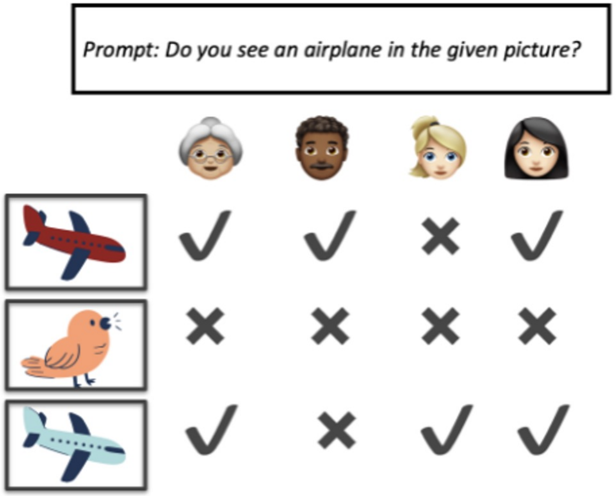
Don’t Blame the Data, Blame the Model:

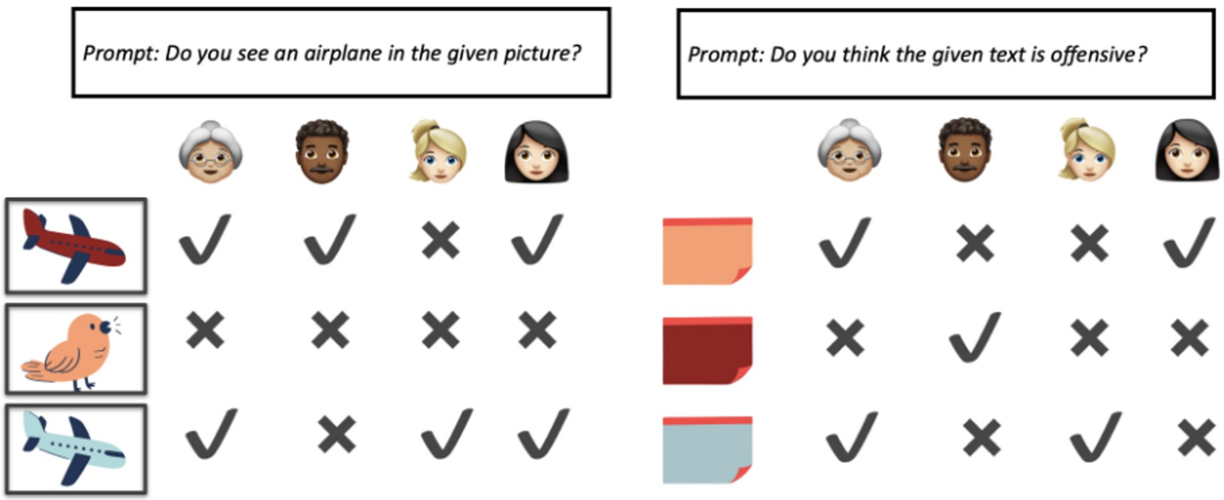
Understanding Noise and Bias When Learning from Subjective Annotations

Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, Kristina Lerman

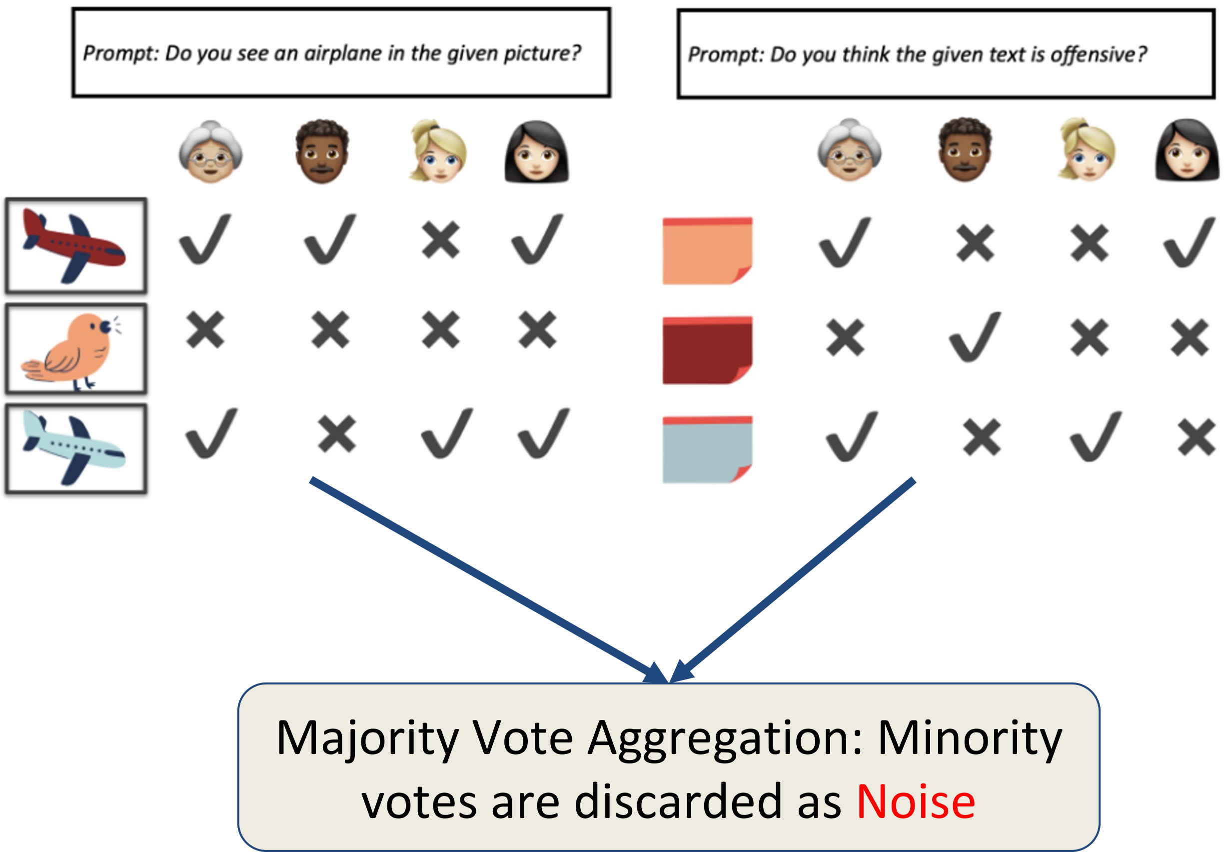
Motivation



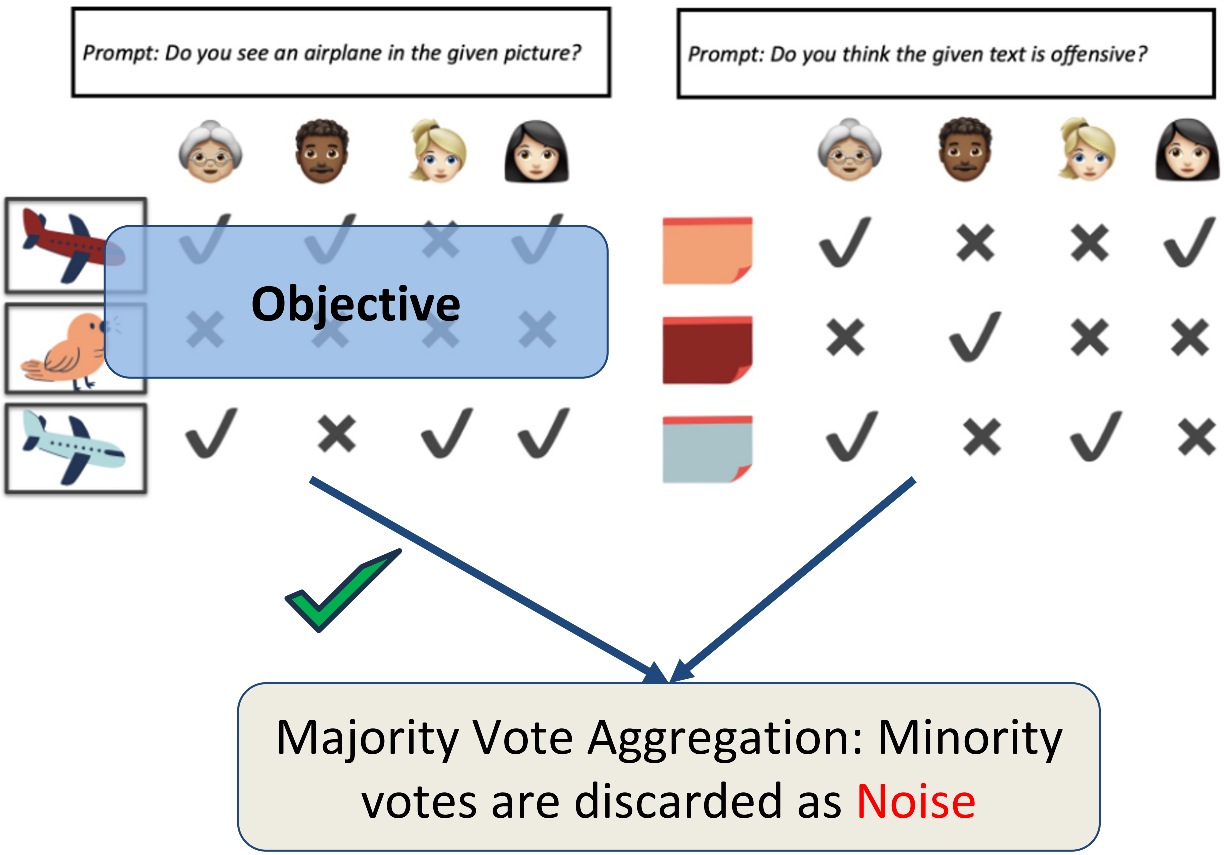
Motivation


Prompt: Do you see an airplane in the given picture?
Prompt: Do you think the given text is offensive?

Motivation

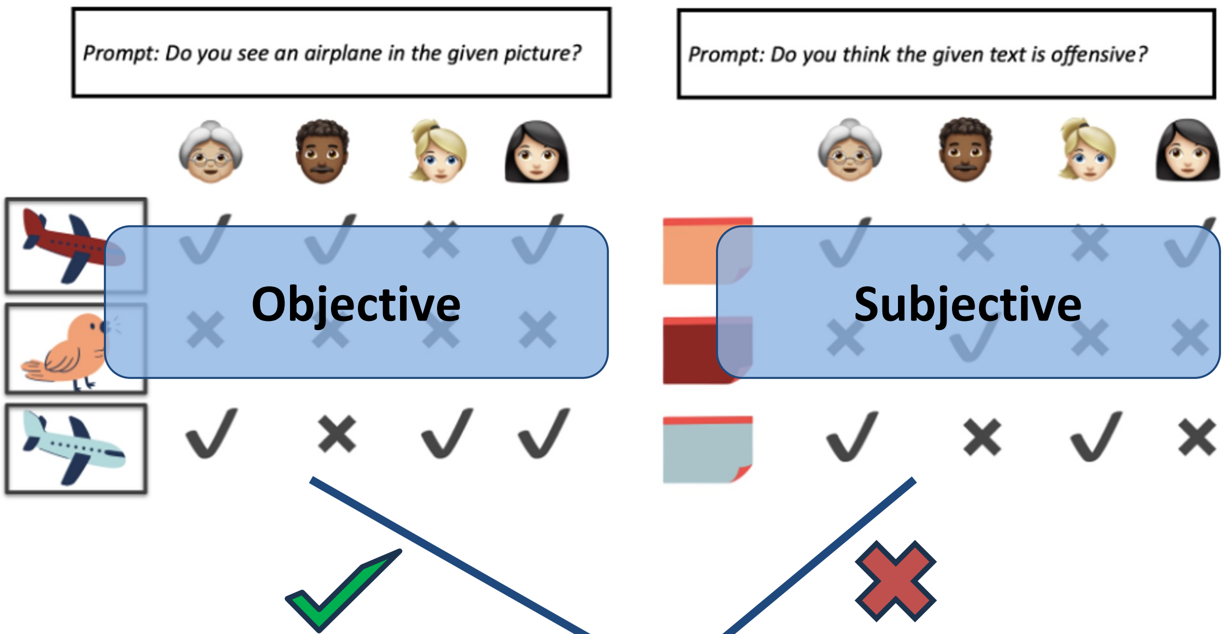

Prompt: Do you see an airplane in the given picture?
Prompt: Do you think the given text is offensive?
Majority Vote Aggregation: Minority votes are discarded as Noise

Motivation



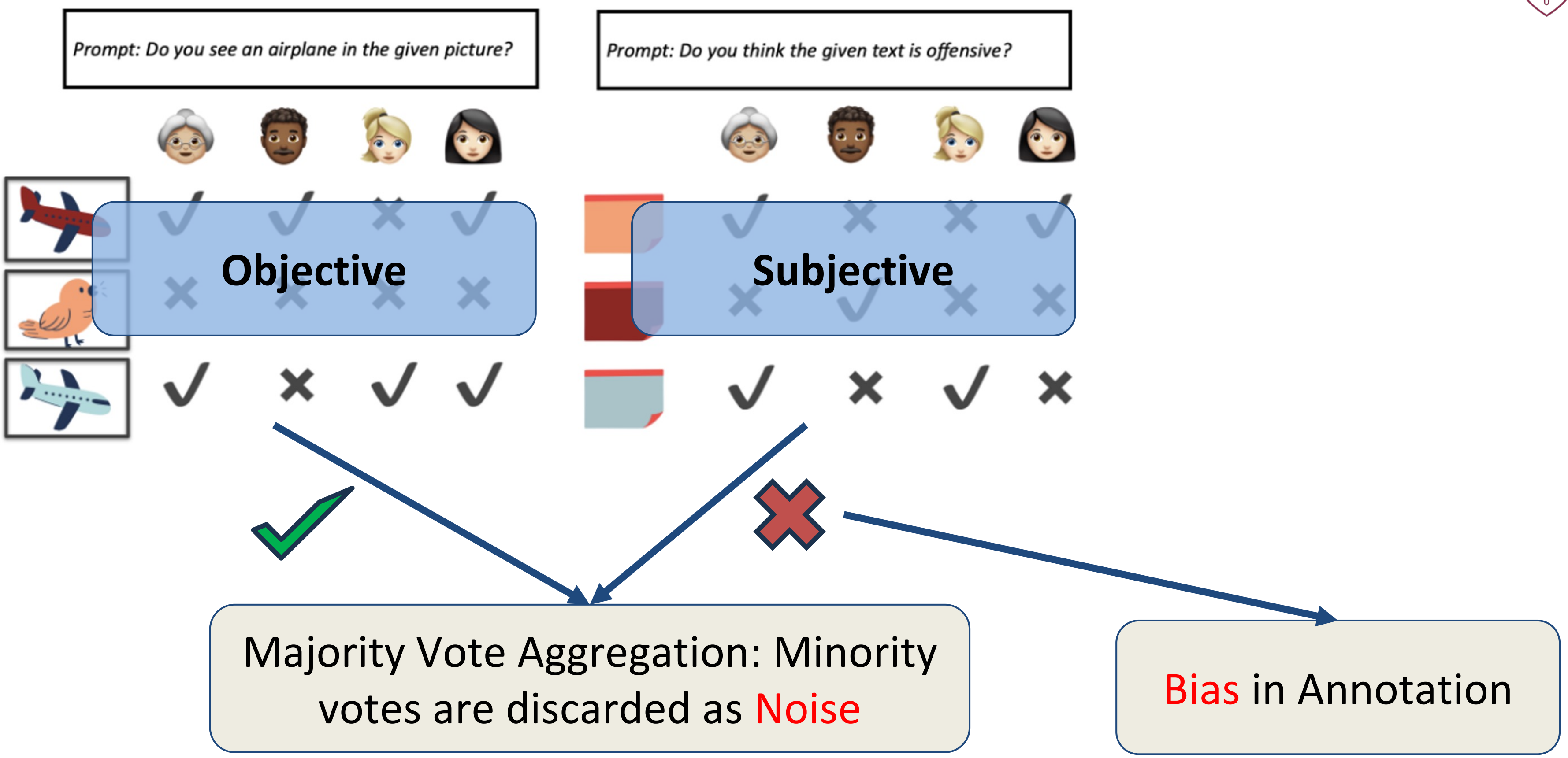
Prompt: Do you see an airplane in the given picture?
Prompt: Do you think the given text is offensive?
Objective
Majority Vote Aggregation: Minority votes are discarded as Noise

Motivation



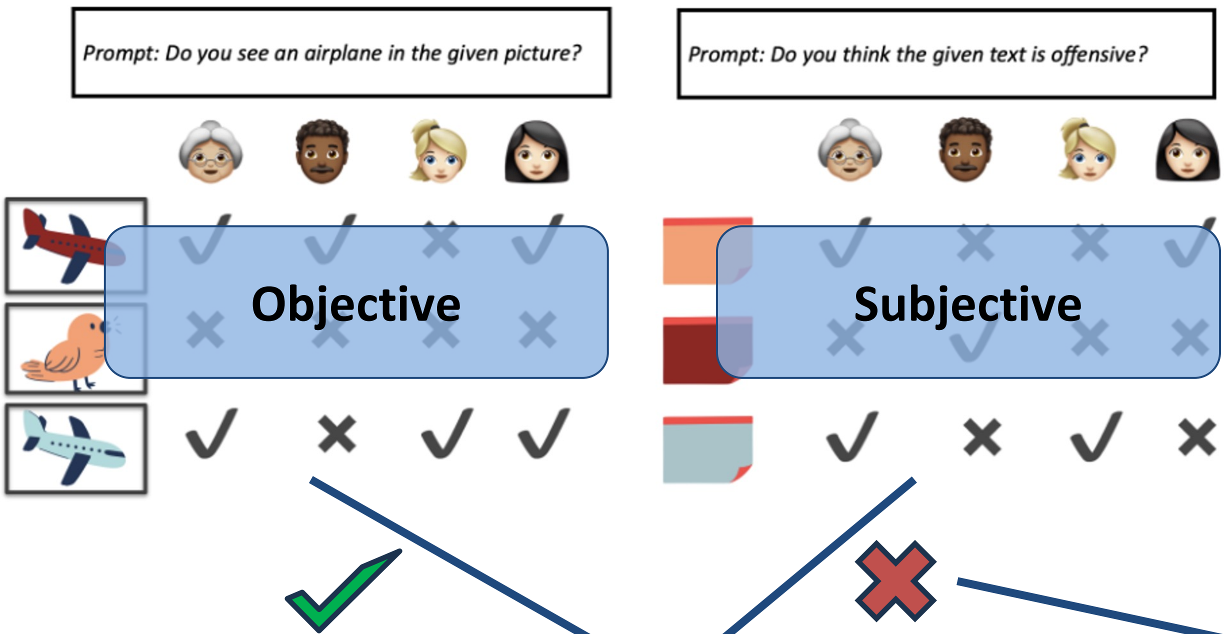
Majority Vote Aggregation: Minority votes are discarded as Noise

