Automatic Scoring of a German WEIT

Automatic Scoring

Input

Multiple-choice answers

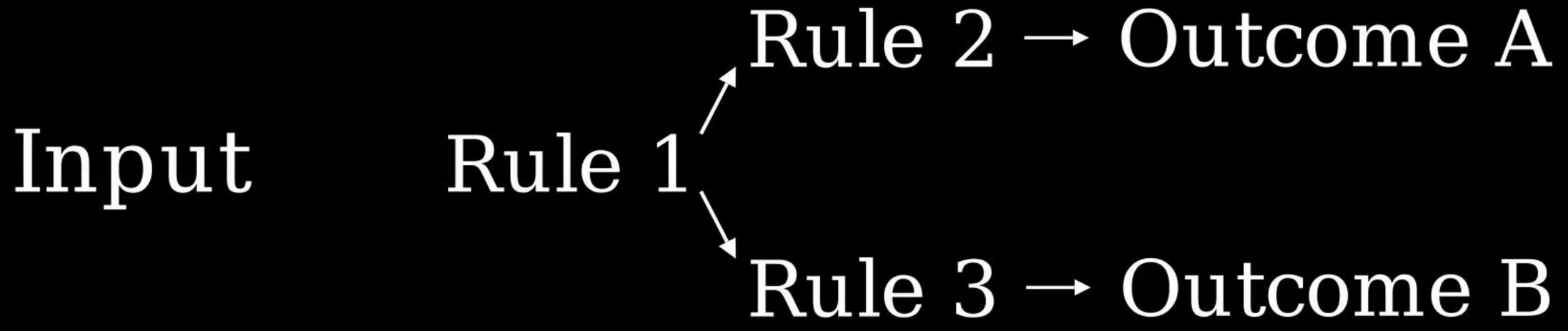
One-word answers

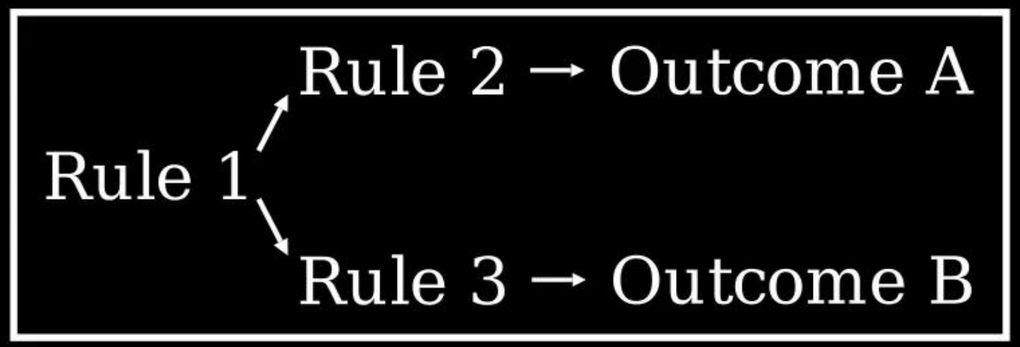
Short text answers

True/False

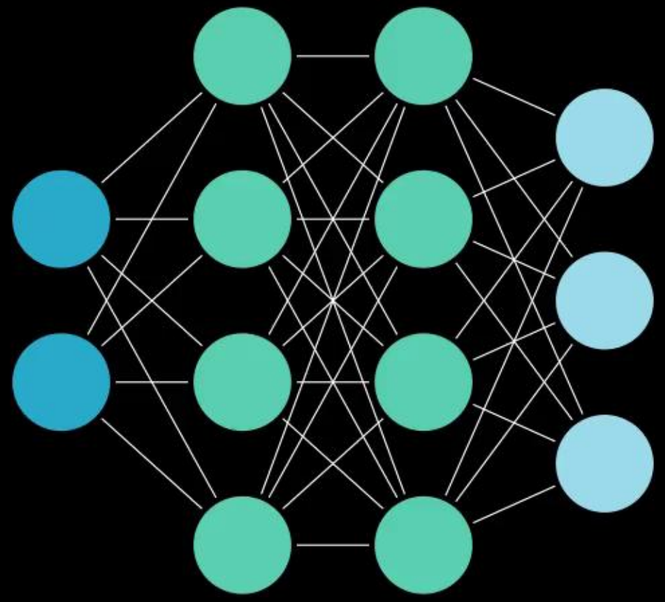
Essays

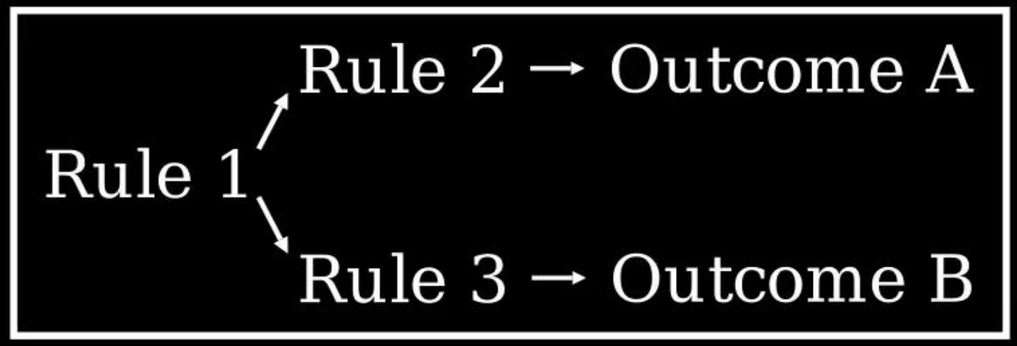
Input


Input
Rule 1
Rule 2 → Outcome A
Rule 3 → Outcome B

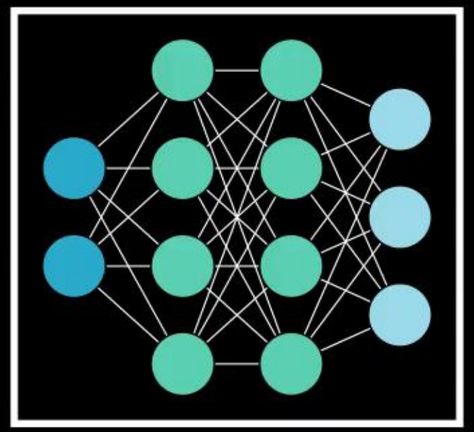


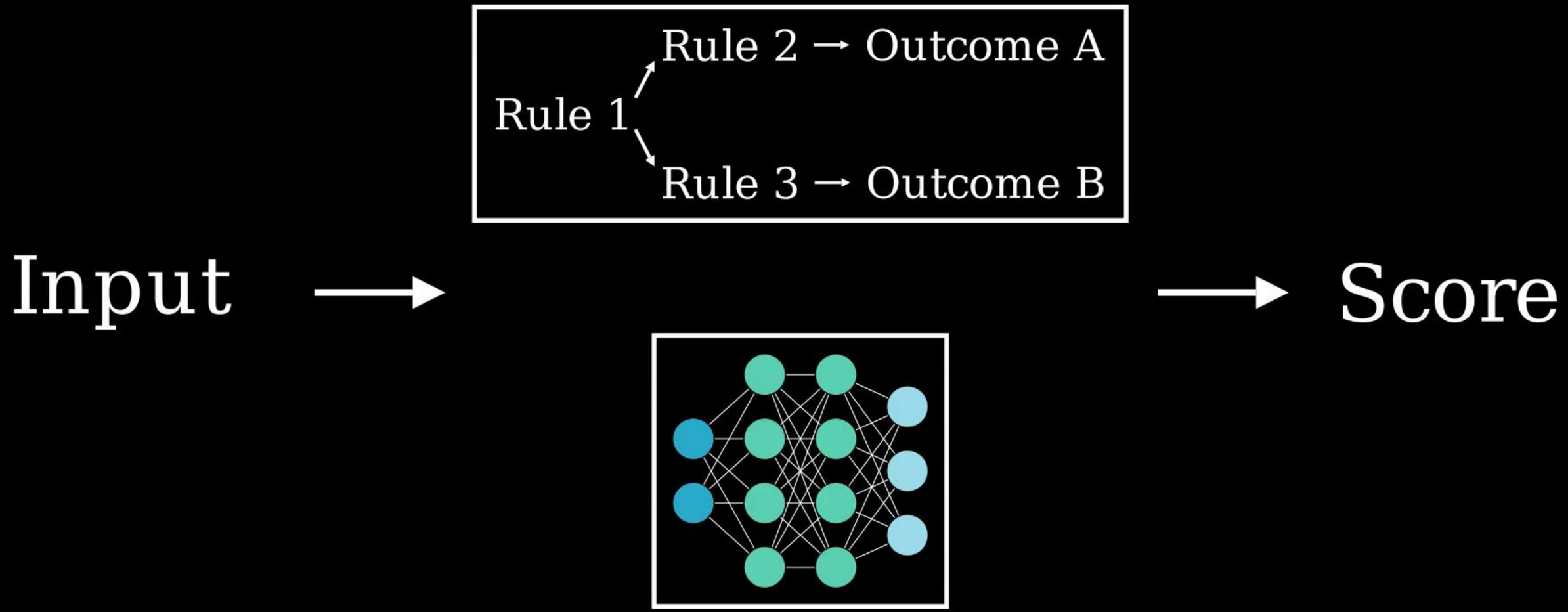
Input





Input




Rule 2 → Outcome A
Rule 1
Rule 3 → Outcome B
Input
Score

Automatic Scoring of a German WEIT

WEIT

Written

Elicited

Imitation

Test

Elicited Imitation Test

Elicited Imitation Test

Stimulus → Imitation

Elicited Imitation Test

Stimulus → Imitation

Elicited Imitation Test

Stimulus	→	Imitation
		
		

The WEIT Scoring Rubric

Original: Bei einem Praktikum lernt man viel

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einem Praktikum lernt man viel

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einem Praktikum lernt man viel

Original: Bei einem Praktikum lernt man viel

Imitation:

	Gold Score
Bei einem Praktikum lernt man viel	4

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einem Praktikum lernt man zu viel

	Gold Score
Bei einem Praktikum lernt man viel	4

Original: Bei einem Praktikum lernt man viel

Imitation:

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einer Praktikum lernt man viel

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3

Original: Bei einem Praktikum lernt man viel

Imitation:

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3
Bei einer Praktikum lernt man viel	2

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einen Praktikum lern man viel

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3
Bei einer Praktikum lernt man viel	2

Original: Bei einem Praktikum lernt man viel

Imitation:

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3
Bei einer Praktikum lernt man viel	2
Bei einen Praktikum lern man viel	1

Original: Bei einem Praktikum lernt man viel

Imitation: praktikum auf viel

	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3
Bei einer Praktikum lernt man viel	2
Bei einen Praktikum lern man viel	1

Original: Bei einem Praktikum lernt man viel

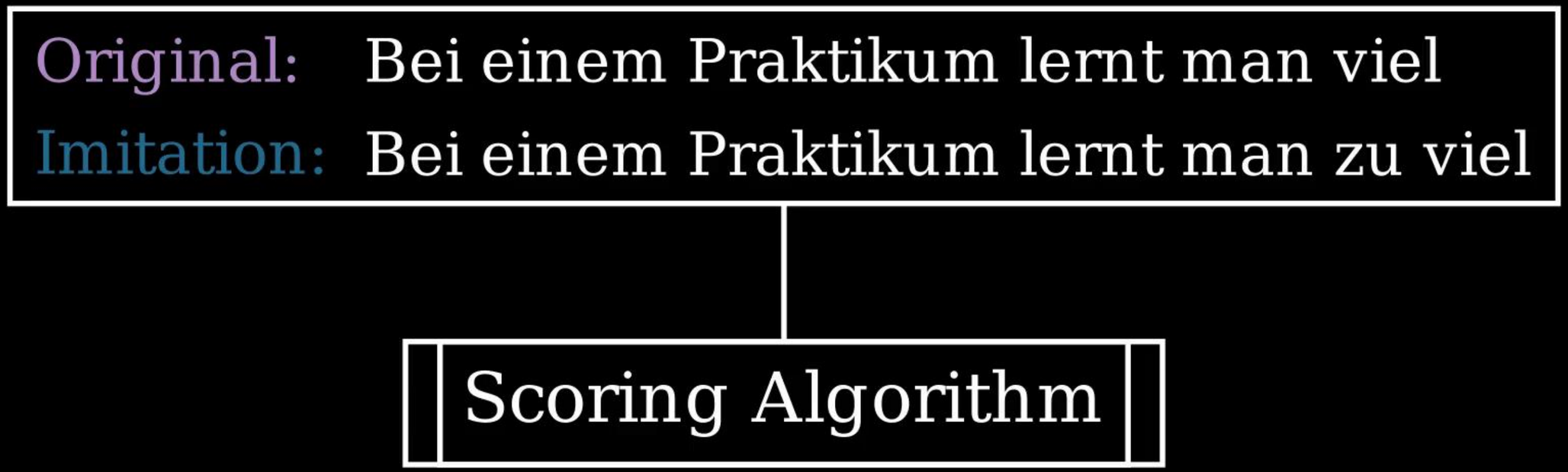
Imitation:

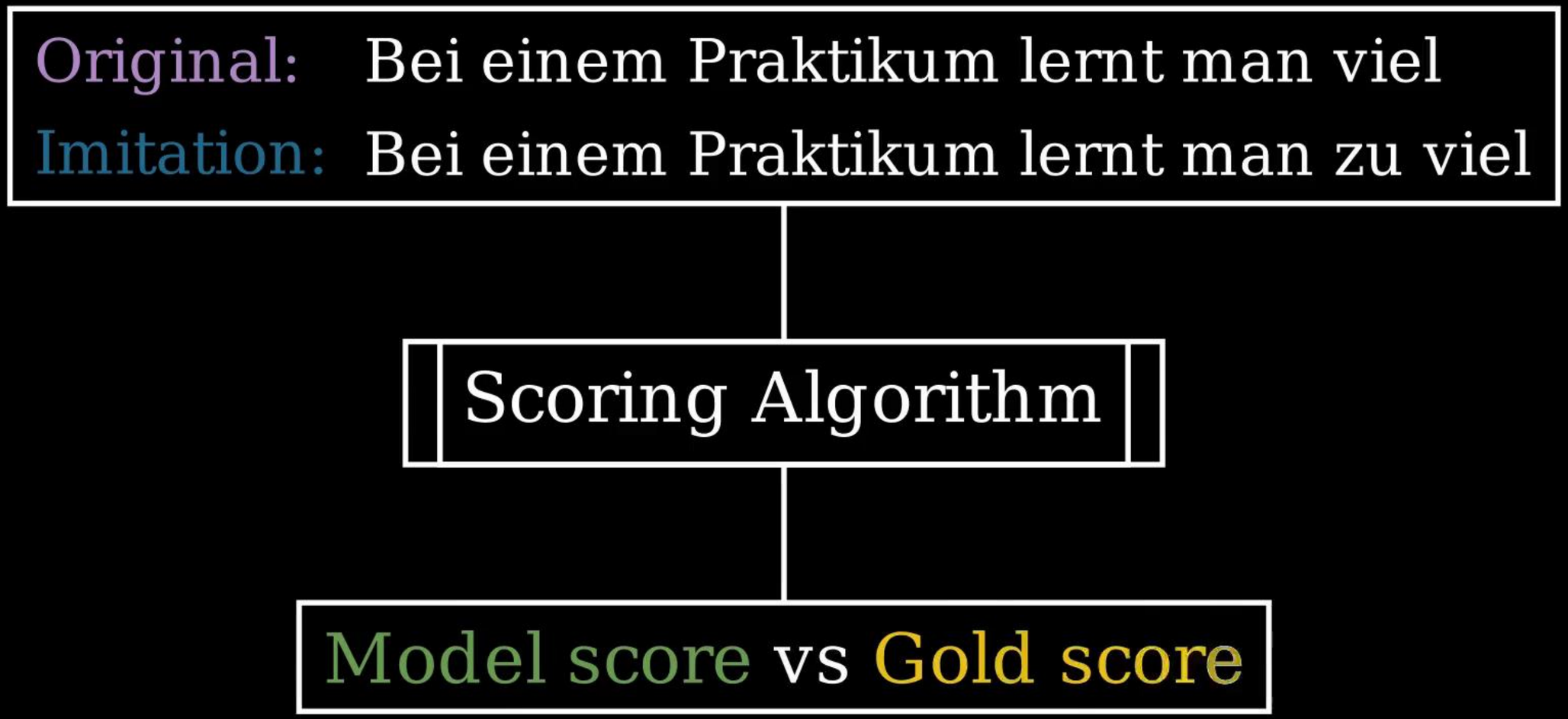
	Gold Score
Bei einem Praktikum lernt man viel	4
Bei einem Praktikum lernt man zu viel	3
Bei einer Praktikum lernt man viel	2
Bei einen Praktikum lern man viel	1
praktikum auf viel	0

