TECHNISCHE
UNIVERSITÄT
DARMSTADT

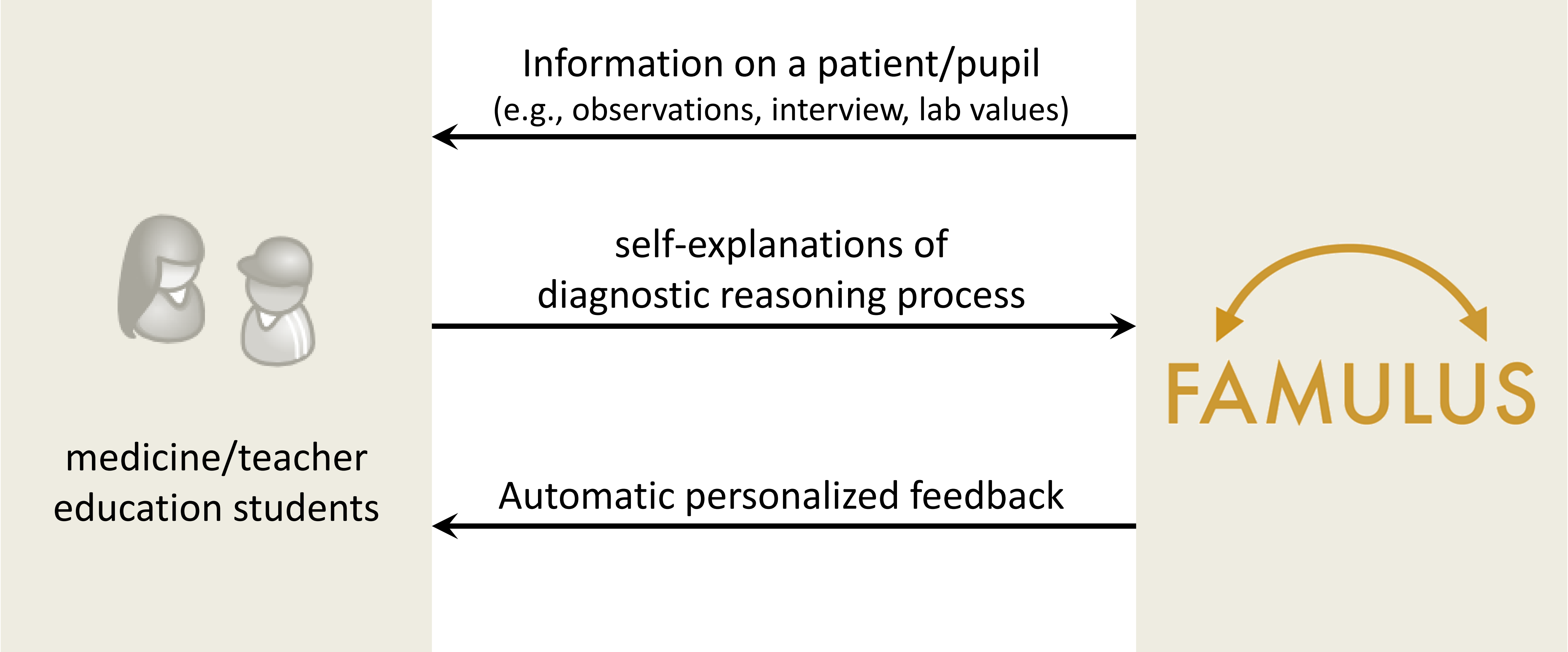
Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains

Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer,
Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych

FAMULUS

http://famulus-project.de

Learning to Diagnose



Information on a patient/pupil
(e.g., observations, interview, lab values)

self-explanations of
diagnostic reasoning process

Automatic personalized feedback

medicine/teacher
education students

Diagnostic Reasoning

The patient reports to be lethargic and feverish. From the anamnesis I learned that he had purulent tonsilitis and is still suffering from symptoms. I first performed some laboratory tests and notice the decreased number of lymphocytes, which can be indicative of a bone marrow disease or an HIV infection. The HIV test is positive. However, the results from the blood cultures are negative, so it is a virus, parasite, or a fungal infection causing the symptoms.

🟦 Hypothesis generation (HG)	🟨 Evidence evaluation (EE)
🟩 Drawing conclusions (DC)	🟪 Evidence generation (EG)

Research Question

The patient reports to be lethargic and feverish. From the anamnesis I learned that he had purulent tonsilitis and is still suffering from symptoms. I first performed some laboratory tests and notice the decreased number of lymphocytes, which can be indicative of a bone marrow disease or an HIV infection. The HIV

How can we improve this hard and time-consuming annotation task?

the symptoms.

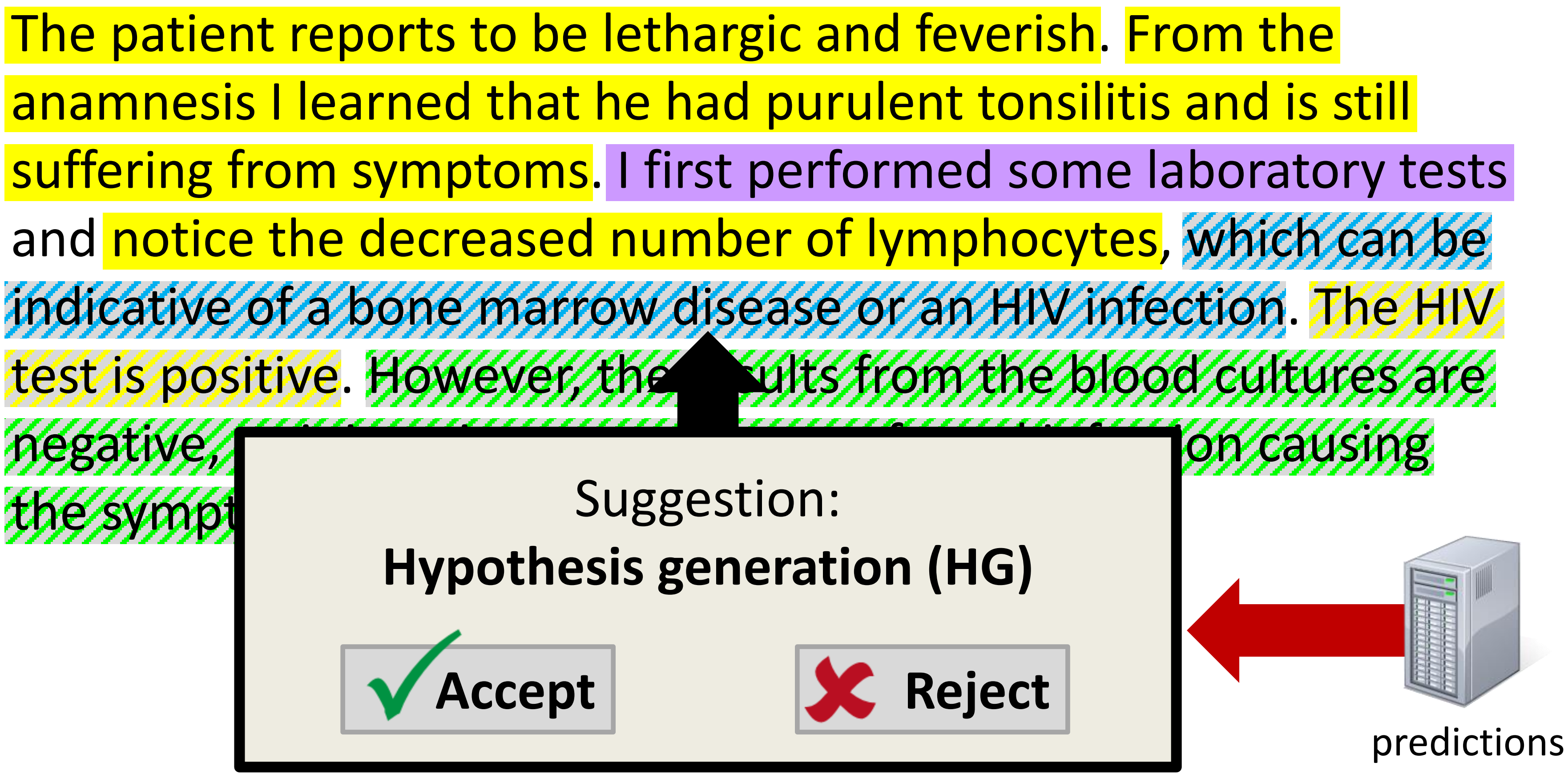
Hypothesis generation (HG)

Evidence evaluation (EE)

Drawing conclusions (DC)

Evidence generation (EG)

Annotation Suggestions



The patient reports to be lethargic and feverish. From the anamnesis I learned that he had purulent tonsilitis and is still suffering from symptoms. I first performed some laboratory tests and notice the decreased number of lymphocytes, which can be indicative of a bone marrow disease or an HIV infection. The HIV test is positive. However, the results from the blood cultures are negative, on causing the sympt

Suggestion:

Hypothesis generation (HG)

