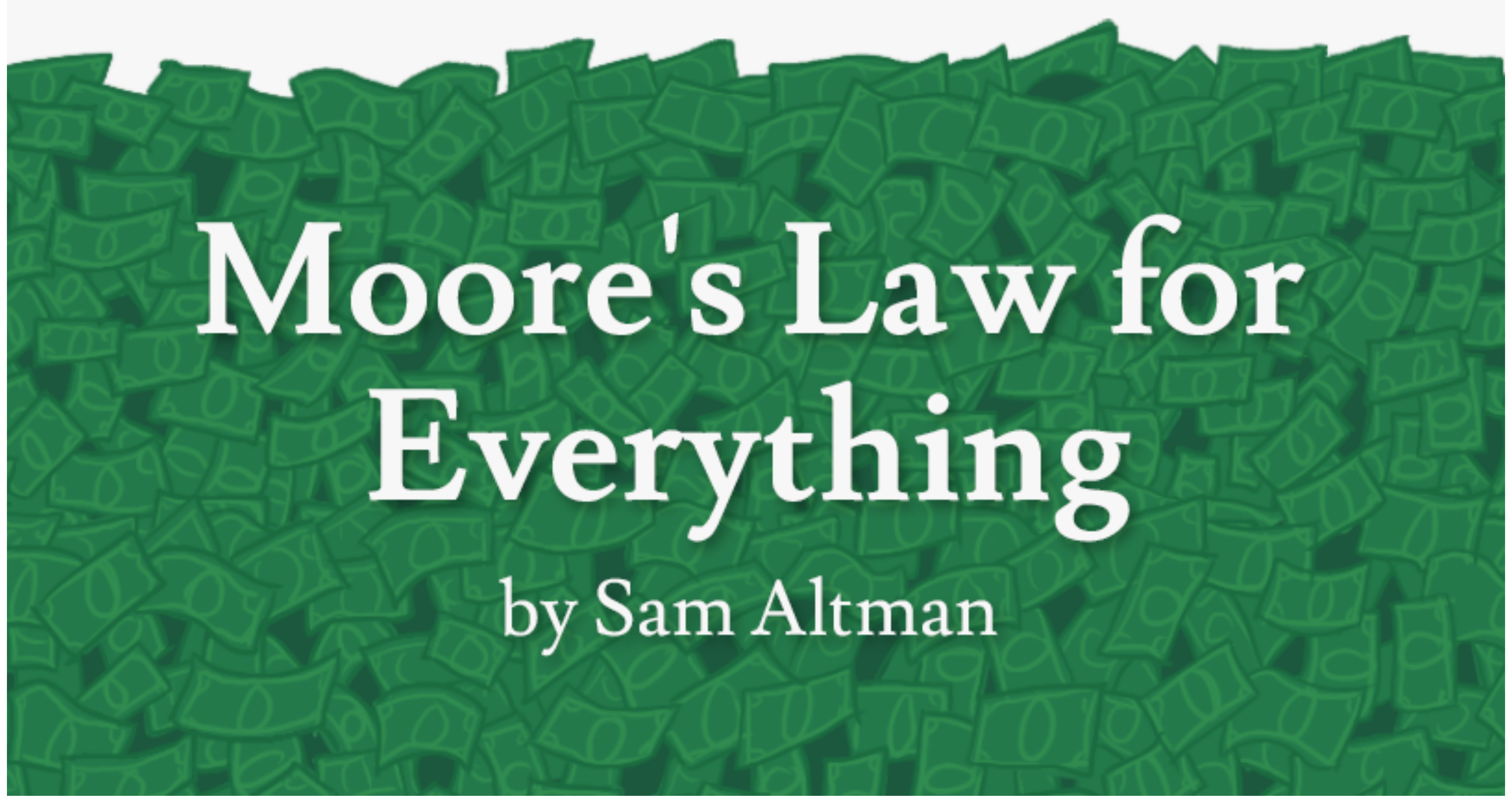
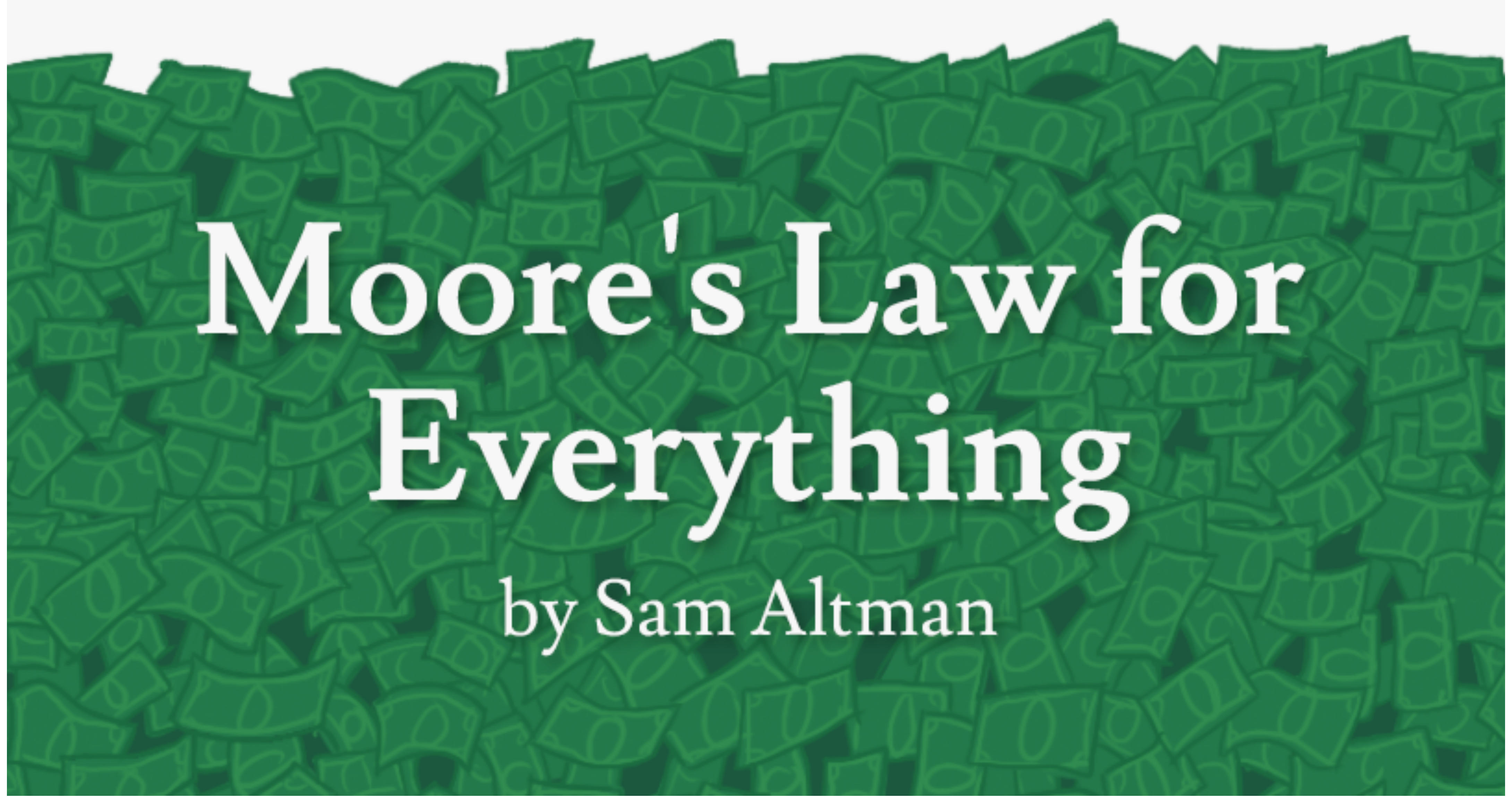
The Devil's in the Data

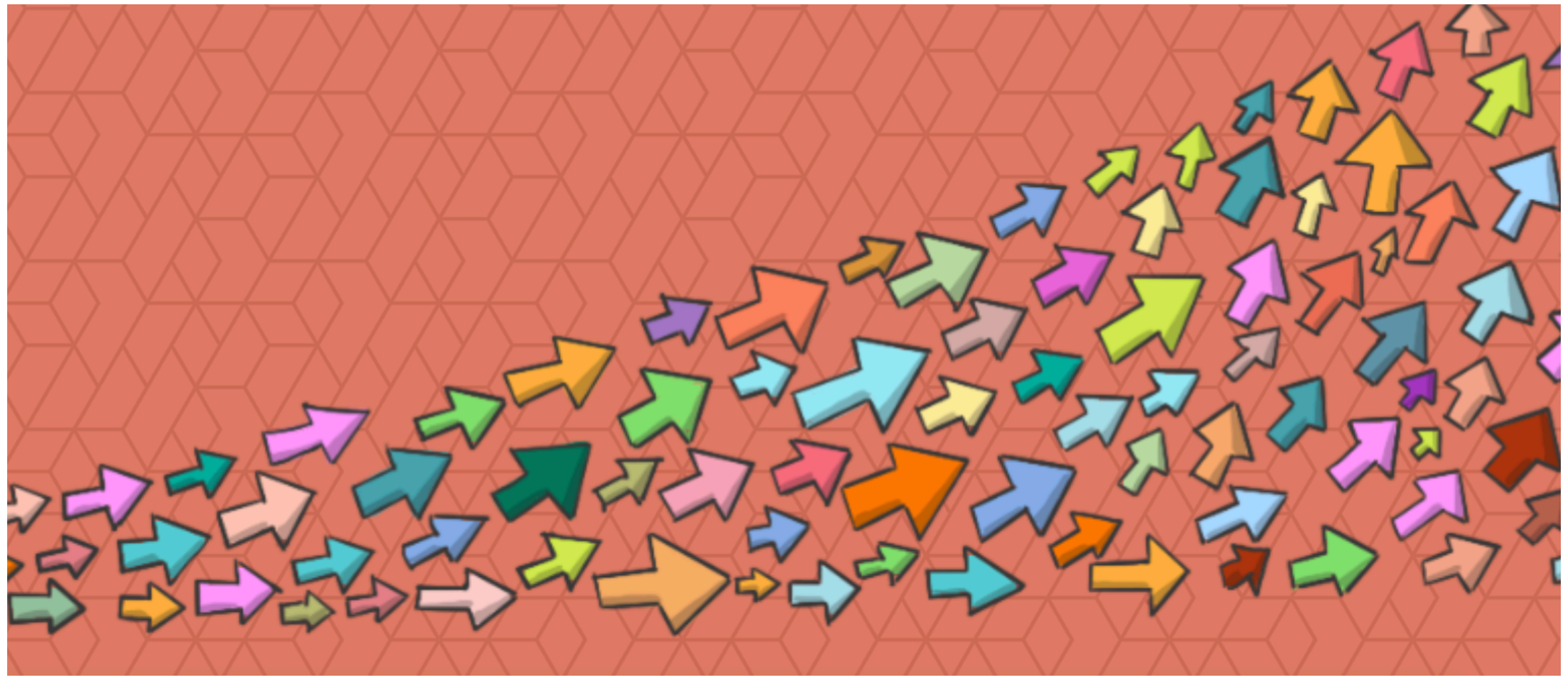
Mapping and Generating Datasets for Robust Generalization

Swabha Swayamdipta
Incoming Asst. Prof, USC CS
Postdoc, Allen Institute for AI
23rd May, 2022

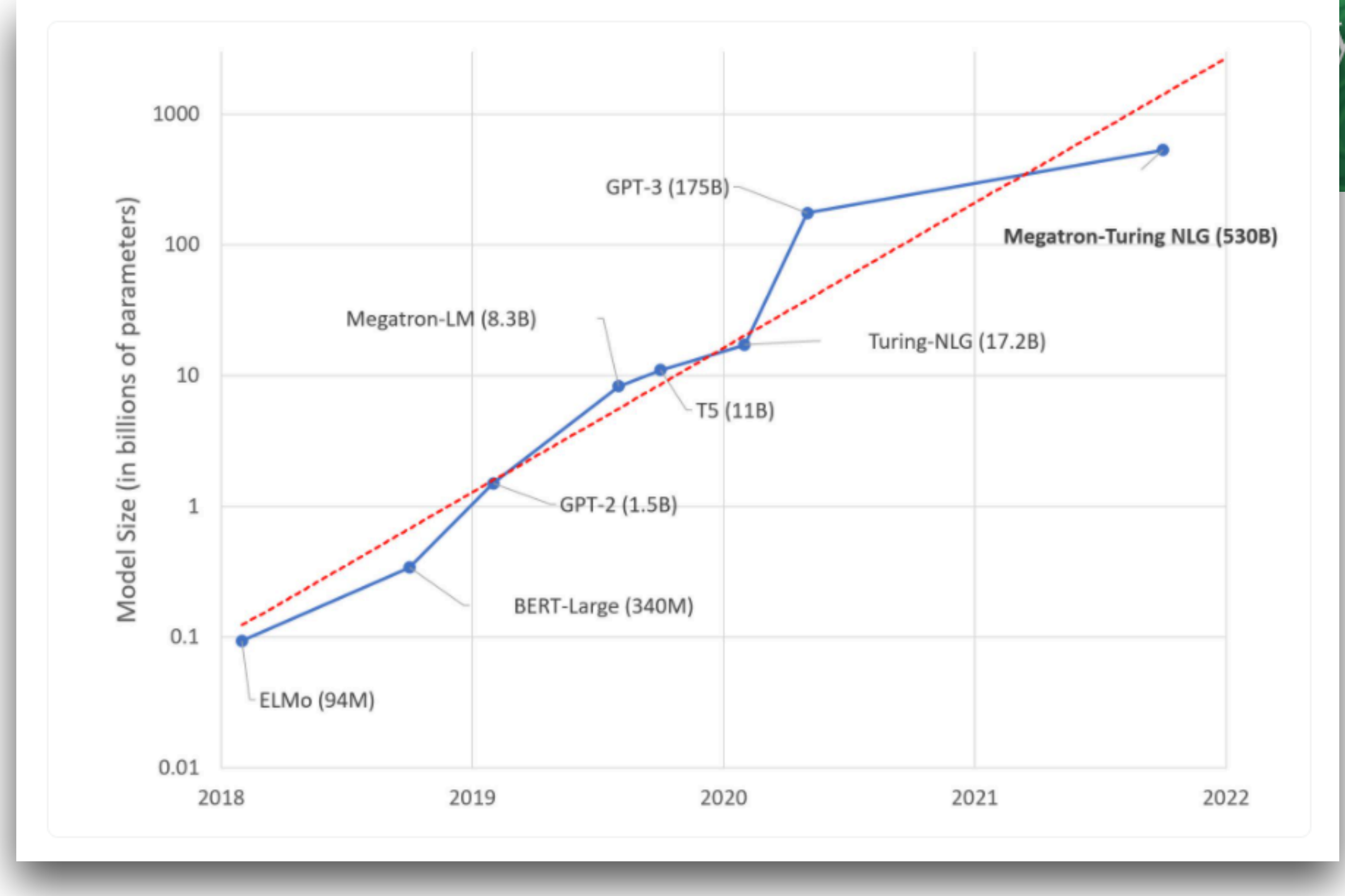


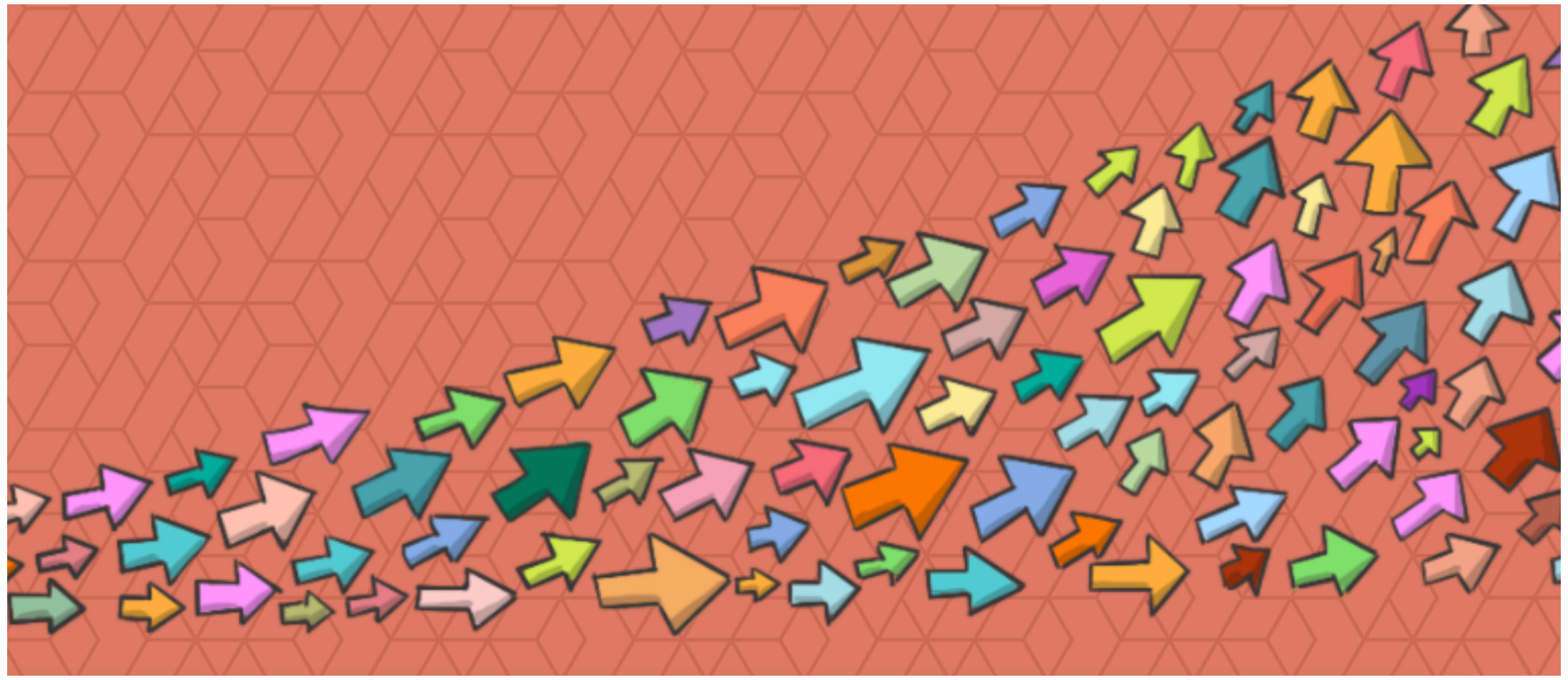
https://moores.samaltman.com/





https://moores.samaltman.com/





c.f. Julien Simon's blog https://huggingface.co/blog/large-language-models

https://moores.samaltman.com/

c.f. Julien Simon's blog https://huggingface.co/blog/large-language-models

https://moores.samaltman.com/

Is data scale really the key to generalization?

c.f. Julien Simon’s blog https://huggingface.co/blog/large-language-models

https://moores.samaltman.com/

Natural Language Inference

Given a premise, is a hypothesis true, false or neither?

Semantic Theory [Katz, 1972]

Natural Language Inference

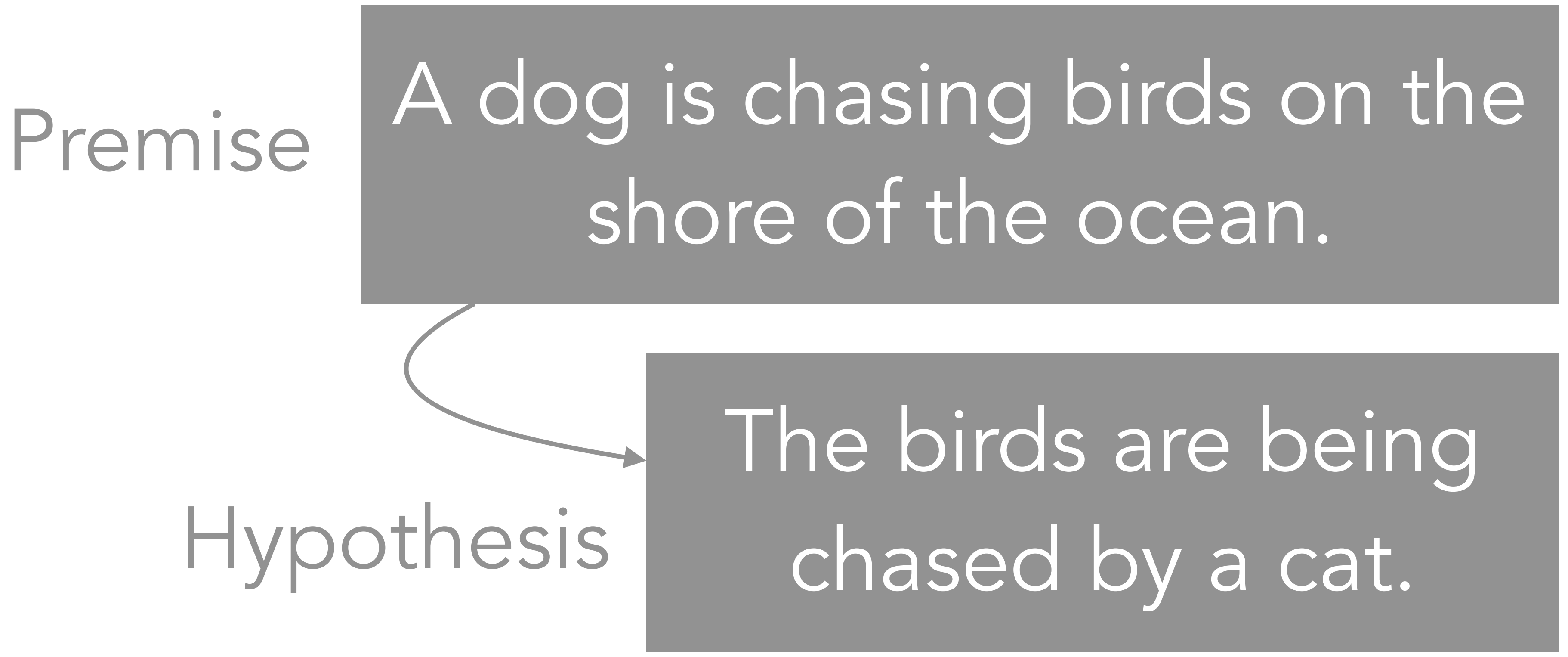
Given a premise, is a hypothesis true, false or neither?

Premise: A dog is chasing birds on the shore of the ocean.

Semantic Theory [Katz, 1972]

Natural Language Inference

Given a premise, is a hypothesis true, false or neither?



Premise

A dog is chasing birds on the shore of the ocean.

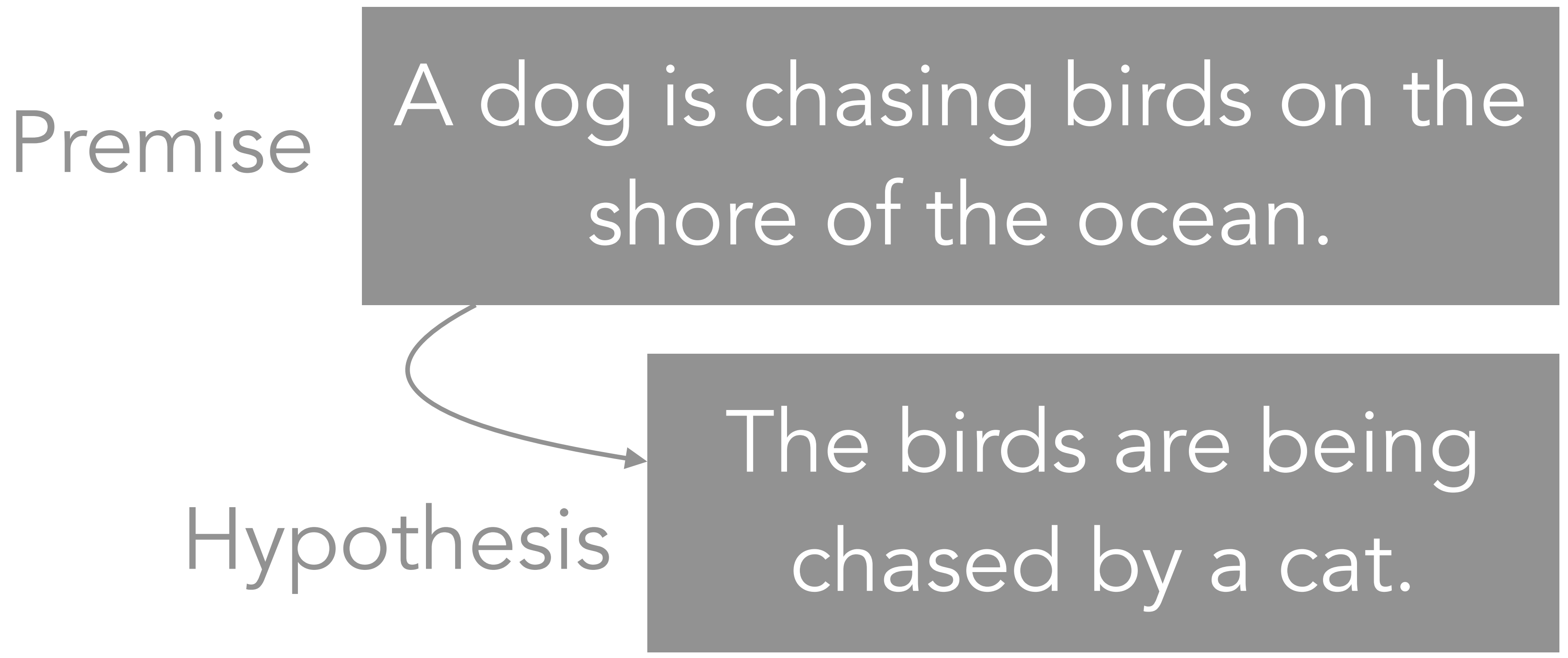
Hypothesis

The birds are being chased by a cat.

Semantic Theory [Katz, 1972]

Natural Language Inference

Given a premise, is a hypothesis true, false or neither?



Premise

A dog is chasing birds on the shore of the ocean.

Hypothesis

The birds are being chased by a cat.

○ True → **Entailment**

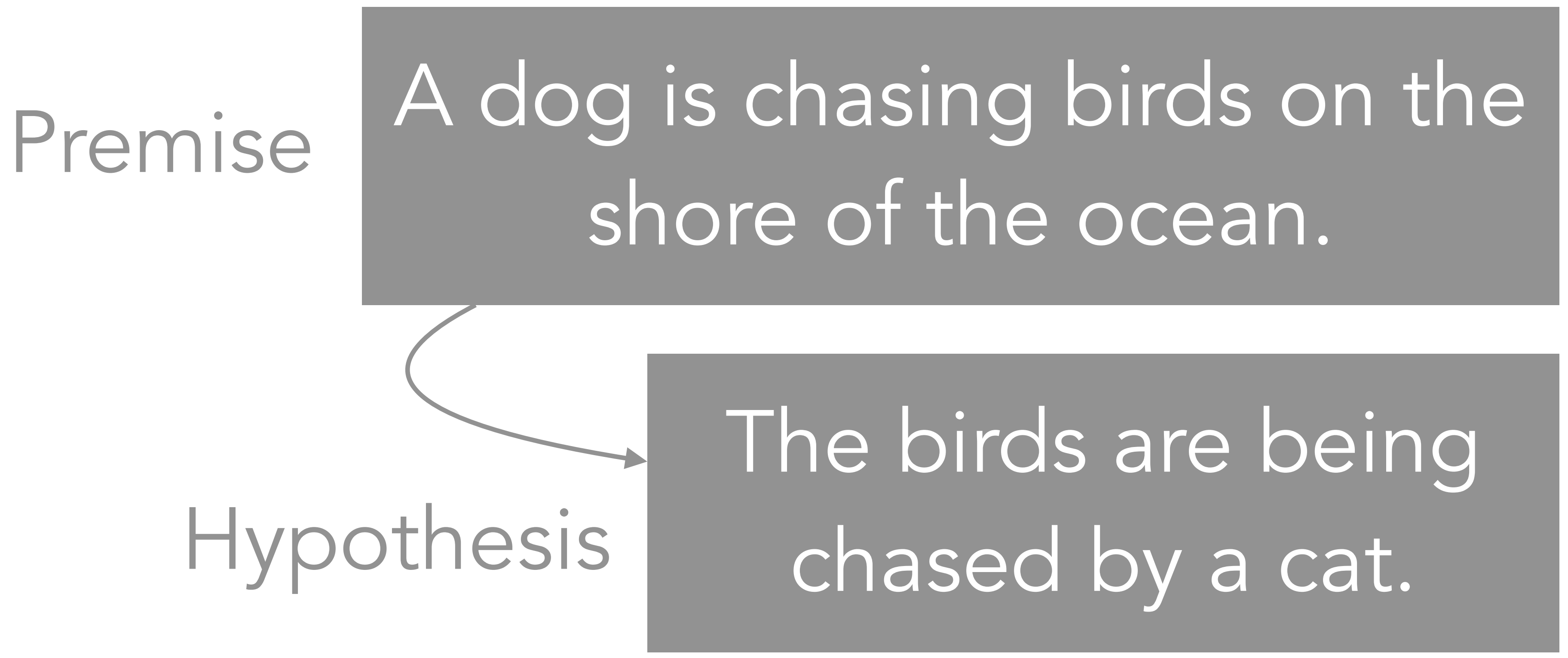
○ False → **Contradiction**

○ Cannot Say → **Neutral**

Semantic Theory [Katz, 1972]

Natural Language Inference

Given a premise, is a hypothesis true, false or neither?



Premise

A dog is chasing birds on the shore of the ocean.

Hypothesis

The birds are being chased by a cat.

○ True → **Entailment**

[✓ False] → **Contradiction**

○ Cannot Say → **Neutral**

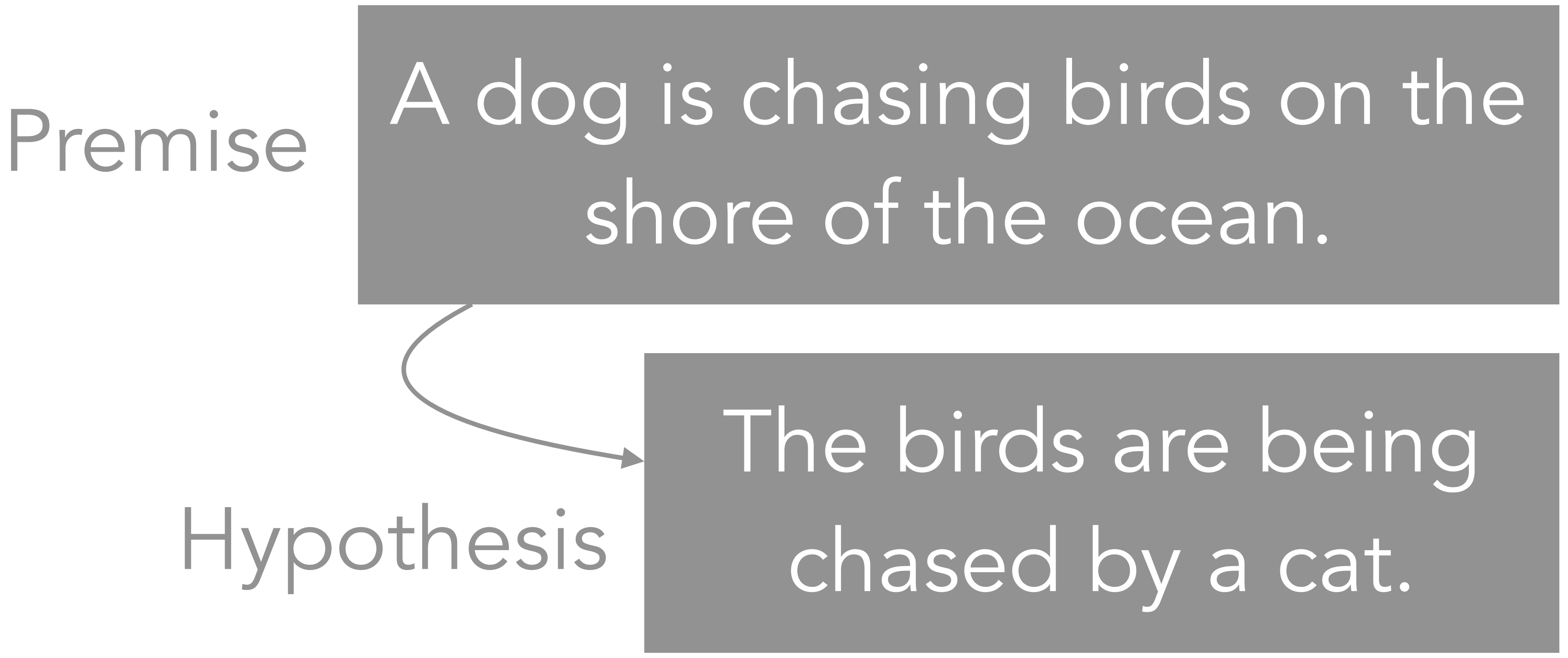
Semantic Theory [Katz, 1972]

Natural Language Inference

Stanford NLI [Bowman et al., 2015]
~0.5m instances

MultiNLI [Williams et al., 2018]
~0.4m instances

Given a premise, is a hypothesis true, false or neither?



Premise: A dog is chasing birds on the shore of the ocean.

Hypothesis: The birds are being chased by a cat.

