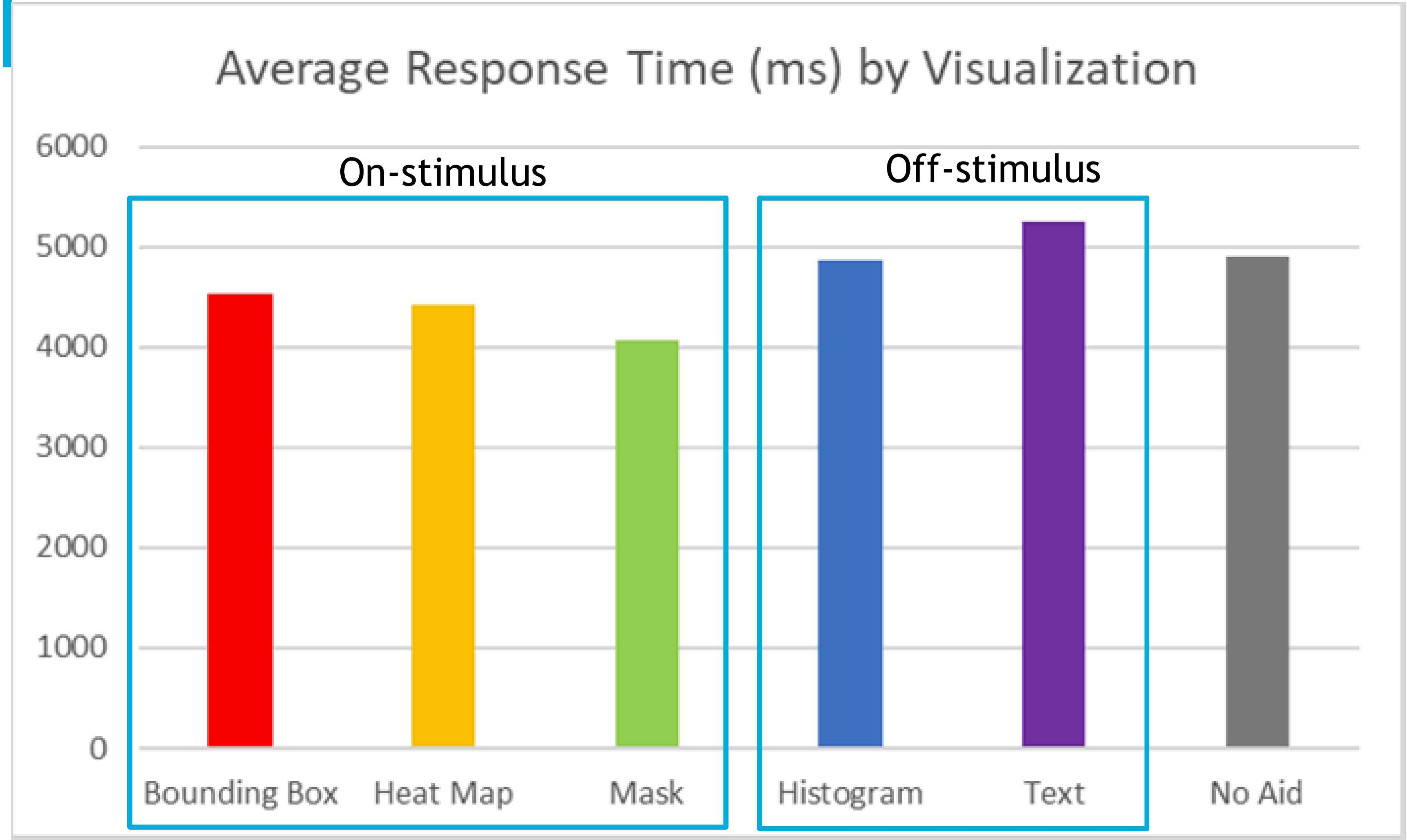
Object Detection Task Performance

Object Detection Task Performance



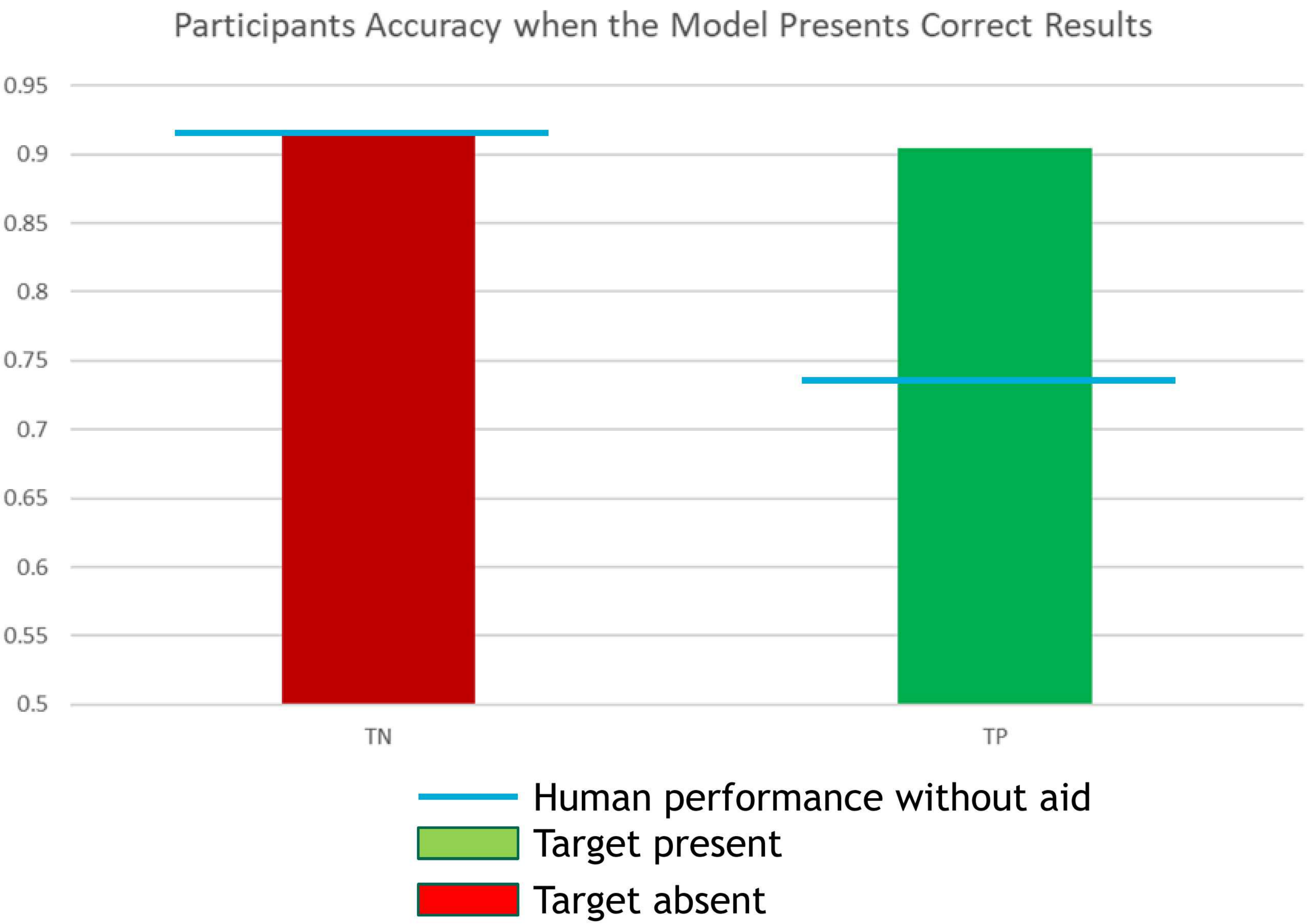
Visualization Results


Average Participant Accuracy by Visualization
On-stimulus
Off-stimulus
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Bounding Box
Heat Map
Mask
Histogram
Text
No Aid


Average Response Time (ms) by Visualization
6000
5000
4000
3000
2000
1000
0
On-stimulus
Off-stimulus
Bounding Box
Heat Map
Mask
Histogram
Text
No Aid

When the model is right….