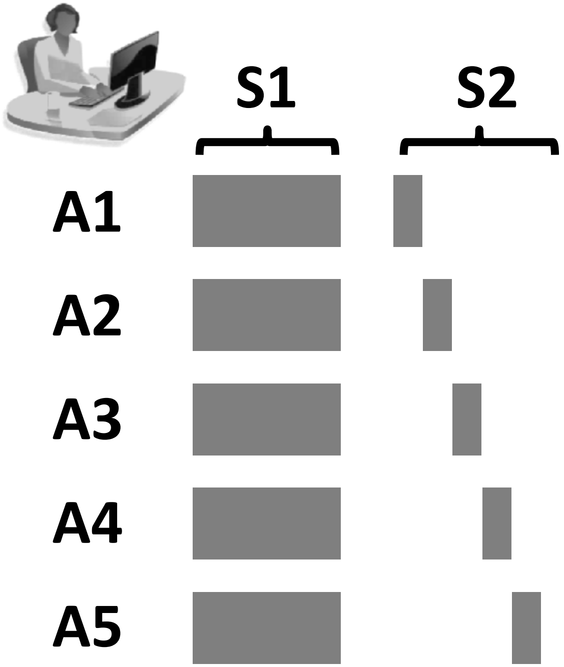
✓ Accept

✗ Reject

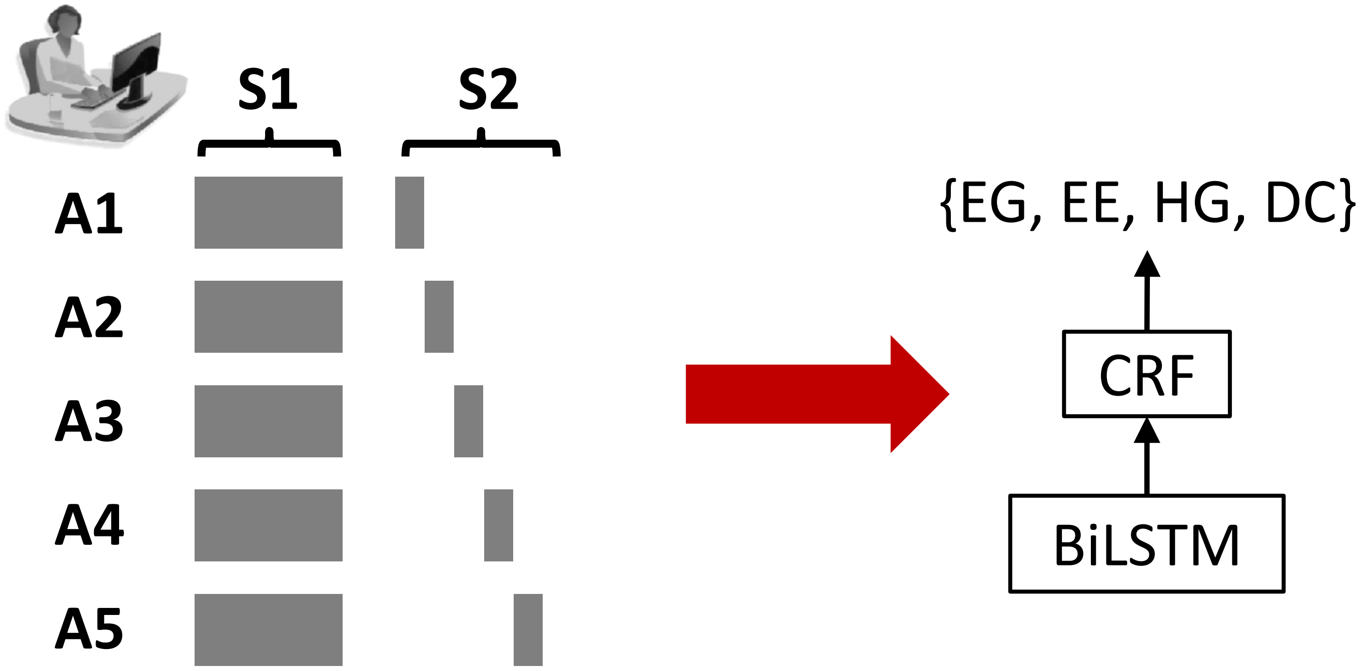
predictions (BiLSTM-CRF)

INCEpTION annotation platform
https://inception-project.github.io

Training Data and Suggestion Quality



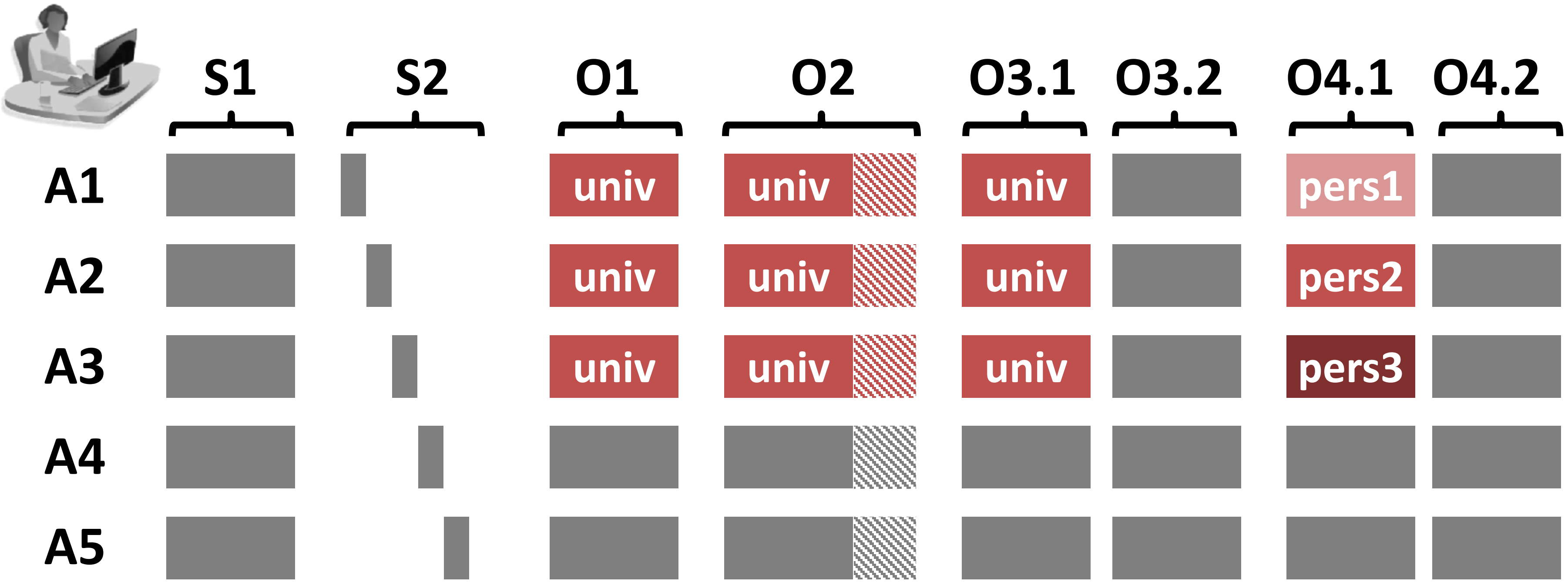
Training Data and Suggestion Quality



univ(ersal) model: $F_1 \approx .63$

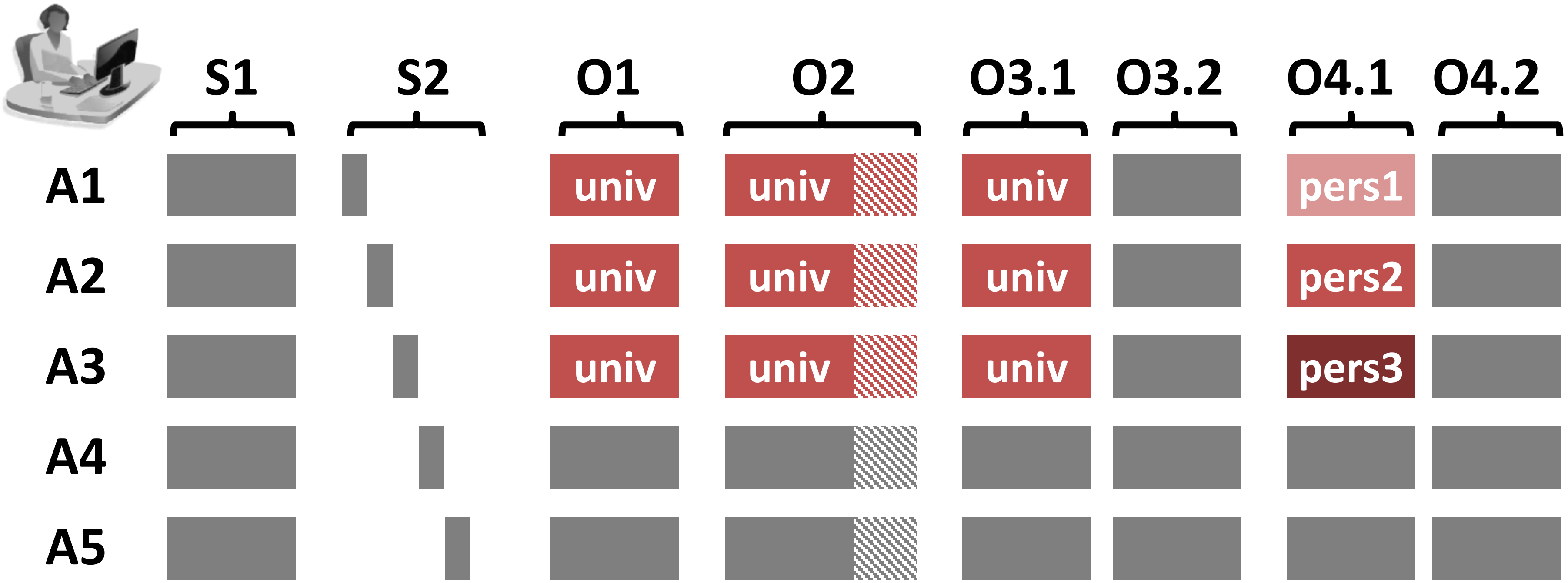
pers(onalized) models: $F_1 \approx .55$

Effectiveness of Annotation Suggestions



	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1			univ	univ	univ		pers1	
A2			univ	univ	univ		pers2	
A3			univ	univ	univ		pers3	
A4								
A5								

Effectiveness of Annotation Suggestions



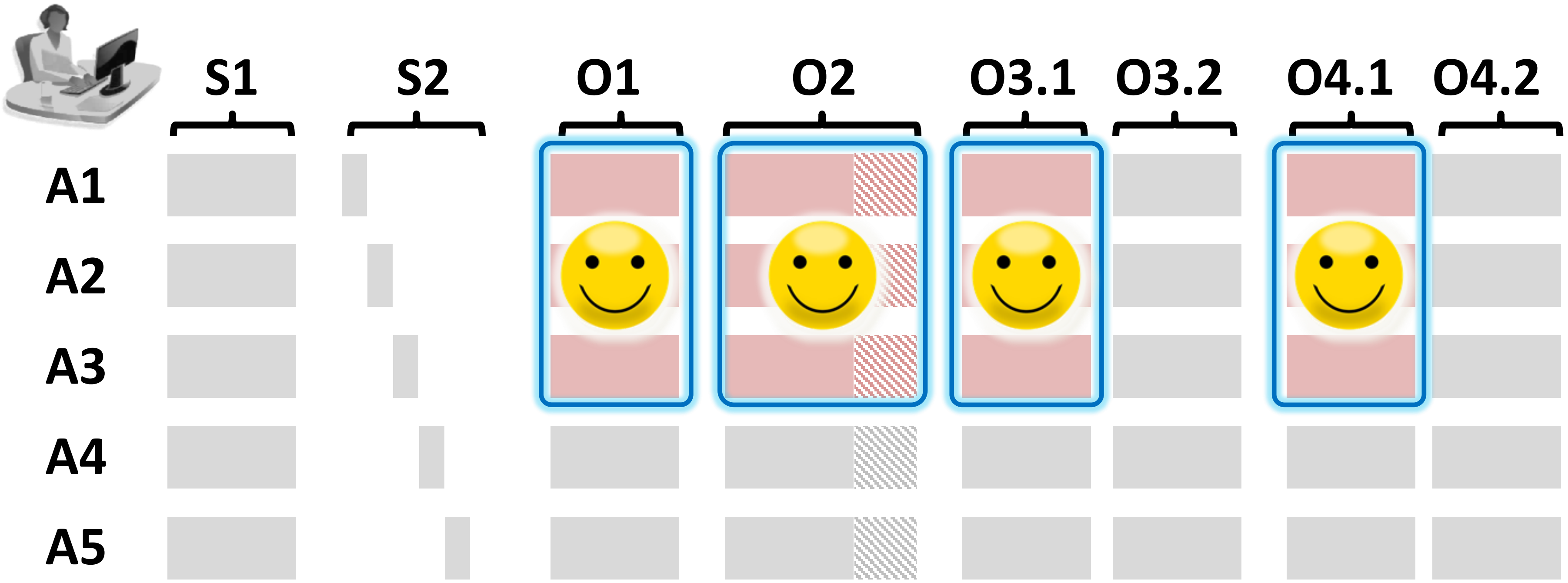
	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1			univ	univ	univ		pers1	
A2			univ	univ	univ		pers2	
A3			univ	univ	univ		pers3	
A4								
A5								

presentation: focus on medical use case

paper: results for same setup in teacher education!