Original: Bei einem Praktikum lernt man viel

Imitation: Bei einem Praktikum lernt man zu viel

Original:	Bei einem Praktikum lernt man viel
Imitation:	Bei einem Praktikum lernt man zu viel


Original: Bei einem Praktikum lernt man viel
Imitation: Bei einem Praktikum lernt man zu viel
Scoring Algorithm


Original: Bei einem Praktikum lernt man viel
Imitation: Bei einem Praktikum lernt man zu viel
Scoring Algorithm
Model score vs Gold score

Deep Learning Model

Data Split

Data Split

$$20\,\texttt{Originals} \cdot 195\,\texttt{Subjects} = 3900\,\texttt{Imitations}$$

Score	Total
0	1232 (32%)
1	843 (22%)
2	665 (17%)
3	388 (10%)
4	772 (20%)
Total	3900

Score	Total
0	1232 (32%)
1	843 (22%)
2	665 (17%)
3	388 (10%)
4	772 (20%)
Total	3900

Score	Train	Val.	Test	Total
0				1232 (32%)
1				843 (22%)
2				665 (17%)
3				388 (10%)
4				772 (20%)
Total			125 + 390	3900

Score	Train	Val.	Test	Total
0			25 + 87	1232 (32%)
1			25 + 92	843 (22%)
2			25 + 62	665 (17%)
3			25 + 78	388 (10%)
4			25 + 71	772 (20%)
Total			125 + 390	3900

Score	Train	Val.	Test	Total
0		25	25 + 87	1232 (32%)
1		25	25 + 92	843 (22%)
2		25	25 + 62	665 (17%)
3		25	25 + 78	388 (10%)
4		25	25 + 71	772 (20%)
Total		125	125 + 390	3900

Score	Train	Val.	Test	Total
0	1095	25	25 + 87	1232 (32%)
1	701	25	25 + 92	843 (22%)
2	553	25	25 + 62	665 (17%)
3	260	25	25 + 78	388 (10%)
4	651	25	25 + 71	772 (20%)
Total	3260	125	125 + 390	3900

Model Choice

Model Choice

- DistilBERT Model

Model Choice

- DistilBERT Model
- Pretrained on German cased data

Model Choice

- DistilBERT Model
- Pretrained on German cased data
- Multi-label classification head

Model Choice

- DistilBERT Model
- Pretrained on German cased data
- Multi-label classification head
- Fine tuning on WEIT Dataset

Model Choice

- DistilBERT Model
- Pretrained on German cased data
- Multi-label classification head
- Fine tuning on WEIT Dataset
- Manual hyperparameter optimization

- Learning rate: 1e-5

- Learning rate: 1e-5
- Epsilon: 1.5e-3 (Adam optimizer)

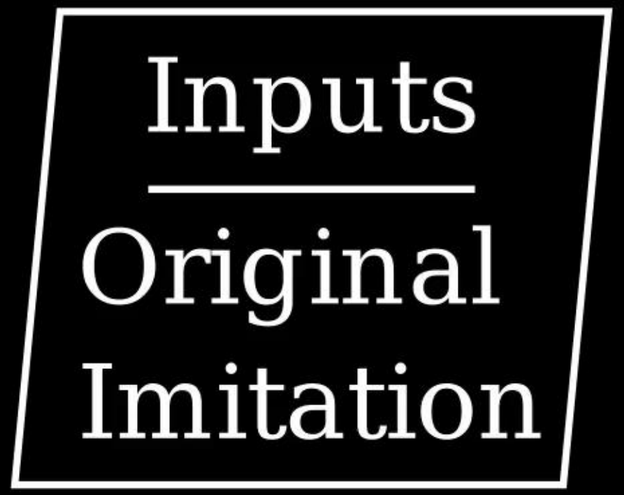
- Learning rate: 1e-5
- Epsilon: 1.5e-3 (Adam optimizer)
- Sparce categorical crossentropy loss functon

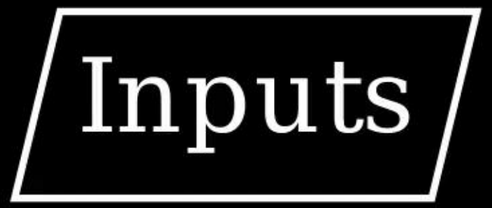
- Learning rate: 1e-5
- Epsilon: 1.5e-3 (Adam optimizer)
- Sparce categorical crossentropy loss functon
- Batch size: 16 (shuffled each iteration)

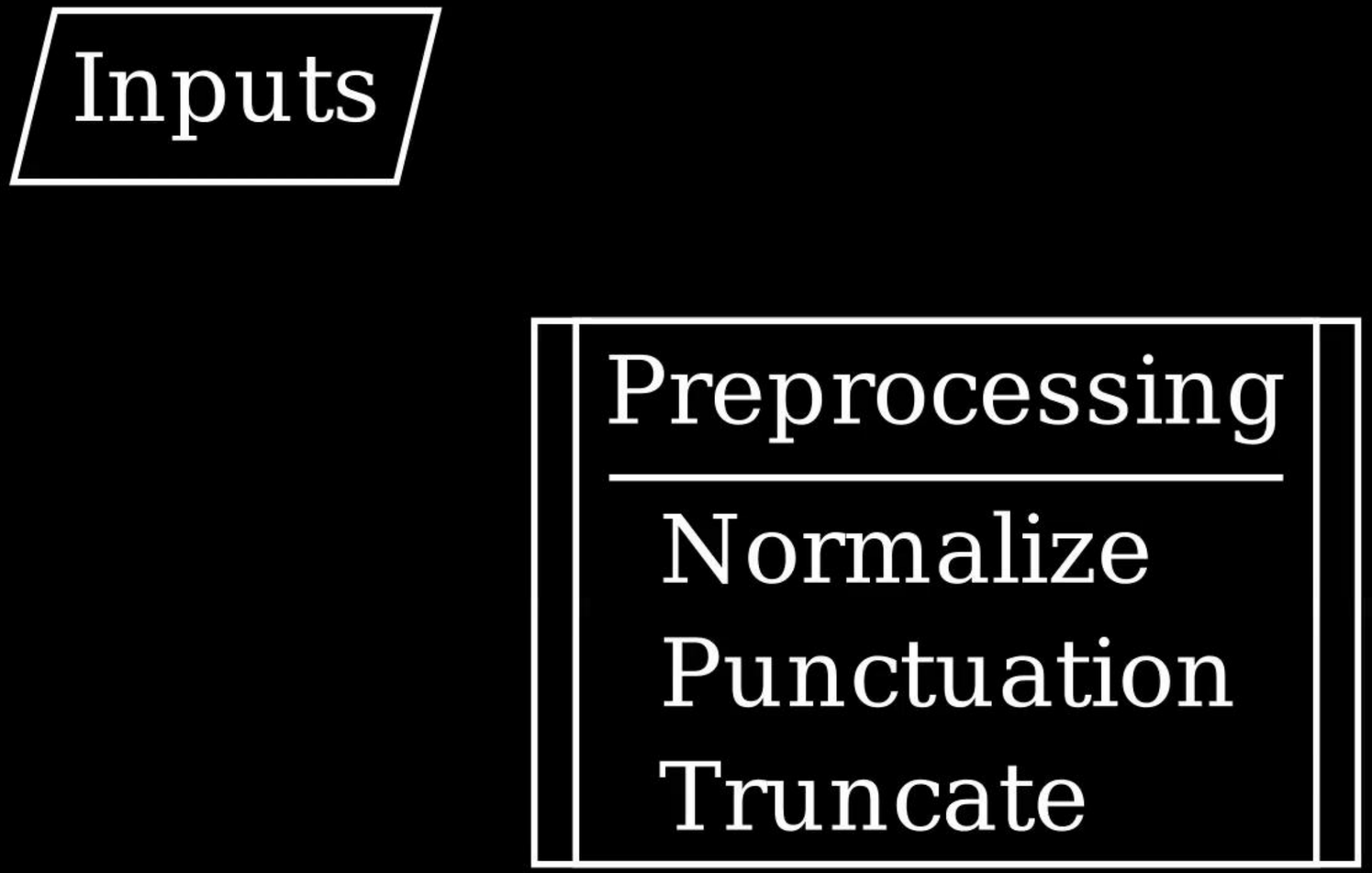
- Learning rate: 1e-5
- Epsilon: 1.5e-3 (Adam optimizer)
- Sparce categorical crossentropy loss functon
- Batch size: 16 (shuffled each iteration)
- 50 epochs with early stopping (after 5)

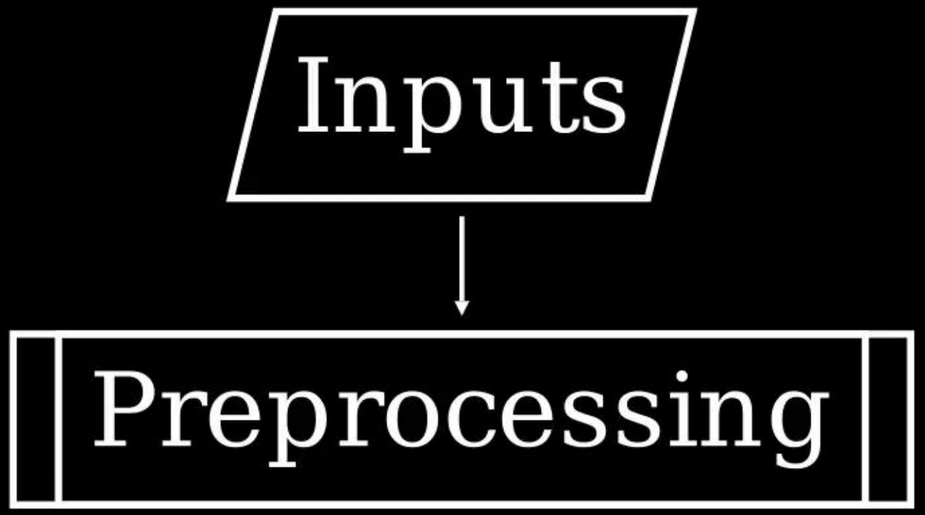
- Learning rate: 1e-5
- Epsilon: 1.5e-3 (Adam optimizer)
- Sparce categorical crossentropy loss functon
- Batch size: 16 (shuffled each iteration)
- 50 epochs with early stopping (after 5)
- Class weights due to unbalanced training set