Motivation



Prompt: Do you see an airplane in the given picture?
Prompt: Do you think the given text is offensive?
Objective
Subjective
Majority Vote Aggregation: Minority votes are discarded as Noise
Bias in Annotation

Motivation

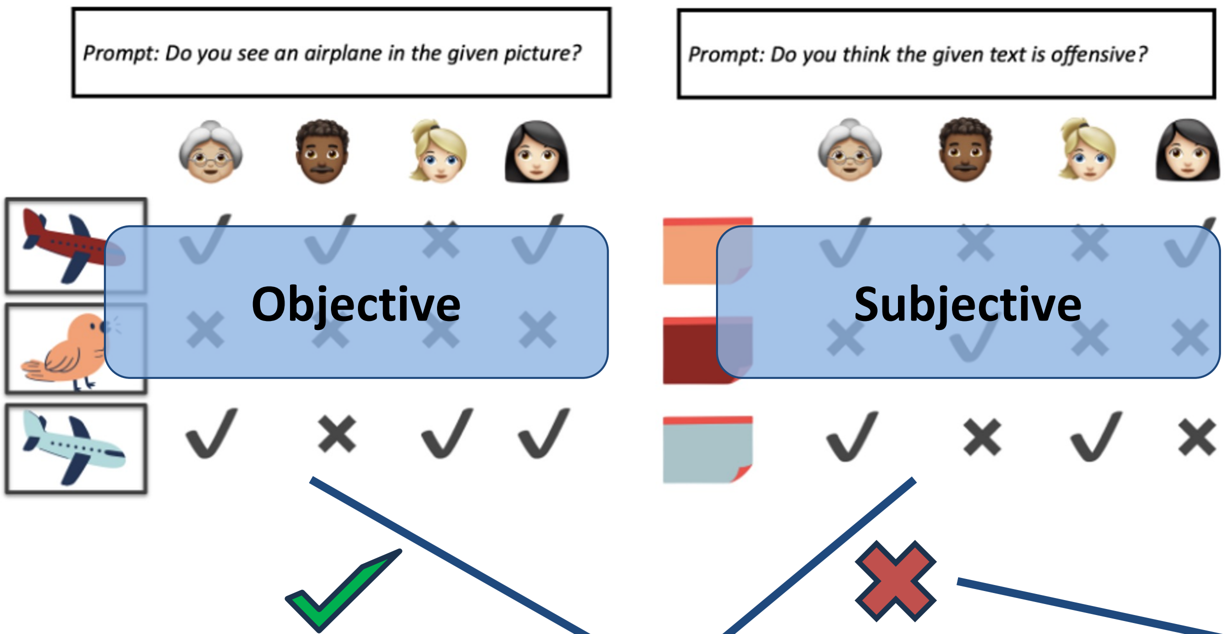


Different perspectives essential for subjective tasks?

Majority Vote Aggregation: Minority votes are discarded as Noise

Bias in Annotation

Motivation



Different perspectives essential for subjective tasks?

Majority Vote Aggregation: Minority votes are discarded as Noise

Bias in Annotation

Noise vs **Bias**?

Research Questions?

Assumption – Single correct label exists

Research Questions?

Assumption – Single correct label exists

Correlation between human disagreement on instances and model's uncertainty in prediction when using majority labels?

Research Questions?

Assumption – Single correct label exists

Correlation between human disagreement on instances and model's uncertainty in prediction when using majority labels?

All annotations available

Research Questions?

Assumption – Single correct label exists

Correlation between human disagreement on instances and model's uncertainty in prediction when using majority labels?

All annotations available

Does learning from raw annotations enhance the model's confidence?

Are perspectivist classification models effective?

Model uncertainty

Model uncertainty

Assumption – Noisy samples lead to uncertainty in modelling.

Model uncertainty

Assumption – Noisy samples lead to uncertainty in modelling.

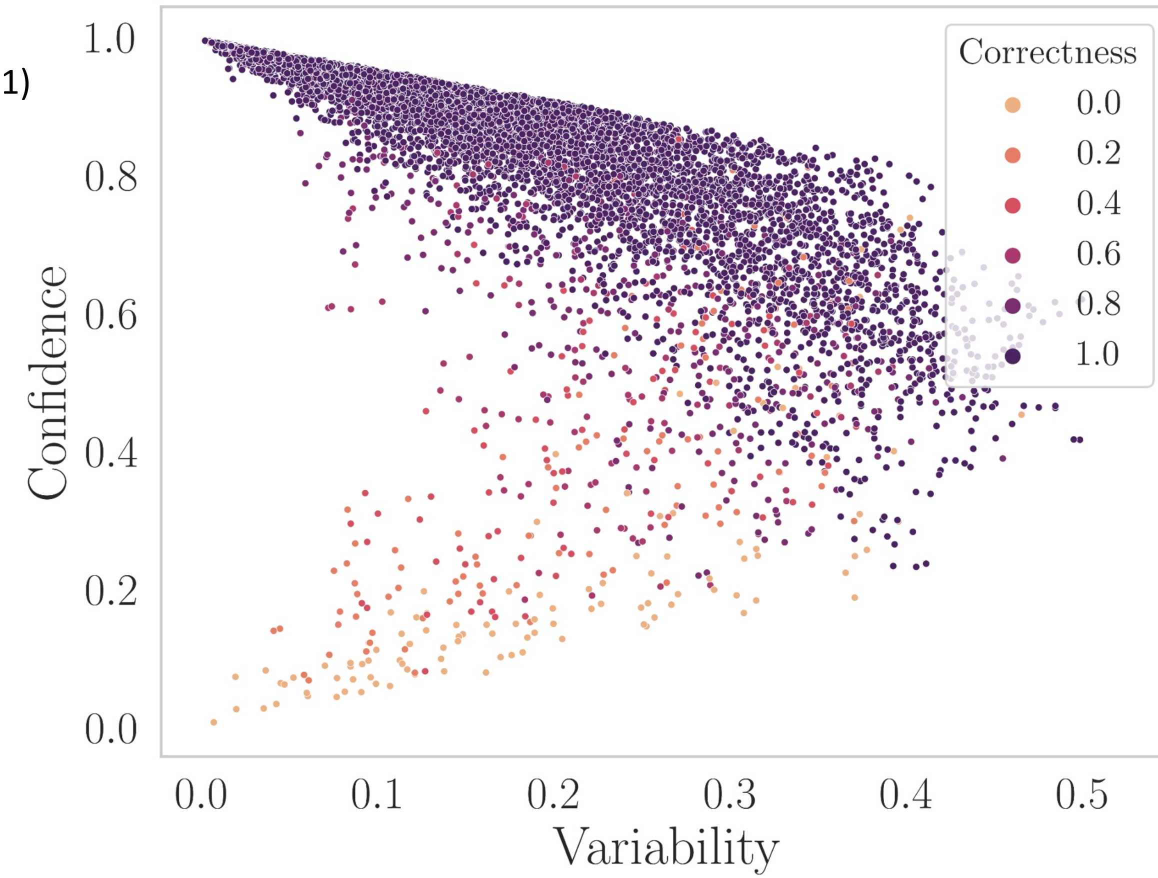
Quantifying uncertainty in modelling - Data Maps (Swayamdipta et al., 2020)

Model uncertainty

Assumption – Noisy samples lead to uncertainty in modelling.

Quantifying uncertainty in modelling - Data Maps (Swayamdipta et al., 2020)

MDA (Leonardelli et al., EMNLP 2021)



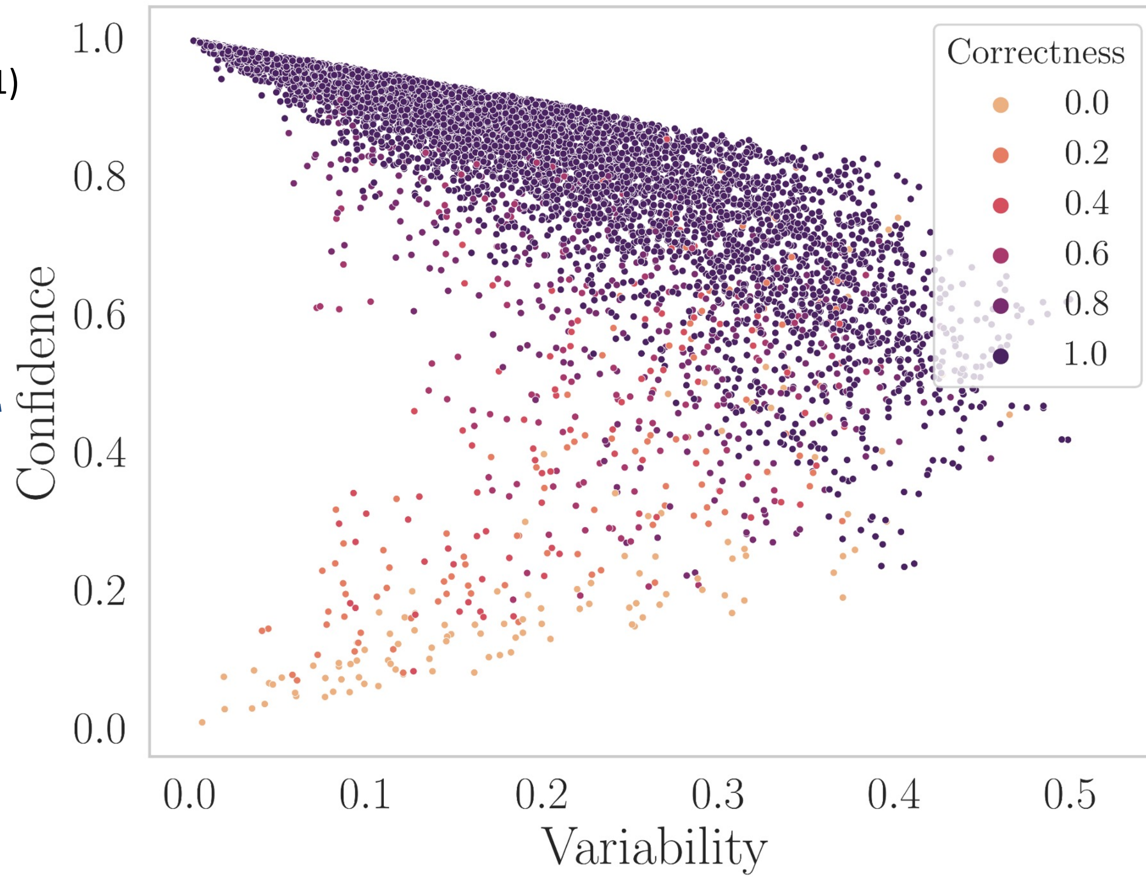
Model uncertainty

Assumption – Noisy samples lead to uncertainty in modelling.

Quantifying uncertainty in modelling - Data Maps (Swayamdipta et al., 2020)

MDA (Leonardelli et al., EMNLP 2021)

Mean of probabilities for gold label across epochs.

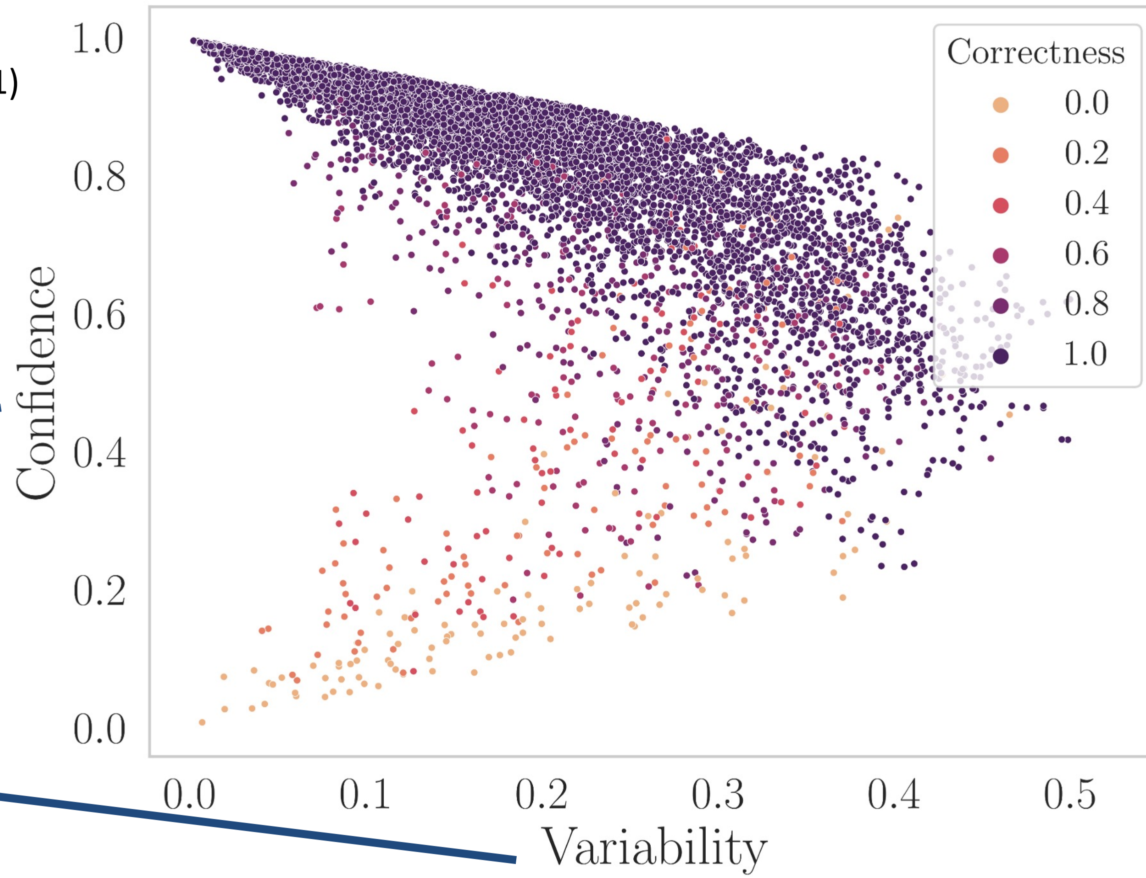


Model uncertainty

Assumption – Noisy samples lead to uncertainty in modelling.

Quantifying uncertainty in modelling - Data Maps (Swayamdipta et al., 2020)

MDA (Leonardelli et al., EMNLP 2021)



Mean of probabilities for gold label across epochs.

Standard Deviation of probabilities for gold label across epochs.

Correlation between human disagreement on instances and model's uncertainty when using majority labels?

Annotator Agreement Level (a_m) : fraction of annotations that align with majority vote for a text sample

Correlation between **human disagreement** on instances and model's uncertainty when using majority labels?

A measure to quantify disagreement between annotators on a label for a given sample.

Annotator Agreement Level (a_m) : fraction of annotations that align with majority vote for a text sample

Correlation between **human disagreement** on instances and model's uncertainty when using majority labels?

A measure to quantify disagreement between annotators on a label for a given sample.

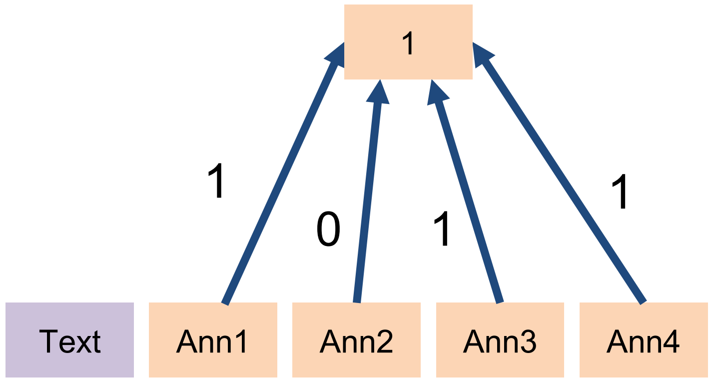
Annotator Agreement Level (a_m) : fraction of annotations that align with majority vote for a text sample

Text	Ann1	Ann2	Ann3	Ann4

Correlation between **human disagreement** on instances and model's uncertainty when using majority labels?

A measure to quantify disagreement between annotators on a label for a given sample.

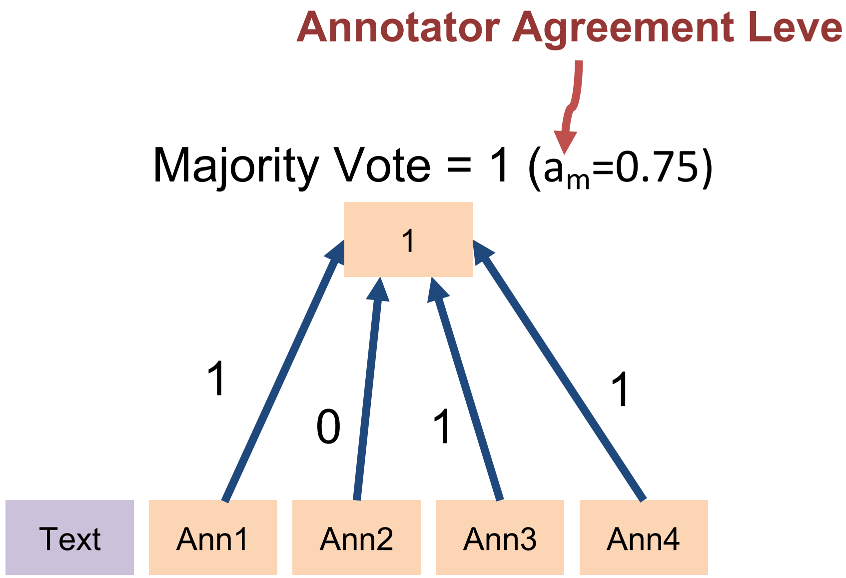
Annotator Agreement Level (a_m) :
fraction of annotations that align with majority vote for a text sample



Correlation between **human disagreement** on instances and model's uncertainty when using majority labels?