Usefulness of Annotations

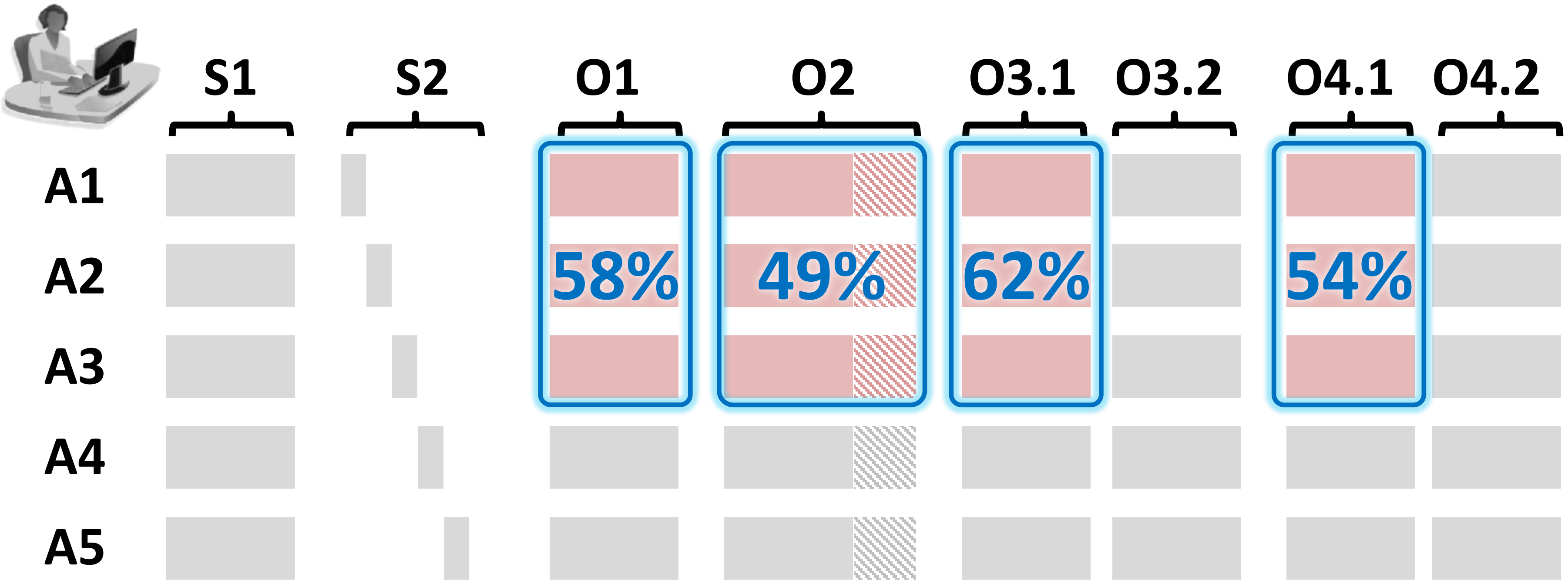
[Annotator happiness]



https://pixabay.com/illustrations/smiley-yellow-happy-smile-emoticon-163510/ (Pixabay license)

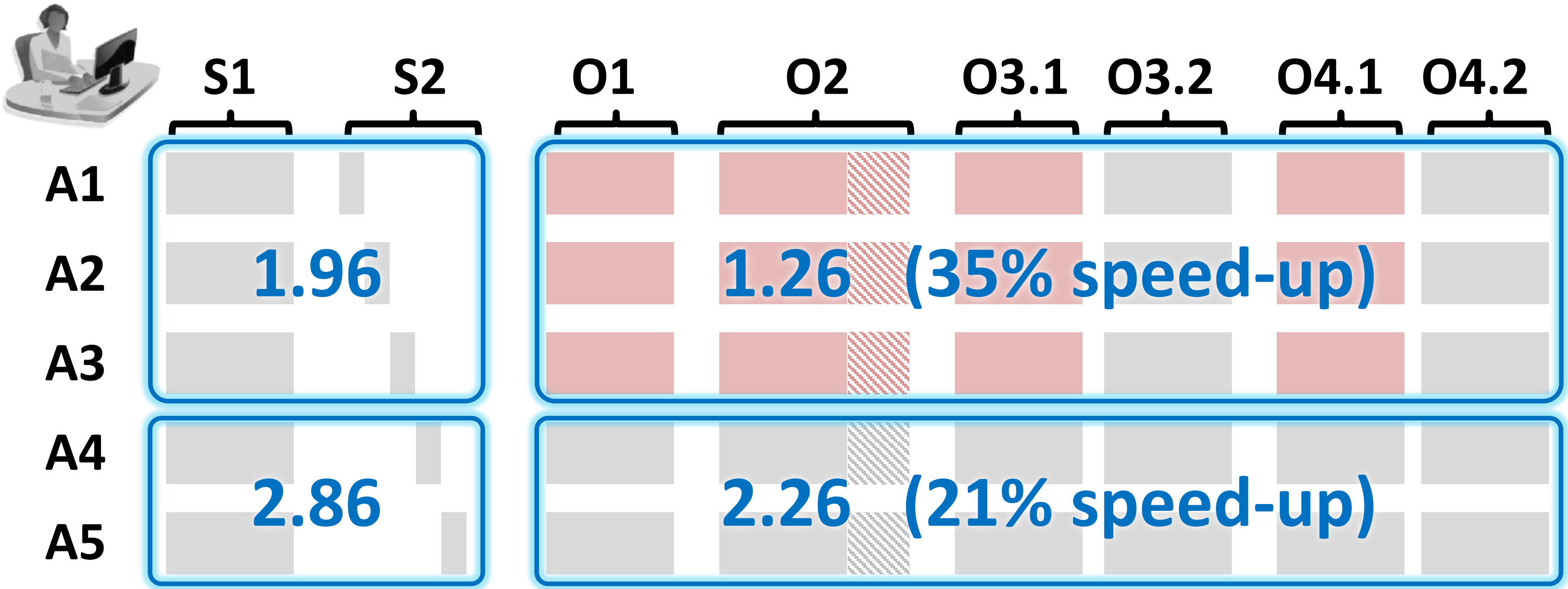
Usefulness of Annotations

[Percentage of accepted suggestions]



Annotation Time

[Minutes per text]



S1 S2 O1 O2 O3.1 O3.2 O4.1 O4.2

A1

A2

A3

1.96

1.26 (35% speed-up)