Rule based algorithm

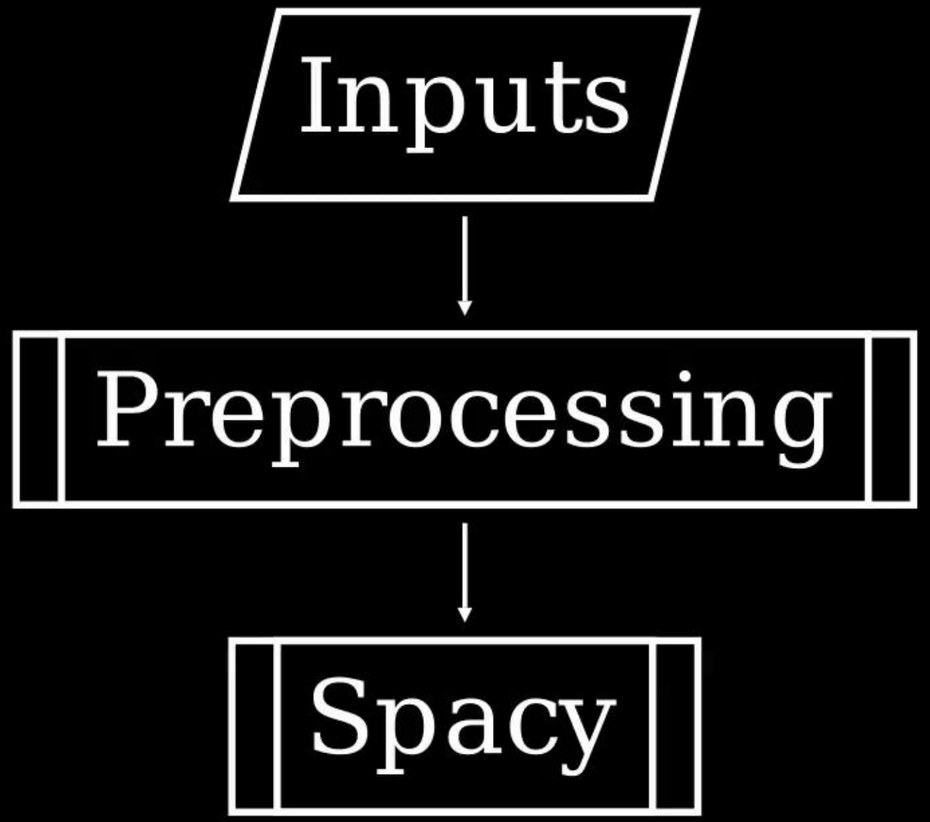

Inputs
Original
Imitation

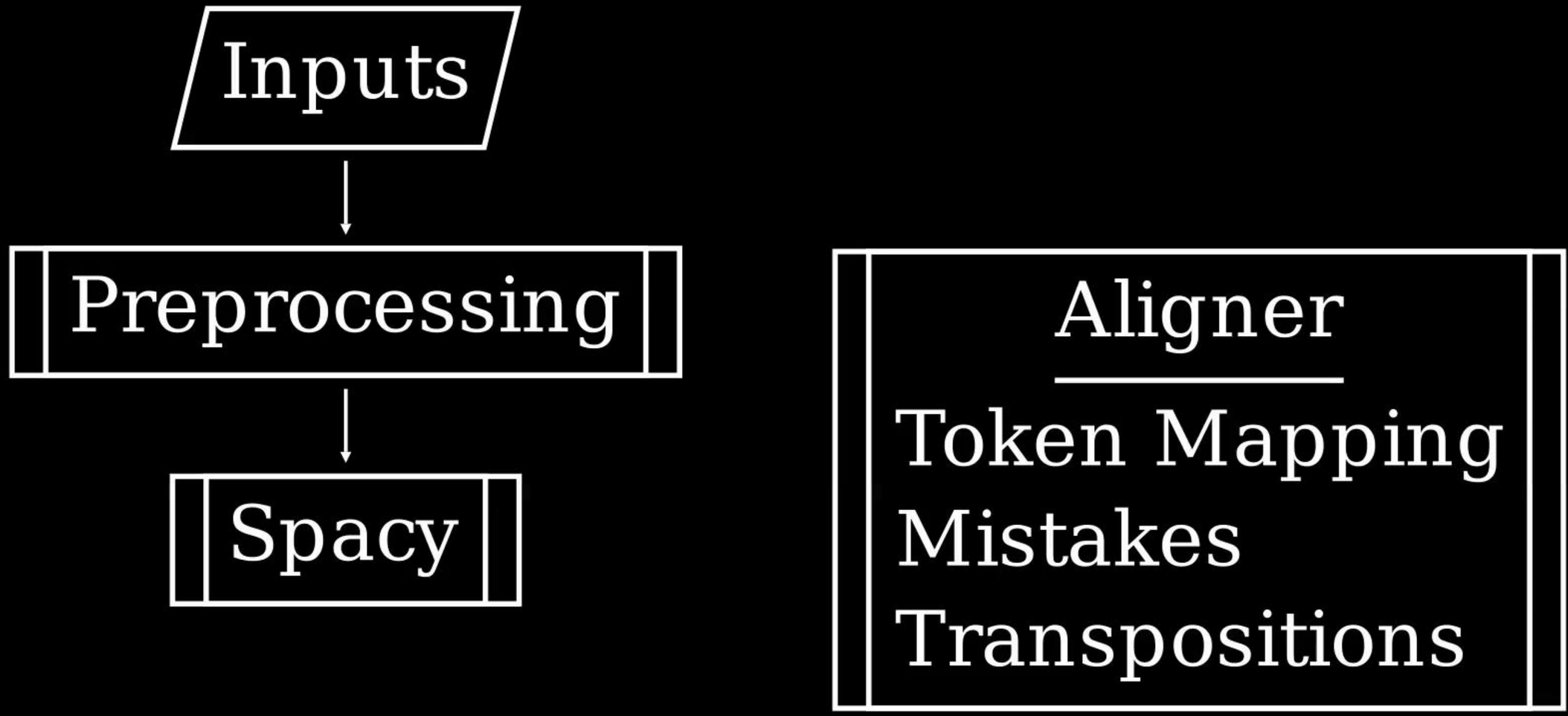

Inputs

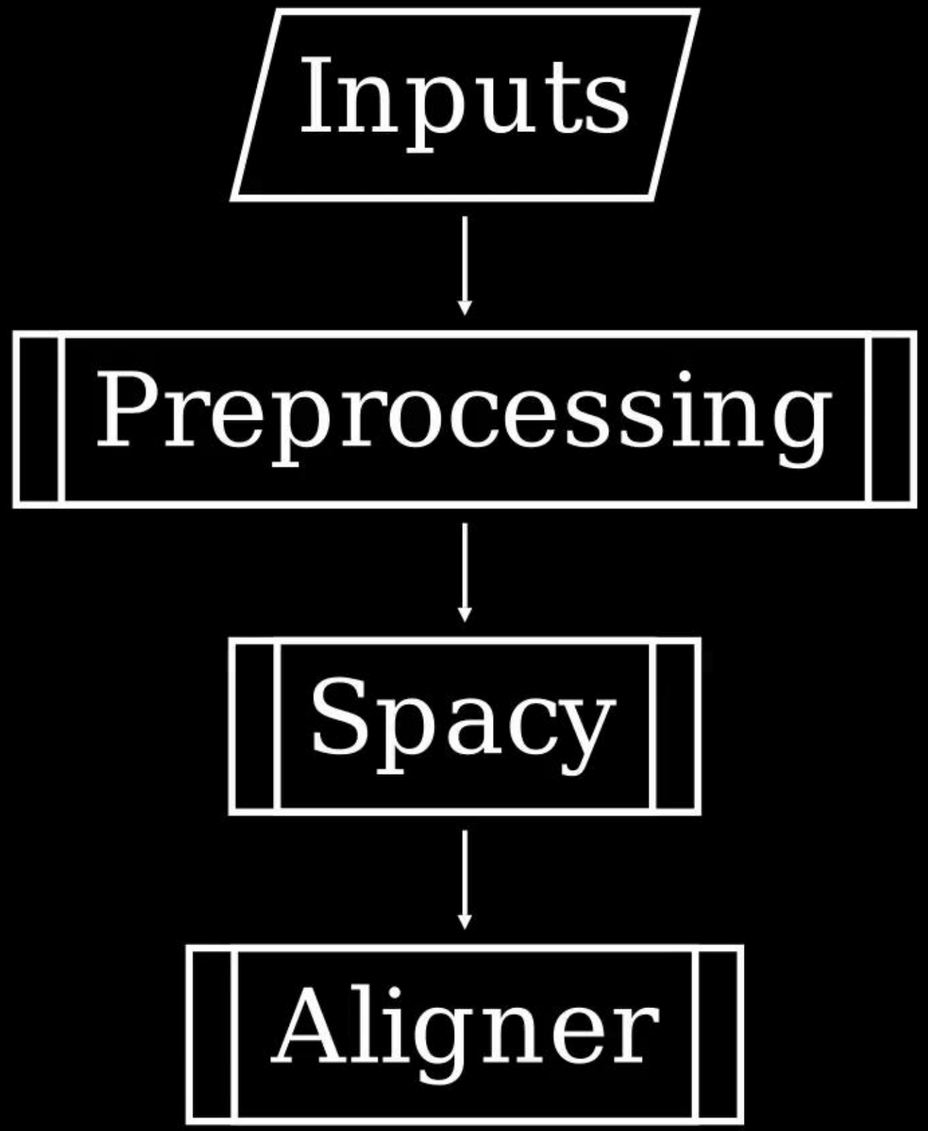

Inputs
Preprocessing
Normalize
Punctuation
Truncate

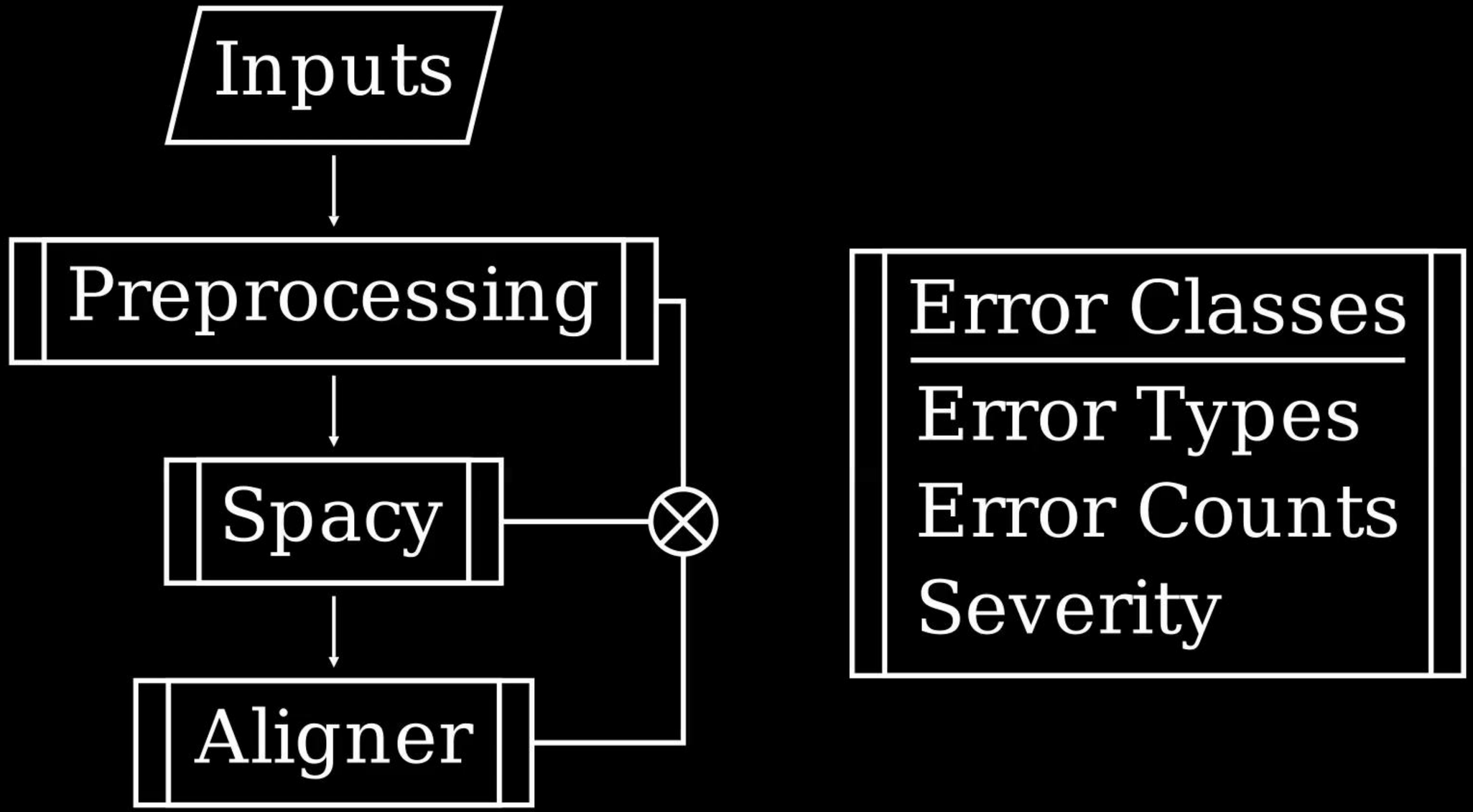

Inputs
Preprocessing

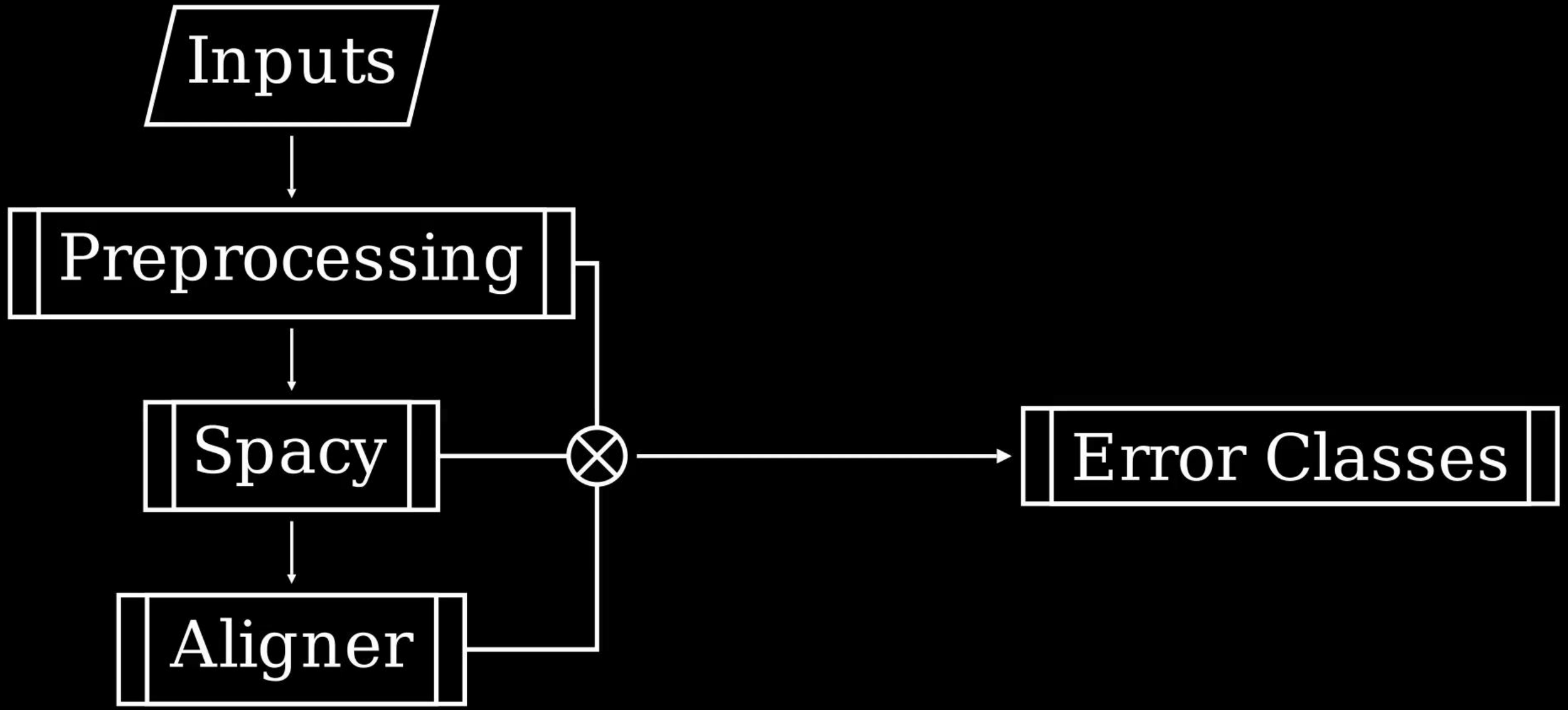

Inputs
Preprocessing
Spacy
Tokenizer
Syllables

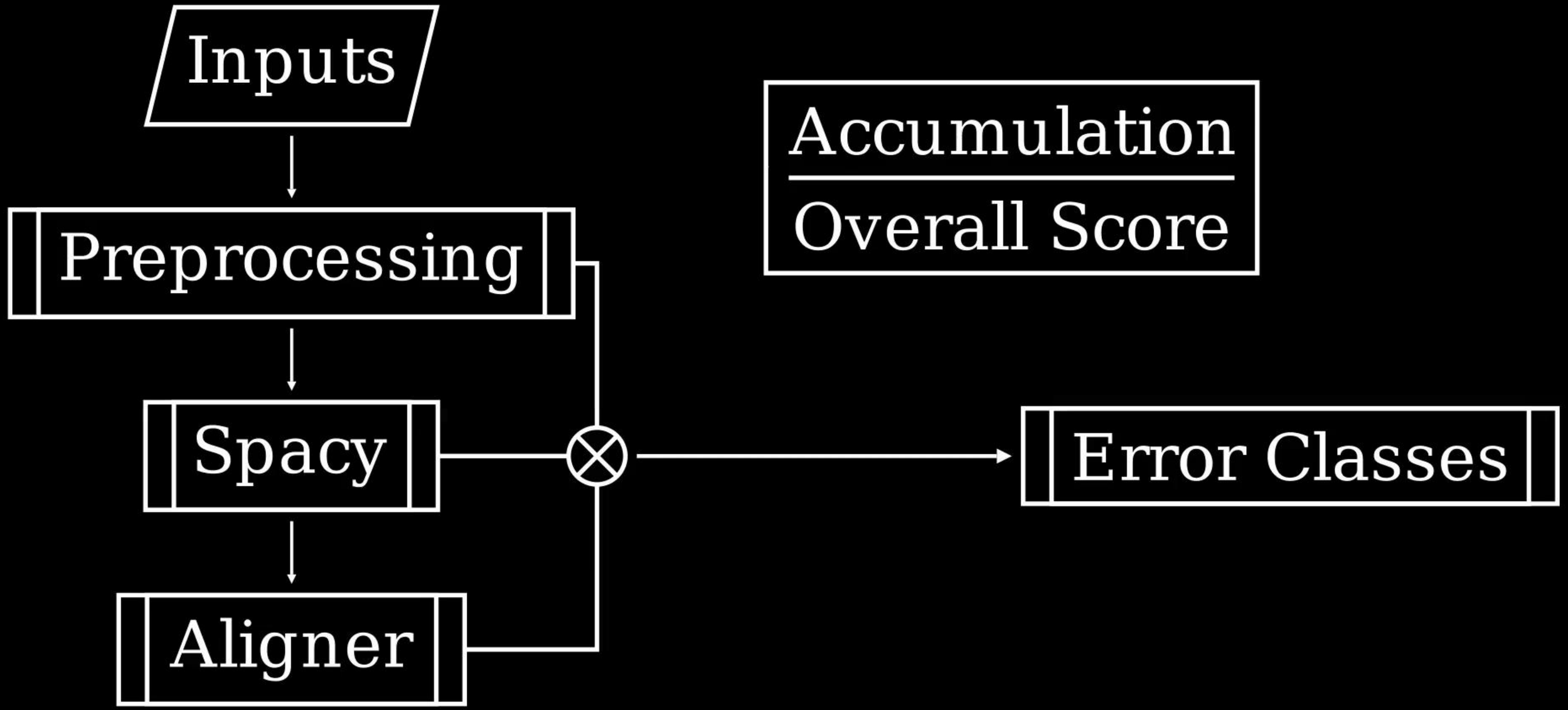

Inputs
Preprocessing
Spacy

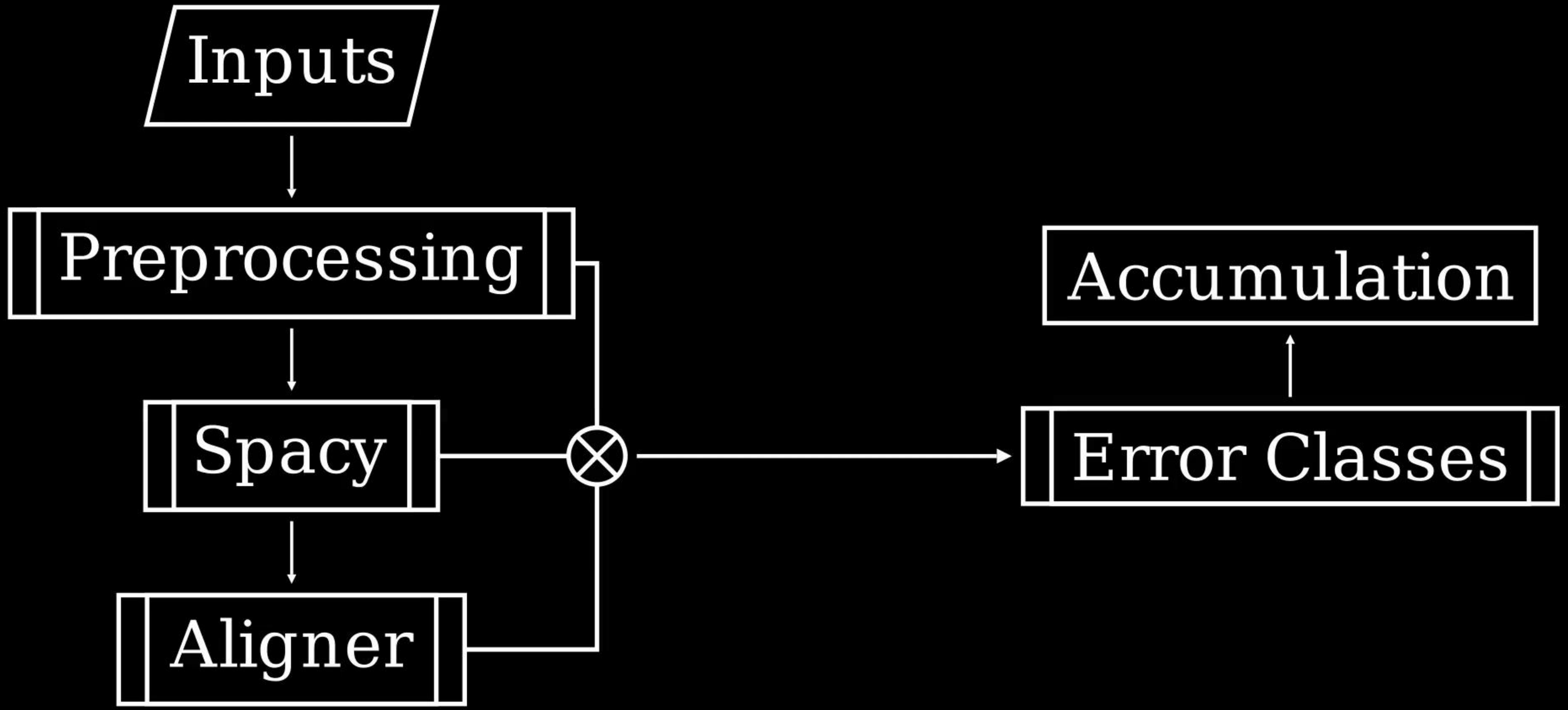

Inputs
Preprocessing
Spacy