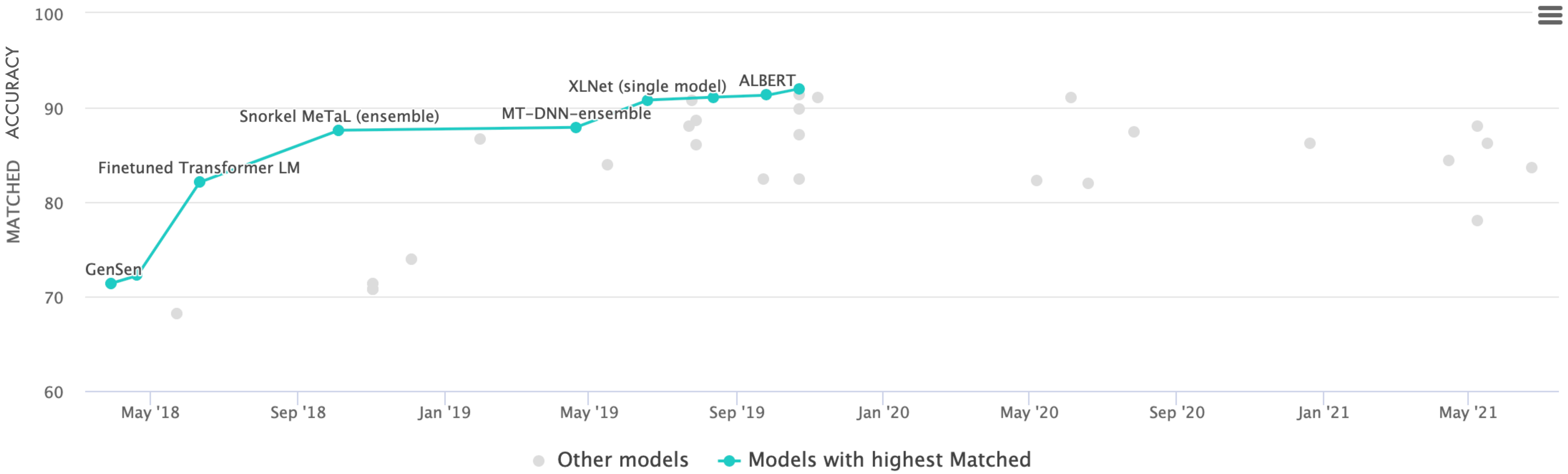
○ True → **Entailment**

[✓ False] → **Contradiction**

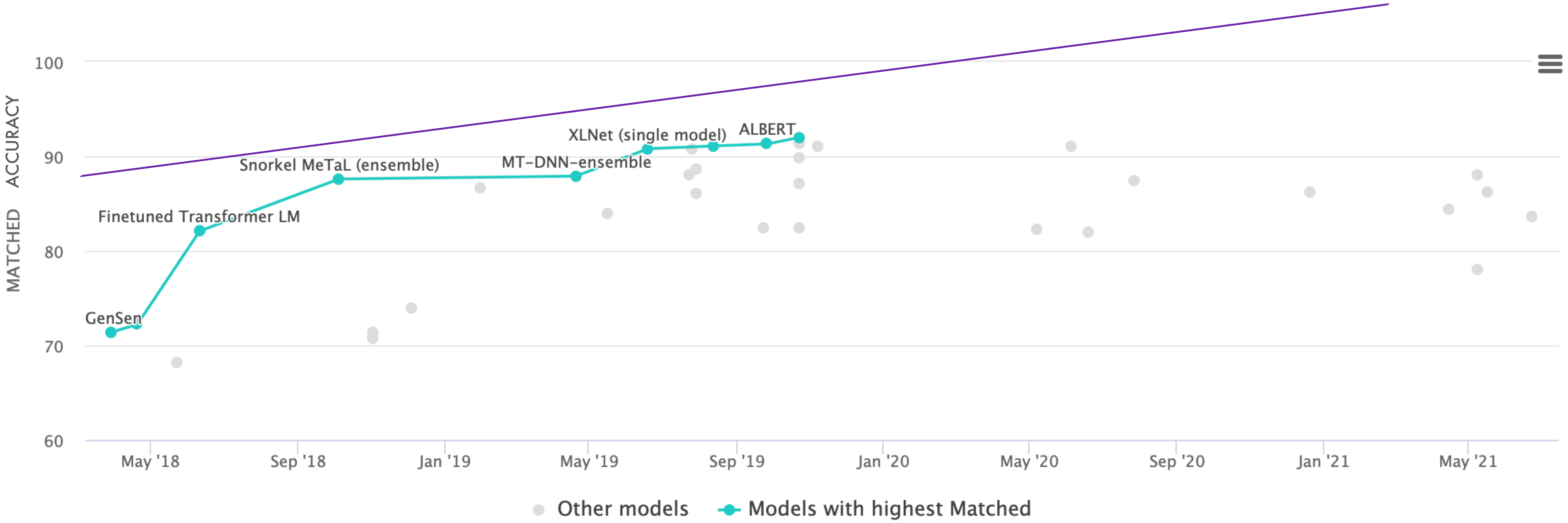
○ Cannot Say → **Neutral**

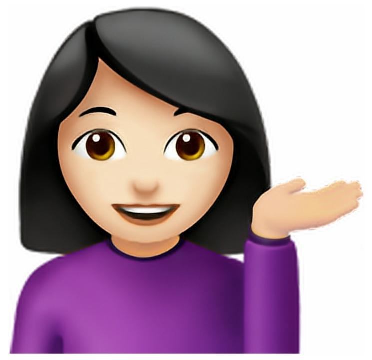
Semantic Theory [Katz, 1972]

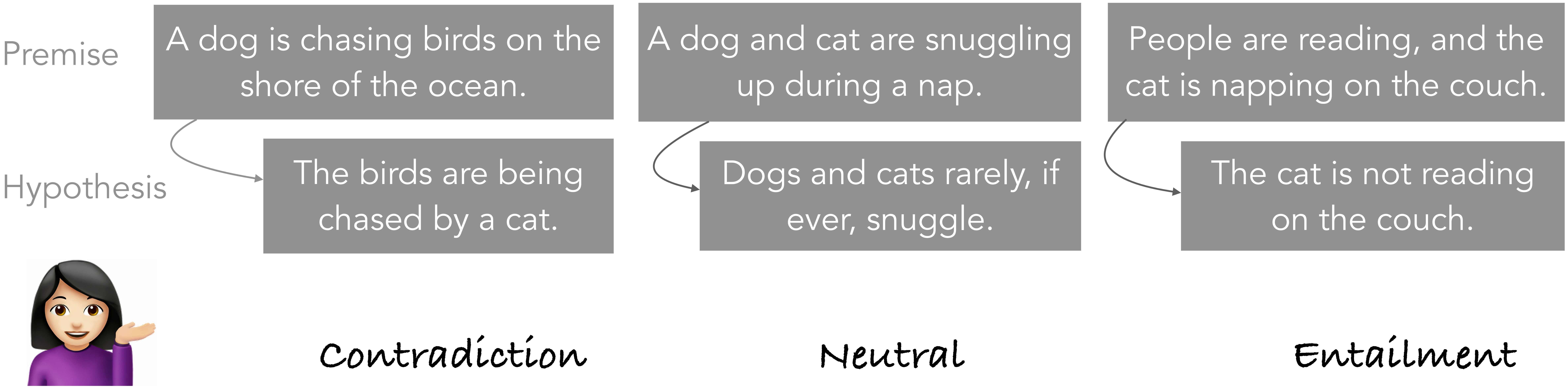
MultiNLI leaderboard results from paperswithcode.com [March 2022]



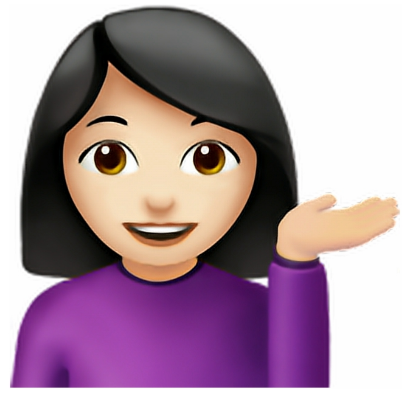
MultiNLI leaderboard results from paperswithcode.com [March 2022]







Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	Contradiction	Neutral	Entailment

Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	Contradiction	Neutral	Entailment

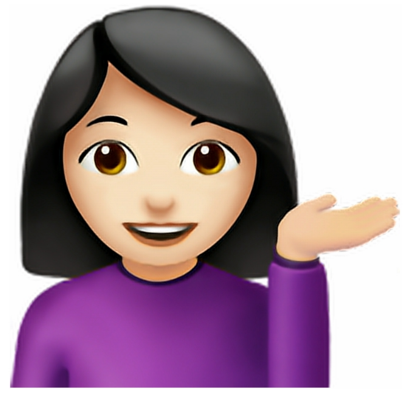
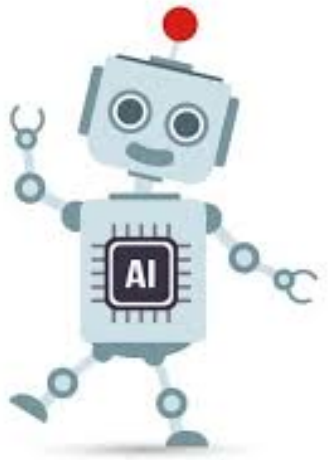
RoBERTa-Large [Liu et al. 2019]

Trained on MultiNLI + SNLI

Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
💁‍♀️	Contradiction	Neutral	Entailment
🤖	**Contradiction**	**Contradiction**	**Contradiction**

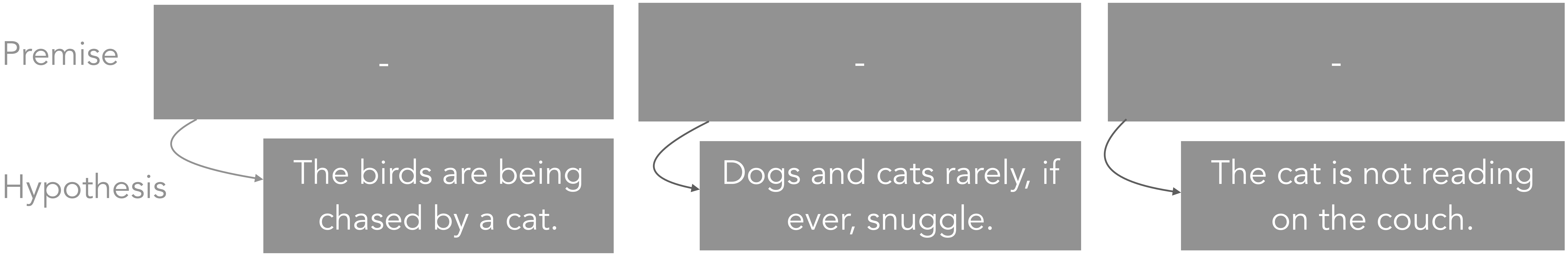
RoBERTa-Large [Liu et al. 2019]

Trained on MultiNLI + SNLI

Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	Contradiction	Neutral	Entailment
	Contradiction ✓	Contradiction ✗	Contradiction ✗

RoBERTa-Large [Liu et al. 2019]

Trained on MultiNLI + SNLI



Premise

-

-

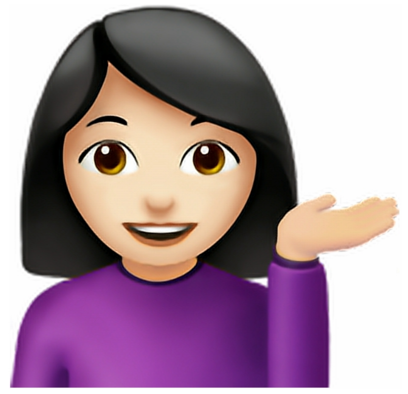
-

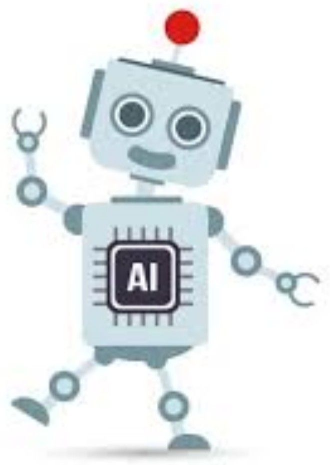
Hypothesis

The birds are being chased by a cat.

Dogs and cats rarely, if ever, snuggle.

The cat is not reading on the couch.

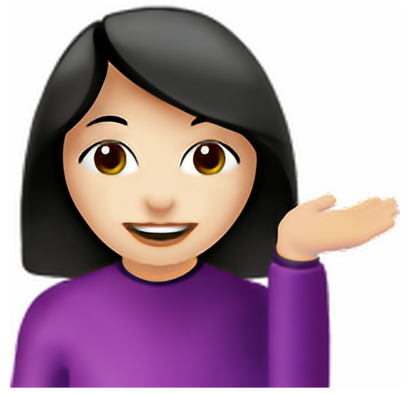


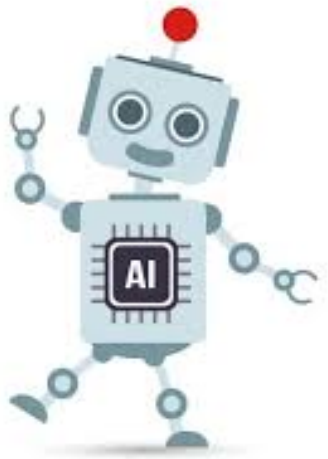


RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]

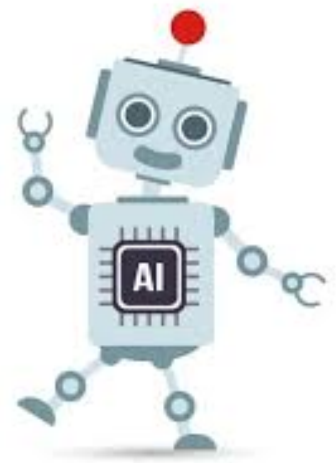
Premise	-	-	-
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	??	??	??



RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]

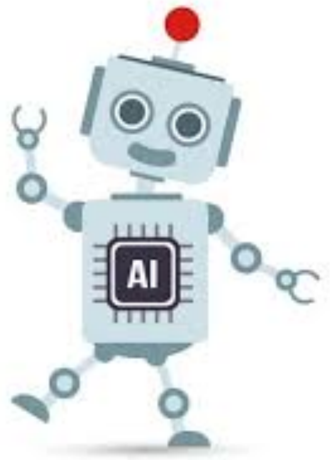
Premise	-	-	-
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	??	??	??
	**Contradiction**	**Contradiction**	**Contradiction**

RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]

Premise	-	-	-
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.



Contradiction

Contradiction

Contradiction

RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]

Premise

-

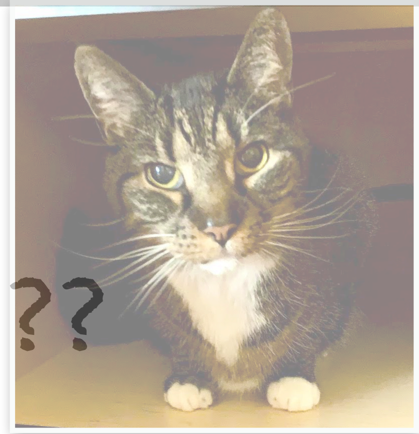
-

Hypothesis

The birds are being chased by a cat.

rely, if

The cat is not reading on the couch.



Contradiction

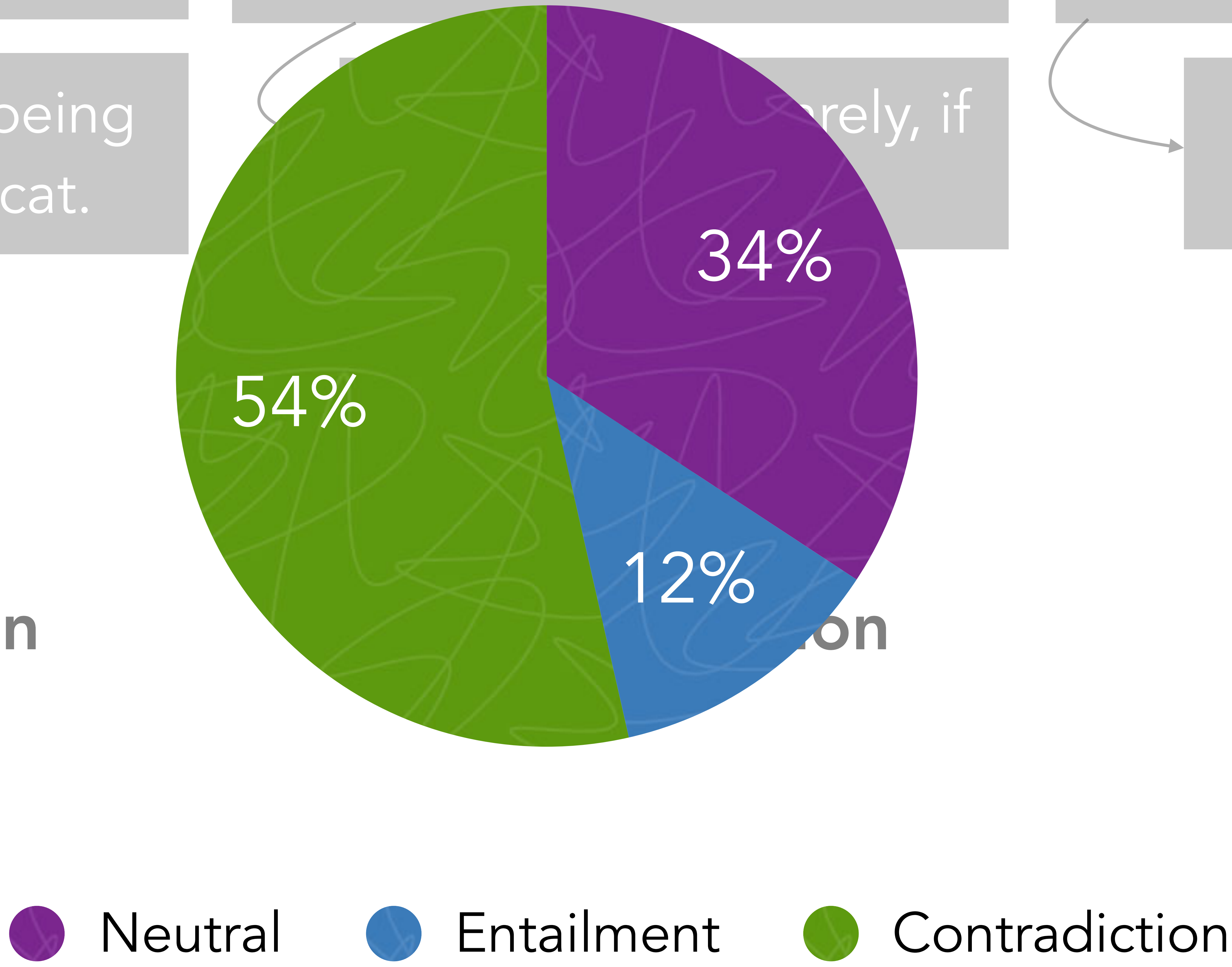
on

Contradiction

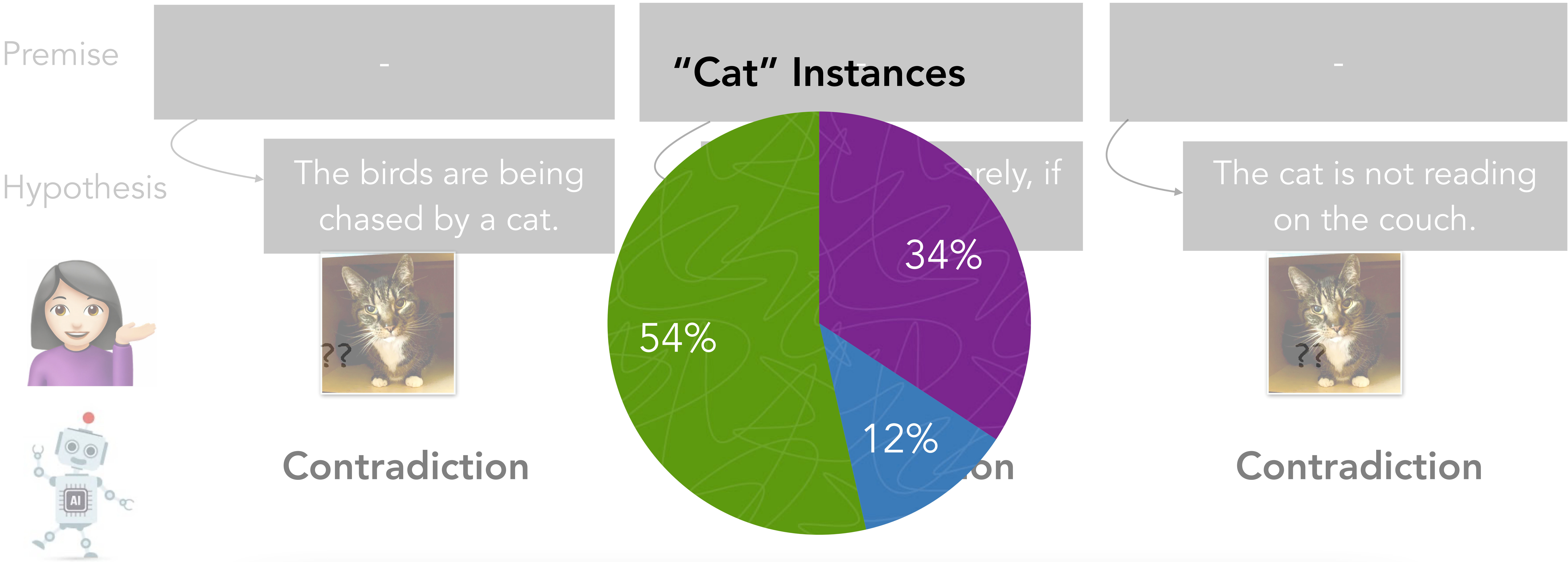
RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

“Cat” Instances



Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]



Premise

-

-

Hypothesis

The birds are being chased by a cat.

rely, if

The cat is not reading on the couch.

Contradiction

on

Contradiction

RoBERTa-Large [Liu et al. 2019]

Trained on SNLI + MultiNLI

State-of-the-art NLP models still succumb to **spurious biases** in data

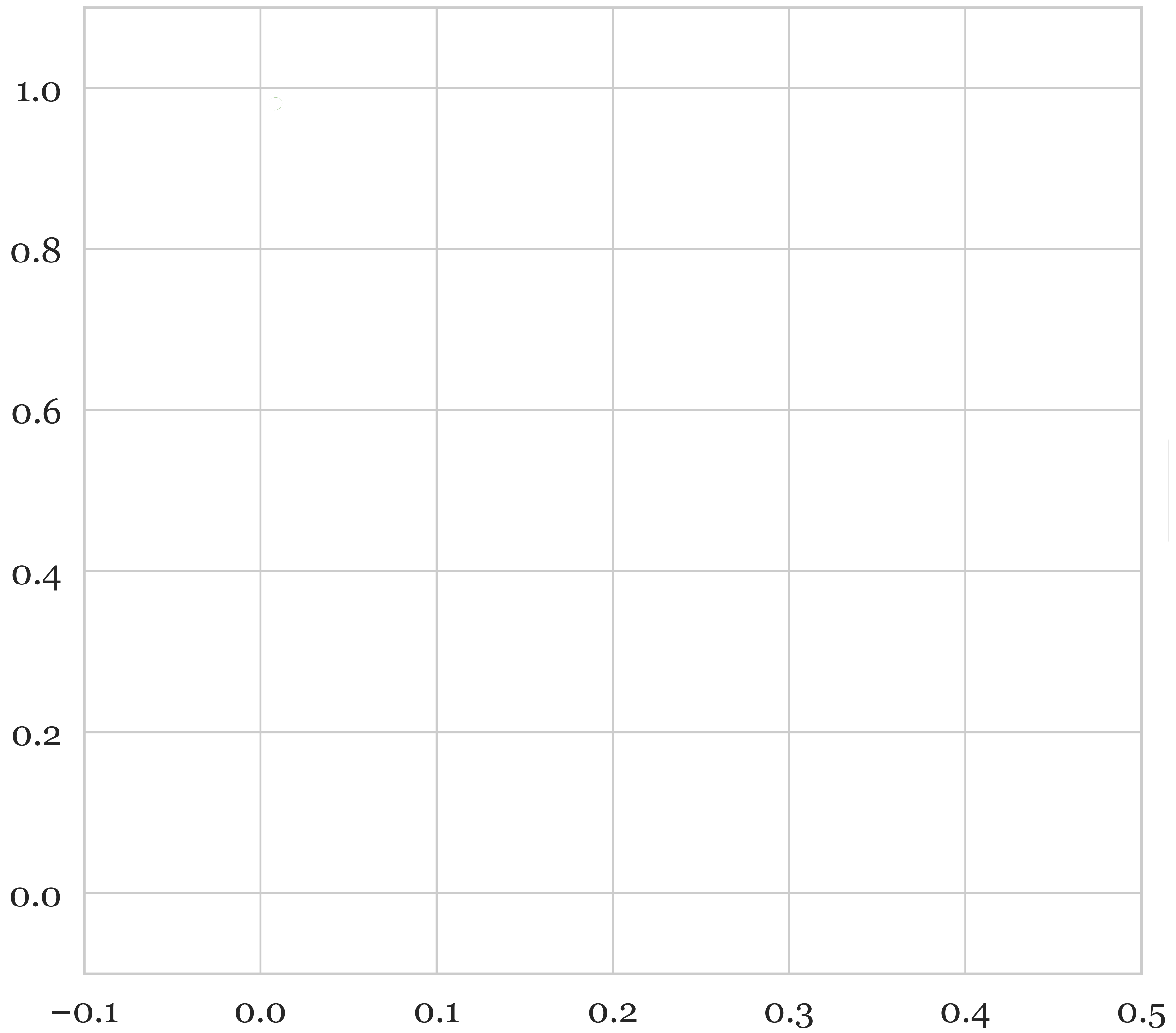
Annotation Artifacts in NLI [G*., **Swayamdipta***, L., S., B., S., NAACL 2018]

How can we better analyze the model-data relationship?

Model Training Dynamics

Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

Model Training Dynamics



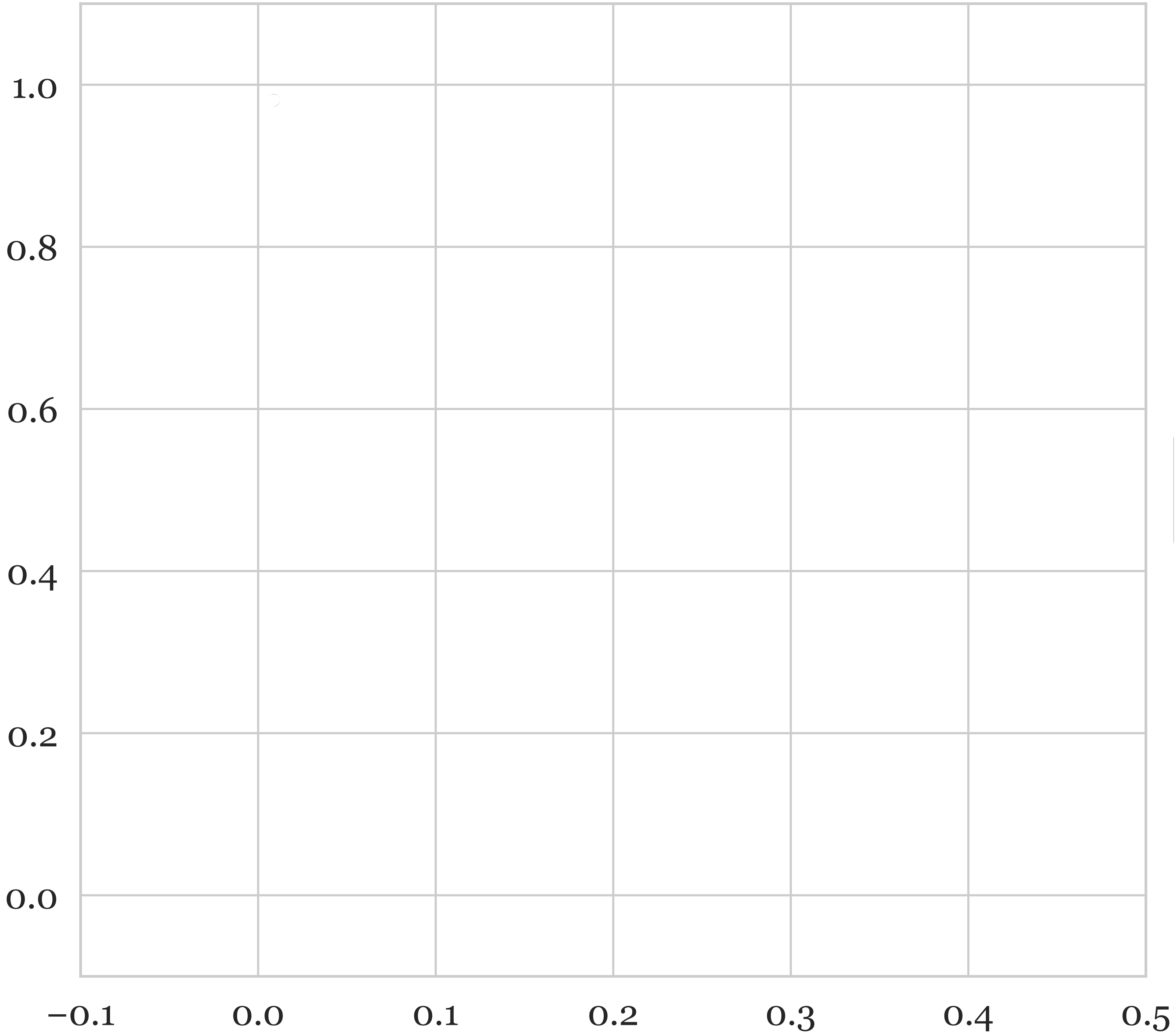
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

Model Training Dynamics

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* \mid x_i)$$

confidence

Mean probability of the **true class**



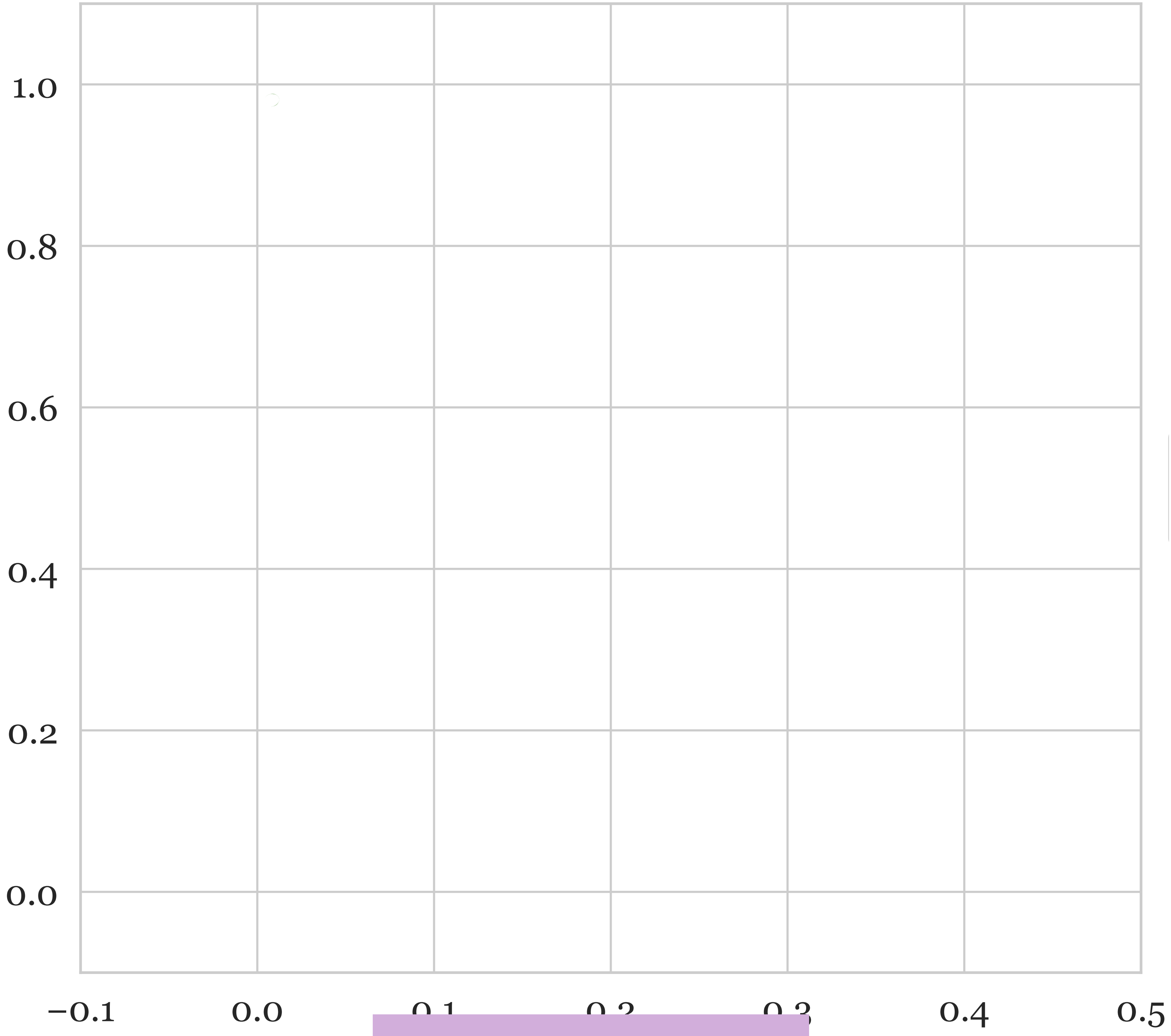
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

Model Training Dynamics

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* \mid x_i)$$

confidence

Mean probability of the **true class**



variability

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta^{(e)}}(y_i^* \mid x_i) - \hat{\mu}_i\right)^2}{E}}$$

Standard deviation of the **true class** probability

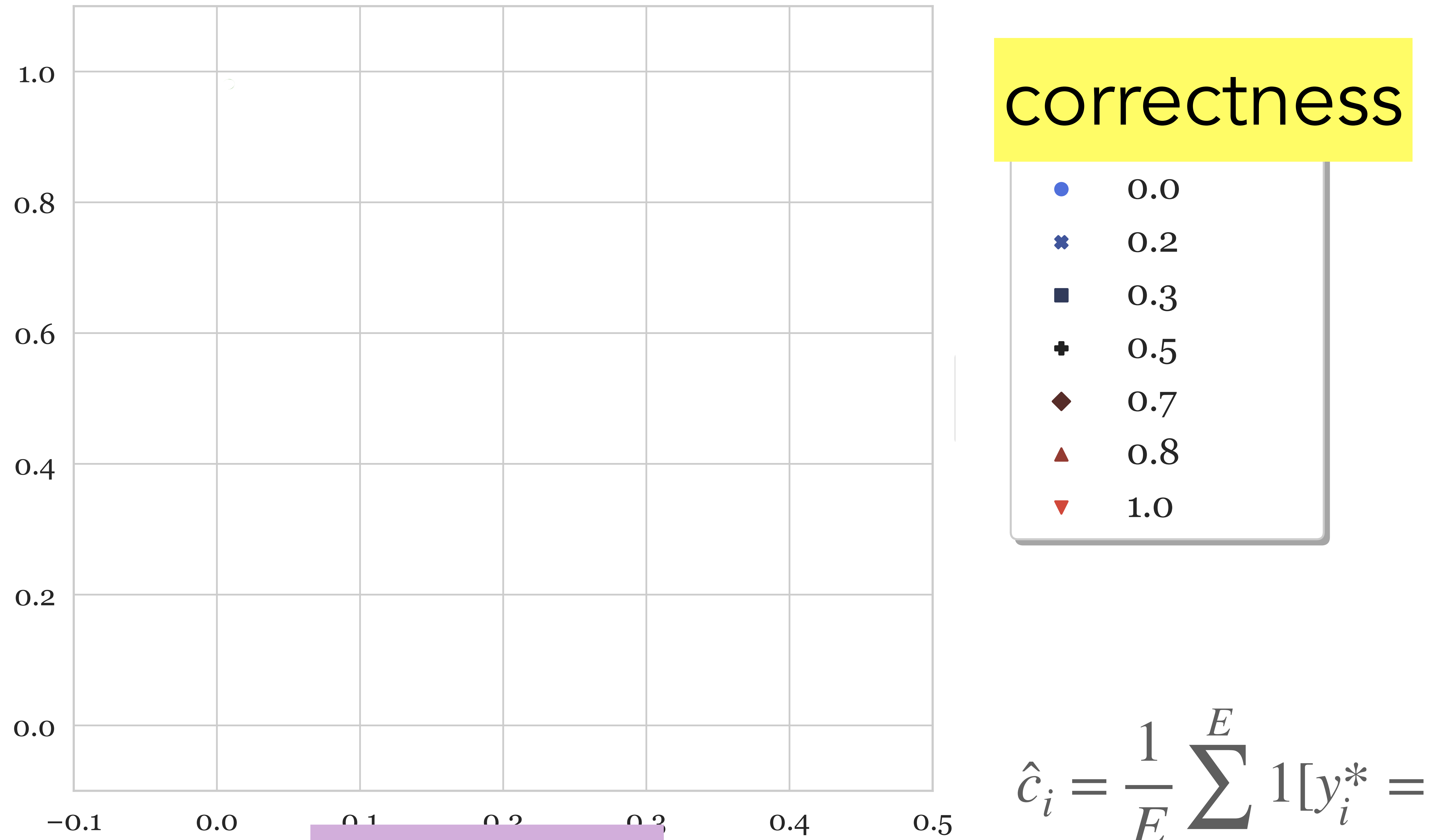
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

Model Training Dynamics

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* \mid x_i)$$

confidence

Mean probability of the **true class**



Ratio at which model prediction matches **true class**

$$\hat{c}_i = \frac{1}{E}\sum_{e=1}^{E} 1[y_i^* = \arg\max_{y} p_{\theta^{(e)}}(y \mid x_i)]$$