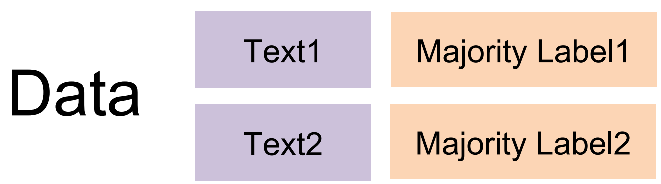
A measure to quantify disagreement between annotators on a label for a given sample.

Annotator Agreement Level (a_m) : fraction of annotations that align with majority vote for a text sample



Correlation between **human disagreement** on instances and model's uncertainty when using majority labels?

Correlation between human disagreement on instances and model's uncertainty when using majority labels?



Correlation between human disagreement on instances and model's uncertainty when using majority labels?

Single-GT model

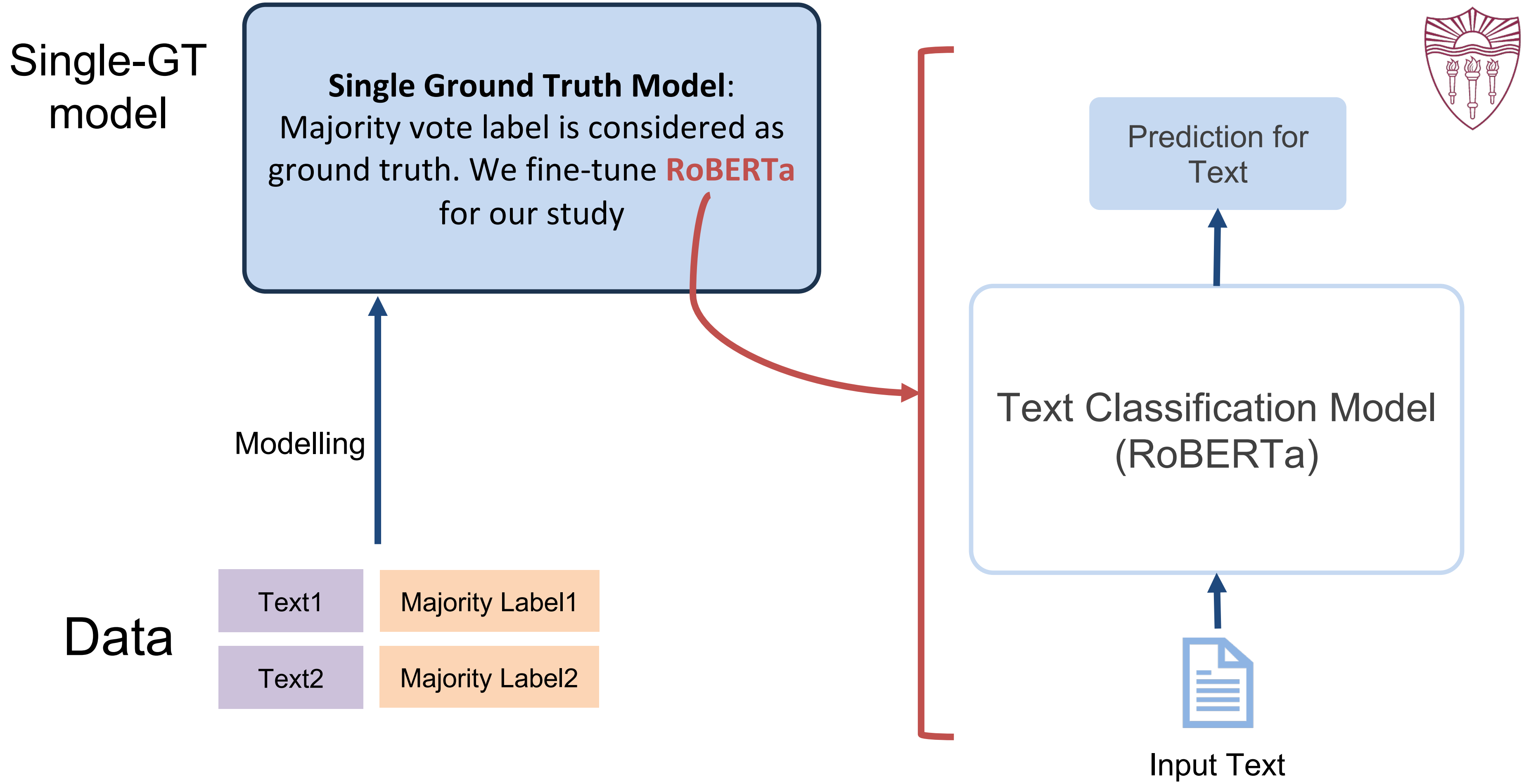
Single Ground Truth Model: Majority vote label is considered as ground truth. We fine-tune RoBERTa for our study

Modelling

Data

Text1	Majority Label1
Text2	Majority Label2

Correlation between human disagreement on instances and model's uncertainty when using majority labels?



Correlation between human disagreement on instances and model's uncertainty when using majority labels?

Datasets

	Toxicity or Hate speech		
	MDA (Leonardelli et al., EMNLP 2021)	**SBIC** (Sap et al., ACL 2020)	**MHS** (Kennedy et al., 2020)
# Annotators	819	307	7,912
# Annotations per annotator	63.7±139	479.3±829.6	17.1±3.8
# Unique texts	10,440	45,318	39,565
# Annotations per text	5	3.2±1.2	2.3±1.0
# Number of labels	2	2	3

Datasets

	Toxicity or Hate speech		
	MDA (Leonardelli et al., EMNLP 2021)	**SBIC** (Sap et al., ACL 2020)	**MHS** (Kennedy et al., 2020)
# Annotators	819	307	7,912
# Annotations per annotator	63.7±139	479.3±829.6	17.1±3.8
# Unique texts	10,440	45,318	39,565
# Annotations per text	5	3.2±1.2	2.3±1.0
# Number of labels	2	2	3

Datasets

	Toxicity or Hate speech		
	MDA (Leonardelli et al., EMNLP 2021)	**SBIC** (Sap et al., ACL 2020)	**MHS** (Kennedy et al., 2020)
# Annotators	819	307	7,912
# Annotations per annotator	63.7±139	479.3±829.6	17.1±3.8
# Unique texts	10,440	45,318	39,565
# Annotations per text	5	3.2±1.2	2.3±1.0
# Number of labels	2	2	3

