Explaining (how to improve) Diagnostic Reasoning

— *The FAMULUS Project* —

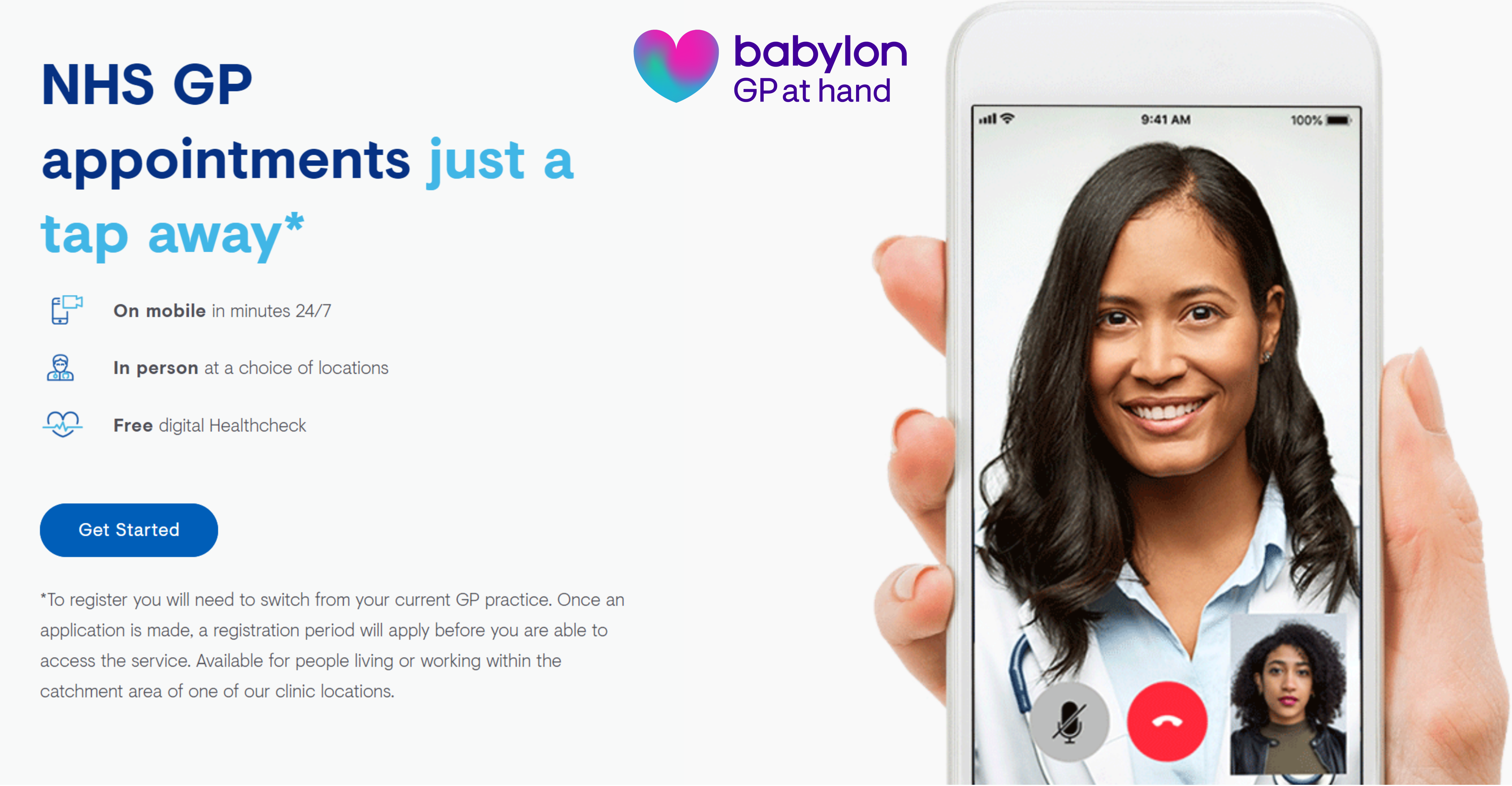
Please don't hesitate to ask questions during the talk!

Claudia Schulz

15/01/2020 XAI Seminar

FAMULUS

http://famulus-project.de



babylon
GP at hand

NHS GP appointments just a tap away*

On mobile in minutes 24/7

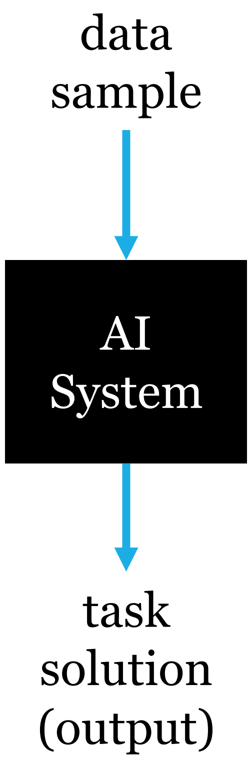
In person at a choice of locations

Free digital Healthcheck

Get Started

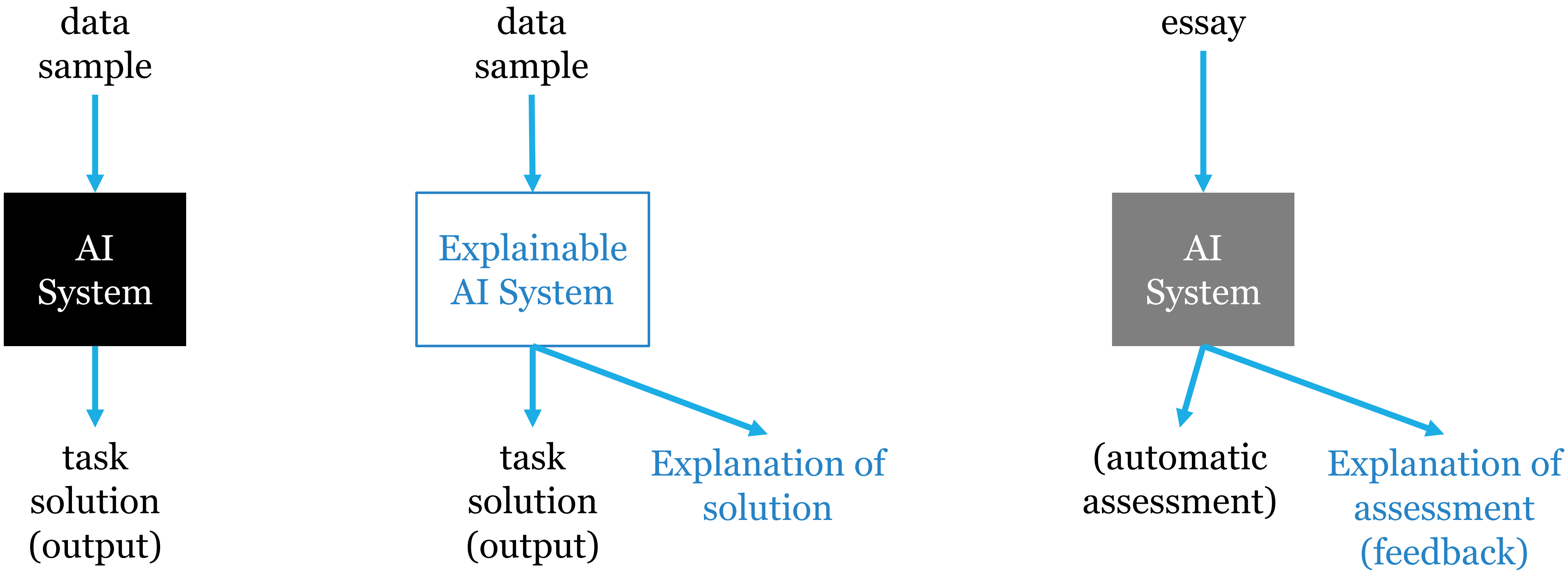
*To register you will need to switch from your current GP practice. Once an application is made, a registration period will apply before you are able to access the service. Available for people living or working within the catchment area of one of our clinic locations.

Explainable AI


data sample
AI System
task solution (output)

People with no idea about AI, telling me my AI will destroy the world
Me wondering why my neural network is classifying a cat as a dog..
Dog

Explainable AI



data sample

AI System

task solution (output)

data sample

Explainable AI System

task solution (output)

Explanation of solution

essay

AI System

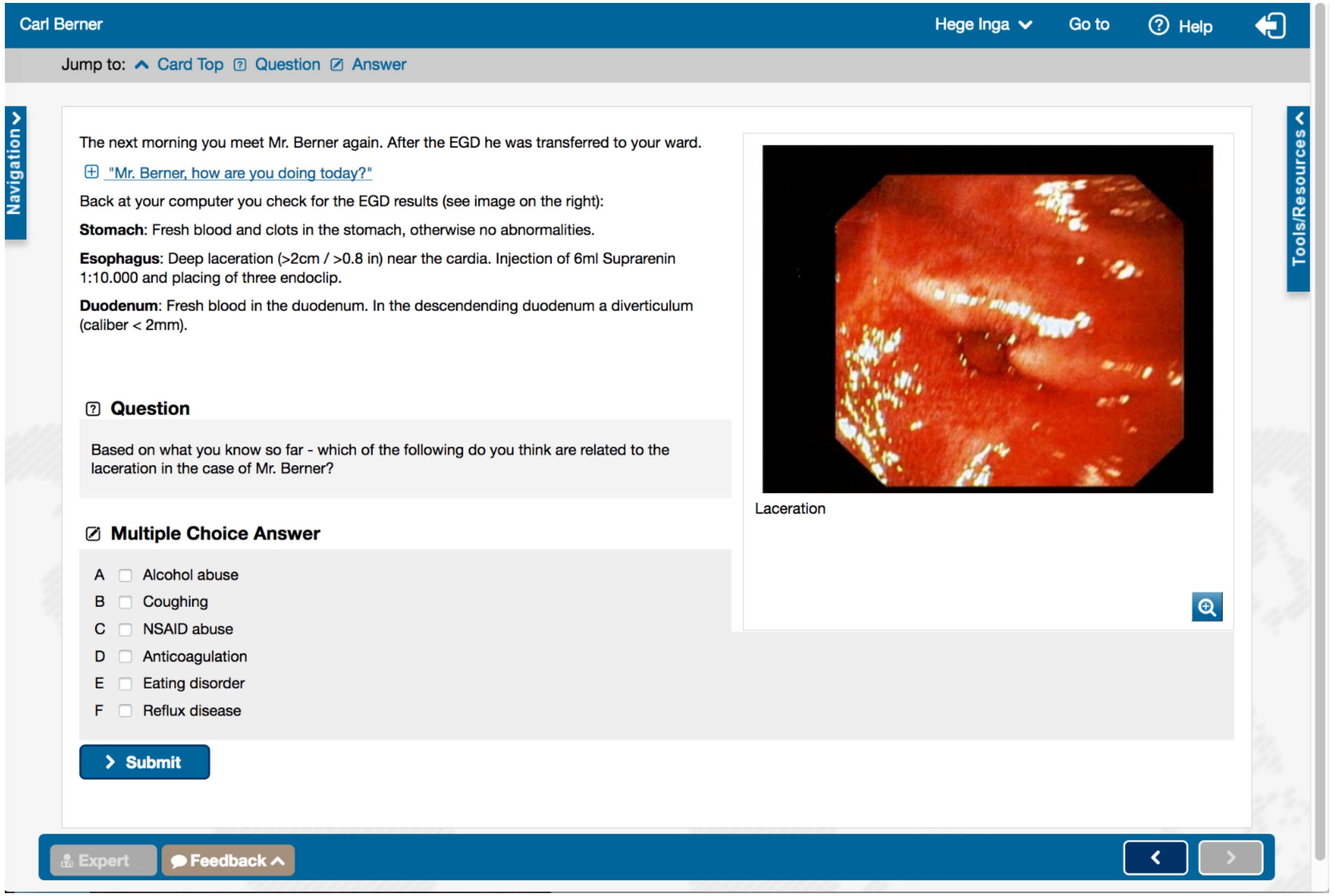
(automatic assessment)

Explanation of assessment (feedback)

Learning Diagnostic Reasoning

Diagnostic Reasoning

Case Simulation



The next morning you meet Mr. Berner again. After the EGD he was transferred to your ward.

"Mr. Berner, how are you doing today?"

Back at your computer you check for the EGD results (see image on the right):

Stomach: Fresh blood and clots in the stomach, otherwise no abnormalities.

Esophagus: Deep laceration (>2cm / >0.8 in) near the cardia. Injection of 6ml Suprarenin 1:10.000 and placing of three endoclip.

Duodenum: Fresh blood in the duodenum. In the descendending duodenum a diverticulum (caliber < 2mm).

Laceration

Question

Based on what you know so far - which of the following do you think are related to the laceration in the case of Mr. Berner?

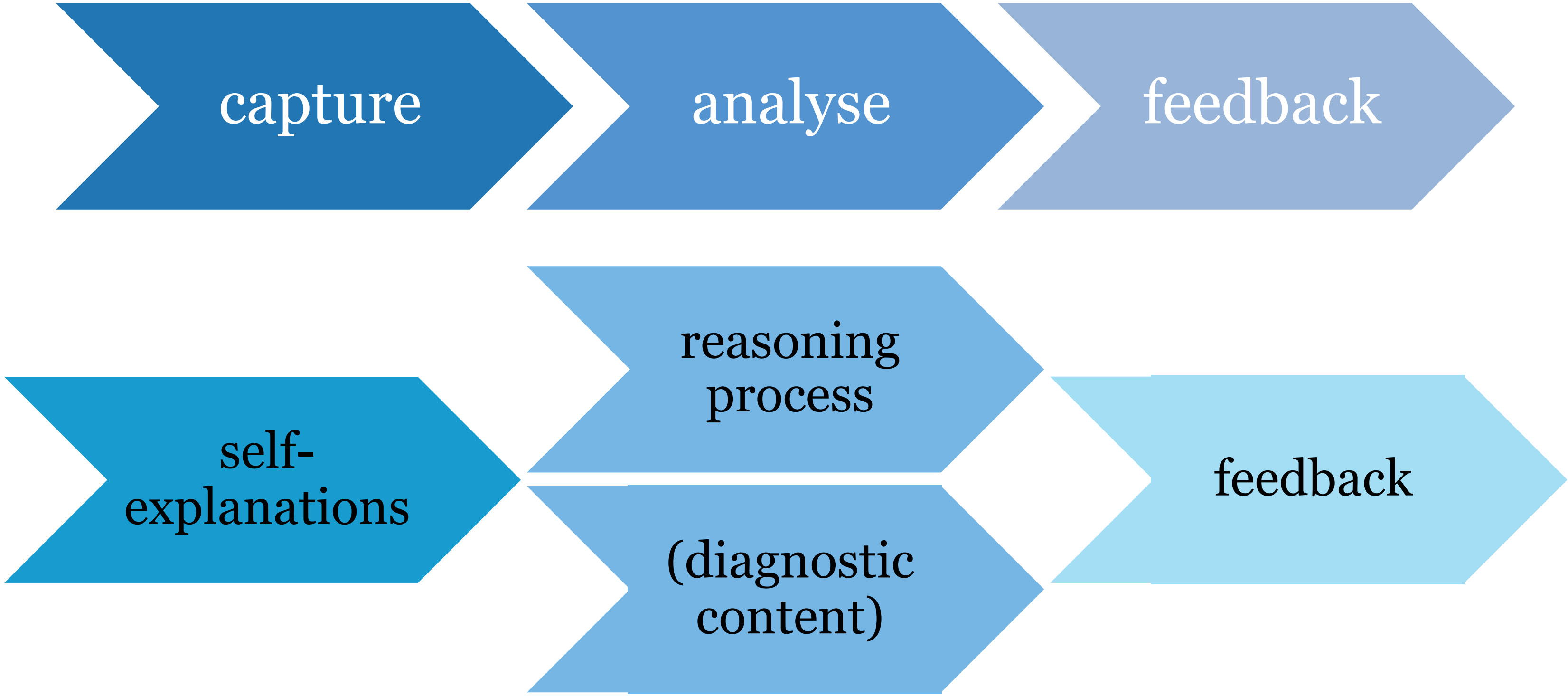
Multiple Choice Answer

A ☐ Alcohol abuse
B ☐ Coughing
C ☐ NSAID abuse
D ☐ Anticoagulation
E ☐ Eating disorder
F ☐ Reflux disease

