UNIREX: A Unified Learning Framework for Language Model Rationale Extraction

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan,
Shaoliang Nie, Xiaochang Peng, Xiang Ren, Hamed Firooz

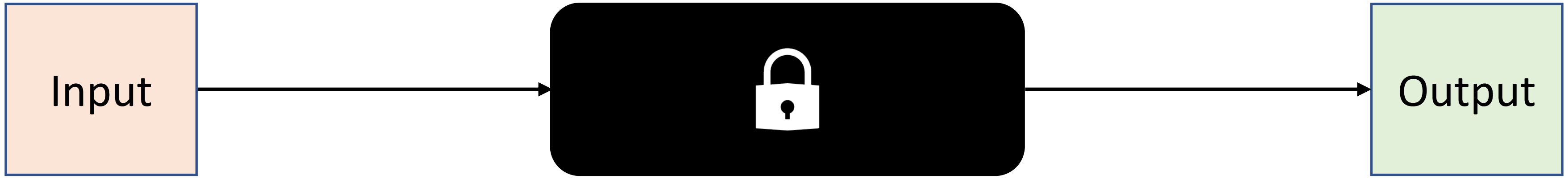
ICML 2022

Input

Neural Language Model (NLM)

Output

Still , this flick is fun , and host to some truly excellent sequences .	✅ **Sentiment Analysis**	**positive**	negative	
<premise> Children smiling and waving at camera <hypothesis> The kids are frowning	✅ **Natural Language Inference**	entailment	neutral	**contradiction**
What happens when someone is playing too aggressively?	✅ **Question Answering**	**have injury**	get hungry	feel happy

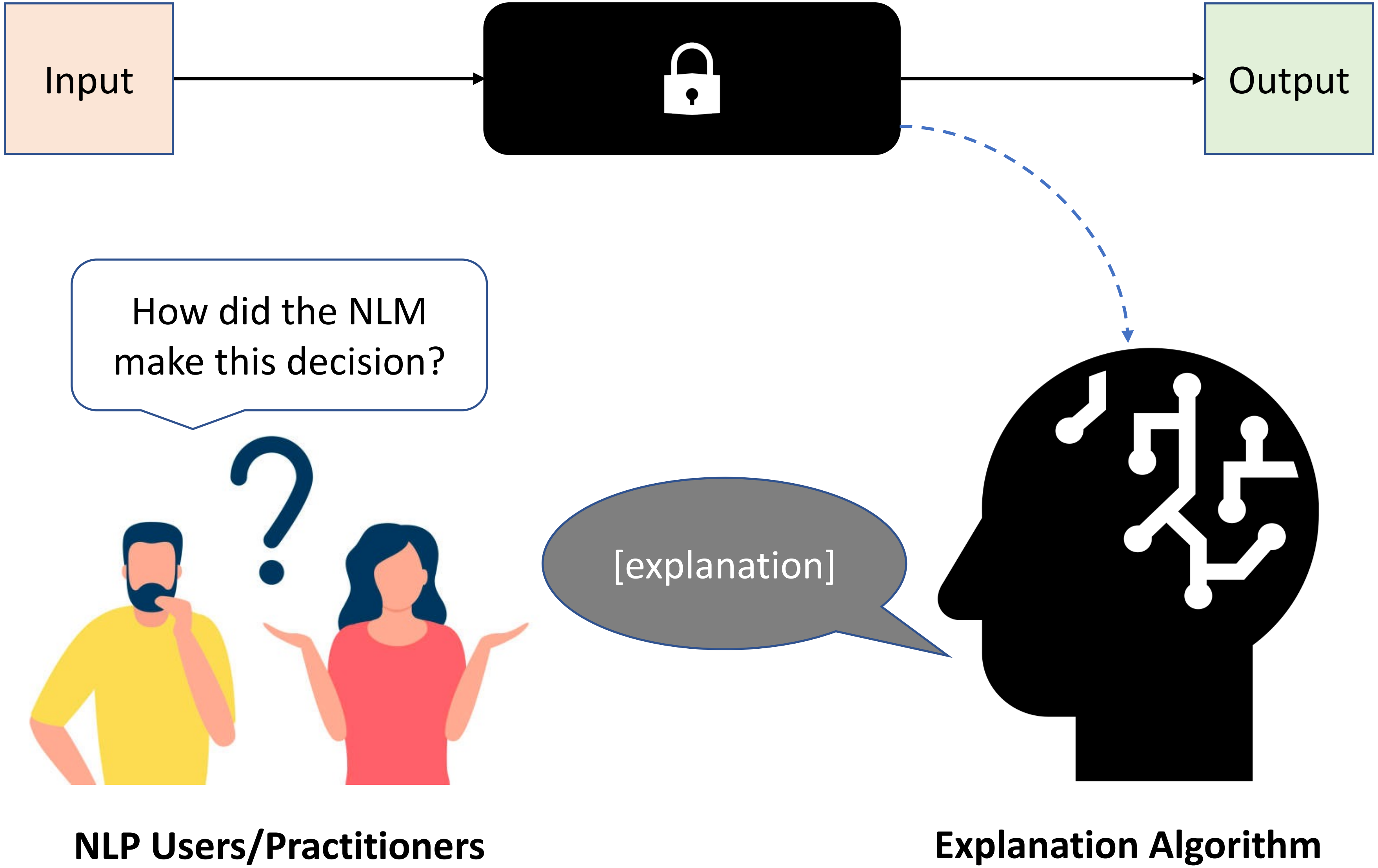


Input

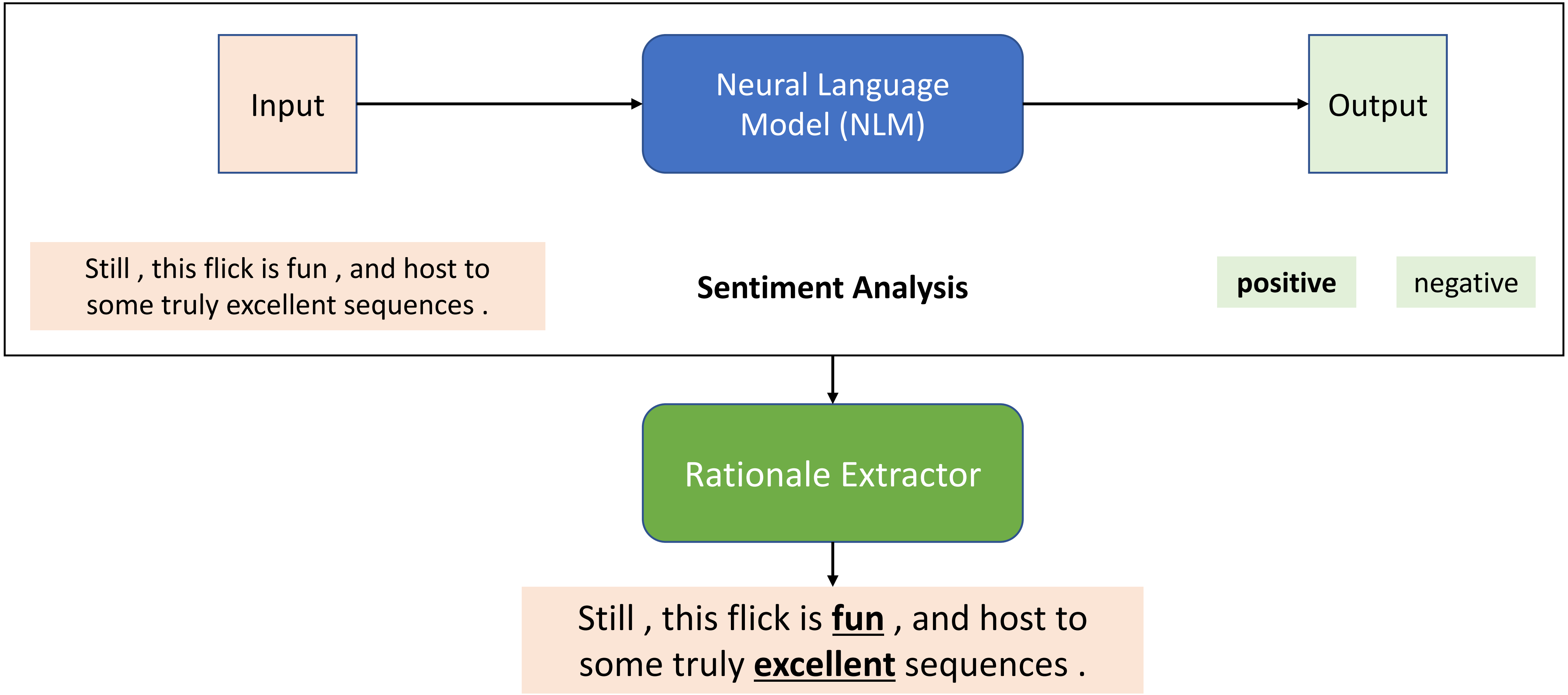
Output

How did the NLM make this decision?