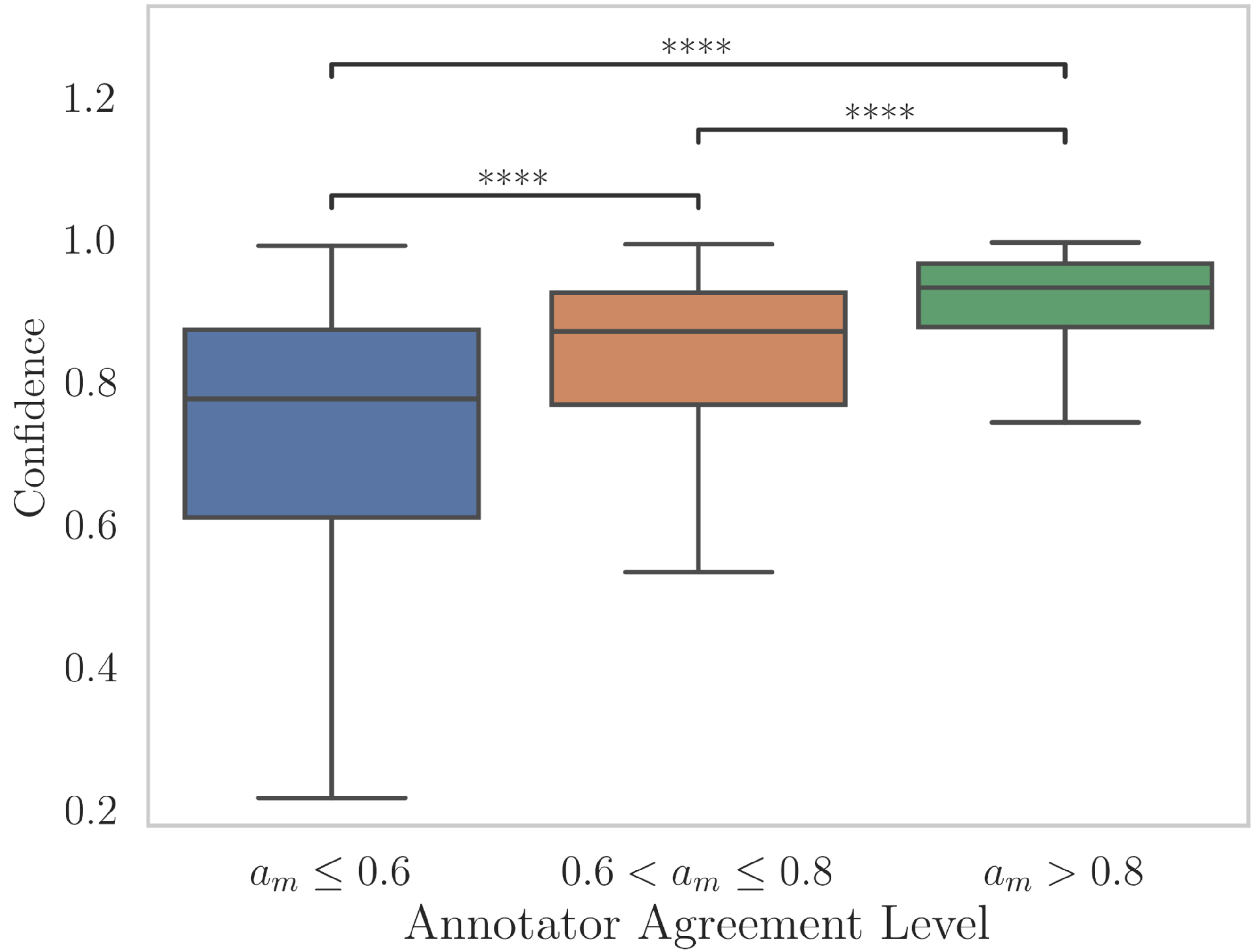
Single-GT model

Confidence vs Annotator Agreement Level

Single-GT model

Confidence vs Annotator Agreement Level

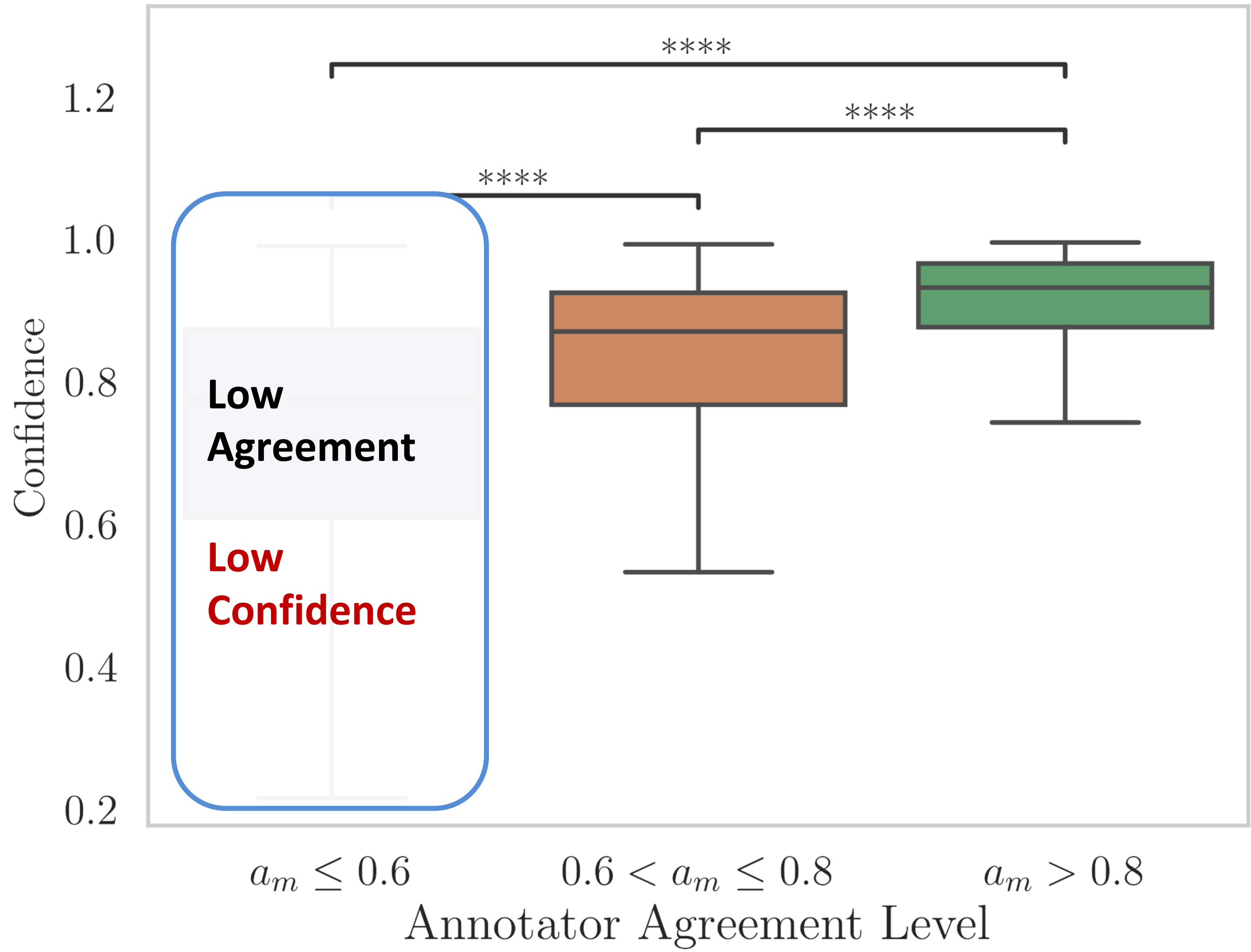
MDA



Single-GT model

Confidence vs Annotator Agreement Level

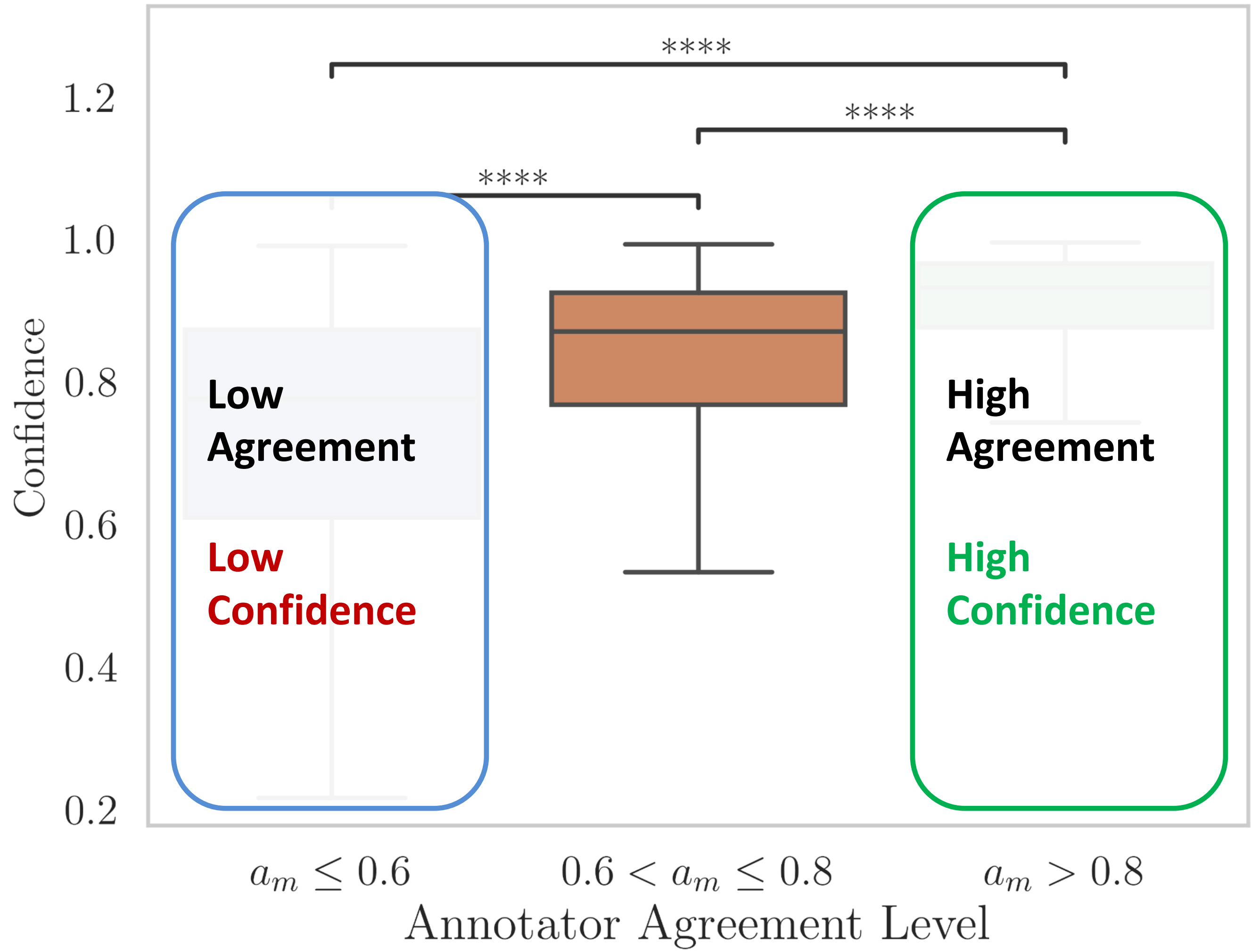
MDA



Single-GT model

Confidence vs Annotator Agreement Level

MDA



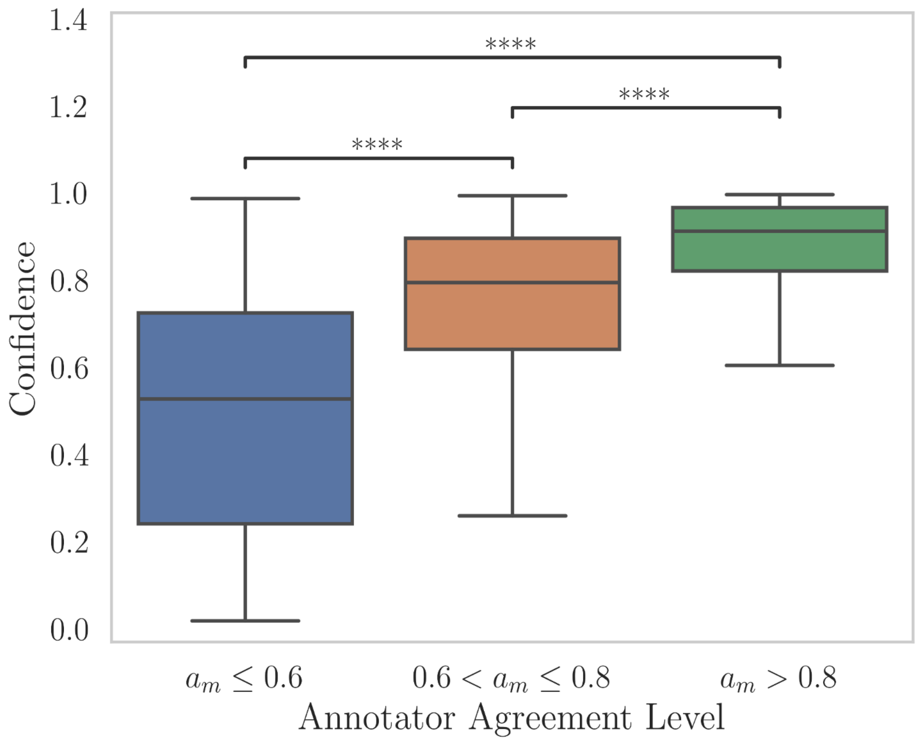
Single-GT model

Confidence vs Annotator Agreement Level

Single-GT model

Confidence vs Annotator Agreement Level

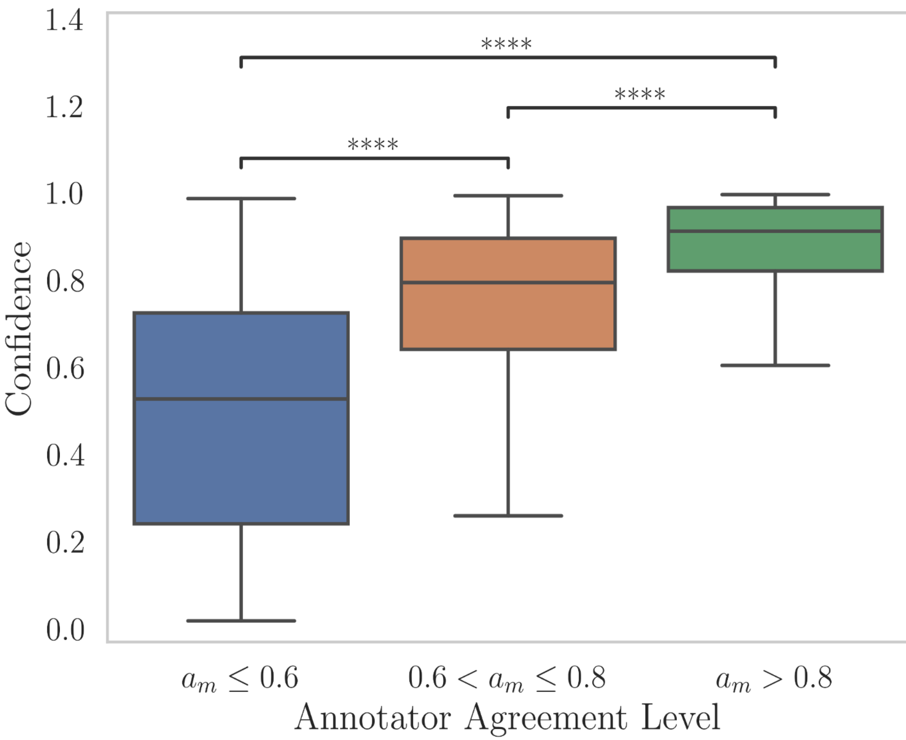
SBIC



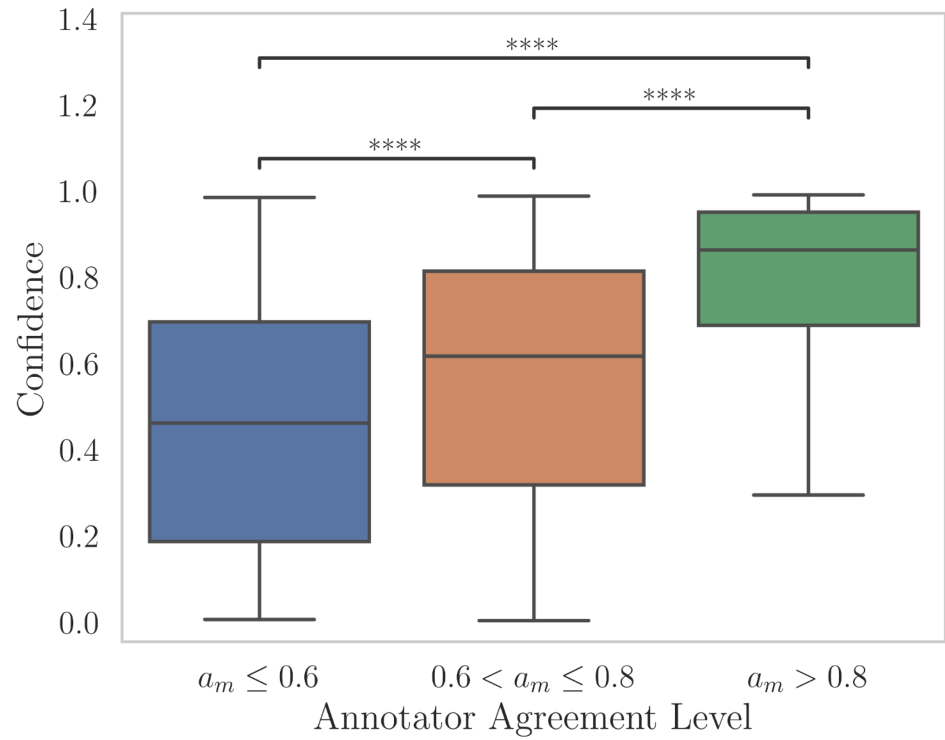
Single-GT model

Confidence vs Annotator Agreement Level

SBIC



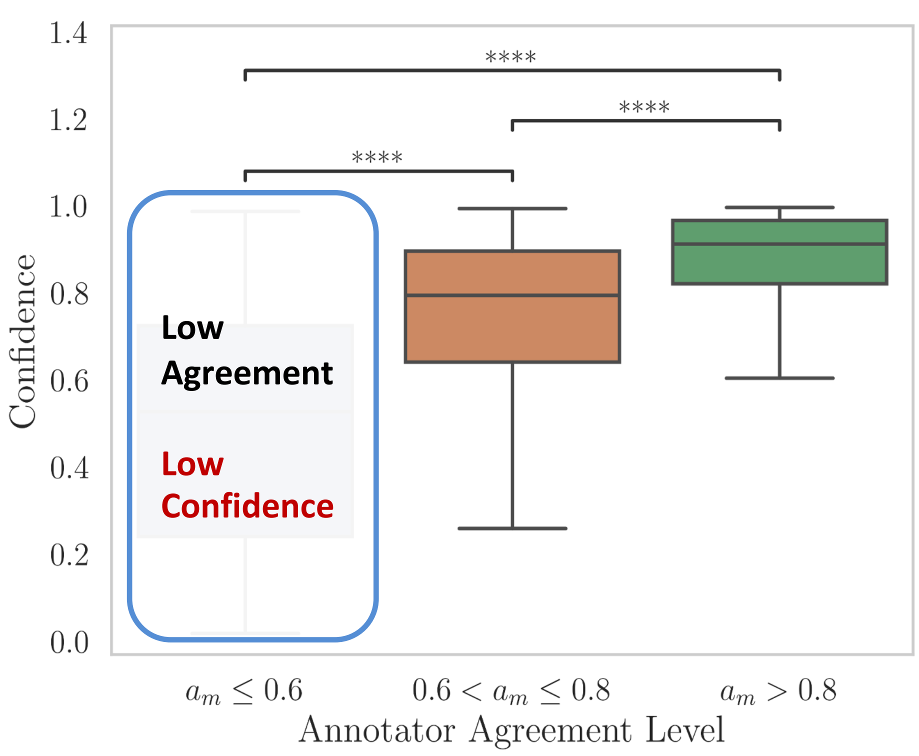
MHS



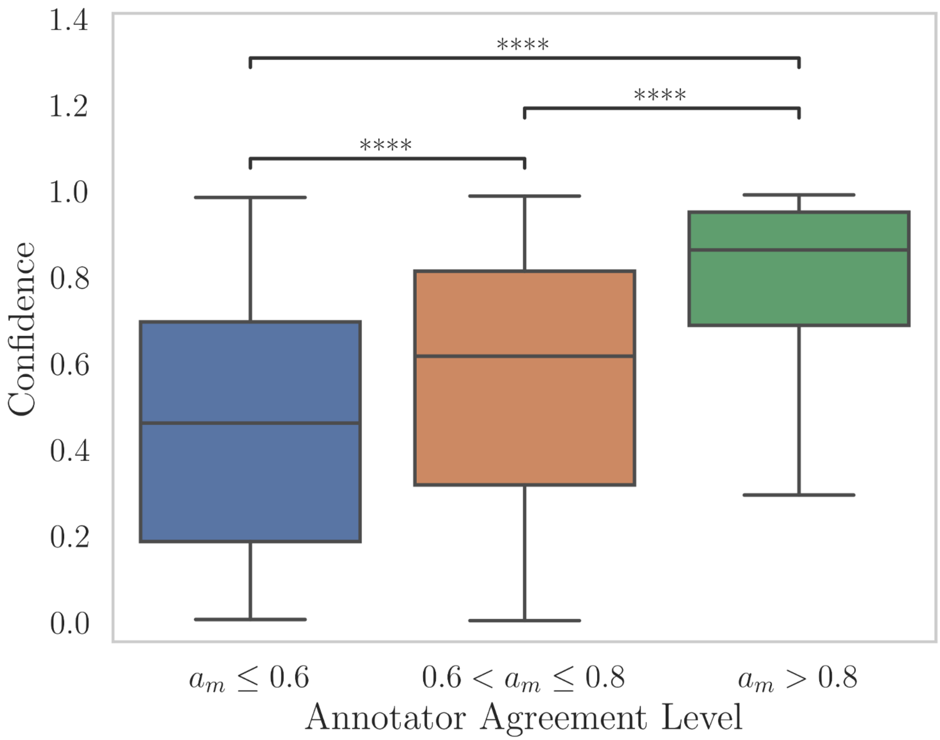
Single-GT model

Confidence vs Annotator Agreement Level

SBIC

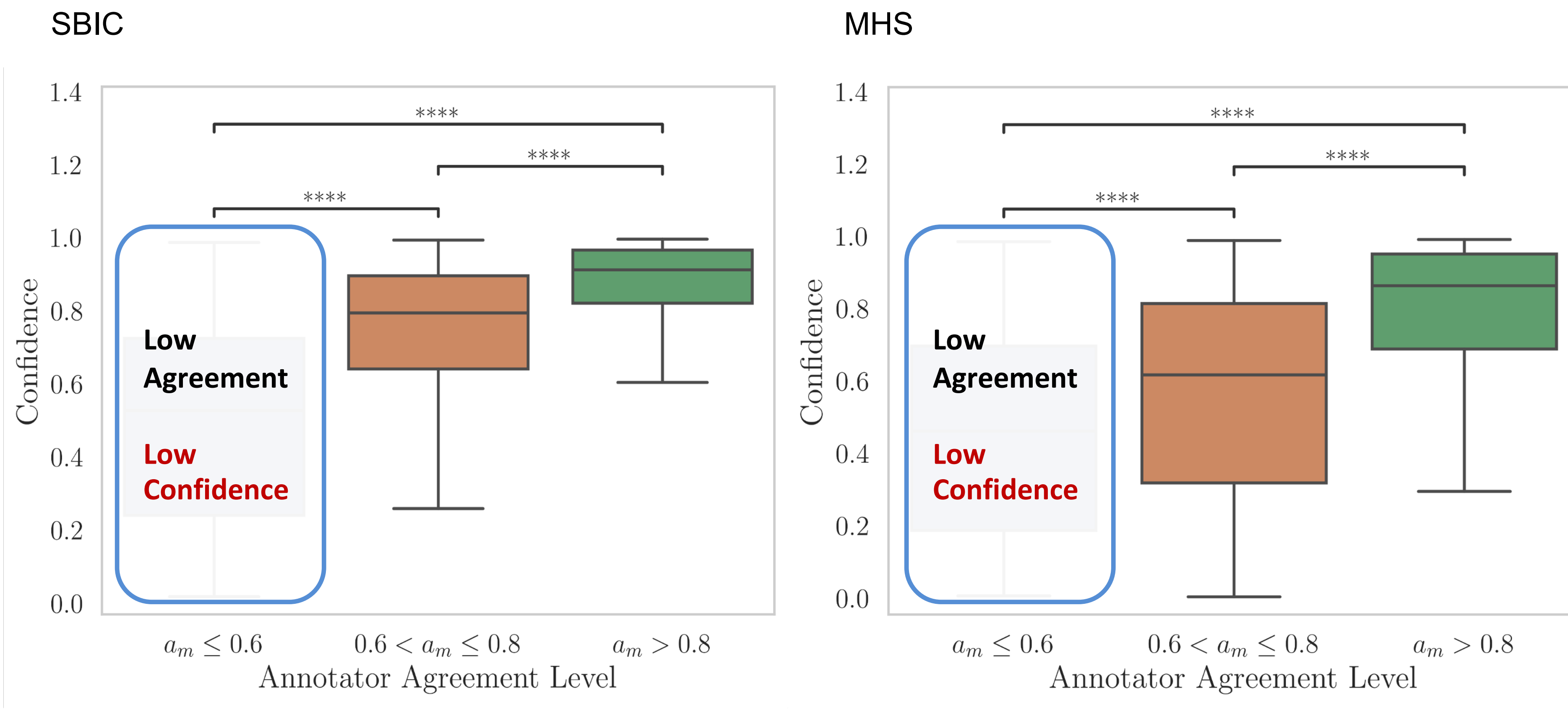


MHS



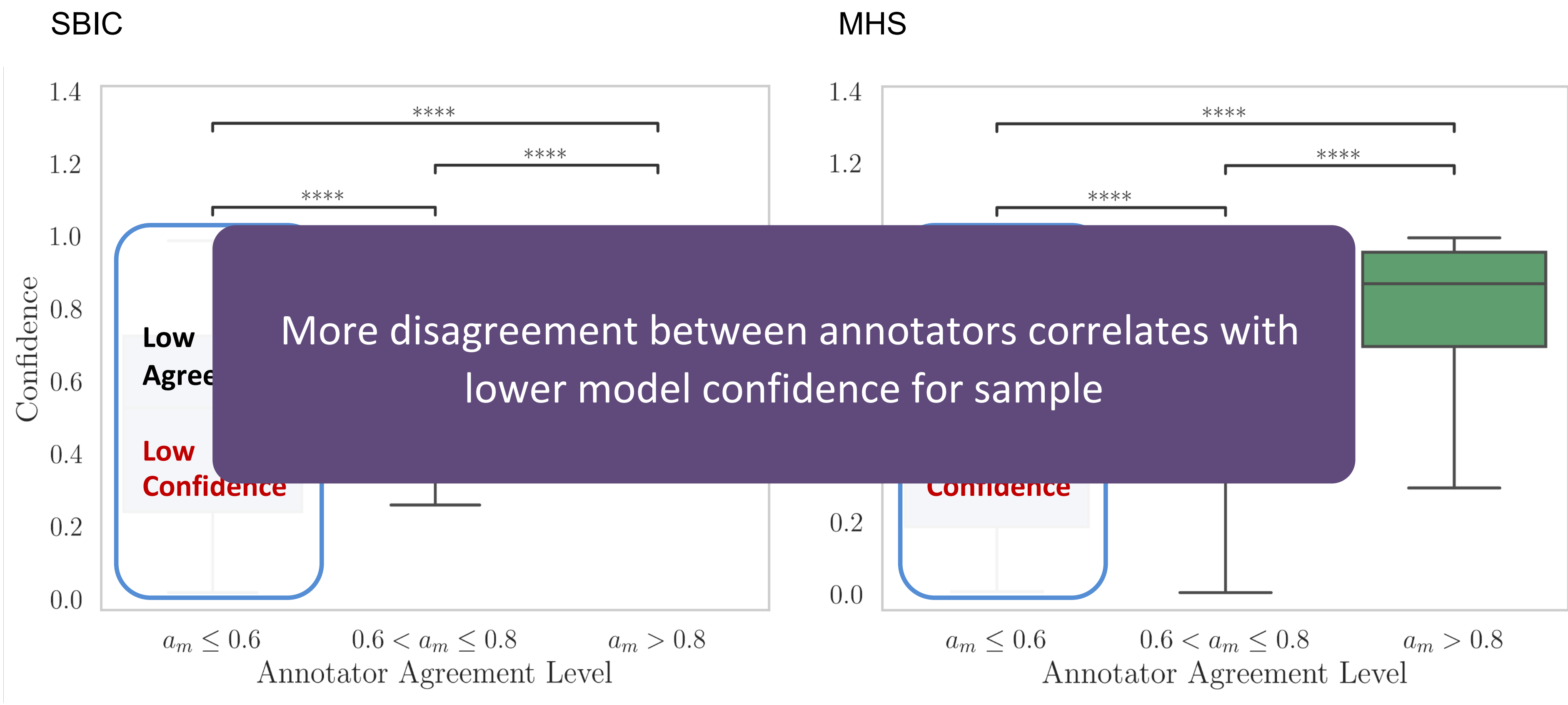
Single-GT model

Confidence vs Annotator Agreement Level



Single-GT model

Confidence vs Annotator Agreement Level



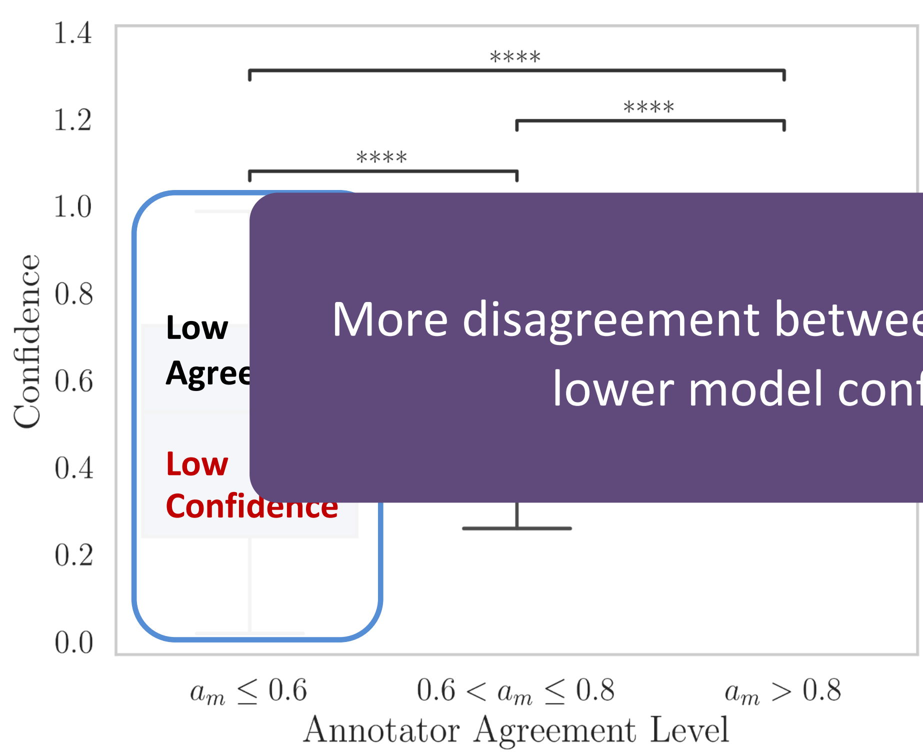
More disagreement between annotators correlates with lower model confidence for sample