Submit

Individual Feedback

Diagnostic Reasoning



capture

analyse

feedback

self-explanations

reasoning process

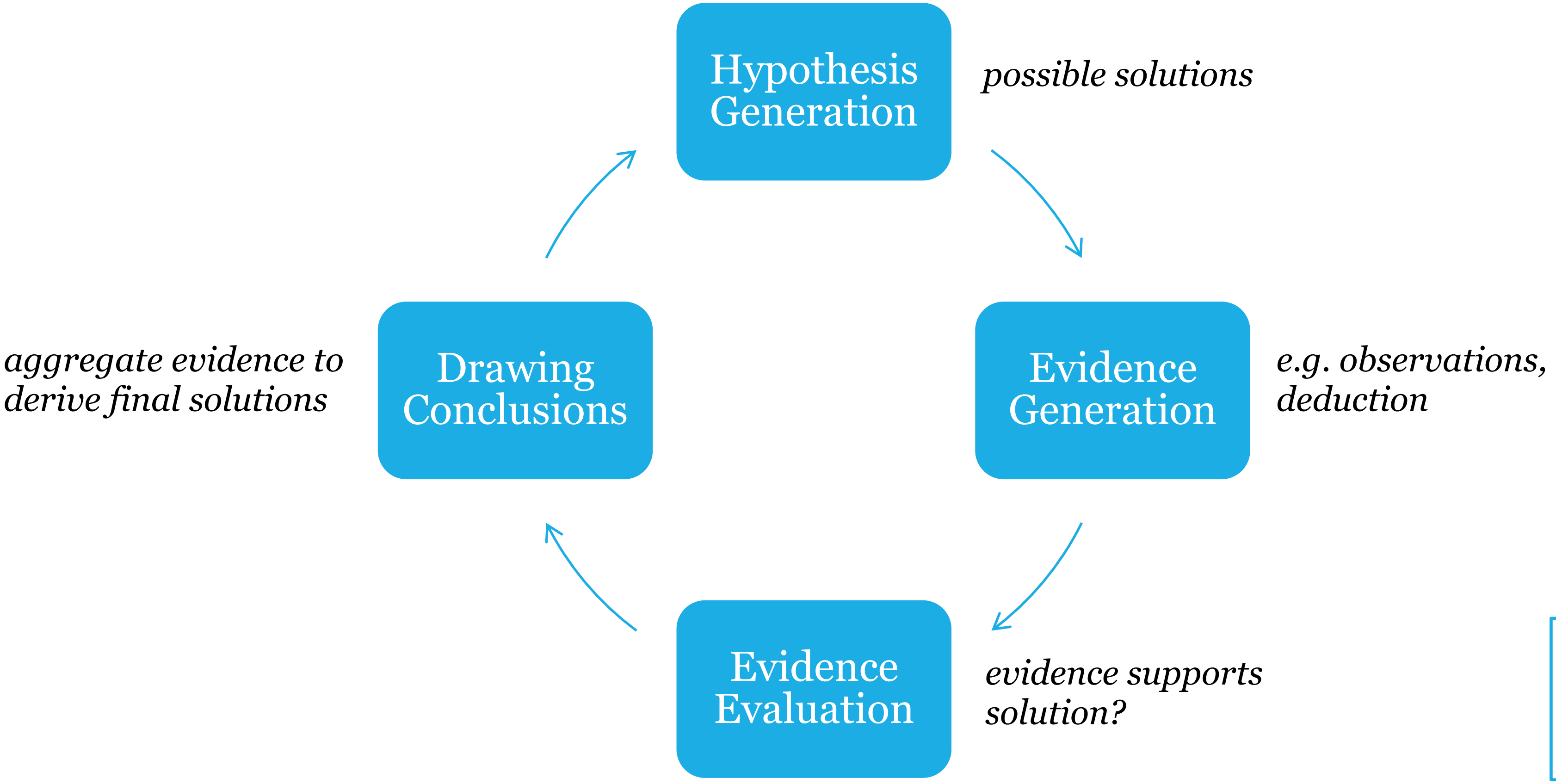
(diagnostic content)

feedback

Capture: Self-Explanation

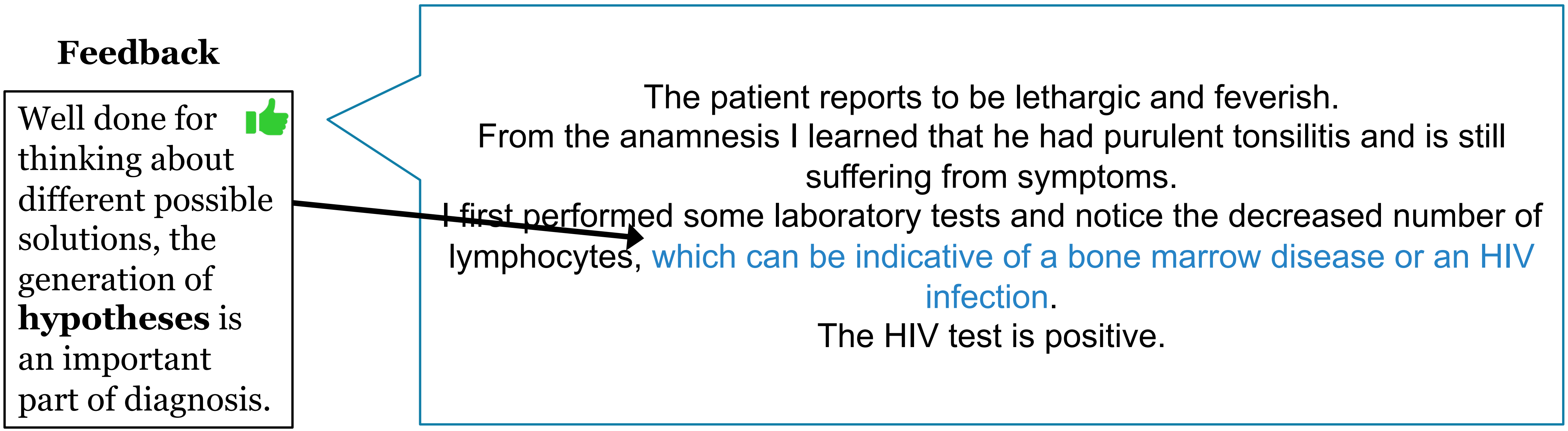
The patient reports to be lethargic and feverish.
From the anamnesis I learned that he had purulent tonsilitis and is still suffering from symptoms.
I first performed some laboratory tests and notice the decreased number of lymphocytes, which can be indicative of a bone marrow disease or an HIV infection.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

Analyse: Reasoning Process


Hypothesis Generation
possible solutions
Evidence Generation
e.g. observations, deduction
Evidence Evaluation
evidence supports solution?
Drawing Conclusions
aggregate evidence to derive final solutions

Diagnostic reasoning steps (epistemic activities)
Fischer et al. 2014

Self-Explanation with Feedback

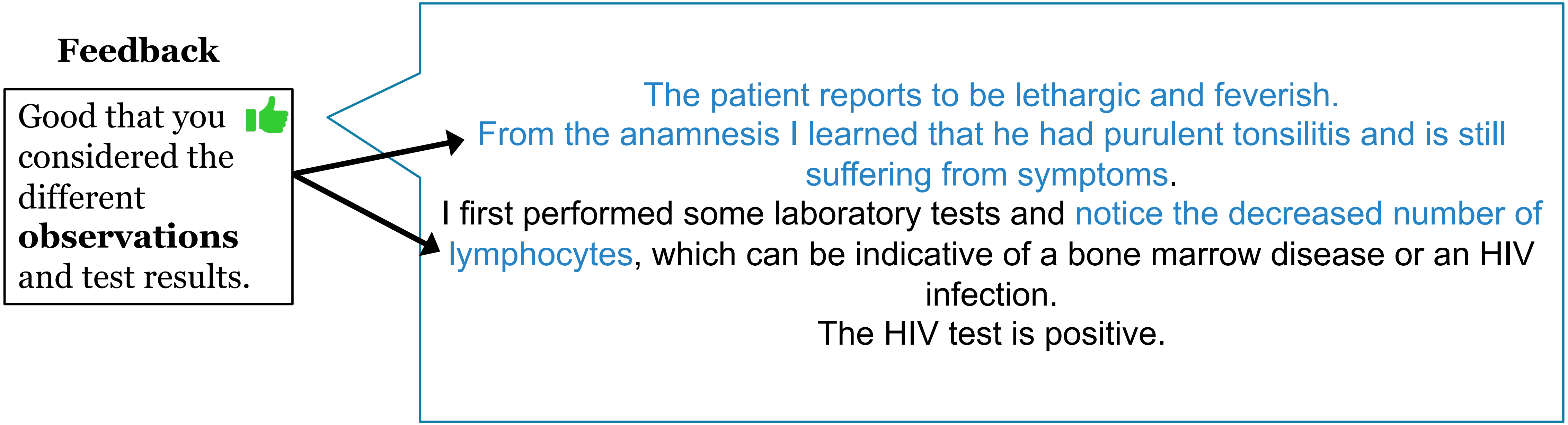


Feedback

Well done for 👍 thinking about different possible solutions, the generation of **hypotheses** is an important part of diagnosis.

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.

Self-Explanation with Feedback



Feedback

Good that you considered the different **observations** and test results. 👍

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.

Self-Explanation with Feedback

Feedback

After collecting 👎 and considering all evidence, you should decide on the most likely **diagnosis**. This is an important duty of a doctor.

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.

Self-Explanation with Reasoning Steps

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   Evidence Generation
Evidence Evaluation   Drawing Conclusions

Detecting Diagnostic Reasoning Steps

1) **Corpus Creation**
2) **Automatic Detection**

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
Evidence Generation
Evidence Evaluation
Drawing Conclusions

Corpus Creation

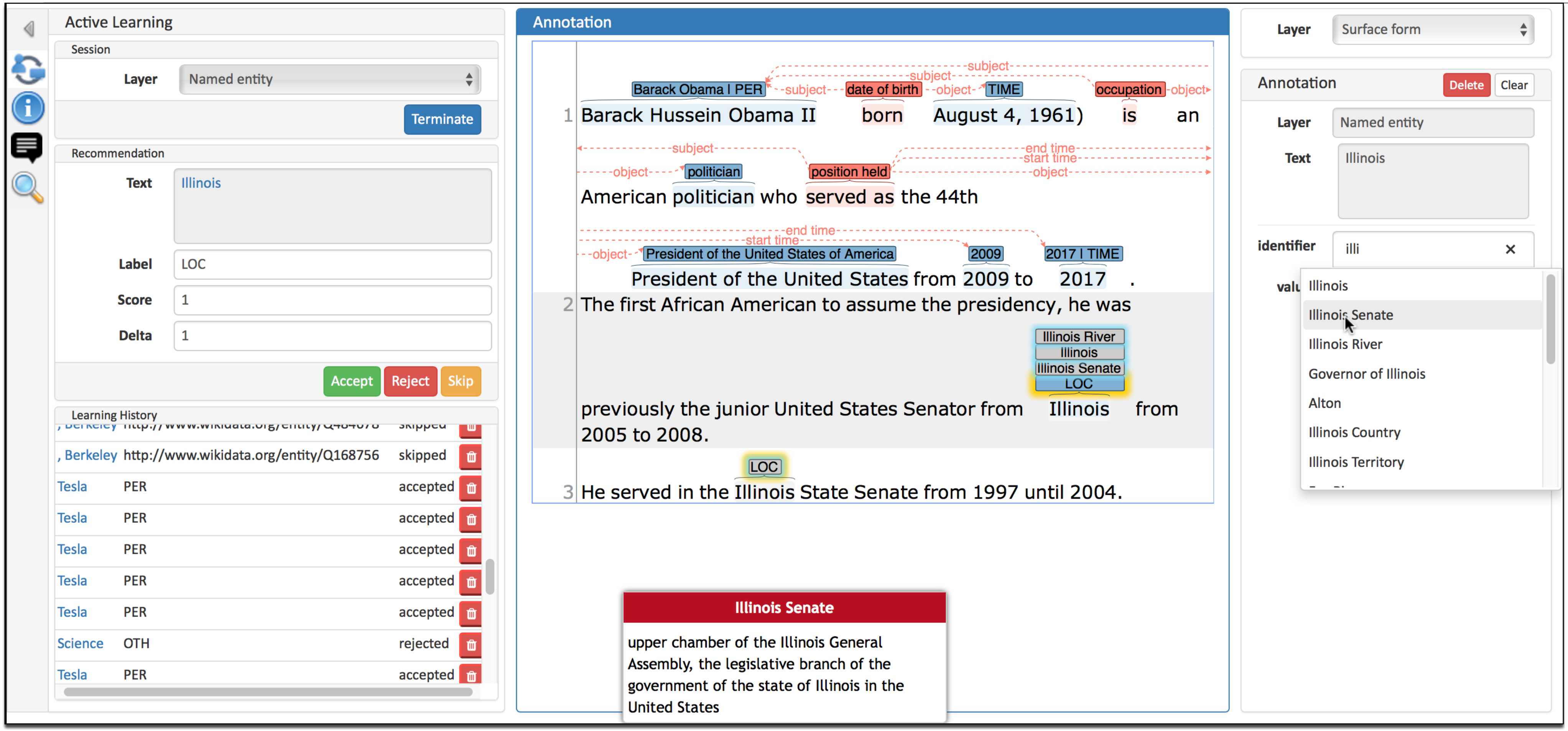
Schulz, Meyer, Gurevych. "Challenges in the Automatic Analysis of Students' Diagnostic Reasoning." *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 2019.

Corpus Creation

- Two domains:
  - Medicine Domain (MeD): 1131 self-explanations → 650 used
  - Teaching Domain (TeD): 976 self-explanations → 550 used
- (Domain) Expert annotators
- Cross-domain annotation scheme
  - Segmentation + classification
  - Easily adaptable to new domains
- German

INCEpTION

https://inception-project.github.io/



Layer: Named entity

Text: Illinois

Label: LOC

Score: 1

Delta: 1

, Berkeley	http://www.wikidata.org/entity/Q484078	skipped
, Berkeley	http://www.wikidata.org/entity/Q168756	skipped
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Tesla	PER	accepted
Science	OTH	rejected
Tesla	PER	accepted

1 Barack Hussein Obama II born August 4, 1961) is an American politician who served as the 44th President of the United States from 2009 to 2017 .

2 The first African American to assume the presidency, he was previously the junior United States Senator from Illinois from 2005 to 2008.

3 He served in the Illinois State Senate from 1997 until 2004.

Illinois Senate

upper chamber of the Illinois General Assembly, the legislative branch of the government of the state of Illinois in the United States

Layer: Surface form

Annotation

Layer: Named entity

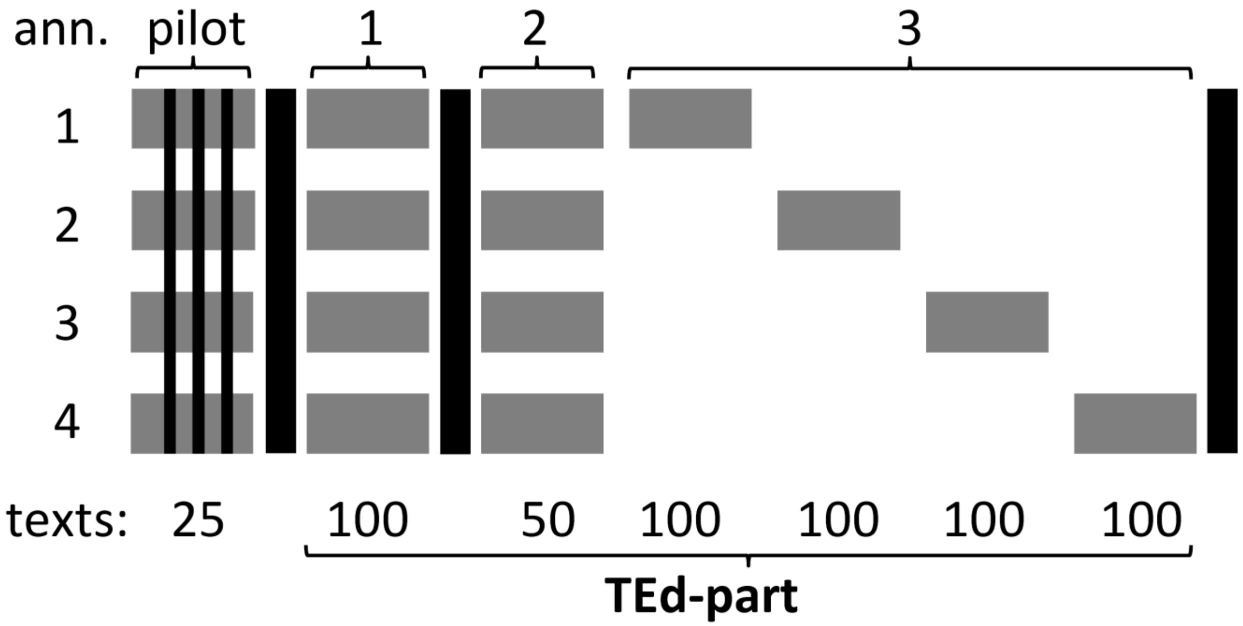
Text: Illinois

identifier: illi

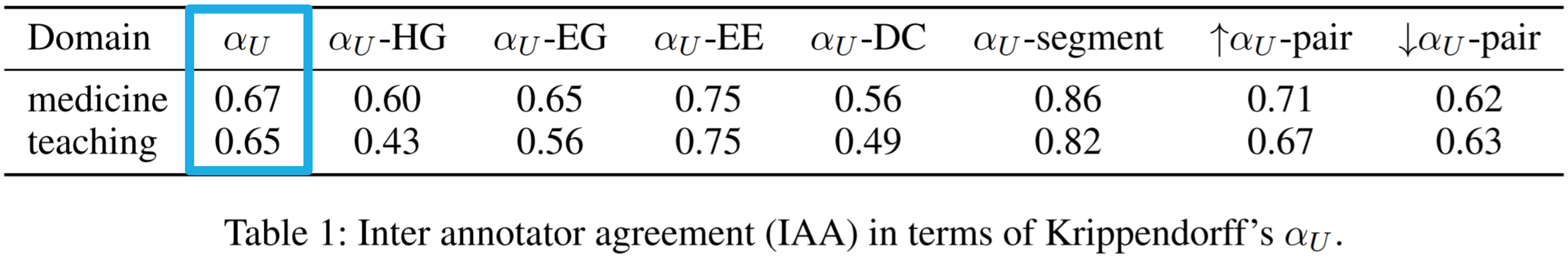
valu

- Illinois
- Illinois Senate
- Illinois River
- Governor of Illinois
- Alton
- Illinois Country
- Illinois Territory

Corpus Creation



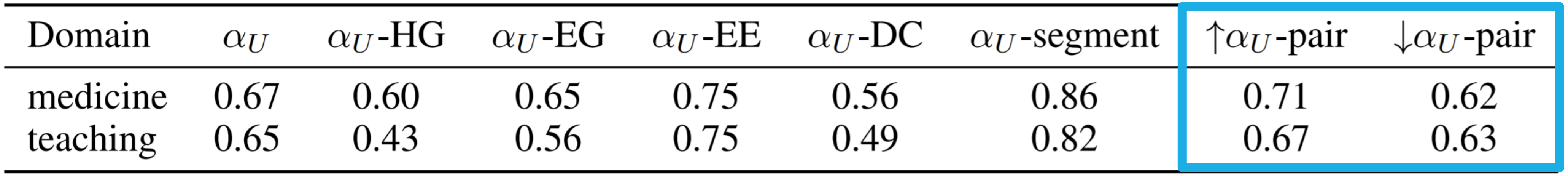
Inter-Annotator Agreement



Domain	α_U	α_U-HG	α_U-EG	α_U-EE	α_U-DC	α_U-segment	$\uparrow\alpha_U$-pair	$\downarrow\alpha_U$-pair
medicine	0.67	0.60	0.65	0.75	0.56	0.86	0.71	0.62
teaching	0.65	0.43	0.56	0.75	0.49	0.82	0.67	0.63

Table 1: Inter annotator agreement (IAA) in terms of Krippendorff's α_U.

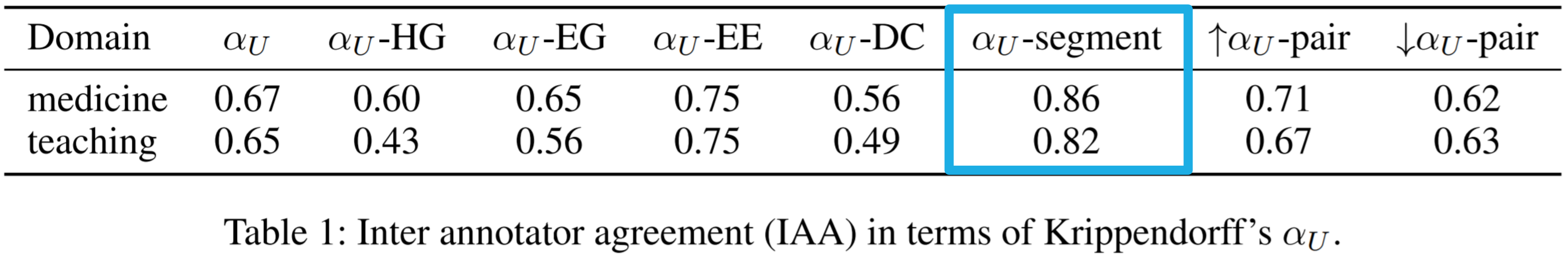
Inter-Annotator Agreement



Domain	α_U	α_U-HG	α_U-EG	α_U-EE	α_U-DC	α_U-segment	$\uparrow\alpha_U$-pair	$\downarrow\alpha_U$-pair
medicine	0.67	0.60	0.65	0.75	0.56	0.86	0.71	0.62
teaching	0.65	0.43	0.56	0.75	0.49	0.82	0.67	0.63

Table 1: Inter annotator agreement (IAA) in terms of Krippendorff's α_U.