NLP Users/Practitioners


Input
Output
How did the NLM make this decision?
[explanation]
NLP Users/Practitioners
Explanation Algorithm

Rationale Extraction



Input

Neural Language Model (NLM)

Output

Still , this flick is fun , and host to some truly excellent sequences .

Sentiment Analysis

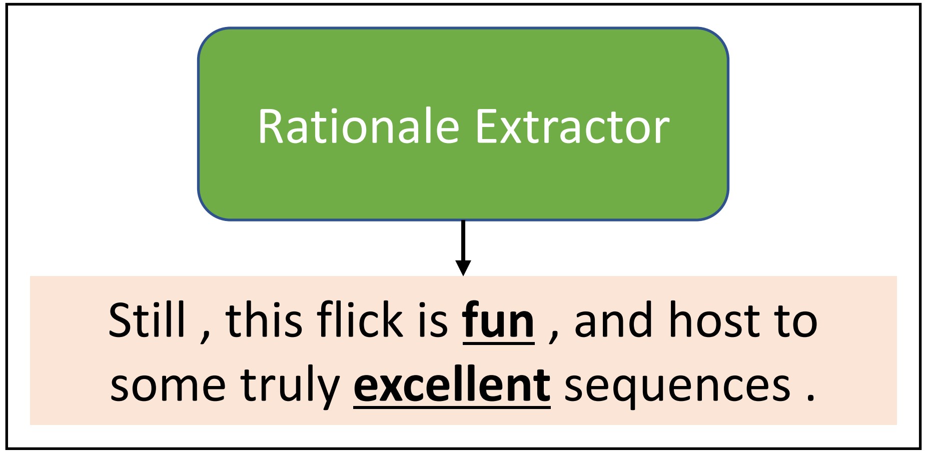
positive

negative

Rationale Extractor

Still , this flick is **fun** , and host to some truly **excellent** sequences .

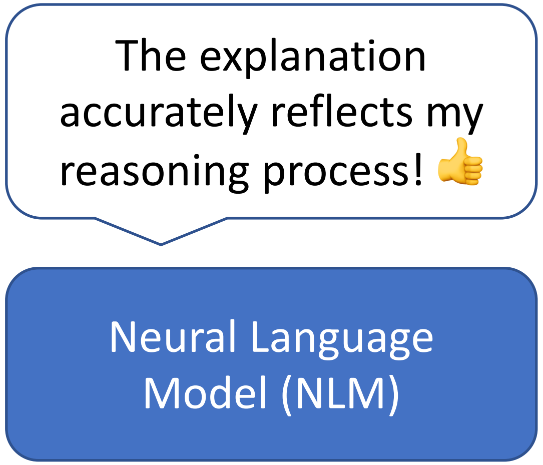
Three Desiderata of Rationale Extraction



Still , this flick is **fun** , and host to some truly **excellent** sequences .

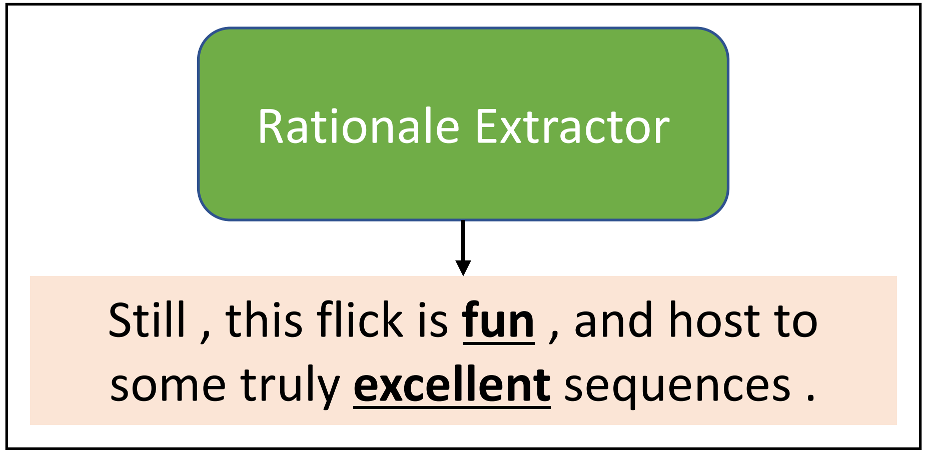
Three Desiderata of Rationale Extraction

1. Faithfulness



The explanation accurately reflects my reasoning process! 👍

Neural Language Model (NLM)

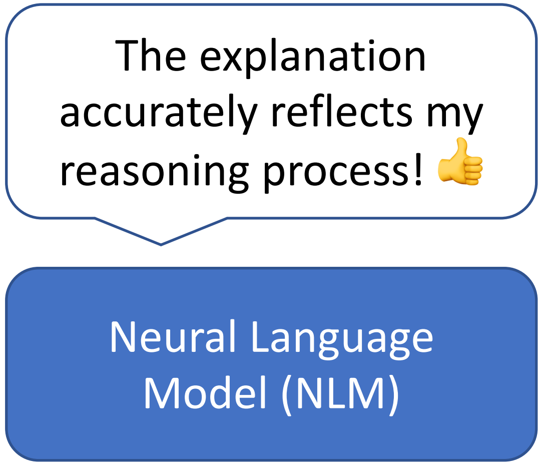


Rationale Extractor

Still , this flick is **fun** , and host to some truly **excellent** sequences .

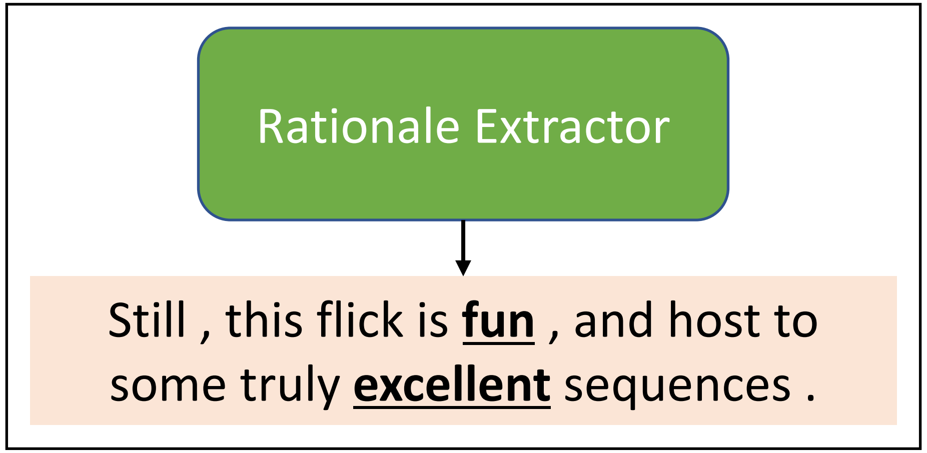
Three Desiderata of Rationale Extraction

1. Faithfulness



The explanation accurately reflects my reasoning process! 👍

Neural Language Model (NLM)



Rationale Extractor

Still , this flick is **fun** , and host to some truly **excellent** sequences .