A4

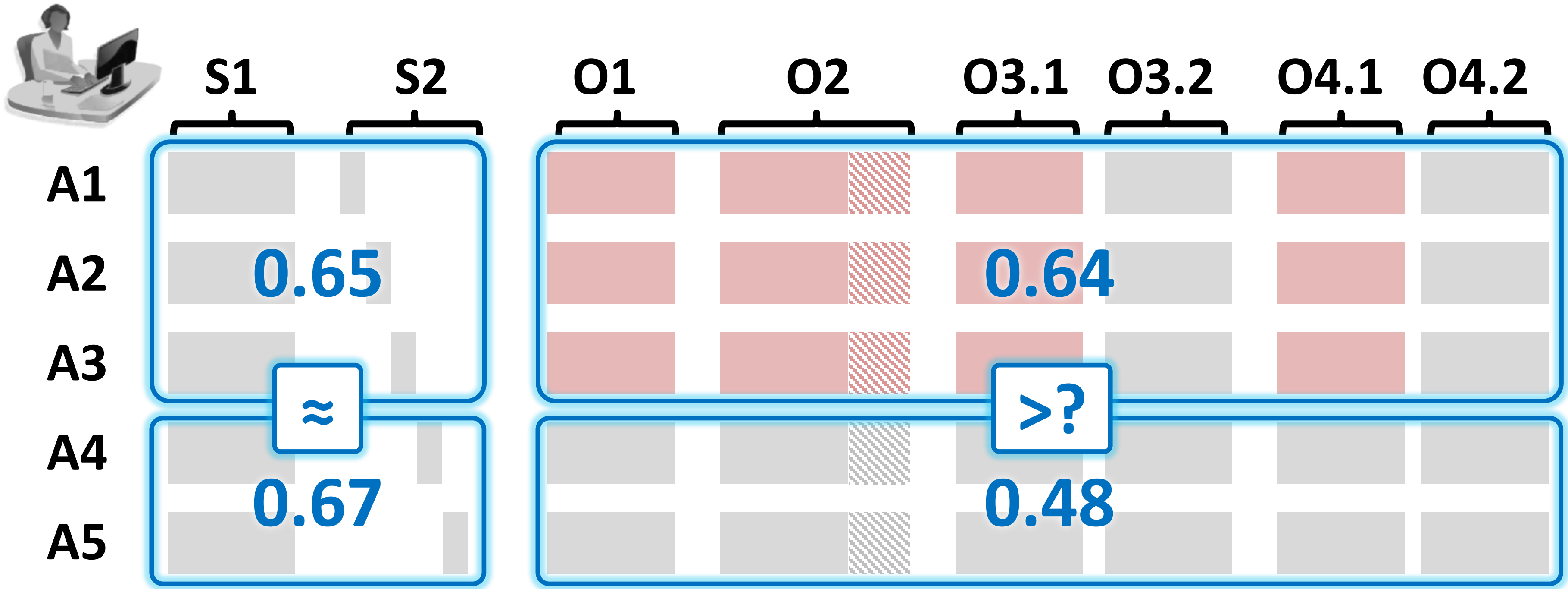
A5

2.86

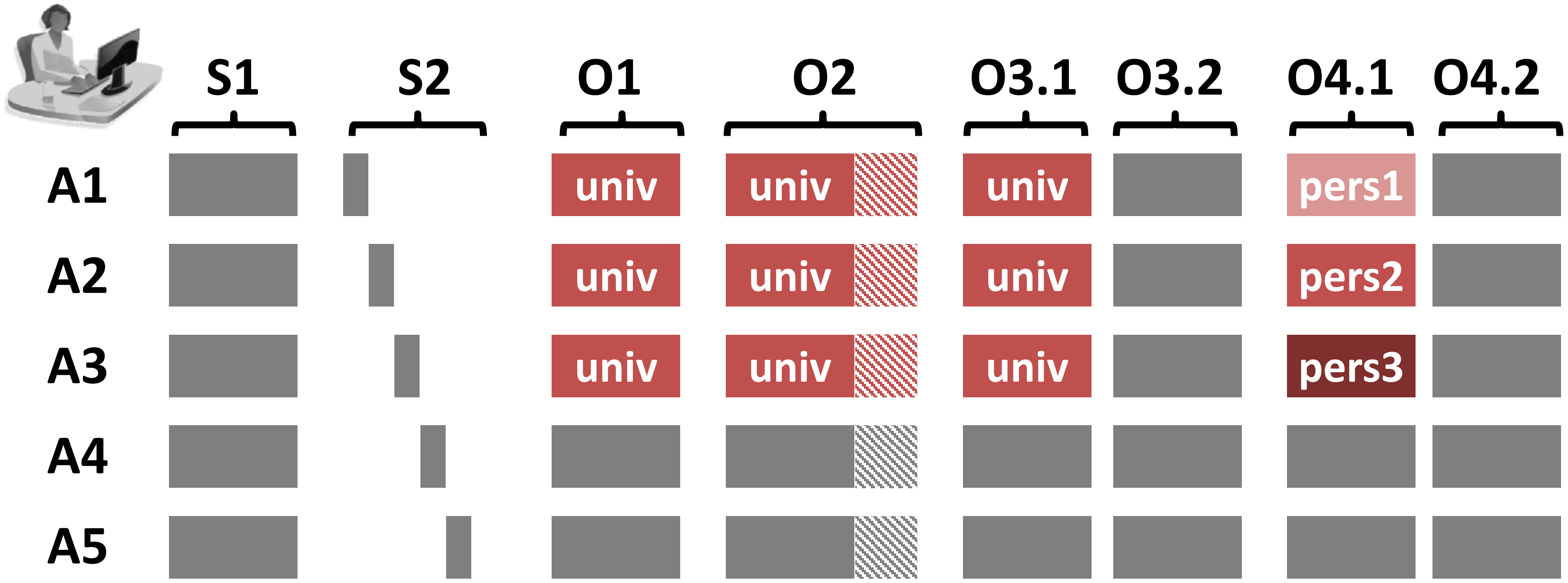
2.26 (21% speed-up)

Reliability of Annotations

[Krippendorff's α]



Effects of Annotation Suggestions



	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1			univ	univ	univ		pers1	
A2			univ	univ	univ		pers2	
A3			univ	univ	univ		pers3	
A4								
A5								

Conclusion 1: Annotation suggestions are helpful for experts and yield faster and (maybe) more reliable annotations!

But: Do predictions bias the decisions?

TECHNISCHE
UNIVERSITÄT
DARMSTADT

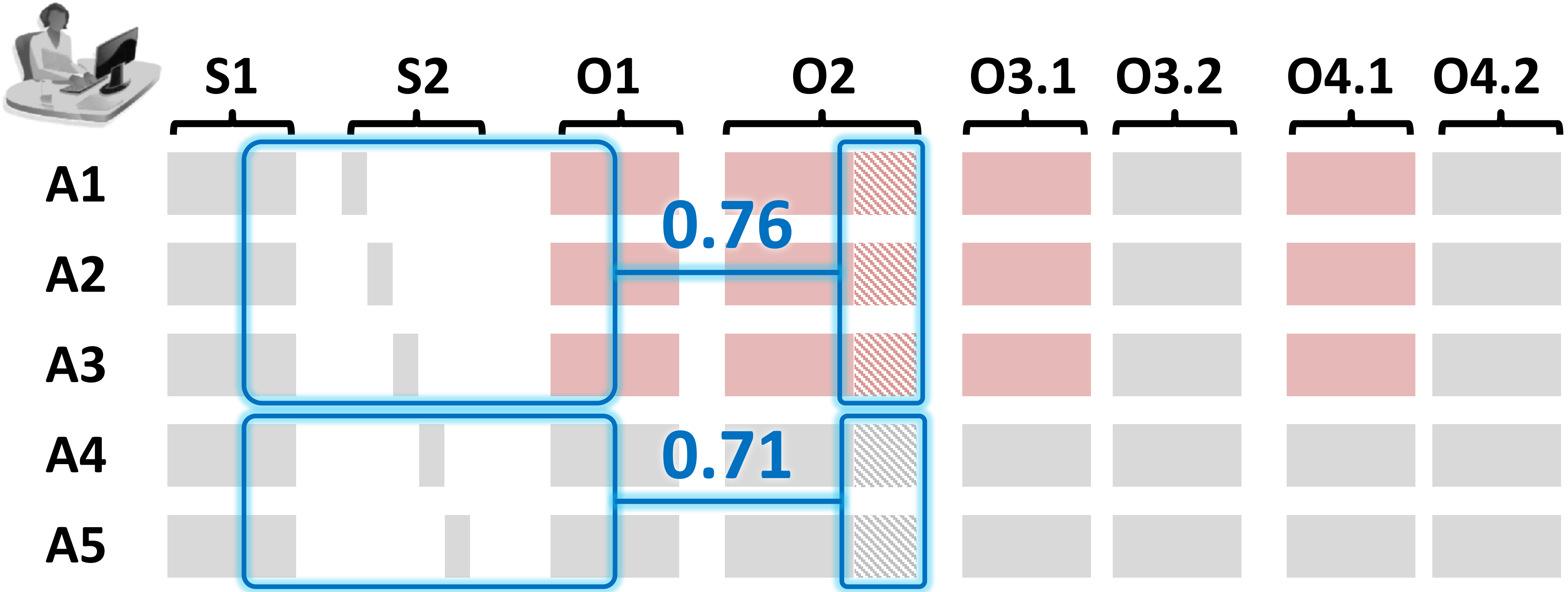
	S1	O4.1	O4.2
A1		pers1	
A2		pers2	
A3		pers3	
A4			
A5			

An perts
and ions!

https://commons.wikimedia.org/wiki/File:FAULES1200.JPG?uselang=de (CC-BY-SA)

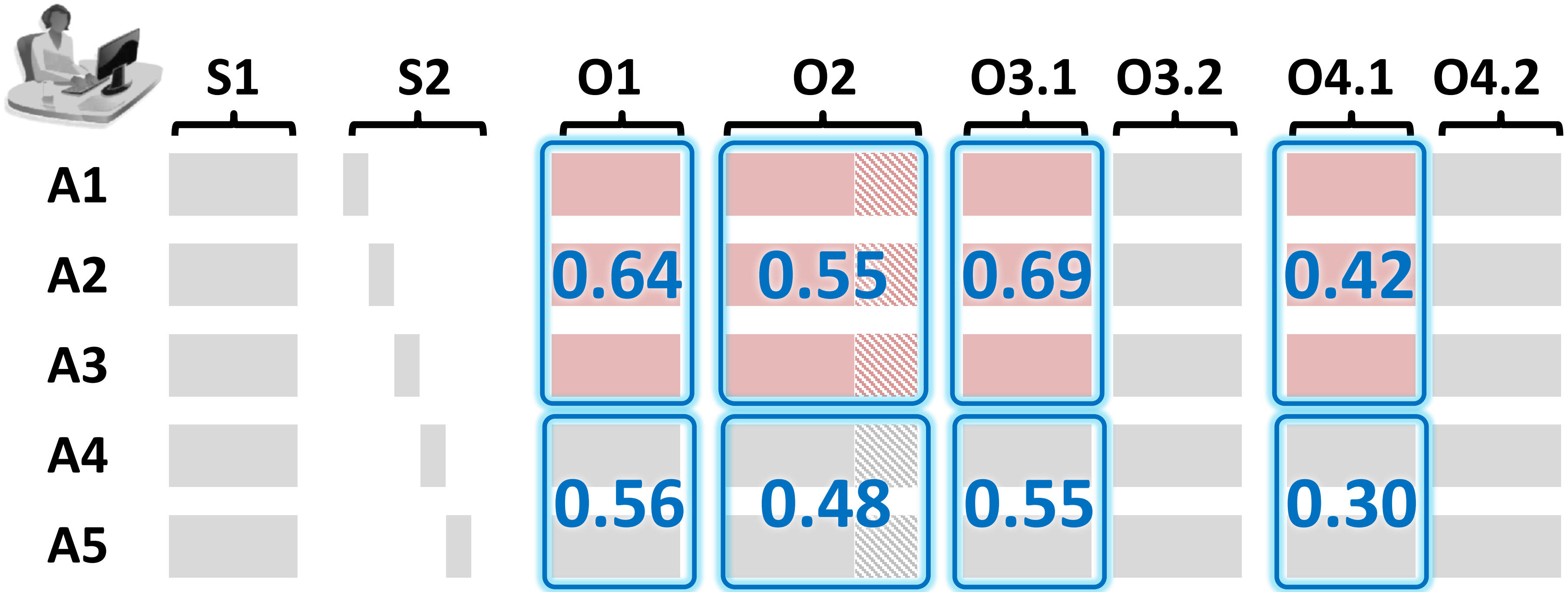
Intra-Annotator Consistency

[Krippendorff's α]



Human–Machine Agreement

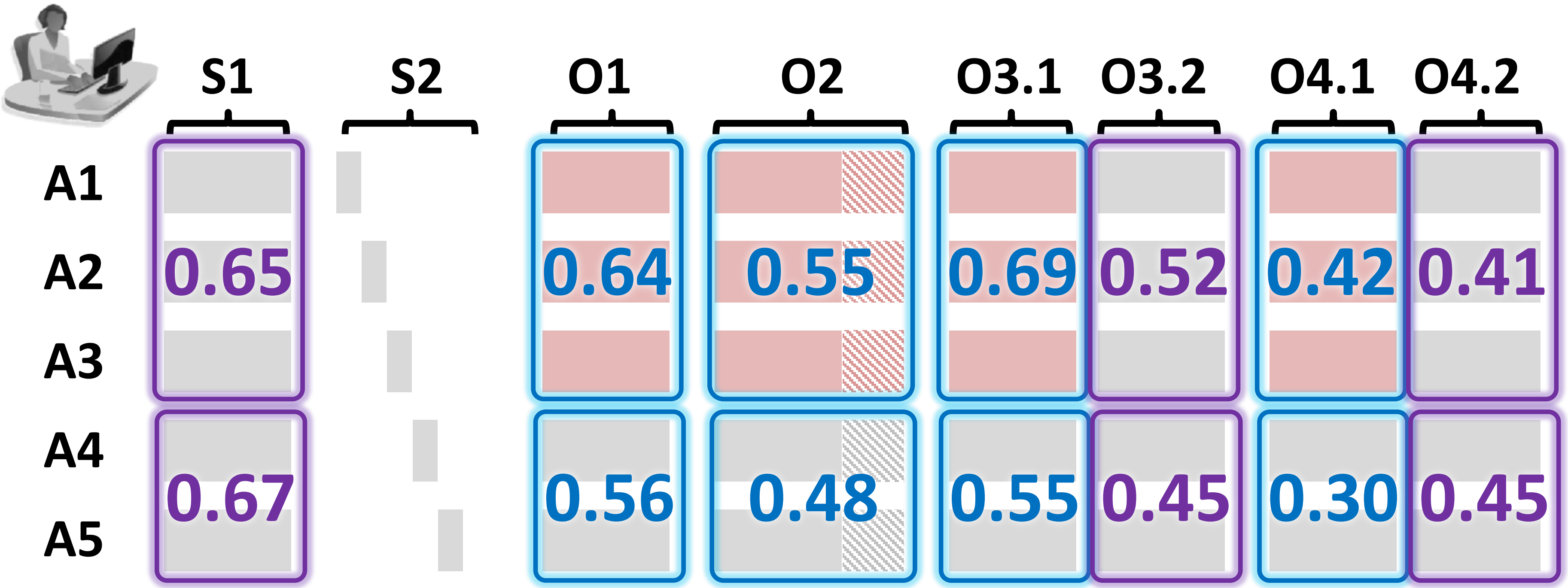
[Krippendorff’s α]