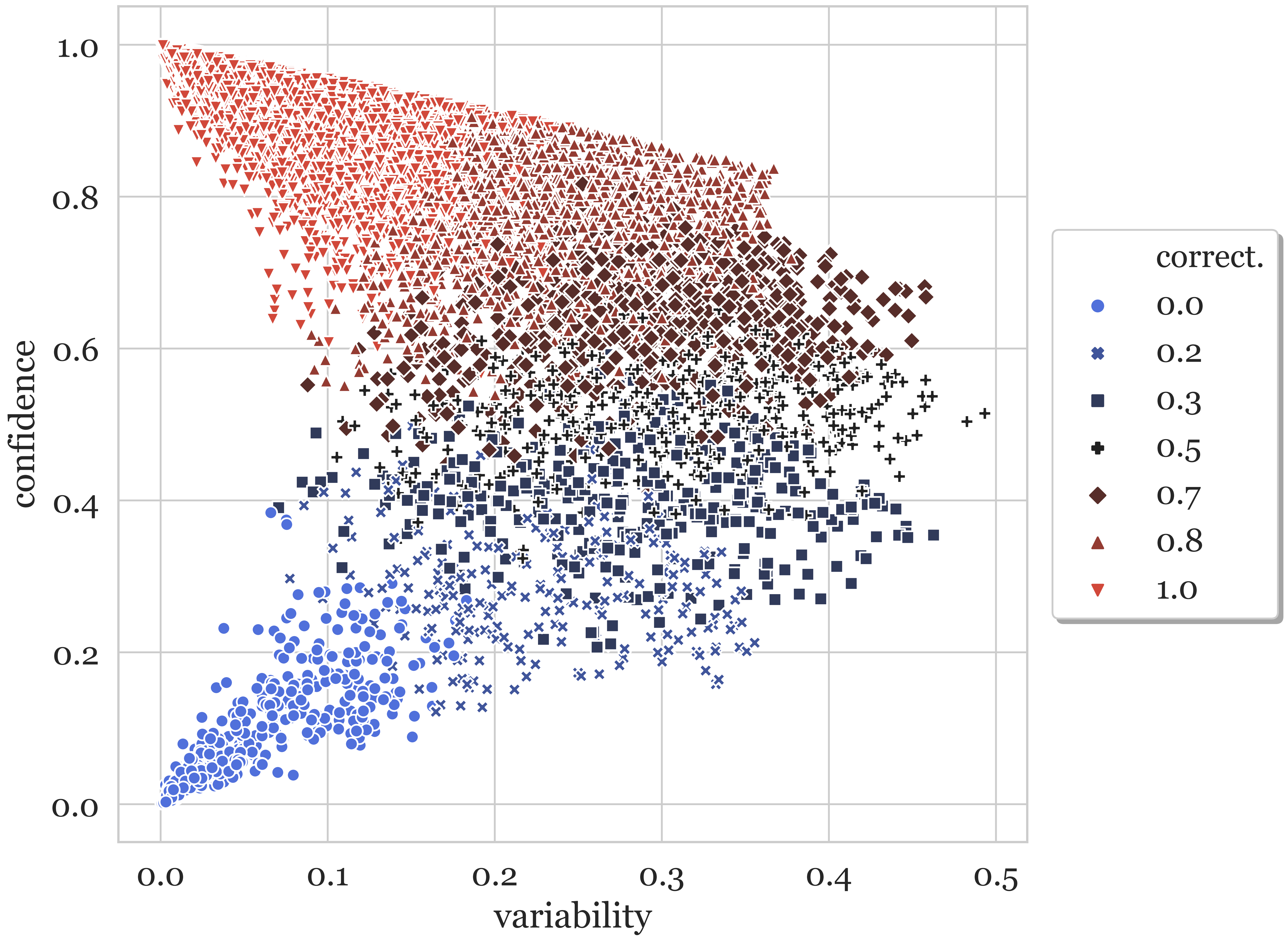
variability

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\theta^{(e)}}(y_i^* \mid x_i) - \hat{\mu}_i\right)^2}{E}}$$

Standard deviation of the **true class** probability

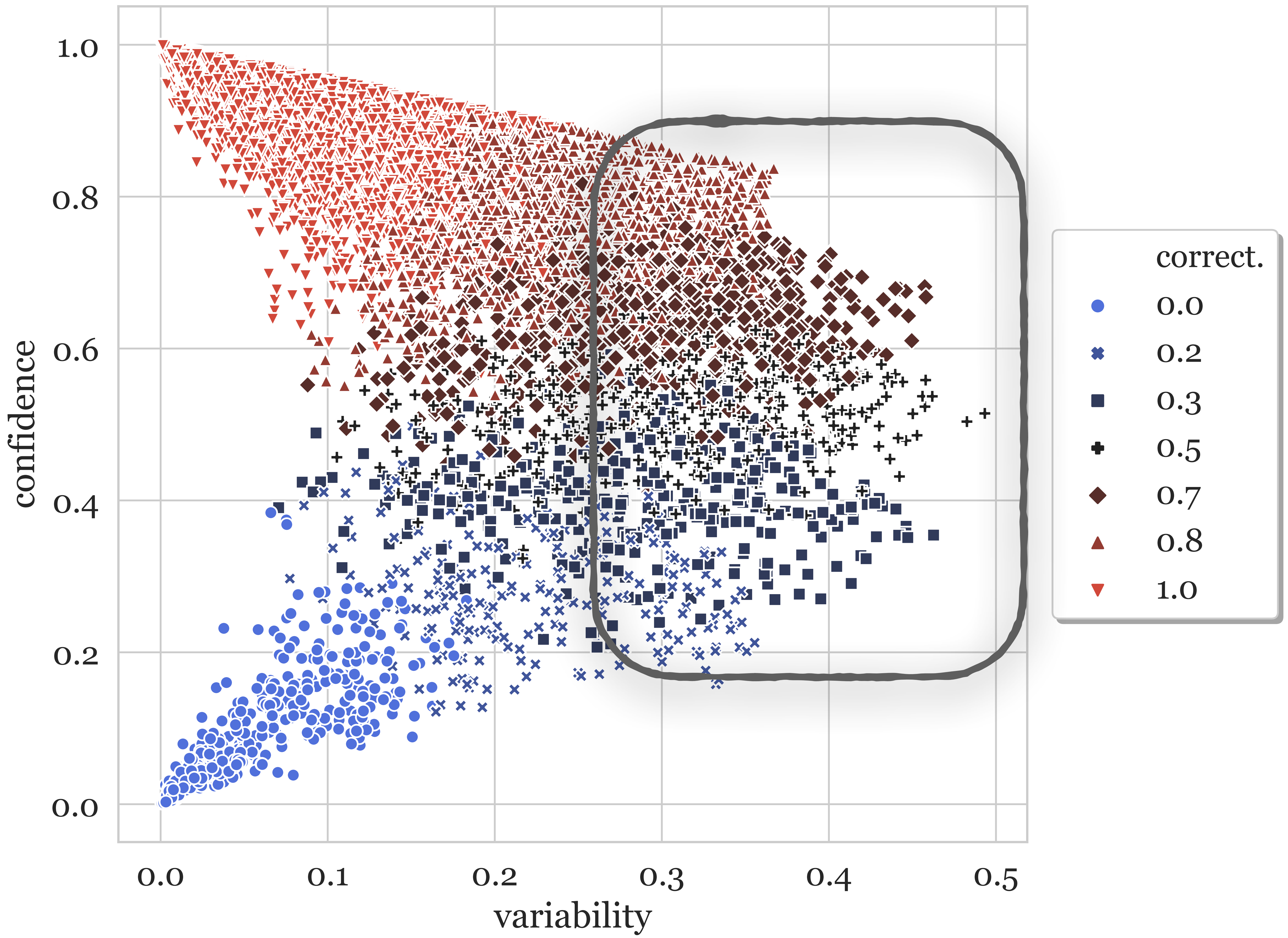
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map



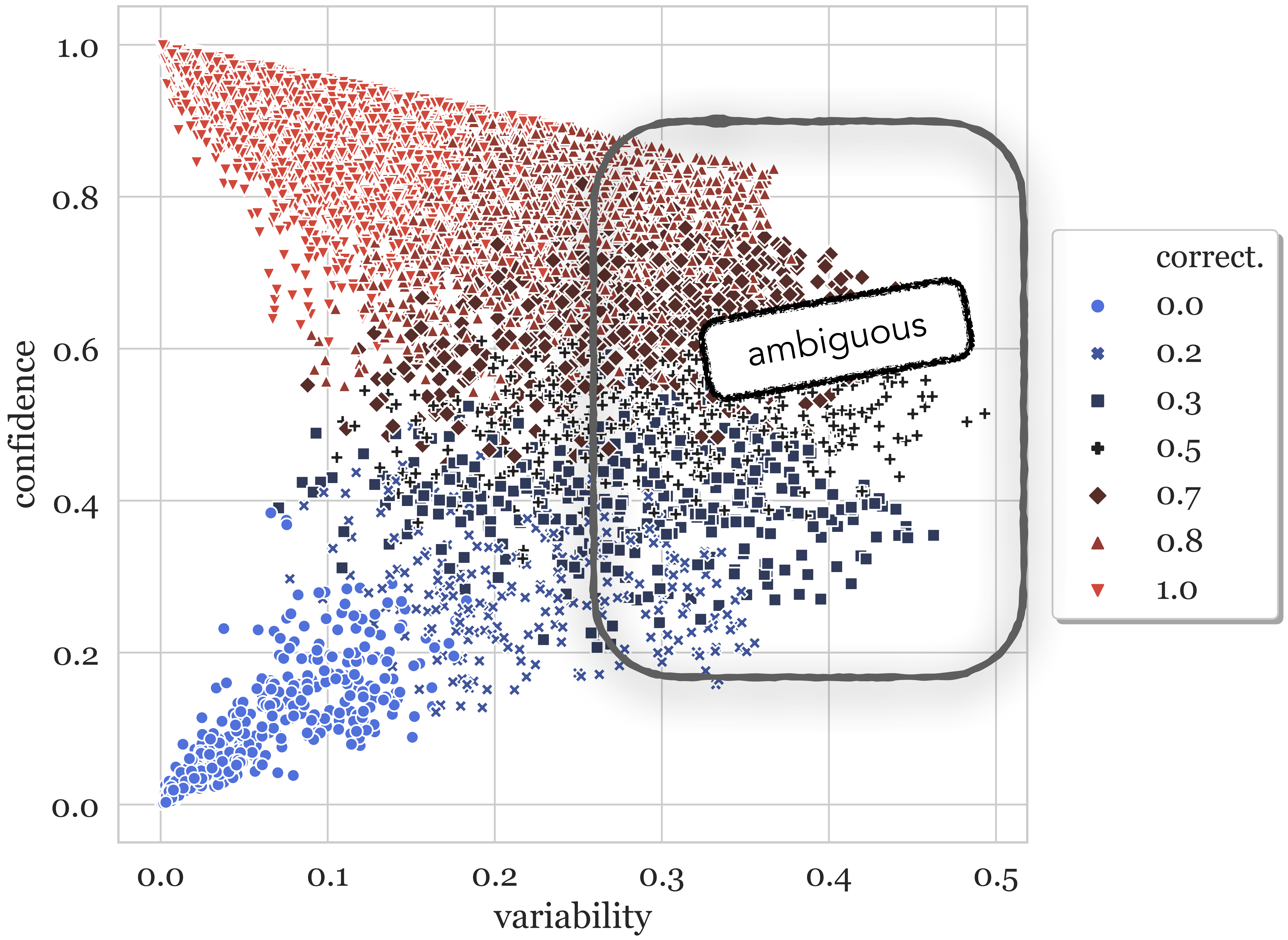
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map



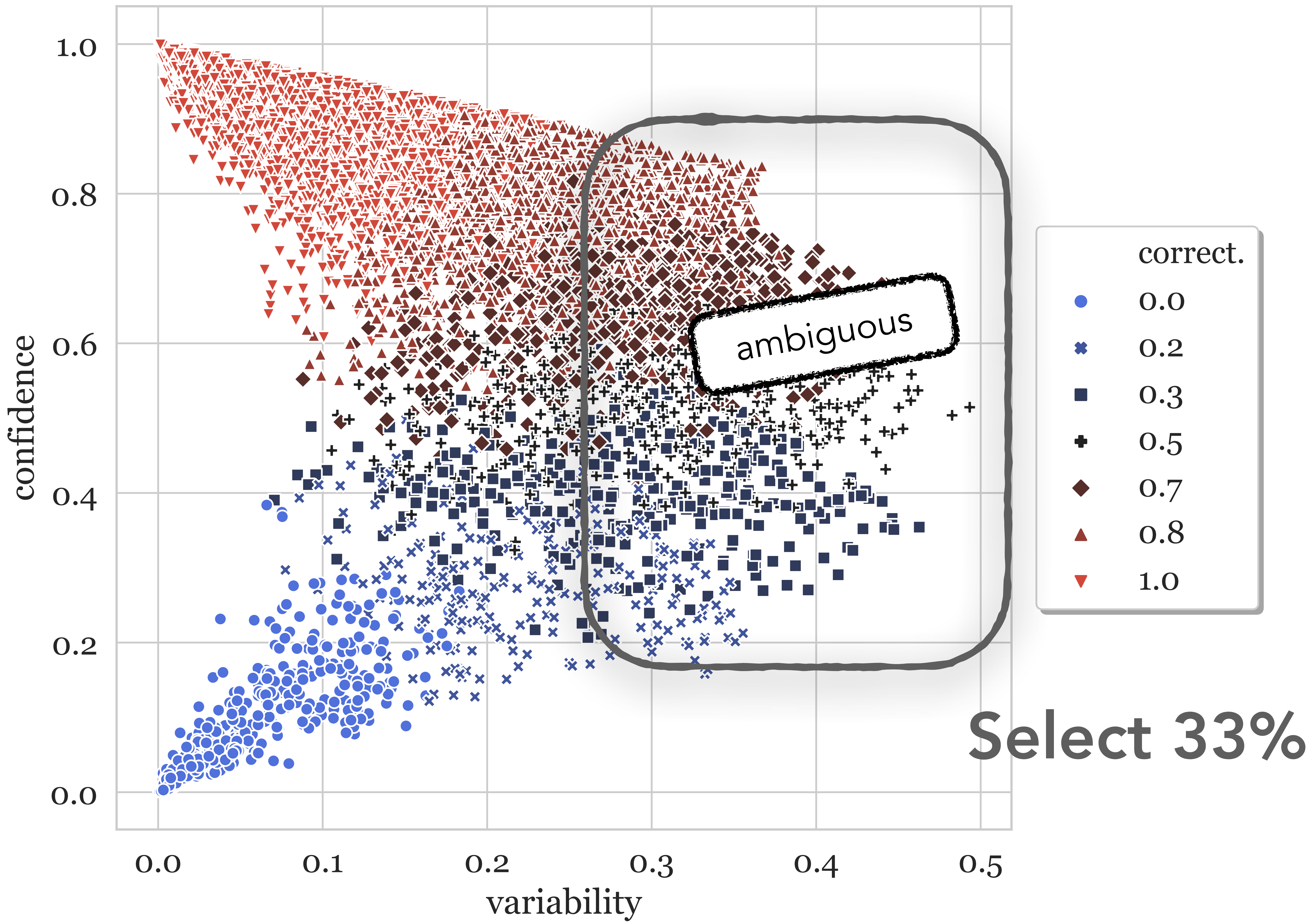
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map



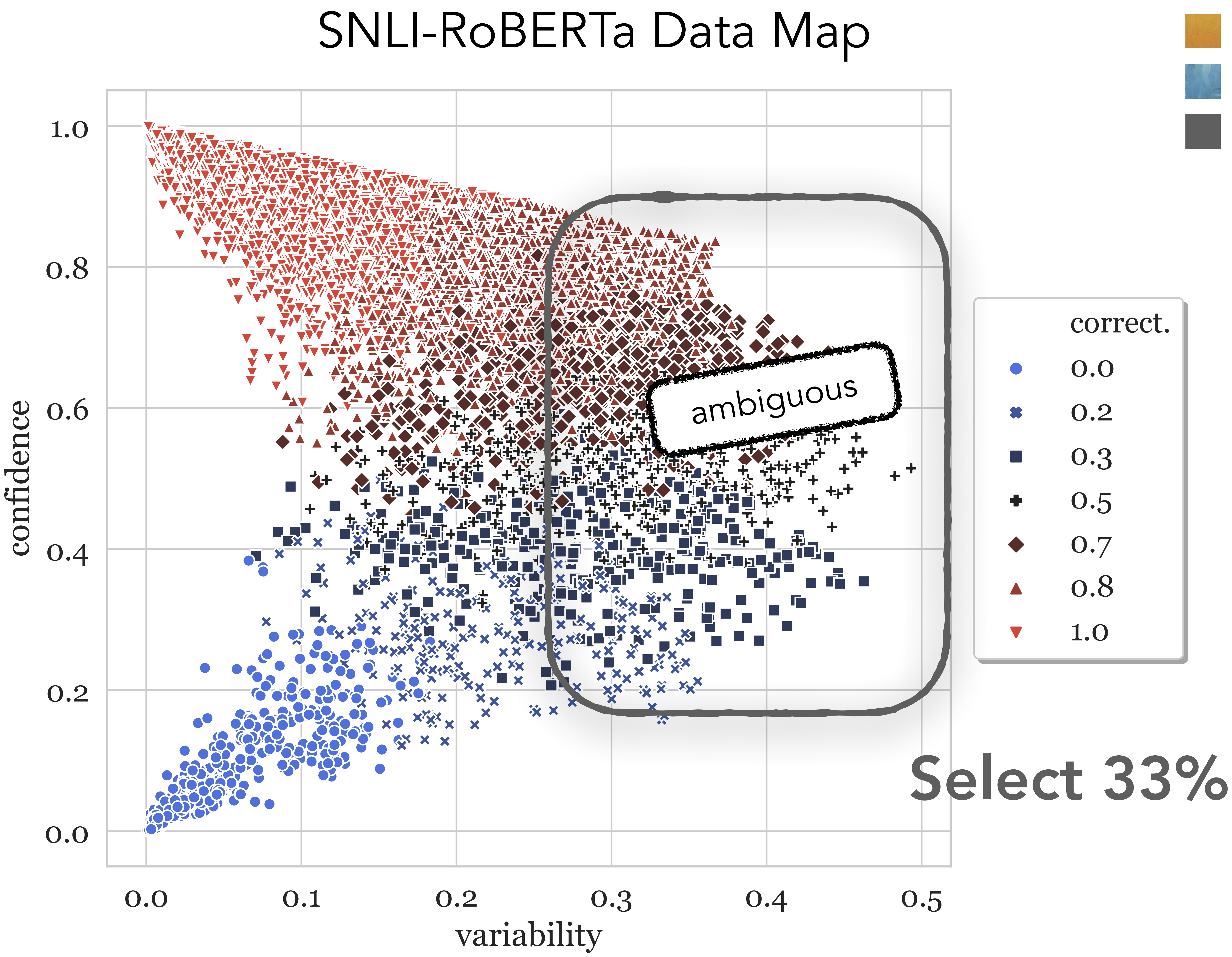
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map

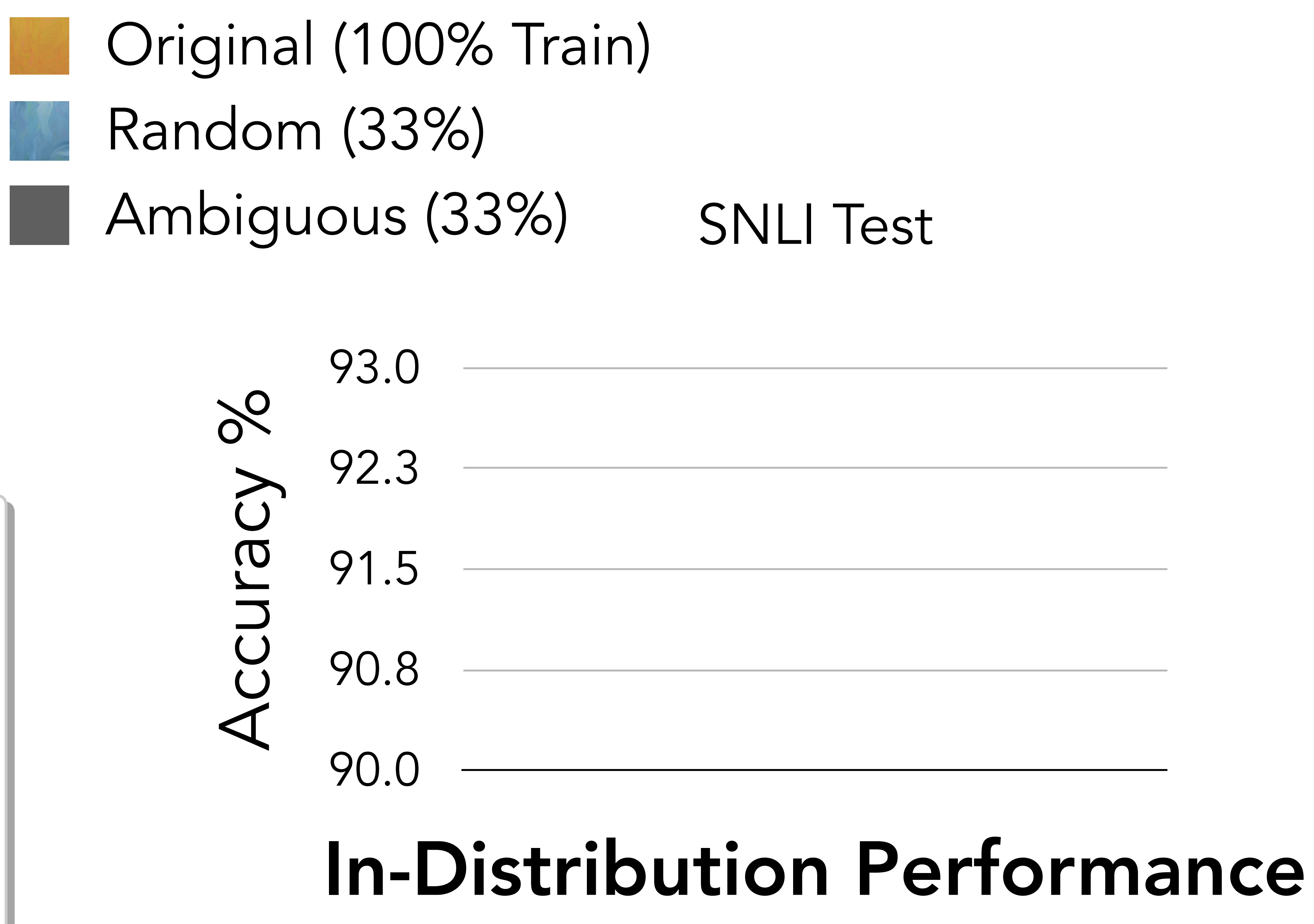


Select 33%

Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

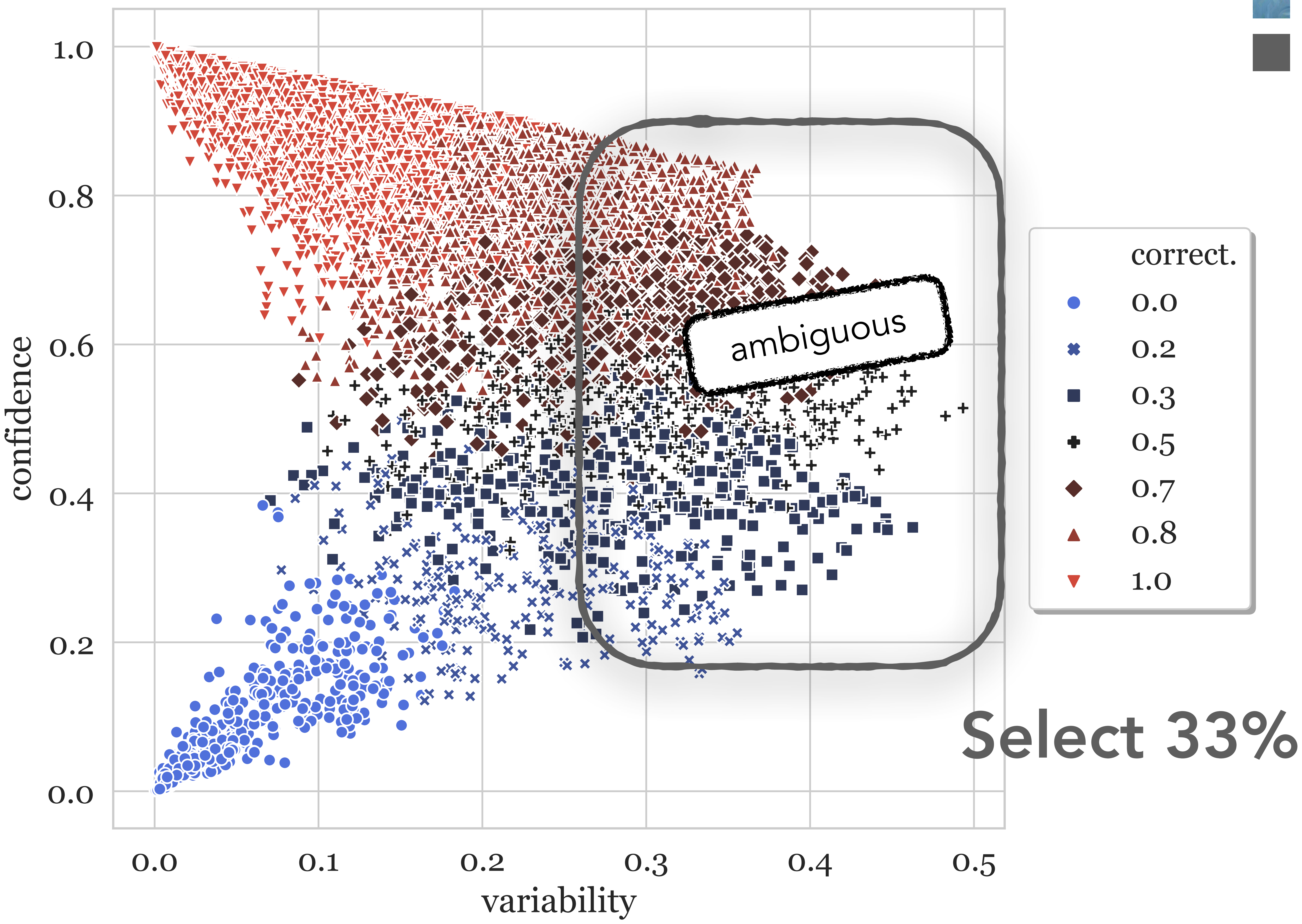


Select 33%

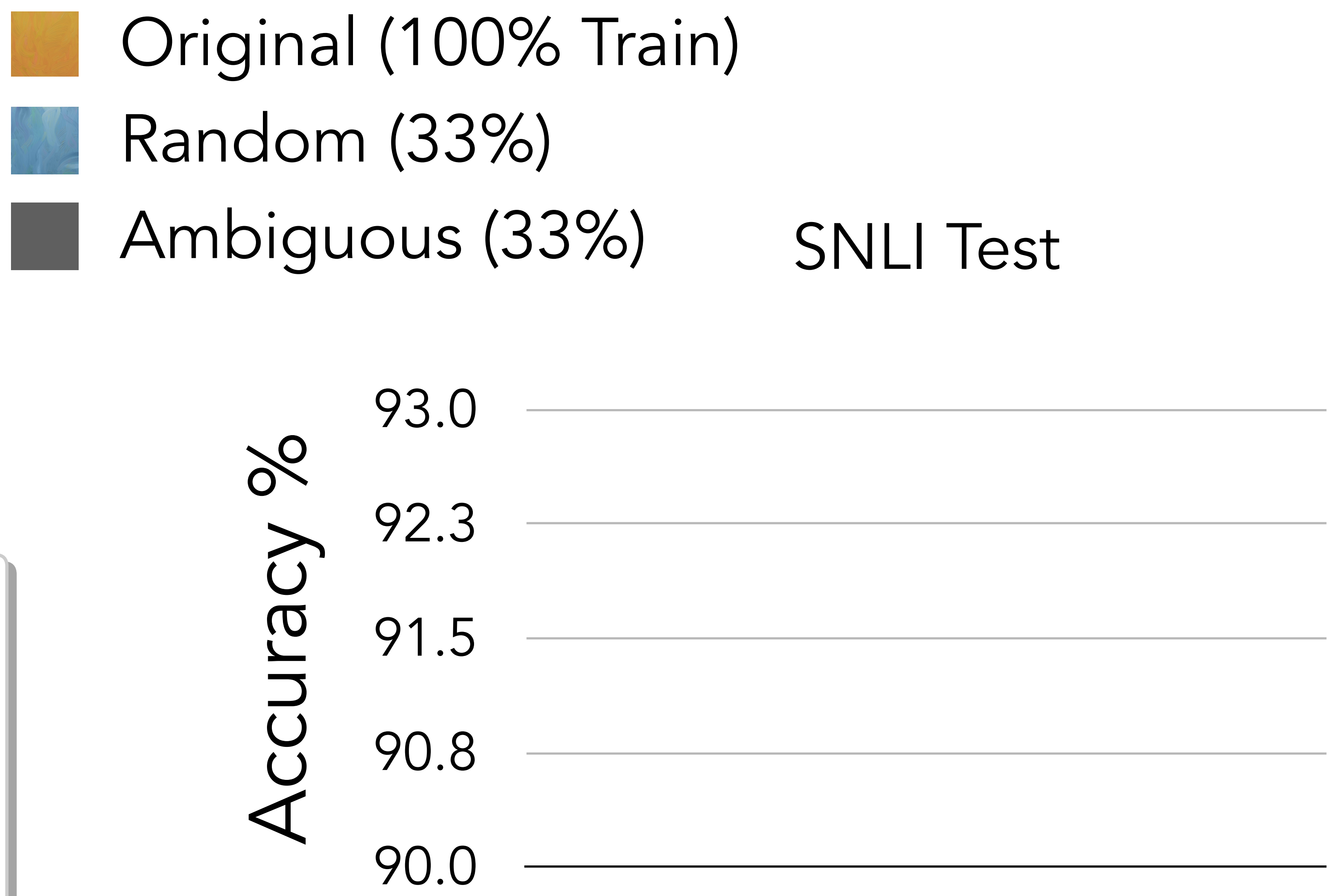


Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map



Select 33%

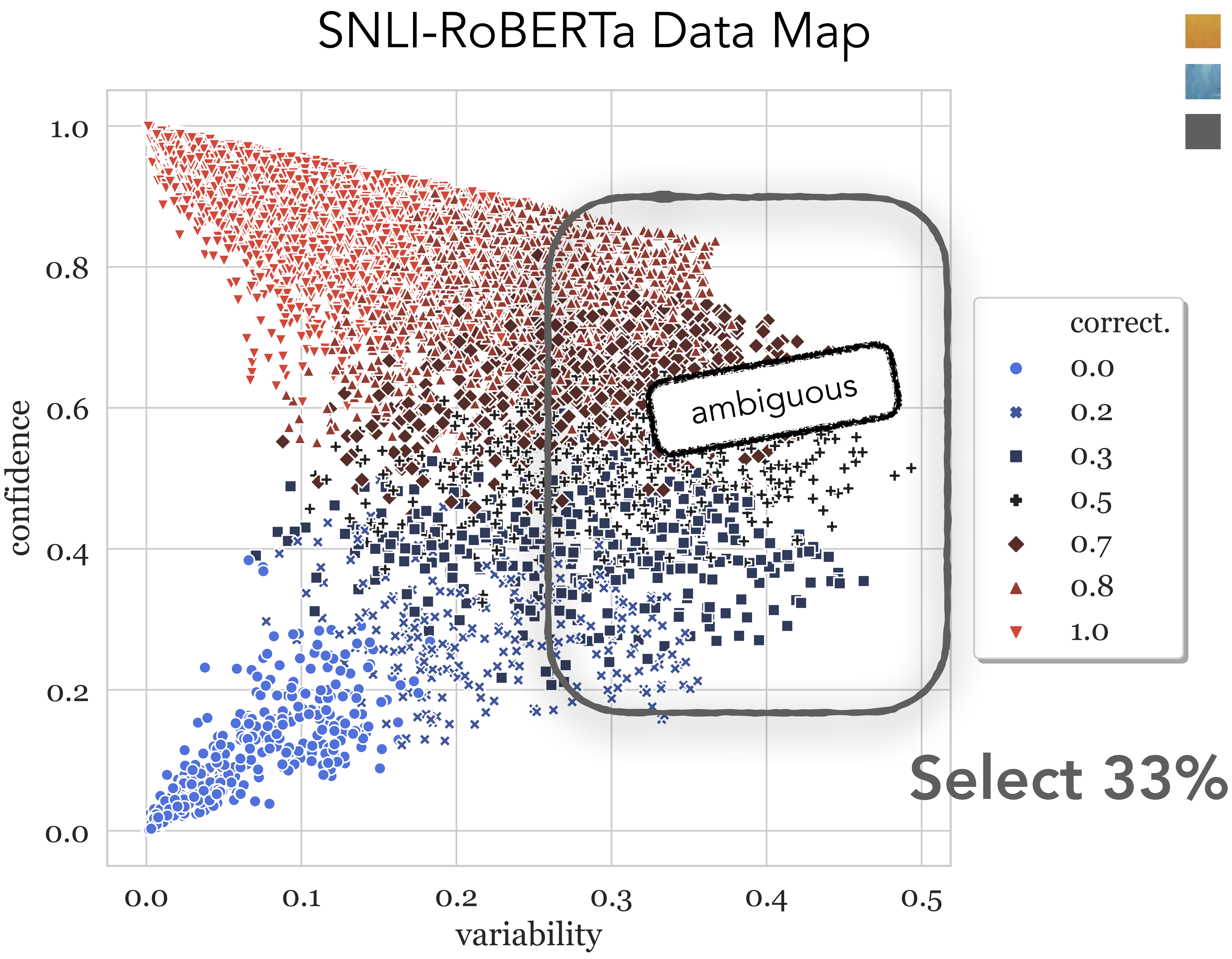


In-Distribution Performance

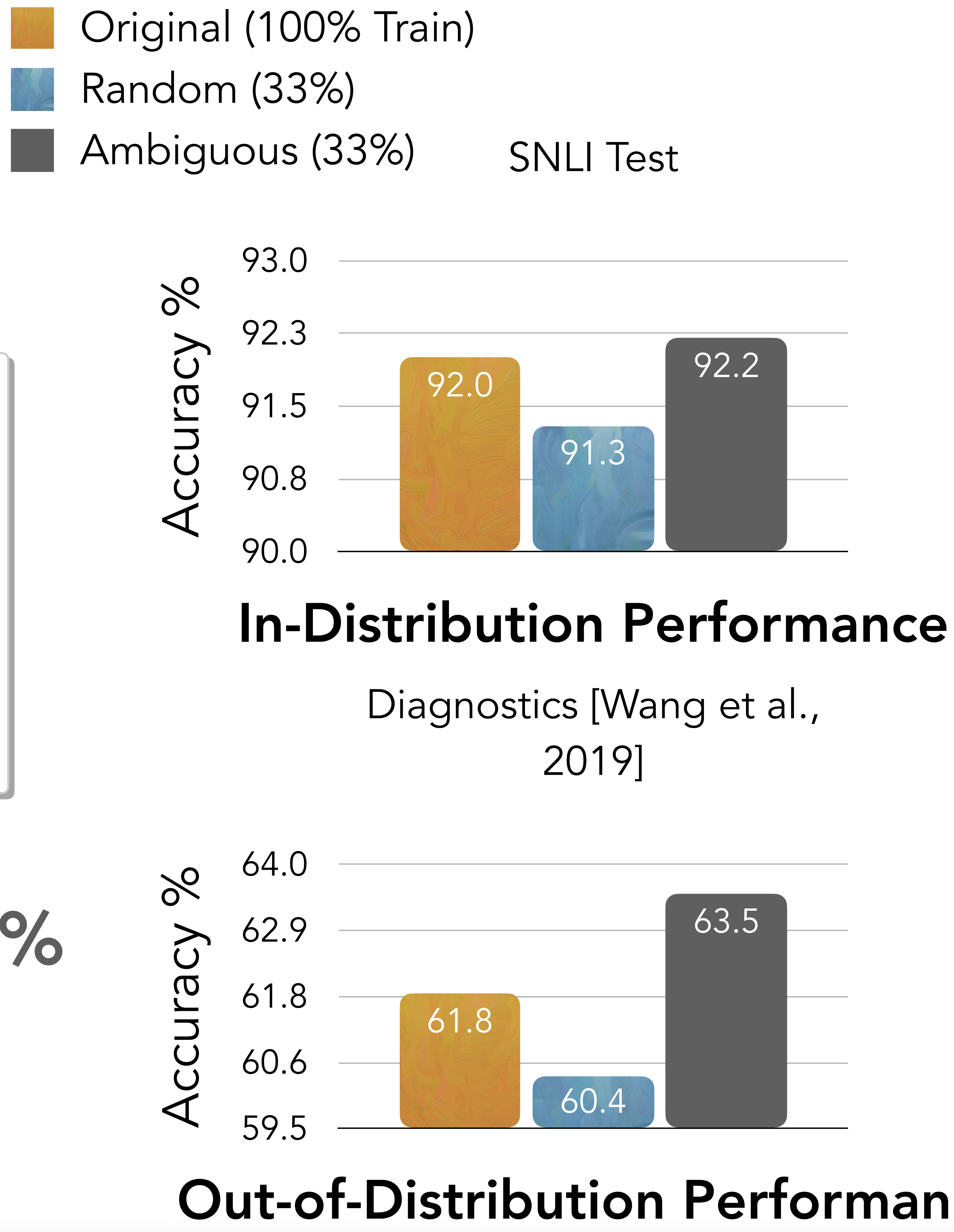
Diagnostics [Wang et al., 2019]