2. Plausibility

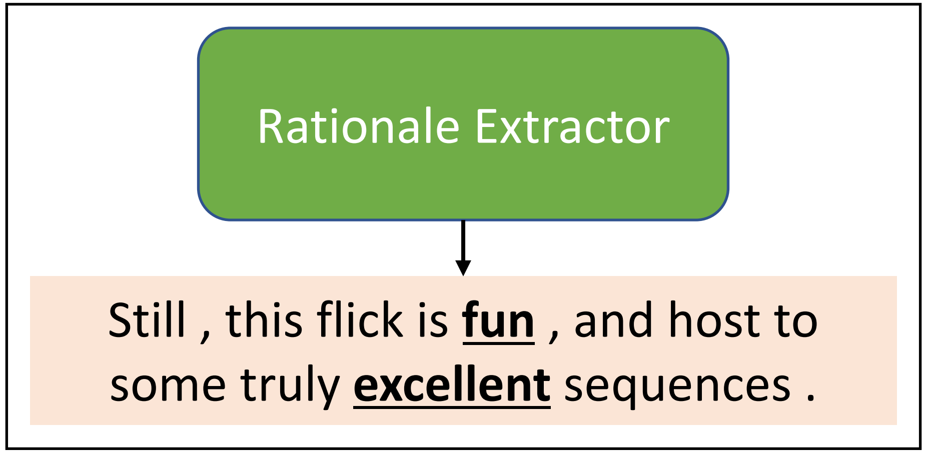
The explanation makes sense to us! 👍

Three Desiderata of Rationale Extraction

1. Faithfulness

The explanation accurately reflects my reasoning process! 👍

Neural Language Model (NLM)



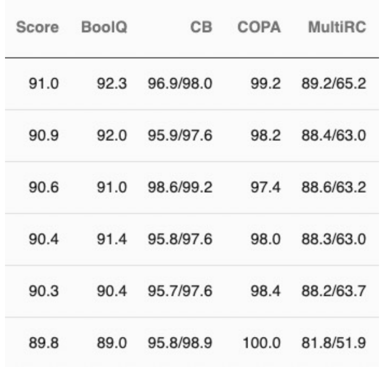
Rationale Extractor

Still , this flick is **fun** , and host to some truly **excellent** sequences .

2. Plausibility

The explanation makes sense to us! 👍

3. Task Performance



Score	BoolQ	CB	COPA	MultiRC
91.0	92.3	96.9/98.0	99.2	89.2/65.2
90.9	92.0	95.9/97.6	98.2	88.4/63.0
90.6	91.0	98.6/99.2	97.4	88.6/63.2
90.4	91.4	95.8/97.6	98.0	88.3/63.0
90.3	90.4	95.7/97.6	98.4	88.2/63.7
89.8	89.0	95.8/98.9	100.0	81.8/51.9

Faithfulness

Plausibility

Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)

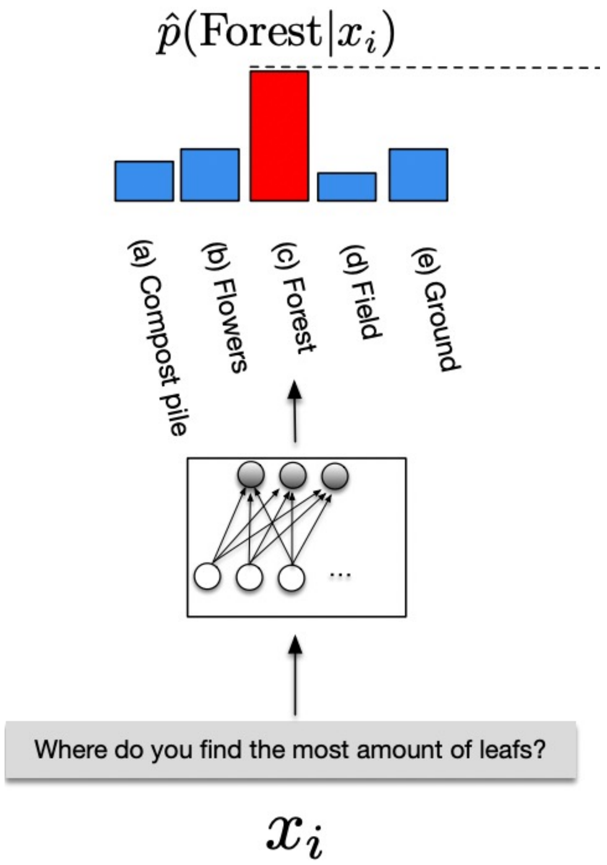
Plausibility

Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)

Plausibility

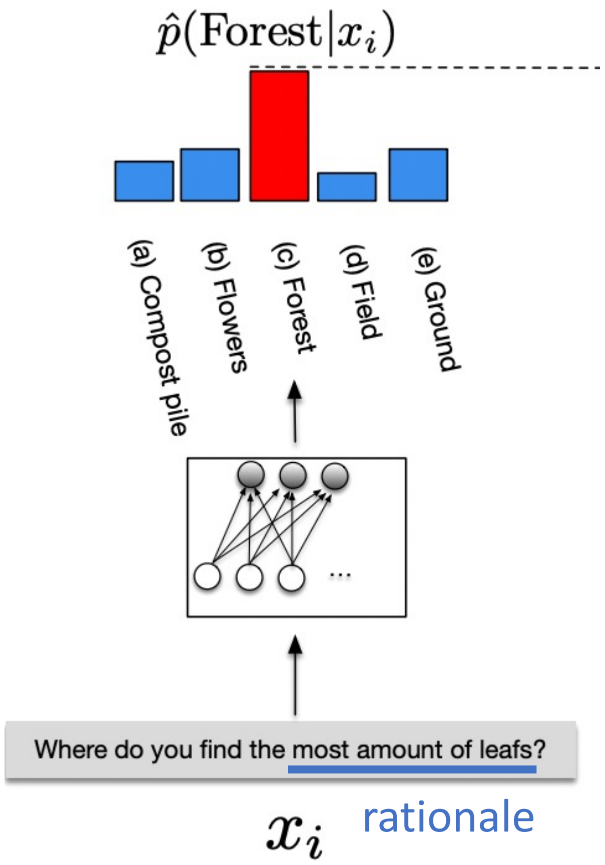

$\hat{p}(\text{Forest}|x_i)$
(a) Compost pile
(b) Flowers
(c) Forest
(d) Field
(e) Ground
...
Where do you find the most amount of leafs?
x_i

Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)