Aligner
Token Mapping
Mistakes
Transpositions

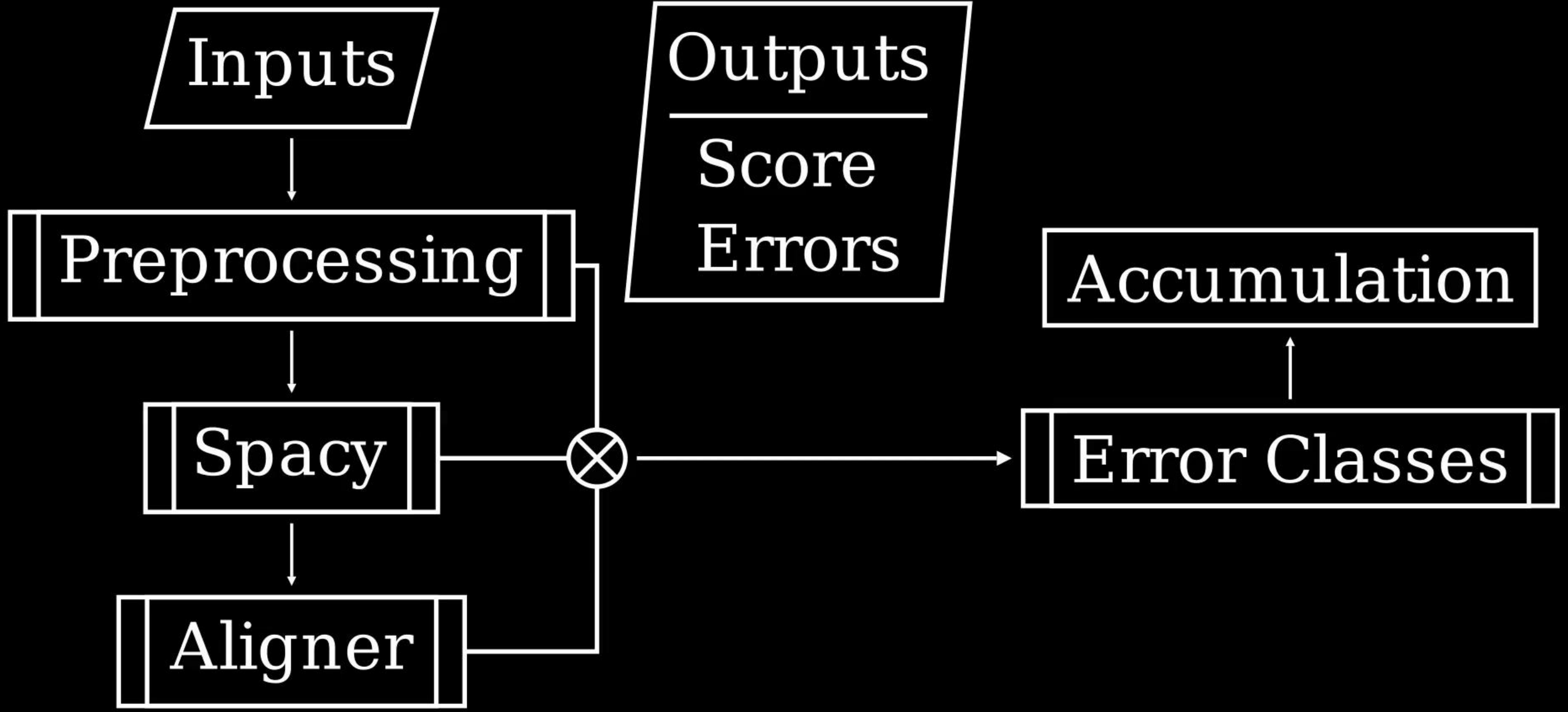

Inputs
Preprocessing
Spacy
Aligner

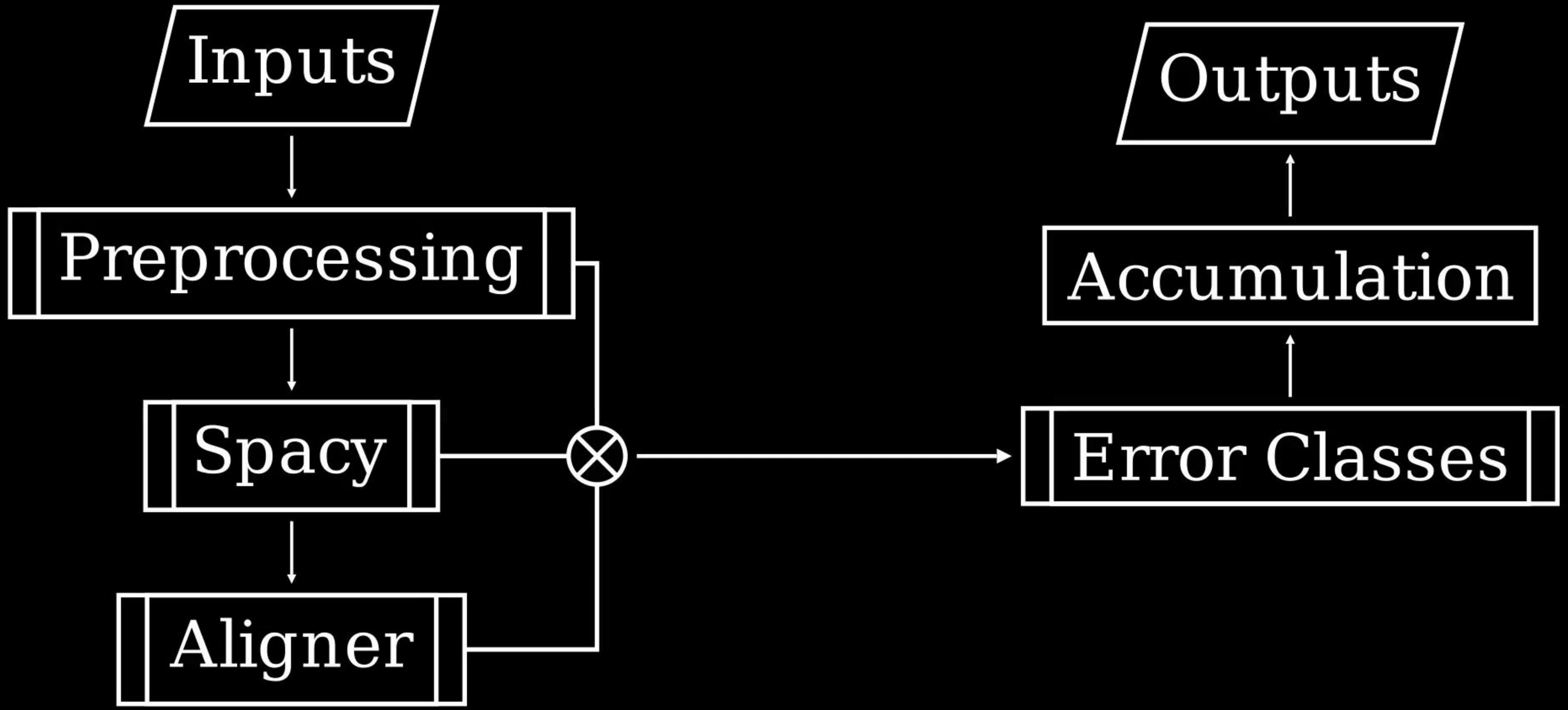

Inputs
Preprocessing
Spacy
Aligner
Error Classes
Error Types
Error Counts
Severity


Inputs
Preprocessing
Spacy
Aligner
Error Classes


Inputs
Preprocessing
Spacy
Aligner
Accumulation
Overall Score
Error Classes


Inputs
Preprocessing
Spacy
Aligner
Accumulation
Error Classes


Inputs
Preprocessing
Spacy
Aligner
Outputs
Score
Errors
Accumulation
Error Classes