Participant Accuracy for Correct Model Indicators

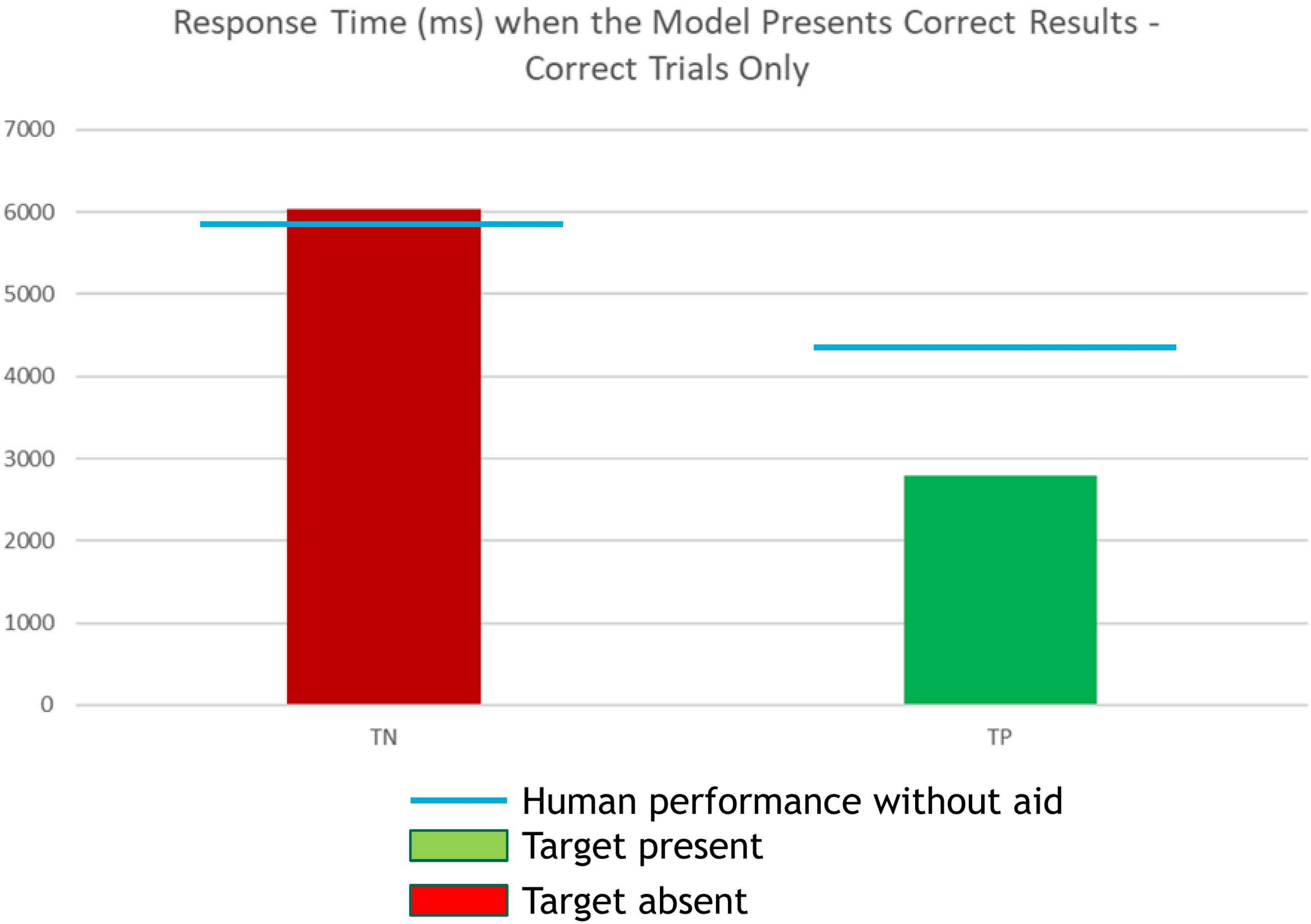


- TP: Accuracy increases from 74% to over 90%

- TN: Accuracy remains about the same at 91%.

- True negative presentation is an absence of indicator for all but text conditions, so similar performance on TN and no aid is expected