Out-of-Distribution Performance

Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

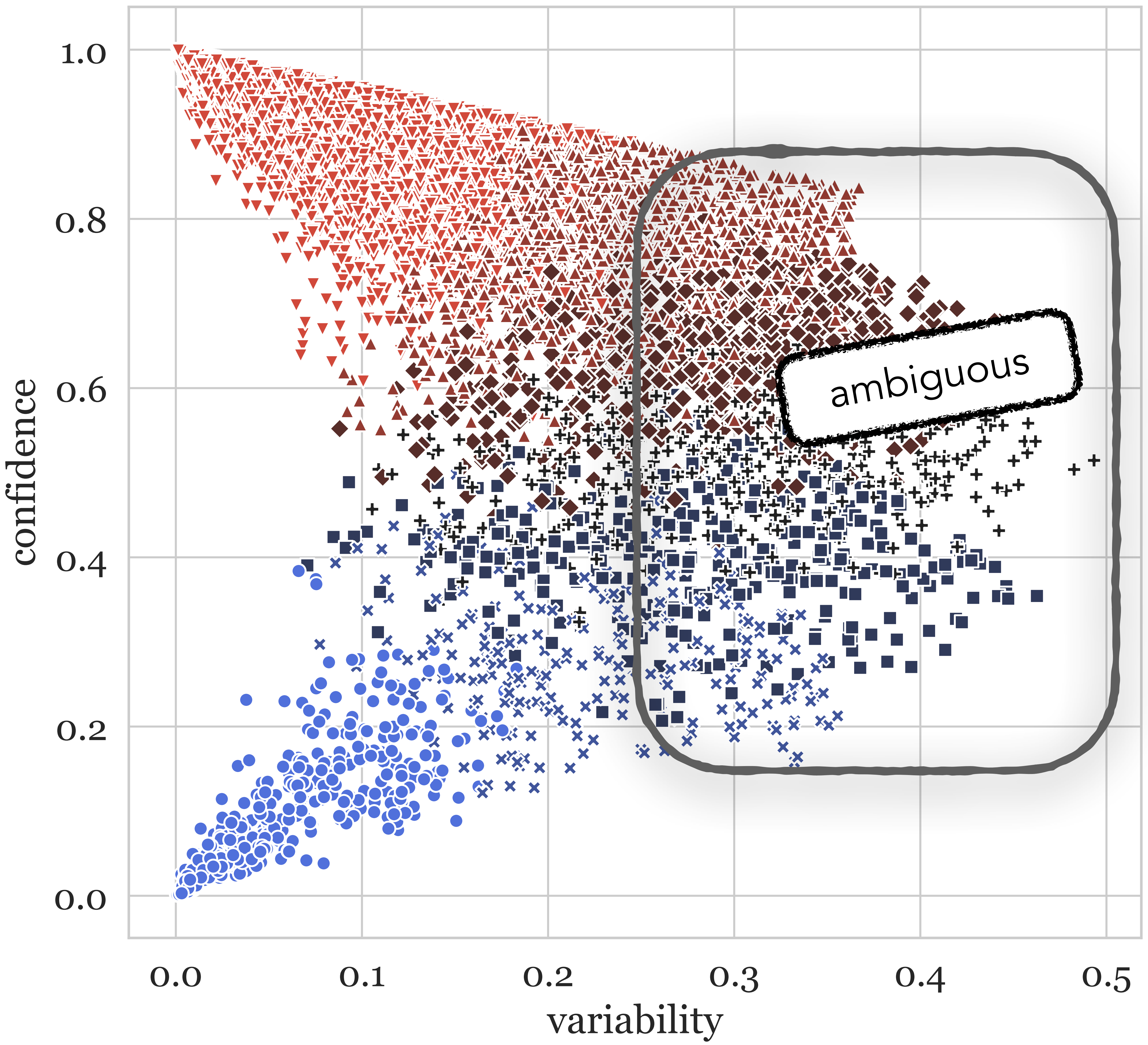


Select 33%



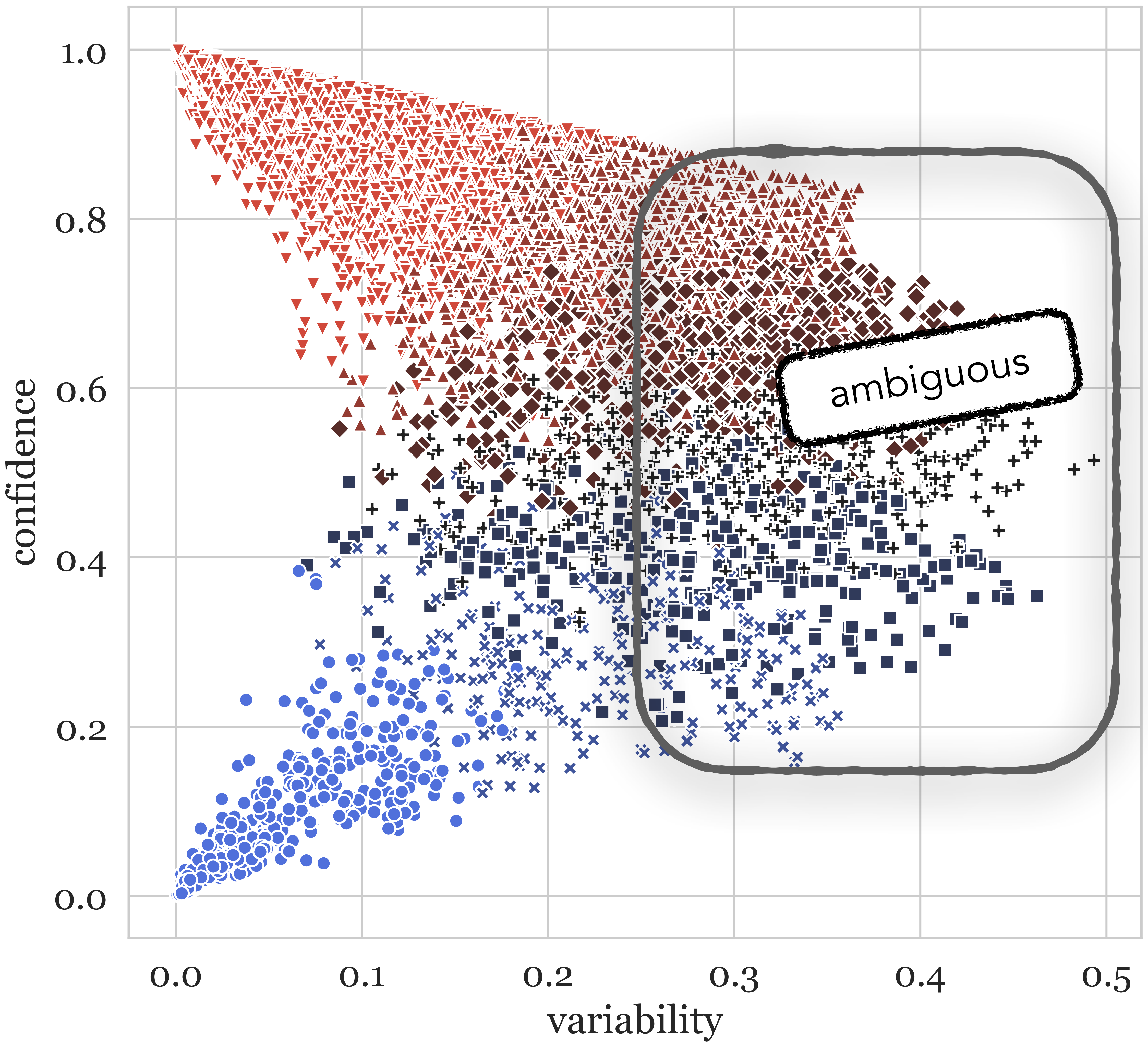
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

SNLI-RoBERTa Data Map



Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

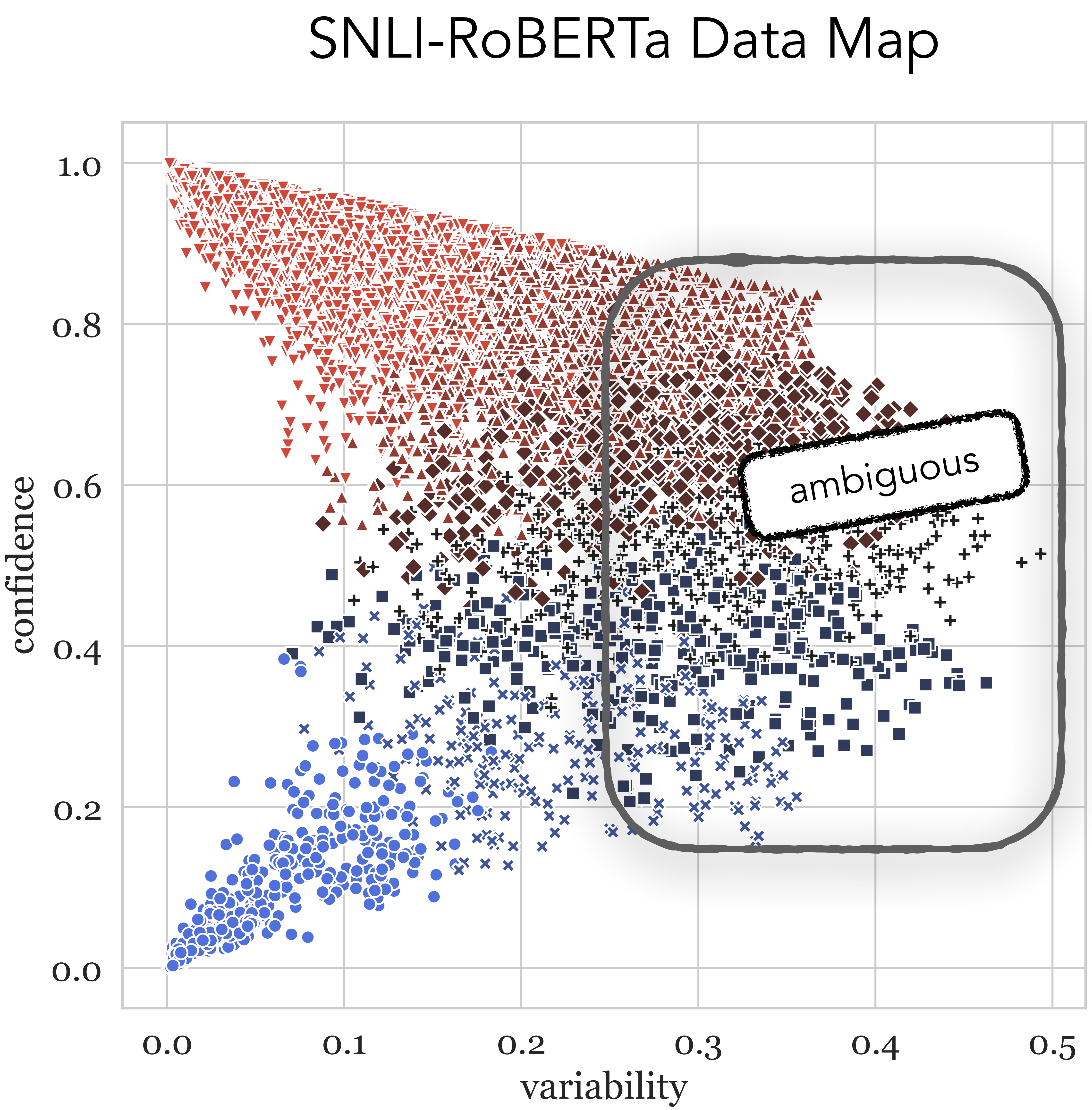
SNLI-RoBERTa Data Map



An expression gathered there that I can only describe as **half puzzled, and half relieved**.

The expression on their face was **puzzled and relieved**.

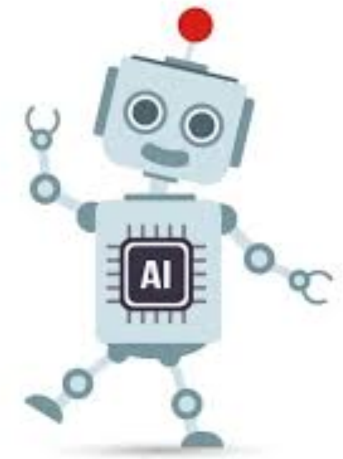
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]



An expression gathered there that I can only describe as **half puzzled, and half relieved**.

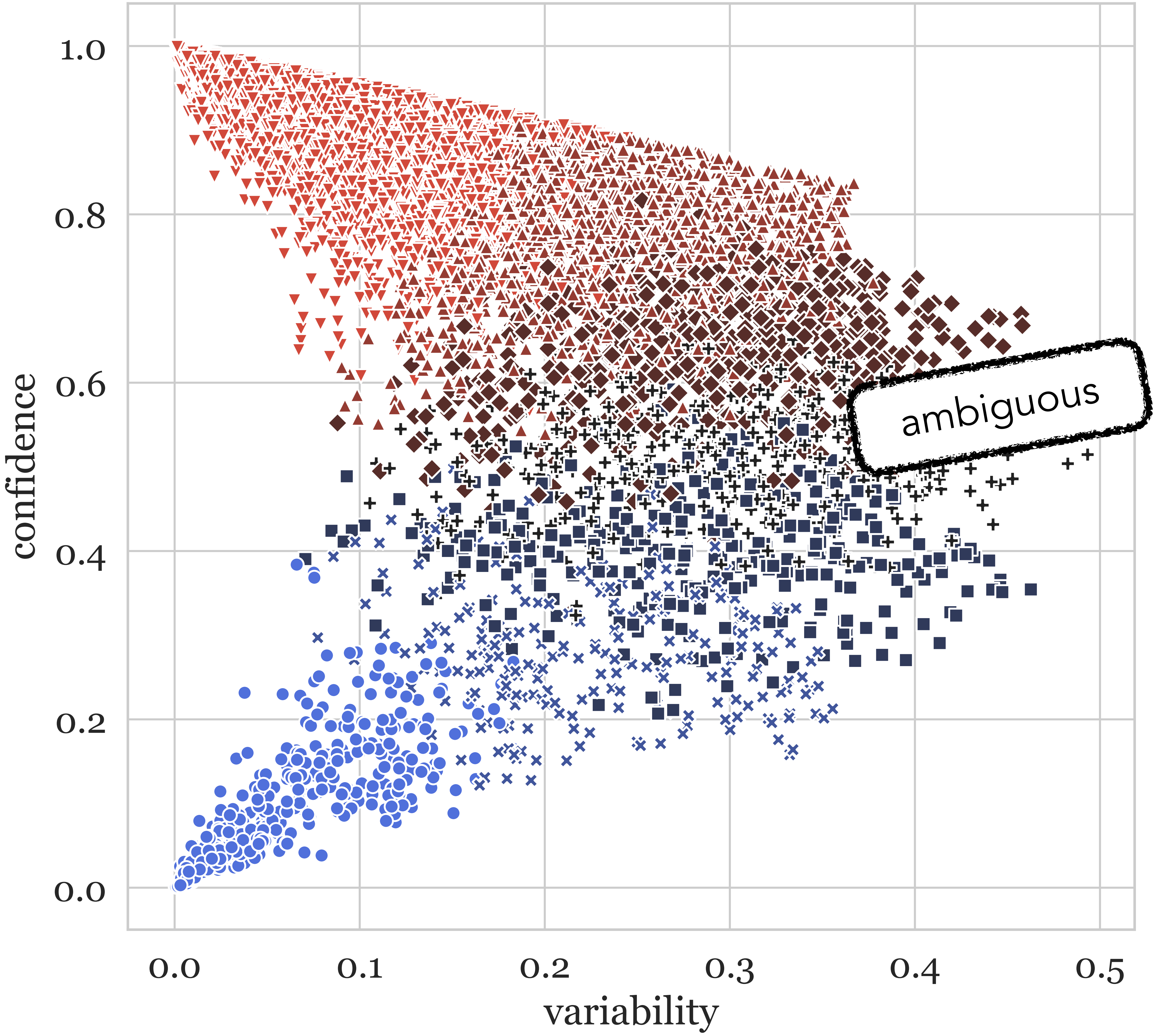
The expression on their face was **puzzled and relieved**.

Neutral

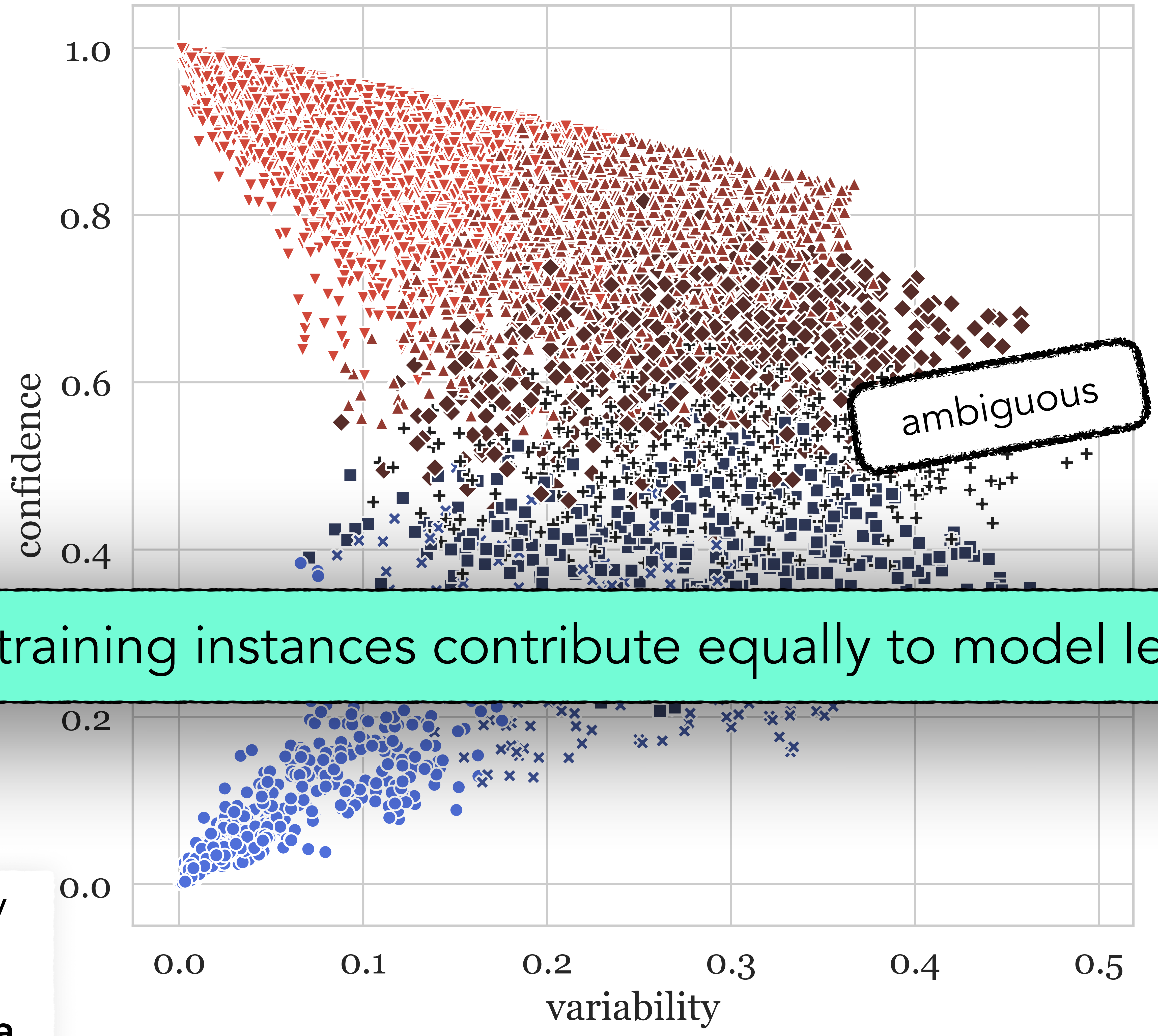


Entailment

Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]



Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

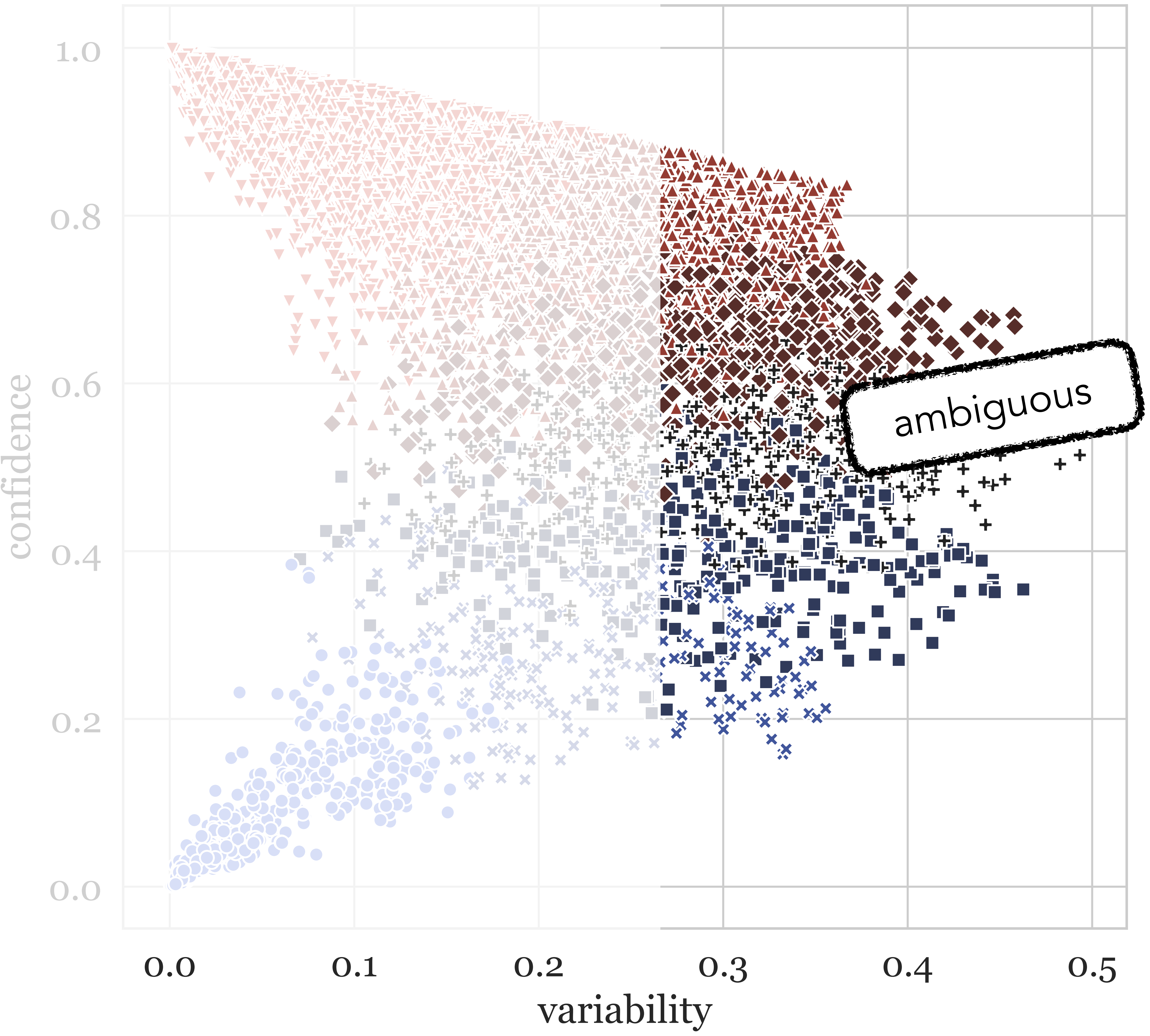


Not all training instances contribute equally to model learning

Also see

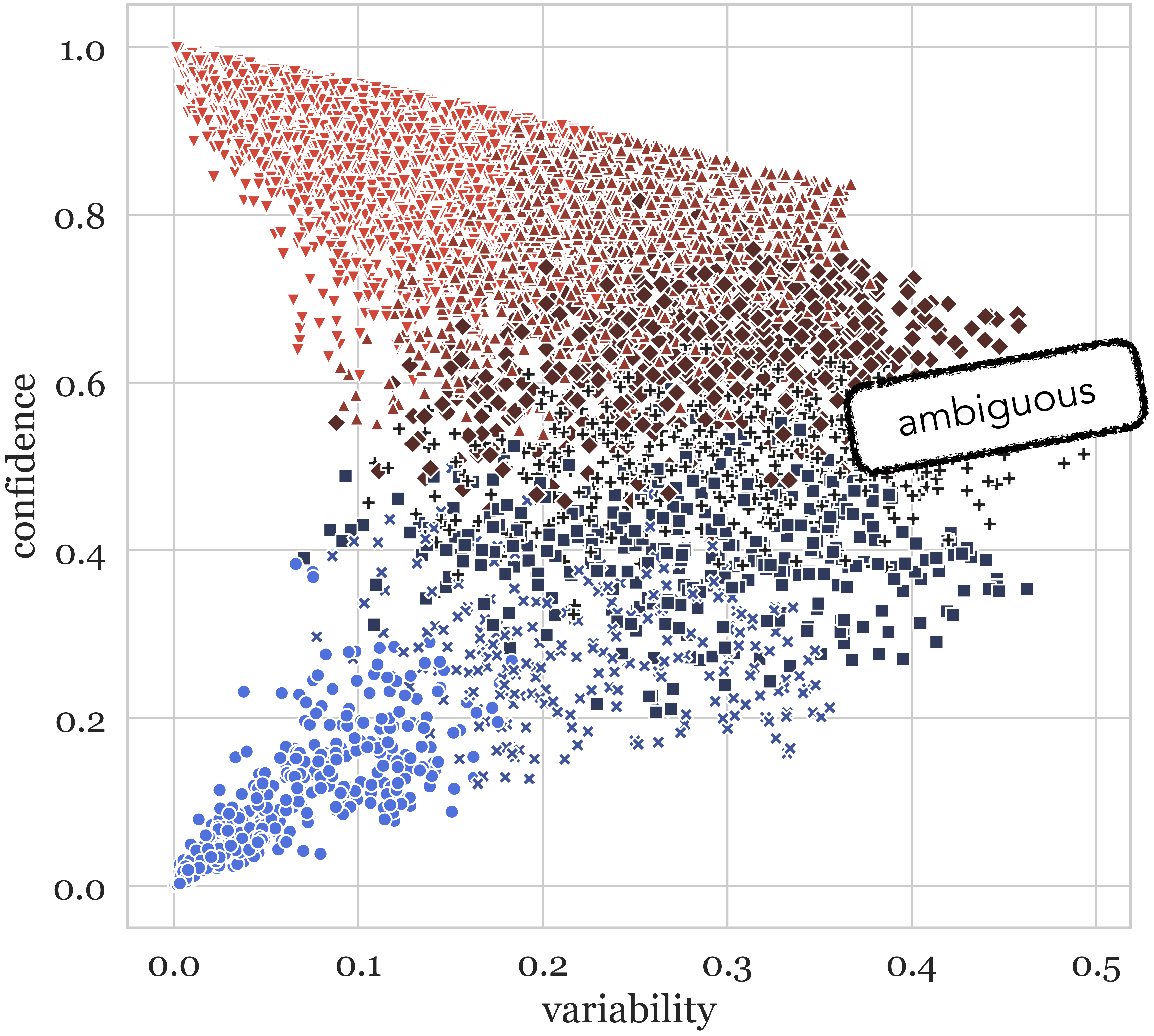
Understanding Dataset Difficulty with 𝒱-Usable Information [Ethayarajh, Choi & **Swayamdipta**, ICML 2022]

Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]



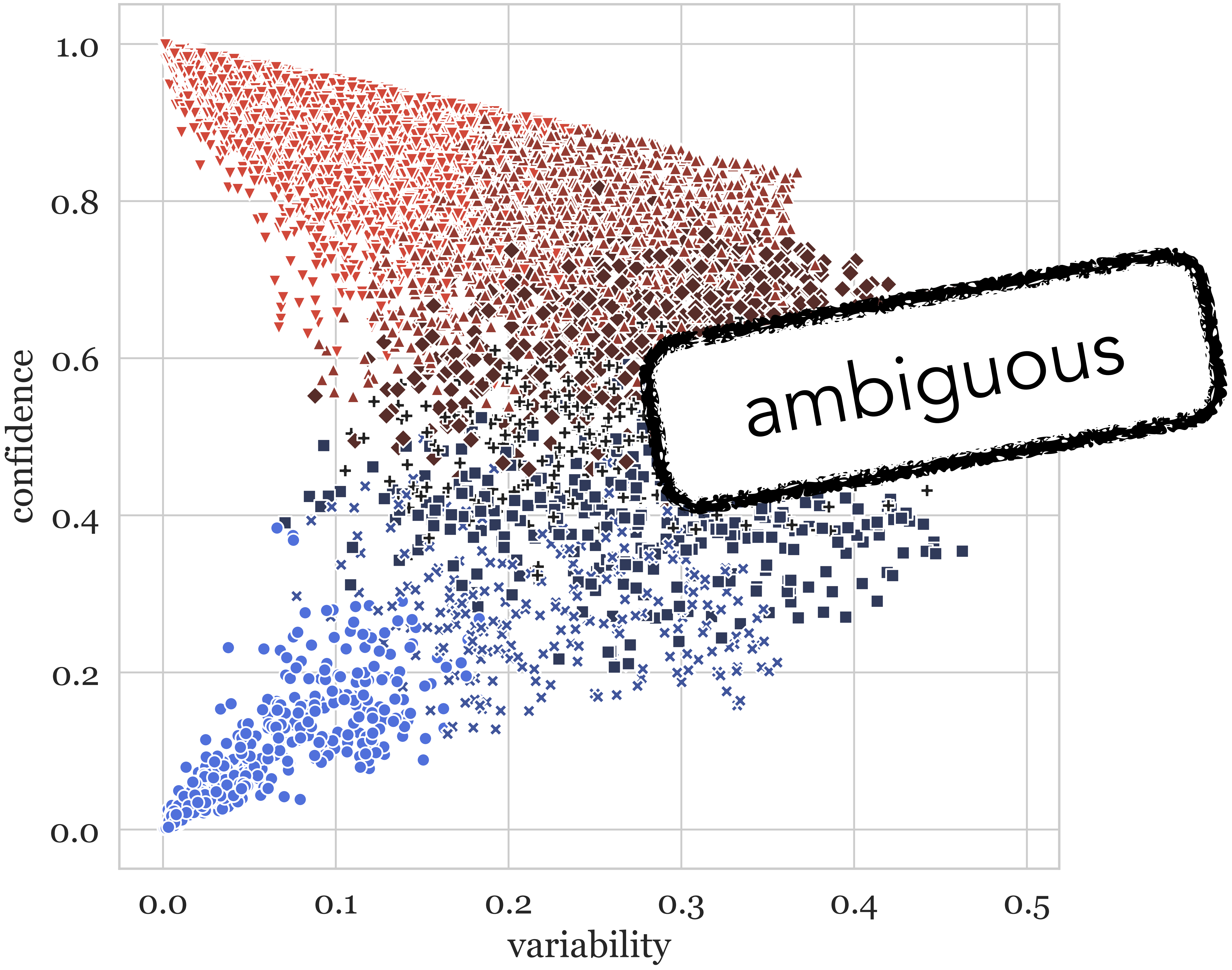
Dataset Cartography [**Swayamdipta** et al., EMNLP 2020]

Can we leverage data maps
to improve dataset collection?