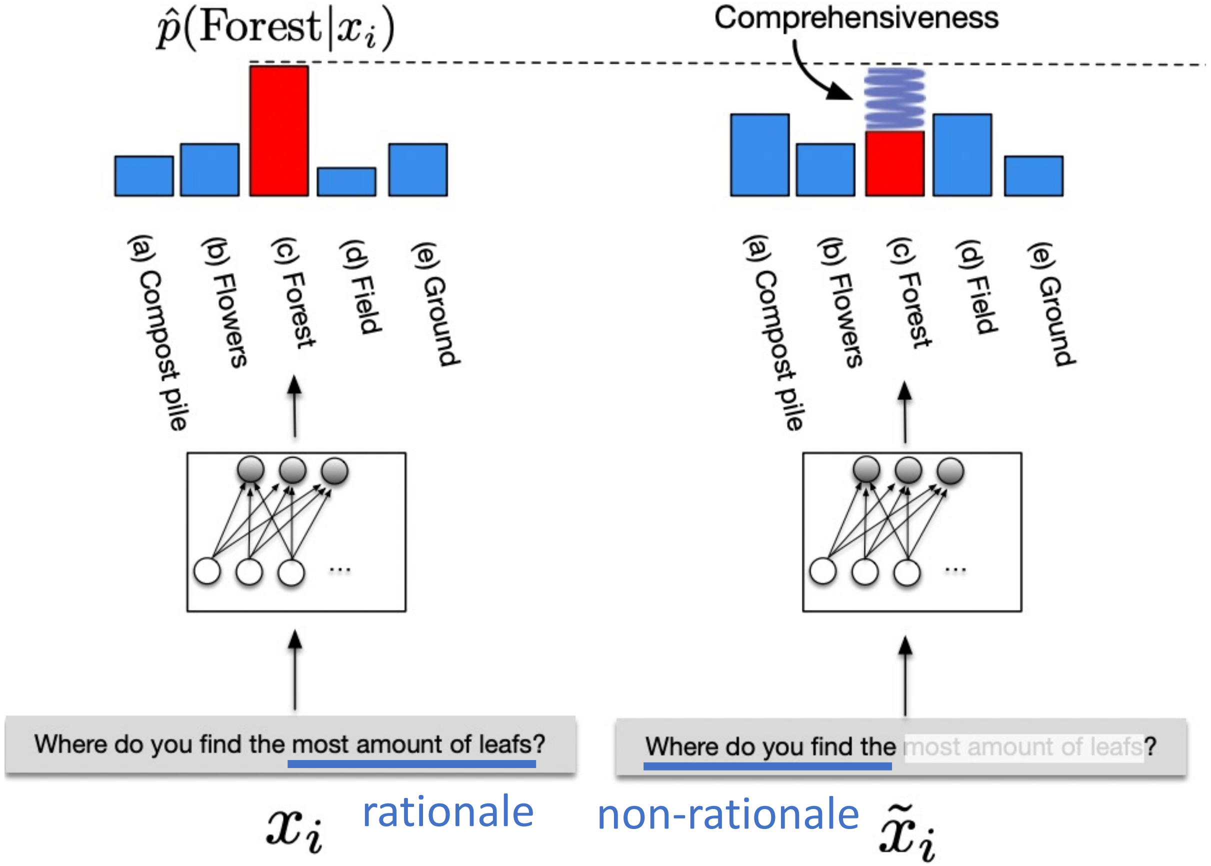
Plausibility



Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)



Where do you find the most amount of leafs?

Where do you find the most amount of leafs?

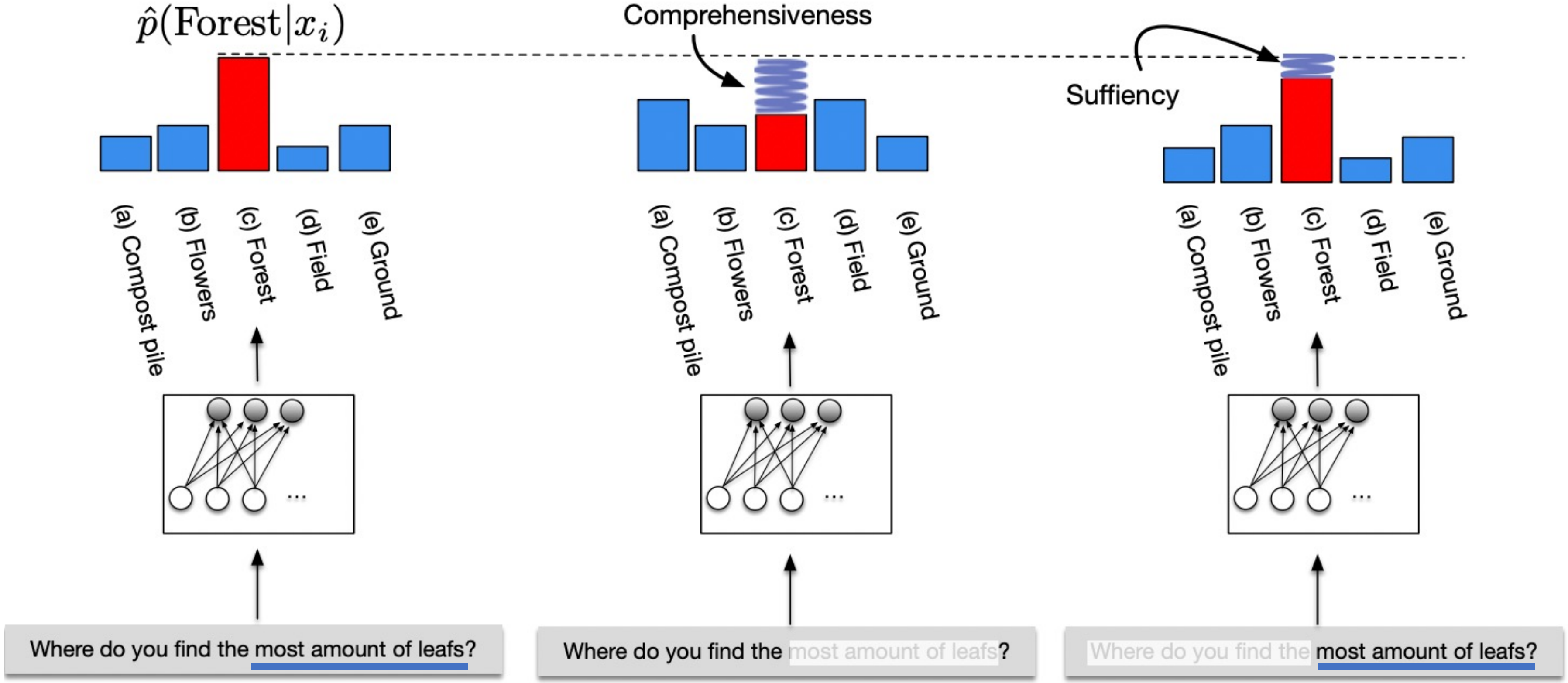
***Higher* Comp is better!**

Plausibility

Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)



Where do you find the most amount of leafs?

Where do you find the most amount of leafs?

Where do you find the most amount of leafs?