Inputs
Preprocessing
Spacy
Aligner
Outputs
Accumulation
Error Classes

Evaluation

Metrics used

Metrics used

Overall metrics Per-score metrics

Metrics used

Overall metrics	Per-score metrics
	Precision
	Recall
	F1 Score

Metrics used

Overall metrics

Accuracy

Cohen's Kappa

Per-score metrics

Precision

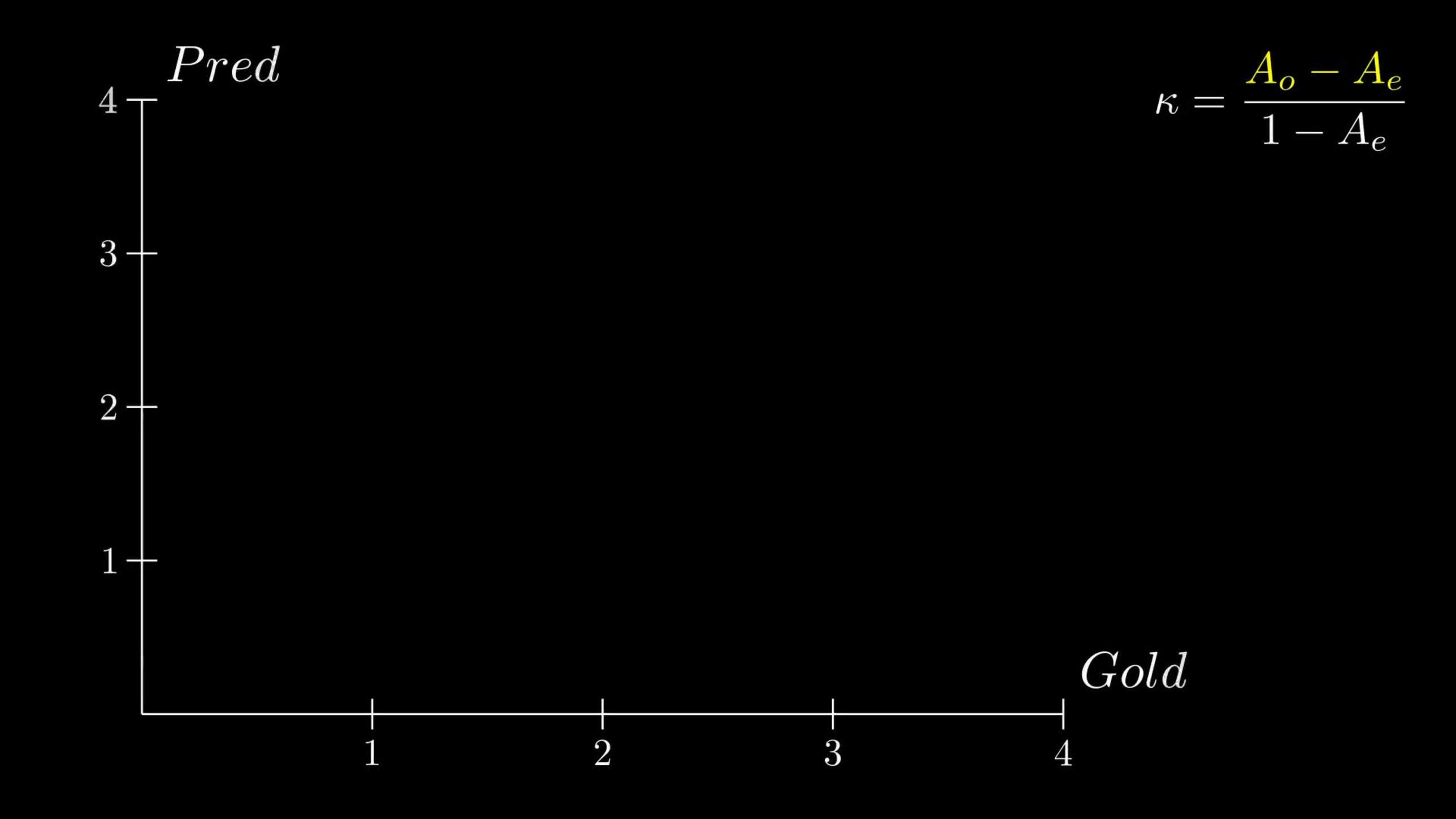
Recall

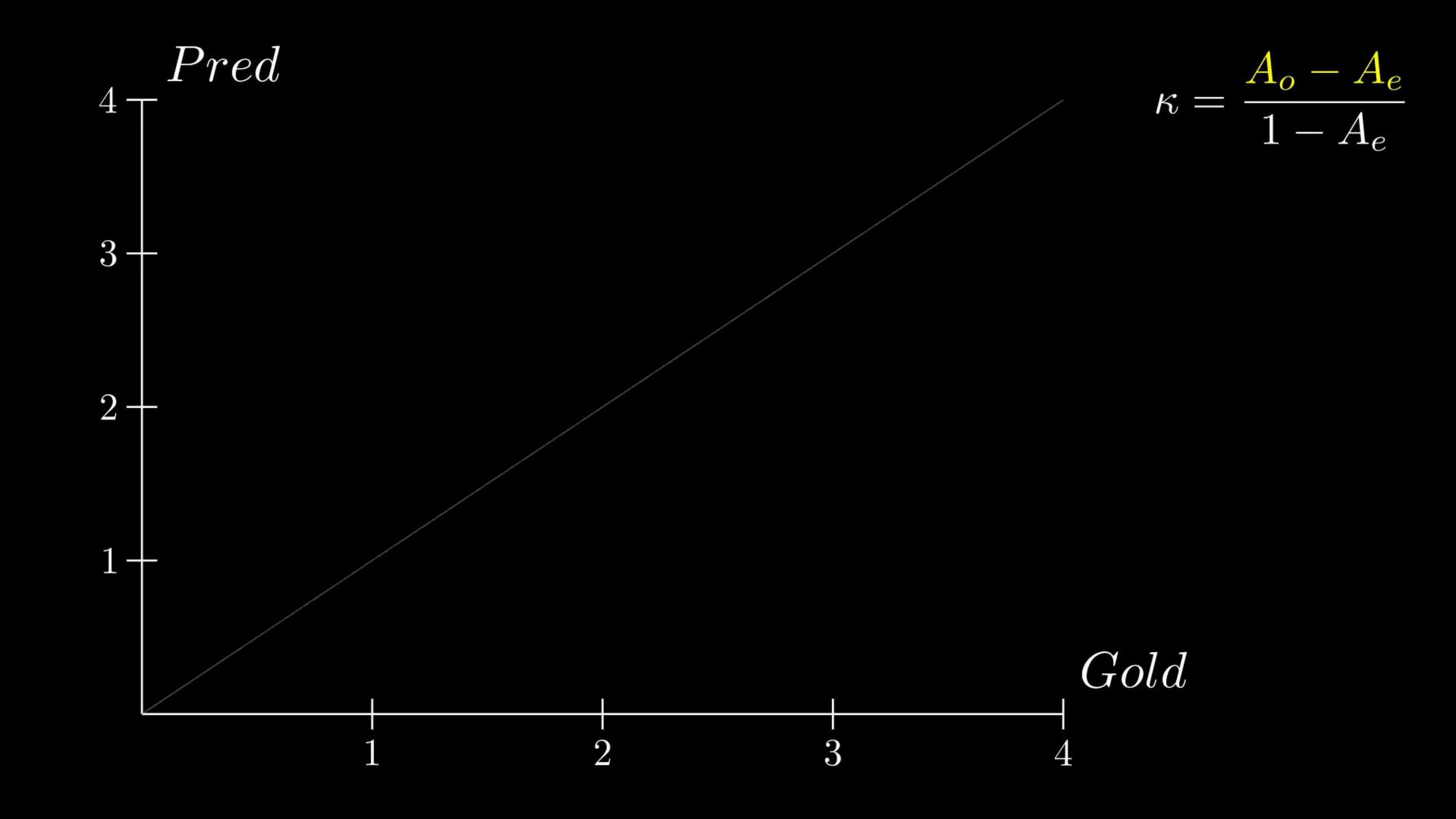
F1 Score

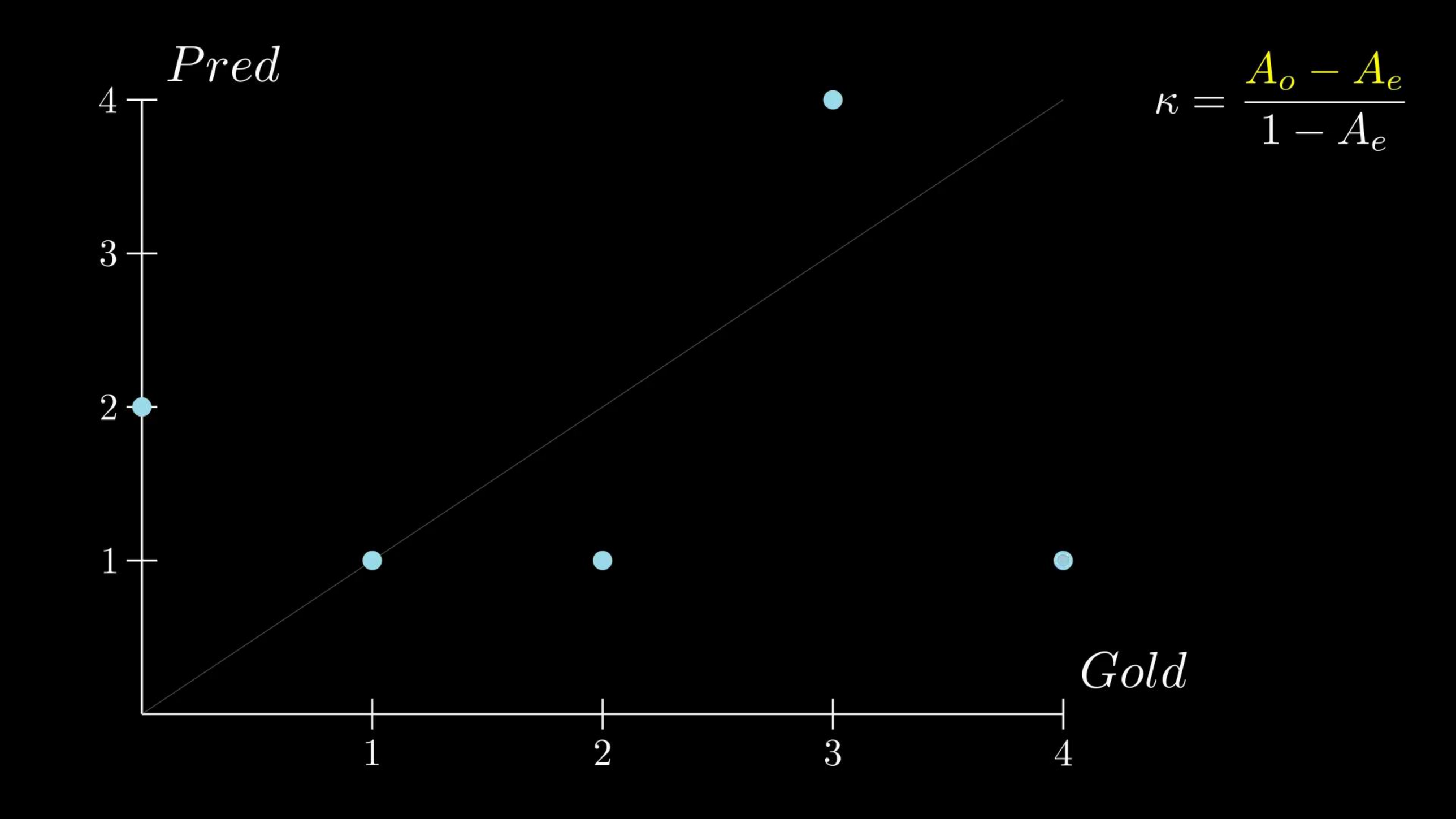
Cohen's Kappa Coefficient (QWK)

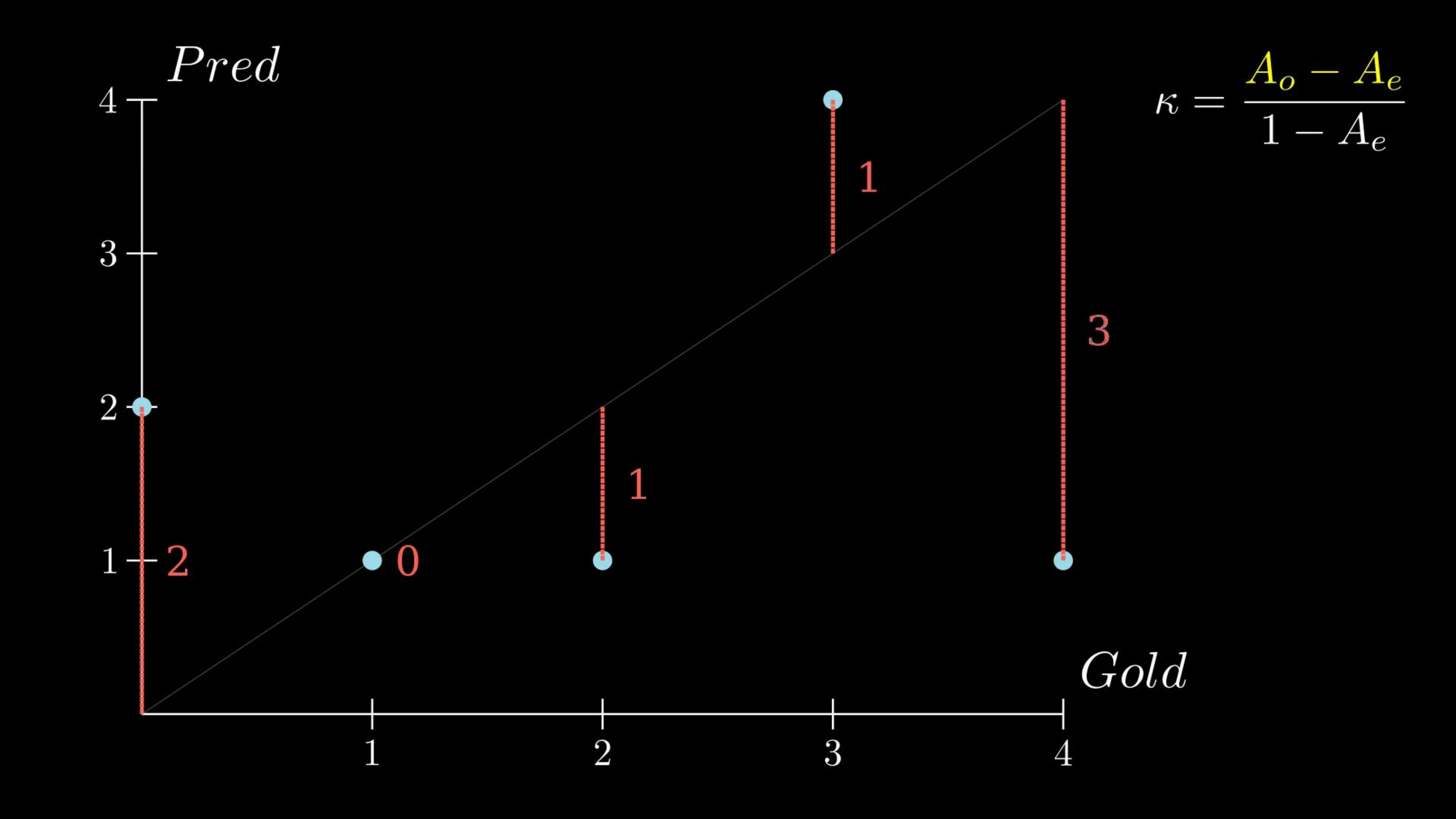
$$\kappa = \frac{A_o - A_e}{\boxed{1 - A_e}}$$