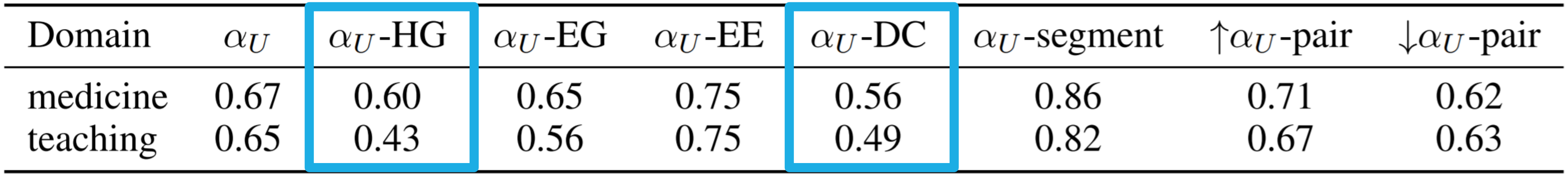
Inter-Annotator Agreement



Domain	α_U	α_U-HG	α_U-EG	α_U-EE	α_U-DC	α_U-segment	$\uparrow\alpha_U$-pair	$\downarrow\alpha_U$-pair
medicine	0.67	0.60	0.65	0.75	0.56	0.86	0.71	0.62
teaching	0.65	0.43	0.56	0.75	0.49	0.82	0.67	0.63

Table 1: Inter annotator agreement (IAA) in terms of Krippendorff's α_U.

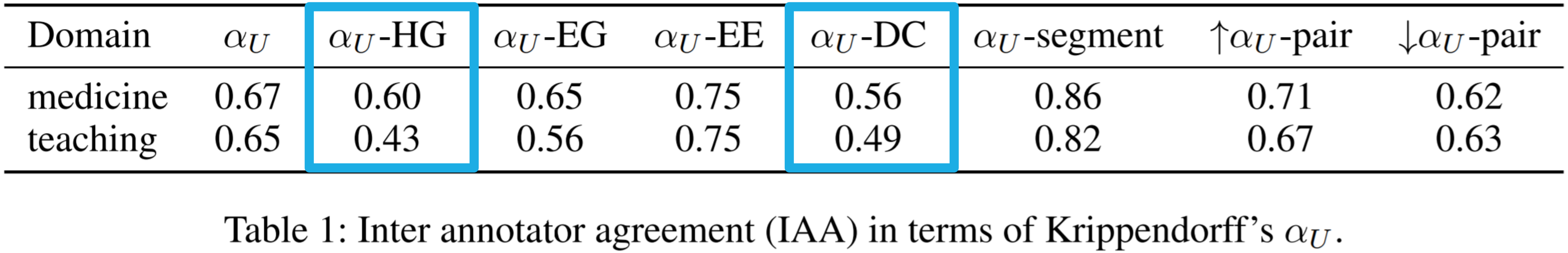
Inter-Annotator Agreement



Domain	α_U	α_U-HG	α_U-EG	α_U-EE	α_U-DC	α_U-segment	$\uparrow\alpha_U$-pair	$\downarrow\alpha_U$-pair
medicine	0.67	0.60	0.65	0.75	0.56	0.86	0.71	0.62
teaching	0.65	0.43	0.56	0.75	0.49	0.82	0.67	0.63

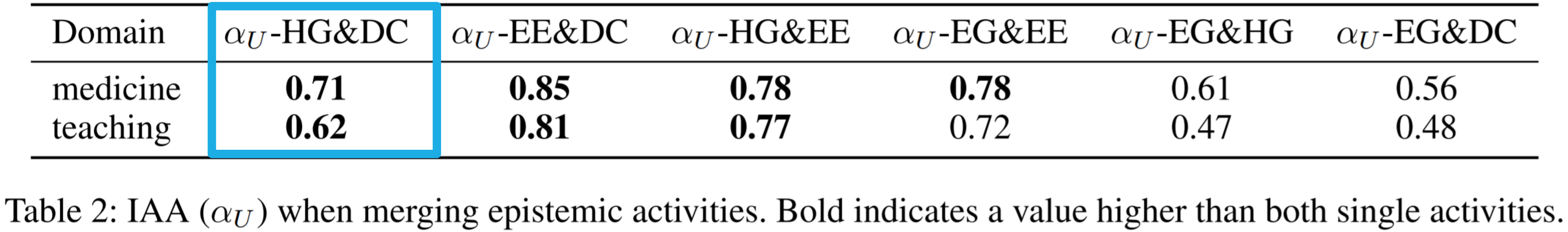
Table 1: Inter annotator agreement (IAA) in terms of Krippendorff's α_U.

Inter-Annotator Agreement



Domain	α_U	α_U-HG	α_U-EG	α_U-EE	α_U-DC	α_U-segment	$\uparrow\alpha_U$-pair	$\downarrow\alpha_U$-pair
medicine	0.67	0.60	0.65	0.75	0.56	0.86	0.71	0.62
teaching	0.65	0.43	0.56	0.75	0.49	0.82	0.67	0.63

Table 1: Inter annotator agreement (IAA) in terms of Krippendorff's α_U.

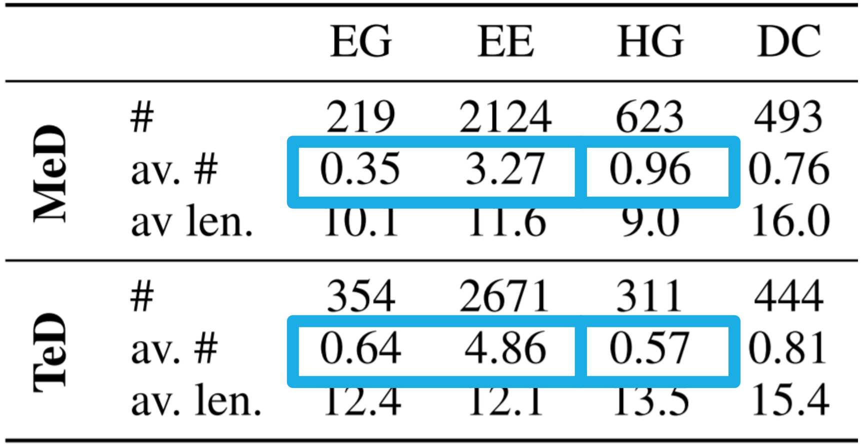


Domain	α_U-HG&DC	α_U-EE&DC	α_U-HG&EE	α_U-EG&EE	α_U-EG&HG	α_U-EG&DC
medicine	**0.71**	**0.85**	**0.78**	**0.78**	0.61	0.56
teaching	**0.62**	**0.81**	**0.77**	0.72	0.47	0.48

Table 2: IAA (α_U) when merging epistemic activities. Bold indicates a value higher than both single activities.

Corpus Statistics

- majority vote (4/5, 3/4) + annotator meeting
- MeD av. length: 63.8 tokens
- TeD av. Length: 100.2 tokens

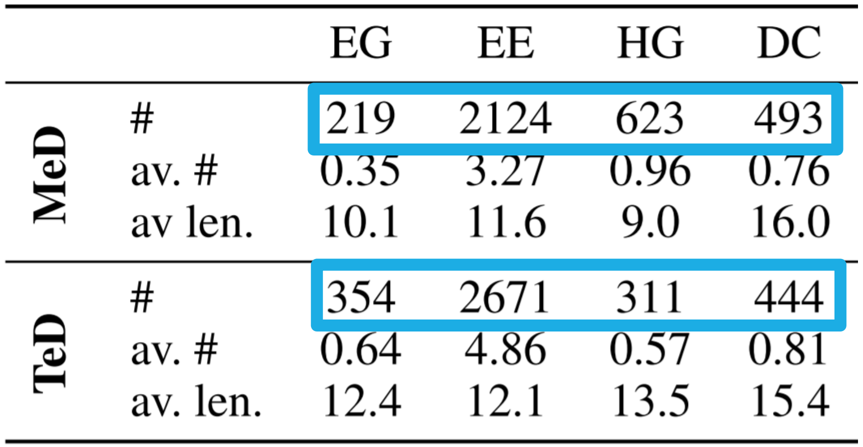


		EG	EE	HG	DC
MeD	#	219	2124	623	493
	av. #	0.35	3.27	0.96	0.76
	av len.	10.1	11.6	9.0	16.0
TeD	#	354	2671	311	444
	av. #	0.64	4.86	0.57	0.81
	av. len.	12.4	12.1	13.5	15.4

Table 3: Corpus statistics in terms of absolute number (#), average number per text (av. #), and average number of tokens (av. len), where EE/EG (and similar) denotes an overlap of an EG and EE segment.

Corpus Statistics

- majority vote (4/5, 3/4) + annotator meeting
- MeD av. length: 63.8 tokens
- TeD av. Length: 100.2 tokens



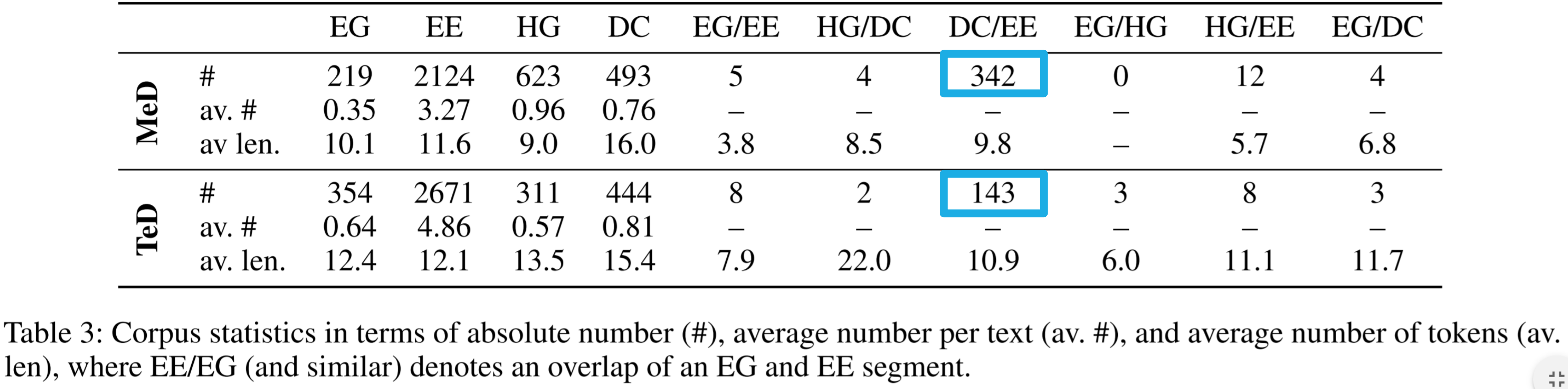
		EG	EE	HG	DC
MeD	#	219	2124	623	493
	av. #	0.35	3.27	0.96	0.76
	av len.	10.1	11.6	9.0	16.0
TeD	#	354	2671	311	444
	av. #	0.64	4.86	0.57	0.81
	av. len.	12.4	12.1	13.5	15.4

Table 3: Corpus statistics in terms of absolute number (#), average number per text (av. #), and average number of tokens (av. len), where EE/EG (and similar) denotes an overlap of an EG and EE segment.

Corpus Statistics

- majority vote (4/5, 3/4) + annotator meeting
- MeD av. length: 63.8 tokens
- TeD av. Length: 100.2 tokens

The x-ray and the subsequent MRI confirmed a vertebral body fracture



		EG	EE	HG	DC	EG/EE	HG/DC	DC/EE	EG/HG	HG/EE	EG/DC
MeD	#	219	2124	623	493	5	4	342	0	12	4
	av. #	0.35	3.27	0.96	0.76	–	–	–	–	–	–
	av len.	10.1	11.6	9.0	16.0	3.8	8.5	9.8	–	5.7	6.8
TeD	#	354	2671	311	444	8	2	143	3	8	3
	av. #	0.64	4.86	0.57	0.81	–	–	–	–	–	–
	av. len.	12.4	12.1	13.5	15.4	7.9	22.0	10.9	6.0	11.1	11.7

Table 3: Corpus statistics in terms of absolute number (#), average number per text (av. #), and average number of tokens (av. len), where EE/EG (and similar) denotes an overlap of an EG and EE segment.

Detecting Diagnostic Reasoning Steps

✓

1) **Corpus Creation**
2) **Automatic Detection**

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 Evidence Generation
Evidence Evaluation Drawing Conclusions

Automatic Detection of Diagnostic Reasoning Steps

Schulz, Meyer, Gurevych. "Challenges in the Automatic Analysis of Students' Diagnostic Reasoning." *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 2019.

Automatic Detection

3 Challenges:

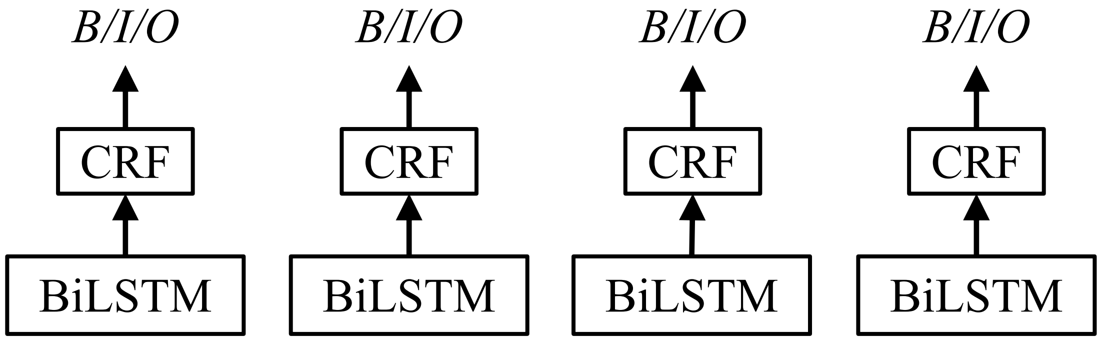
1. segments of arbitrary length (C1),
2. distinguishing different epistemic activity types (C2)
3. overlapping epistemic activity segments (C3)

Multi-class sequence labelling
$C = (\{B, I\} \times A) \cup \{O\}$
$A = \{HG, EG, EE, DC\}$

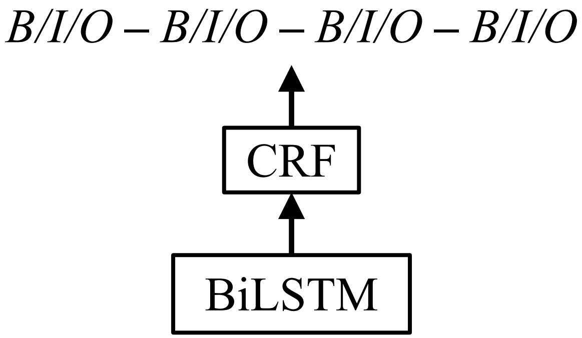
→ **multi-label problem:** $C' \subset C$

Approach: 3 problem transformations

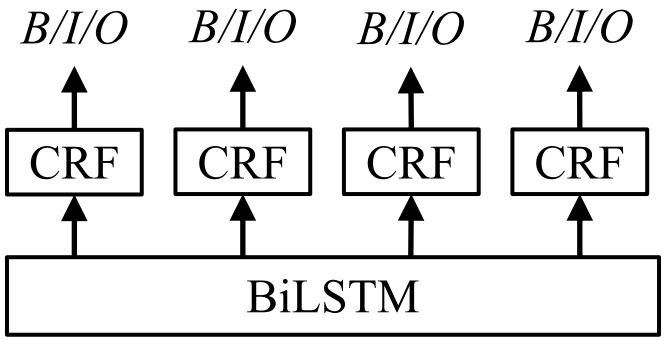
Problem Transformations



SEPARATE:
multiple (single-label)
multi-class problems

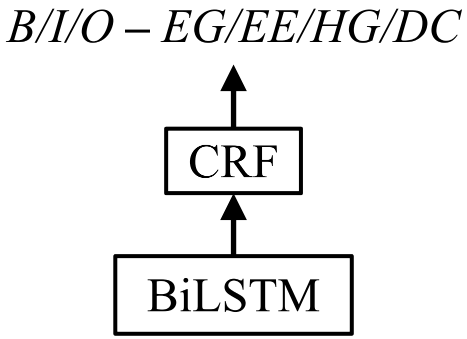


CONCAT:
unique (single-label)
multi-class problem



MULTI-OUTPUT:
Multidimensional
classification problem

Baseline Transformations



1. PREF-BASELINE: unique (single-label) multi-class problem
   - Without overlaps
   - Using preference order: DC > HG > EG > EE

2. MAJ-BASELNE: I-EE for all tokens

Evaluation Metrics

- Hamming Loss

$$HL = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{1}{|C|} \sum_{c \in C} \mathbf{xor}(y_{x,c},\ \hat{y}_{x,c})$$

$$y_{x,c} = \begin{cases} 1 & if\ token\ x\ has\ label\ c \\ 0 & otherwise \end{cases}$$

- C1 (Segmentation)

$$M_S(a) = macro\text{-}F1(C_a, \mathcal{X})$$

$$for\ a\ \in A = \{HG, EG, EE, DC\}$$

$$C_{HG} = \{B - HG, I - HG, O - HG\}$$

- C2 (Type Distinction)

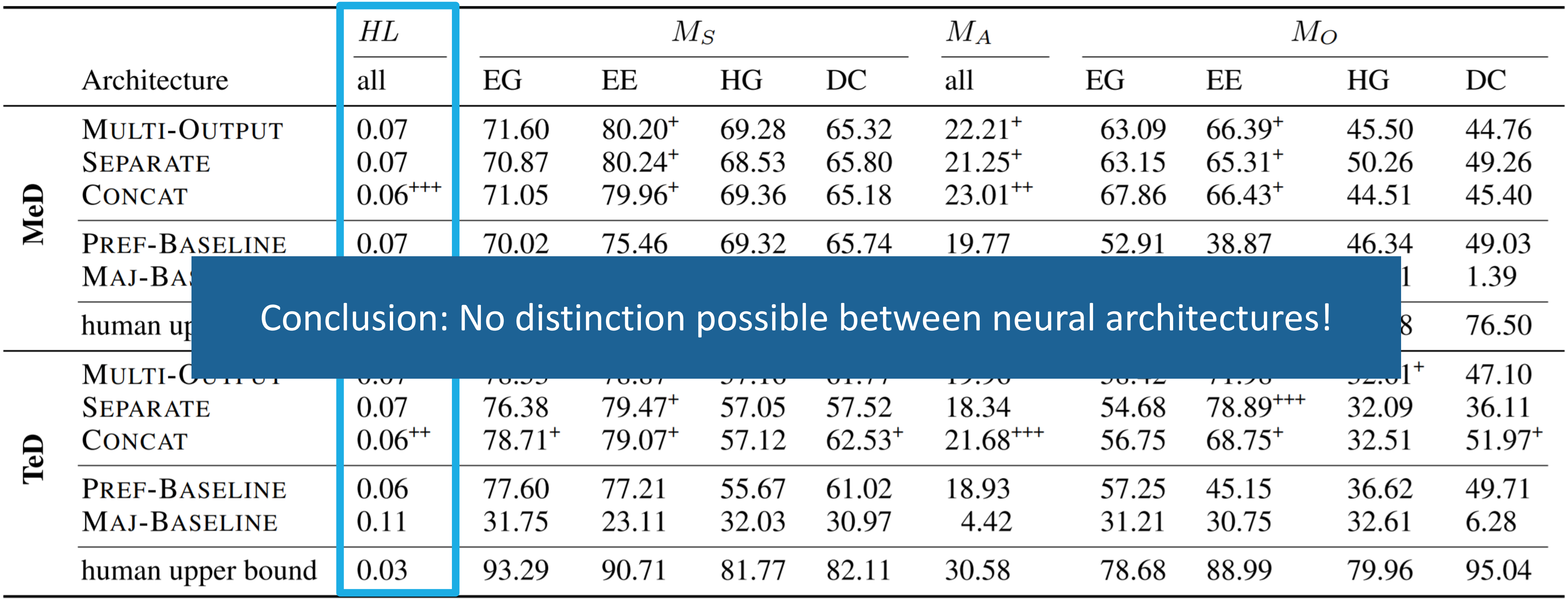
$$M_A = macro\text{-}F1(\mathscr{P}(A), \mathcal{X})$$