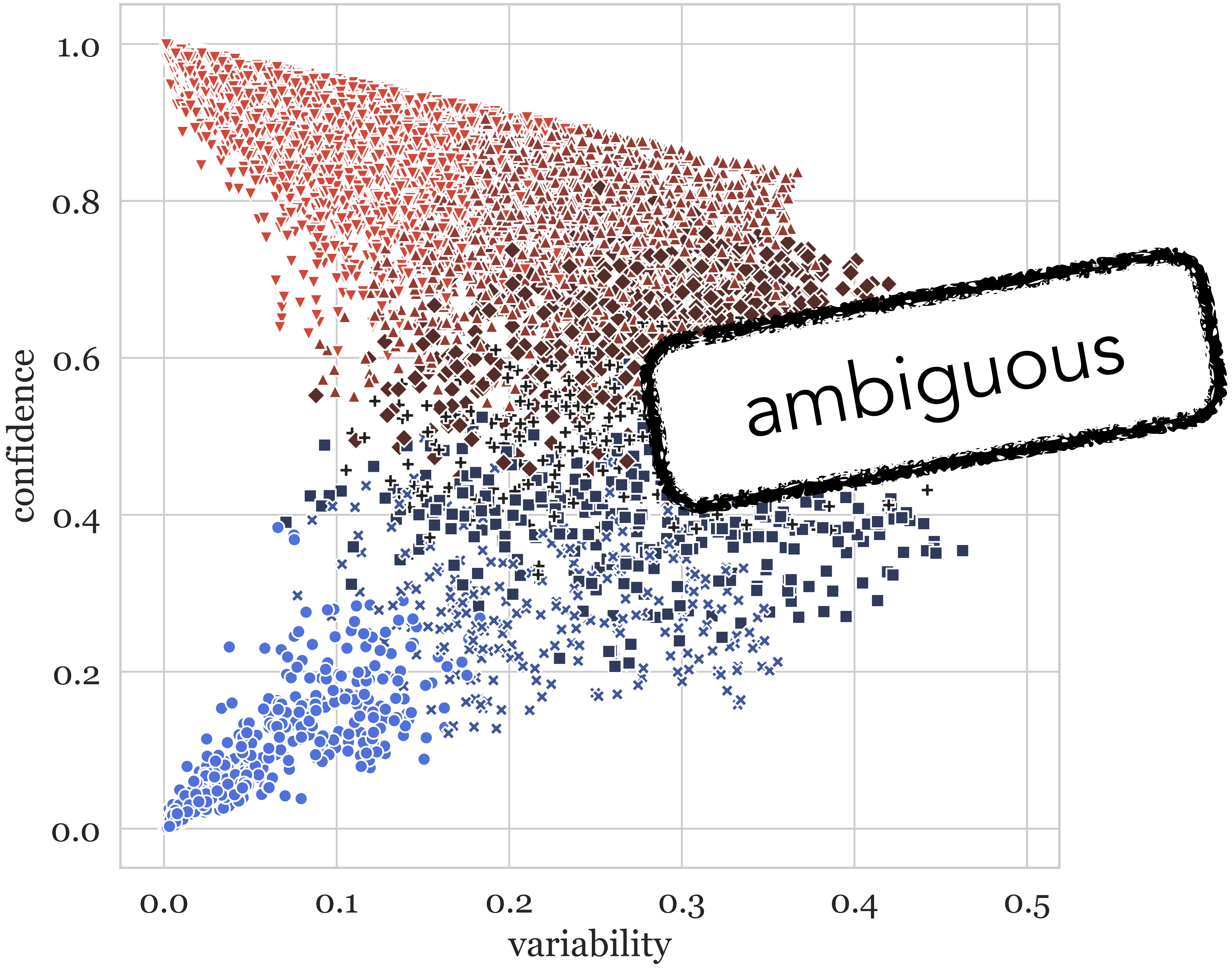
ambiguous

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

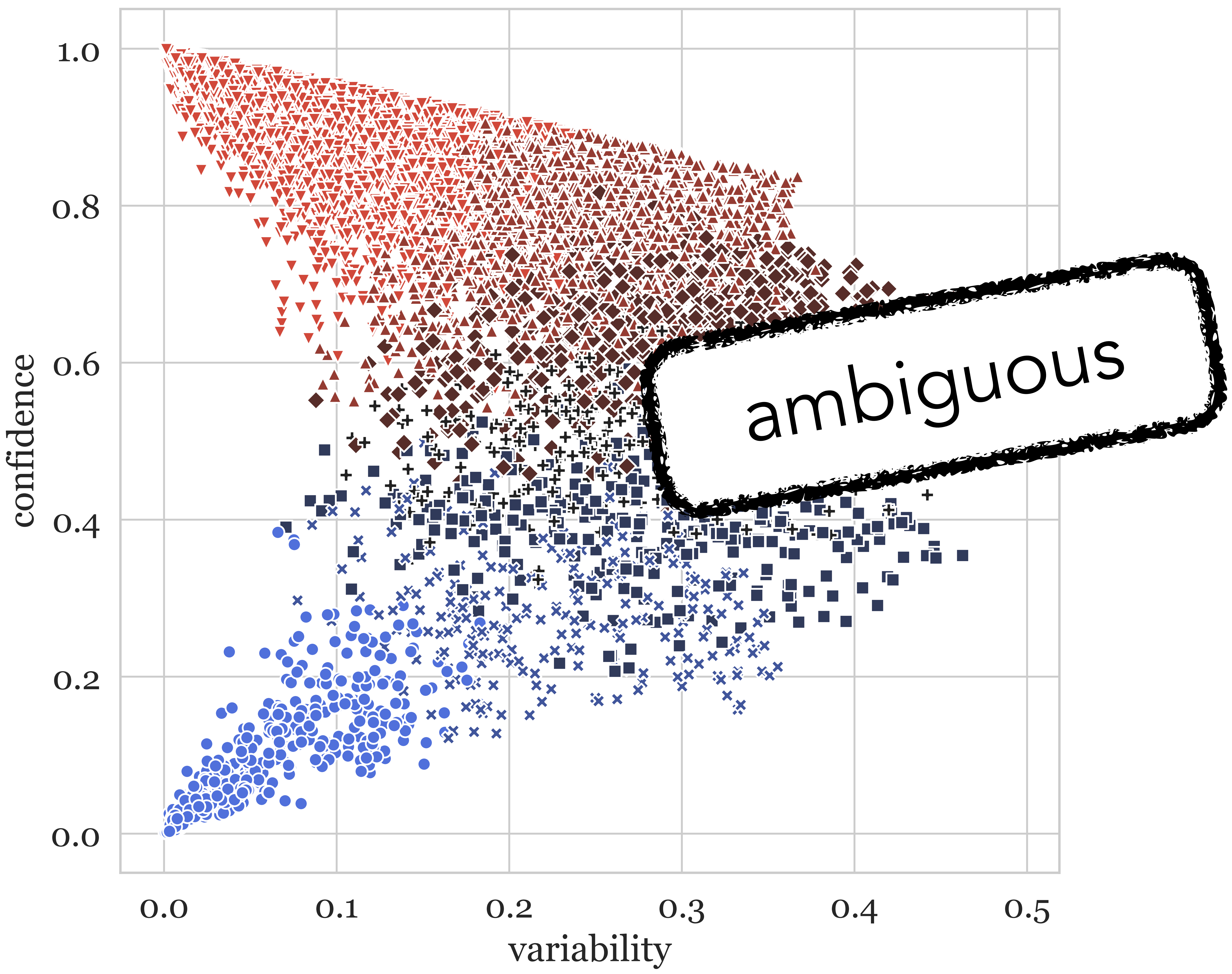


ambiguous

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

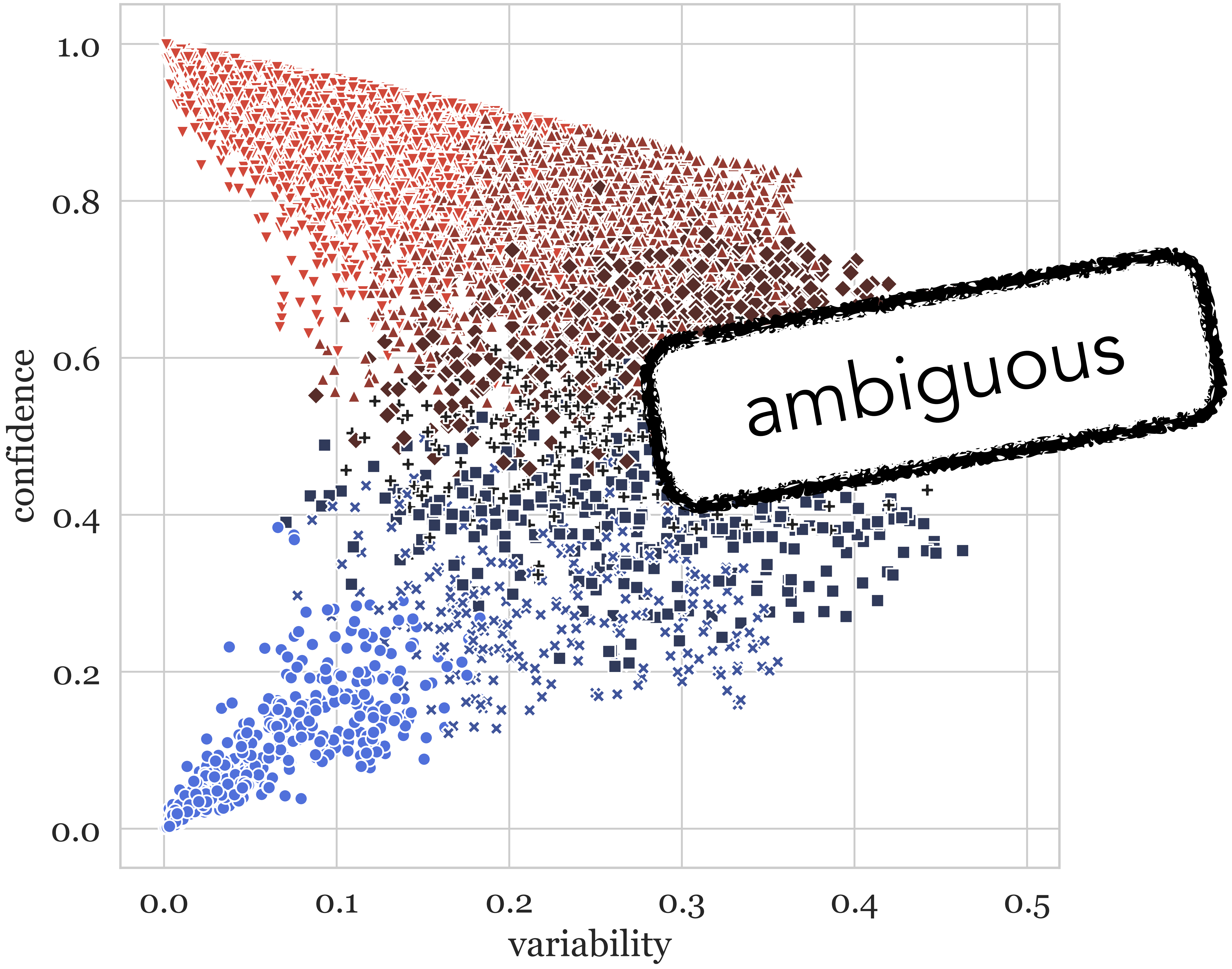


WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Might introduce heuristics leading to annotation artifacts

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



GPT-3

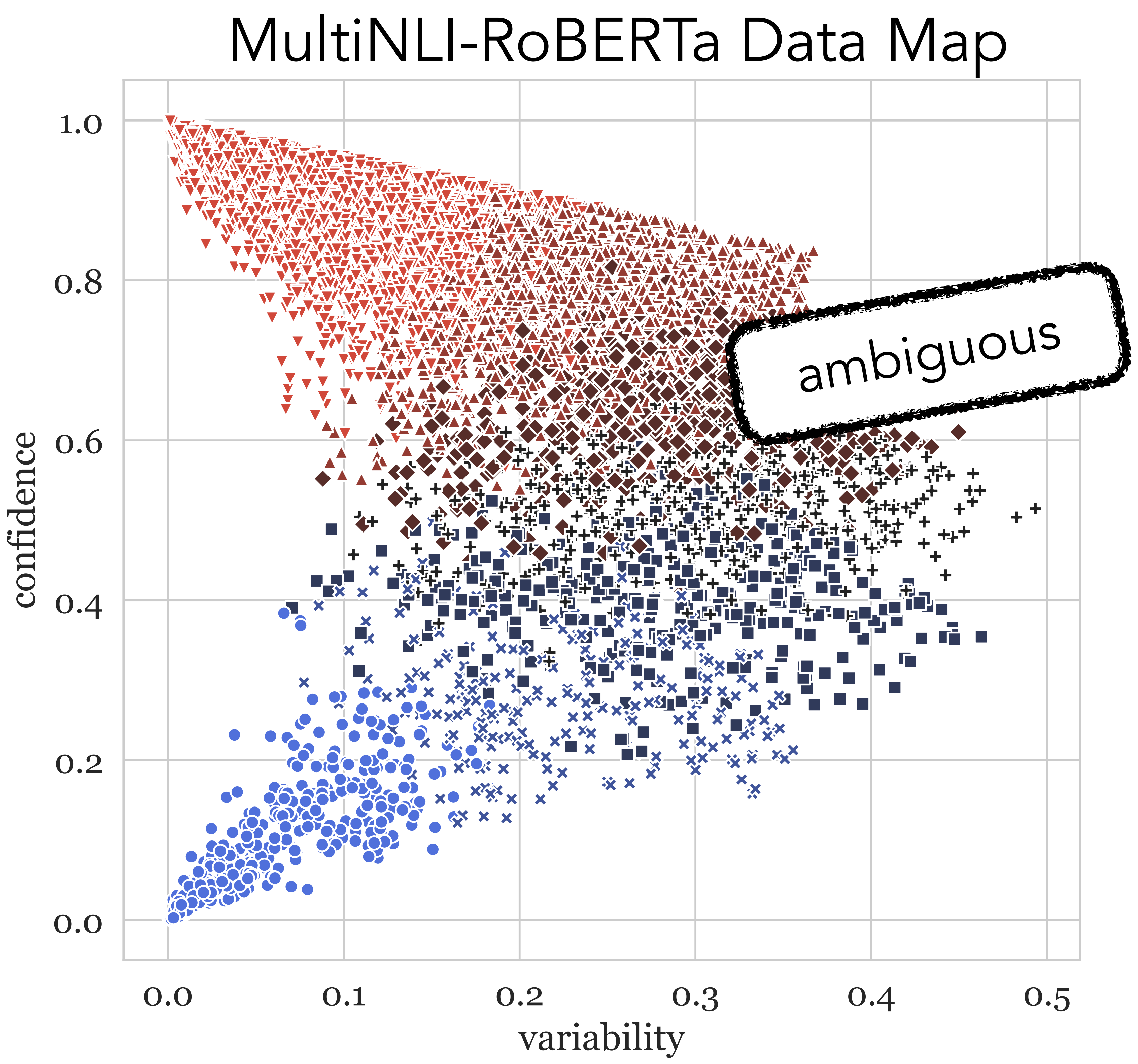
Can be easily modified for diverse generations

[Schick & Schütze, 2021; Meng et al. 2022; West et al., 2021; Lee et al., 2021; Bartolo et al., 2021]

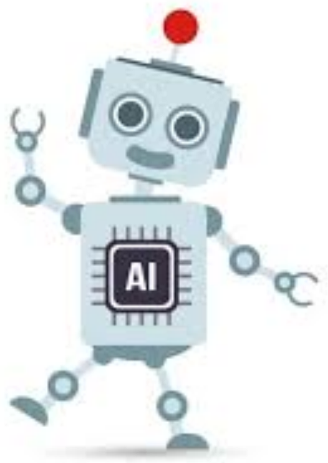
G-DAUG: Generative Data Augmentation for Commonsense Reasoning [Y. M., F., **Swayamdipta**, L., W., B., C., D EMNLP-findings, 2020]

Also see

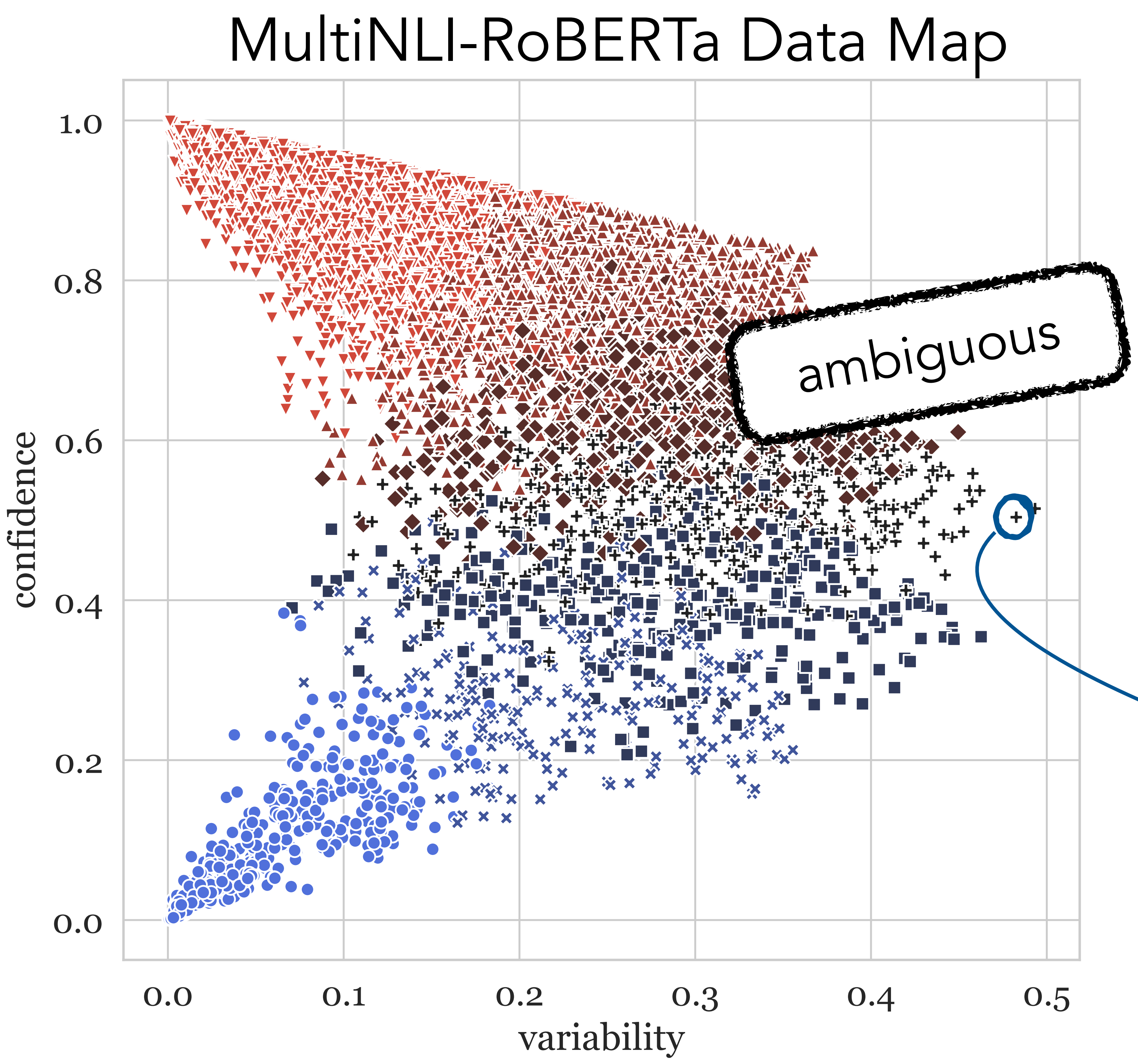
WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



GPT-3



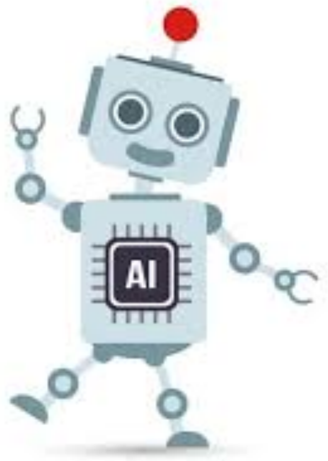
WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



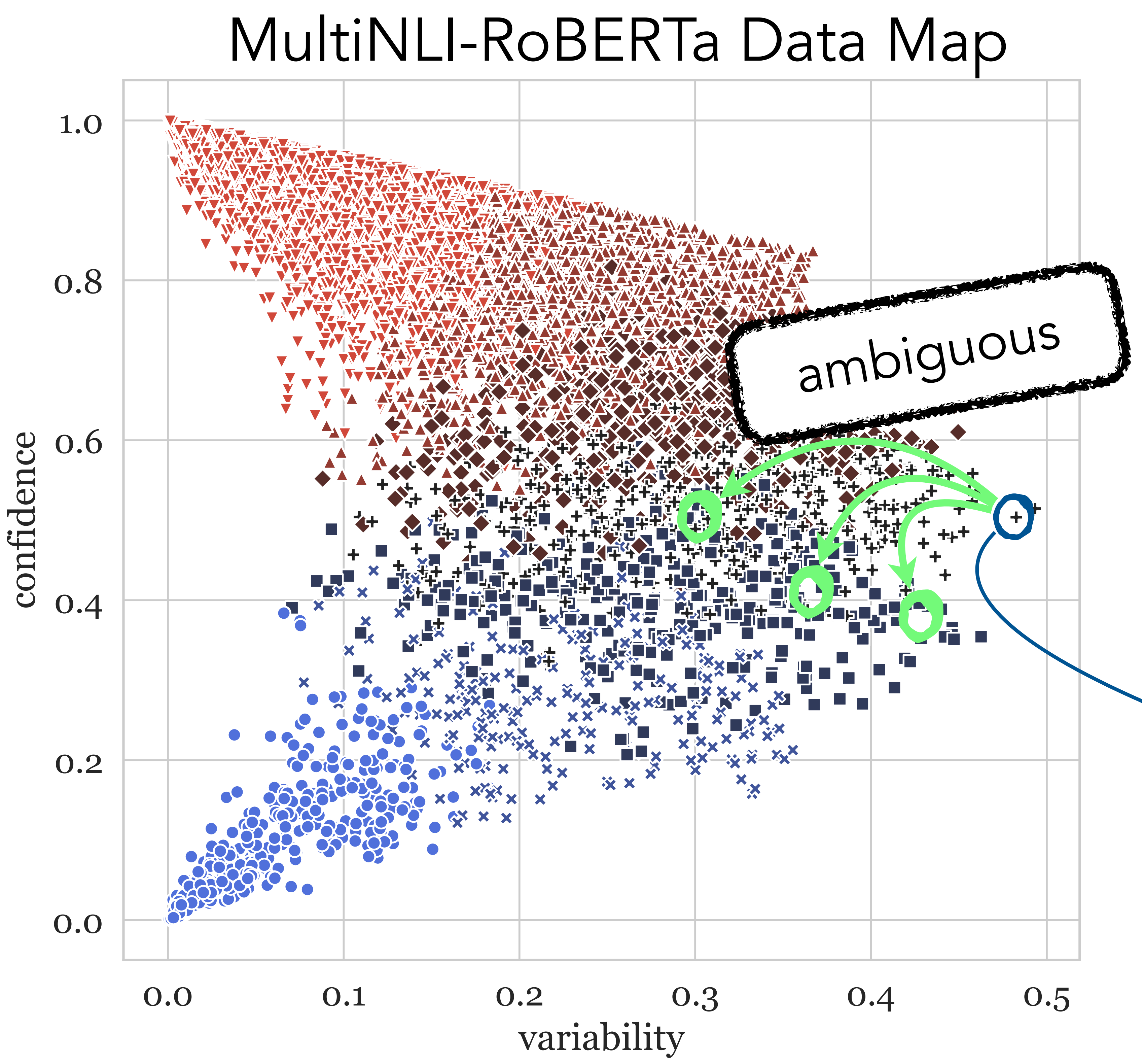
5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

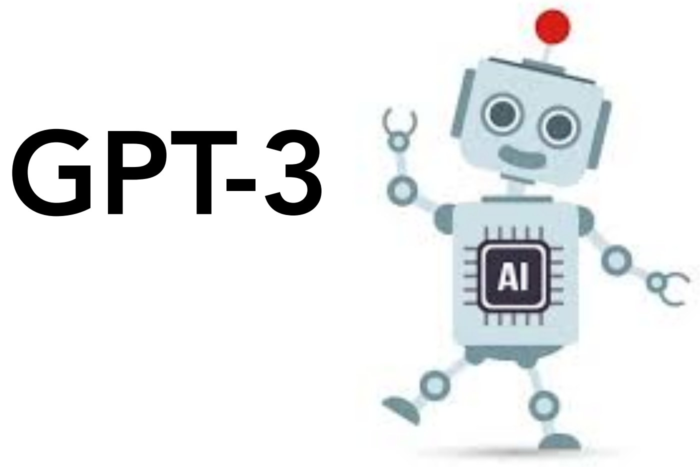
5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

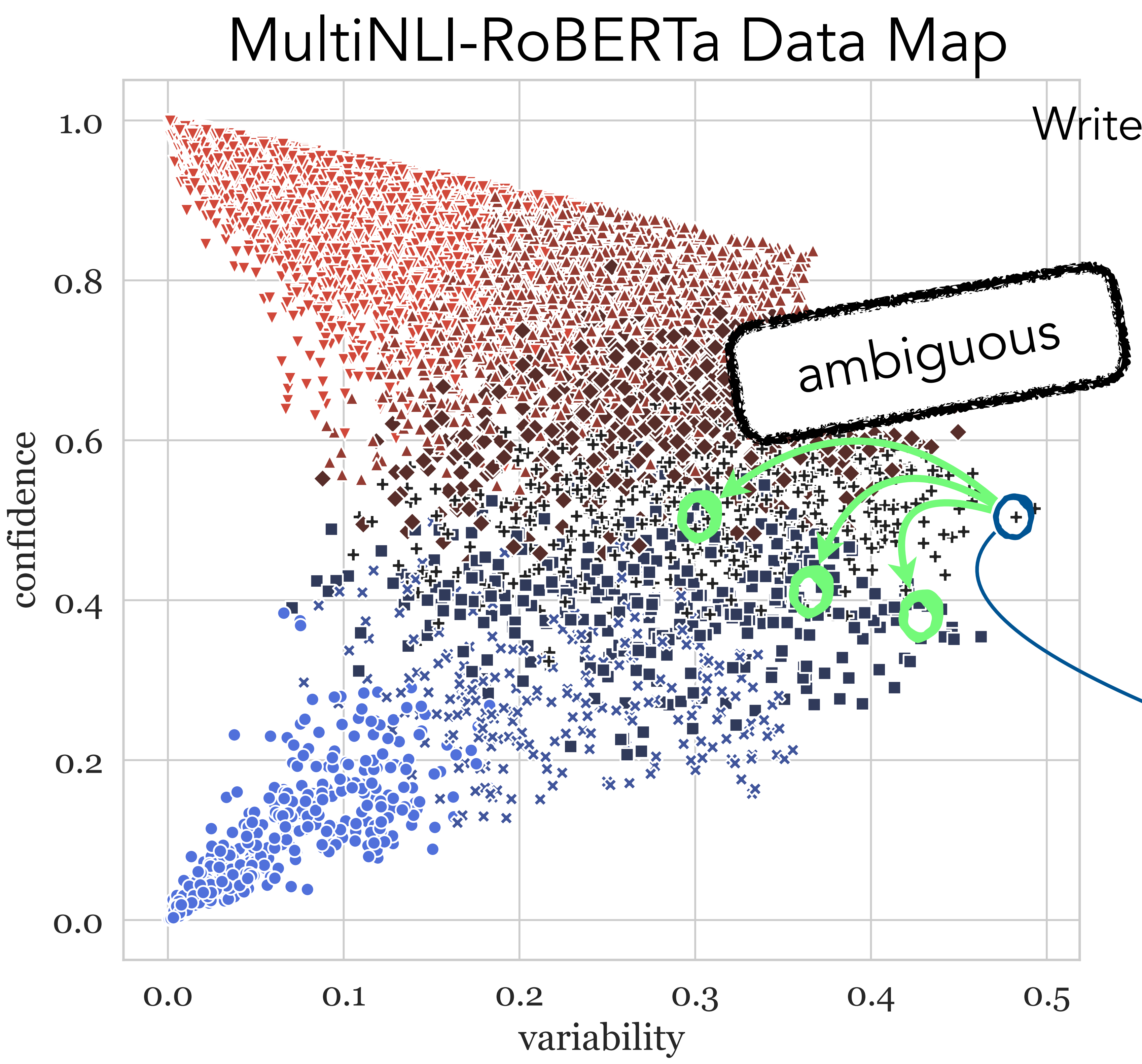
nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

seed ambiguous example from MultiNLI - RoBERTa



WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

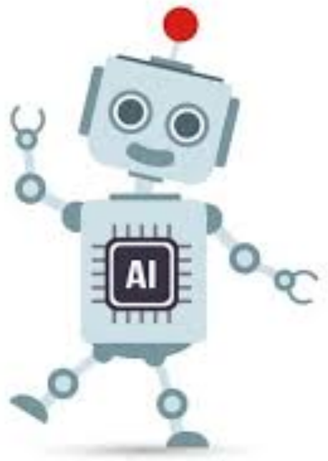
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

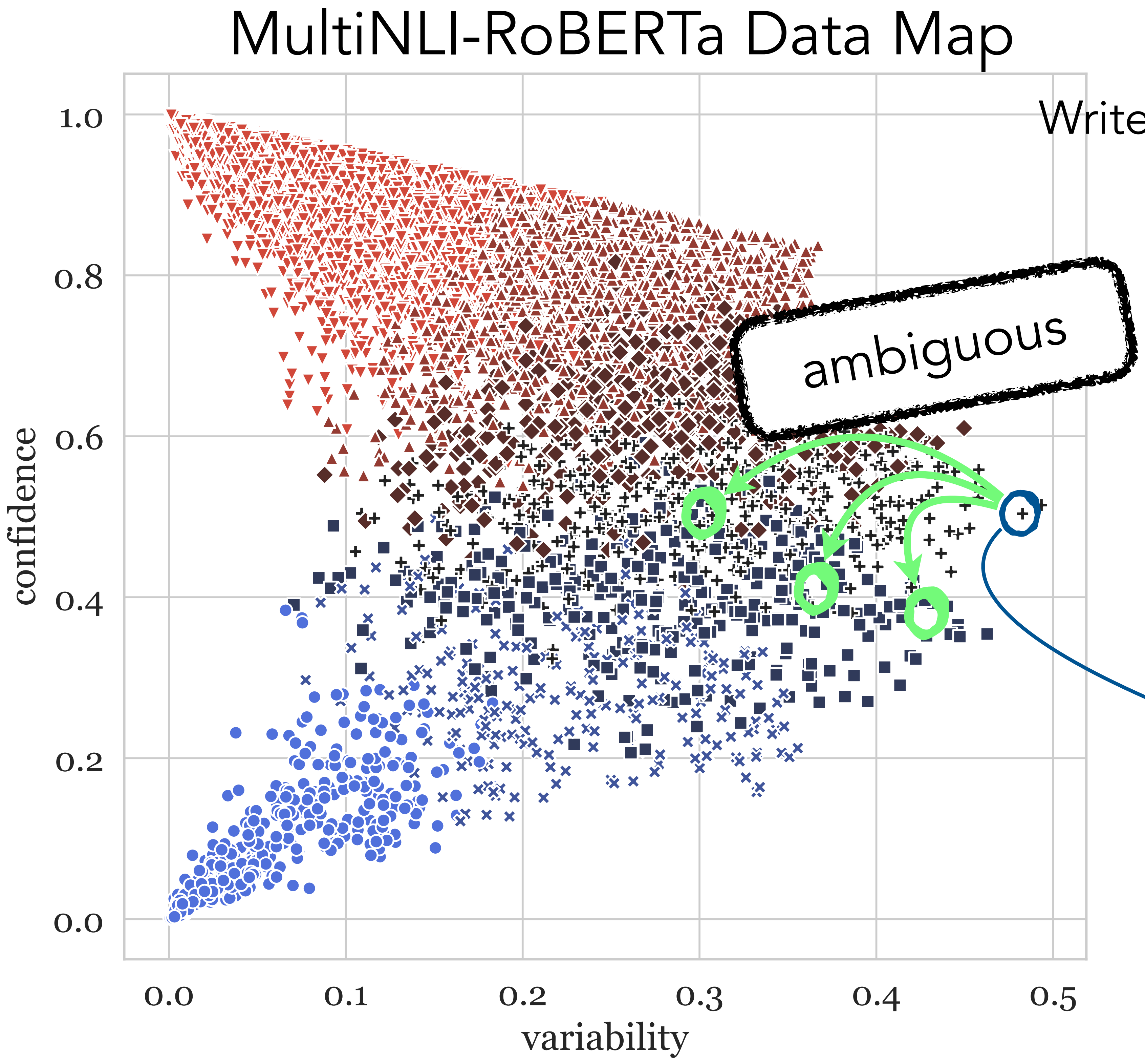
5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

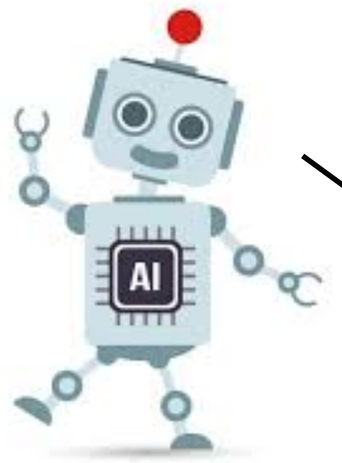
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

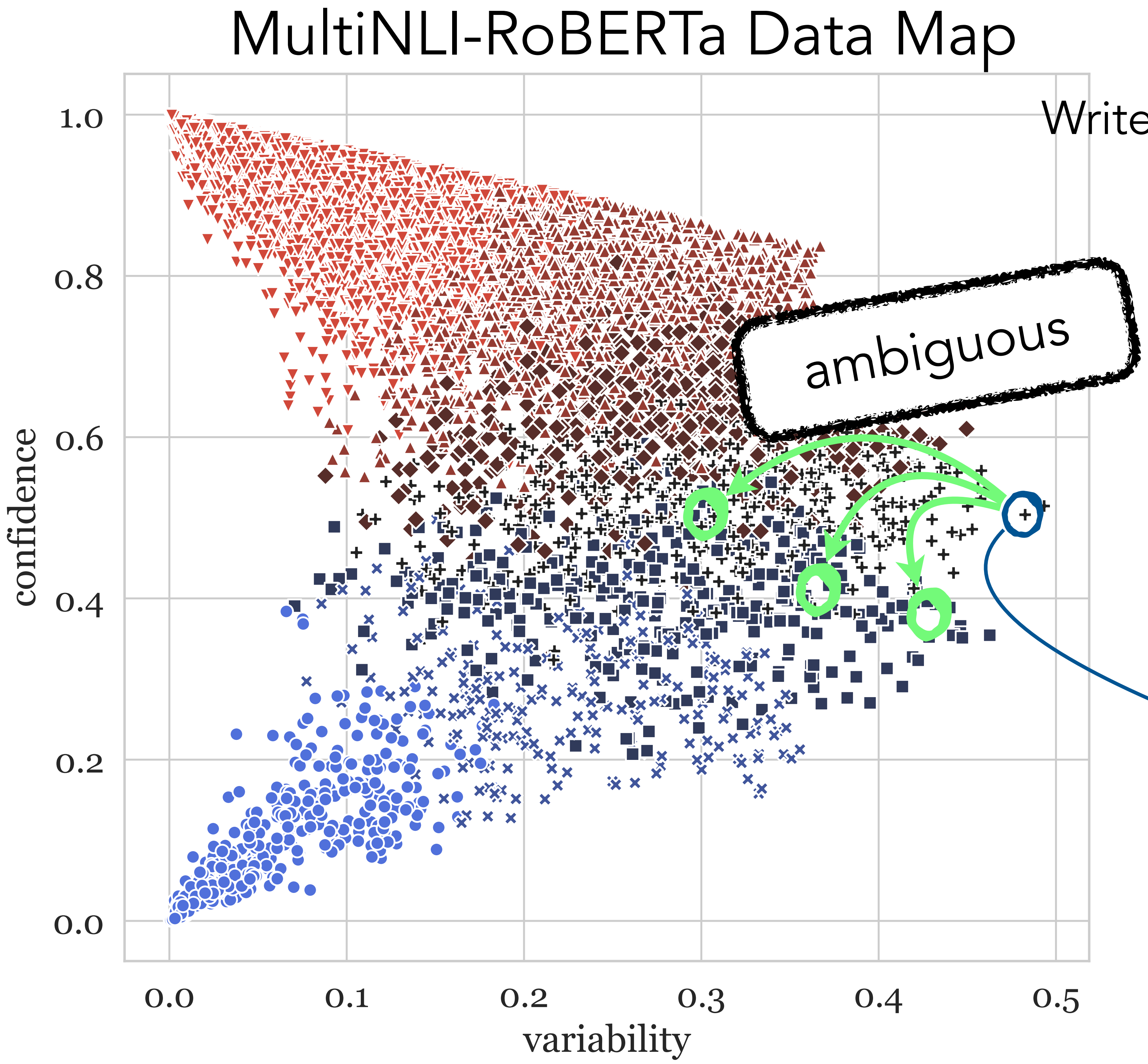
} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter****, when the big sightseeing destinations are far less crowded.*
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

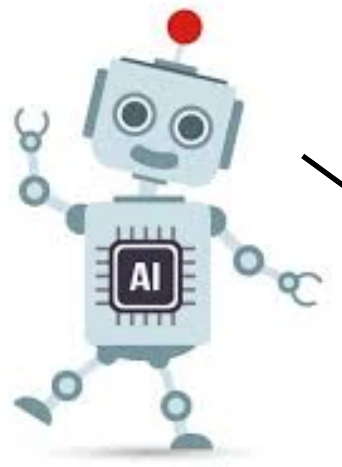
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

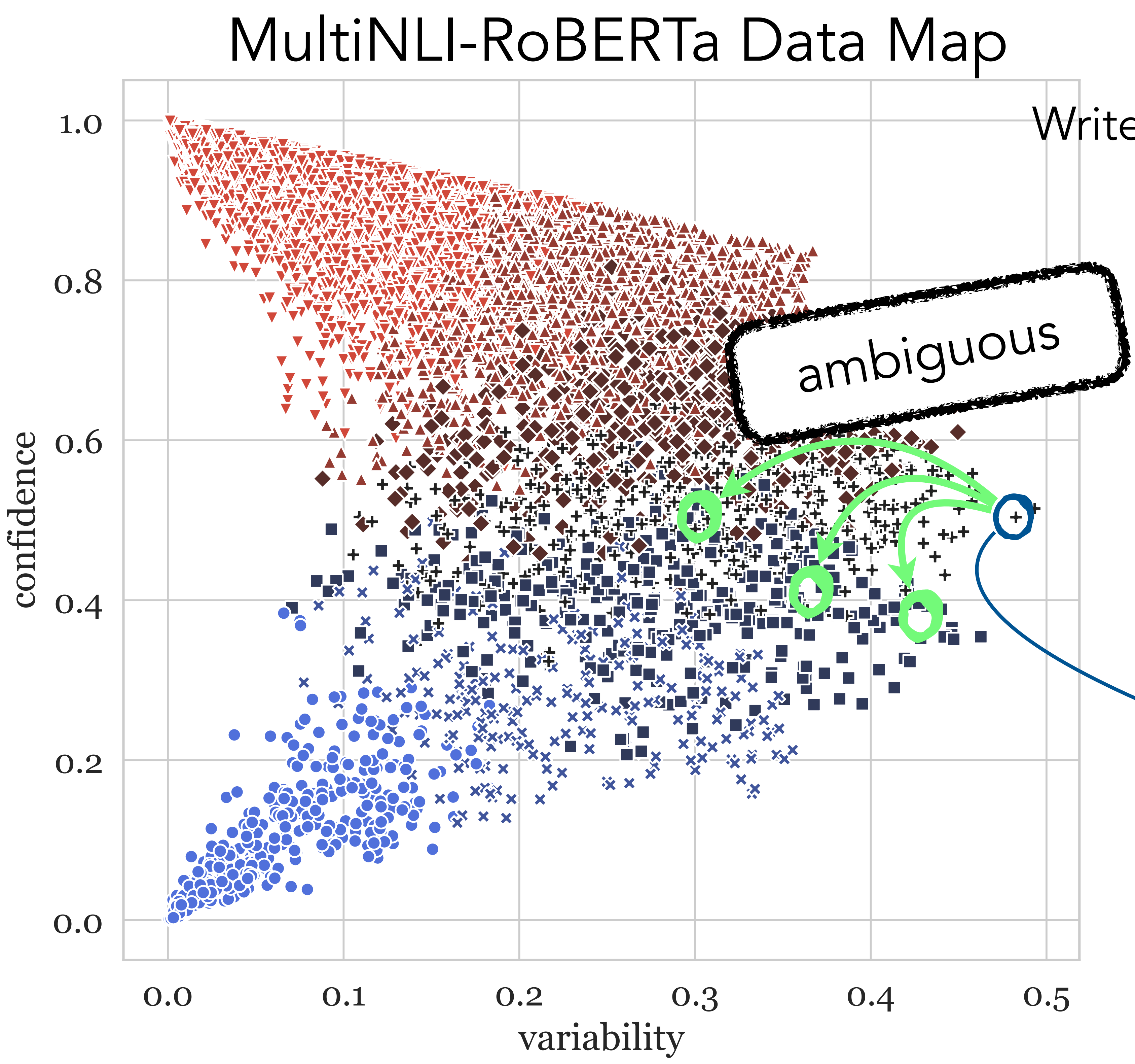
} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied. ✓