x_i rationale

$\tilde{x}_i$

r_i rationale

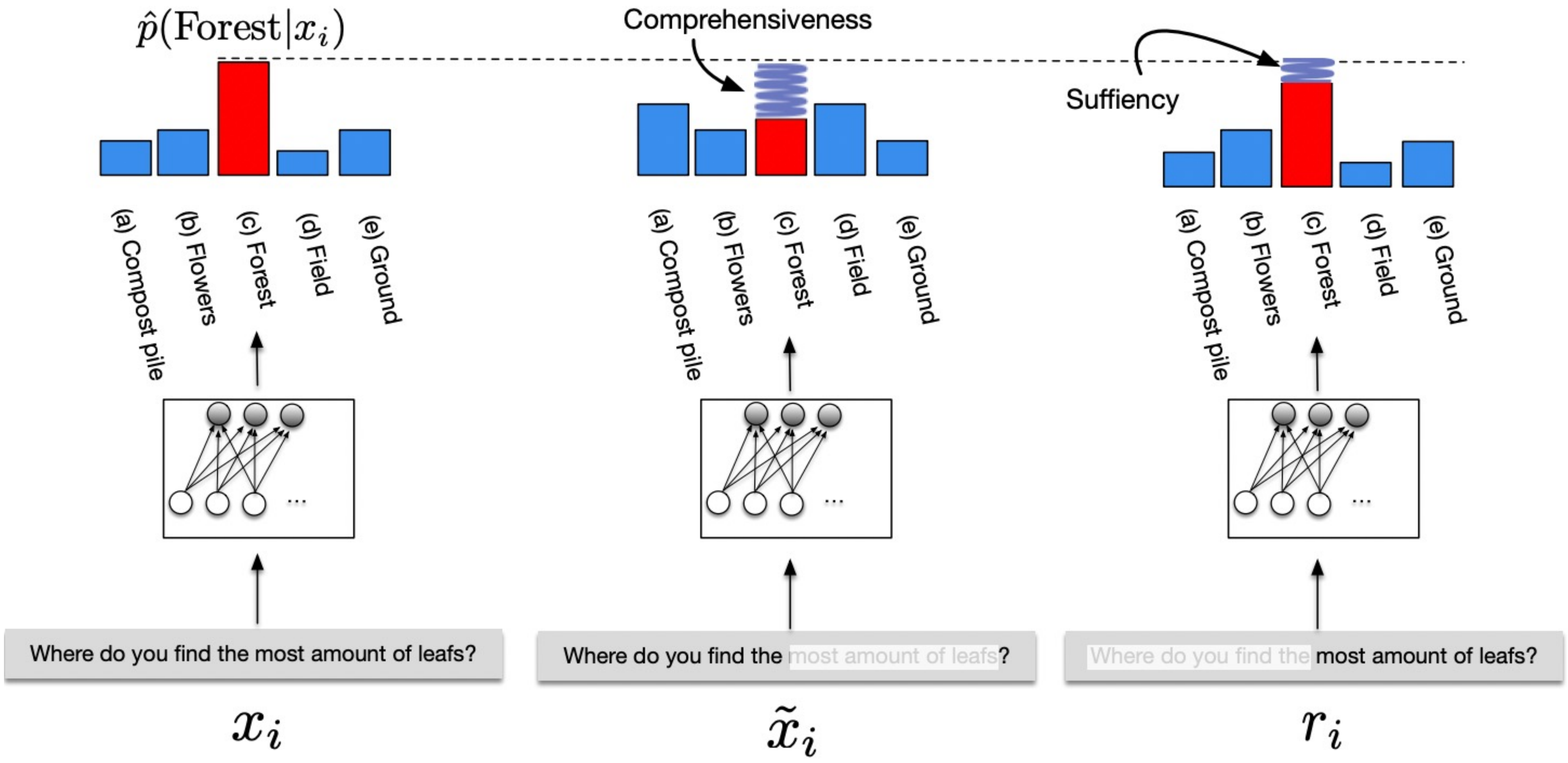
***Lower* Suff is better!**

Plausibility

Faithfulness

Comprehensiveness (Comp)

Sufficiency (Suff)



Where do you find the most amount of leafs?

Where do you find the most amount of leafs?

Where do you find the most amount of leafs?

Plausibility

Similarity to Gold Rationales

Predicted Rationale

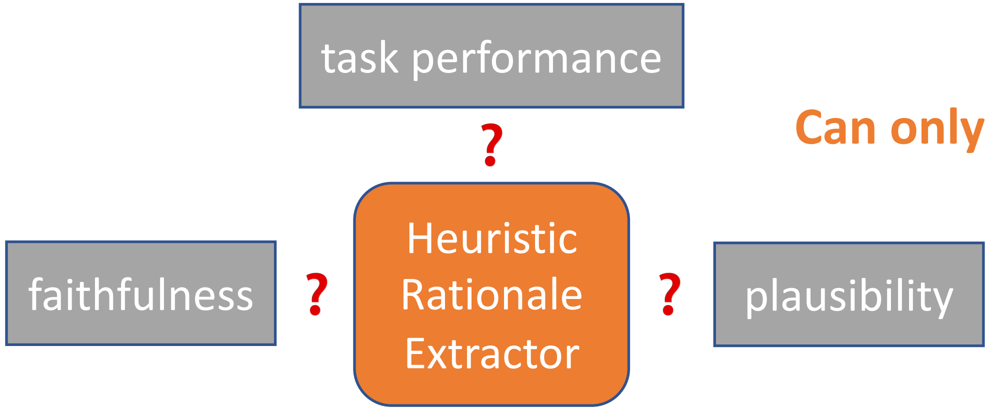
Still , this **flick** is **fun** , and **host** to **some** truly excellent sequences .

Gold Rationale

Still , this **flick** is **fun** , and host to some **truly excellent sequences** .

UNIREX: **UNI**fied Learning Framework for **R**ationale **EX**traction

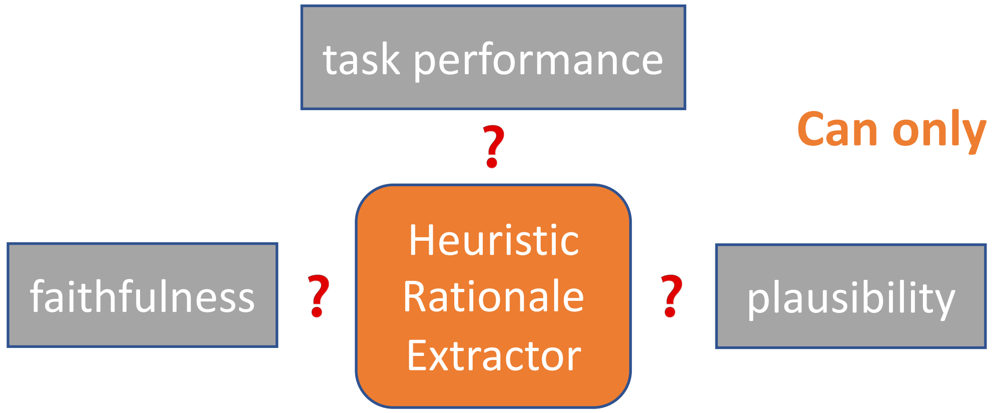
Existing Works



Can only choose up to two!

UNIREX: UNIfied Learning Framework for Rationale EXtraction

Existing Works



task performance

?

faithfulness

?

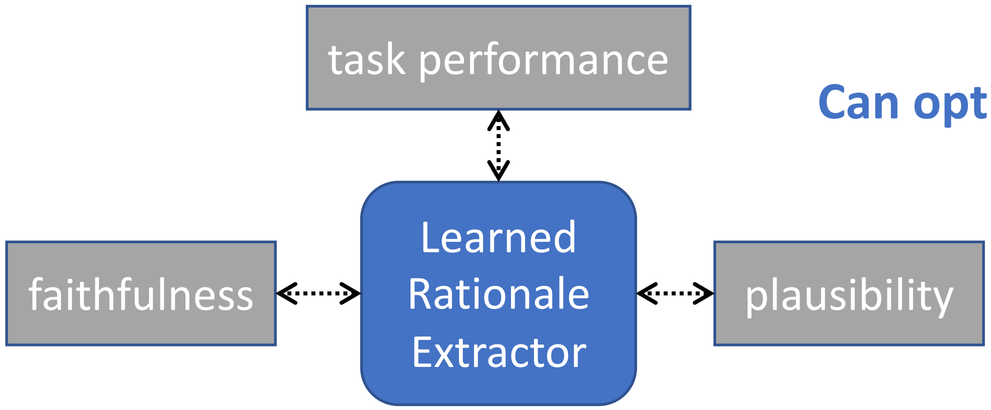
Heuristic Rationale Extractor

?

plausibility

Can only choose up to two!