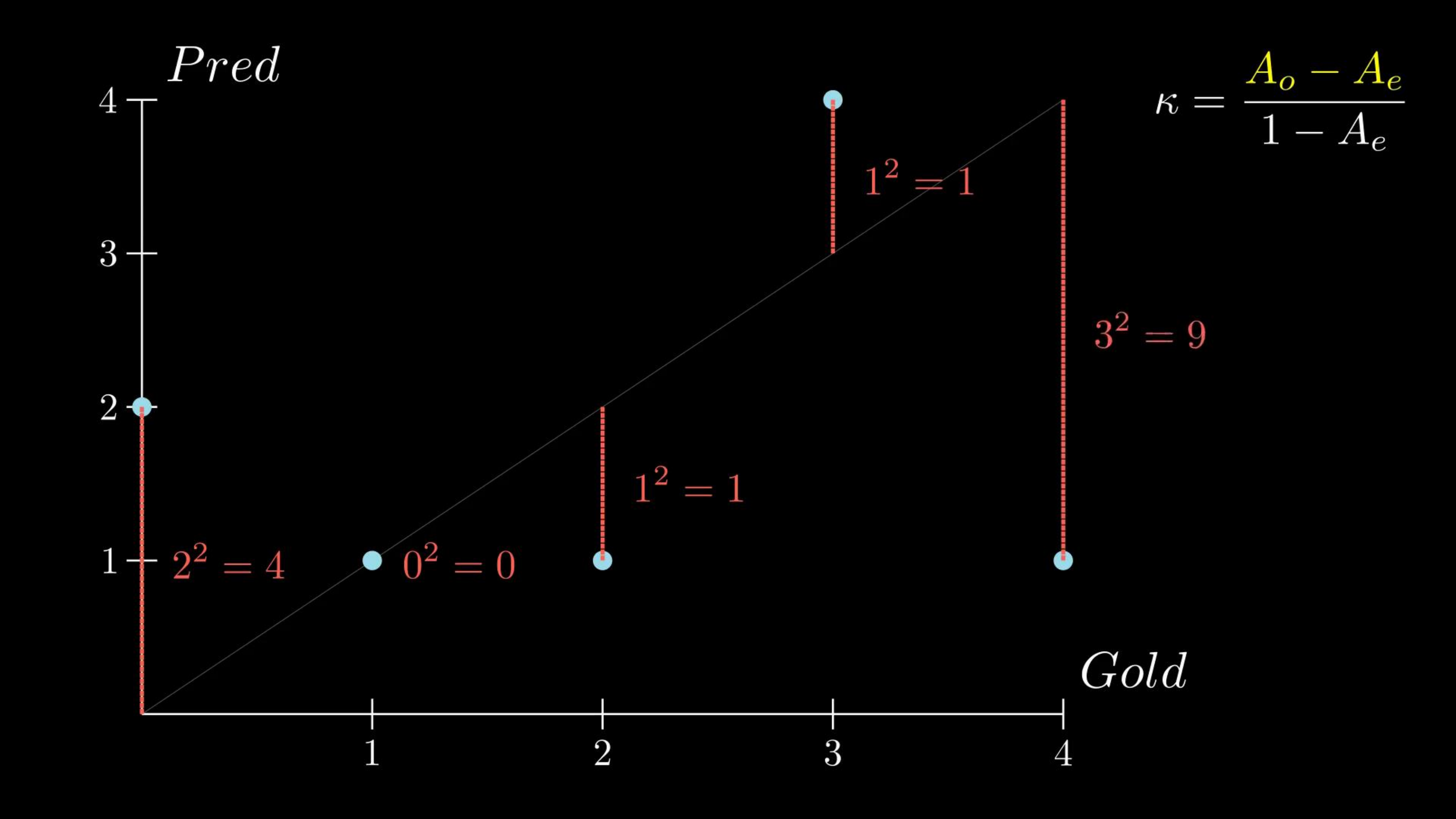
$$\kappa = \frac{\boxed{A_o - A_e}}{1 - A_e}$$


Pred
4
3
2
1
1
2
3
4
Gold
$\kappa = \frac{A_o - A_e}{1 - A_e}$


Pred
4
3
2
1
Gold
1
2
3
4
$\kappa = \frac{A_o - A_e}{1 - A_e}$


Pred
4
3
2
1
1
2
3
4
Gold
$\kappa = \frac{A_o - A_e}{1 - A_e}$

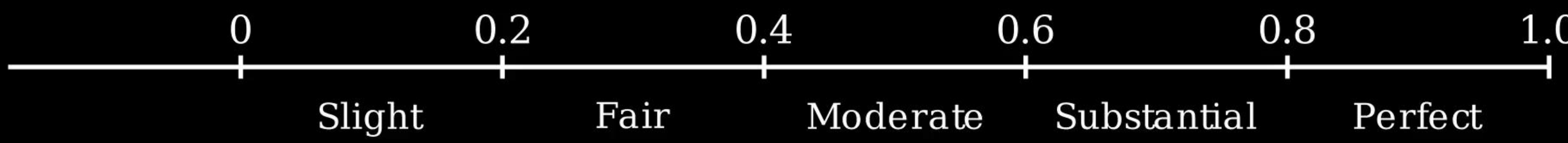

Pred
4
3
2
1
2
0
1
1
3
Gold
1
2
3
4
$\kappa = \frac{A_o - A_e}{1 - A_e}$


Pred
4
3
2
1
$2^2 = 4$
$0^2 = 0$
$1^2 = 1$
$1^2 = 1$
$3^2 = 9$
Gold
1
2
3
4
$\kappa = \frac{A_o - A_e}{1 - A_e}$

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

$$-1 \geq \kappa \geq 1$$

$$-1 \geq \kappa \geq 1$$

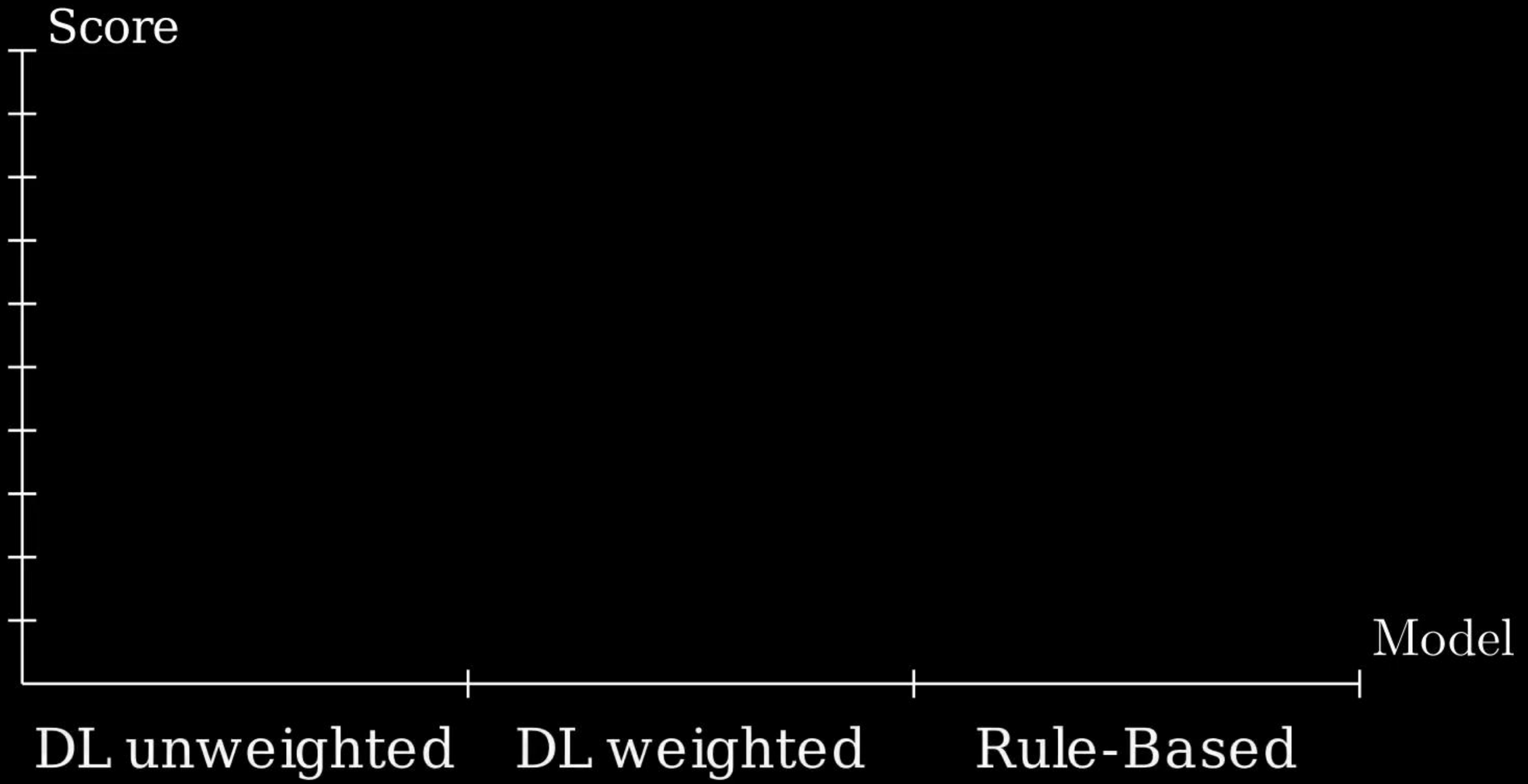


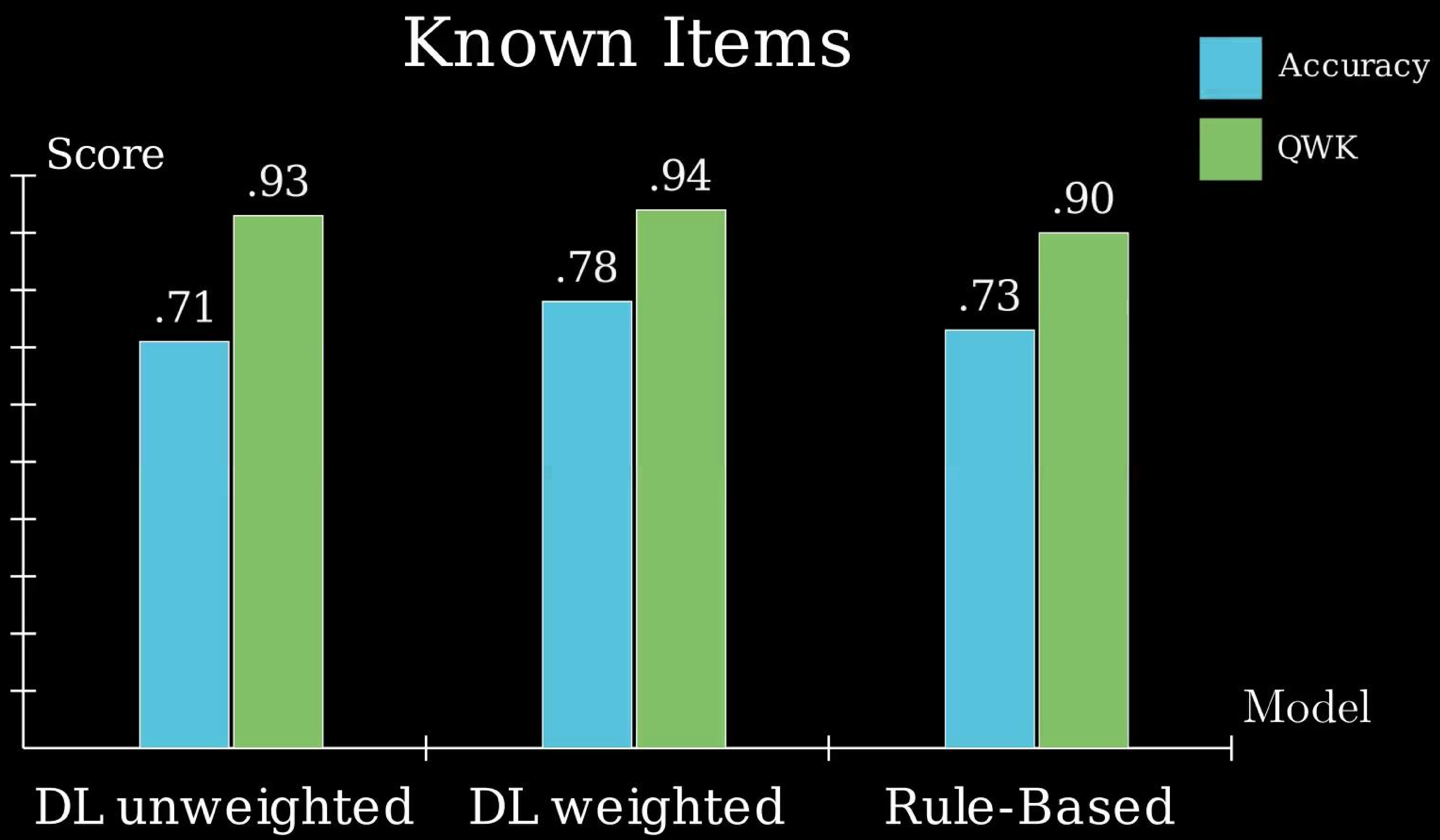
Agreement levels: Landis and Koch (1977)