- C3 (Overlaps)

$$M_O(a) = macro\text{-}F1(C_a, \mathcal{X}_{\text{overlap}})$$

Automatic Detection: Results

$$HL = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{1}{|C|} \sum_{c \in C} \mathbf{xor}(y_{x,c},\ \hat{y}_{x,c})$$

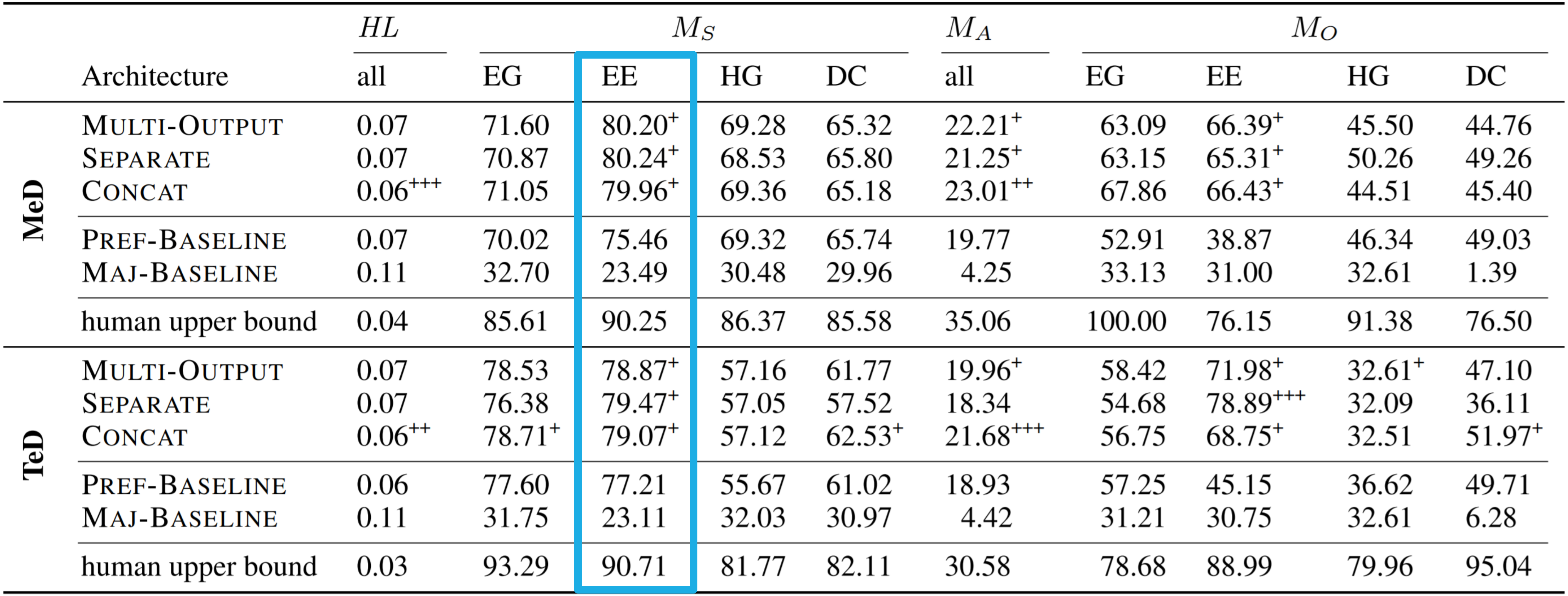


		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-BA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1.39
	human up	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	8	76.50
TeD	MULTI-O	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1^{+}	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Conclusion: No distinction possible between neural architectures!

Automatic Detection: Results

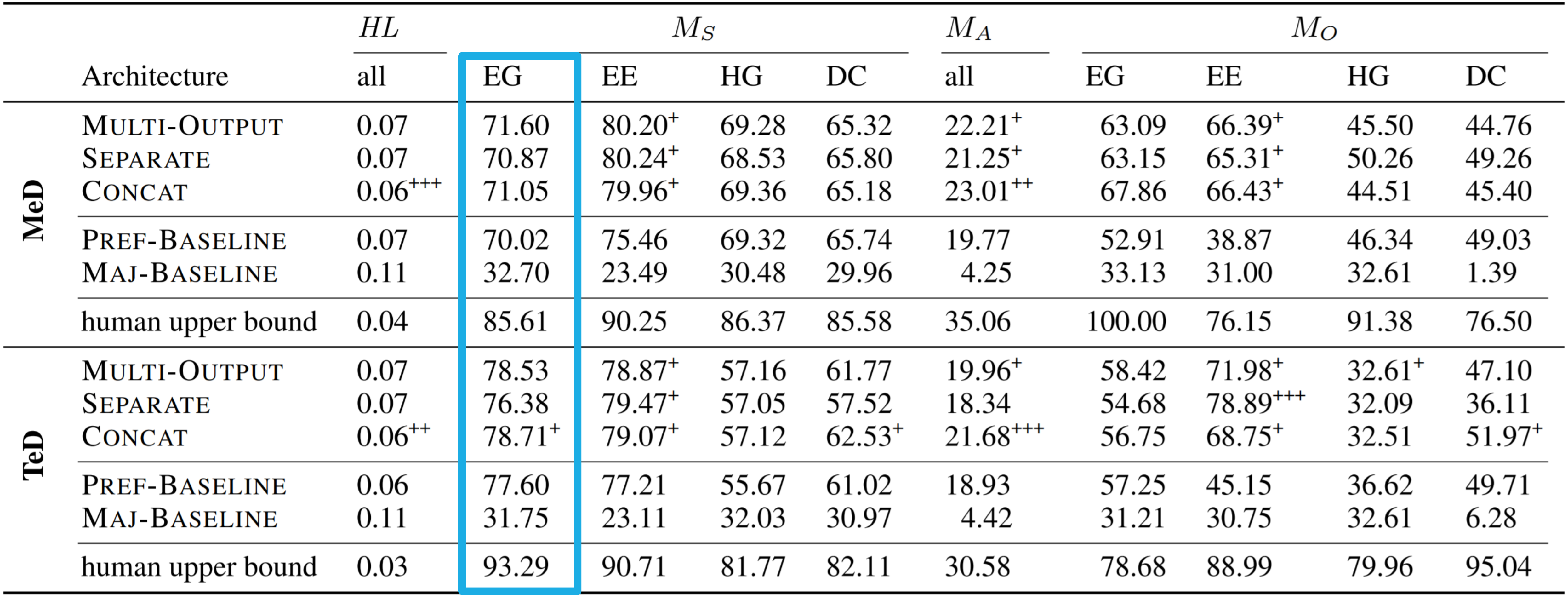
$$M_S(a) = \textit{macro-F1}(C_a, \mathcal{X})$$



		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-BASELINE	0.11	32.70	23.49	30.48	29.96	4.25	33.13	31.00	32.61	1.39
	human upper bound	0.04	85.61	90.25	86.37	85.58	35.06	100.00	76.15	91.38	76.50
TeD	MULTI-OUTPUT	0.07	78.53	78.87^{+}	57.16	61.77	19.96^{+}	58.42	71.98^{+}	32.61^{+}	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Automatic Detection: Results

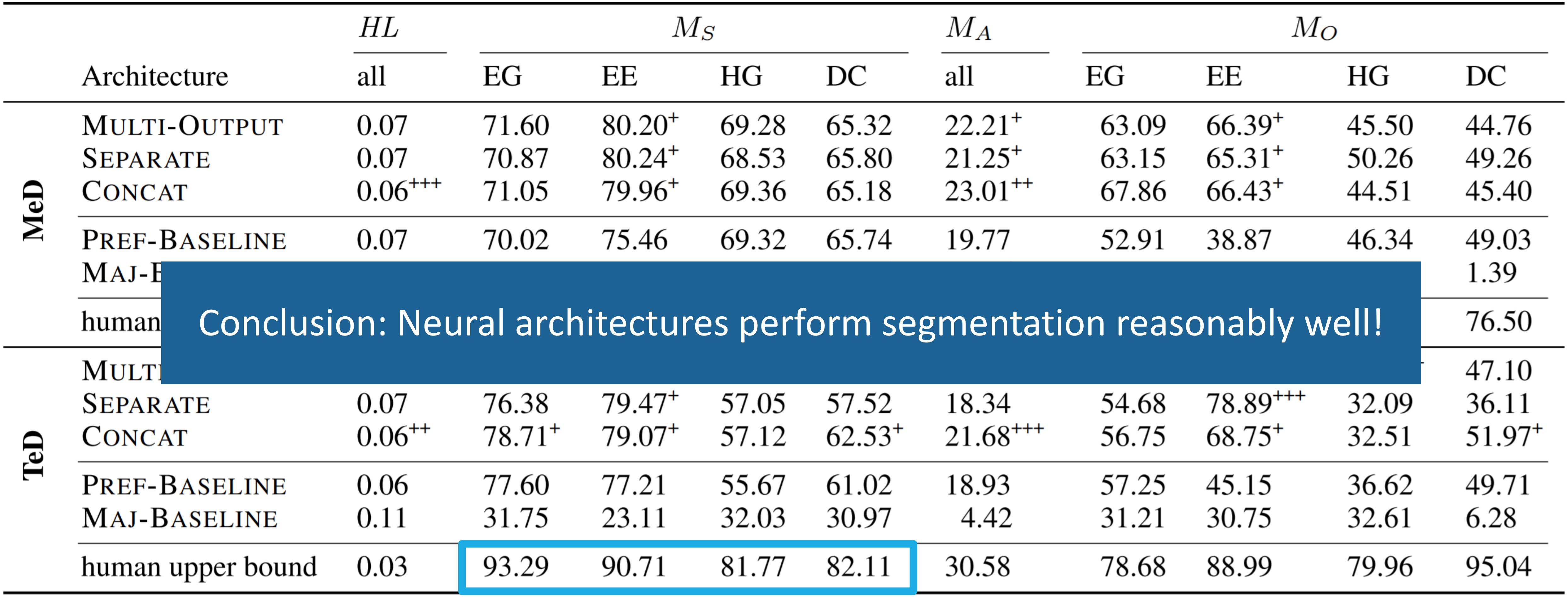
$$M_S(a) = \textit{macro-F1}(C_a, \mathcal{X})$$



		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	Multi-Output	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	Separate	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	Concat	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	Pref-Baseline	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	Maj-Baseline	0.11	32.70	23.49	30.48	29.96	4.25	33.13	31.00	32.61	1.39
	human upper bound	0.04	85.61	90.25	86.37	85.58	35.06	100.00	76.15	91.38	76.50
TeD	Multi-Output	0.07	78.53	78.87^{+}	57.16	61.77	19.96^{+}	58.42	71.98^{+}	32.61^{+}	47.10
	Separate	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	Concat	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	Pref-Baseline	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	Maj-Baseline	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Automatic Detection: Results

$$M_S(a) = \textit{macro-F1}(C_a, \mathcal{X})$$



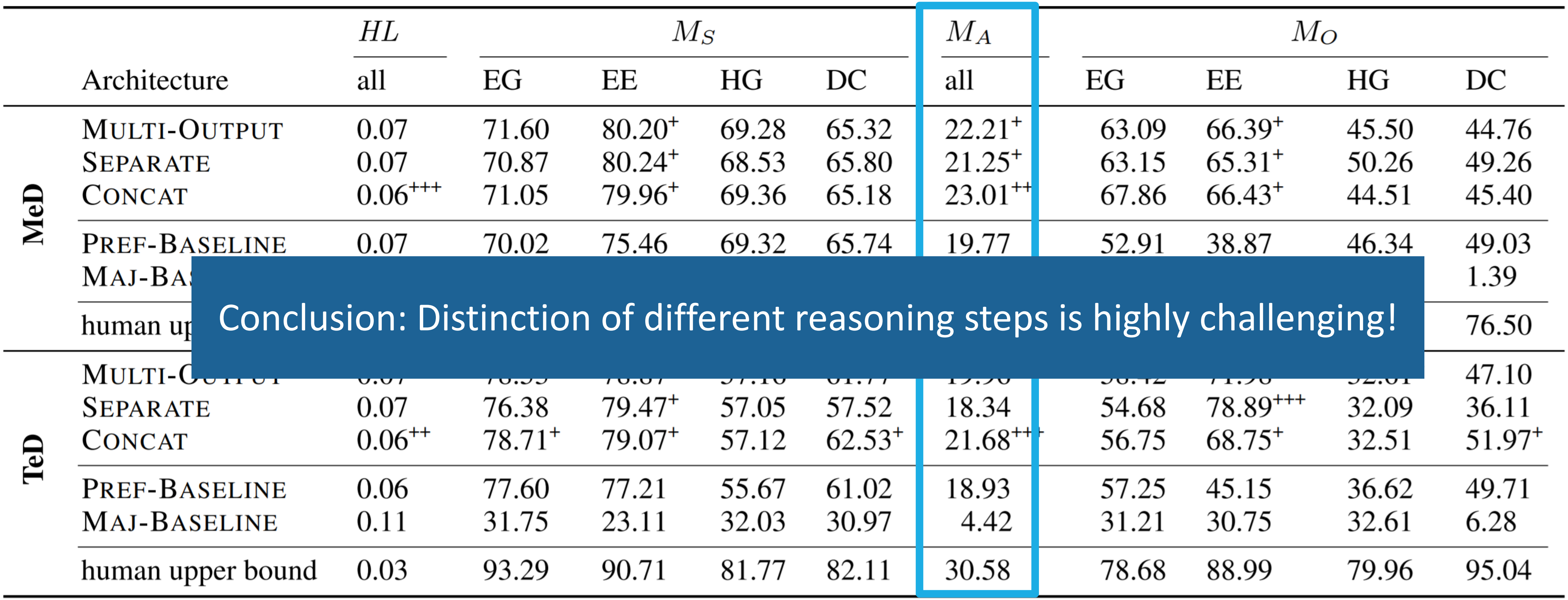
		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.39
	human	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	76.50
TeD	MULTI	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Conclusion: Neural architectures perform segmentation reasonably well!

Upper Bound: 62.5

$M_A = macro\text{-}F1(\mathscr{P}(A), \mathcal{X})$

Automatic Detection: Results



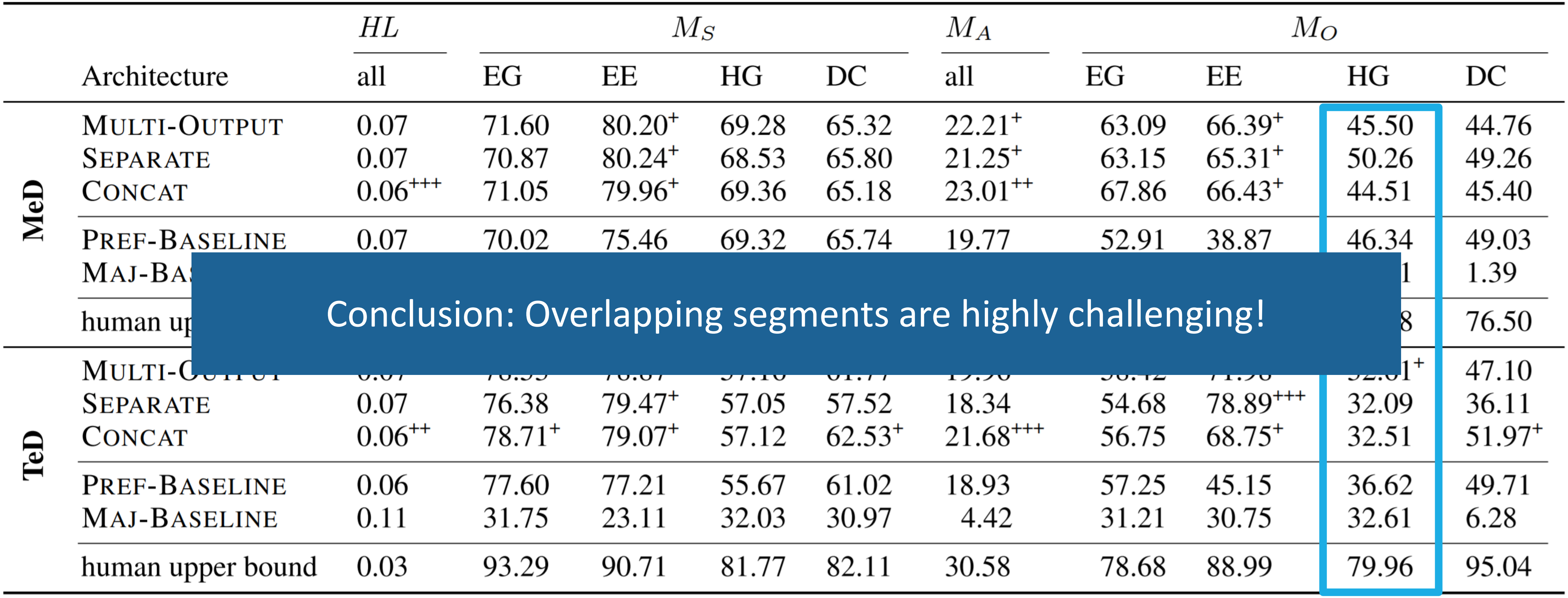
		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-BA[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.39
	human up[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	76.50
TeD	MULTI-O[illegible]	0.07	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Conclusion: Distinction of different reasoning steps is highly challenging!