	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1			0.64	0.55	0.69		0.42	
A2								
A3								
A4			0.56	0.48	0.55		0.30	
A5								

Human–Machine Agreement

[Krippendorff’s α]



	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1								
A2	0.65		0.64	0.55	0.69	0.52	0.42	0.41
A3								
A4	0.67		0.56	0.48	0.55	0.45	0.30	0.45
A5								

Further Analysis of Annotation Bias

- **Pairwise agreement between the A1–A3 and the A4–A5 groups**
  → A1–A3 do not behave differently than A4–A5
- **Distribution of labels**
  → no systematic difference

Further Analysis of Annotation Bias

- **Pairwise agreement between the A1–A3 and the A4–A5 groups**
  → A1–A3 do not behave differently than A4–A5
- **Distribution of labels**
  → no systematic difference
- **Distribution of disagreements**
  → only small differences

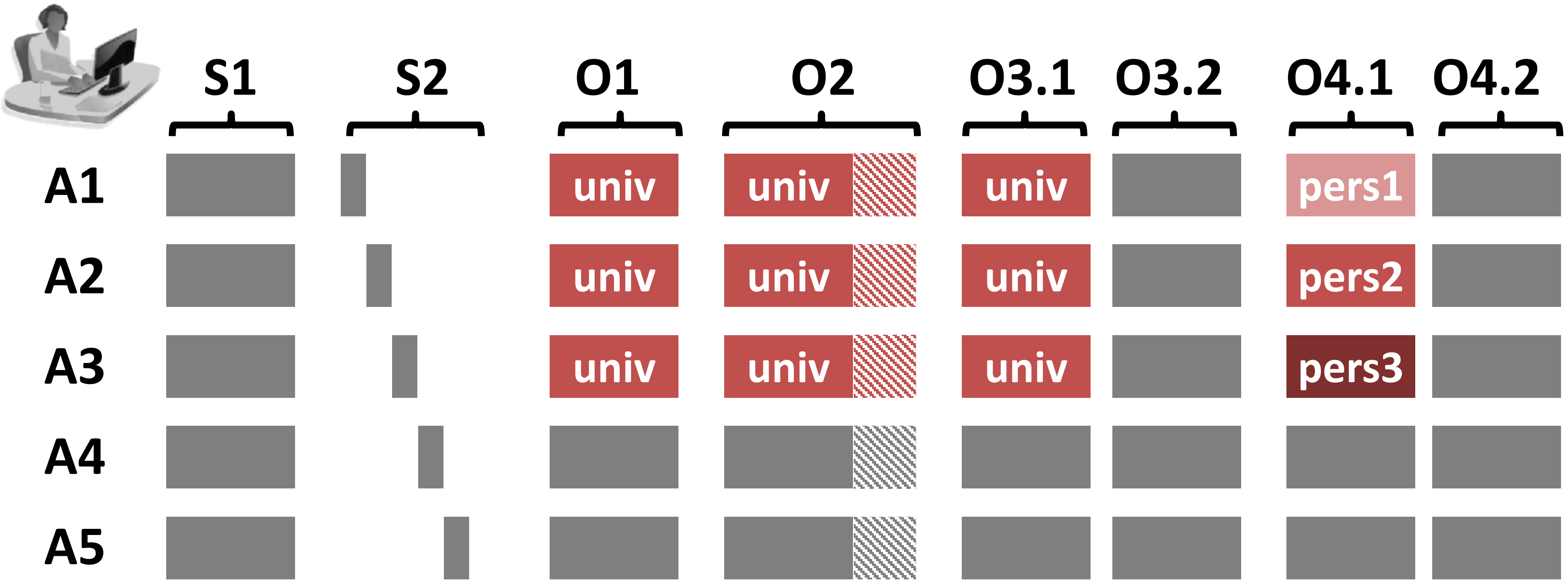
	EG	EE	DC	HG
EG	-	7%	1%	0%
EE	7%	-	22%	13%
DC	1%	22%	-	7%
HG	0%	13%	7%	-

with suggestions A1–A3

	EG	EE	DC	HG
EG	-	5%	1%	2%
EE	5%	-	21%	14%
DC	1%	21%	-	8%
HG	2%	14%	8%	-

without suggestions A4–A5

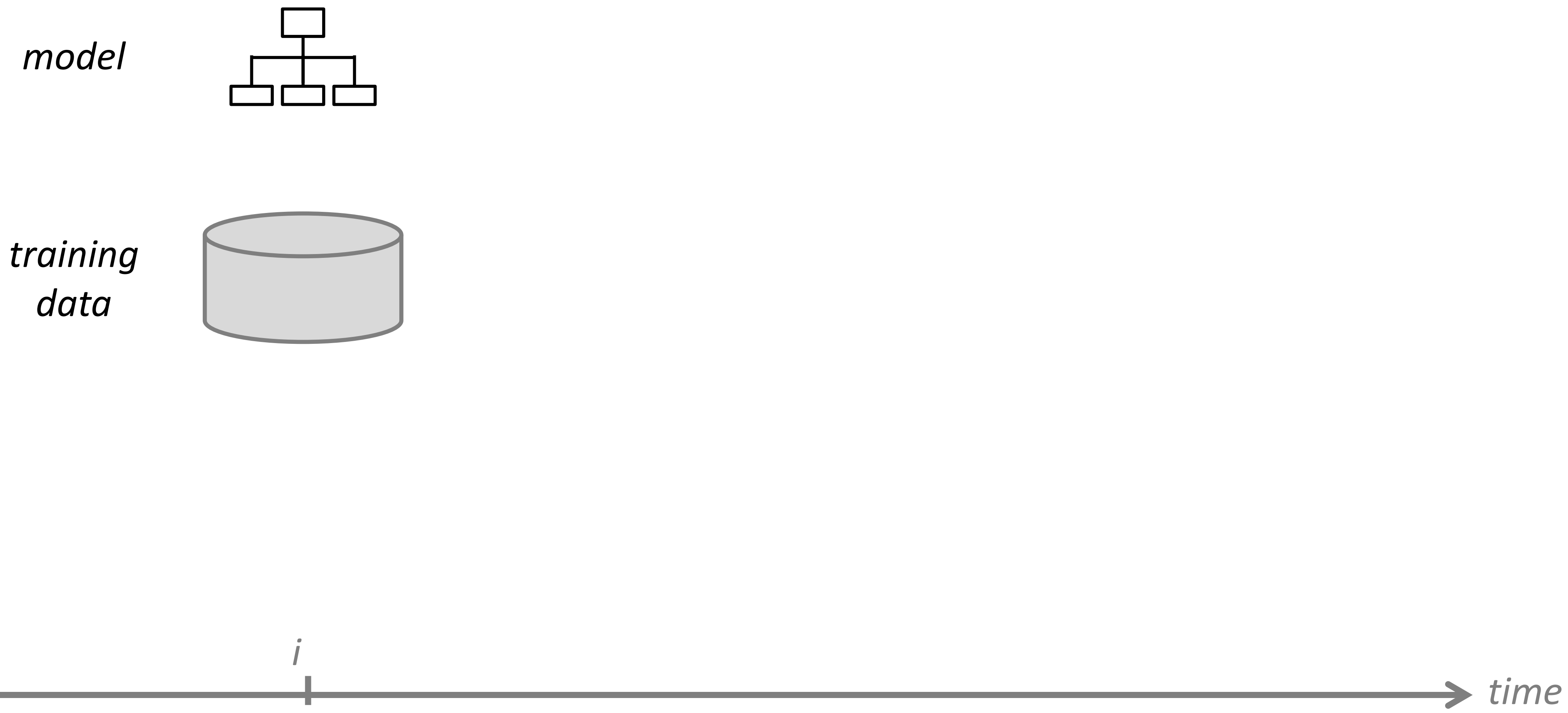
Effects of Annotation Suggestions



	S1	S2	O1	O2	O3.1	O3.2	O4.1	O4.2
A1			univ	univ	univ		pers1	
A2			univ	univ	univ		pers2	
A3			univ	univ	univ		pers3	
A4								
A5								

Conclusion 2: Some evidence for annotation bias, but negligible,
as no systematic discrepancy compared to the control setup!