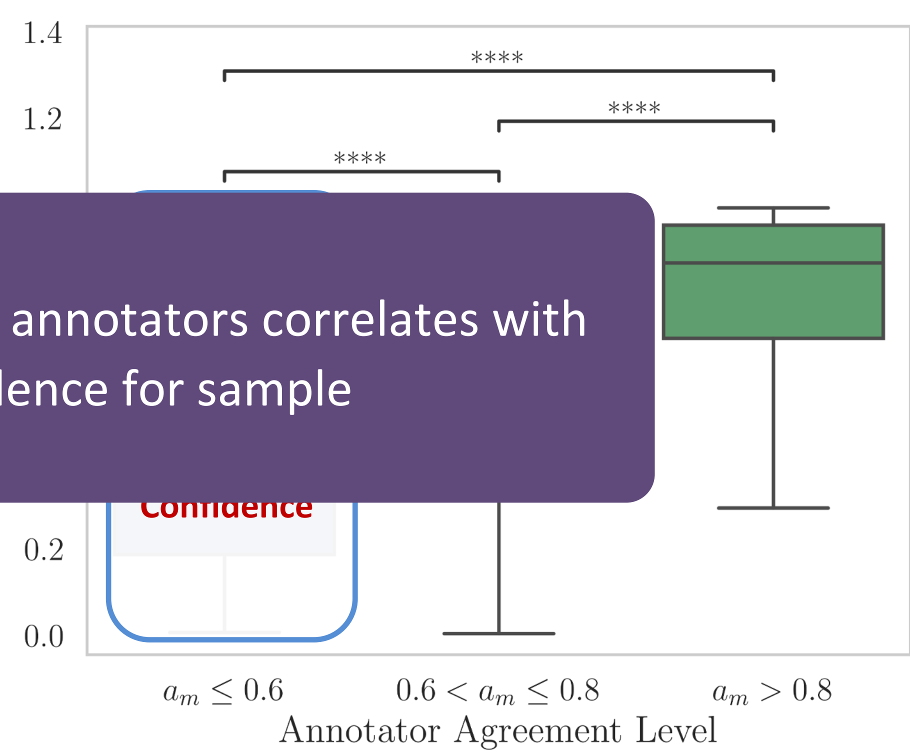
Single-GT model

Confidence vs Annotator Agreement Level

SBIC

MHS





More disagreement between annotators correlates with lower model confidence for sample

Does learning from raw annotations enhance the model's confidence for the high disagreement instances?

Does learning from raw annotations enhance the model's confidence for the high disagreement instances?

Data	Text1	Ann1	Ann2	Ann3	Ann4
	Text2	Ann1	Ann2	Ann3	Ann4

Does learning from **raw annotations** enhance the model's confidence for the high disagreement instances?

Requirement: Models that take raw annotations as input

Data

Text1	Ann1	Ann2	Ann3	Ann4
Text2	Ann1	Ann2	Ann3	Ann4

Does learning from **raw annotations** enhance the model's confidence for the high disagreement instances?

Multi-GT model

Multiple Ground Truth Model: Each annotation by an annotator is considered a ground truth. We consider DISCO (Weerasooriya et al., 2023) for our study

Requirement: Models that take raw annotations as input

Data

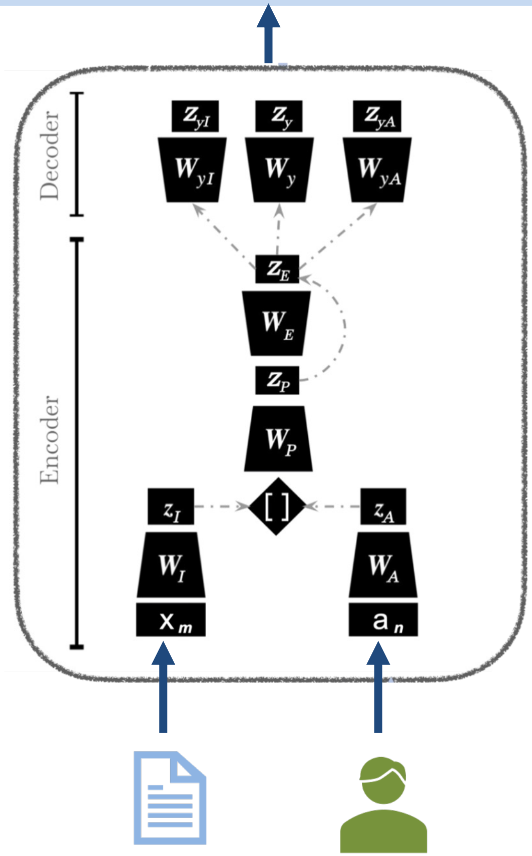
Text1	Ann1	Ann2	Ann3	Ann4
Text2	Ann1	Ann2	Ann3	Ann4

Does learning from **raw annotations** enhance the model's confidence for the high disagreement instances?

Multi-GT model

Multiple Ground Truth Model: Each annotation by an annotator is considered a ground truth. We consider **DISCO** (Weerasooriya et al., 2023) for our study

Prediction for Text, Annotator Id



Input Text

Annotator Id

Requirement: Models that take raw annotations as input

Data

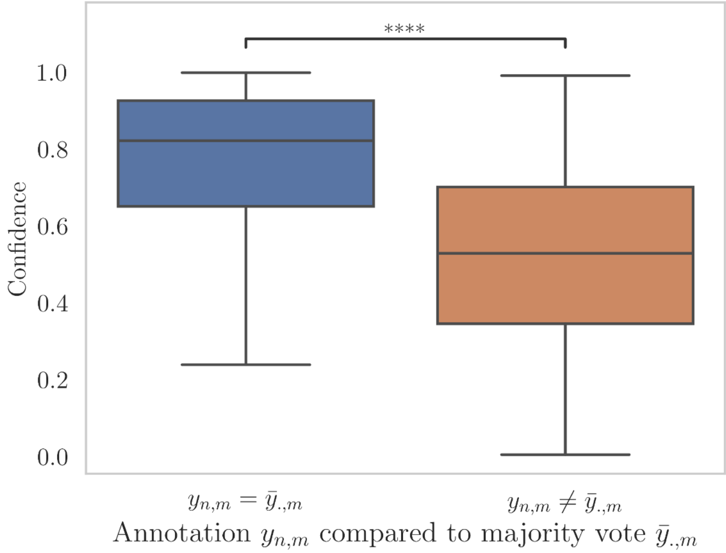
Text1	Ann1	Ann2	Ann3	Ann4
Text2	Ann1	Ann2	Ann3	Ann4

Does learning from **raw annotations** enhance the model's confidence for the high disagreement samples?