UNIREX



task performance

faithfulness

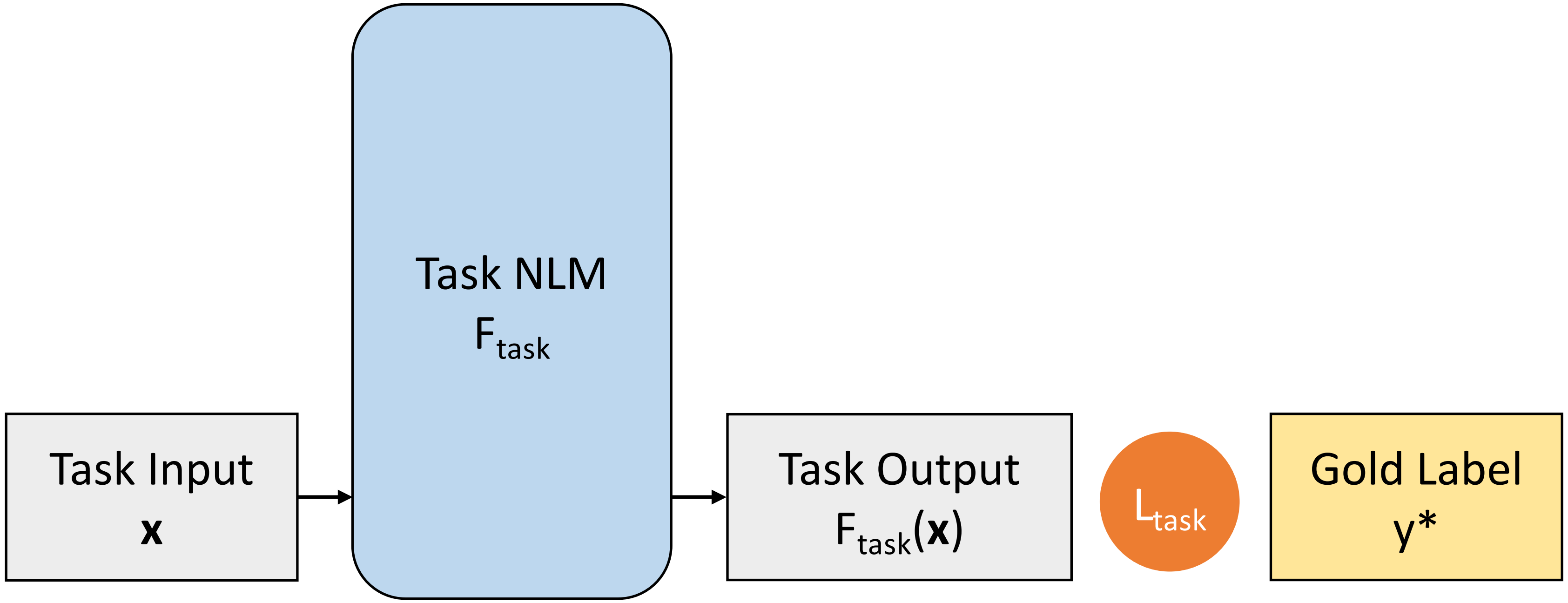
Learned Rationale Extractor

plausibility

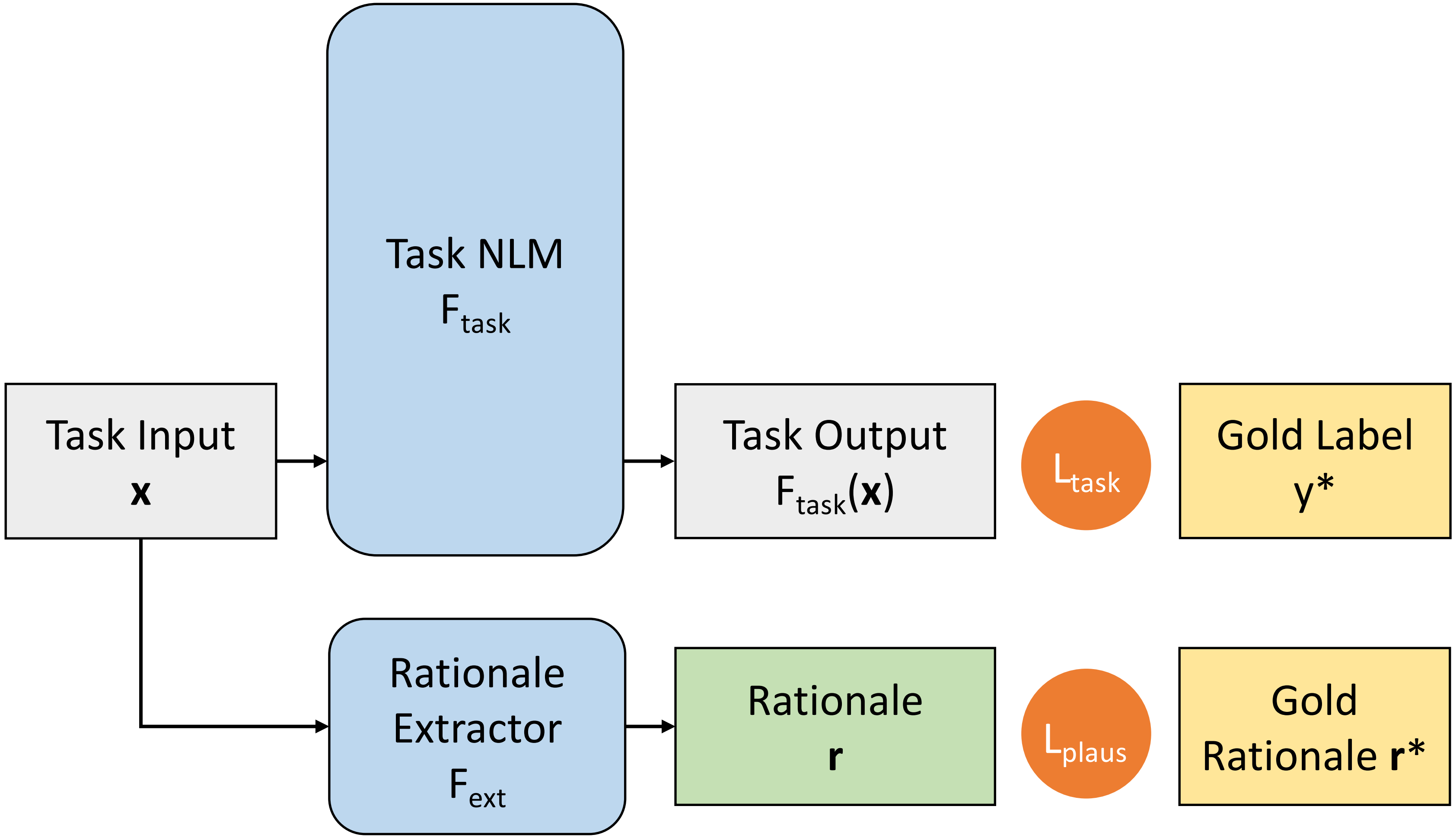
Can optimize for all three!