Interactive Model Training



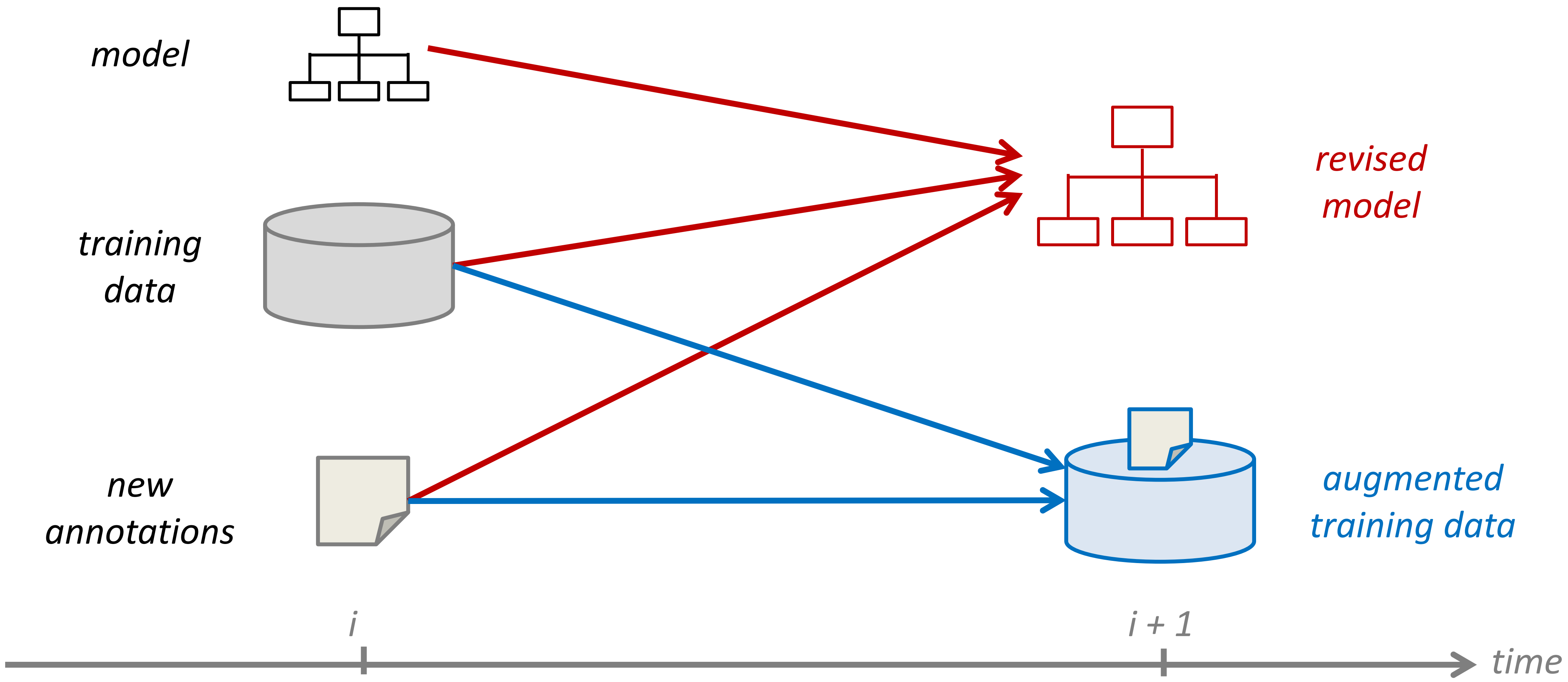
model

training data

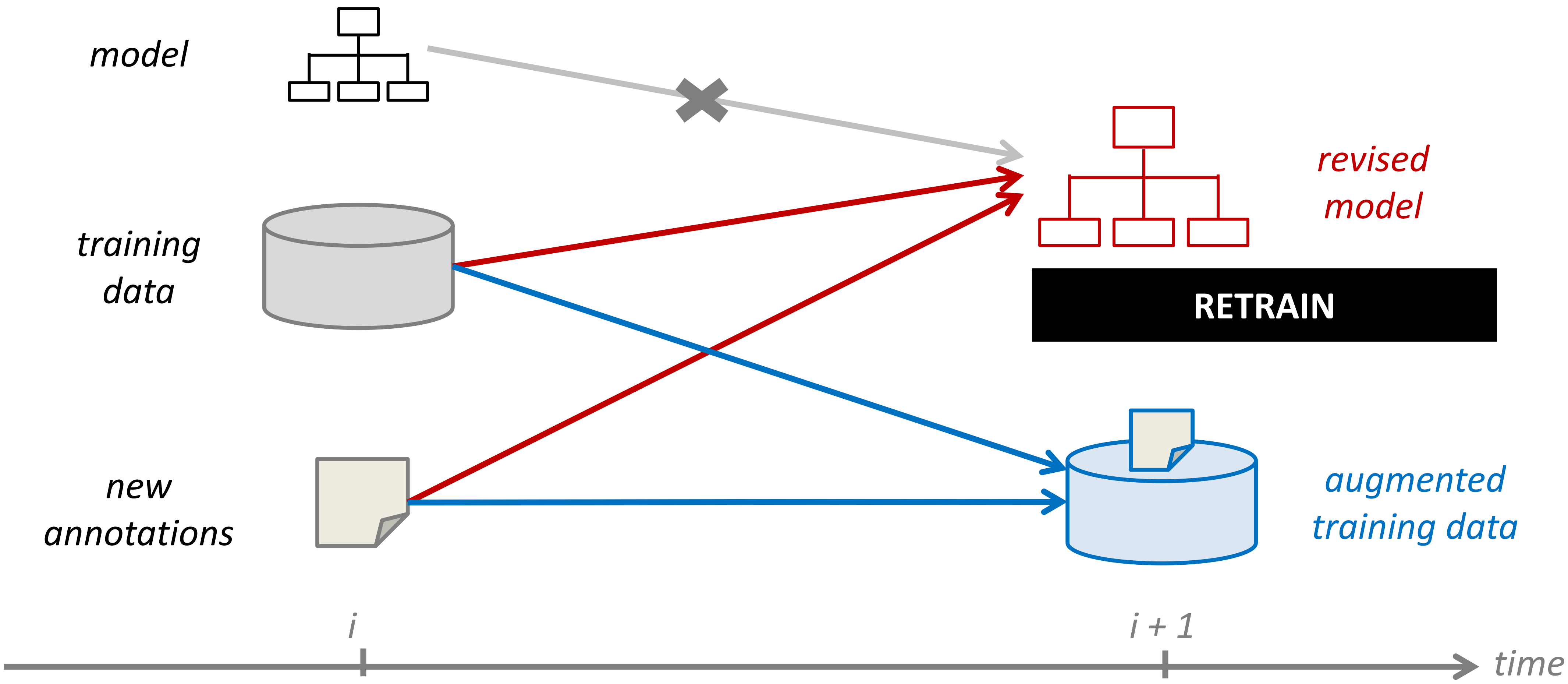
i

time

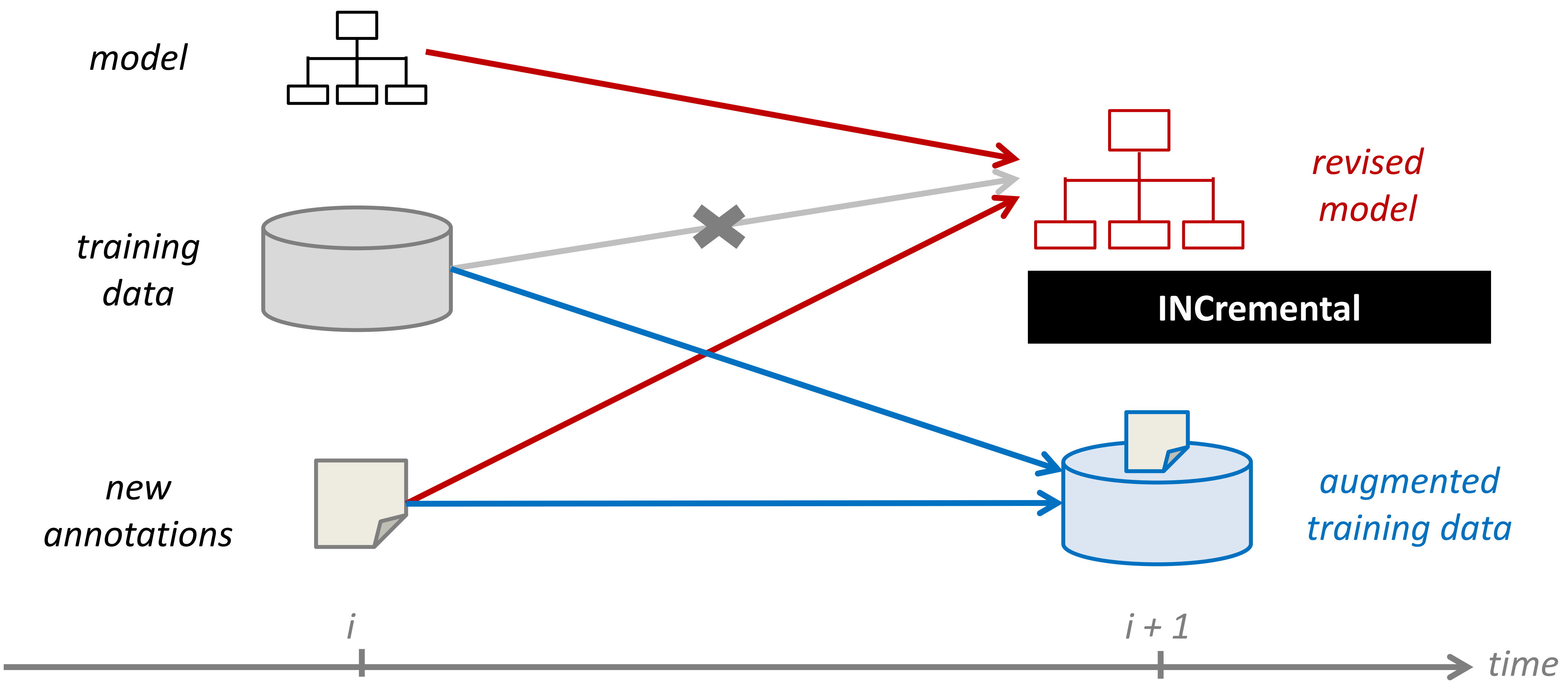
Interactive Model Training



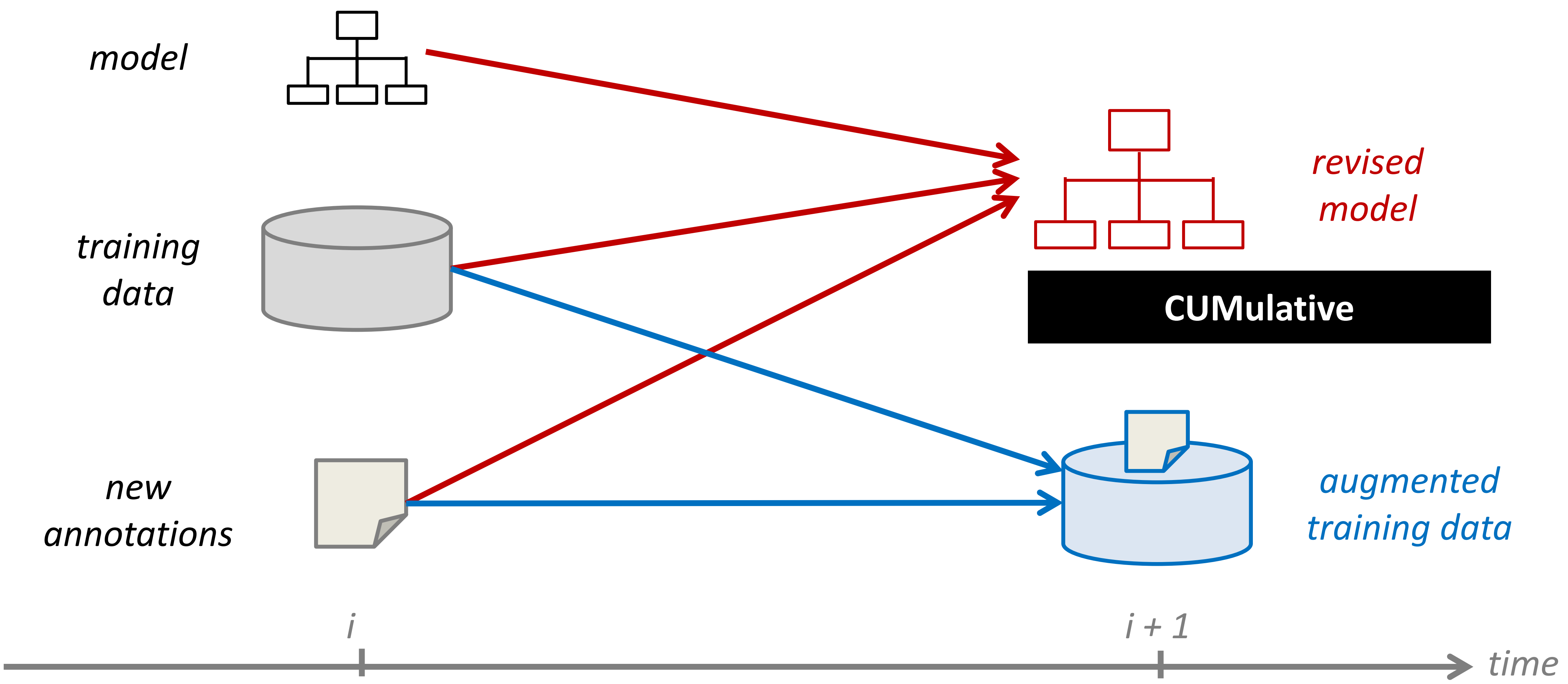
Interactive Model Training



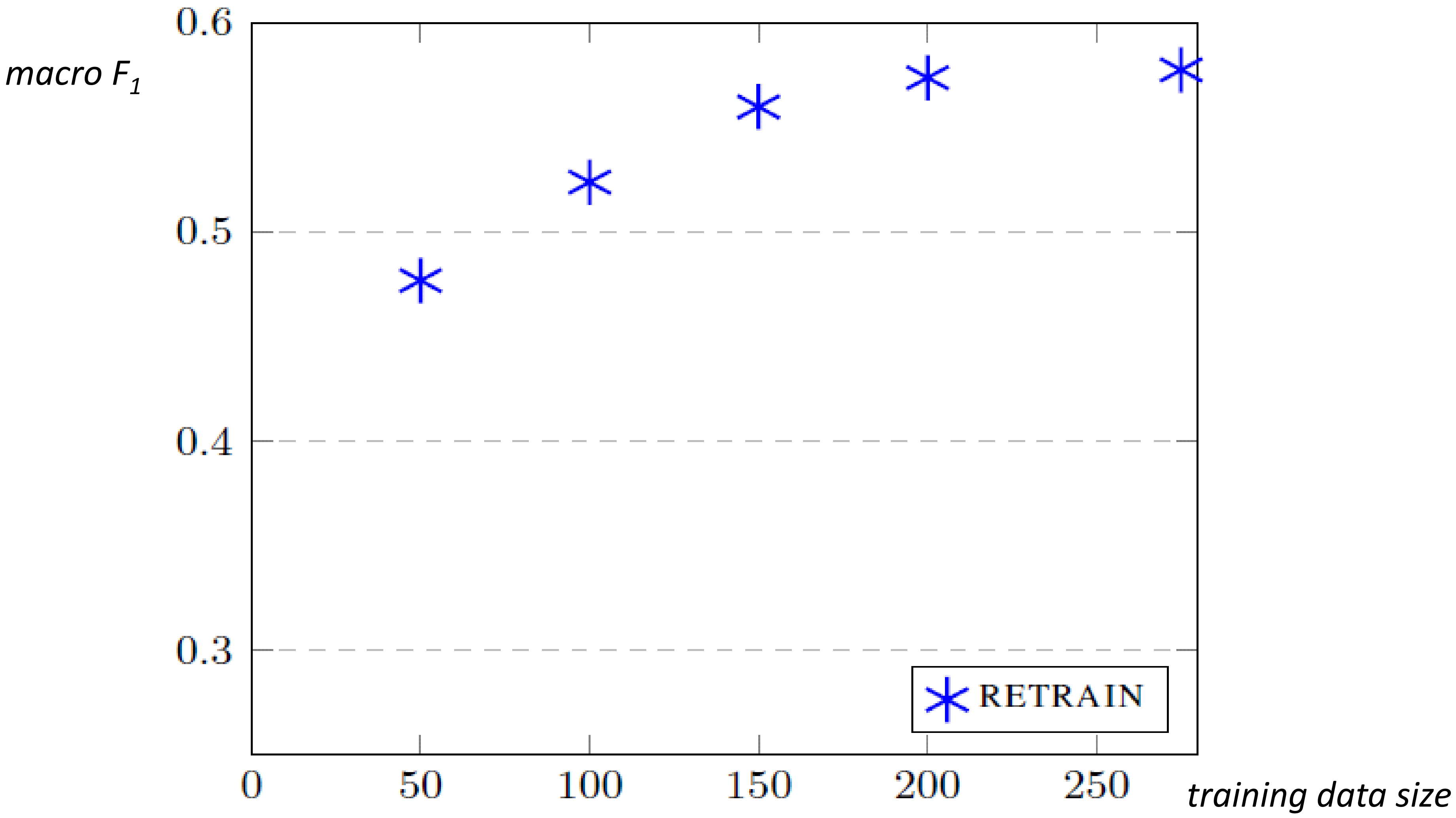
Interactive Model Training



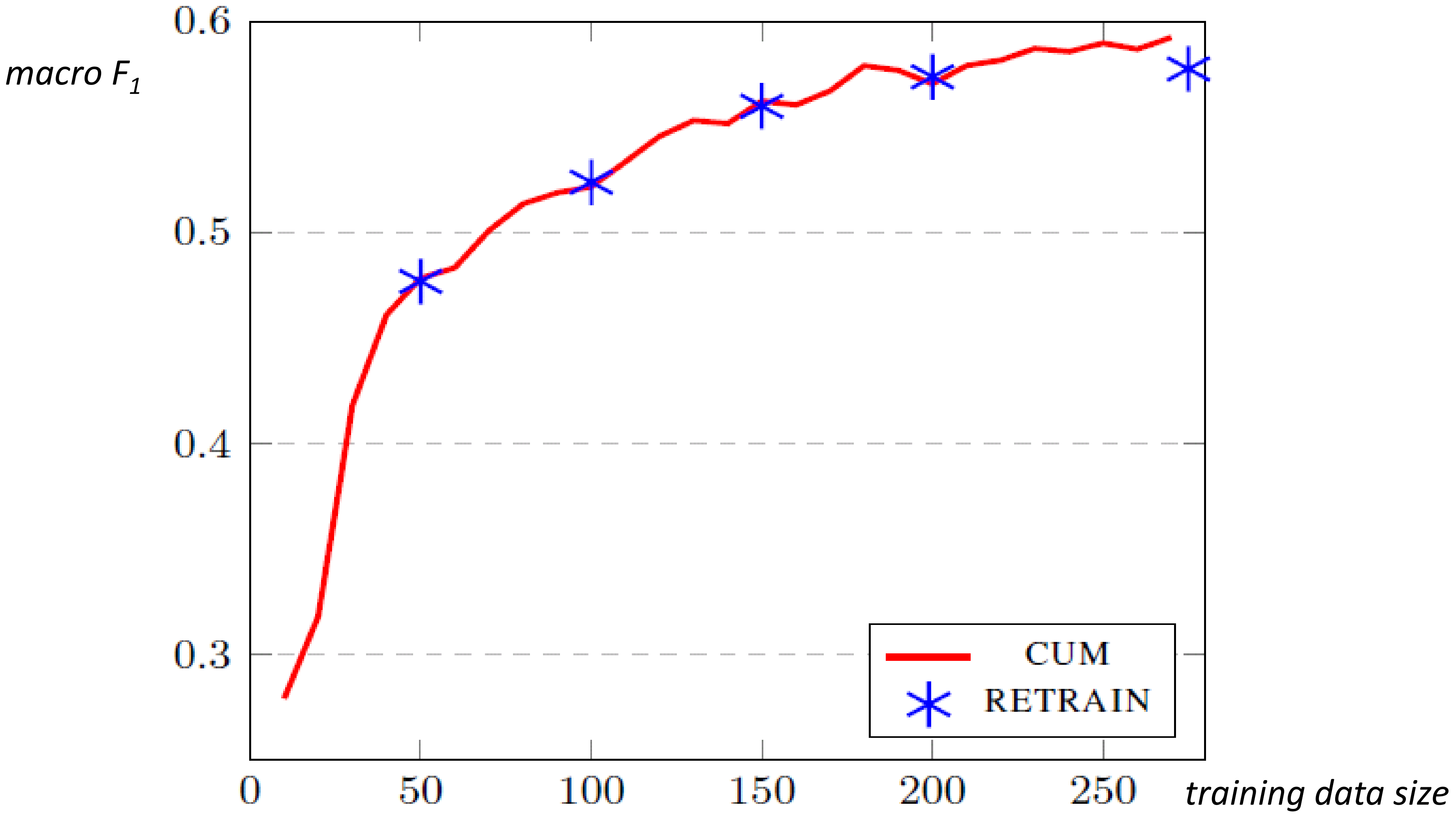
Interactive Model Training



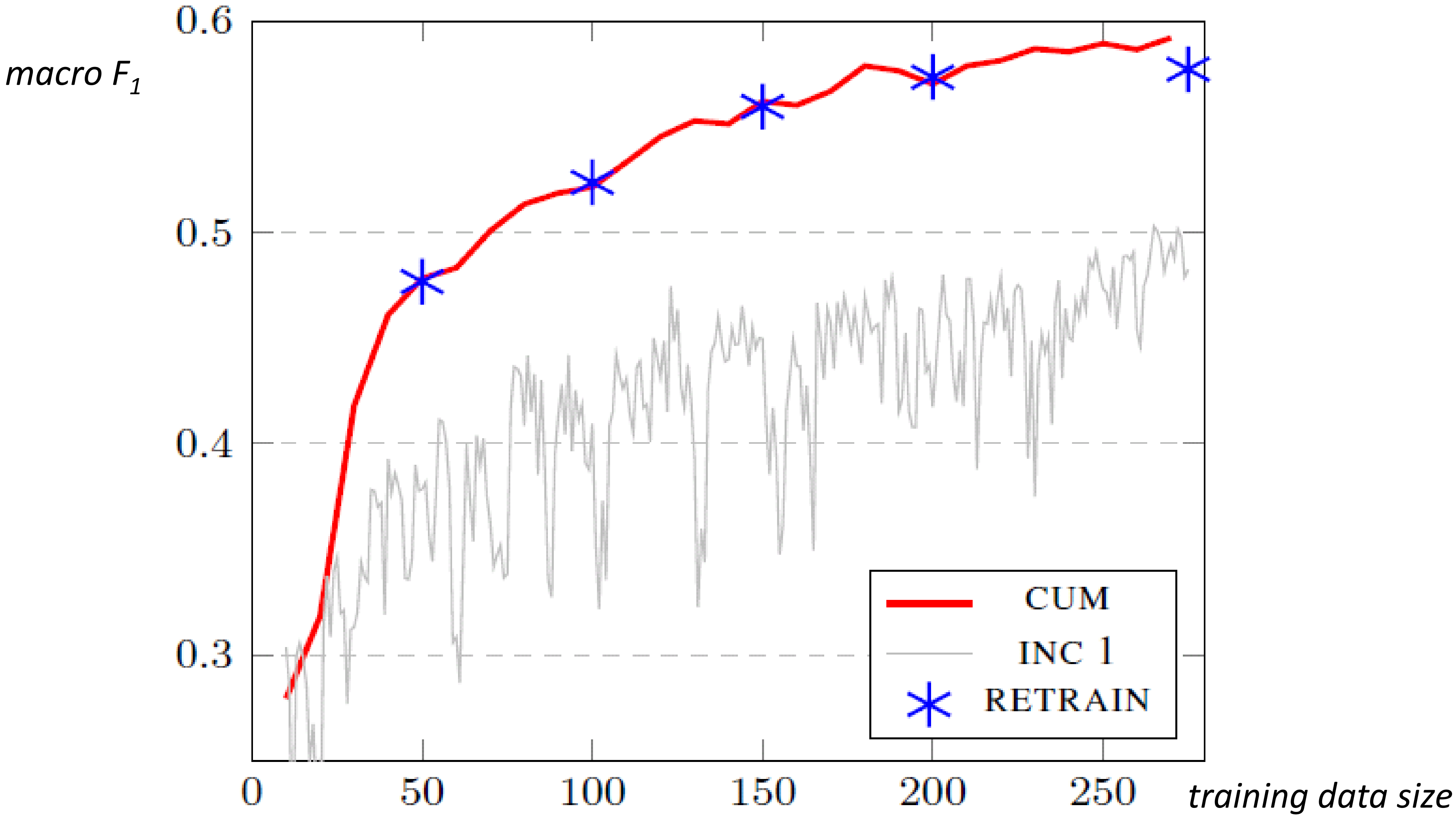
Model Performance



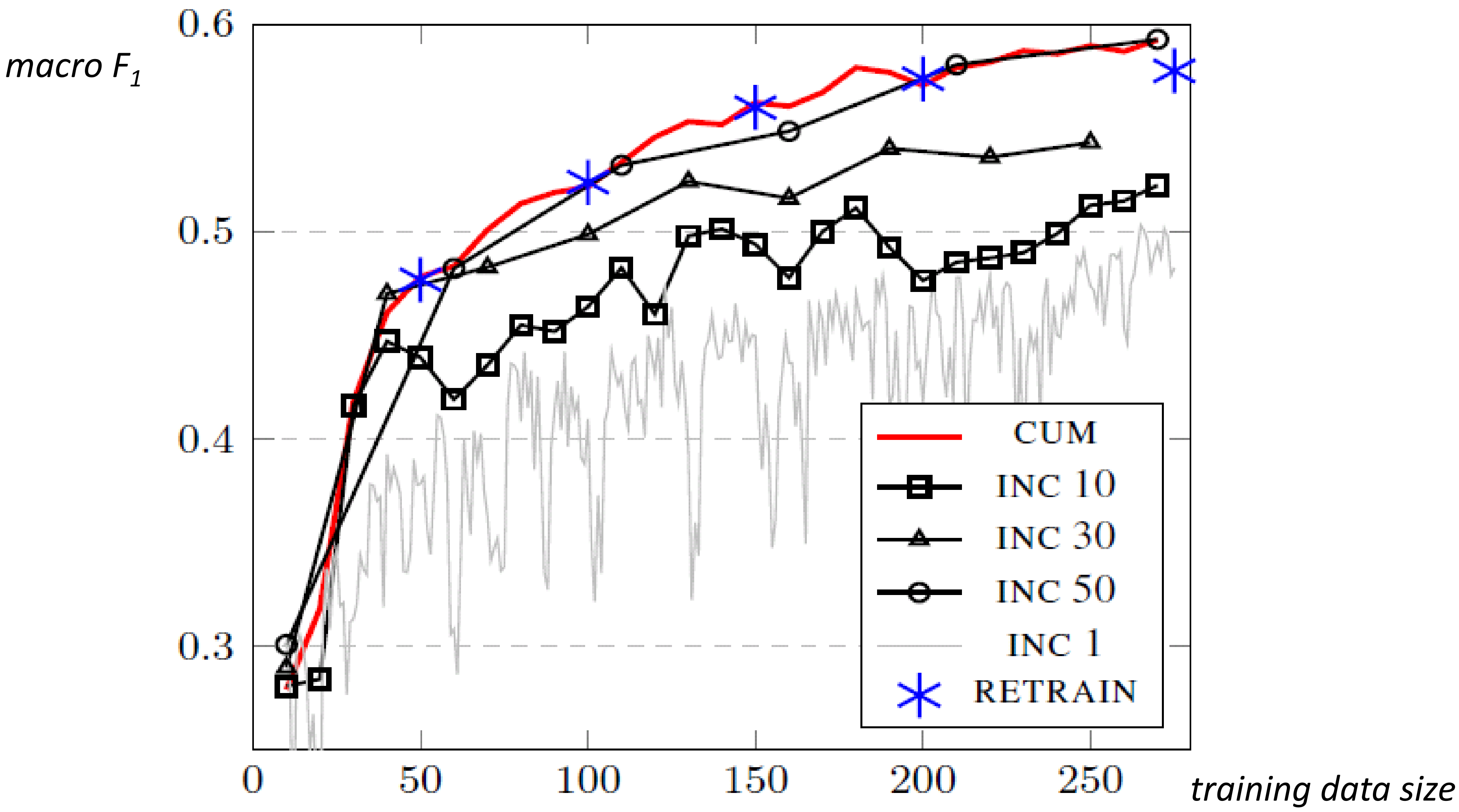
Model Performance



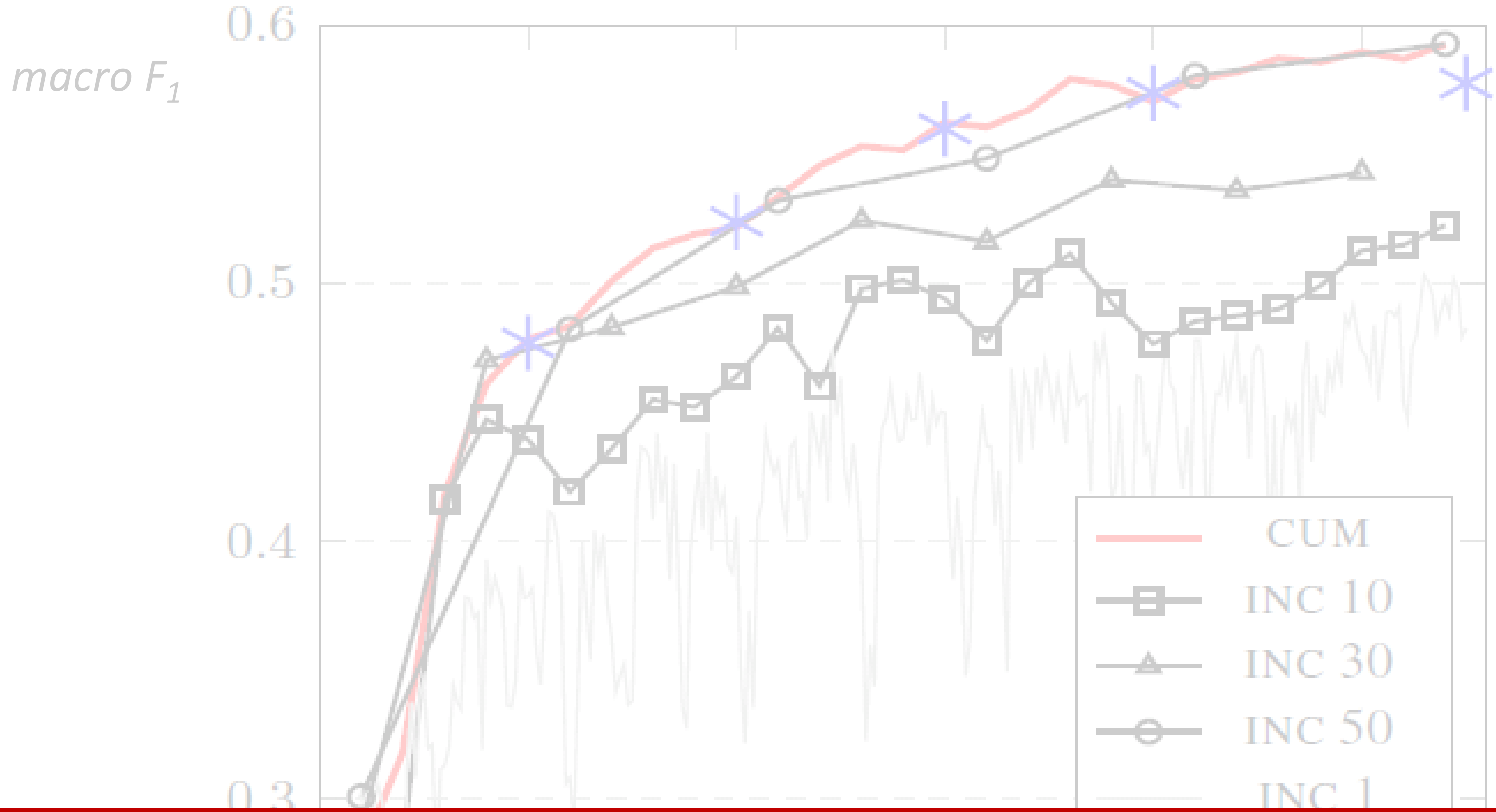
Model Performance



Model Performance



Interactively Trained Suggestions



Conclusion 3: Interactive Model Training yields good performance and allows for time–quality trade-offs!

Summary

Conclusion 1: Annotation suggestions are helpful for experts and yield faster and (maybe) more reliable annotations!

Conclusion 2: Some evidence for annotation bias, but negligible, as no systematic discrepancy compared to the control setup!