Automatic Detection: Results

Upper Bound: 62.5

$M_A = \textit{macro-F1}(\mathscr{P}(A), \mathcal{X})$



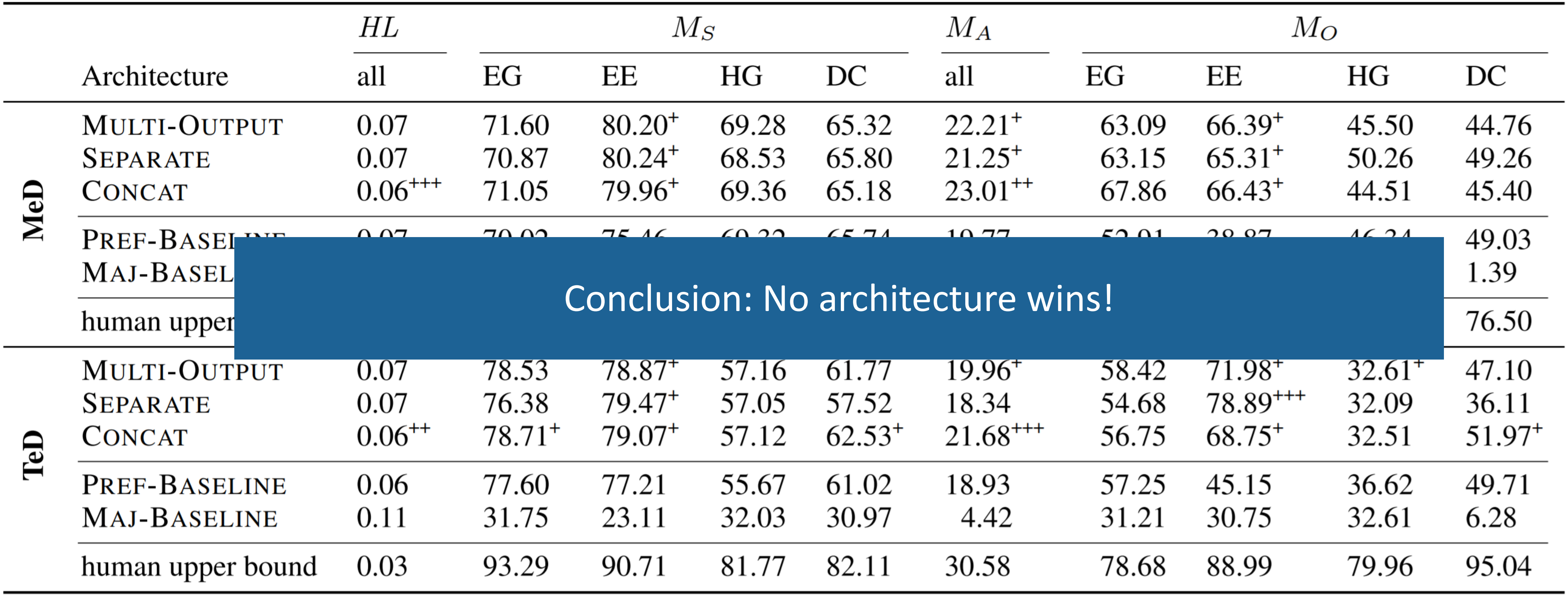
	Architecture	*HL*	M_S				M_A	M_O			
		all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-BA[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]1	1.39
	human up[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]8	76.50
TeD	MULTI-O[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]1^{+}	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Conclusion: Overlapping segments are highly challenging!

Automatic Detection: Results

Upper Bound: 62.5

$M_A = macro\text{-}F1(\mathscr{P}(A), \mathcal{X})$



		HL	M_S				M_A	M_O			
	Architecture	all	EG	EE	HG	DC	all	EG	EE	HG	DC
MeD	MULTI-OUTPUT	0.07	71.60	80.20^{+}	69.28	65.32	22.21^{+}	63.09	66.39^{+}	45.50	44.76
	SEPARATE	0.07	70.87	80.24^{+}	68.53	65.80	21.25^{+}	63.15	65.31^{+}	50.26	49.26
	CONCAT	0.06^{+++}	71.05	79.96^{+}	69.36	65.18	23.01^{++}	67.86	66.43^{+}	44.51	45.40
	PREF-BASELINE	0.07	70.02	75.46	69.32	65.74	19.77	52.91	38.87	46.34	49.03
	MAJ-BASEL	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.39
	human upper	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	76.50
TeD	MULTI-OUTPUT	0.07	78.53	78.87^{+}	57.16	61.77	19.96^{+}	58.42	71.98^{+}	32.61^{+}	47.10
	SEPARATE	0.07	76.38	79.47^{+}	57.05	57.52	18.34	54.68	78.89^{+++}	32.09	36.11
	CONCAT	0.06^{++}	78.71^{+}	79.07^{+}	57.12	62.53^{+}	21.68^{+++}	56.75	68.75^{+}	32.51	51.97^{+}
	PREF-BASELINE	0.06	77.60	77.21	55.67	61.02	18.93	57.25	45.15	36.62	49.71
	MAJ-BASELINE	0.11	31.75	23.11	32.03	30.97	4.42	31.21	30.75	32.61	6.28
	human upper bound	0.03	93.29	90.71	81.77	82.11	30.58	78.68	88.99	79.96	95.04

Conclusion: No architecture wins!

Detecting Diagnostic Reasoning Steps

✓ 1) **Corpus Creation**

✓ 2) **Automatic Detection**

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.

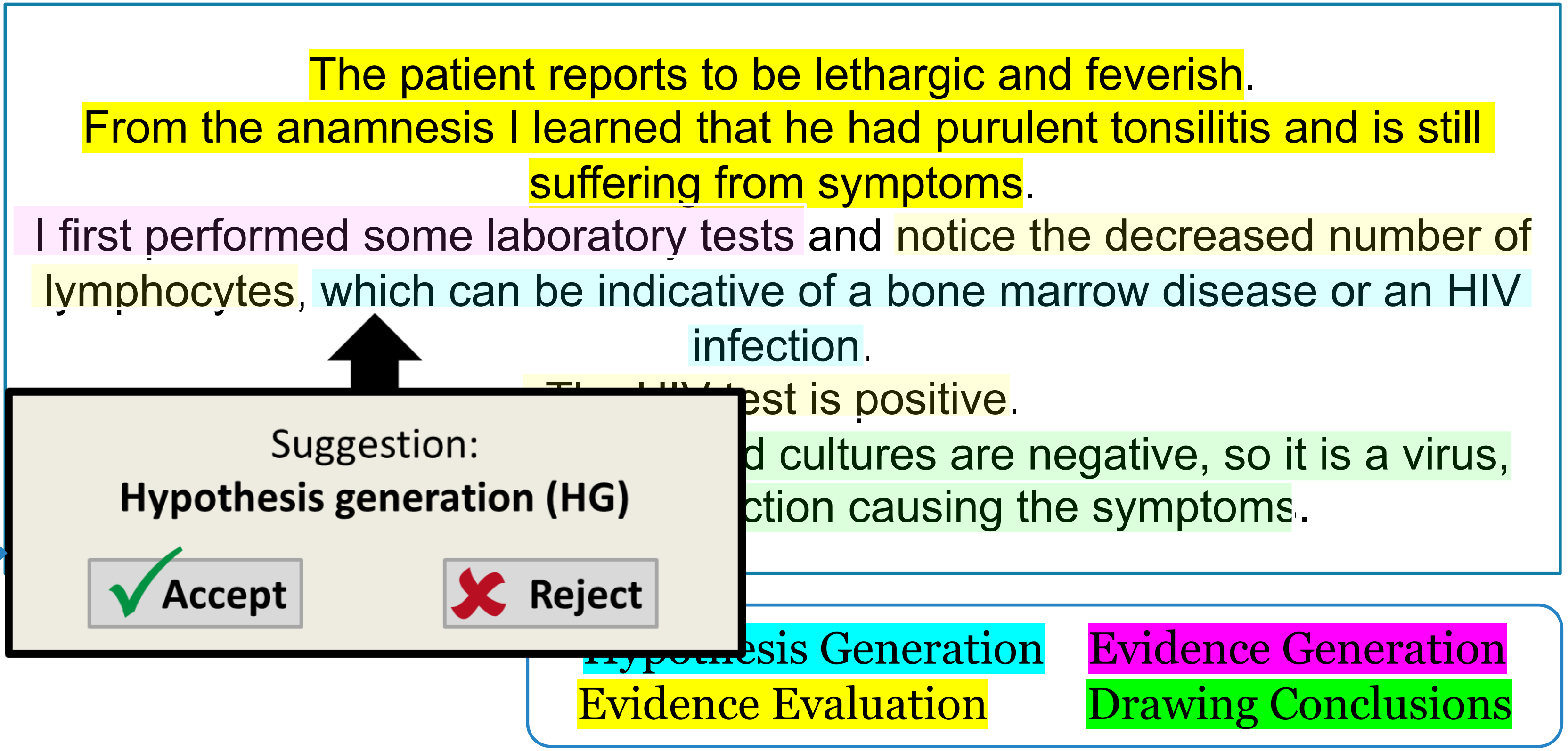
Hypothesis Generation
Evidence Generation
Evidence Evaluation
Drawing Conclusions

Annotating the 2nd half of self-explanations

Schulz, et al. "Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains."
Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

Extending the Corpus

INCEpTION



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.
d cultures are negative, so it is a virus,
ction causing the symptoms.

Suggestion:
Hypothesis generation (HG)

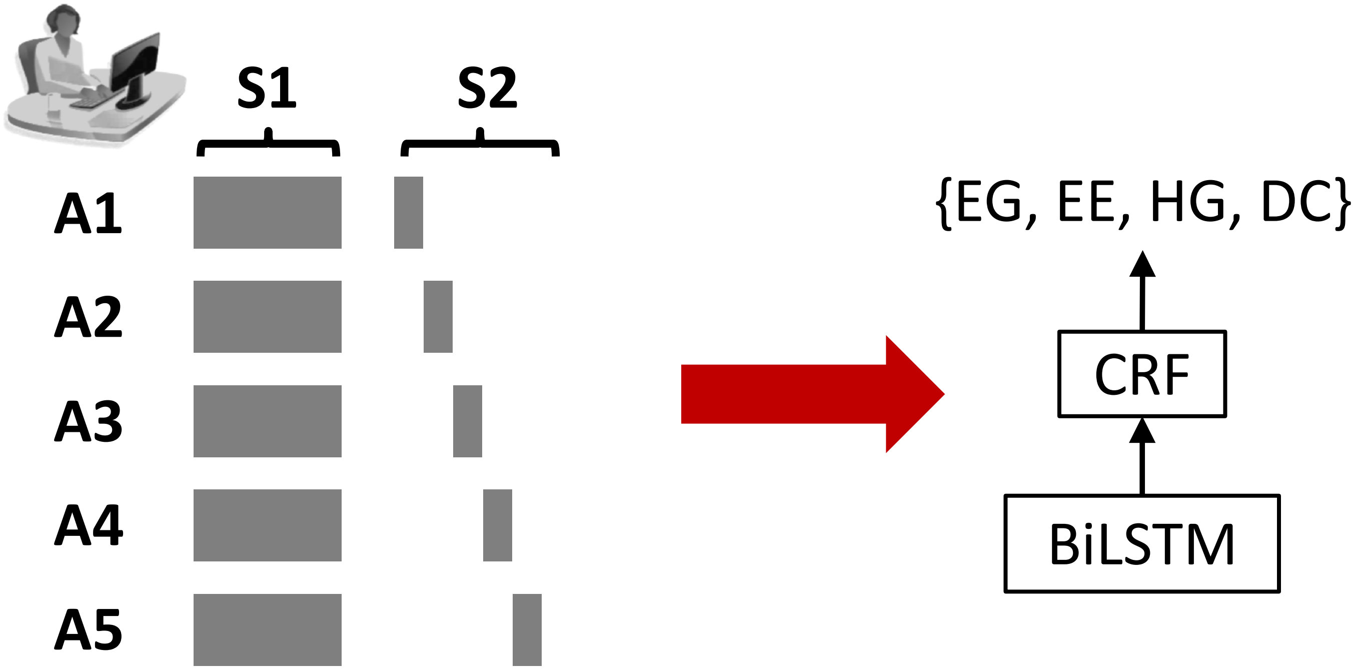
✓ Accept

✗ Reject

Hypothesis Generation
Evidence Generation
Evidence Evaluation
Drawing Conclusions

predictions
(PREF-BASELINE)

Training Data and Suggestion Quality



S1

S2

A1

A2

A3

A4

A5

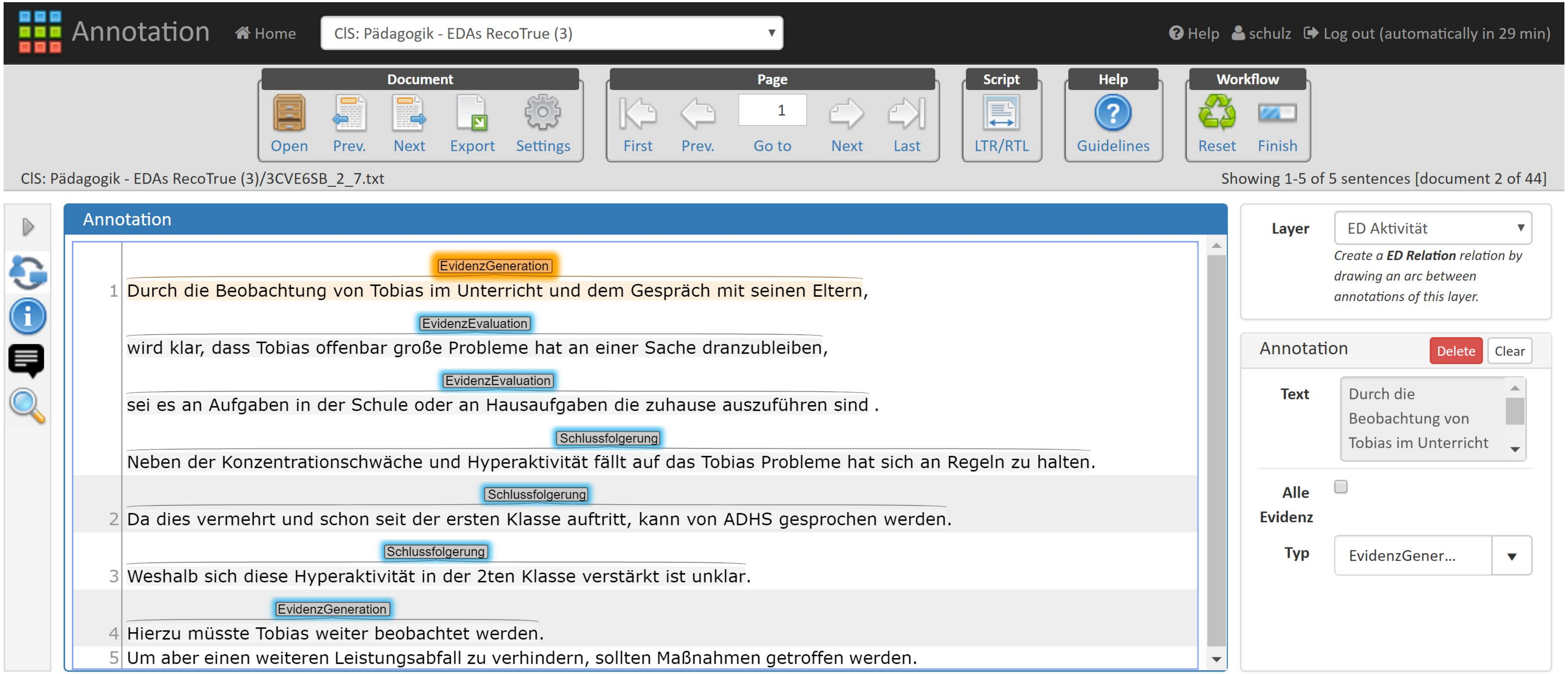
{EG, EE, HG, DC}

CRF

BiLSTM

univ(ersal) model: $F_1 \approx .63$
pers(onalized) models: $F_1 \approx .55$

Annotation Suggestions in INCEpTION



CIS: Pädagogik - EDAs RecoTrue (3)/3CVE6SB_2_7.txt

Showing 1-5 of 5 sentences [document 2 of 44]

Annotation

EvidenzGeneration
1 Durch die Beobachtung von Tobias im Unterricht und dem Gespräch mit seinen Eltern,

EvidenzEvaluation
wird klar, dass Tobias offenbar große Probleme hat an einer Sache dranzubleiben,

EvidenzEvaluation
sei es an Aufgaben in der Schule oder an Hausaufgaben die zuhause auszuführen sind .

Schlussfolgerung
Neben der Konzentrationschwäche und Hyperaktivität fällt auf das Tobias Probleme hat sich an Regeln zu halten.

Schlussfolgerung
2 Da dies vermehrt und schon seit der ersten Klasse auftritt, kann von ADHS gesprochen werden.

Schlussfolgerung
3 Weshalb sich diese Hyperaktivität in der 2ten Klasse verstärkt ist unklar.

EvidenzGeneration
4 Hierzu müsste Tobias weiter beobachtet werden.

5 Um aber einen weiteren Leistungsabfall zu verhindern, sollten Maßnahmen getroffen werden.

Layer: ED Aktivität

*Create a **ED Relation** relation by drawing an arc between annotations of this layer.*

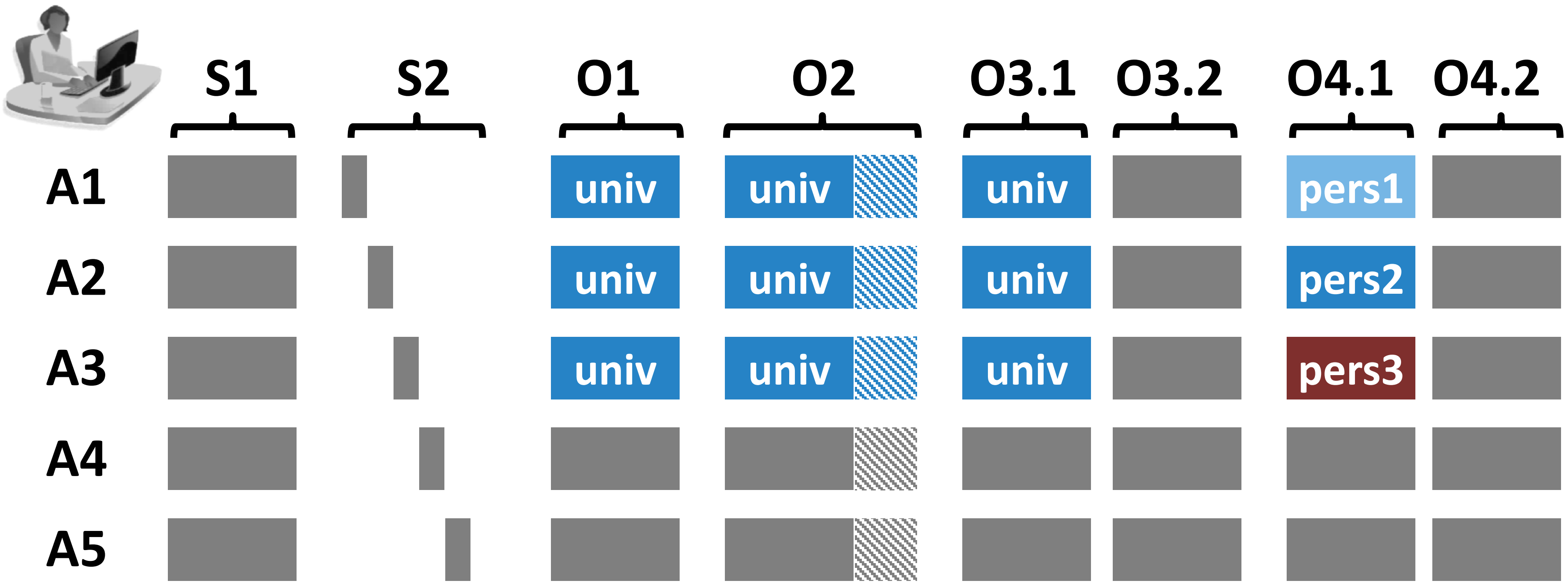
Annotation Delete Clear

Text: Durch die Beobachtung von Tobias im Unterricht

Alle Evidenz: □

Typ: EvidenzGener...