Response Times for Correct Model Indicators

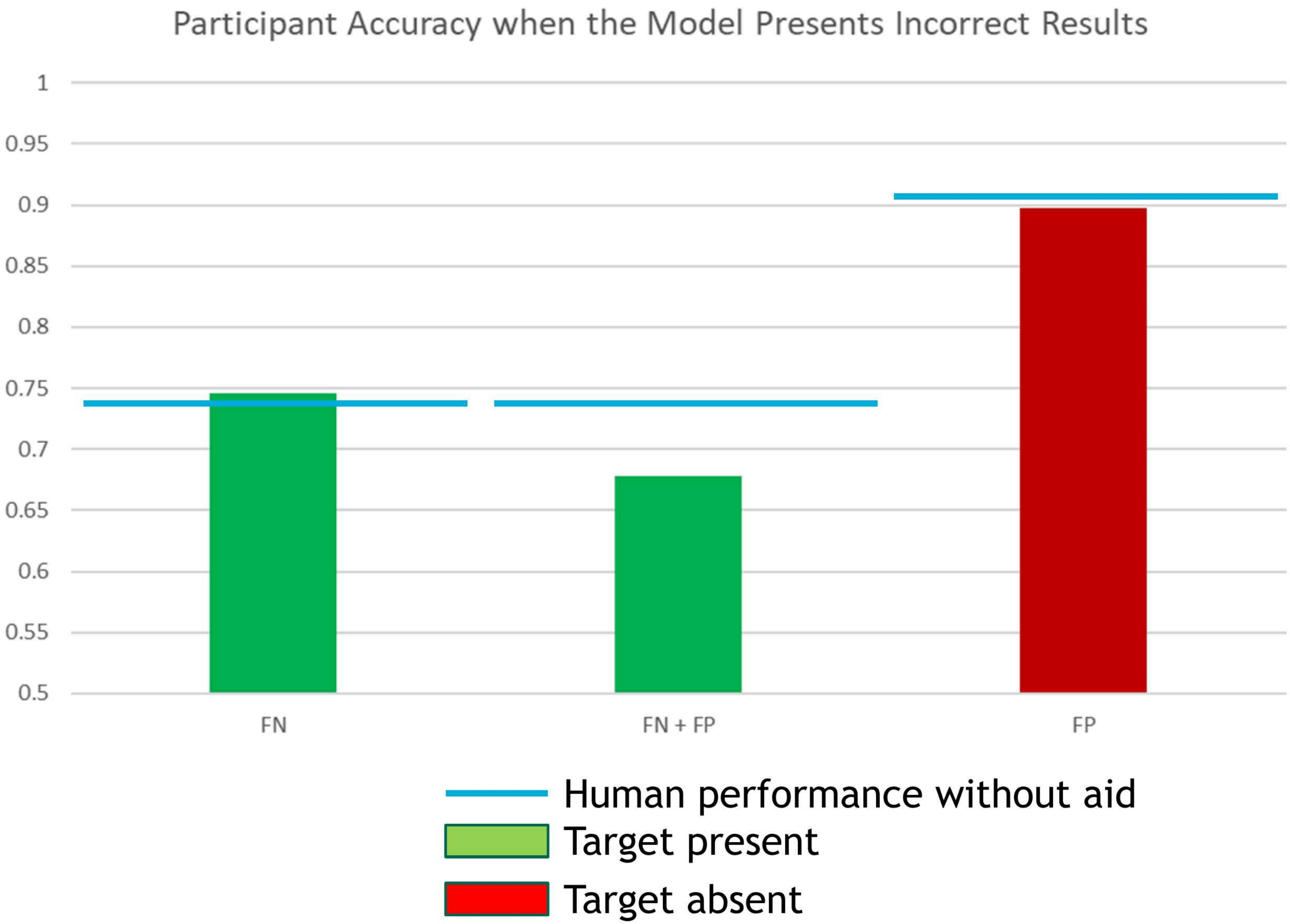


- TP: Response time drops from 4300 ms without and aid to less than 2900ms with an indicator.

- TN: Response time goes up slightly from 5850ms to 6000ms, a difference of 0.15 s.