Multi-GT model

Model Confidence vs Agreement with Majority vote

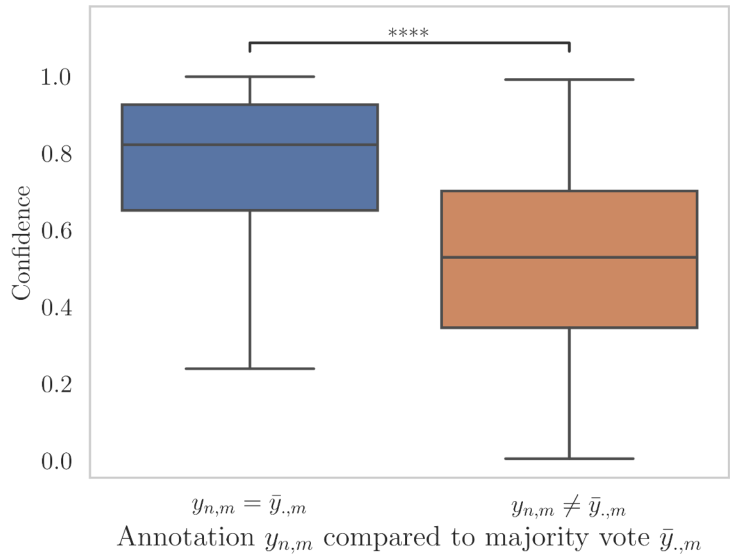
MDA



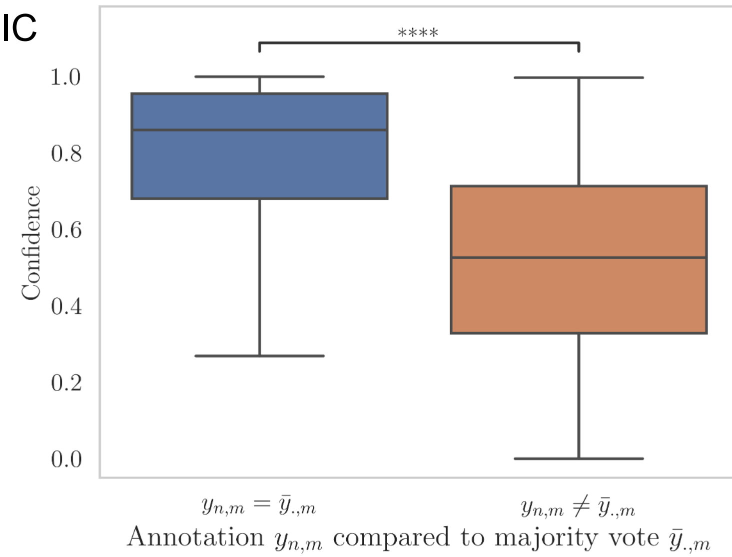
Multi-GT model

Model Confidence vs Agreement with Majority vote

MDA



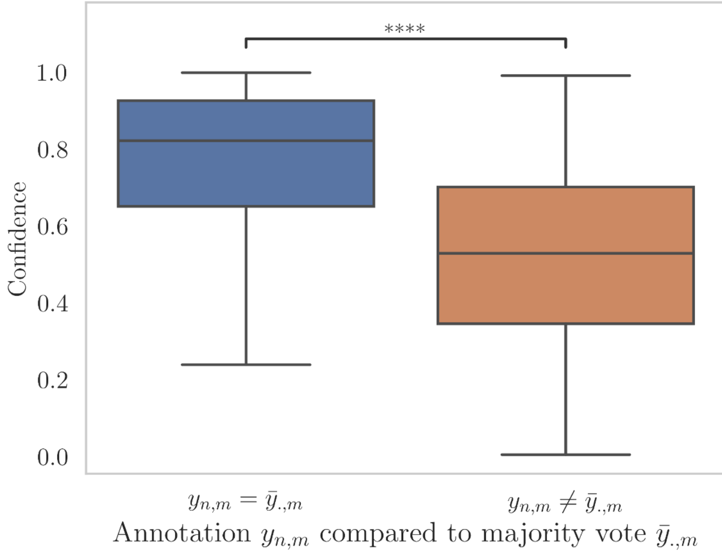
SBIC



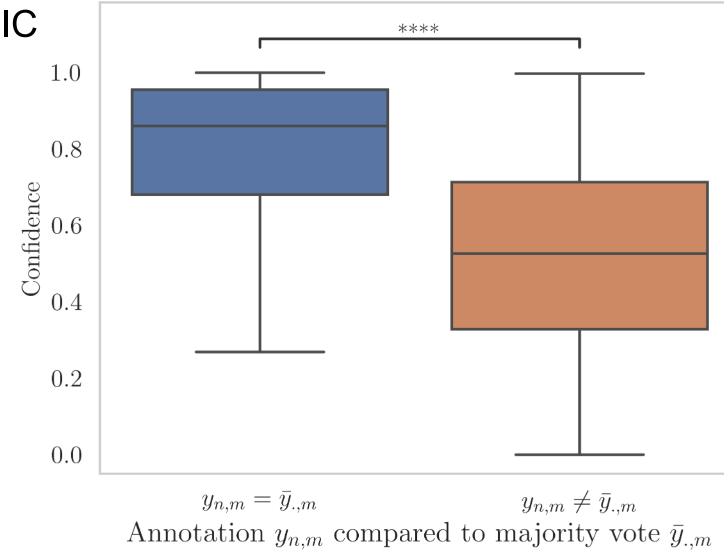
Multi-GT model

Model Confidence vs Agreement with Majority vote

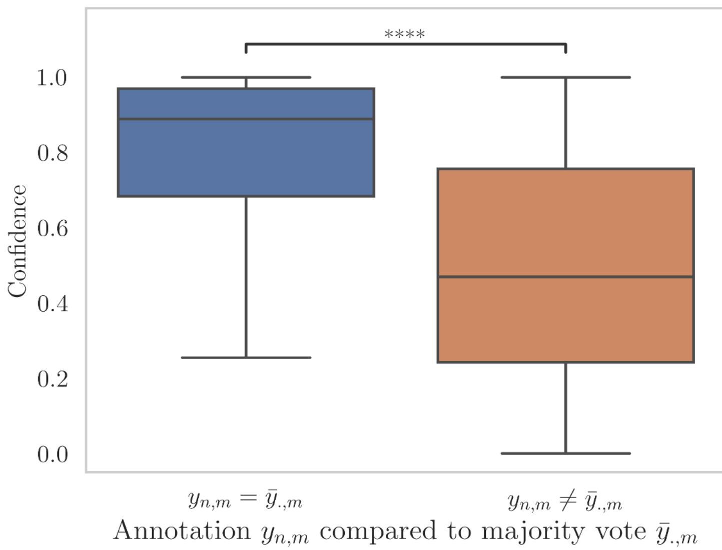
MDA



SBIC



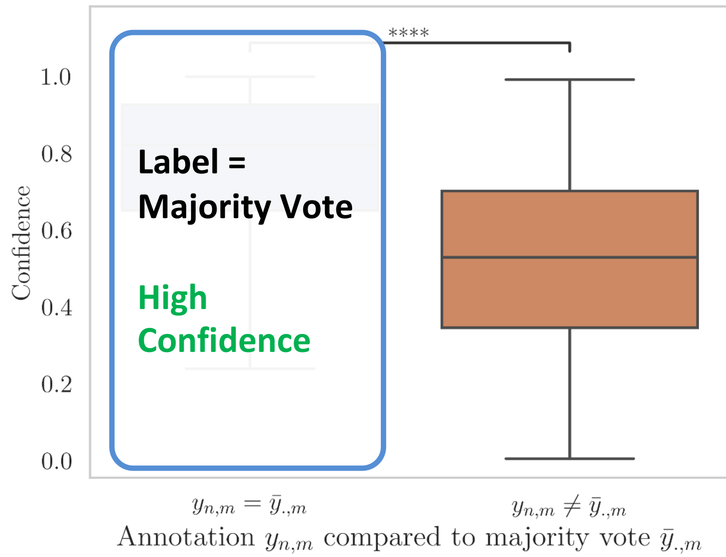
MHS



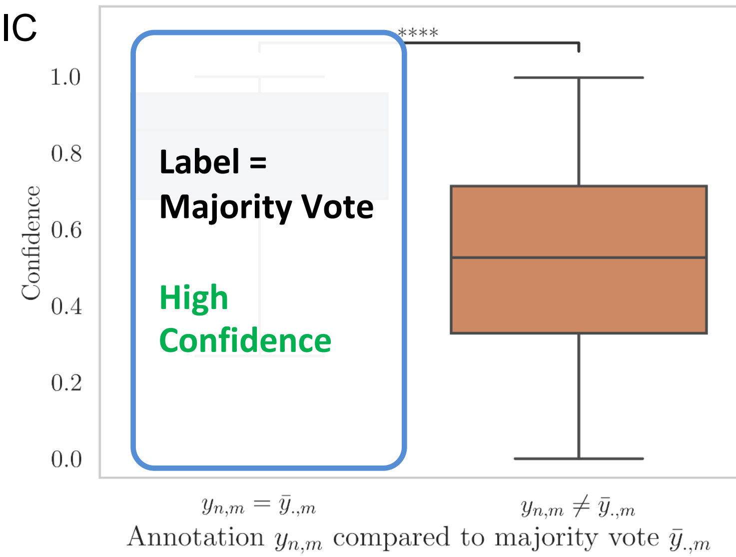
Multi-GT model

Model Confidence vs Agreement with Majority vote

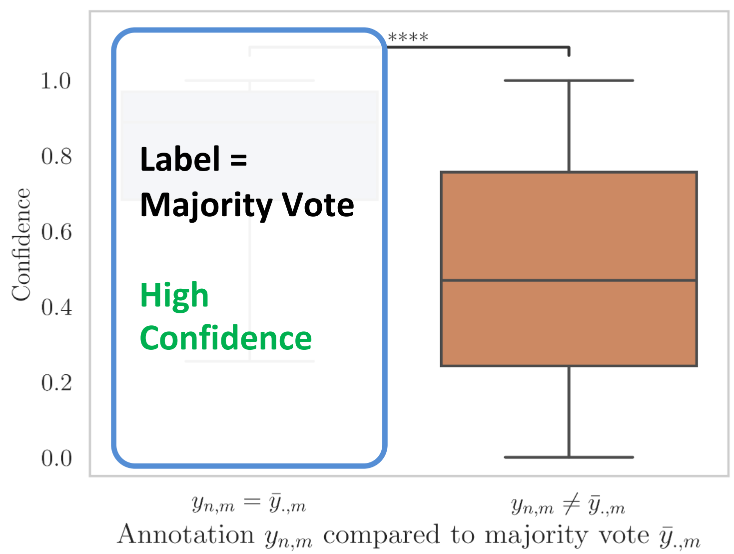
MDA



SBIC



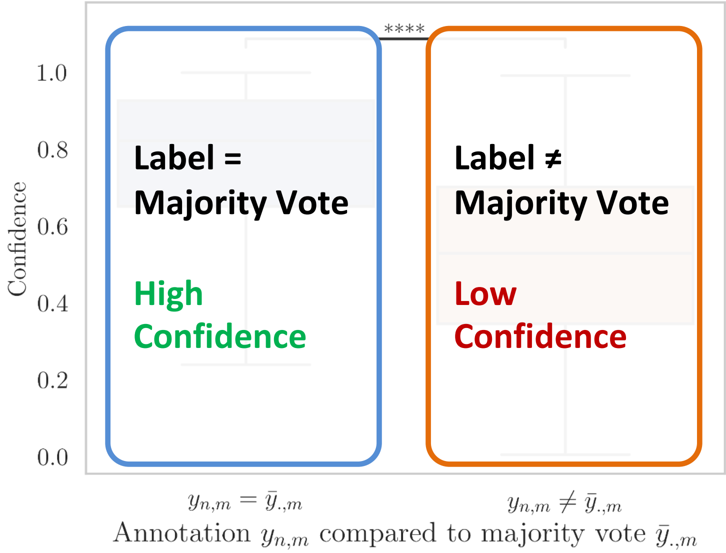
MHS



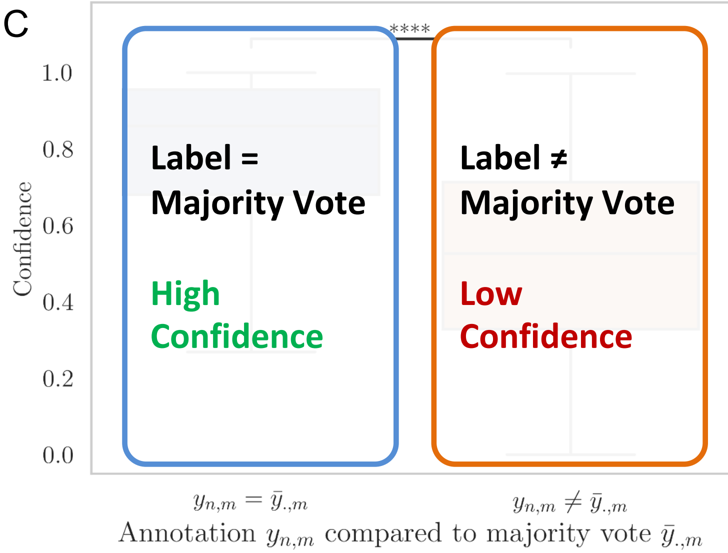
Multi-GT model

Model Confidence vs Agreement with Majority vote

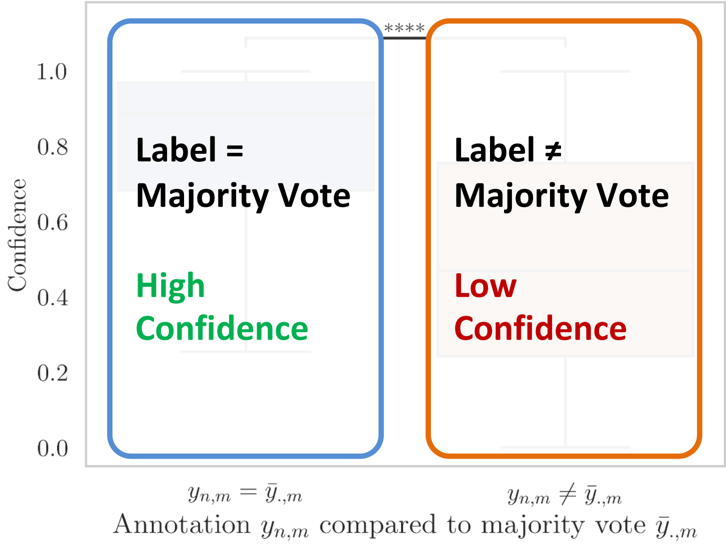
MDA



SBIC



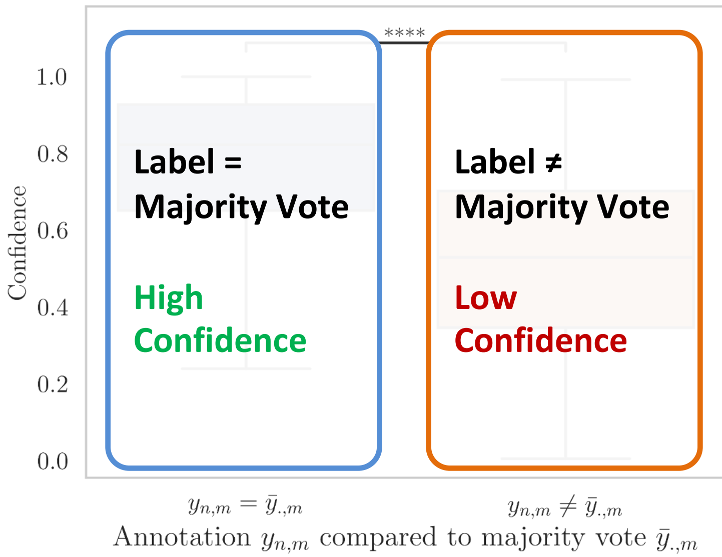
MHS



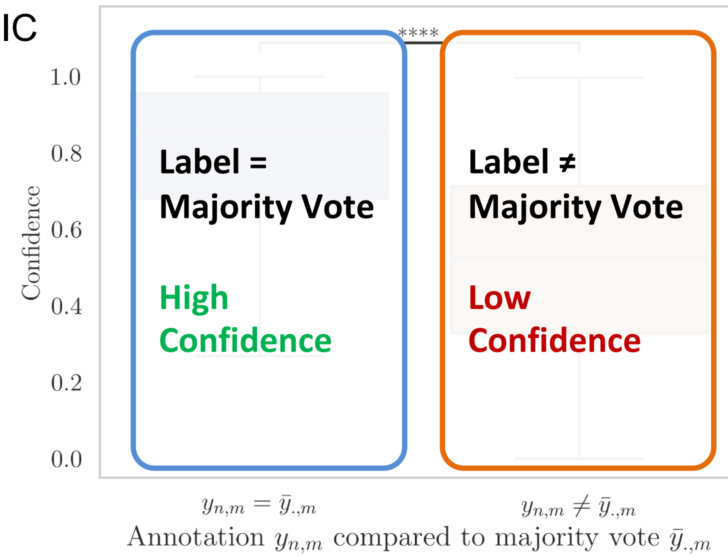
Multi-GT model

Model Confidence vs Agreement with Majority vote

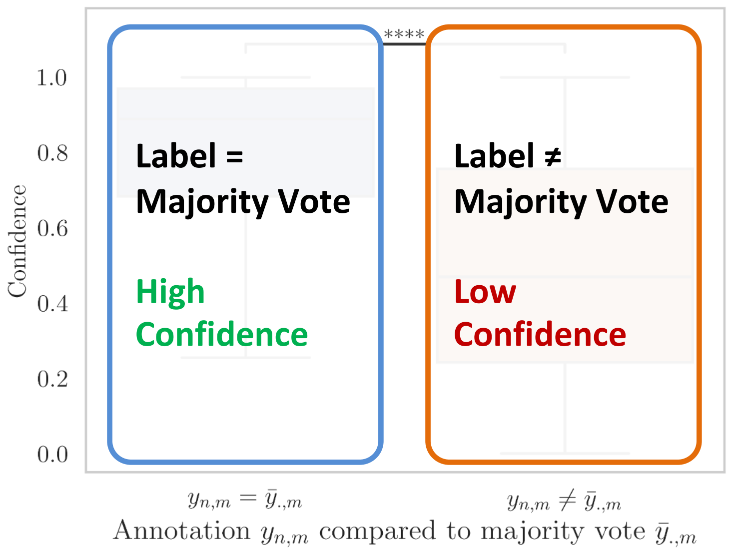
MDA



SBIC



MHS

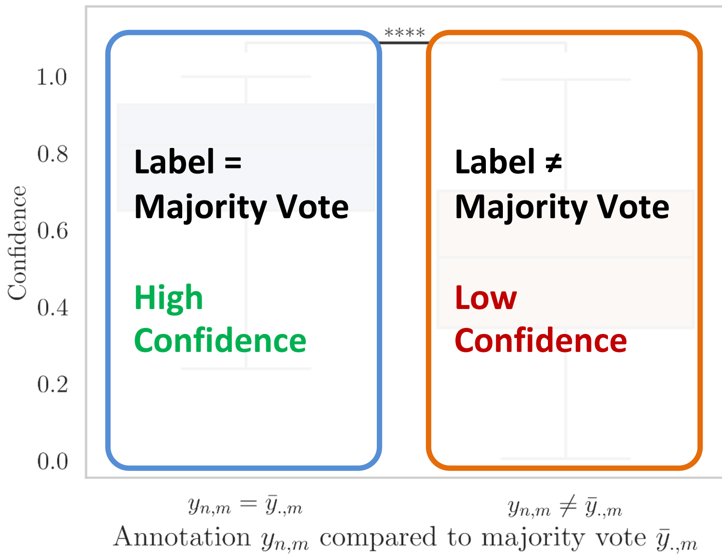


Consistent with observed trends in Single-GT model, higher agreement correlates with higher confidence

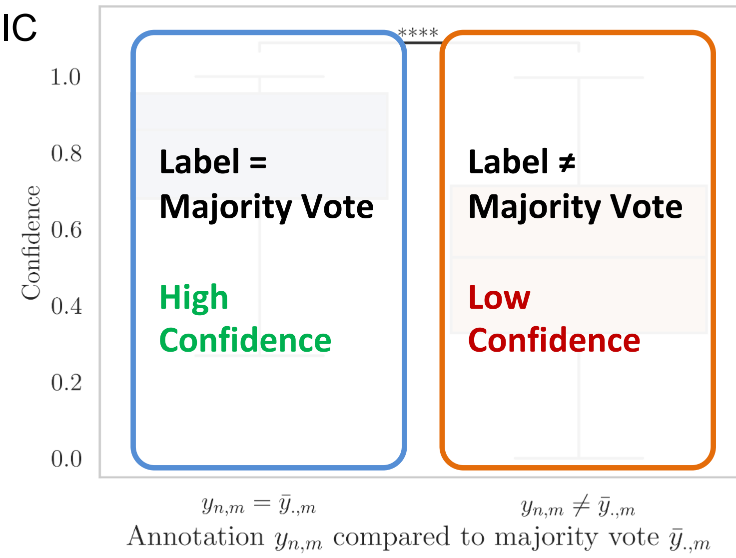
Multi-GT model

Model Confidence vs Agreement with Majority vote

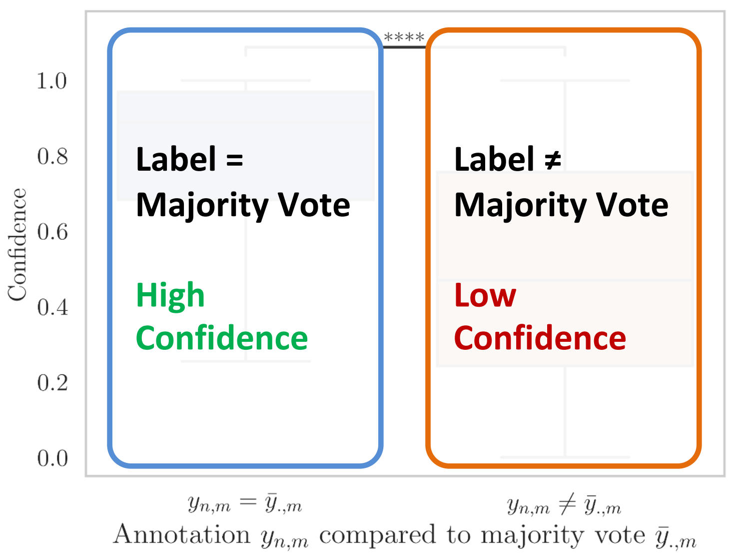
MDA



SBIC



MHS

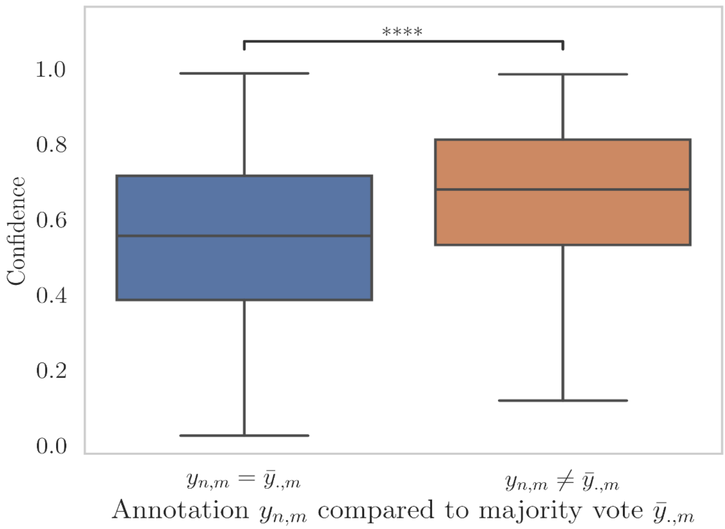


Consistent with observed trends in Single-GT model, higher agreement correlates with higher confidence

Can Multi-GT improve on high disagreement samples from Single-GT?

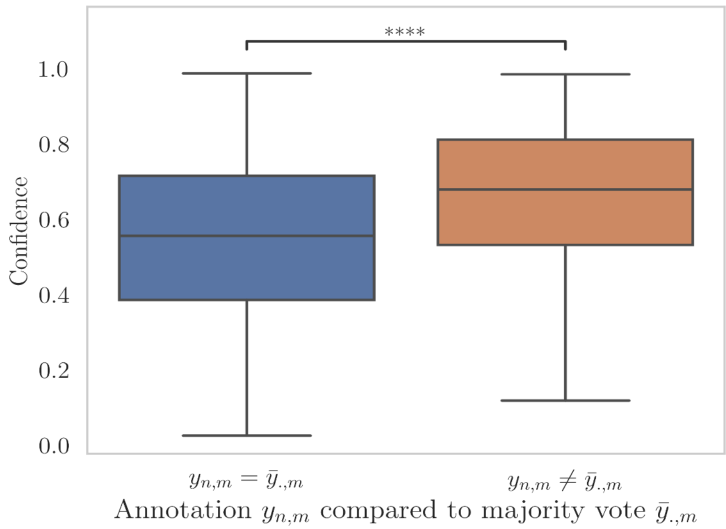
Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