Overall Performance by Model

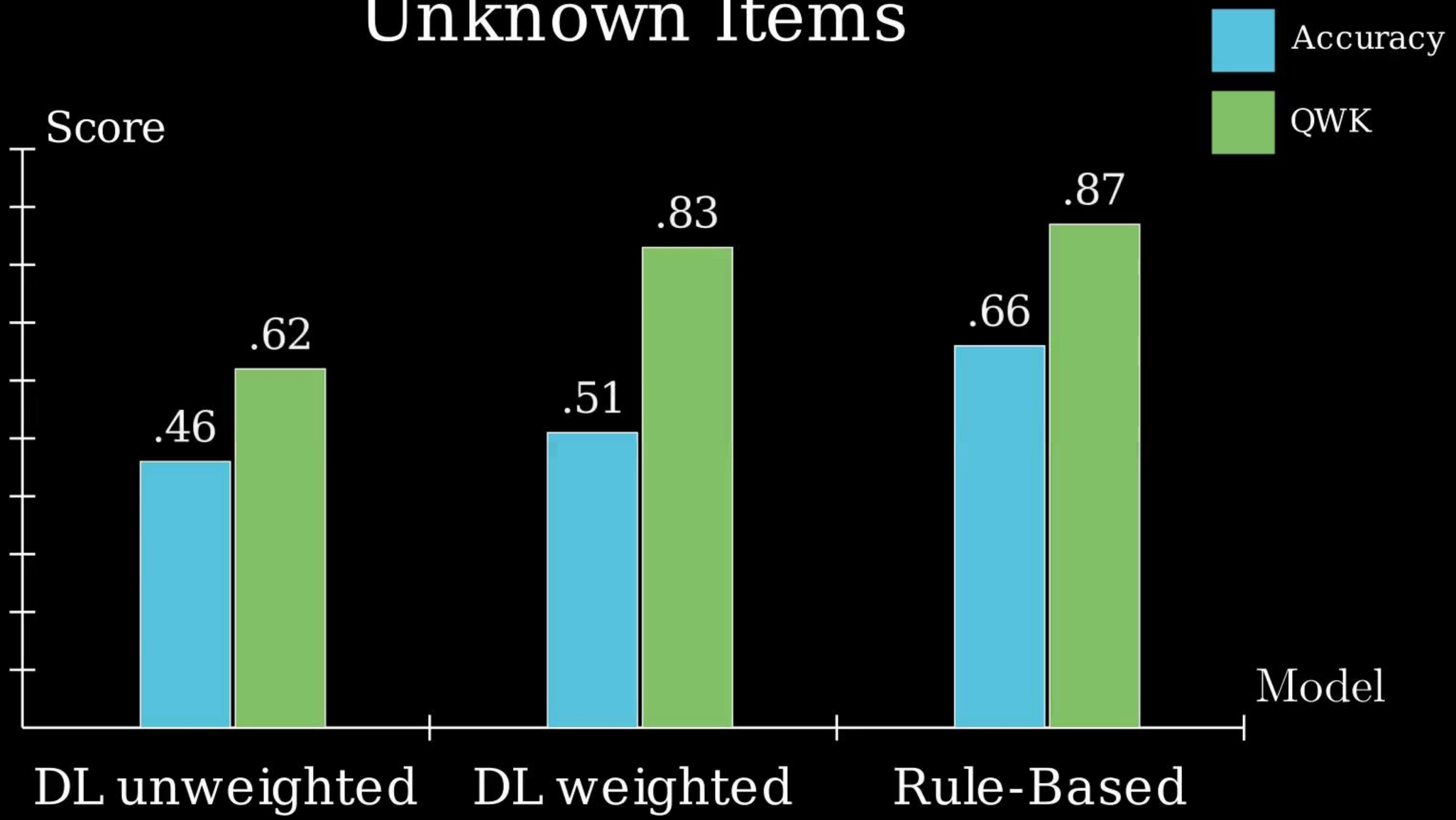
Known Items

Known Items

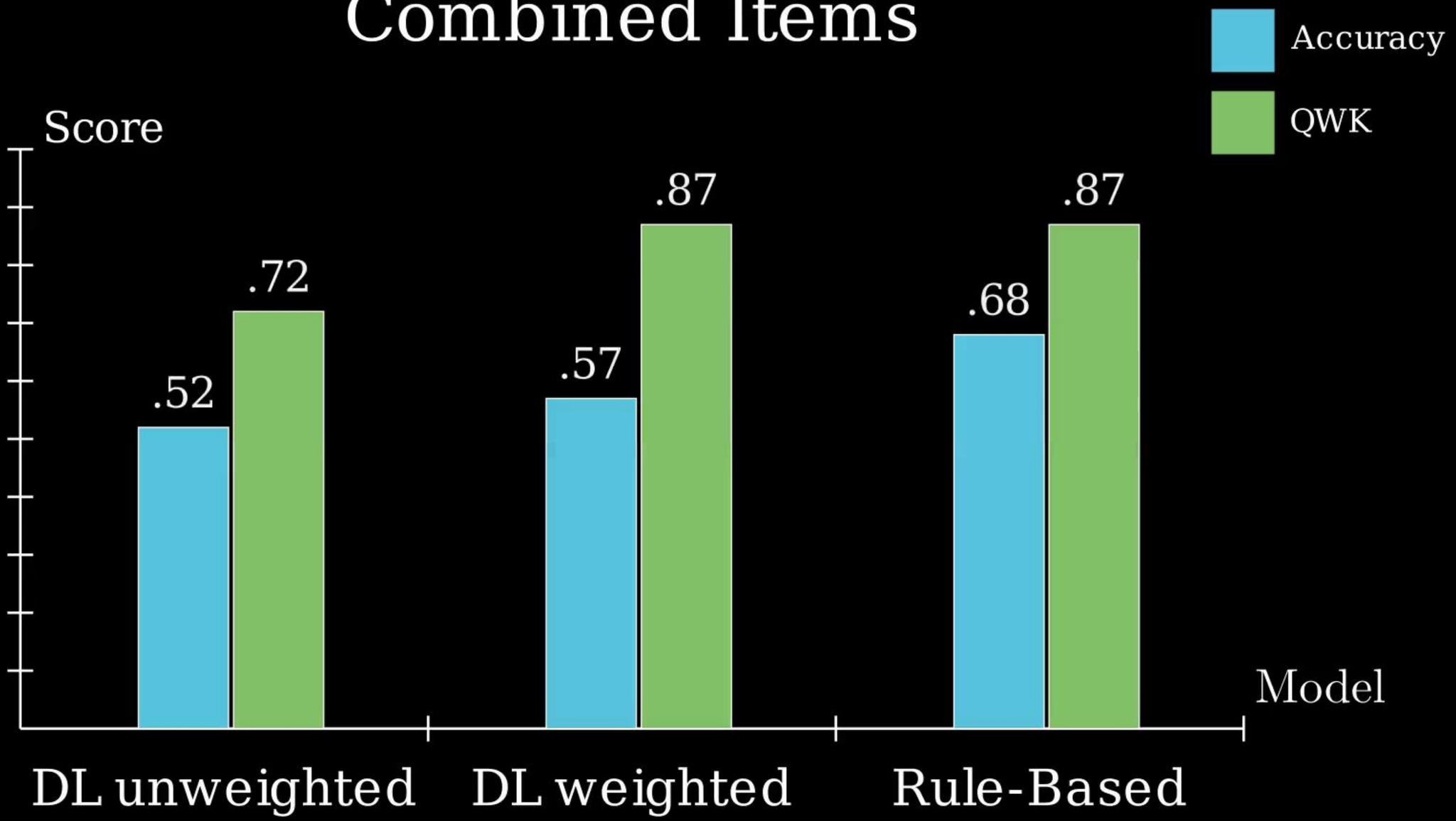



Known Items
Accuracy
QWK
Score
.93
.94
.90
.71
.78
.73
Model
DL unweighted
DL weighted
Rule-Based