UNIREX

UNIREX

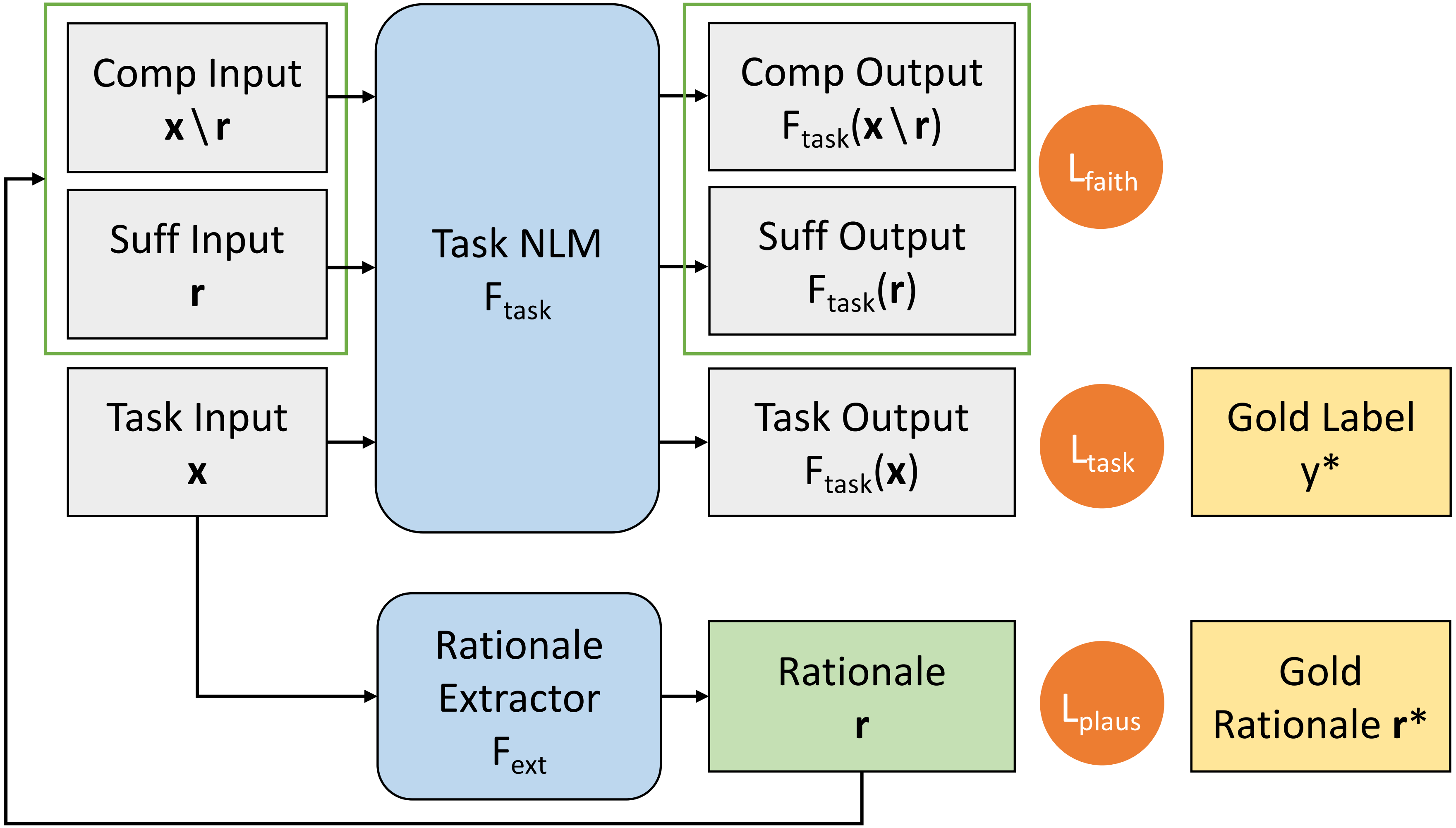

Task Input
x
Task NLM
F_{task}
Task Output
$F_{task}(\mathbf{x})$
L_{task}
Gold Label
y*

UNIREX



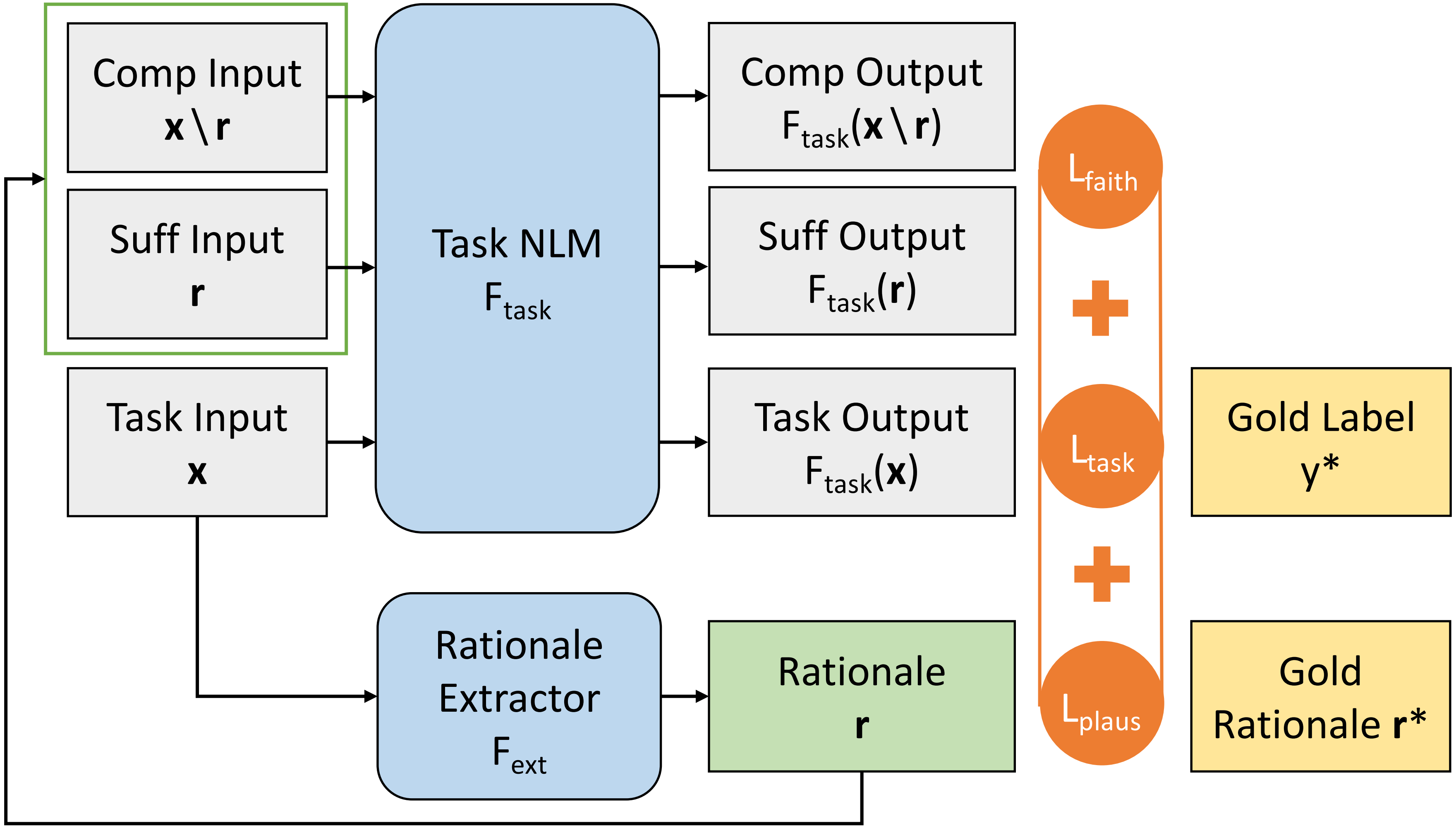
Task NLM
F_{task}
Task Input
x
Task Output
$F_{task}(\mathbf{x})$
L_{task}
Gold Label
y*
Rationale
Extractor
F_{ext}
Rationale
r
L_{plaus}
Gold
Rationale **r***

UNIREX

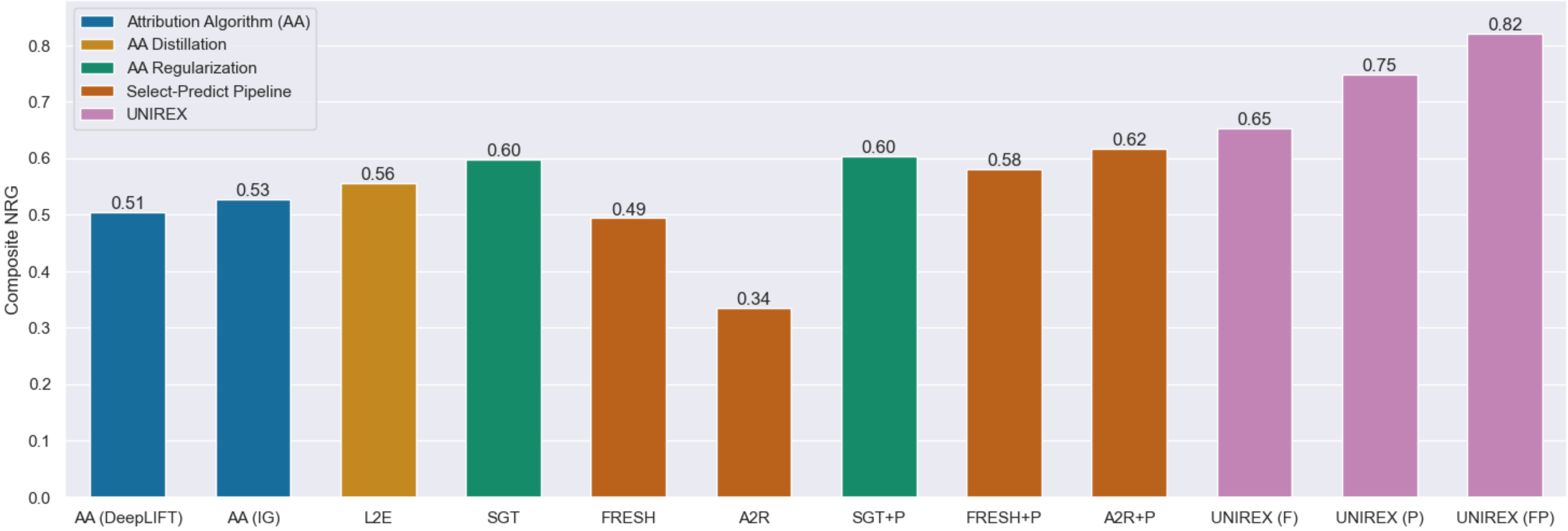


Comp Input
$\mathbf{x} \setminus \mathbf{r}$
Suff Input
$\mathbf{r}$
Task Input
$\mathbf{x}$
Task NLM
F_{task}
Comp Output
$F_{task}(\mathbf{x} \setminus \mathbf{r})$
Suff Output
$F_{task}(\mathbf{r})$
Task Output
$F_{task}(\mathbf{x})$
L_{faith}
L_{task}
Gold Label
y^*
Rationale Extractor
F_{ext}
Rationale
$\mathbf{r}$
L_{plaus}
Gold
Rationale $\mathbf{r}^*$