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

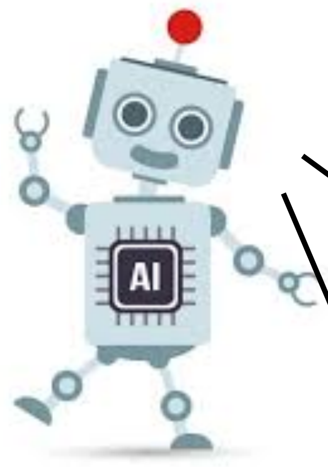
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3

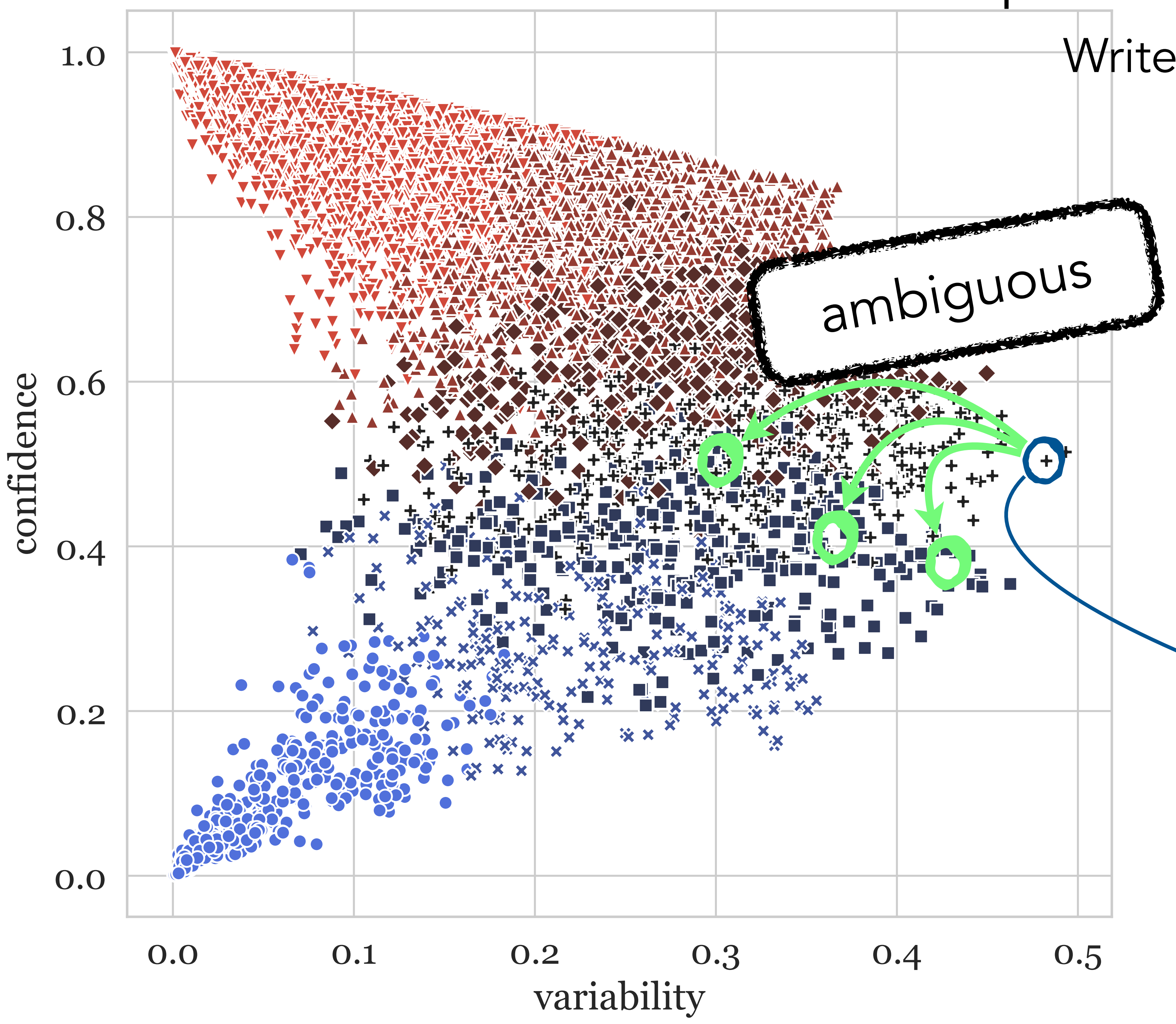


1 percent of the seats were vacant. ✓
Implication: **99 percent** of the seats were occupied.

About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

MultiNLI-RoBERTa Data Map



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

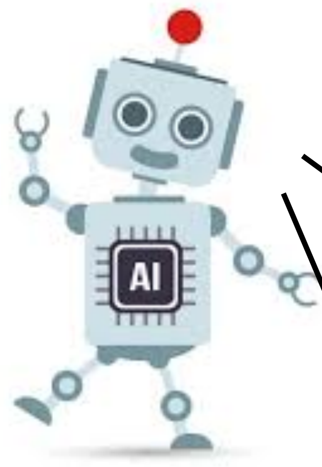
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3

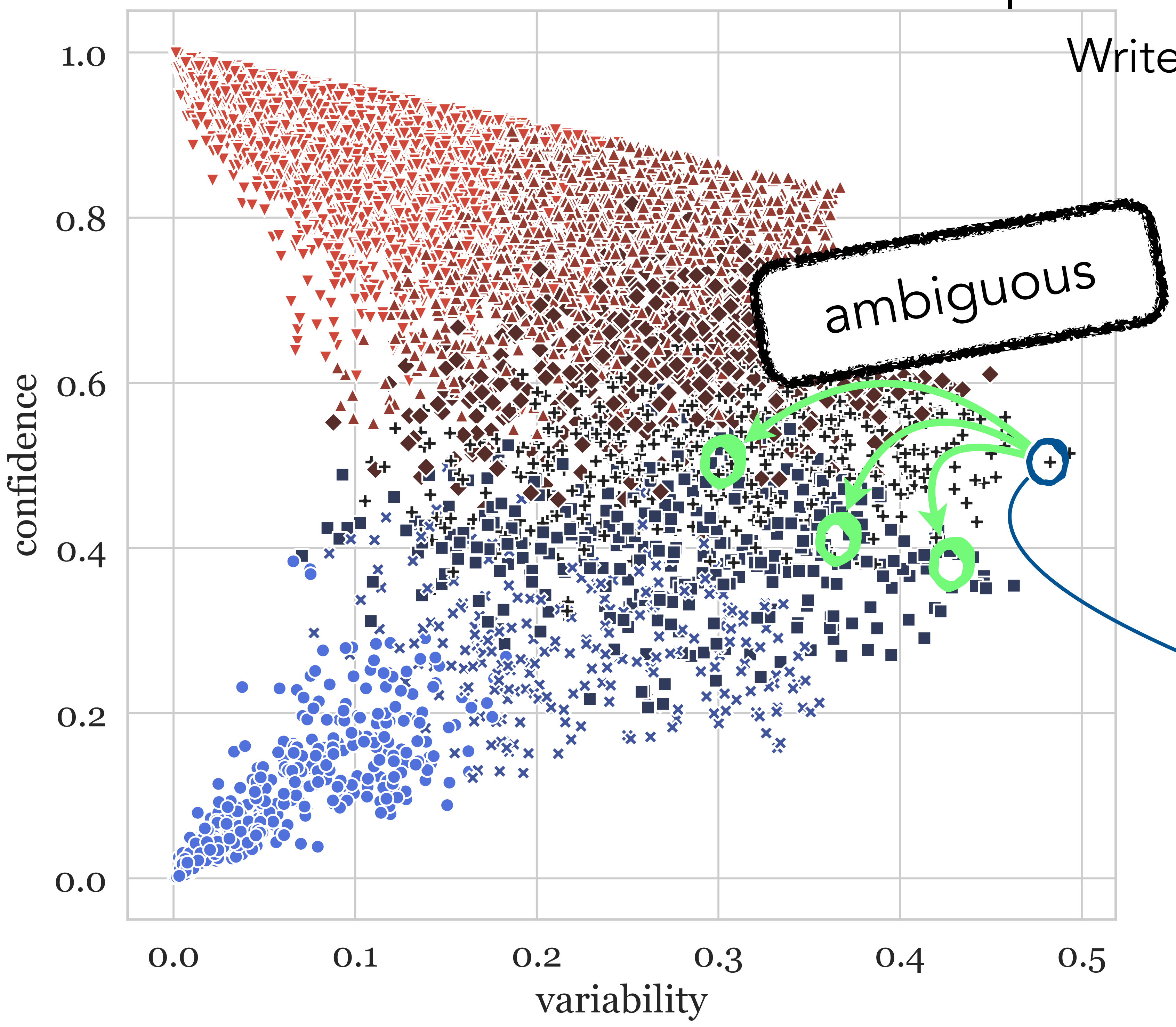


1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied. ✓

About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year. ⊖

Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

MultiNLI-RoBERTa Data Map



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

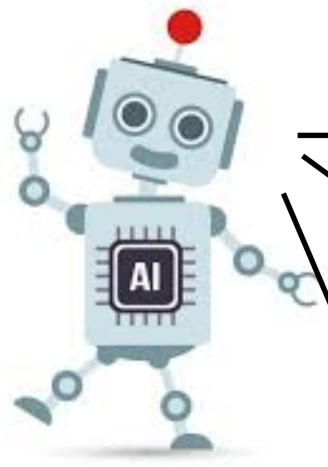
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



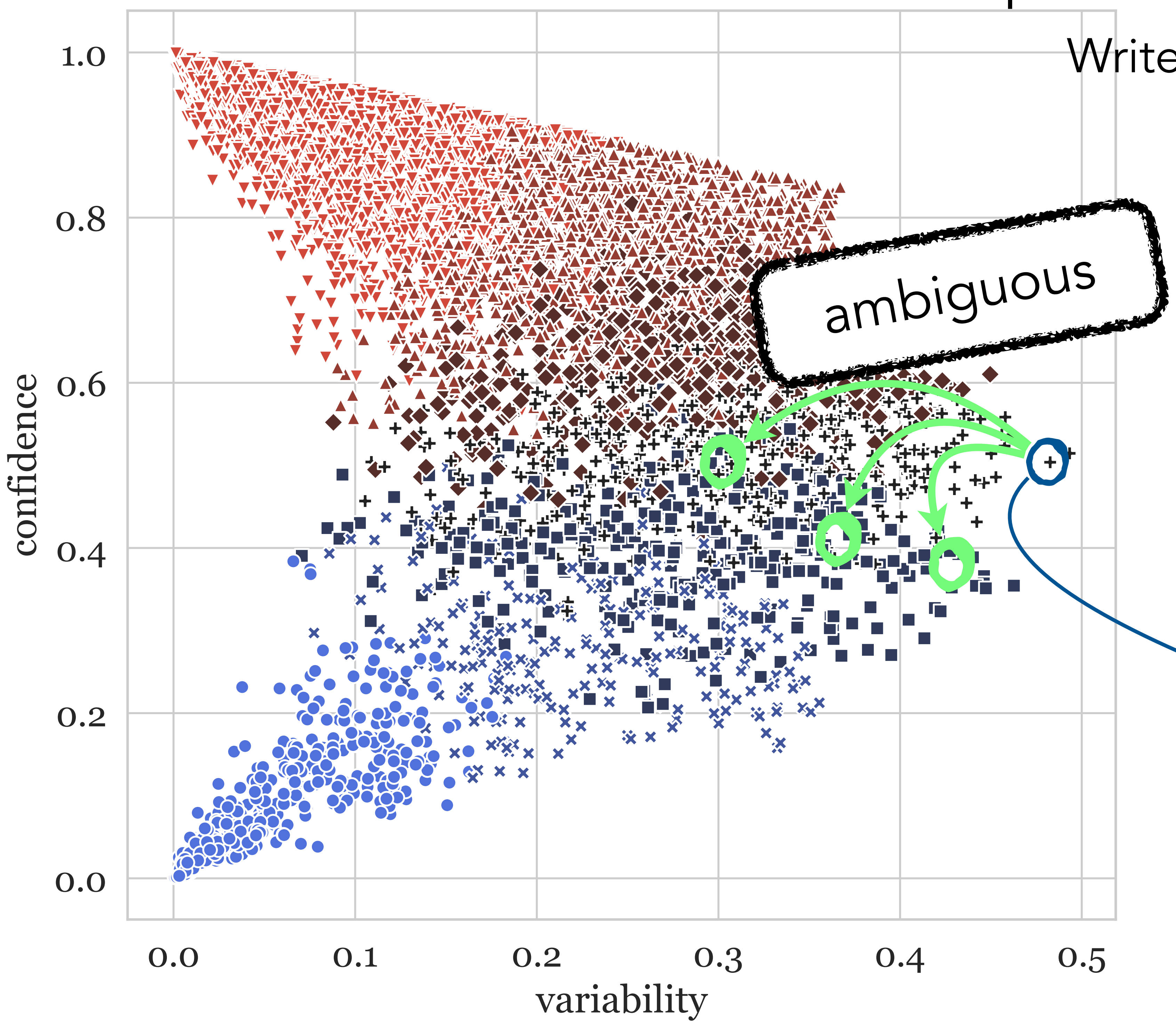
He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied. ✓

About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year. ⊖

Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

MultiNLI-RoBERTa Data Map



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 percent*** of routes are operating at either profit or break-even.

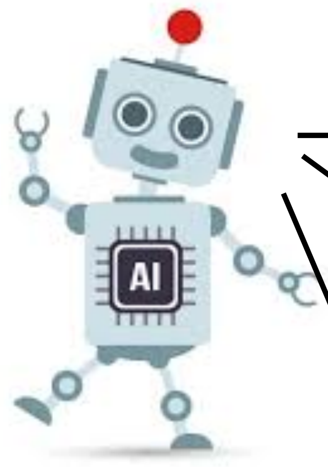
30 About ***10 percent*** of households did not
Implication: Roughly ***ninety percent*** of households did this thing.

} nearest neighbors to seed example

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3

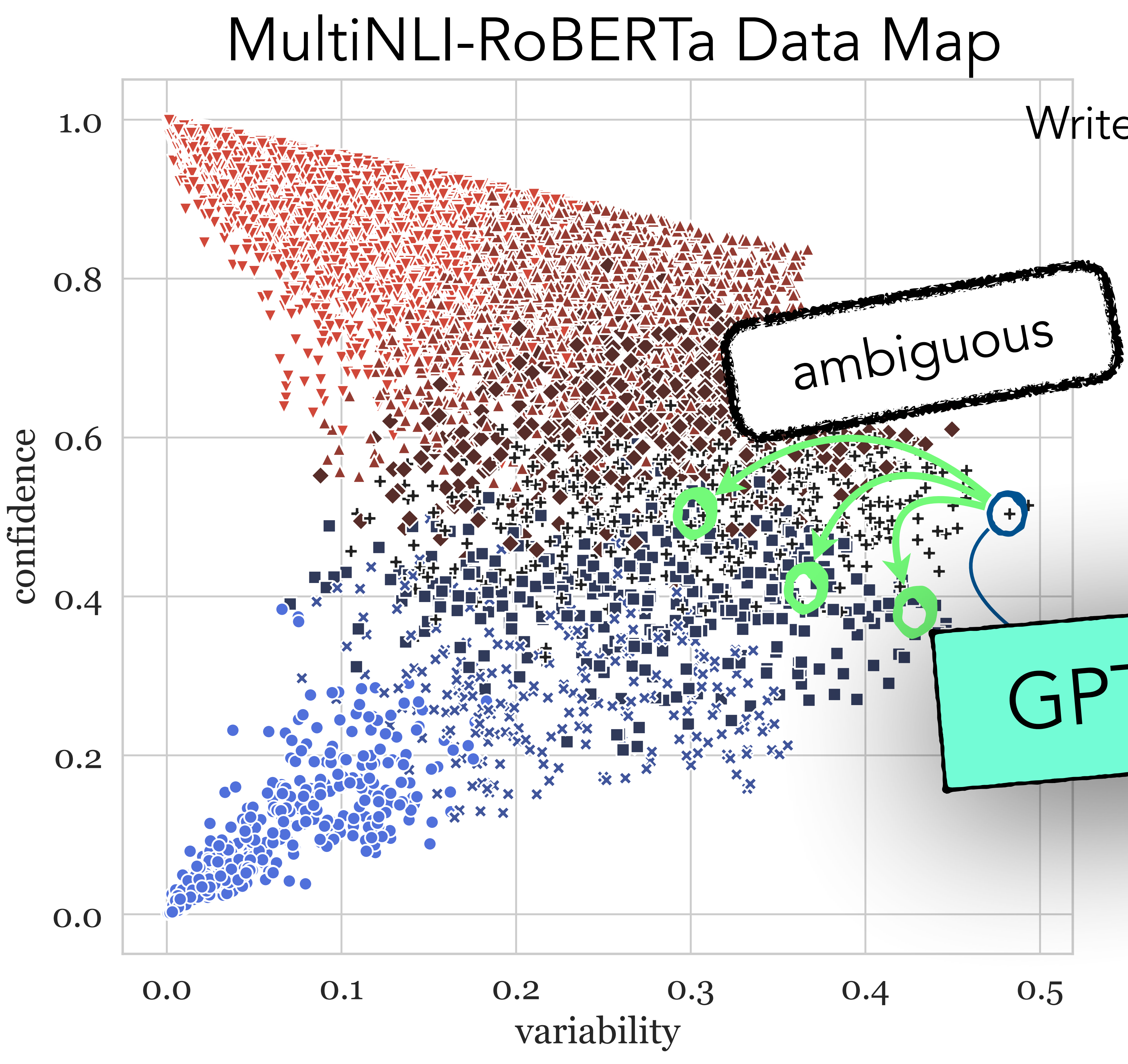


He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank. ✗

1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied. ✓

About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year. ⊖

Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Write a pair of sentences that have the same relationship as the previous examples. Examples: } instruction

But if it's at all possible, plan your visit for the ***spring, autumn, or even the winter***, when the big sightseeing destinations are far less crowded.
Implication: This destination is most crowded in the ***summer***.

5 percent of the routes operating at a loss.
Implication: ***95 pe*** … fit or break-even.

… not
… nplication: Roughly ***ninety percent*** of households did this thing.

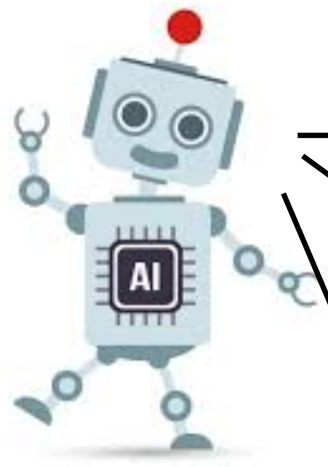
} nearest neighbors to seed example

GPT-3 generations are not always reliable

5 percent probability that each part will be defect free.
Implication: Each part has a ***95 percent*** chance of having a defect.

} seed ambiguous example from MultiNLI - RoBERTa

GPT-3



✗
He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

✓
1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

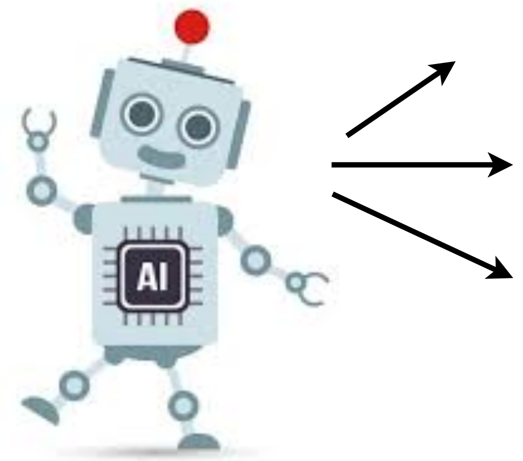
⊖
About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

iu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

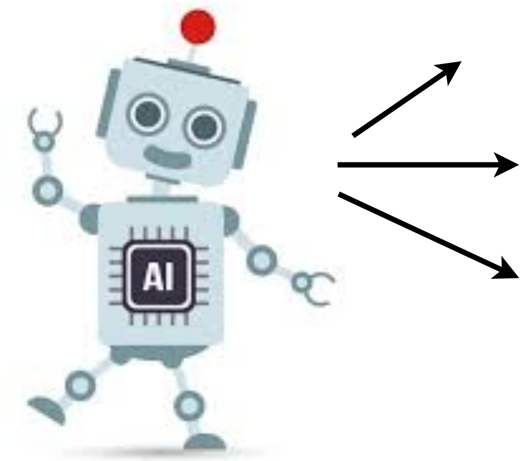
✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

Also see

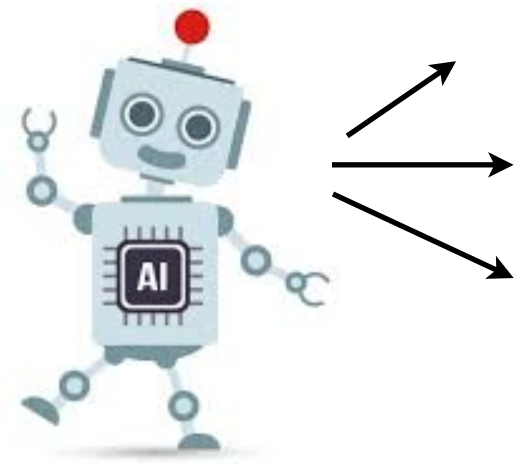
Reframing Human-AI for Generating Free-Text Explanations
[Wiegreffe, Hessel, **Swayamdipta**, Riedel & Choi, NAACL 2022]

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



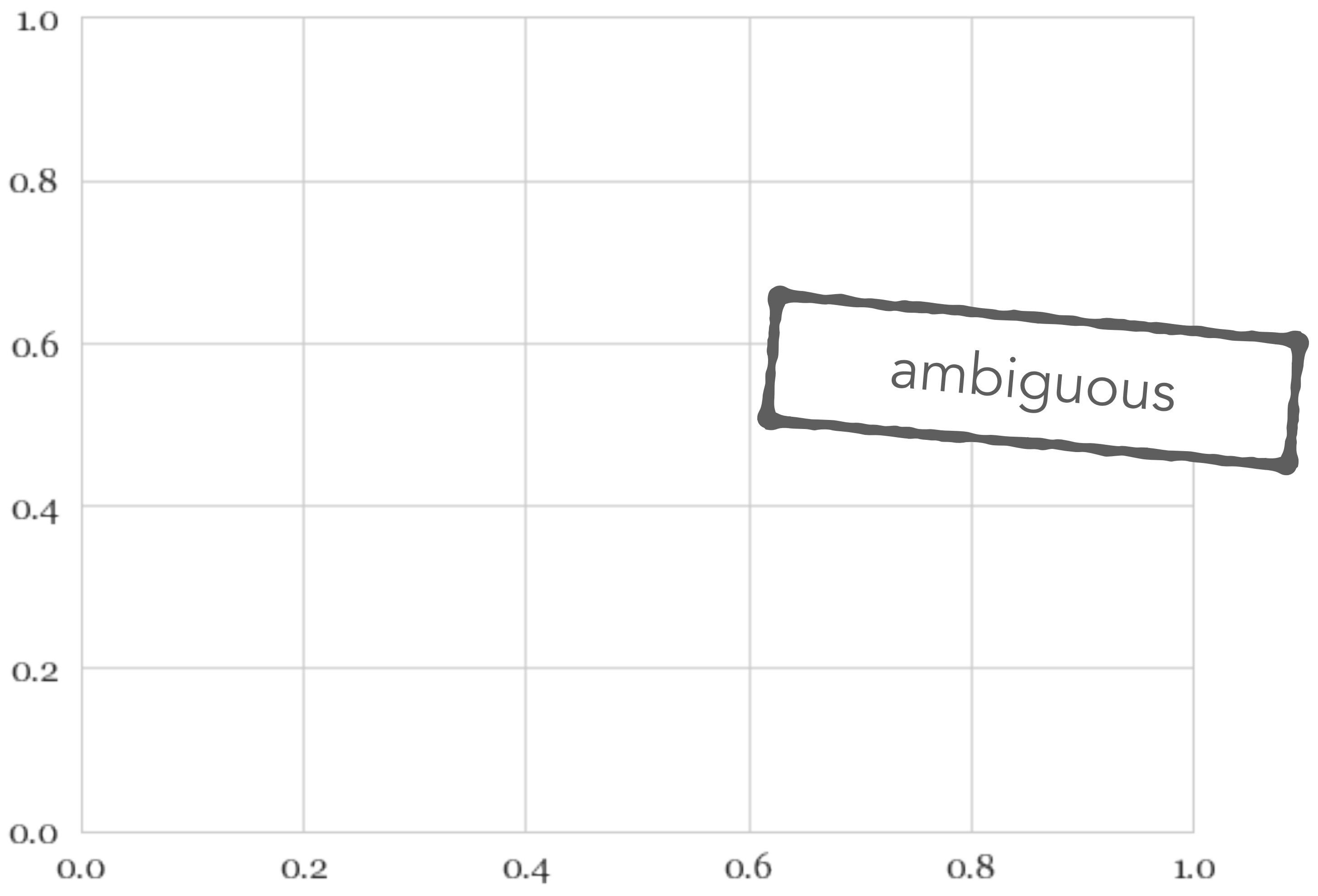
GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

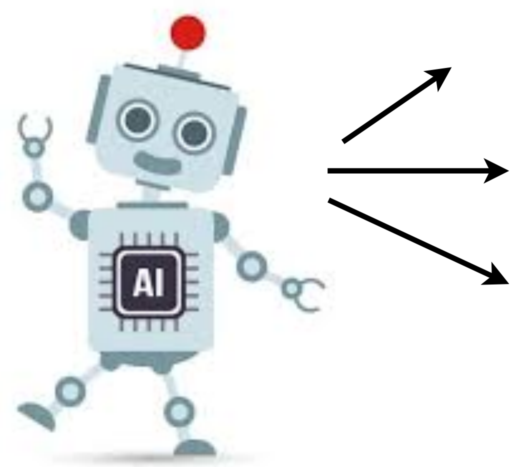


WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



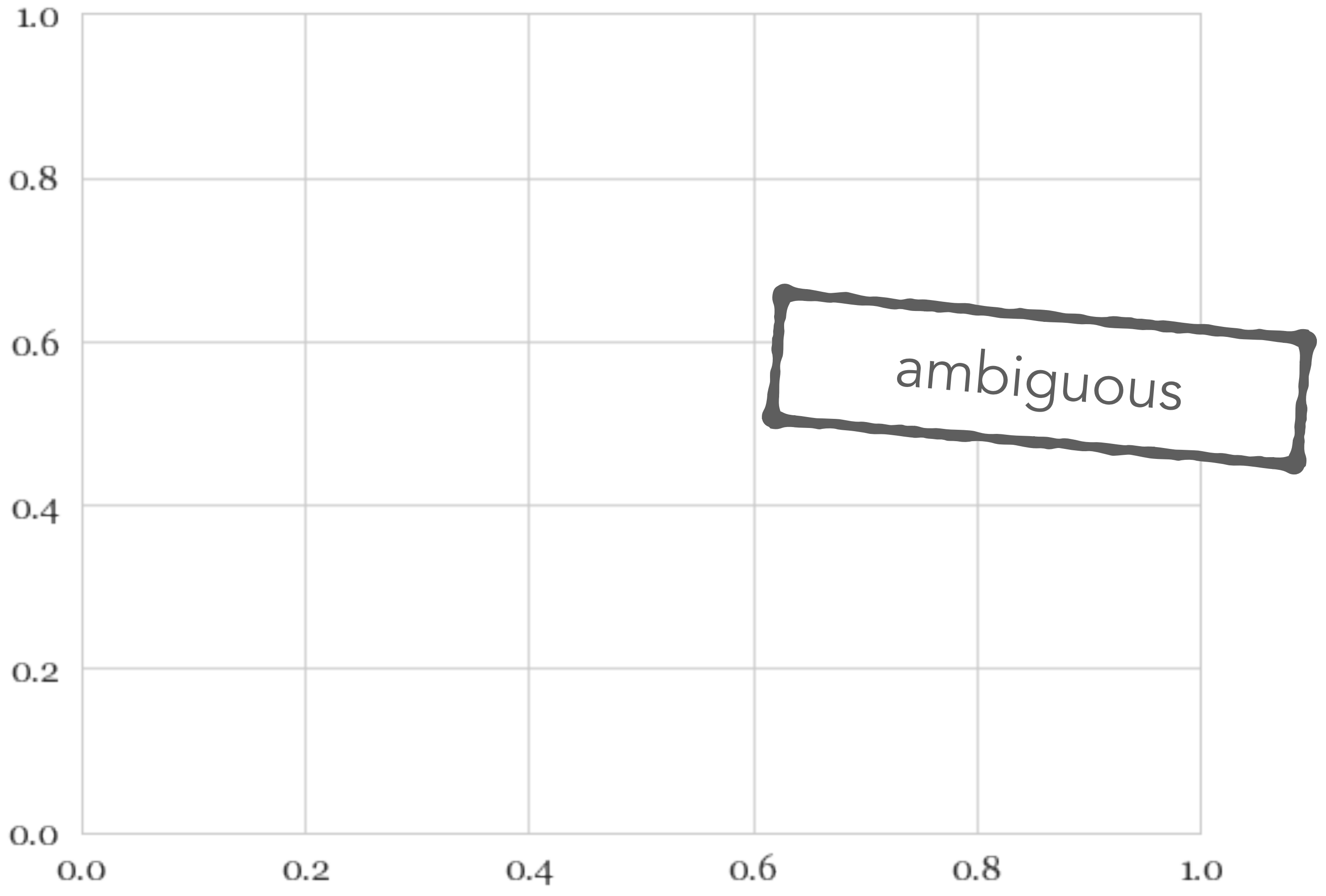
GPT-3

About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.



variability

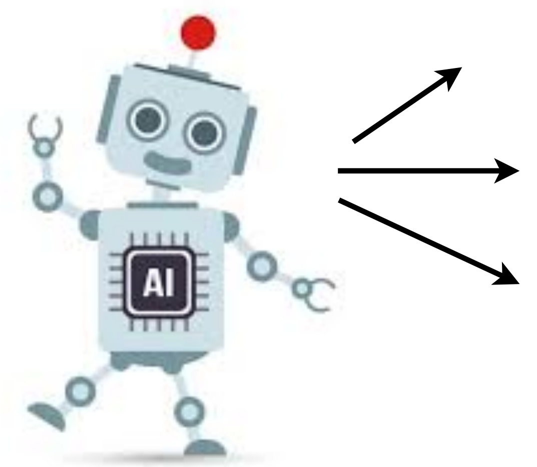
Standard deviation of the **true class** probability

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



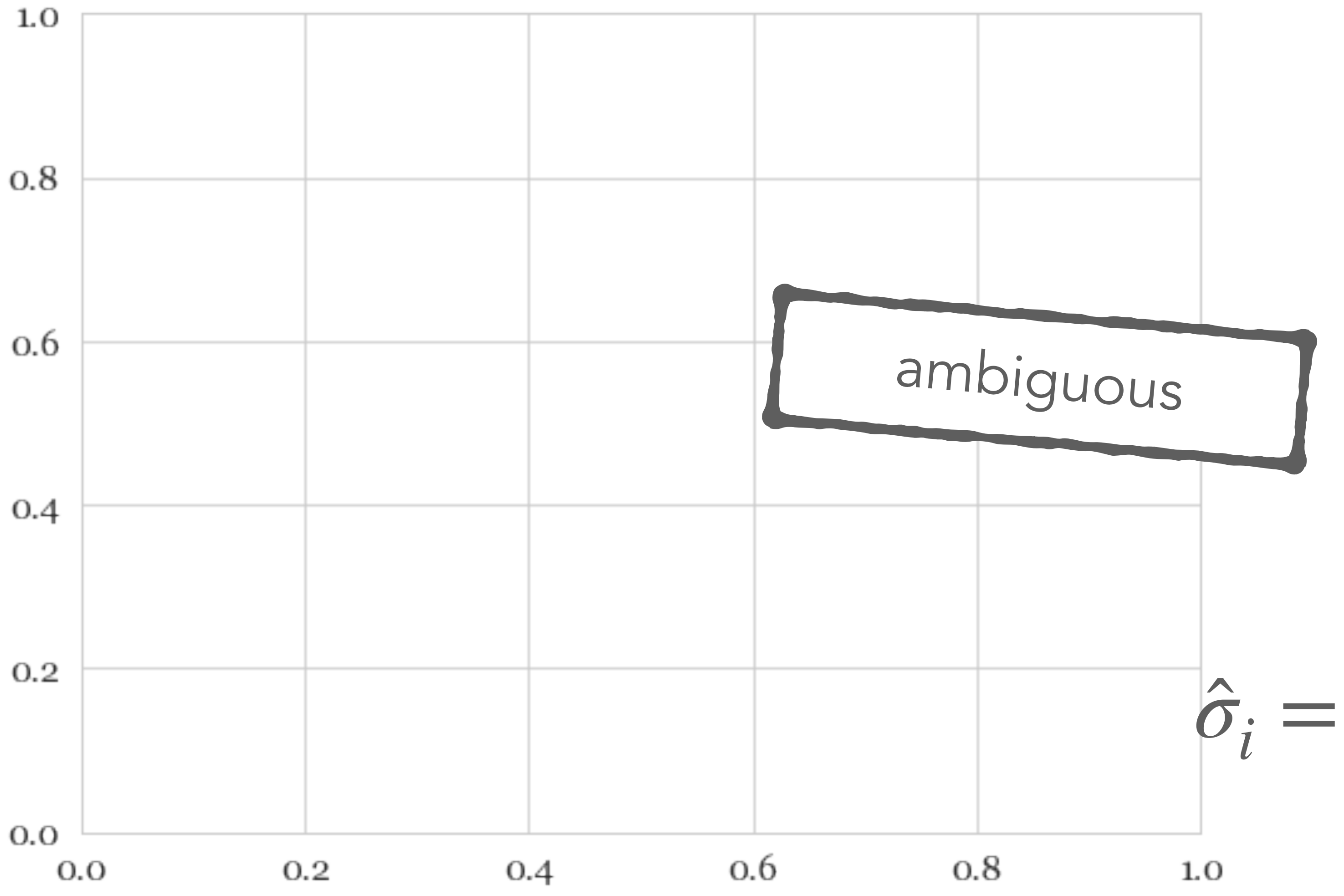
GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.



variability

$$\hat{\sigma}_i = \max_{y\in\mathcal{Y}} \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\boldsymbol{\theta}^{(e)}}(y \mid x_i) - \hat{\mu}_{i,y}\right)^2}{E}}$$