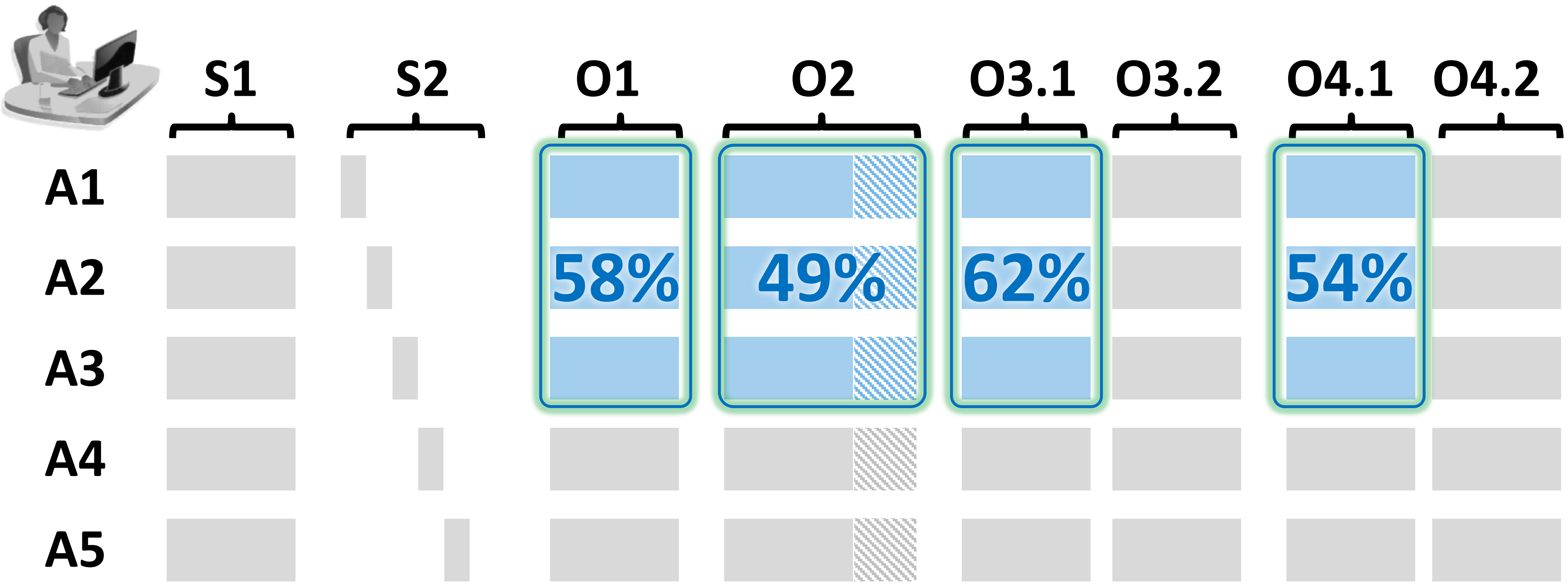
Annotation Suggestions - Setup



	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	■	■	■	■ ▨	■	■	■	■

Usefulness of Annotations

[Percentage of accepted suggestions]



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

A1

A2

A3

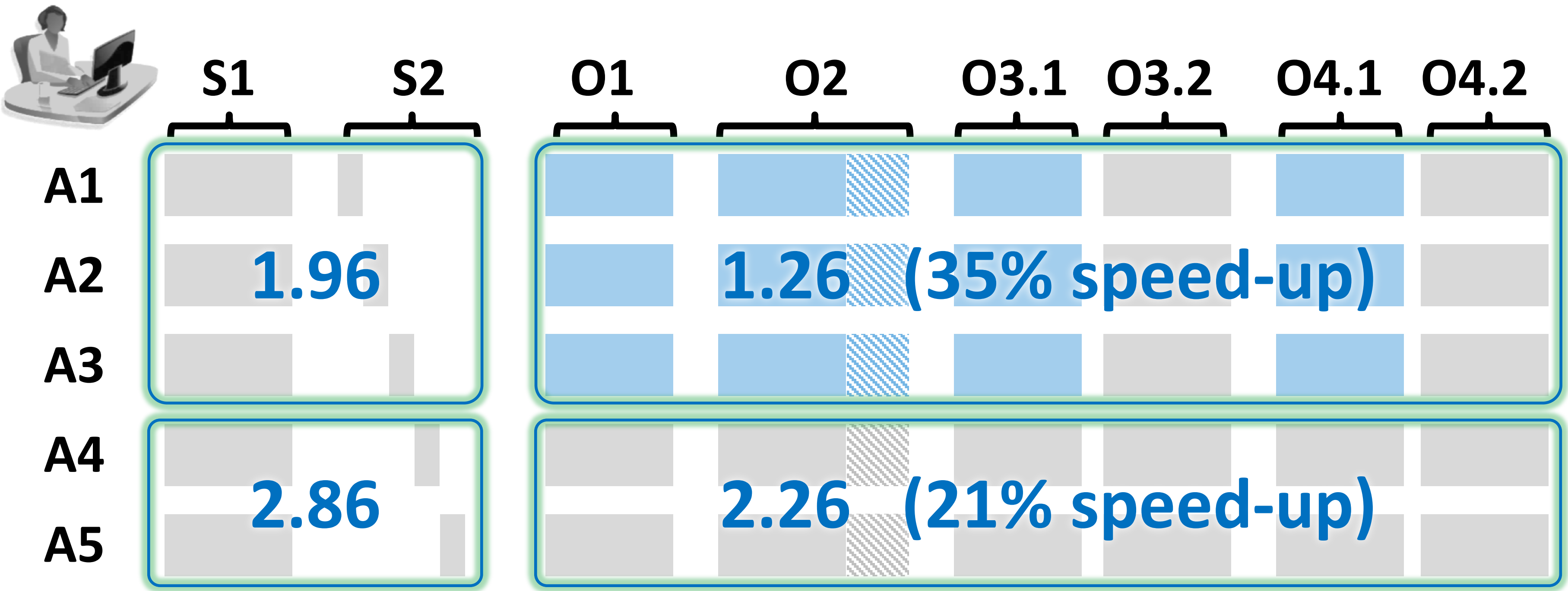
A4

A5

58% 49% 62% 54%

Annotation Time

[Minutes per text]



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

A1

A2

A3

A4

A5

1.96

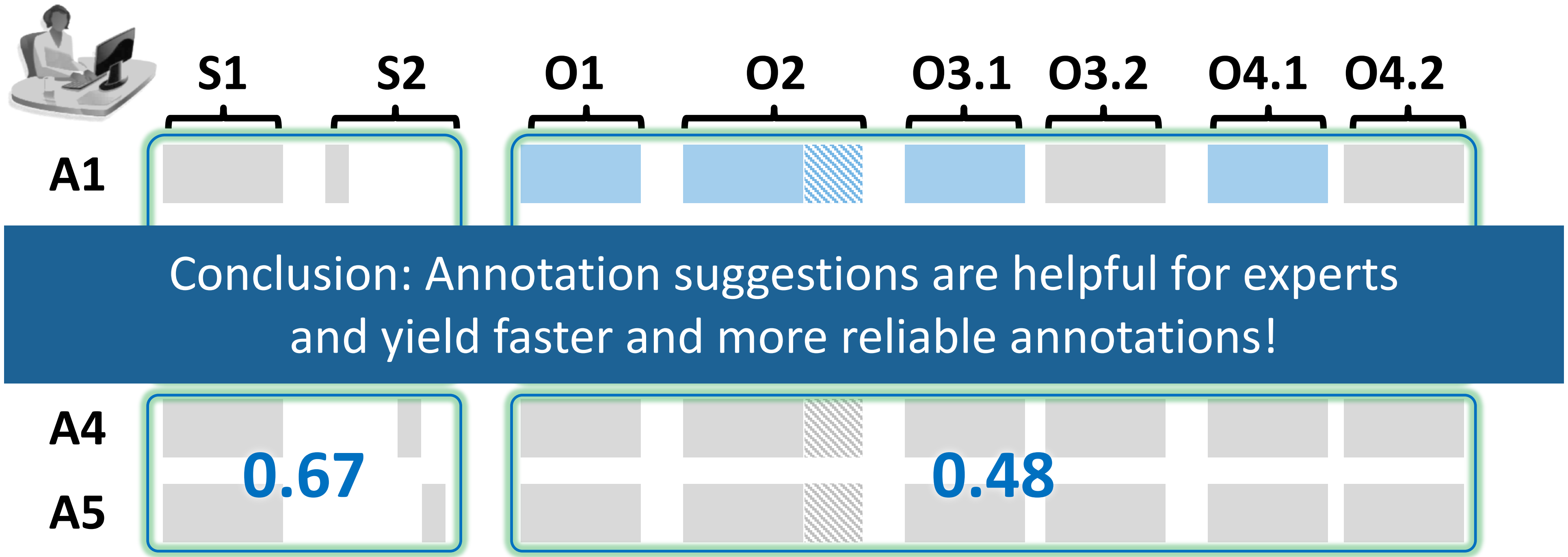
1.26 (35% speed-up)

2.86

2.26 (21% speed-up)

Reliability of Annotations

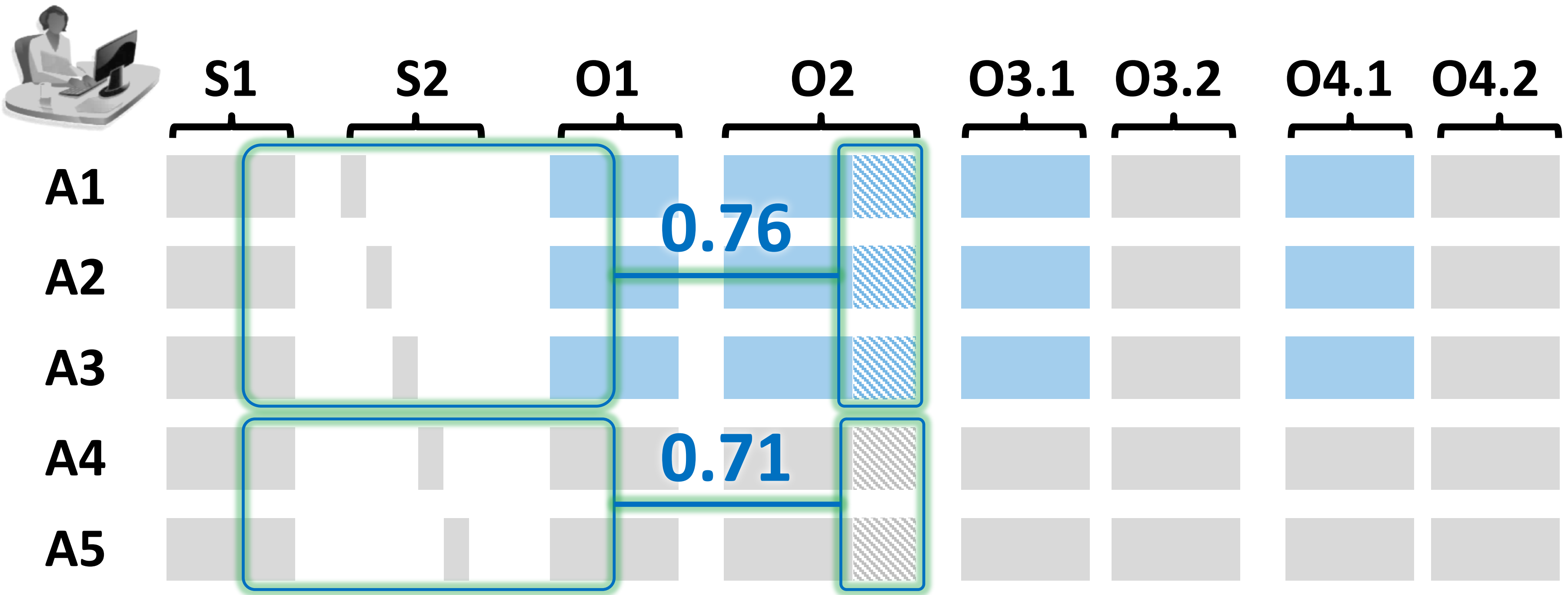
[Krippendorff’s α]



Conclusion: Annotation suggestions are helpful for experts and yield faster and more reliable annotations!

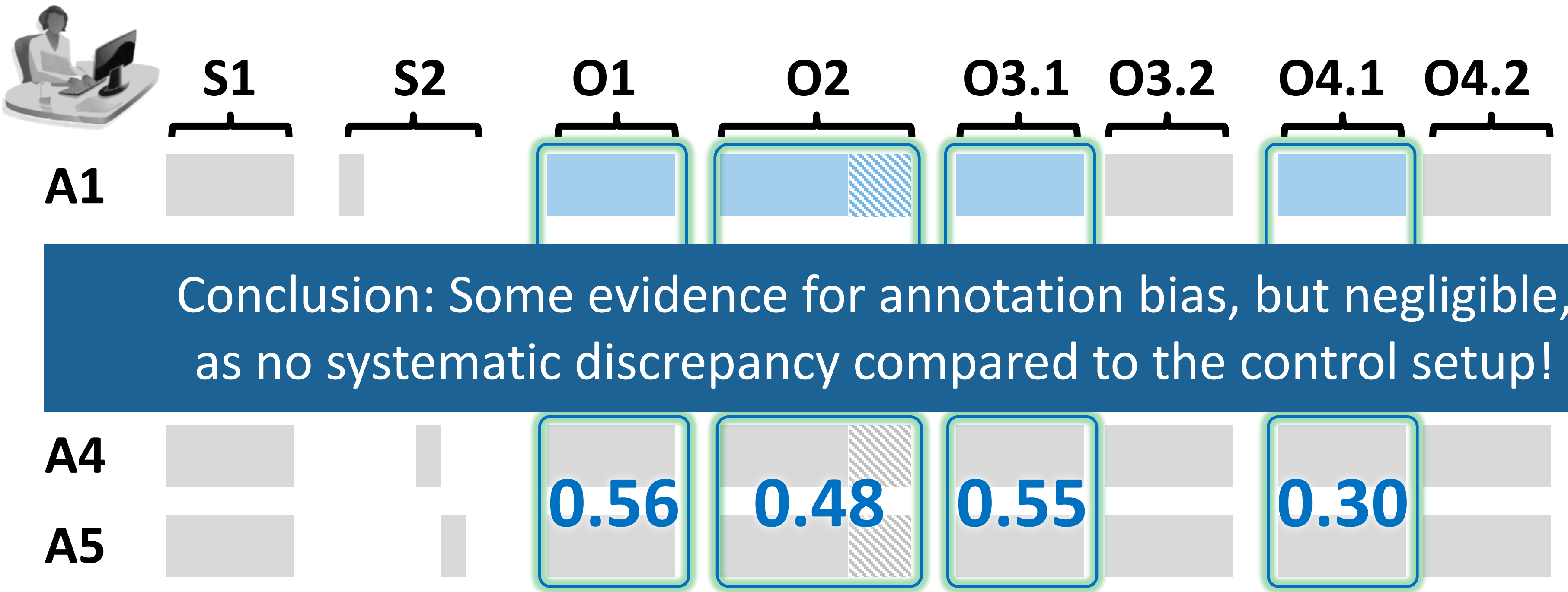
Intra-Annotator Consistency

[Krippendorff's α]



Human / Suggestion Model Agreement

[Krippendorff’s α]



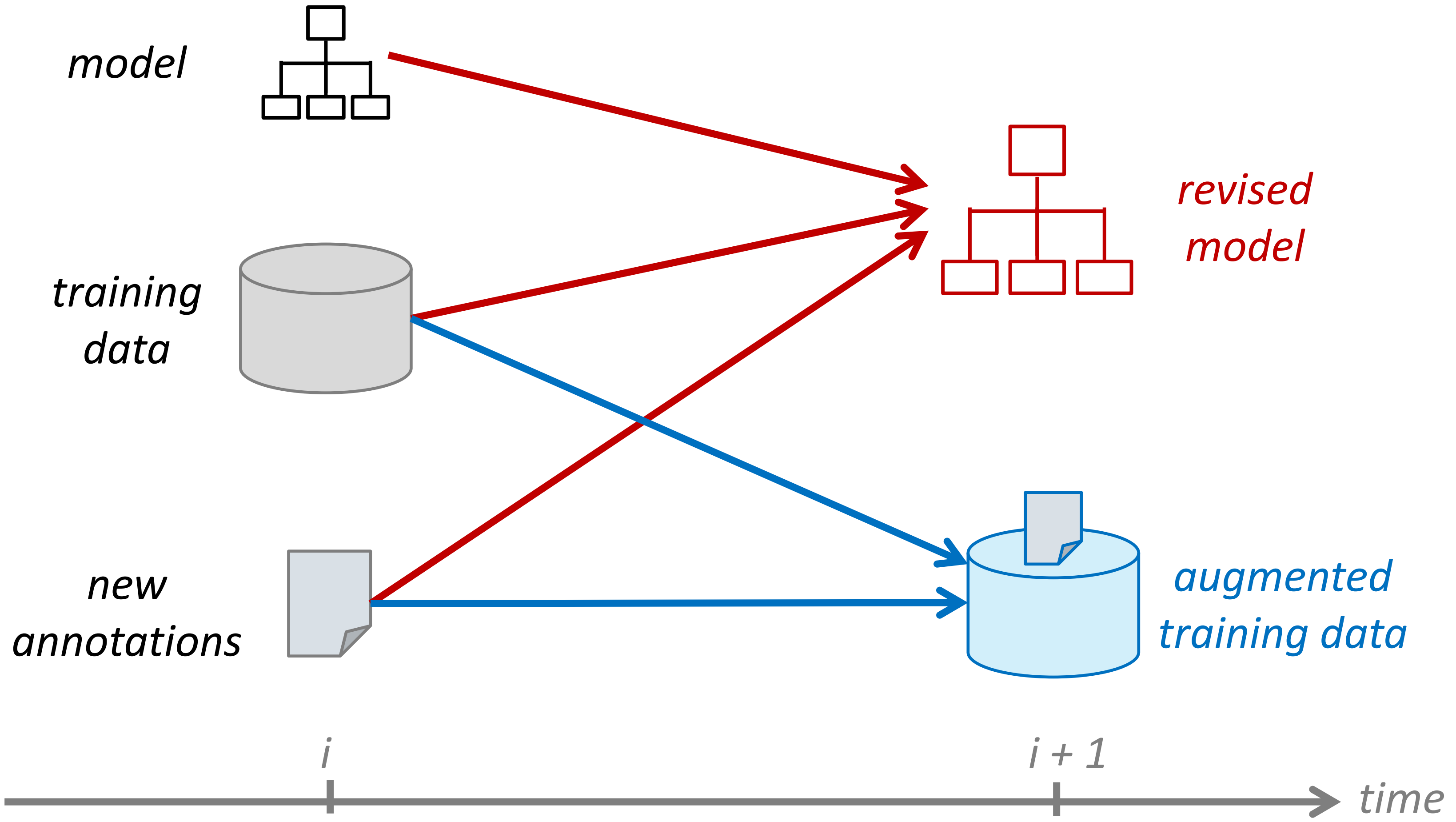
Conclusion: Some evidence for annotation bias, but negligible, as no systematic discrepancy compared to the control setup!