Unknown Items



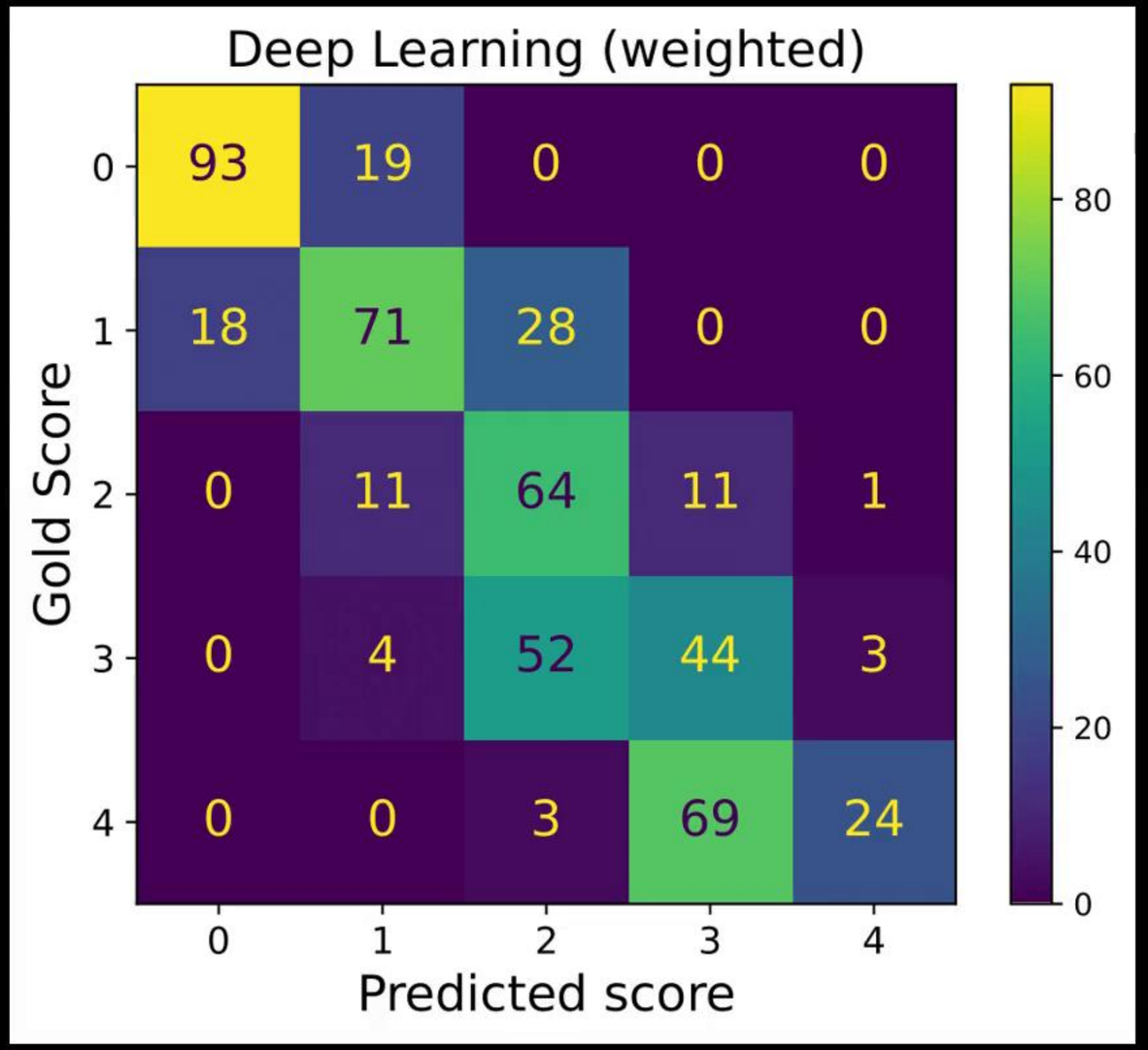
Combined Items

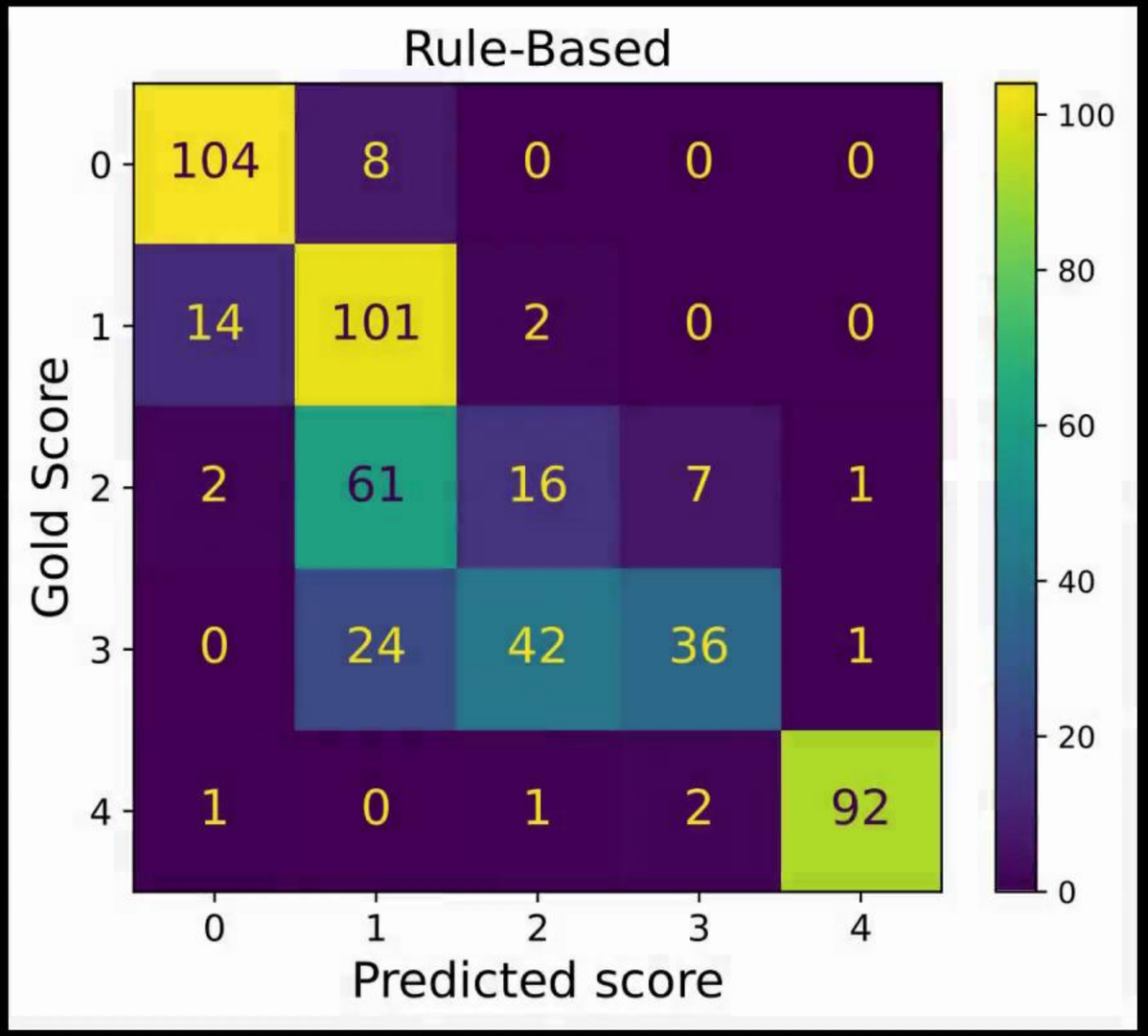


Per-Score Performance by Model

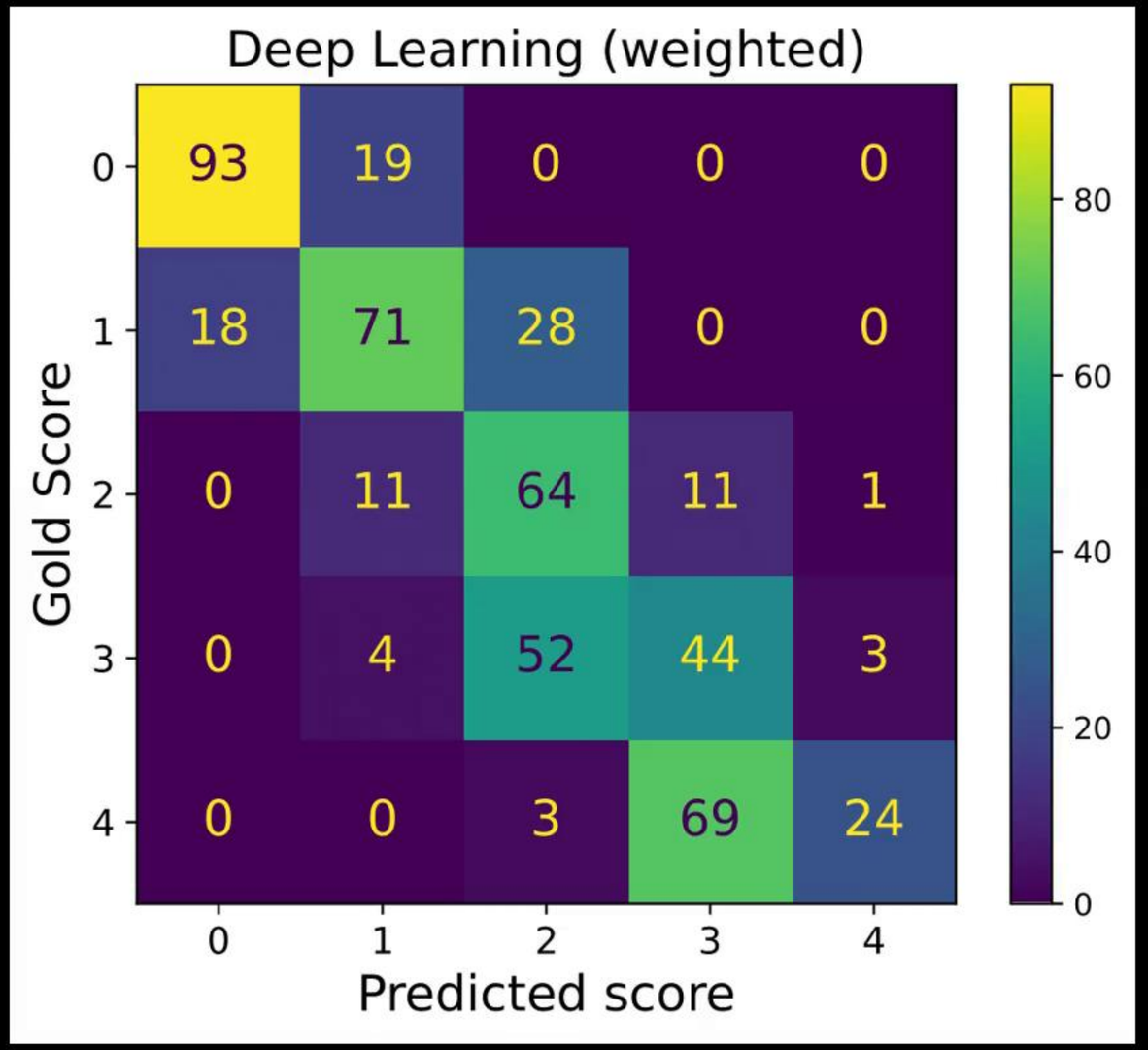
Confusion Matrices

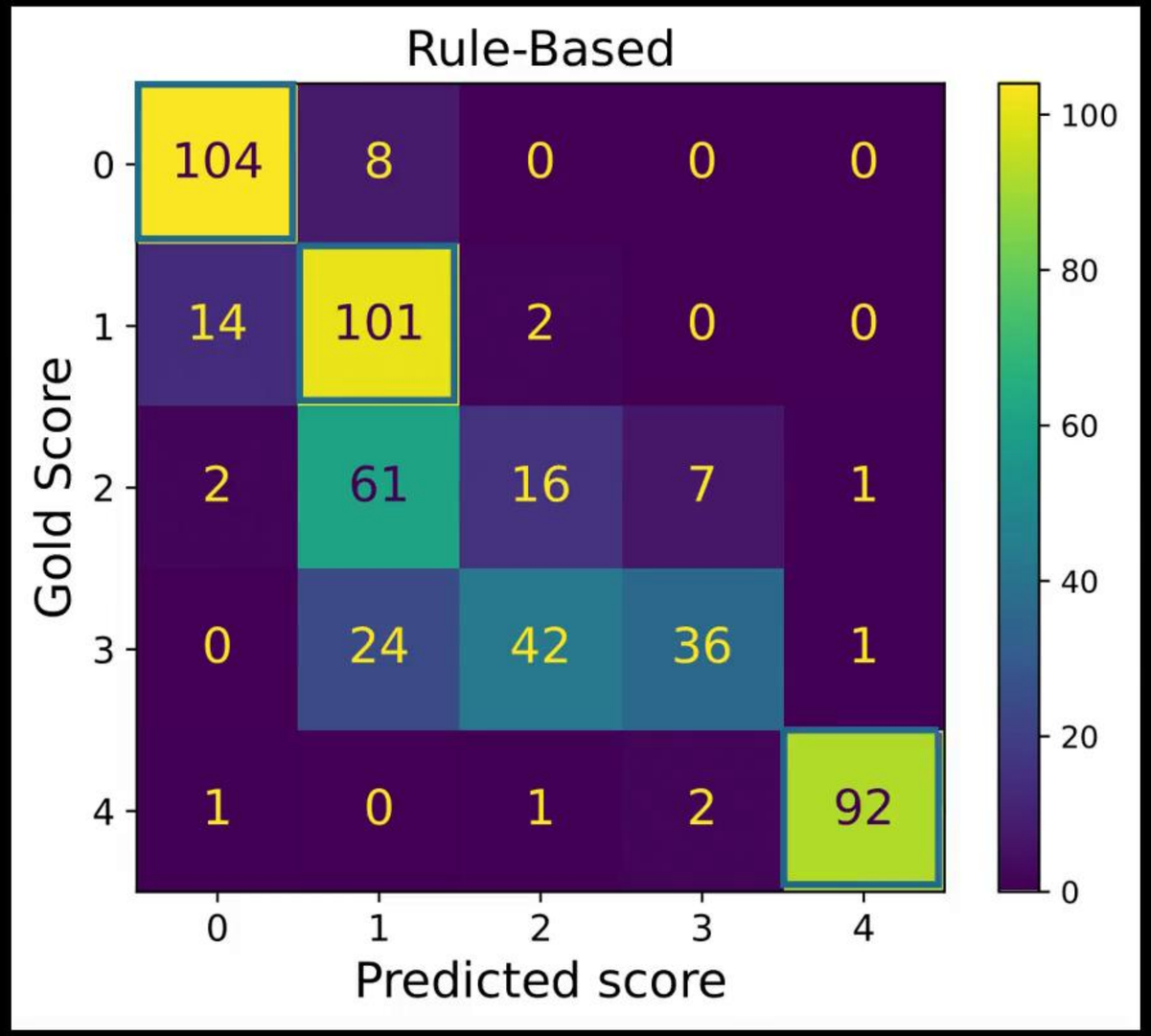
Confusion Matrices



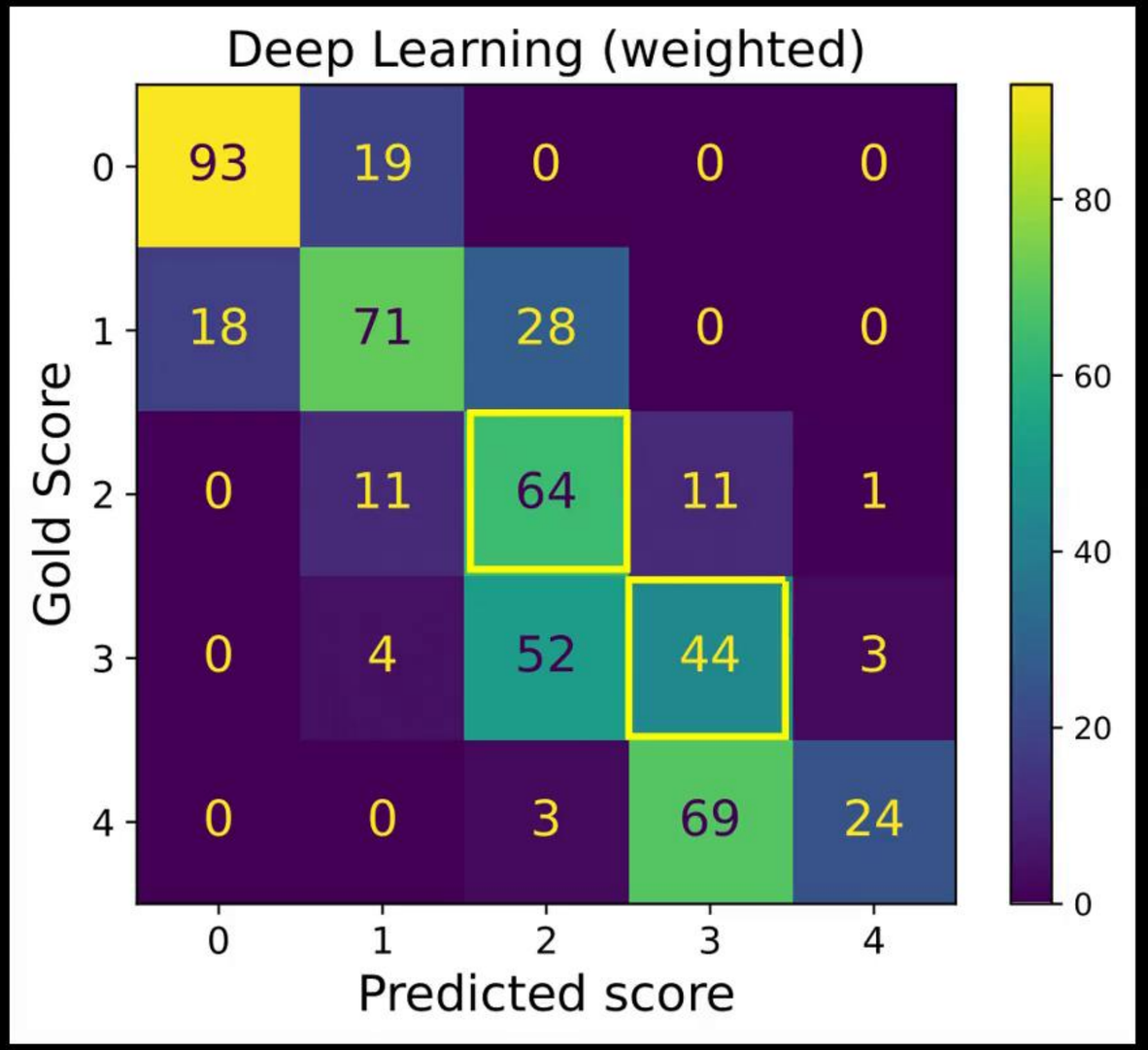


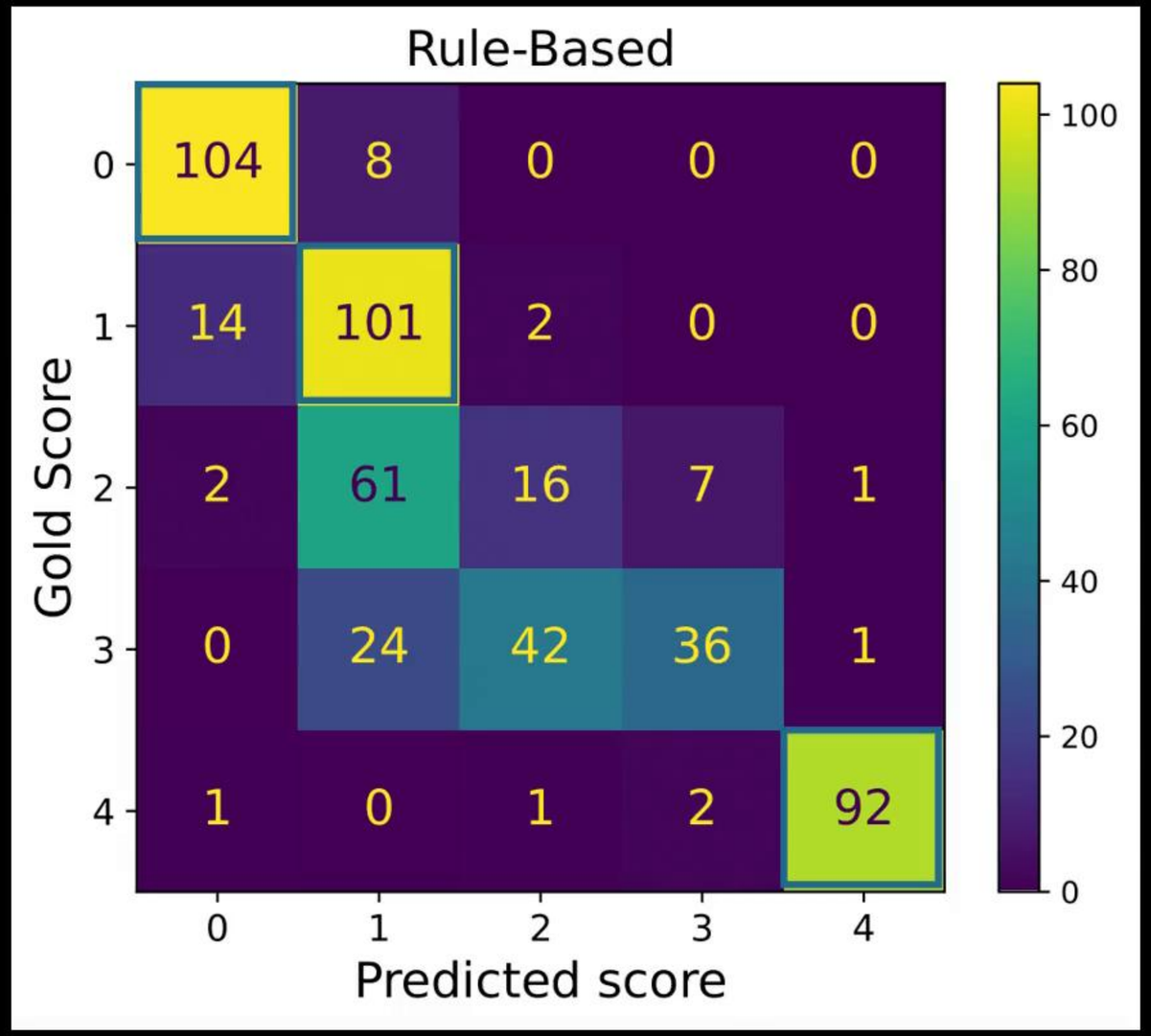
Confusion Matrices





Confusion Matrices





Per-Score Precision, Recall, F1

Per-Score Precision, Recall, F1

Score	Precision		Recall		F1	
	DL	RB	DL	RB	DL	RB
0	.84	**.86**	.83	**.93**	.83	**.89**
1	**.68**	.52	.61	**.86**	.64	**.65**
2	**.44**	.26	**.74**	.18	**.55**	.22
3	.35	**.80**	**.43**	.35	.39	**.49**
4	.86	**.98**	.25	**.96**	.39	**.97**
macro avg	.63	**.68**	.57	**.66**	.56	**.64**
micro avg	.64	**.69**	.57	**.68**	.57	**.66**

Takeaways & Outlook

Takeaways & Outlook

✔ Promising performances with room for improvement

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths
- ✔ RB: Strong on extremes and unseen items

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths
- ✔ RB: Strong on extremes and unseen items
- ✔ DL: Better in the mid-range, where rules are fuzzier

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths
- ✔ RB: Strong on extremes and unseen items
- ✔ DL: Better in the mid-range, where rules are fuzzier

➡ Hybrid approach as a 'best of both worlds' solution

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths
- ✔ RB: Strong on extremes and unseen items
- ✔ DL: Better in the mid-range, where rules are fuzzier

➡ Hybrid approach as a 'best of both worlds' solution

➡ Adaptive Scoring

Takeaways & Outlook

- ✔ Promising performances with room for improvement
- ✔ Both models: High QWK values, complementary strengths
- ✔ RB: Strong on extremes and unseen items
- ✔ DL: Better in the mid-range, where rules are fuzzier

➡ Hybrid approach as a 'best of both worlds' solution

➡ Adaptive Scoring

➡ Integrate LLMs

Thanks for listening!

Let’s dive into your questions!

Mihail Chifligarov (www.chifligarov.dev)