MDA

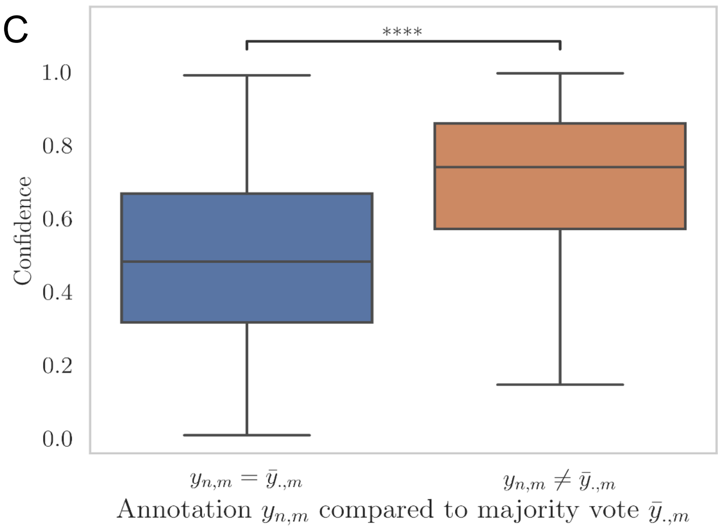


Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

MDA

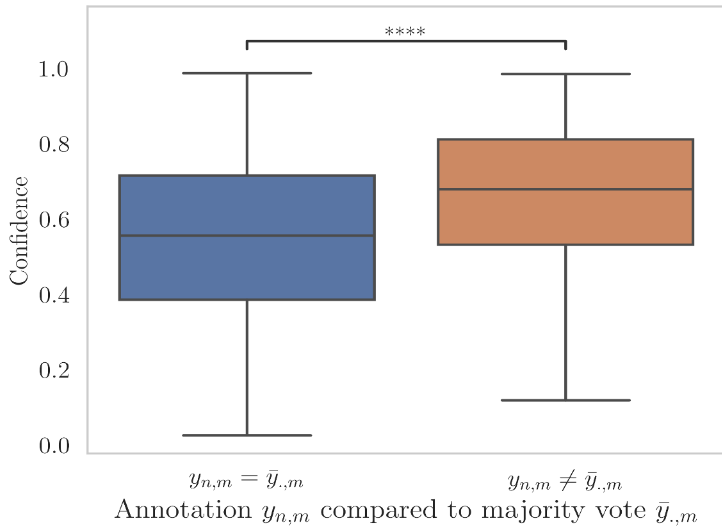


SBIC

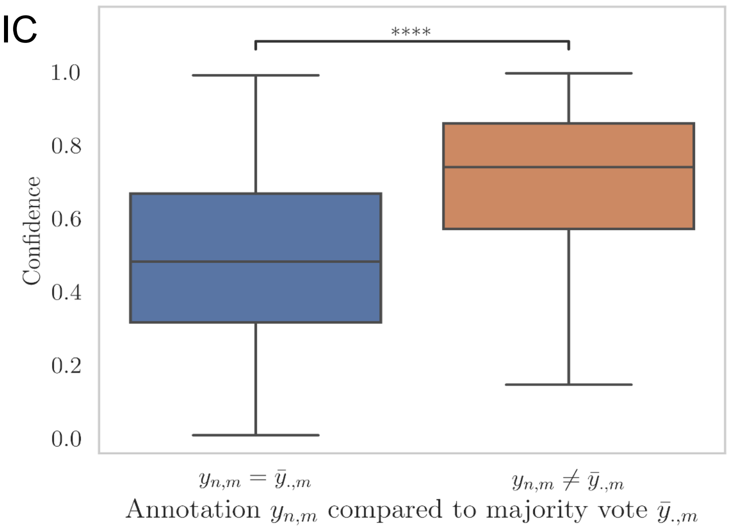


Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

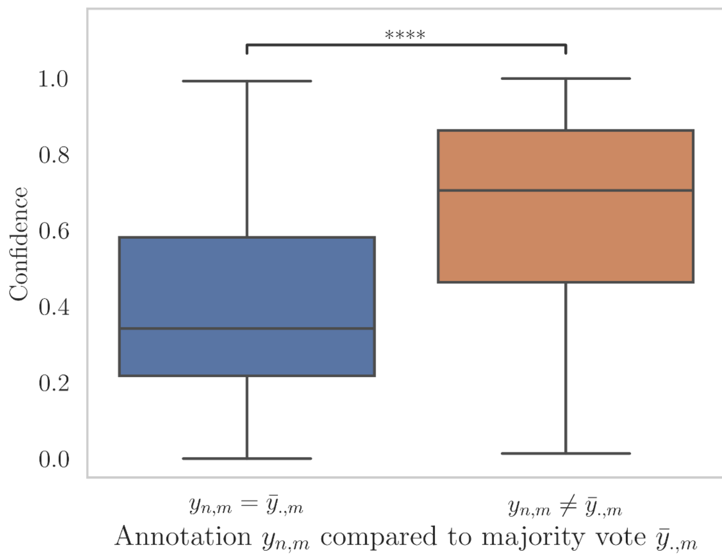
MDA



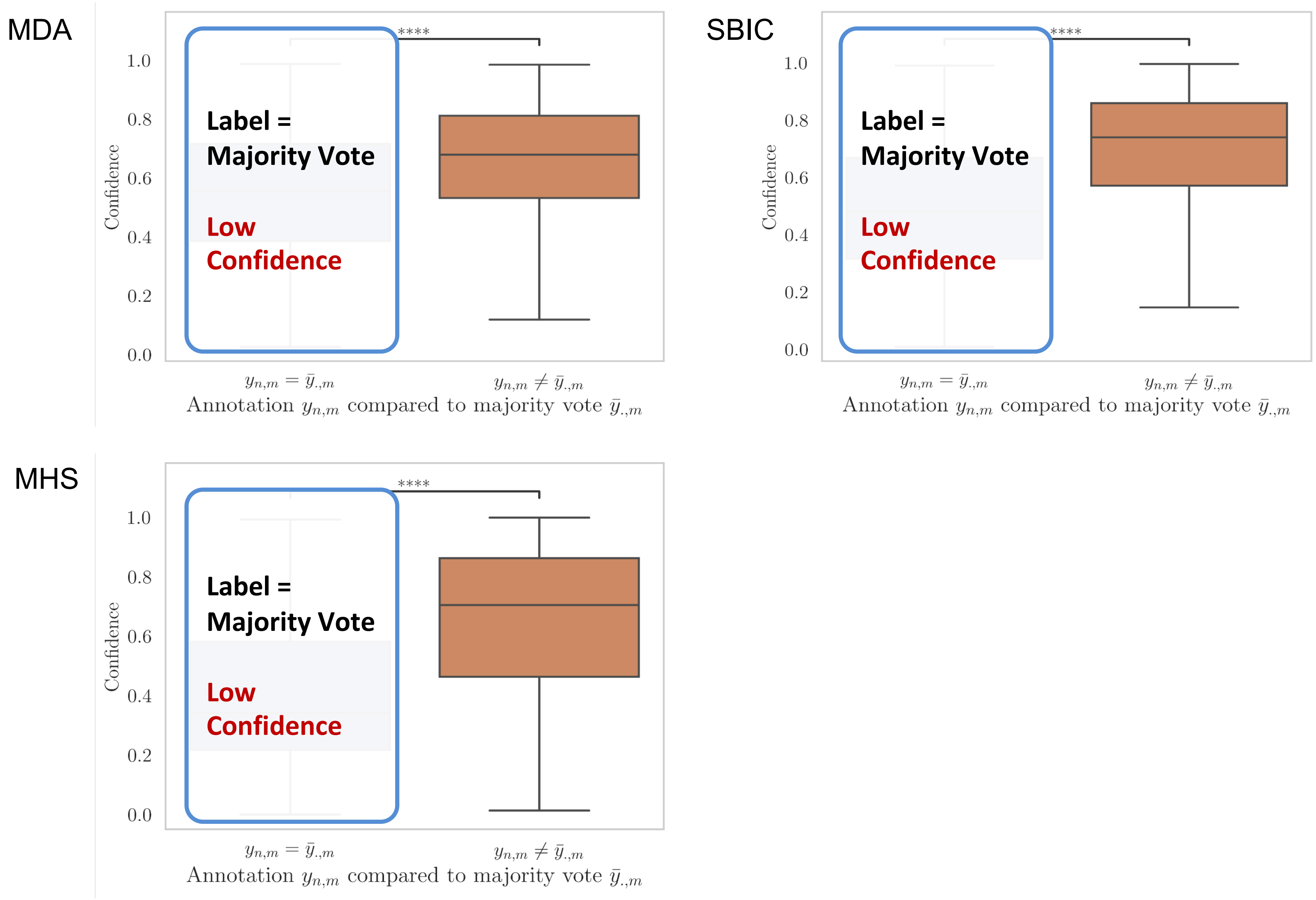
SBIC



MHS

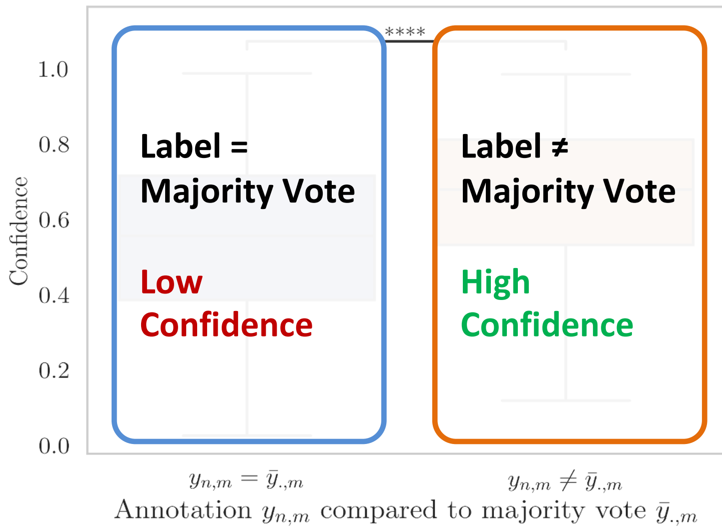


Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

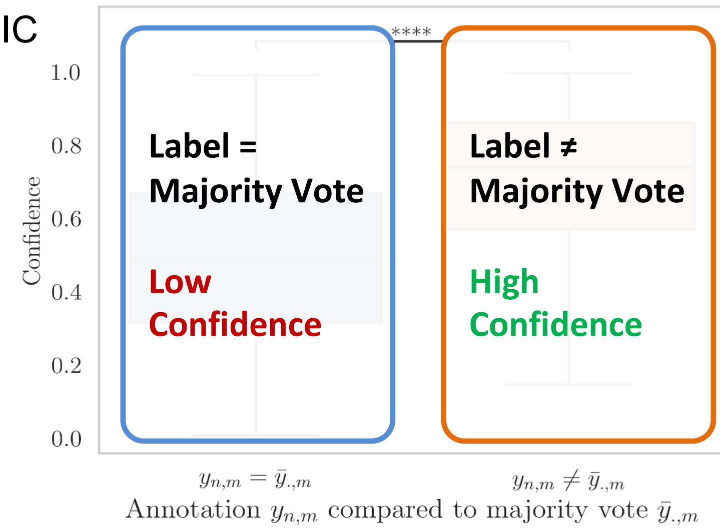


Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

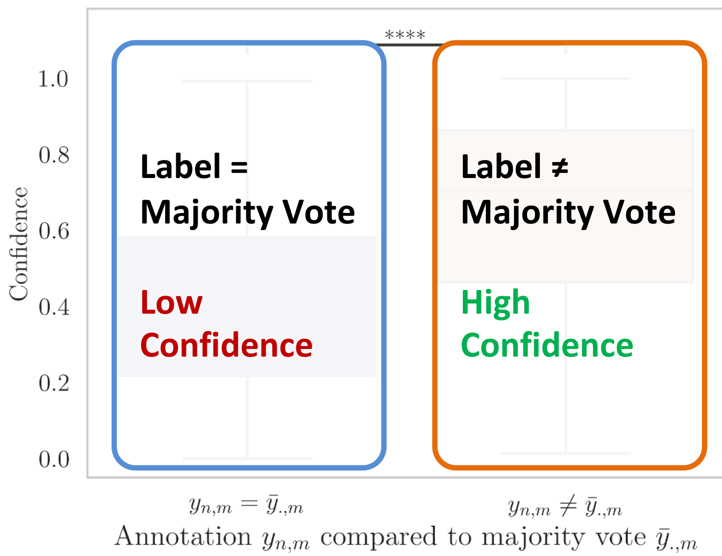
MDA



SBIC

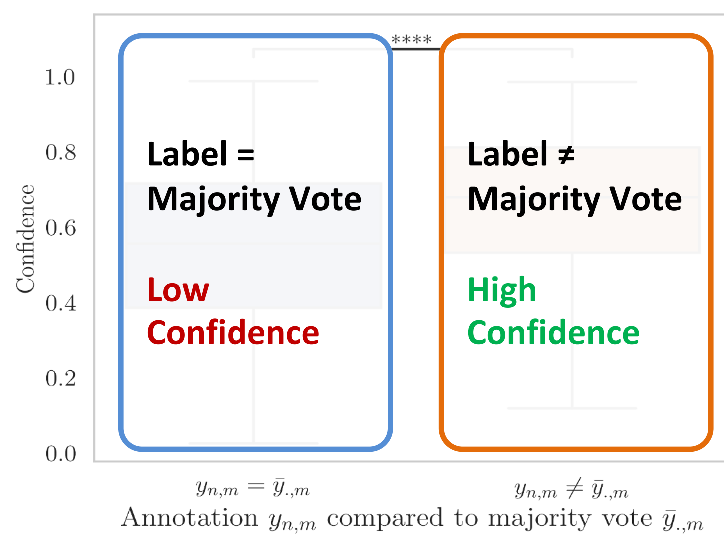


MHS

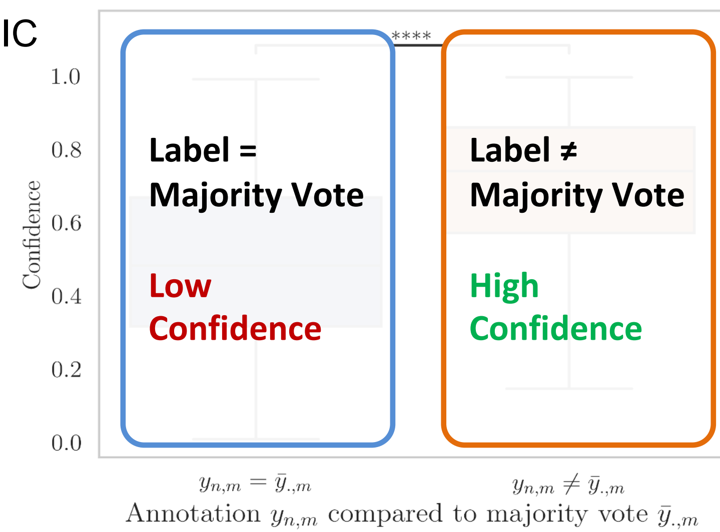


Multi-GT model (Low confidence (< 0.5) samples in Single-GT)

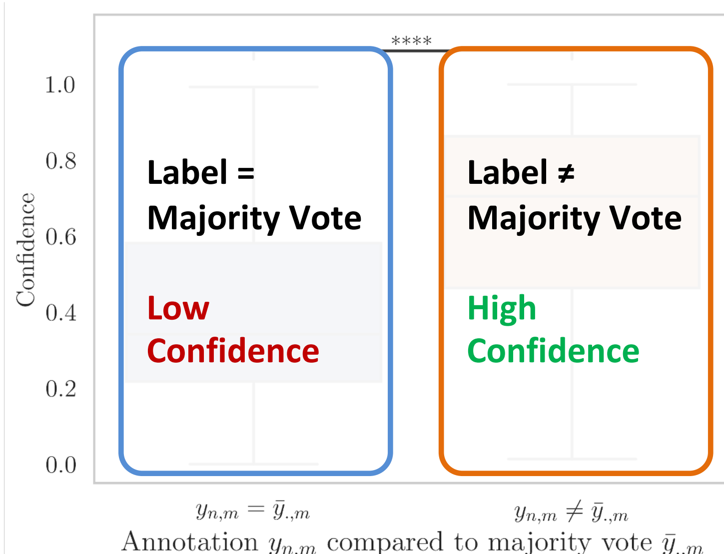
MDA



SBIC

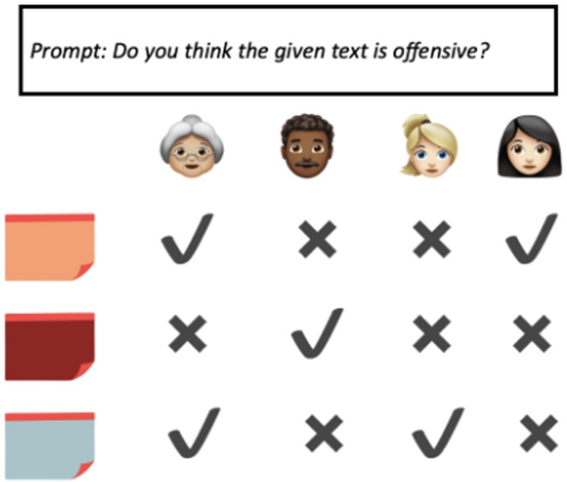


MHS

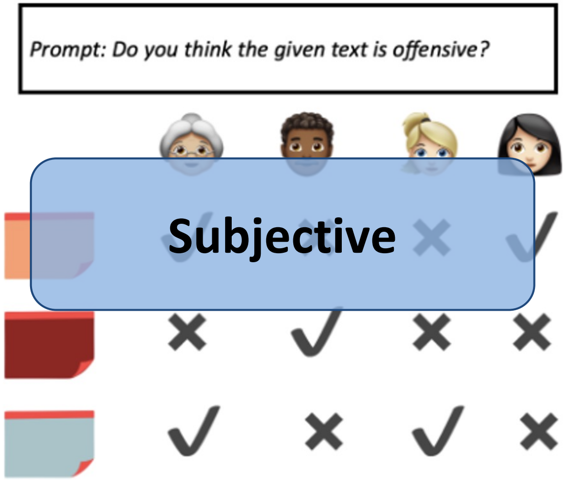


Provided the annotations that were discarded as noise, DISCO learns valuable signals boosting confidence on these samples

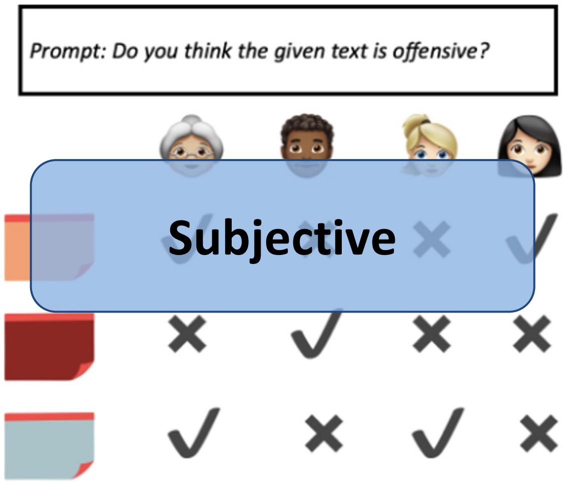
Multi-GT model



Multi-GT model

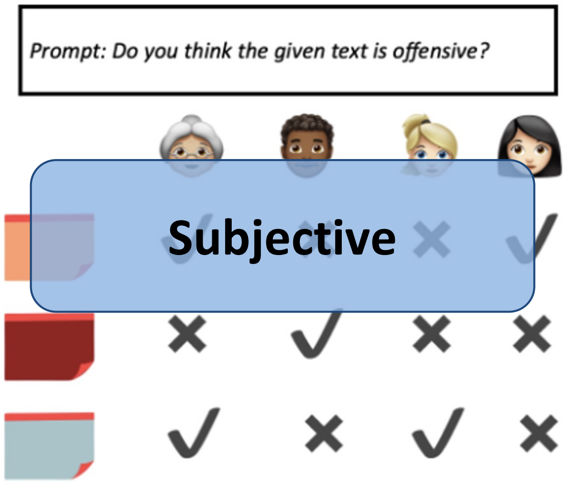


Multi-GT model



Different perspectives essential for subjective tasks

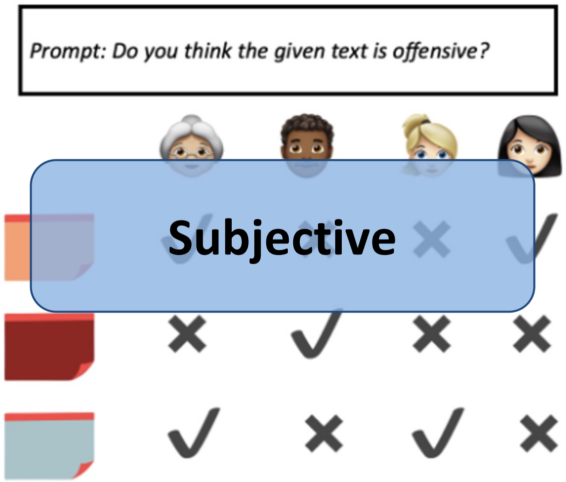
Multi-GT model



Different perspectives essential for subjective tasks

Can the model learn multiple annotators' perspectives?