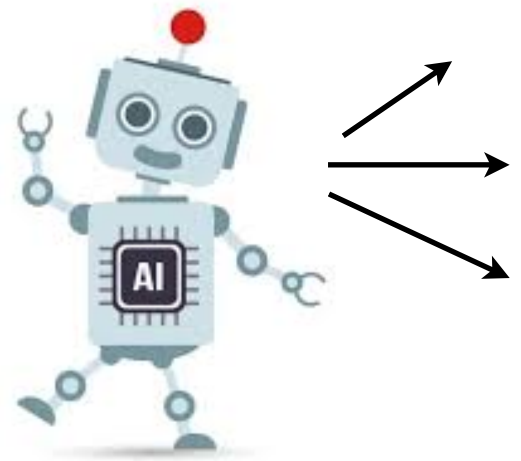
Standard deviation of the **max** ~~**true**~~ **class** probability

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



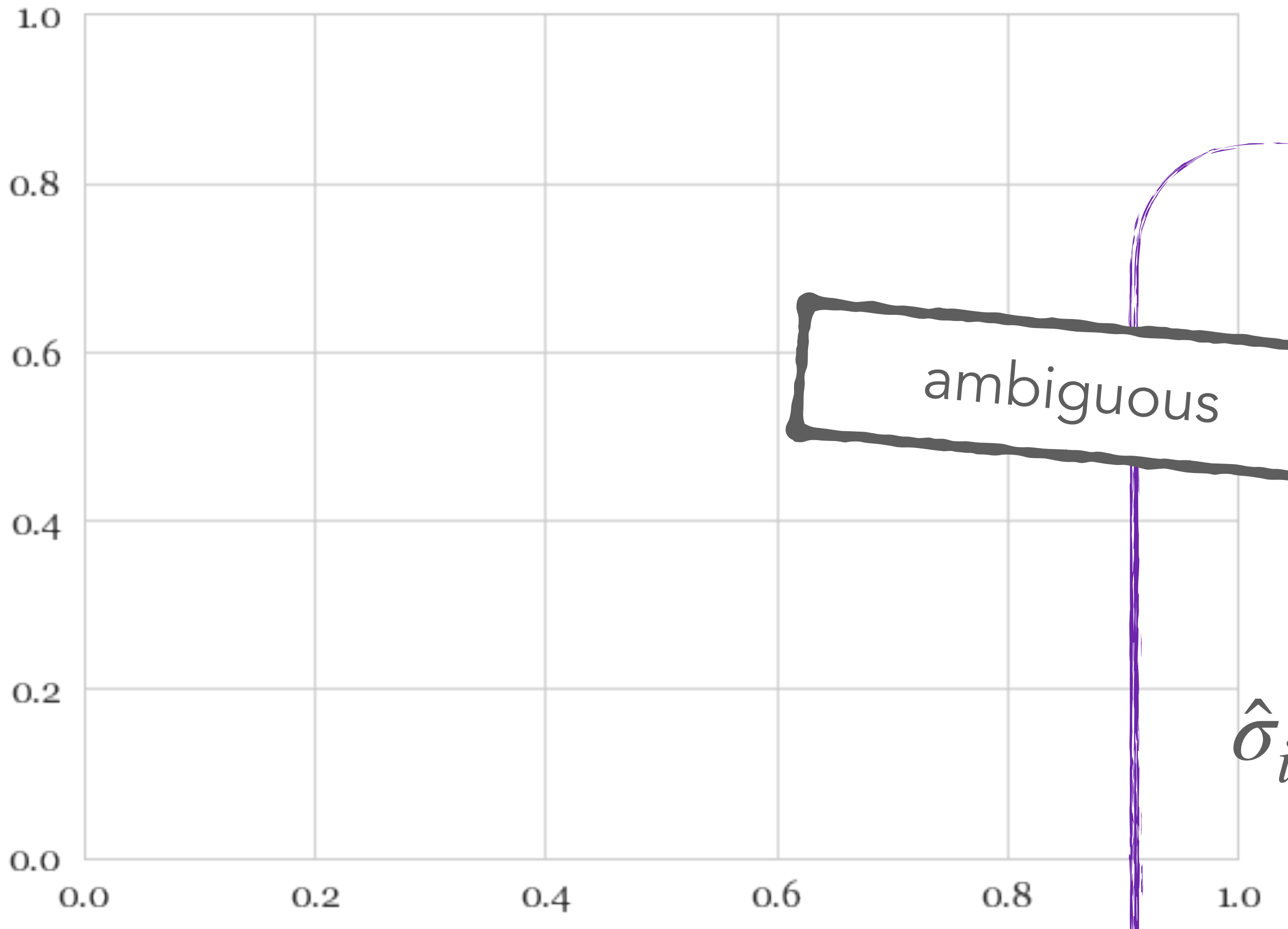
GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.



expected worst-case ambiguity

$$\hat{\sigma}_i = \max_{y\in\mathcal{Y}} \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\boldsymbol{\theta}^{(e)}}(y \mid x_i) - \hat{\mu}_{i,y}\right)^2}{E}}$$

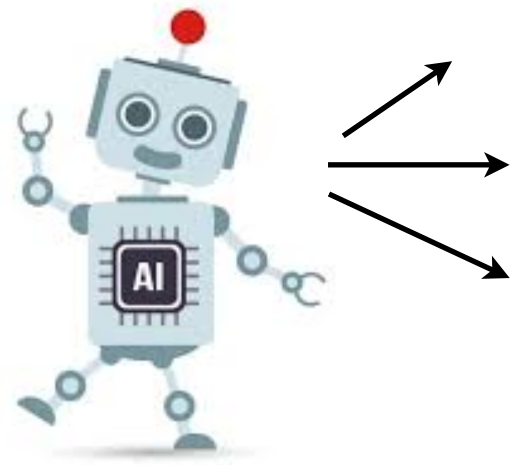
Standard deviation of the **max** ~~**true**~~ **class** probability

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



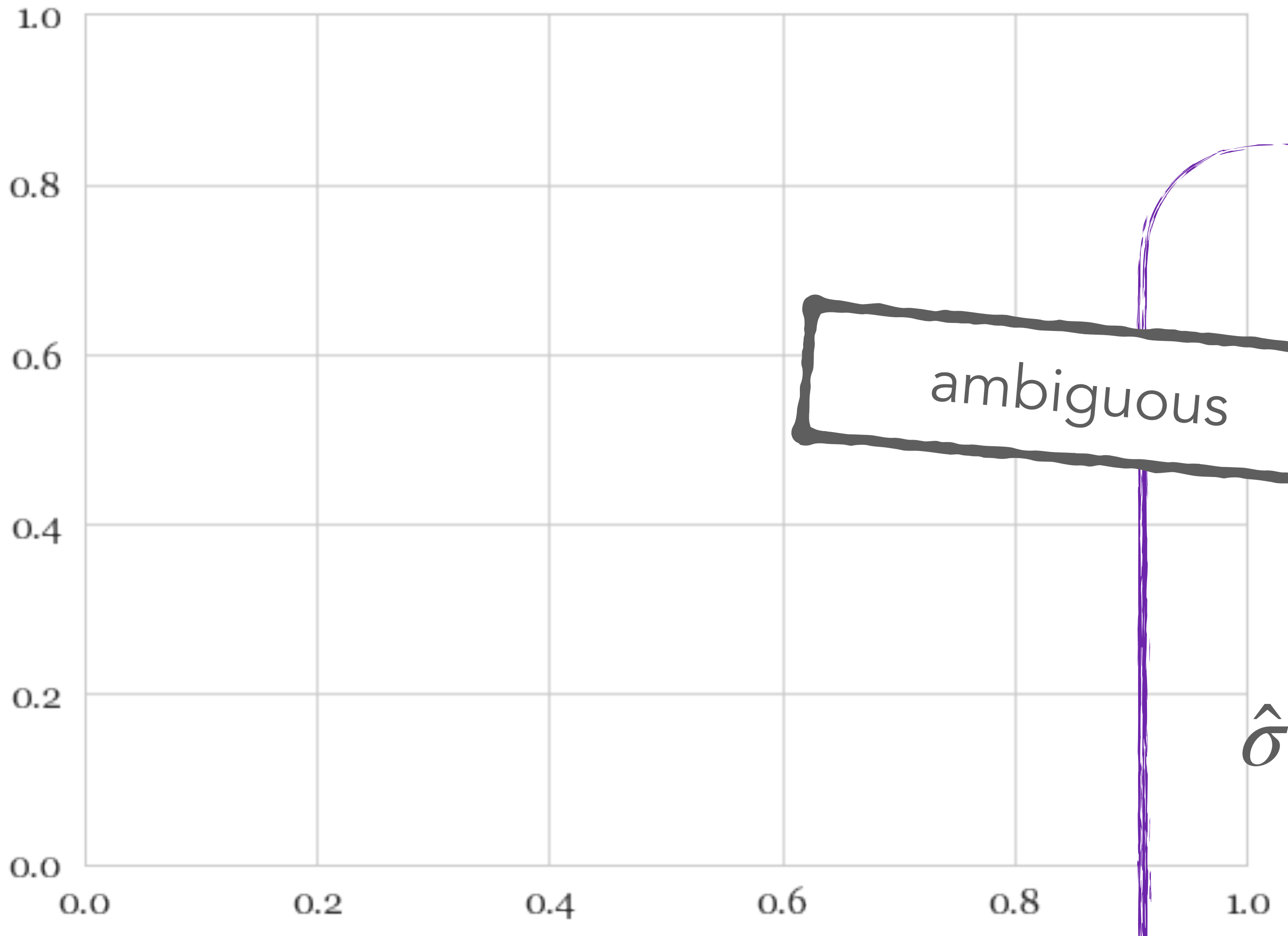
GPT-3

⊖ About **1,000** people are diagnosed with chronic myeloid leukemia each year.
Implication: About **9,000** people are not diagnosed with chronic myeloid leukemia each year.

✗ He has never smoked, and he doesn't drink.
Implication: He has smoked and he has drank.

Filter

✓ **1 percent** of the seats were vacant.
Implication: **99 percent** of the seats were occupied.



expected worst-case ambiguity

$$\hat{\sigma}_i = \max_{y \in \mathcal{Y}} \sqrt{\frac{\sum_{e=1}^{E} \left(p_{\boldsymbol{\theta}^{(e)}}(y \mid x_i) - \hat{\mu}_{i,y}\right)^2}{E}}$$

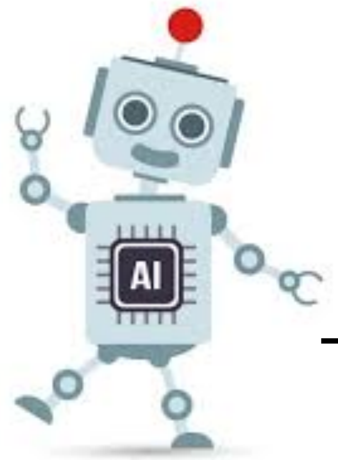
Standard deviation of the **max** ~~**true**~~ **class** probability

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



GPT-3

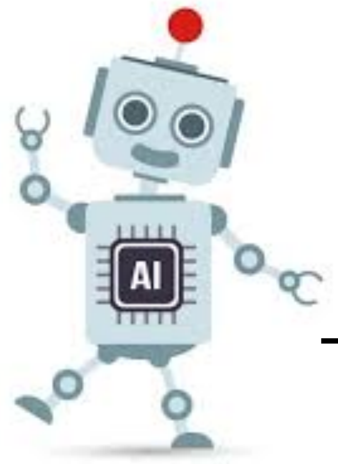
1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa



GPT-3

1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

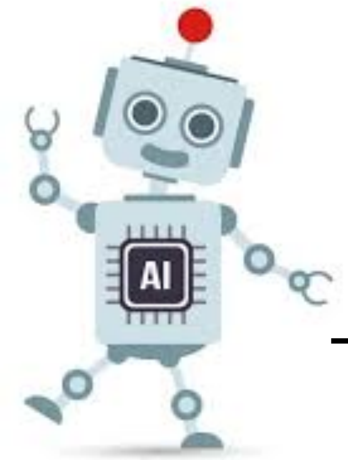
Labels?

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

instruction

nearest neighbors to seed example

seed ambiguous example from MultiNLI - RoBERTa

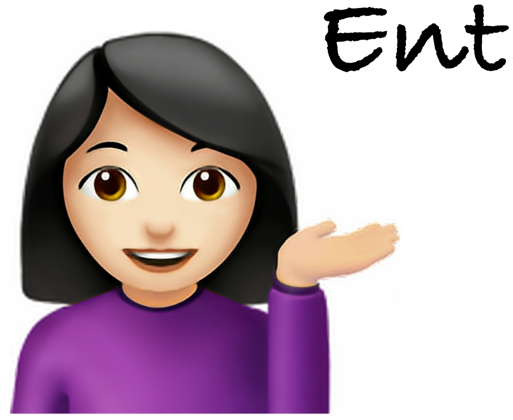


GPT-3

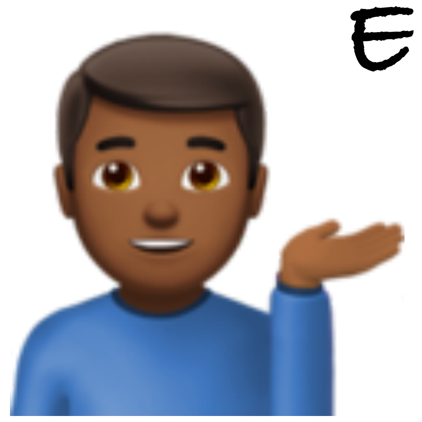
1 percent of the seats were vacant.
Implication: **99 percent** of the seats were occupied.

Labels?

Reliable and trustworthy!



Entailment



Entailment

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

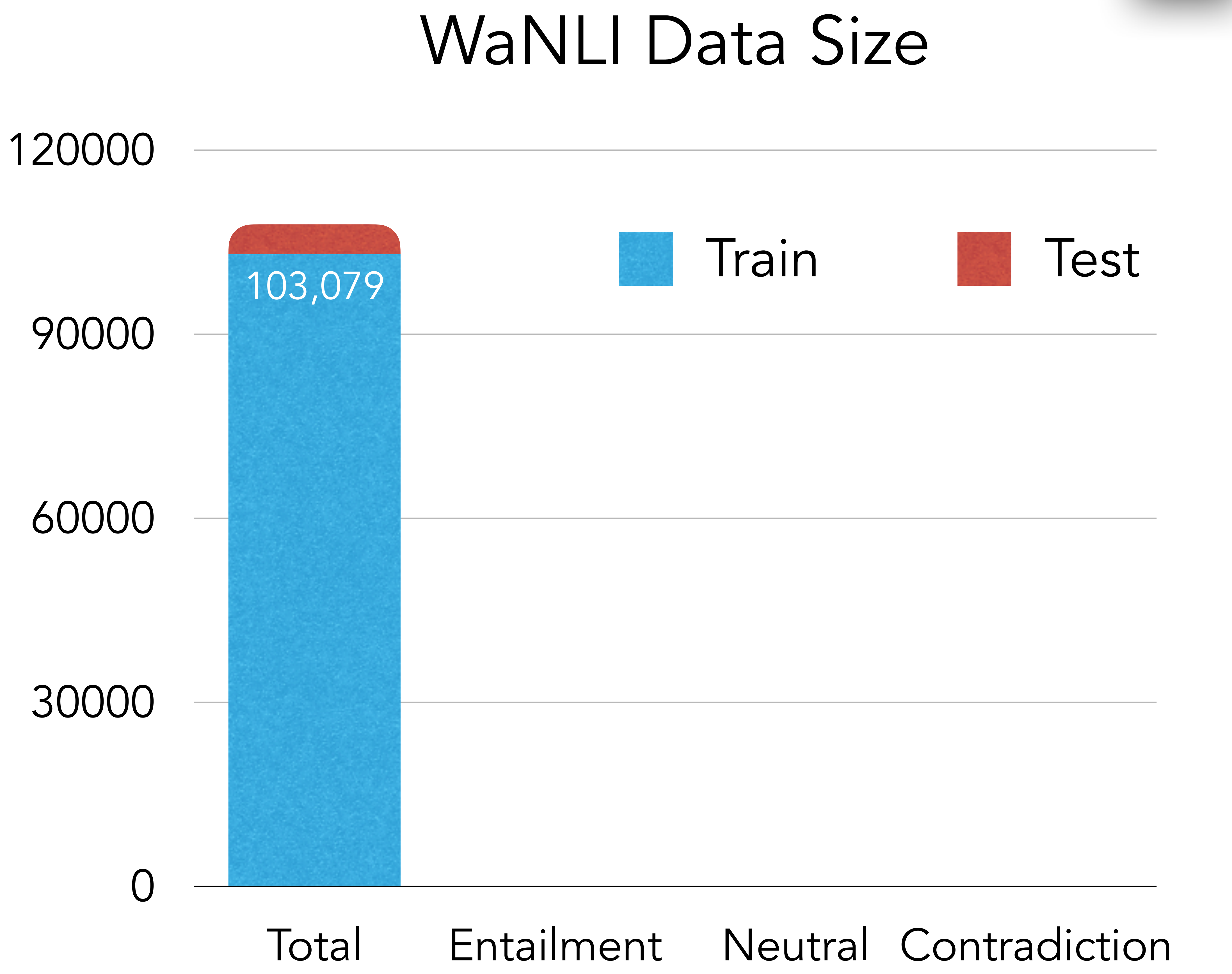
Worker-AI Collaborative NLI: WANLI

万理 Ten thousand reasoning

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

Worker-AI Collaborative NLI: WANLI

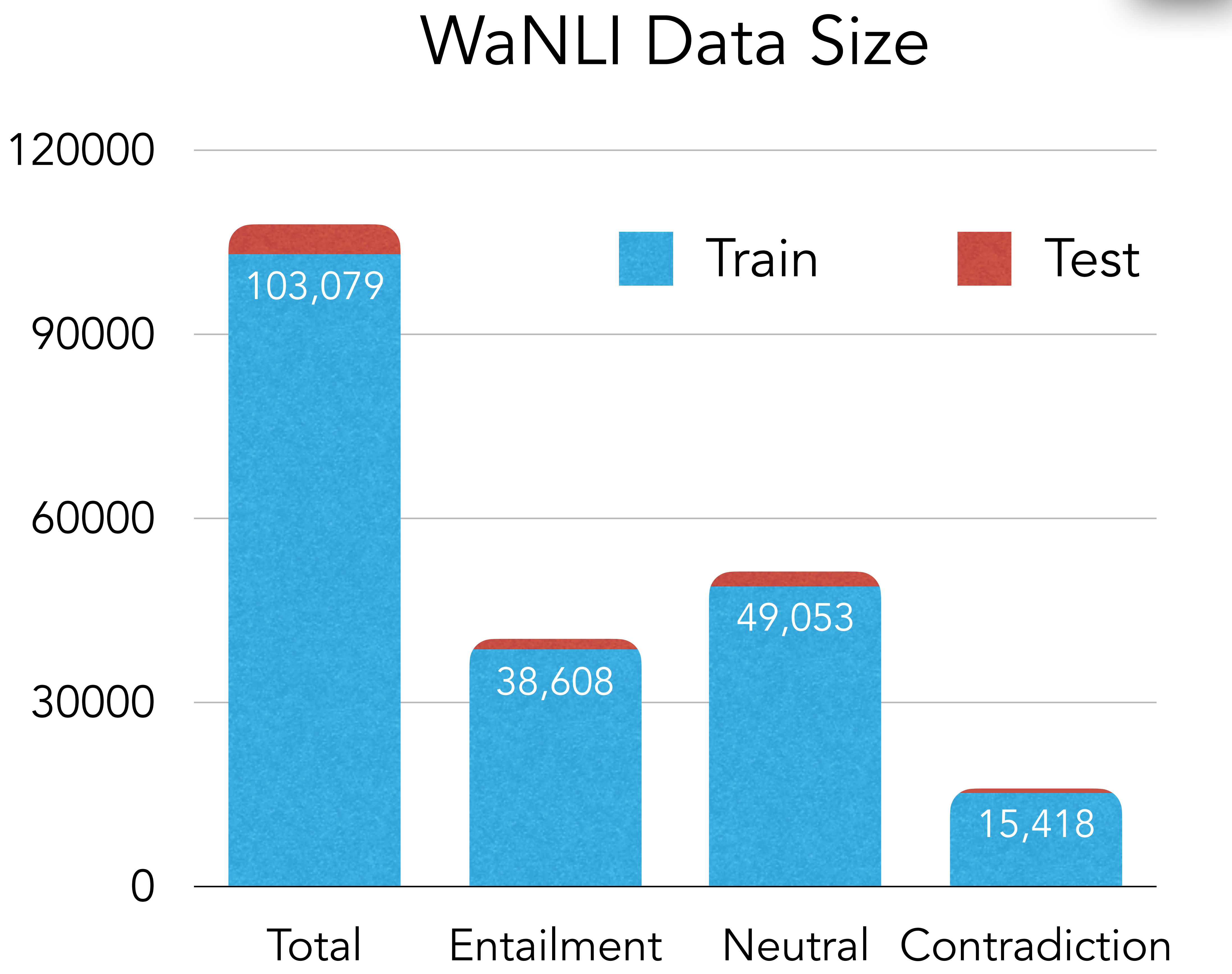
万理 Ten thousand reasoning



WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

Worker-AI Collaborative NLI: WANLI

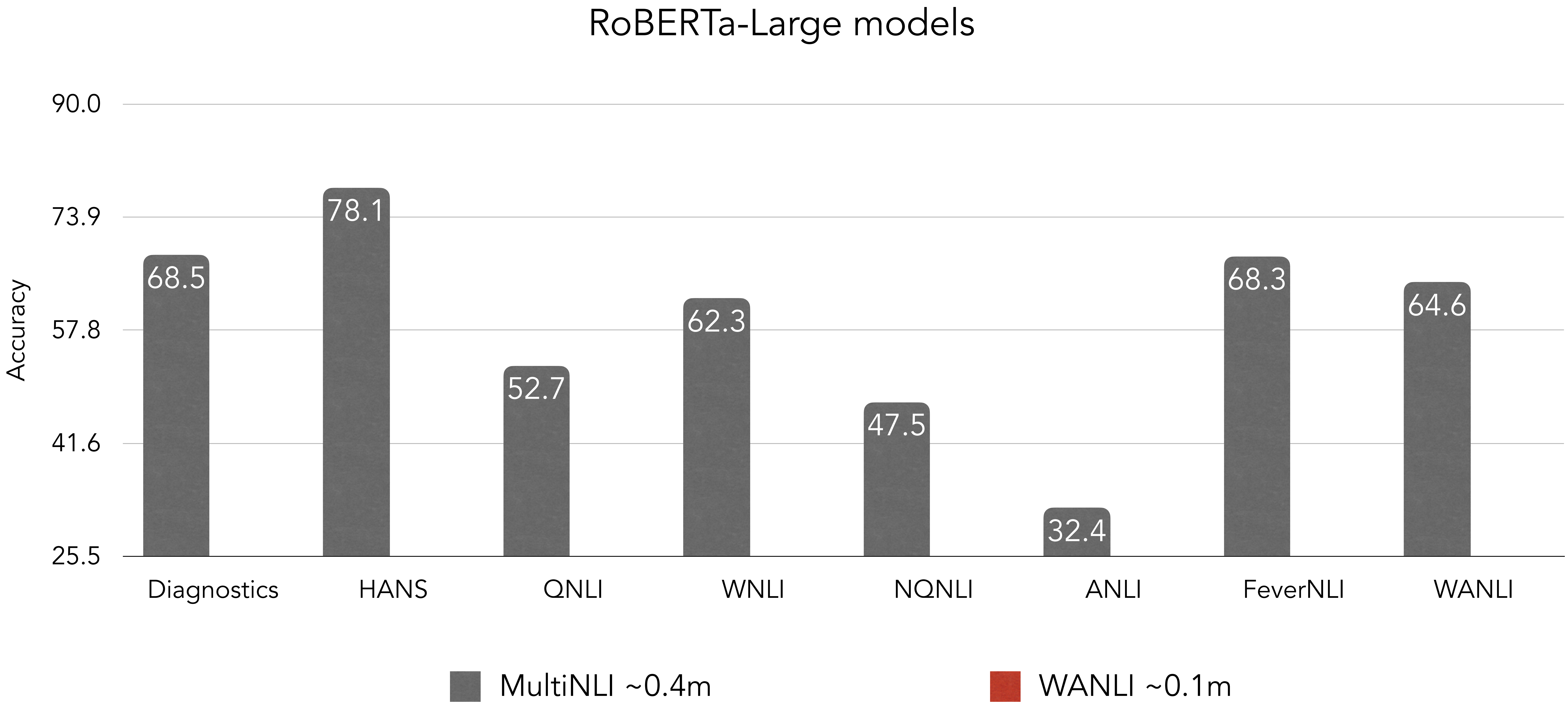
万理 Ten thousand reasoning



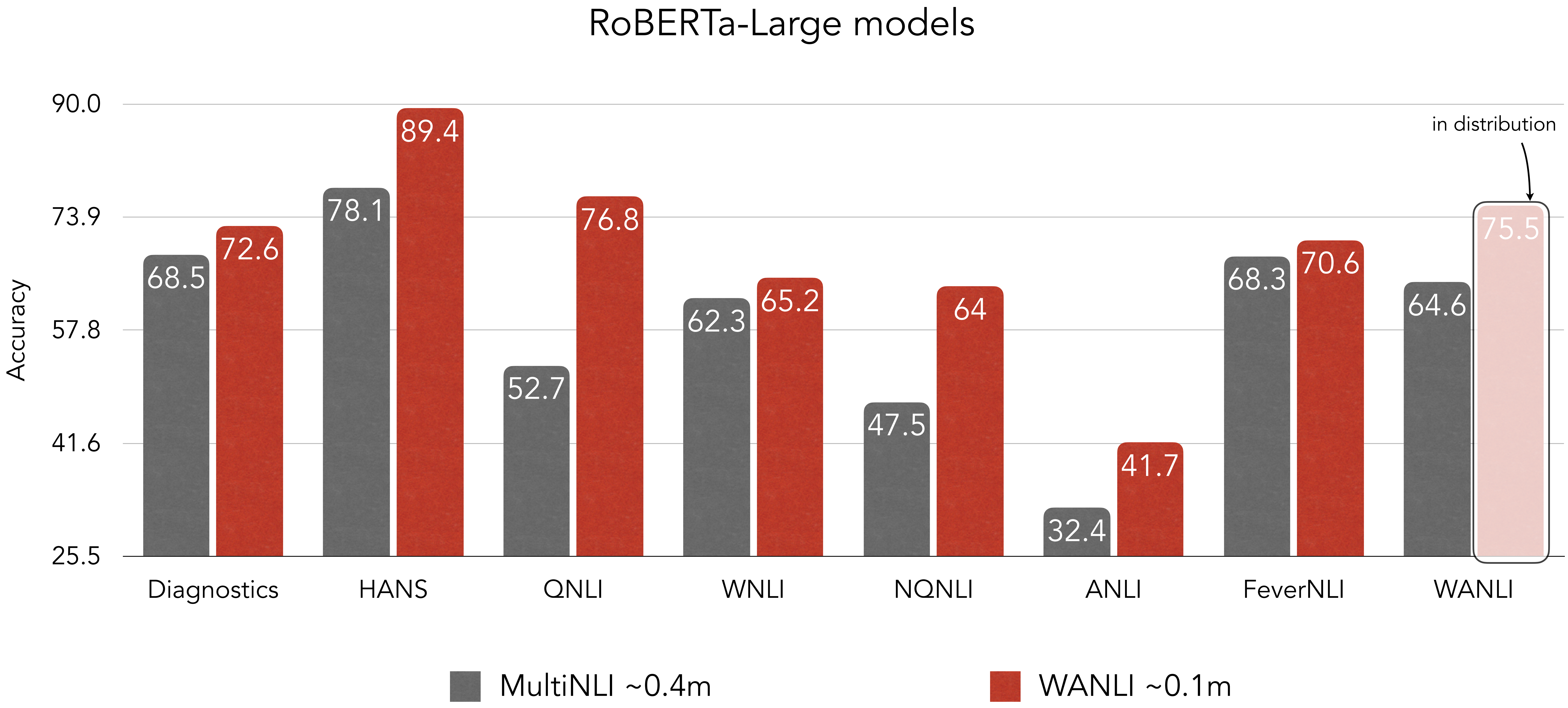
WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

RoBERTa-Large models

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

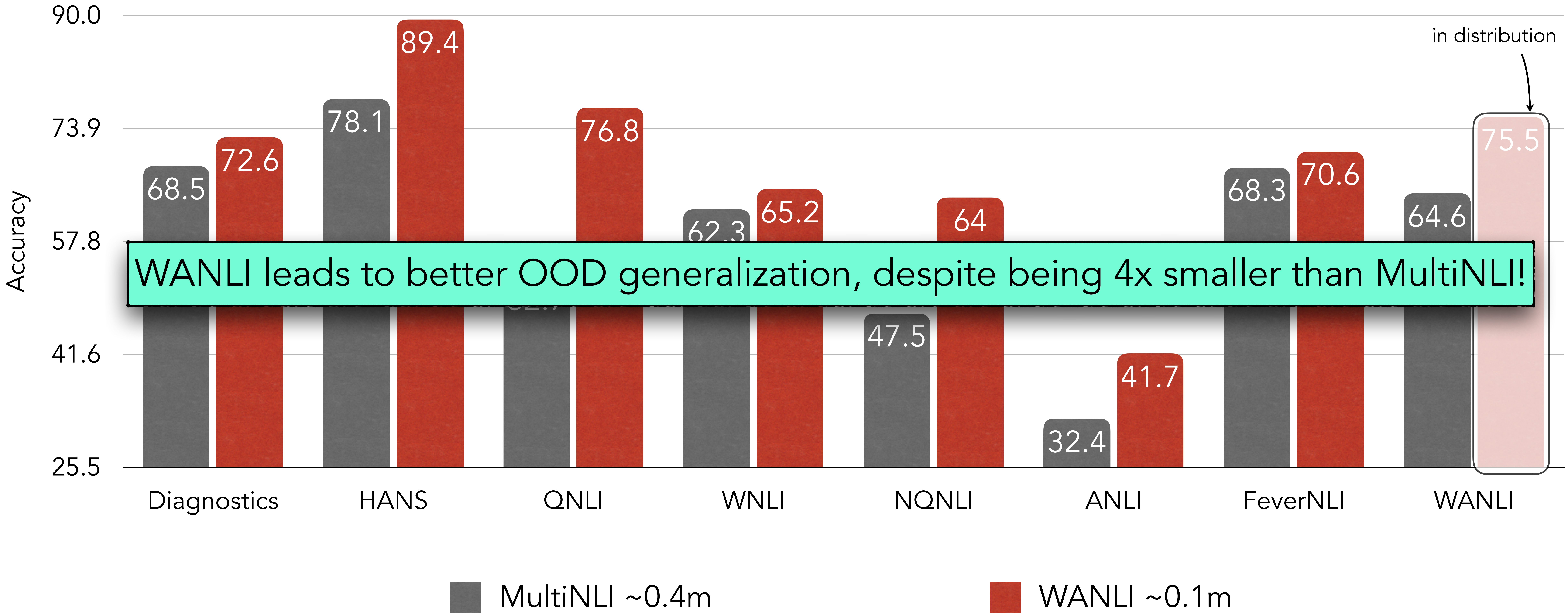


WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

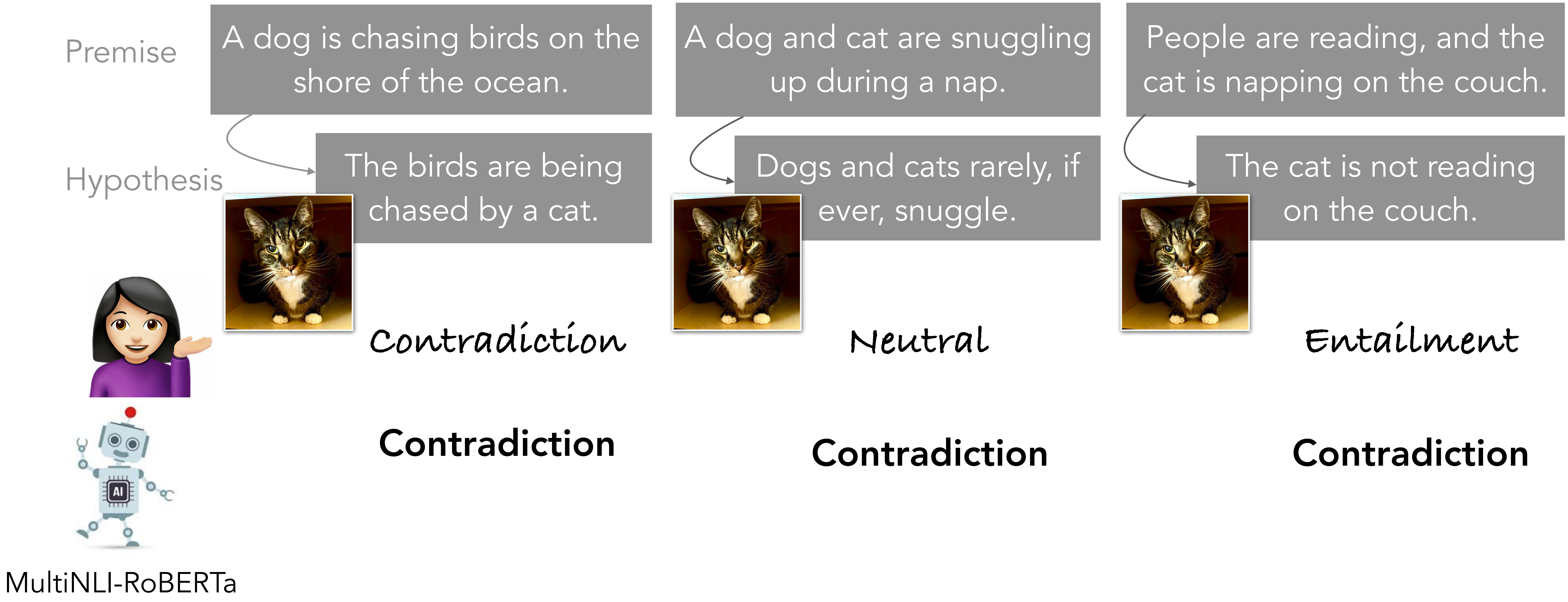
RoBERTa-Large models



WANLI leads to better OOD generalization, despite being 4x smaller than MultiNLI!