Conclusion 3: Interactive Model Training yields good performance and allows for time–quality trade-offs!

Reproducibility

data: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2001
model: https://github.com/UKPLab/aaai19-diagnostic-reasoning

Thank you for your attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

Conclusion 1: Annotation suggestions are helpful for experts and yield faster and (maybe) more reliable annotations!

Conclusion 2: Some evidence for annotation bias, but negligible, as no systematic discrepancy compared to the control setup!

Conclusion 3: Interactive Model Training yields good performance and allows for time–quality trade-offs!

Reproducibility

data: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2001
model: https://github.com/UKPLab/aaai19-diagnostic-reasoning

Kontakt / Contact

Dr. Christian M. Meyer
Technische Universität Darmstadt
Ubiquitous Knowledge Processing Lab

📪 Hochschulstr. 10, 64289 Darmstadt, Germany
☎ +49 (0)6151 16–25293
📠 +49 (0)6151 16–25295
✉ meyer (at) ukp.informatik.tu-darmstadt.de

Rechtliche Hinweise

Die Folien sind für den persönlichen Gebrauch der Vortragsteilnehmer gedacht. Im Vortrag verwendete Photographien, Illustrationen, Wort- und Bildmarken sind Eigentum der jeweiligen Rechteinhaber oder Lizenzgeber. Um Missverständnisse zu vermeiden, wäre eine kurze Kontaktaufnahme vor Weitergabe oder -nutzung der Vortragsmaterialien empfehlenswert. Sofern Sie Ihre Rechte verletzt sehen, bitte ich ebenfalls um Kontaktaufnahme zur Klärung der Sachlage.

Legal Issues

The slides are intended for personal use by the audience of the talk. Photographies, illustrations, tradedmarks, or logos are property of the holder of rights. To avoid any misconceptions, I would strongly recommend to get in touch before reusing or redistributing the slides or any additional material of the talk. The same applies if you consider your rights infringed – please let me know to initiate further clarification.