Multi-GT model

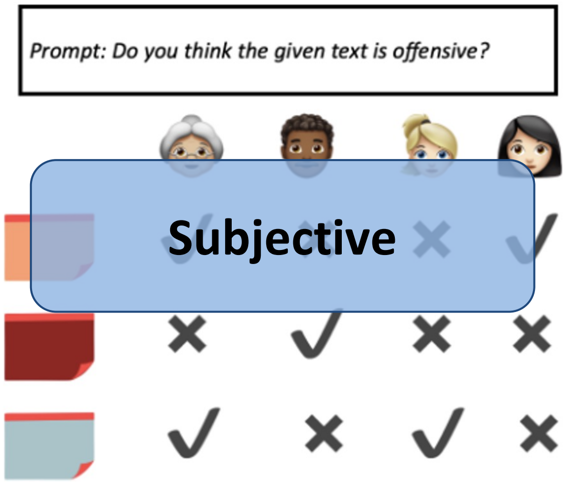


Different perspectives essential for subjective tasks

Can the model learn multiple annotators' perspectives?

Ability to predict different labels for the same text and different annotators?

Multi-GT model



Different perspectives essential for subjective tasks

Can the model learn multiple annotators' perspectives?

Ability to predict different labels for the same text and different annotators?

Only focus on samples with $a_m < 1.0$

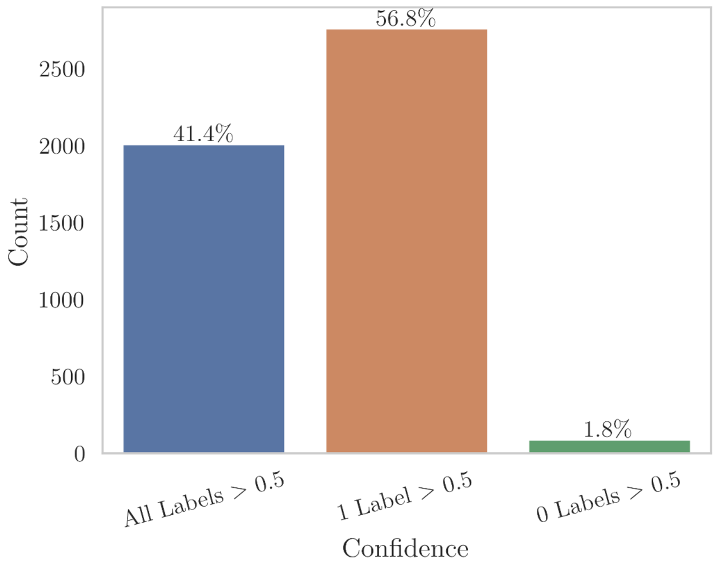
Multi-GT model

Can it learn multiple annotators' perspectives?

Multi-GT model

Can it learn multiple annotators' perspectives?

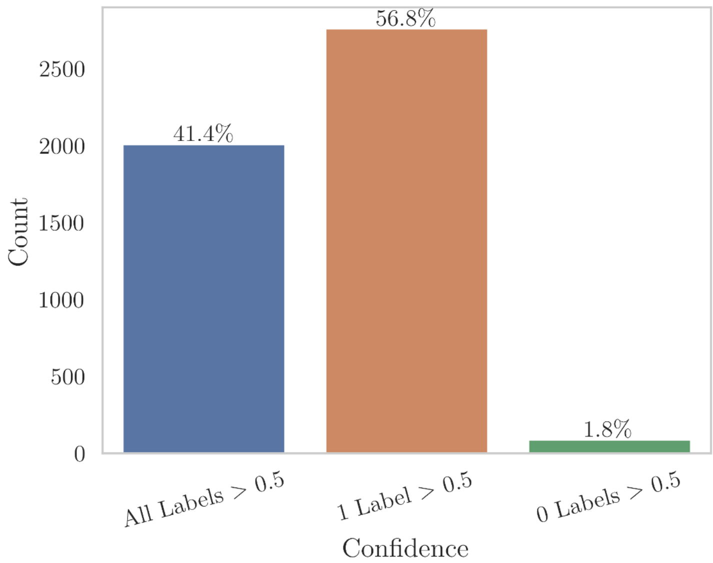
MDA



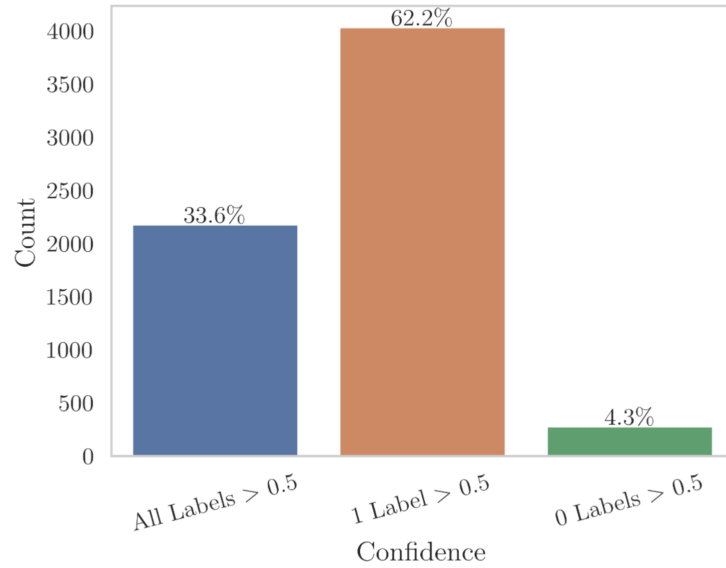
Multi-GT model

Can it learn multiple annotators' perspectives?

MDA



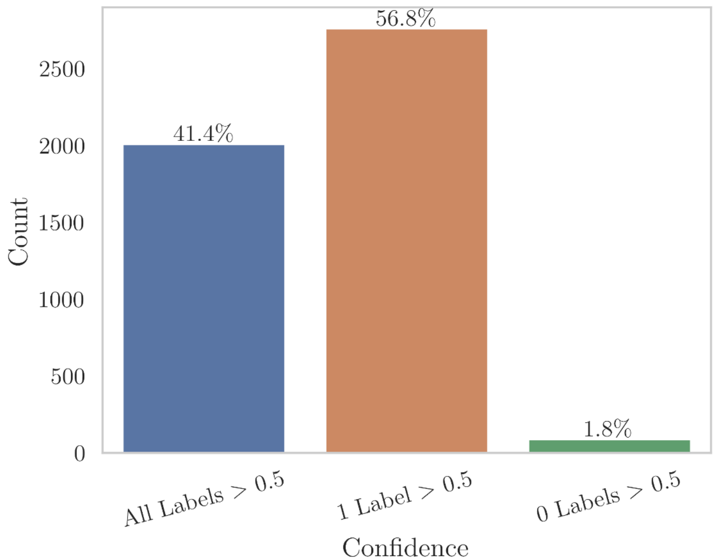
SBIC



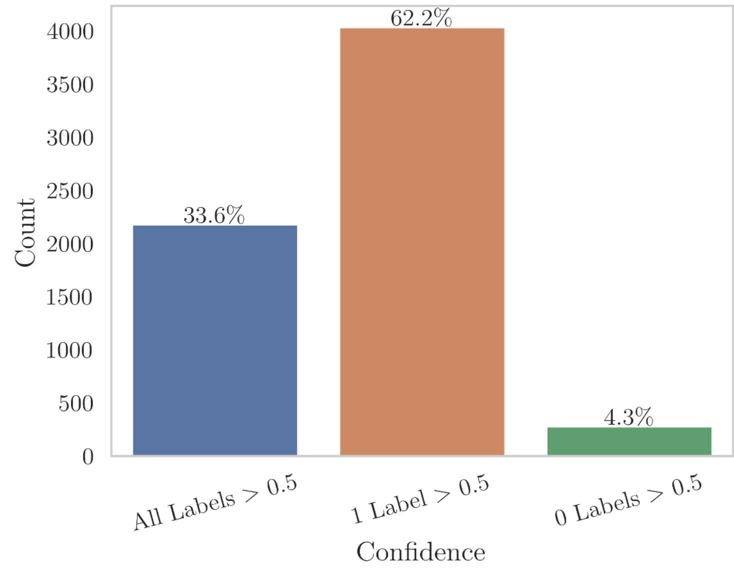
Multi-GT model

Can it learn multiple annotators' perspectives?

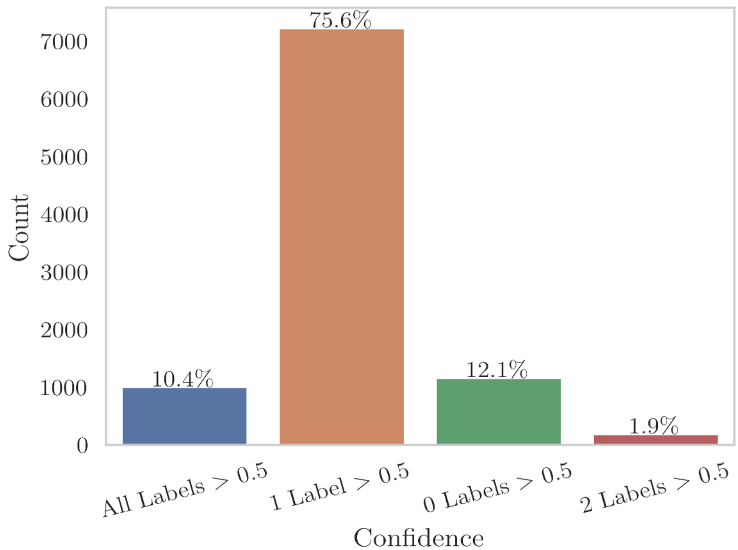
MDA



SBIC



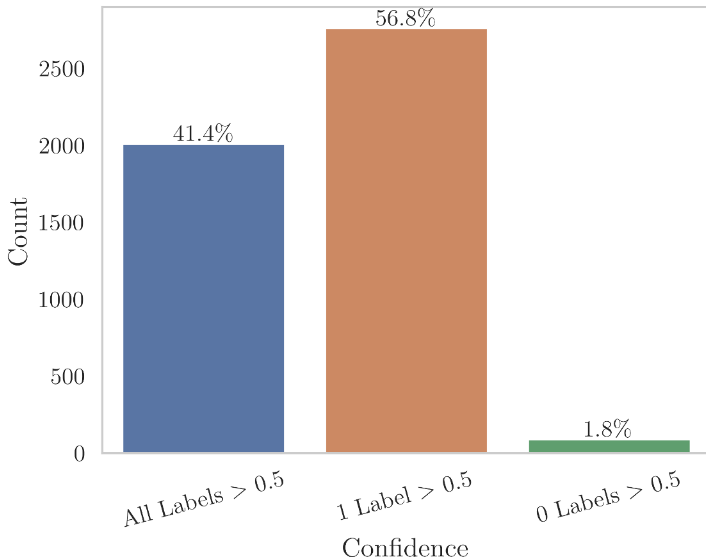
MHS



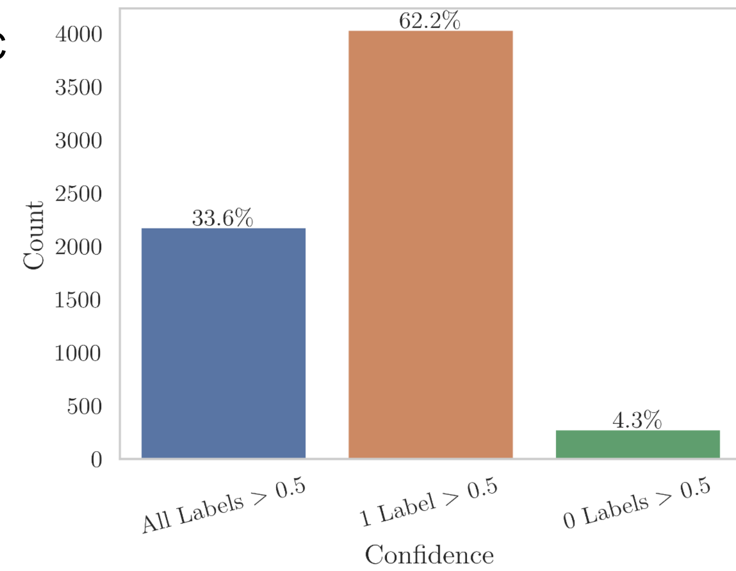
Multi-GT model

Can it learn multiple annotators' perspectives?

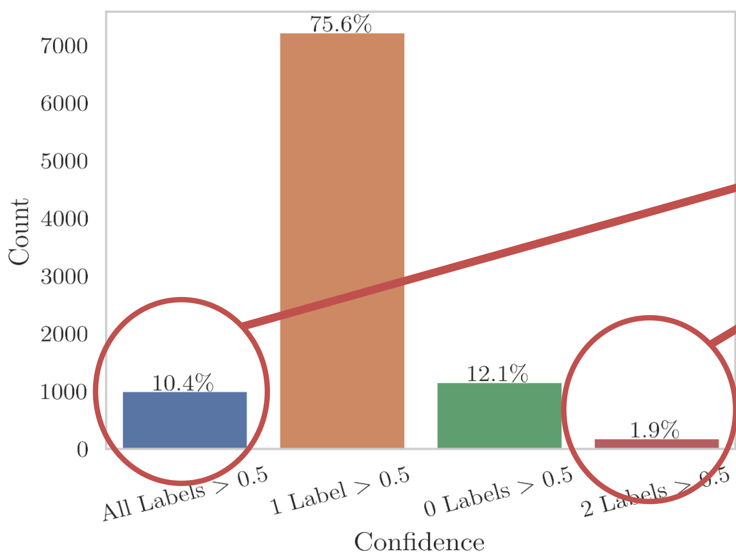
MDA



SBIC



MHS

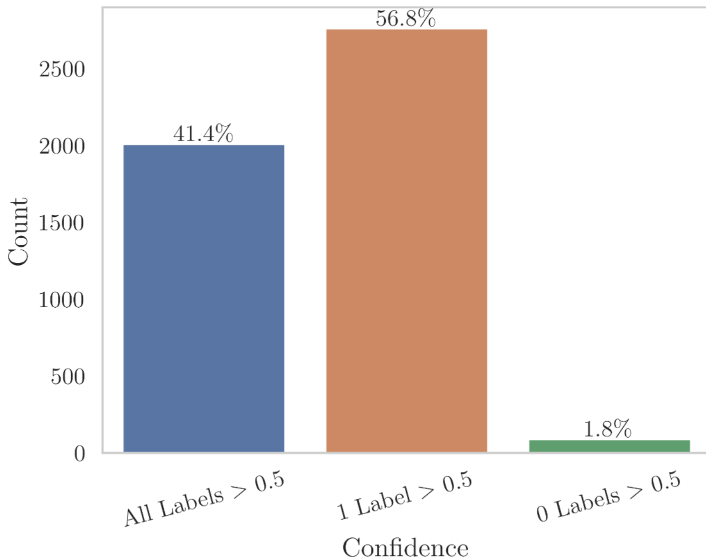


Finds it difficult to learn multiple labels for MHS

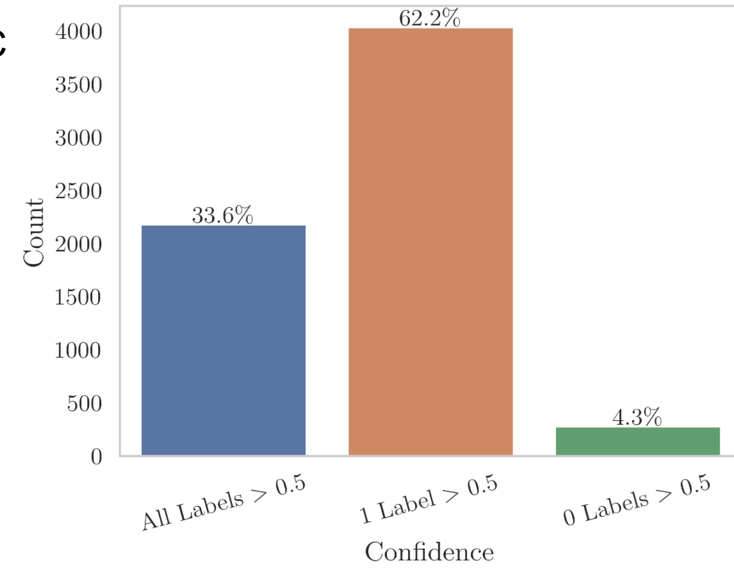
Multi-GT model

Can it learn multiple annotators' perspectives?

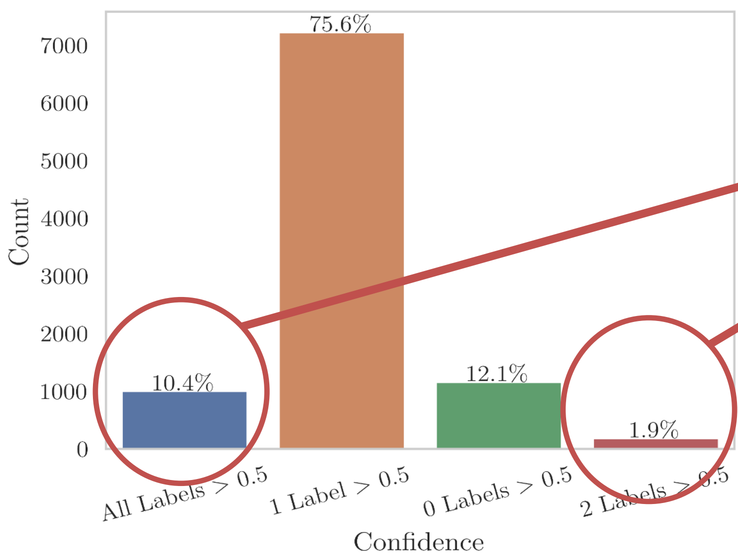
MDA



SBIC



MHS



Finds it difficult to learn multiple labels for MHS

This dataset is extra challenging because avg number of annotations per annotator is ~ 17!

Takeaways

Noise vs **Bias**?

More disagreement between annotators correlates with low model confidence

Takeaways

Noise vs **Bias**?

More disagreement between annotators correlates with low model confidence

Majority vote captures just a single perspective - insufficient for subjective tasks

Takeaways

Noise vs **Bias**?

More disagreement between annotators correlates with low model confidence

Majority vote captures just a single perspective - insufficient for subjective tasks

Multi-Gt model effectively utilizes minority vote annotations that are usually discarded as noise

Noise vs **Bias**?

More disagreement between annotators correlates with low model confidence

Majority vote captures just a single perspective - insufficient for subjective tasks

Multi-Gt model effectively utilizes minority vote annotations that are usually discarded as noise

Number of annotations per annotator important in modelling their perspective