Iterative Model Training

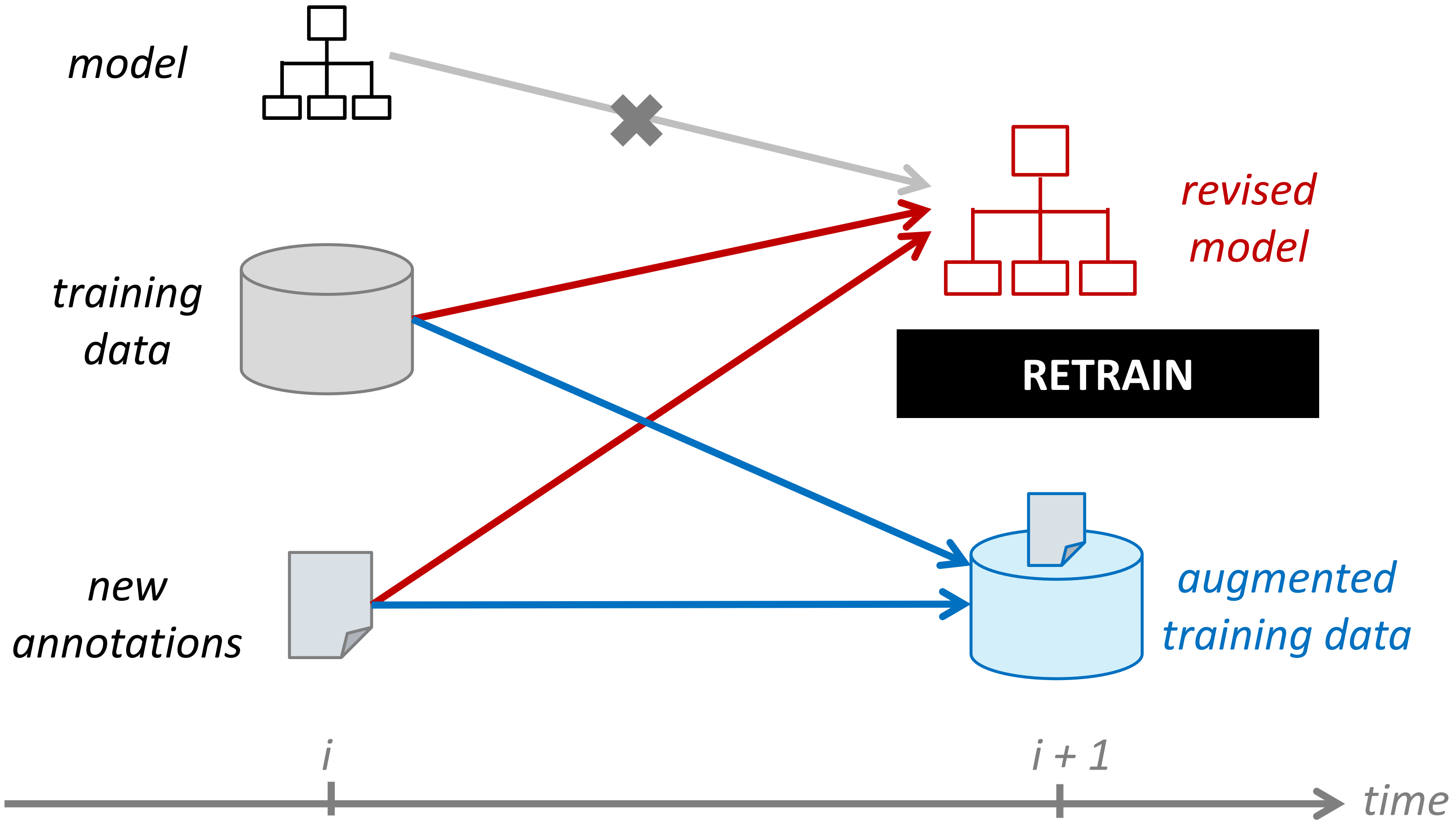
Schulz, et al. "Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains."

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

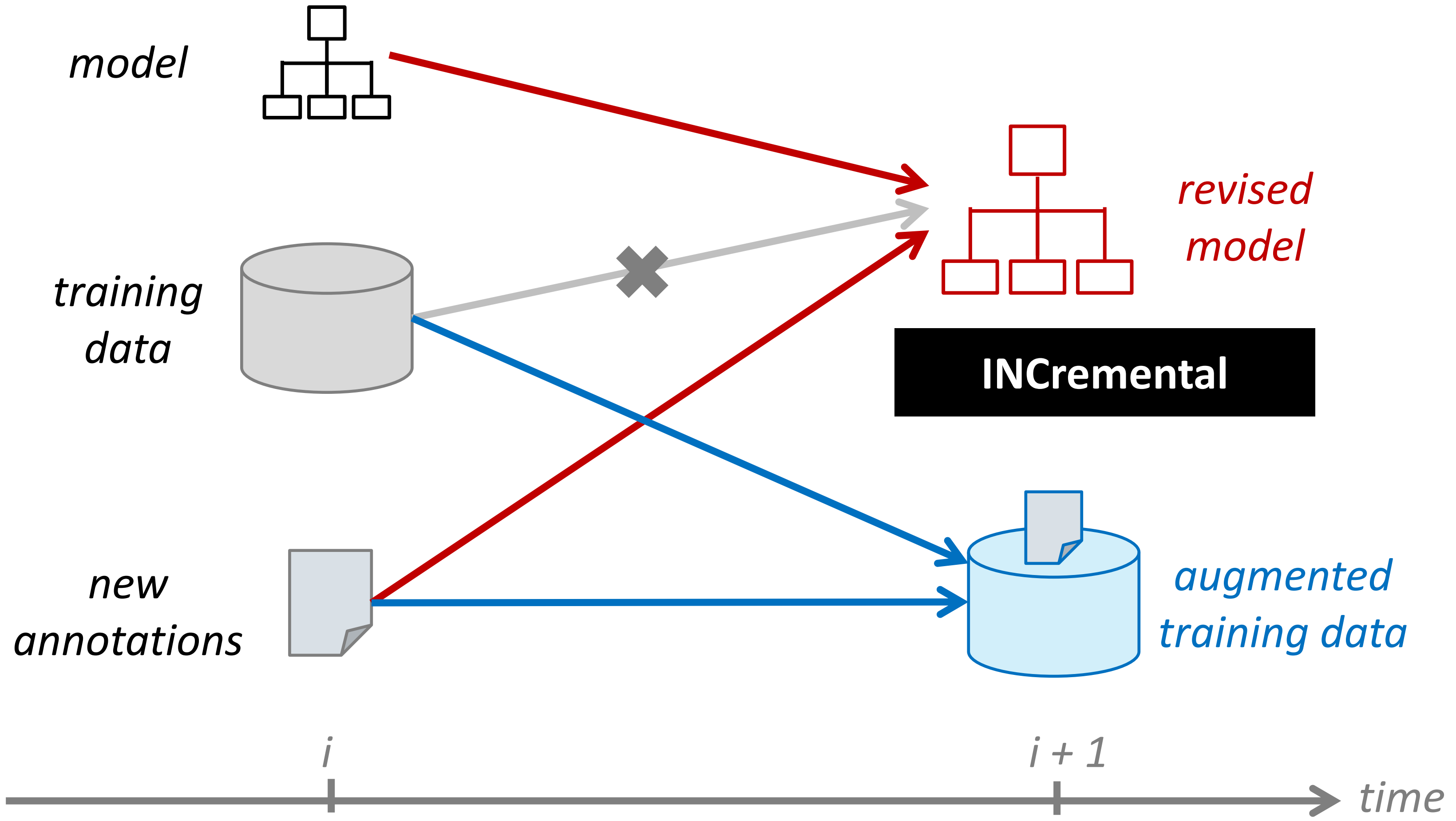
Iterative Model Training



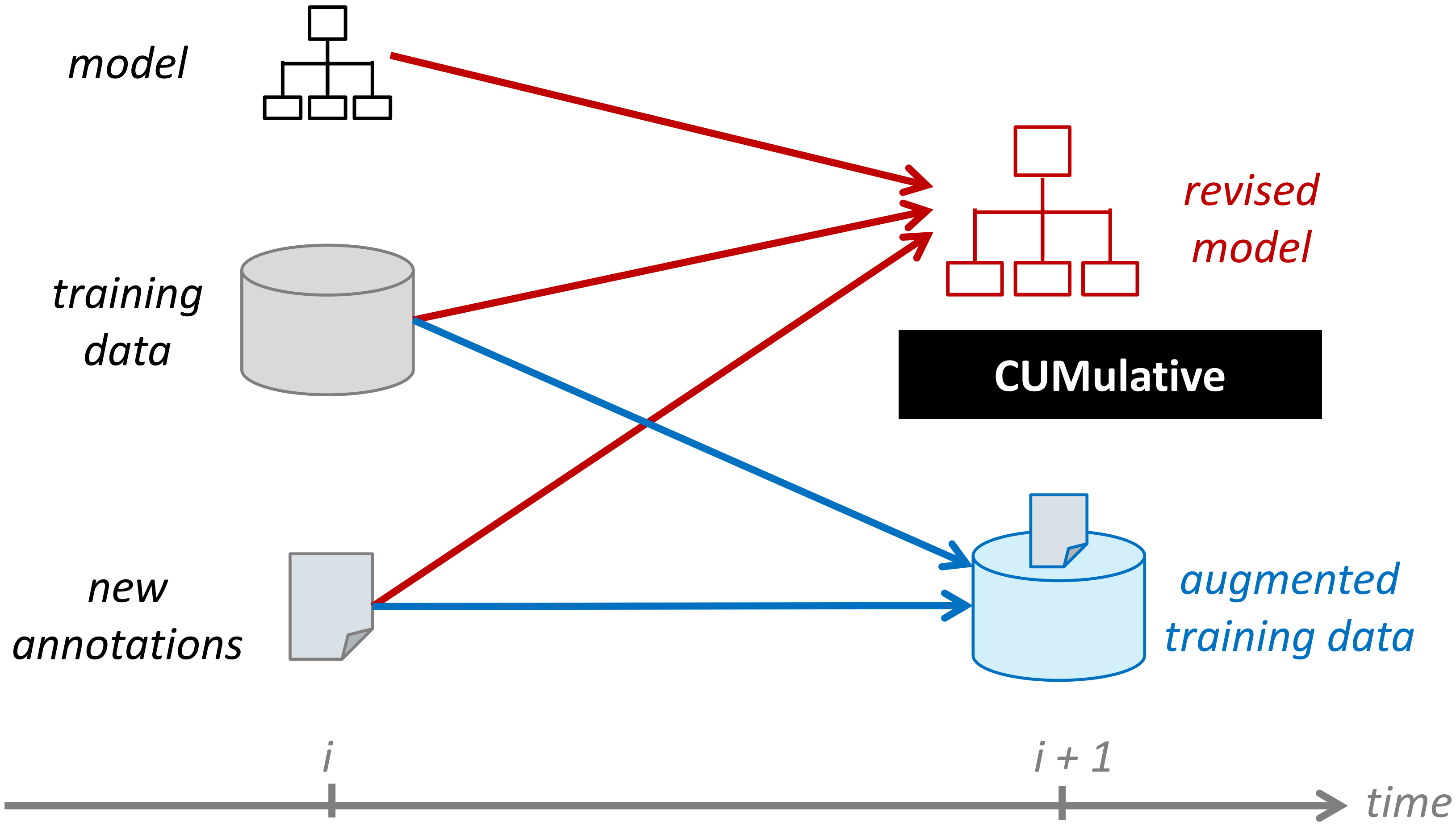
Iterative Model Training



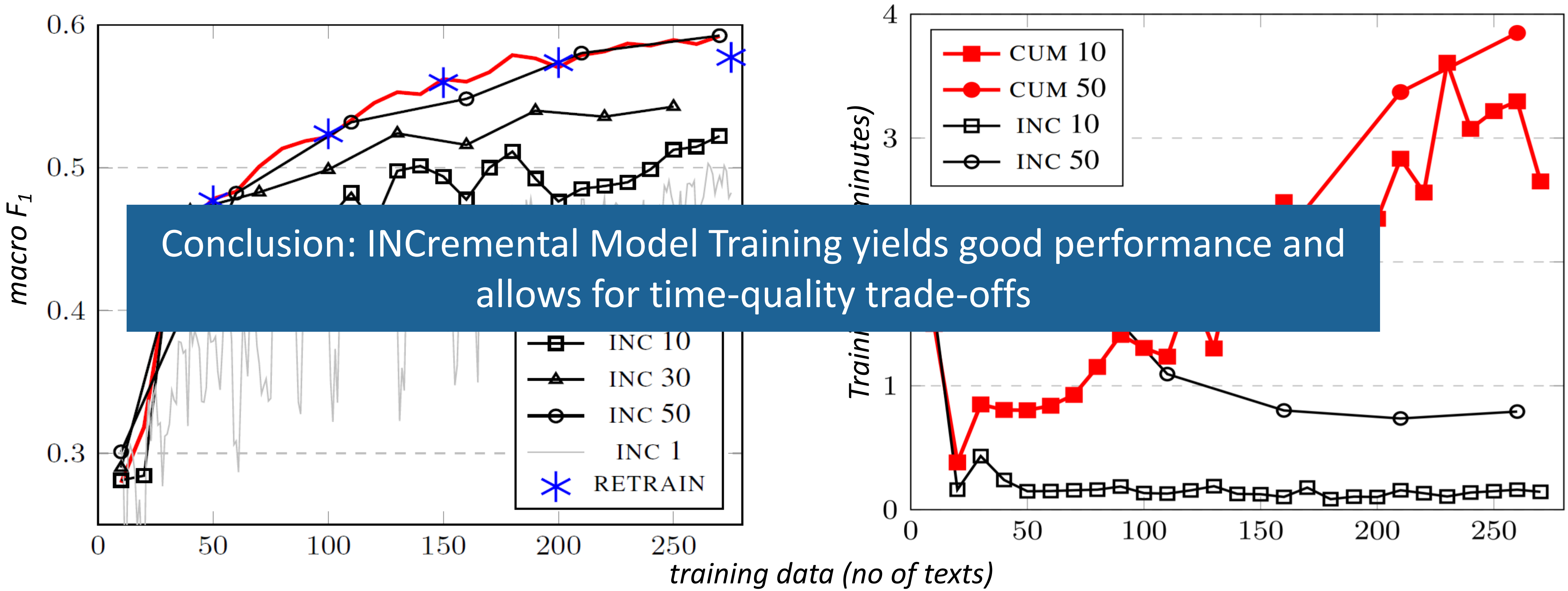
Iterative Model Training



Iterative Model Training



Model Performance



Conclusion: INCremental Model Training yields good performance and allows for time-quality trade-offs

Automatic Feedback on Diagnostic Reasoning

Detecting Diagnostic Reasoning Steps

✓ 1) **Corpus Creation**

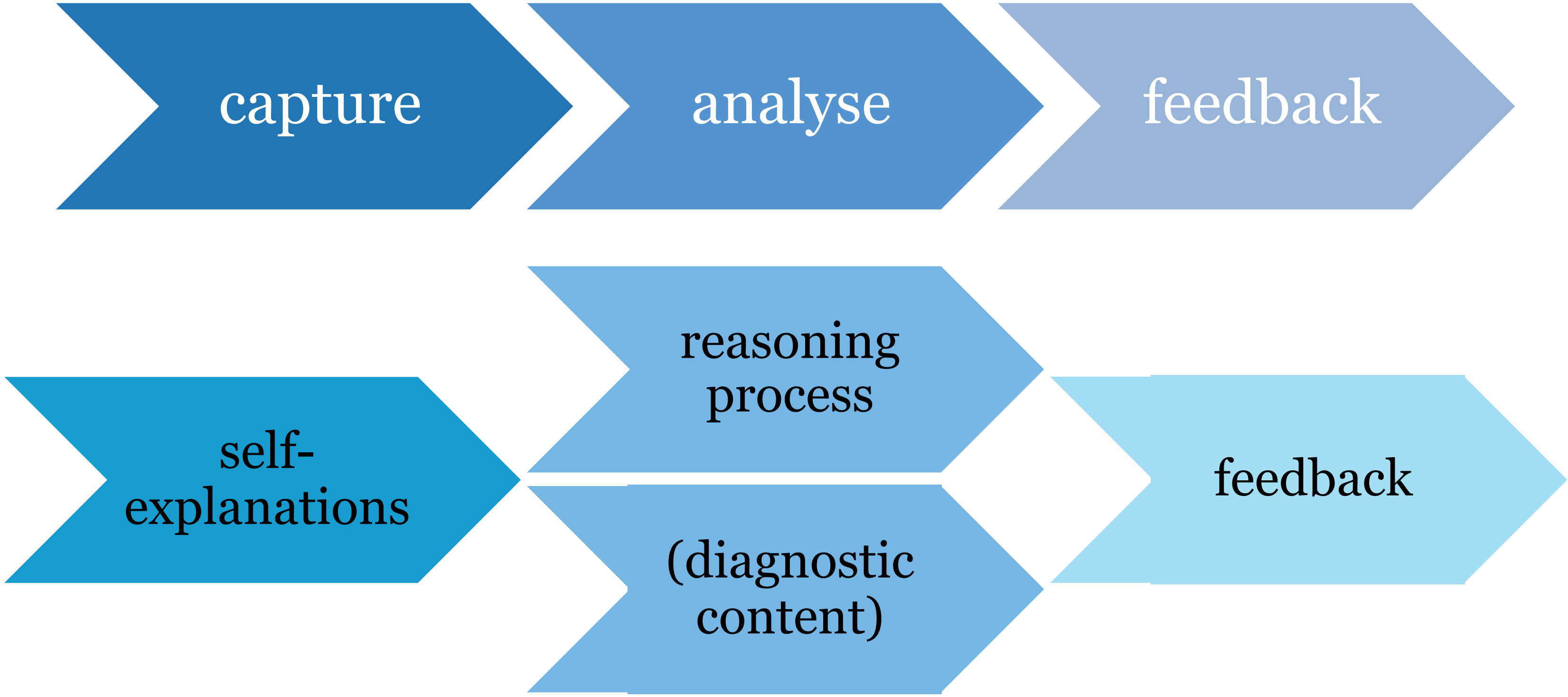
✓ 2) **Automatic Detection**

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
Evidence Generation
Evidence Evaluation
Drawing Conclusions