When the model is wrong…

Human accuracy when the model is wrong

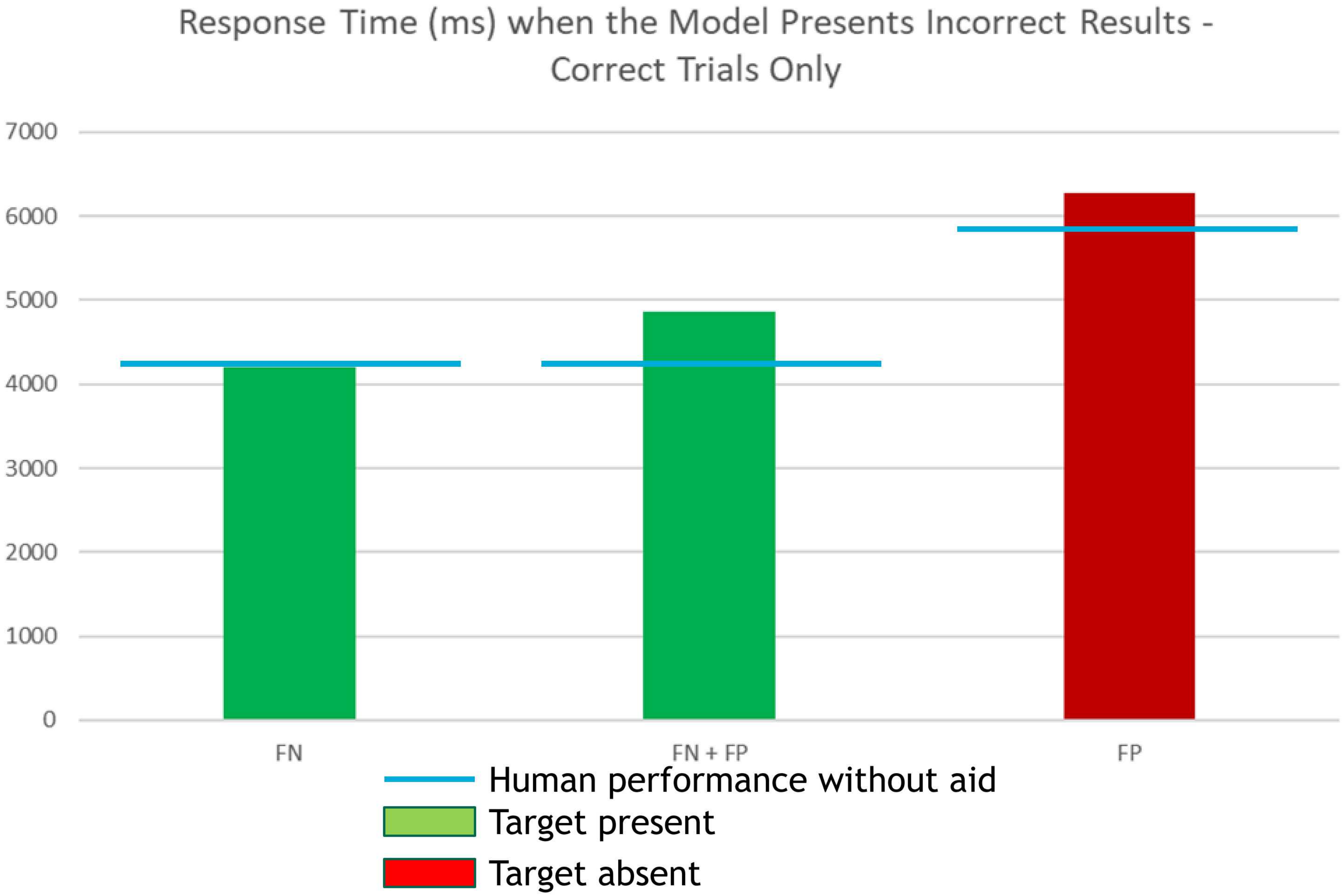


- FN: Accuracy remains about the same, up slightly from 74% no aid to 75%.

- FP: Results stay about the same as no aid to identify the absence of a target

- FN + FP: Performance decreases from the no aid condition, from 71% to about 68%.

Response time when the model is wrong



- FN: RT is about the same as no aid to find a target, which makes sense because most FN responses had no indicator.

FP: RT increases from 5850ms no aid to 6200 ms. This accounts for participant time to interpret the indicator and conduct their search.

- FN + FP: Response time is slower than no aid, increasing from 4300 ms to 4900 ms.

Discussion and Next Steps

Discussion

Participants did better on identifying stimuli without targets, than those with targets.

Participants performed the same or better with MLDL model visualizations than without, even with models with poor performance.

For the most part, visualizations that were on the stimulus provided more benefit to accuracy and RT than visualizations on the side.

The exception was a visualization with two types of errors: a FN + FP.

- We are planning a follow-up experiment to explicitly test the impact of error type.
- Prevalence of the FP + FN condition will vary by model and domain application.
- This condition may be a factor of model threshold for response. We will look at thresholds in later research.

Next Steps

- Implementation
  - Confidence - probability of target for each predicated target
  - Transparency (domain-relevant) how the model works
  - Explainability (domain-relevant) - information used to inform model predictions
- Error rate and threshold
  - Rate - the frequency of MLDL errors
  - Threshold - a combination of the performance of the model, requirements of the domain, and the specific users. We will examine ratios of false negatives : false positives that users are willing to tolerate.
- Error type (domain general and domain-specific)
  - type of error presented to the user, including obvious false positives and false negatives. We will include subtle errors that could potentially indicate bias or corruption of the model, and compare them to more explicit or obvious error types.

Acknowledgements

This work was funded by Sandia National Laboratories Laboratory-Directed Research & Development program's Computing and Information Sciences Investment Area, under LDRD project #218306.

Zoe Gastelum
zgastel@sandia.gov
(505) 401-6959