Please see paper for more comparisons

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

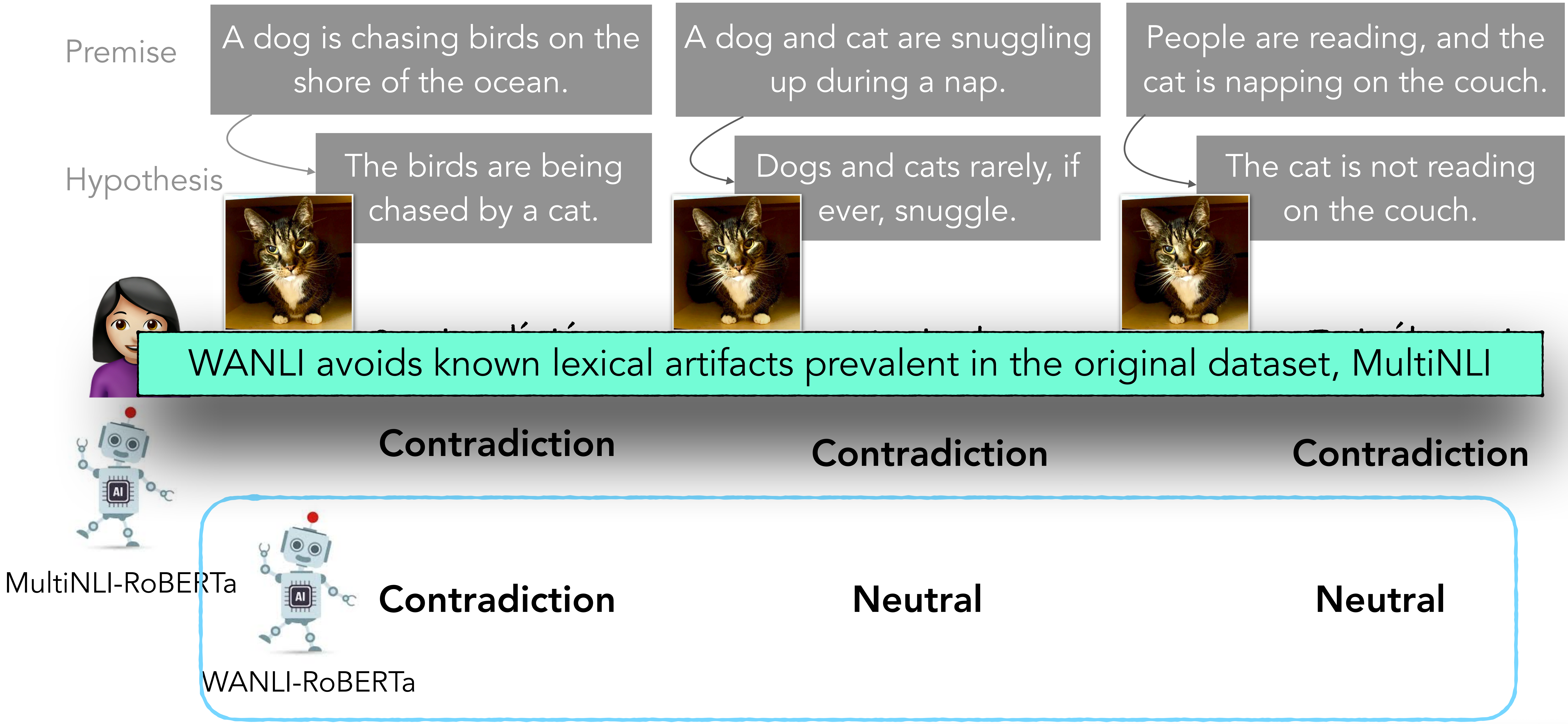


Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
	Contradiction	Neutral	Entailment
MultiNLI-RoBERTa	**Contradiction**	**Contradiction**	**Contradiction**

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
💁‍♀️	Contradiction	Neutral	Entailment
MultiNLI-RoBERTa	**Contradiction**	**Contradiction**	**Contradiction**
WANLI-RoBERTa	**Contradiction**	**Neutral**	**Neutral**

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



Premise	A dog is chasing birds on the shore of the ocean.	A dog and cat are snuggling up during a nap.	People are reading, and the cat is napping on the couch.
Hypothesis	The birds are being chased by a cat.	Dogs and cats rarely, if ever, snuggle.	The cat is not reading on the couch.
MultiNLI-RoBERTa	Contradiction	Contradiction	Contradiction
WANLI-RoBERTa	Contradiction	Neutral	Neutral

WANLI avoids known lexical artifacts prevalent in the original dataset, MultiNLI

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

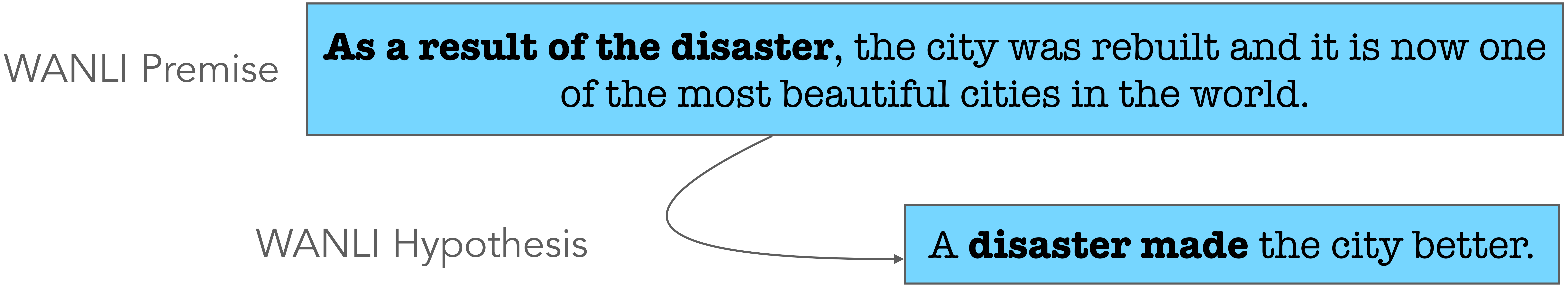
WANLI Premise

As a result of the disaster, the city was rebuilt and it is now one of the most beautiful cities in the world.

WANLI Hypothesis

A **disaster made** the city better.

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]



WANLI Premise

As a result of the disaster, the city was rebuilt and it is now one of the most beautiful cities in the world.

WANLI Hypothesis

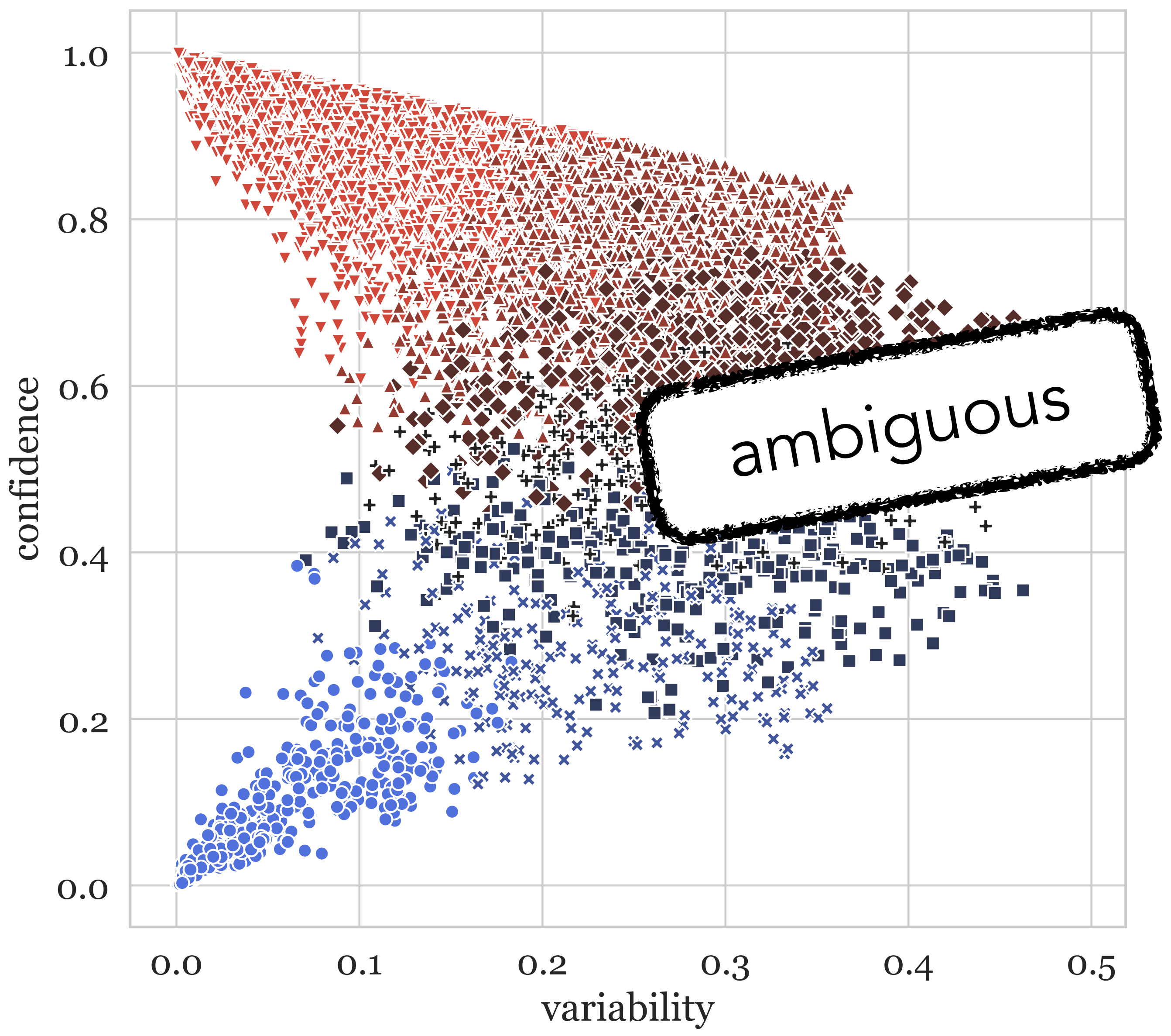
A **disaster made** the city better.

💁🏾‍♂️ Neutral 💁🏽‍♀️ Contradiction 💁🏼 Entailment

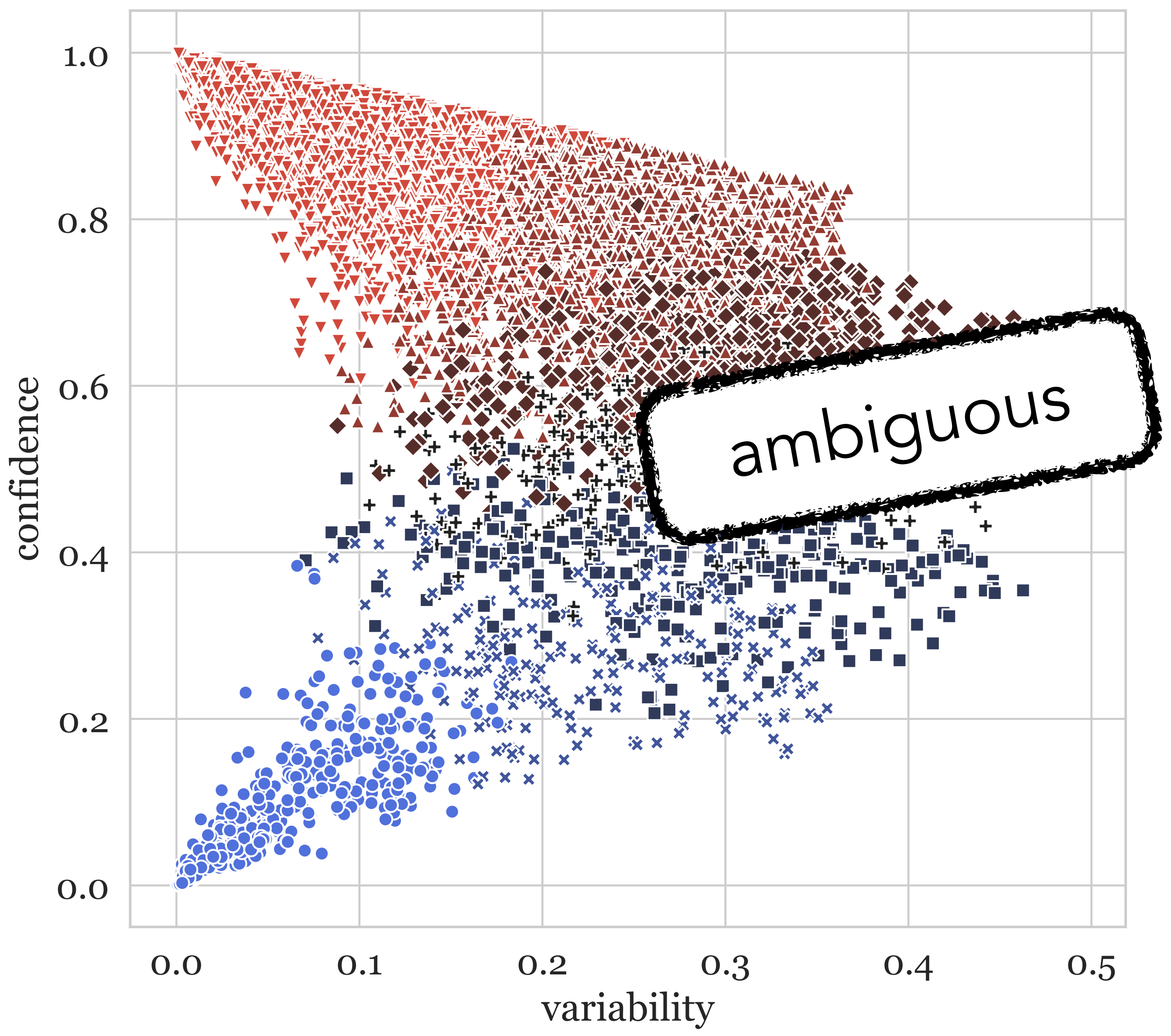
Also see

[Pavlick & Kwiatkowski, 2019; Chen et al., 2020; Zhou et al., 2022; Davani et al., 2021]

WANLI [Liu., **Swayamdipta**, Smith and Choi, ArXiV 2022]

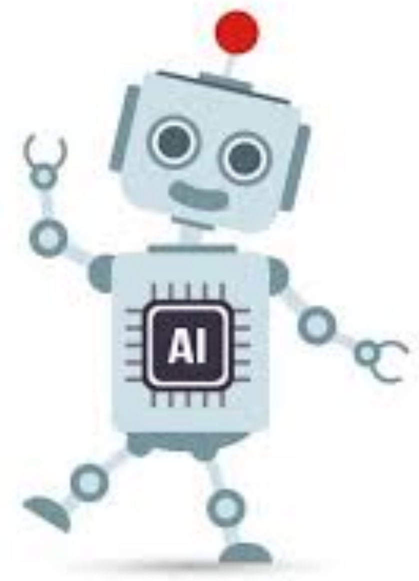


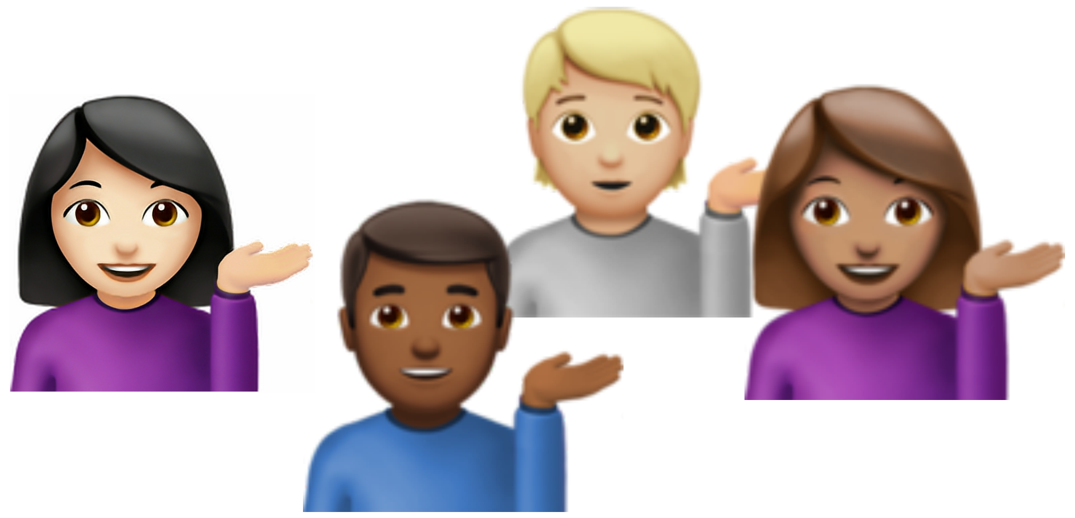
Mapping large datasets to discover regions which are **challenging** to models



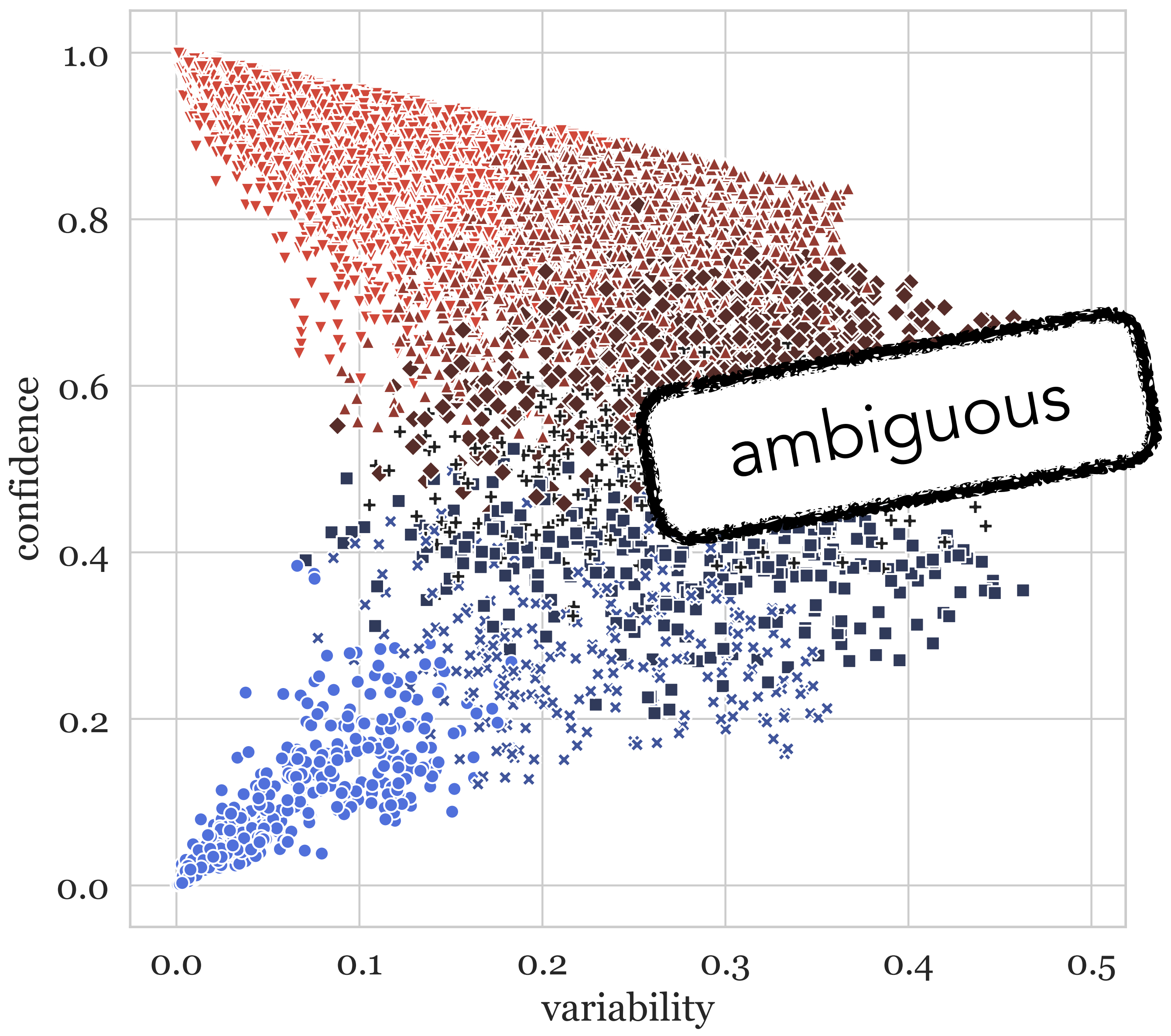
Mapping large datasets to discover regions which are **challenging** to models

GPT-3



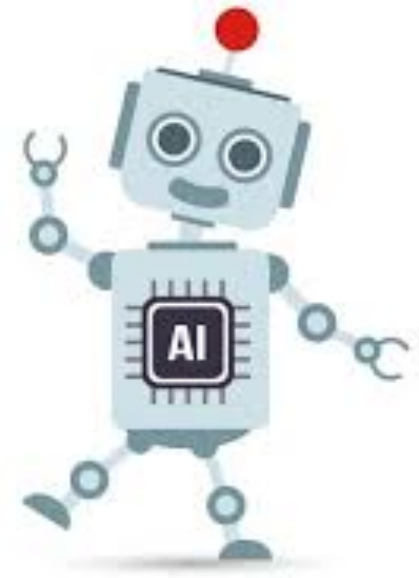


Generating new challenging instances via a collaboration of **humans and models**

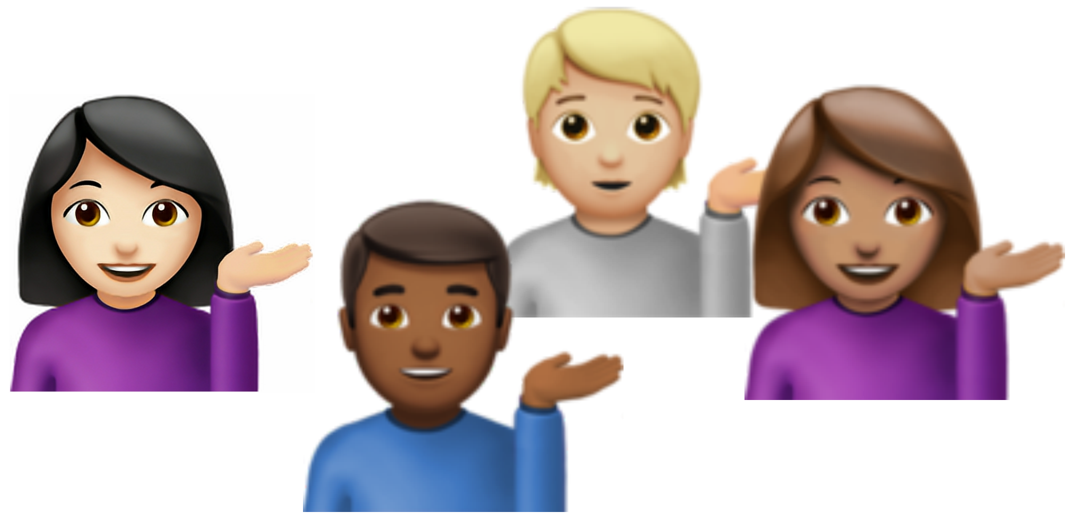


+

GPT-3



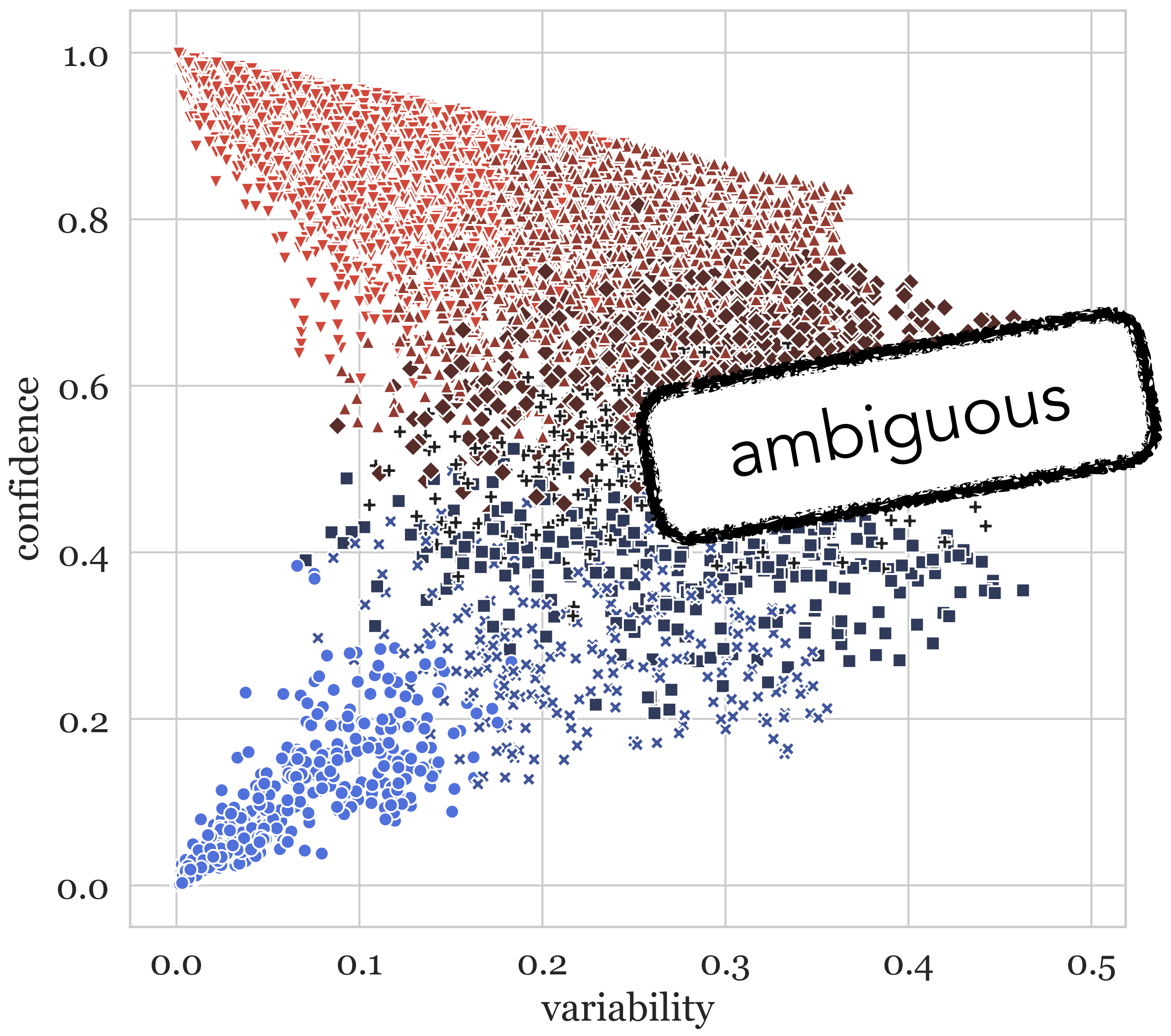
+



Generating new challenging instances via a collaboration of **humans and models**

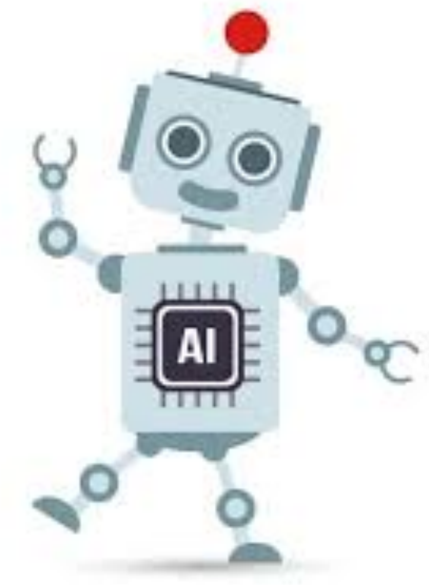
Mapping large datasets to discover regions which are **challenging** to models

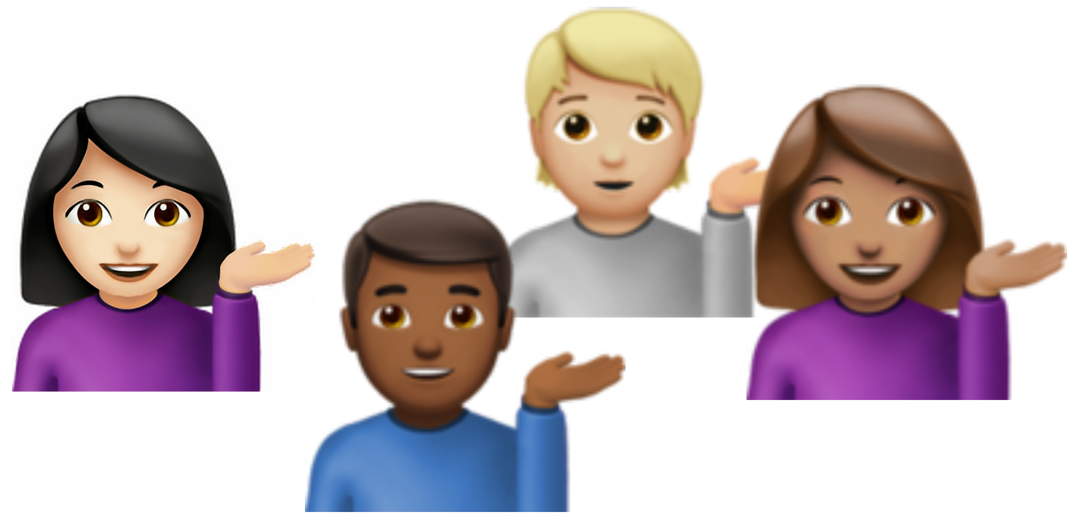
Rethinking data by **shifting the focus to data quality** over quantity



Mapping large datasets to discover regions which are **challenging** to models

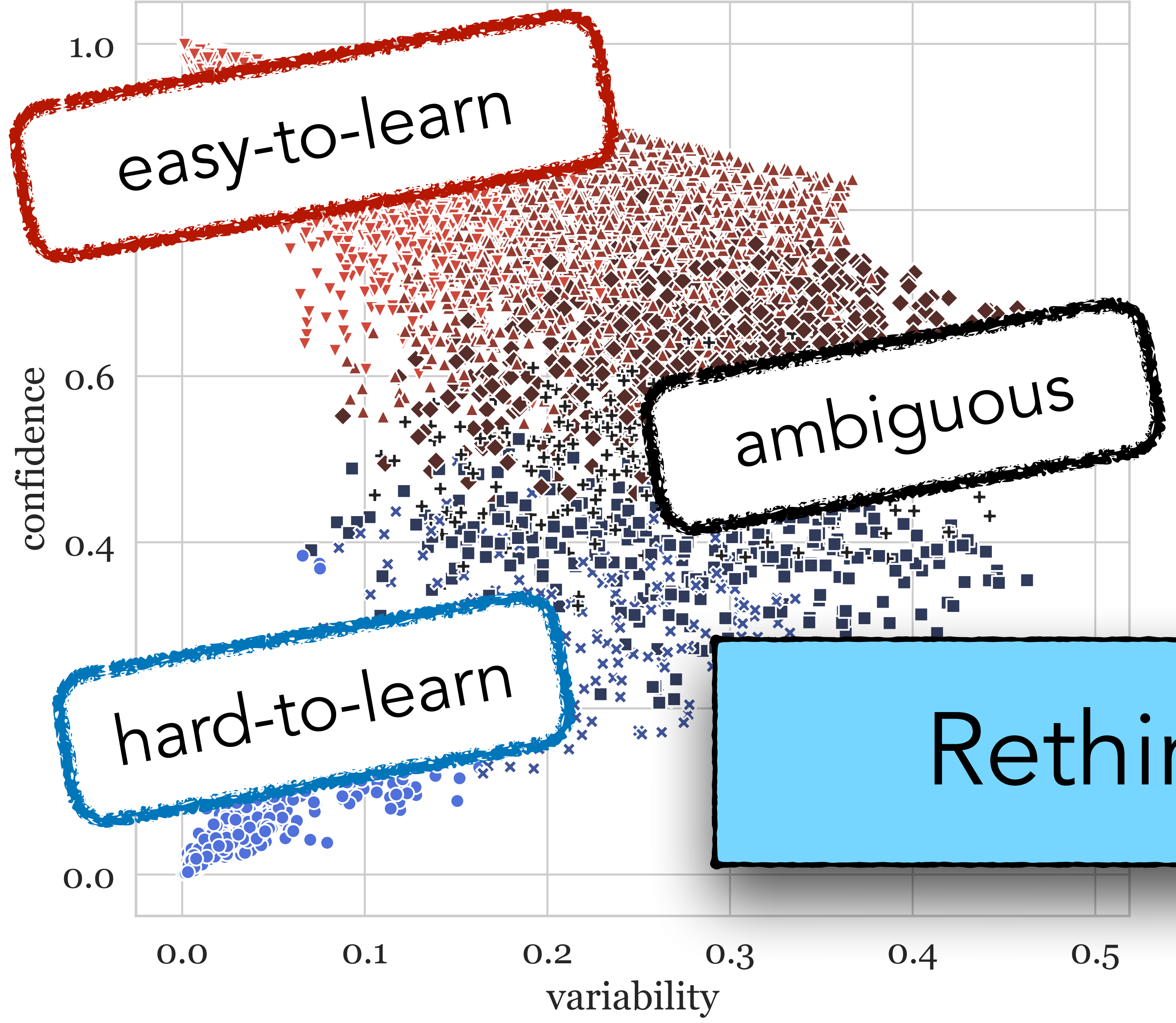
GPT-3



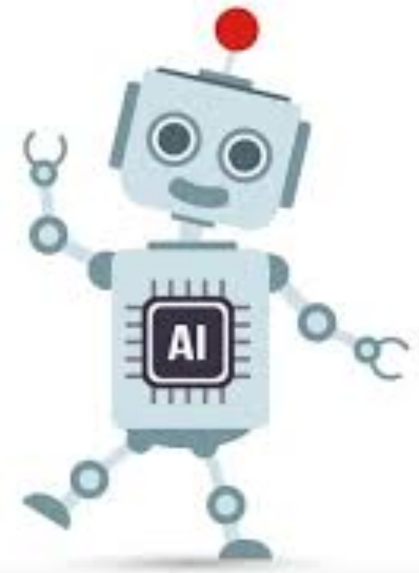


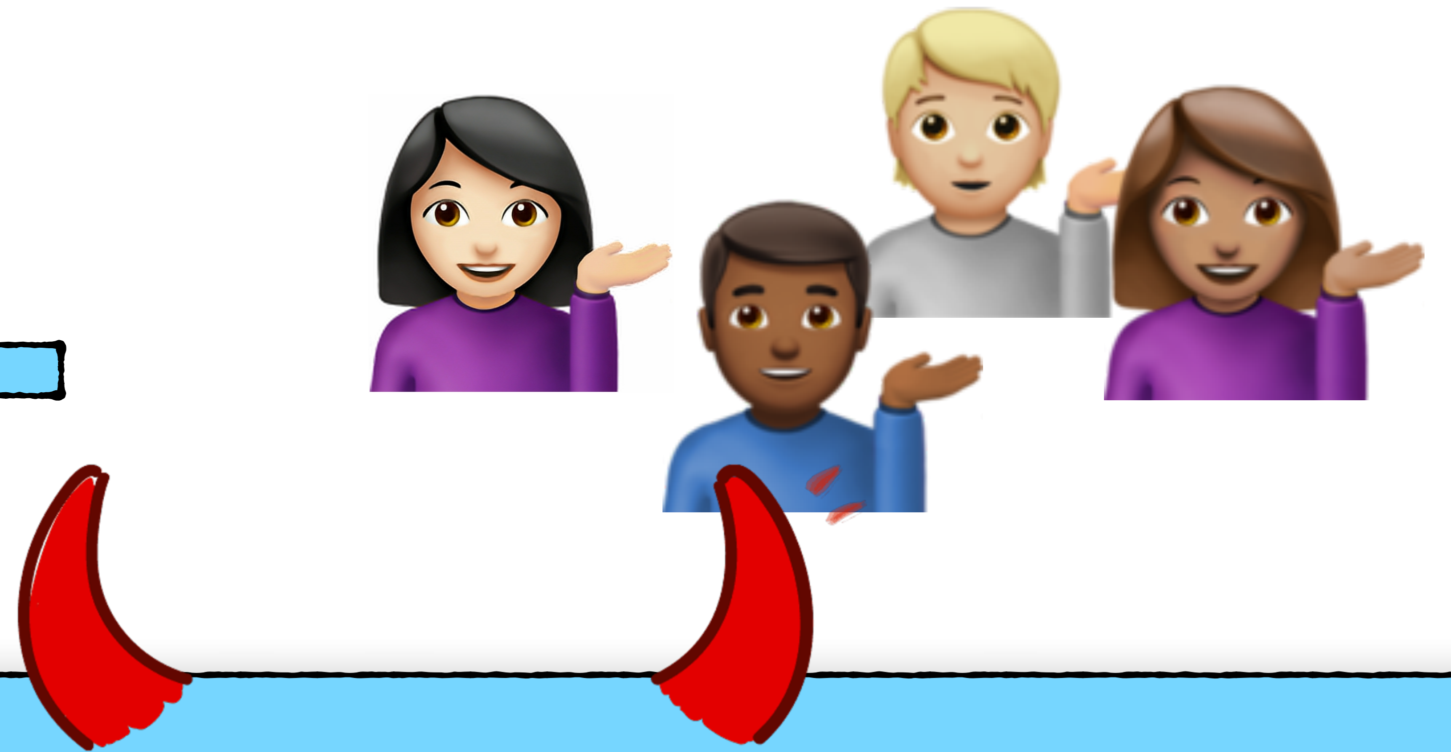
Generating new challenging instances via a collaboration of **humans and models**

Rethinking data by **shifting the focus to data quality** over quantity



GPT-3





Rethinking data by **shifting the focus to data quality** over quantity

WANLI

Cartography

Alisa Liu

Roy Schwartz

Yizhong Wang

Nicholas Lourie

Hannaneh Hajishirzi

Noah A. Smith

Yejin Choi