Adaptive Feedback

Diagnostic Reasoning



capture

analyse

feedback

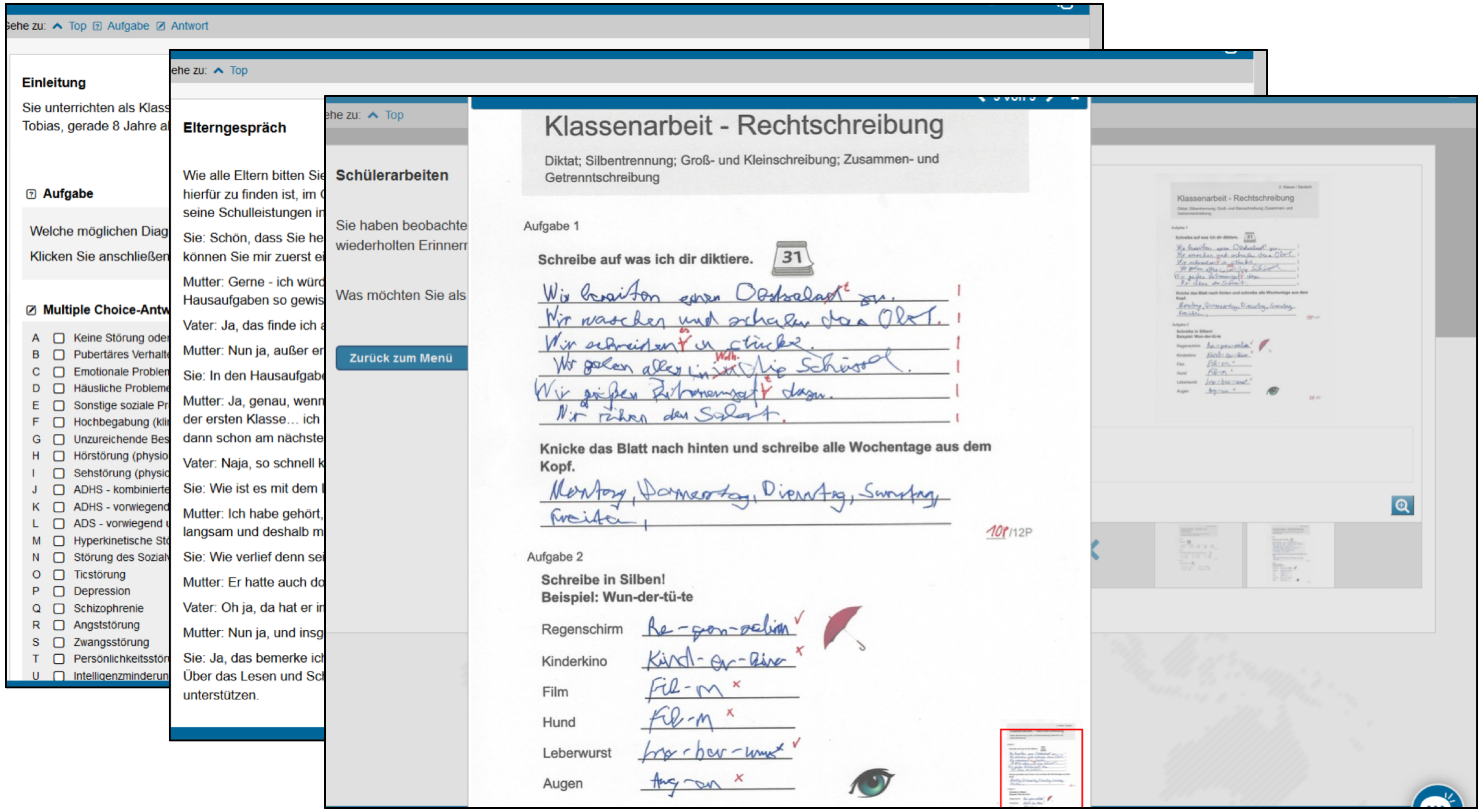
self-explanations

reasoning process

(diagnostic content)

feedback

eLearning Platform CASUS®



Gehe zu: Top Aufgabe Antwort

Einleitung

Sie unterrichten als Klass
Tobias, gerade 8 Jahre al

Aufgabe

Welche möglichen Diag
Klicken Sie anschließen

Multiple Choice-Antw

- A □ Keine Störung oder
- B □ Pubertäres Verhalte
- C □ Emotionale Problem
- D □ Häusliche Probleme
- E □ Sonstige soziale Pr
- F □ Hochbegabung (klin
- G □ Unzureichende Bes
- H □ Hörstörung (physio
- I □ Sehstörung (physio
- J □ ADHS - kombinierte
- K □ ADHS - vorwiegend
- L □ ADS - vorwiegend u
- M □ Hyperkinetische Stö
- N □ Störung des Sozialv
- O □ Ticstörung
- P □ Depression
- Q □ Schizophrenie
- R □ Angststörung
- S □ Zwangsstörung
- T □ Persönlichkeitsstör
- U □ Intelligenzminderun

ehe zu: Top

Elterngespräch

Wie alle Eltern bitten Sie
hierfür zu finden ist, im C
seine Schulleistungen in

Sie: Schön, dass Sie he
können Sie mir zuerst ei

Mutter: Gerne - ich würd
Hausaufgaben so gewis

Vater: Ja, das finde ich a

Mutter: Nun ja, außer er

Sie: In den Hausaufgabe

Mutter: Ja, genau, wenn
der ersten Klasse… ich
dann schon am nächste

Vater: Naja, so schnell k

Sie: Wie ist es mit dem I

Mutter: Ich habe gehört,
langsam und deshalb m

Sie: Wie verlief denn sei

Mutter: Er hatte auch do

Vater: Oh ja, da hat er in

Mutter: Nun ja, und insg

Sie: Ja, das bemerke ich
Über das Lesen und Sc
unterstützen.

ehe zu: Top

Schülerarbeiten

Sie haben beobachte
wiederholten Erinnern

Was möchten Sie als

Zurück zum Menü

Klassenarbeit - Rechtschreibung

Diktat; Silbentrennung; Groß- und Kleinschreibung; Zusammen- und Getrenntschreibung

Aufgabe 1

Schreibe auf was ich dir diktiere.

Wir braiten einen Obssalat zu.
Wir waschen und schalen das Obst.
Wir schreiden es in stücke.
Wir geben alles in die Schüssel.
Wir gießen Zitronensaft dazu.
Wir rühren den Salat.

Knicke das Blatt nach hinten und schreibe alle Wochentage aus dem Kopf.

Montag, Donnerstag, Dienstag, Sonntag, Freitag,

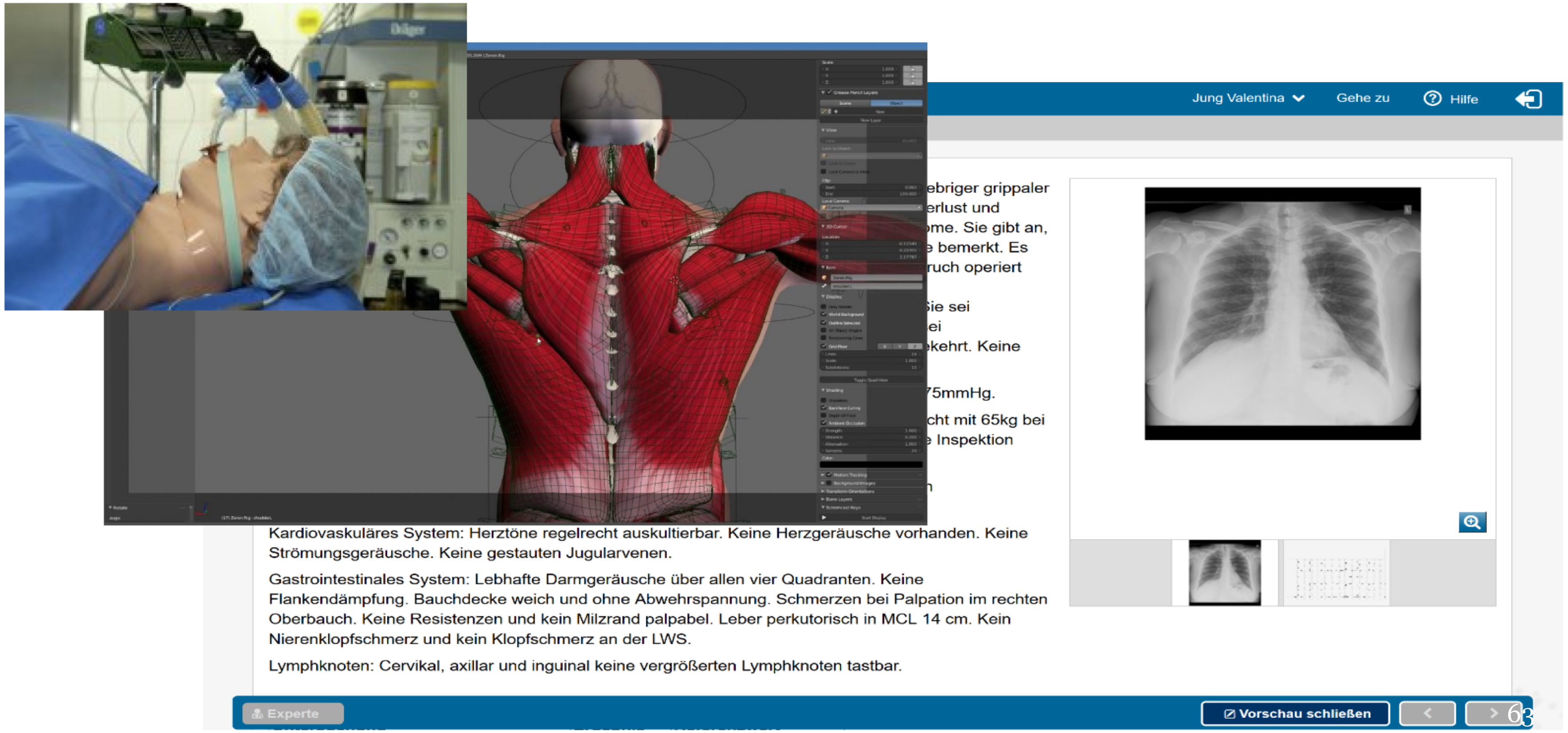
10/12P

Aufgabe 2

Schreibe in Silben!
Beispiel: Wun-der-tü-te

Regenschirm Re-gen-schim ✓
Kinderkino Kind-er-kino ✗
Film Fil-m ✗
Hund Fil-m ✗
Leberwurst Le-ber-wurst ✓
Augen Aug-en ✗

eLearning Platform CASUS®



ebriger grippaler
erlust und
ome. Sie gibt an,
e bemerkt. Es
ruch operiert

Sie sei
ei
kehrt. Keine

75mmHg.

cht mit 65kg bei
e Inspektion

n

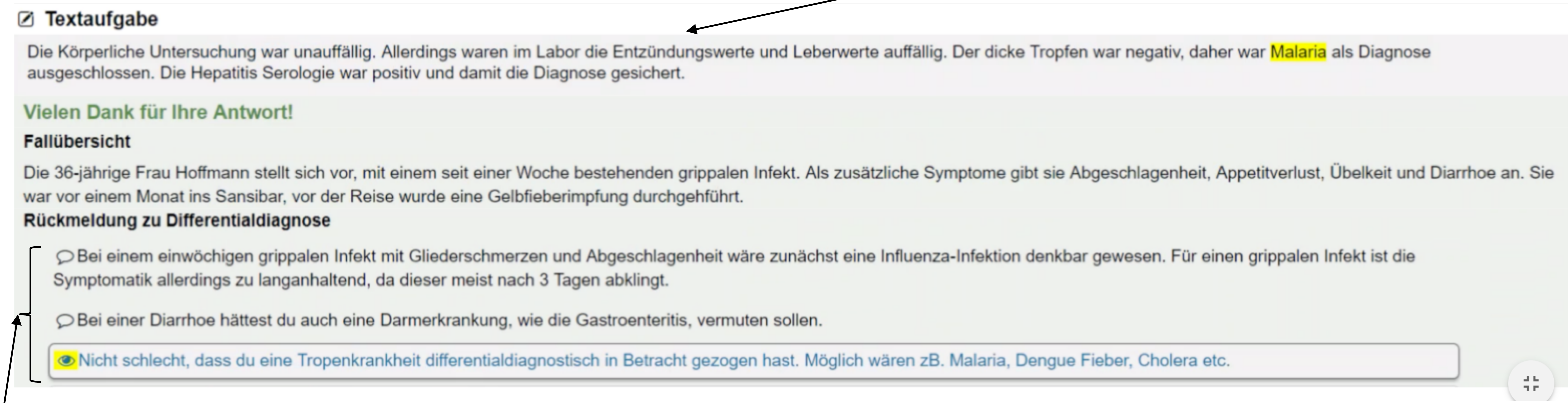
Kardiovaskuläres System: Herztöne regelrecht auskultierbar. Keine Herzgeräusche vorhanden. Keine Strömungsgeräusche. Keine gestauten Jugularvenen.

Gastrointestinales System: Lebhafte Darmgeräusche über allen vier Quadranten. Keine Flankendämpfung. Bauchdecke weich und ohne Abwehrspannung. Schmerzen bei Palpation im rechten Oberbauch. Keine Resistenzen und kein Milzrand palpabel. Leber perkutorisch in MCL 14 cm. Kein Nierenklopfschmerz und kein Klopfschmerz an der LWS.

Lymphknoten: Cervikal, axillar und inguinal keine vergrößerten Lymphknoten tastbar.

Automatic Feedback

student's self-explanation



Textaufgabe

Die Körperliche Untersuchung war unauffällig. Allerdings waren im Labor die Entzündungswerte und Leberwerte auffällig. Der dicke Tropfen war negativ, daher war Malaria als Diagnose ausgeschlossen. Die Hepatitis Serologie war positiv und damit die Diagnose gesichert.

Vielen Dank für Ihre Antwort!

Fallübersicht

Die 36-jährige Frau Hoffmann stellt sich vor, mit einem seit einer Woche bestehenden grippalen Infekt. Als zusätzliche Symptome gibt sie Abgeschlagenheit, Appetitverlust, Übelkeit und Diarrhoe an. Sie war vor einem Monat ins Sansibar, vor der Reise wurde eine Gelbfieberimpfung durchgehführt.

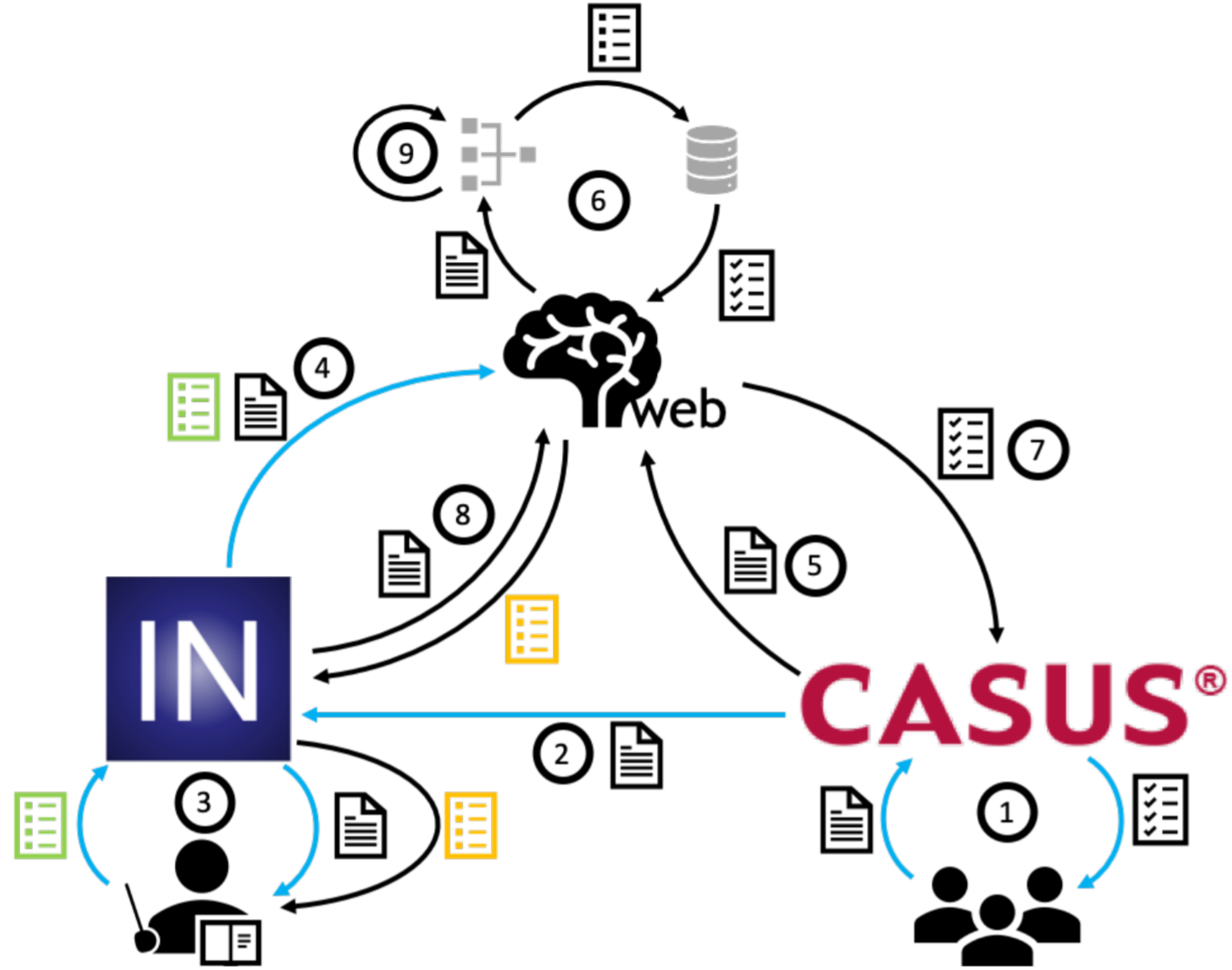
Rückmeldung zu Differentialdiagnose

Bei einem einwöchigen grippalen Infekt mit Gliederschmerzen und Abgeschlagenheit wäre zunächst eine Influenza-Infektion denkbar gewesen. Für einen grippalen Infekt ist die Symptomatik allerdings zu langanhaltend, da dieser meist nach 3 Tagen abklingt.

Bei einer Diarrhoe hättest du auch eine Darmerkrankung, wie die Gastroenteritis, vermuten sollen.

Nicht schlecht, dass du eine Tropenkrankheit differentialdiagnostisch in Betracht gezogen hast. Möglich wären zB. Malaria, Dengue Fieber, Cholera etc.

automatic adaptive feedback


9
6
4
web
8
7
5
IN
CASUS®
2
3
1

Wrapping Up

Explaining how to improve Diagnostic Reasoning

- Collect Self-Explanations
- Annotate Diagnostic Reasoning Steps
  - Train model for annotation suggestions → ease and speed-up
- Train Model for Detecting Reasoning Steps
- Use Model for Automatic Feedback

My publications

http://www.famulus-project.de/

https://inception-project.github.io/

For more questions, contact me:
clauschulz1812@gmail.com