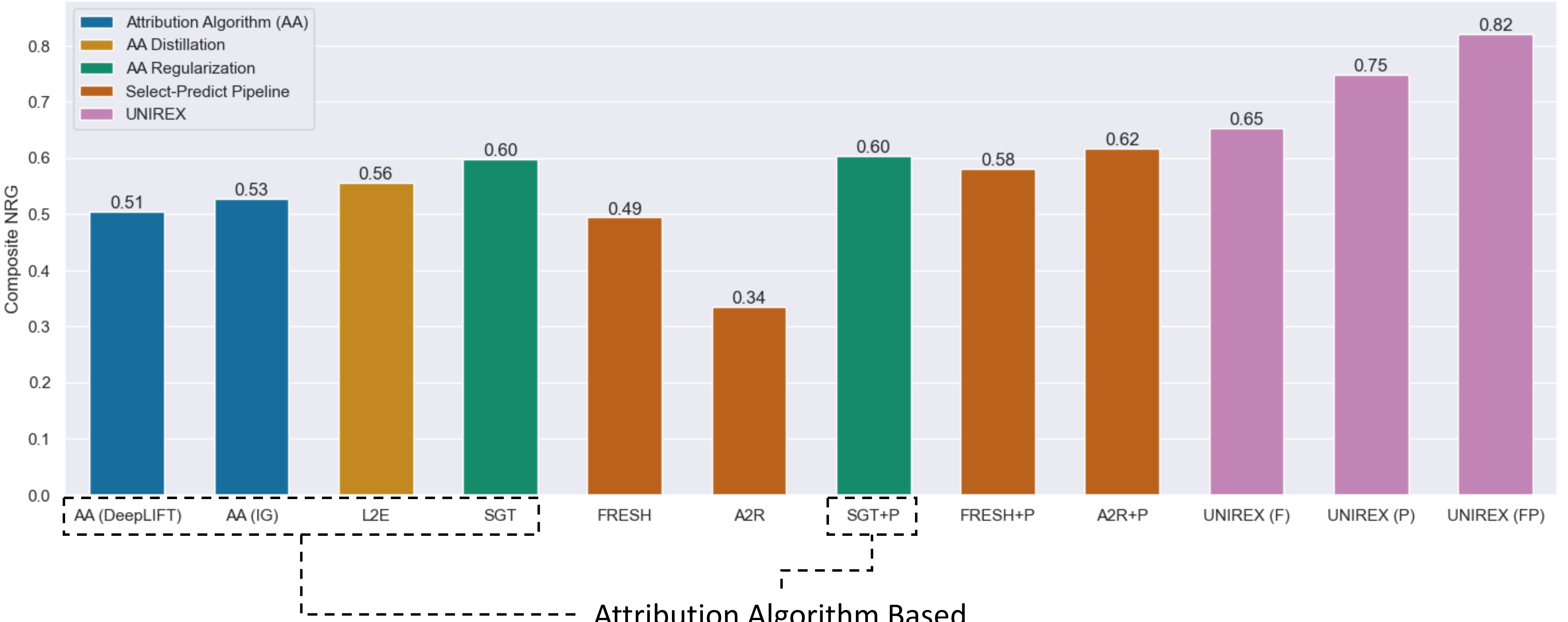
UNIREX


Comp Input
$\mathbf{x} \setminus \mathbf{r}$
Suff Input
$\mathbf{r}$
Task Input
$\mathbf{x}$
Task NLM
F_{task}
Comp Output
$F_{task}(\mathbf{x} \setminus \mathbf{r})$
Suff Output
$F_{task}(\mathbf{r})$
Task Output
$F_{task}(\mathbf{x})$
Rationale Extractor
F_{ext}
Rationale
$\mathbf{r}$
L_{faith}
+
L_{task}
+
L_{plaus}
Gold Label
y^*
Gold Rationale $\mathbf{r}^*$

Experiments: Main Results

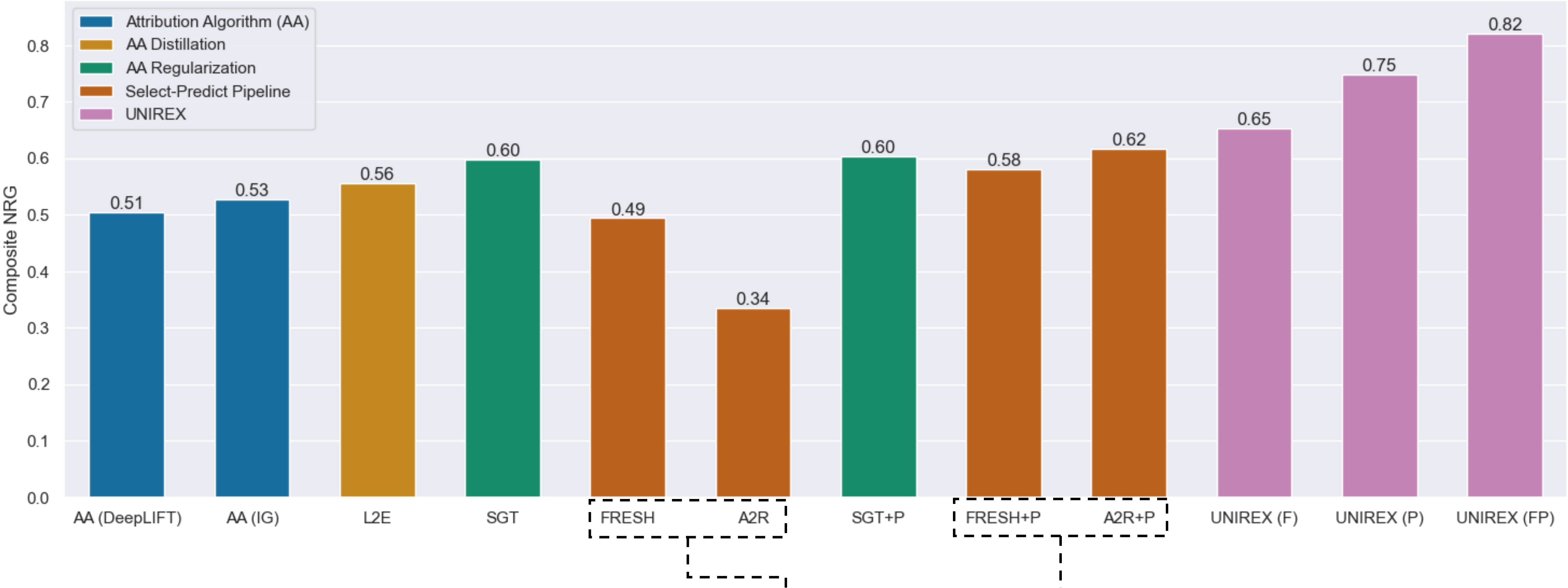


Experiments: Main Results



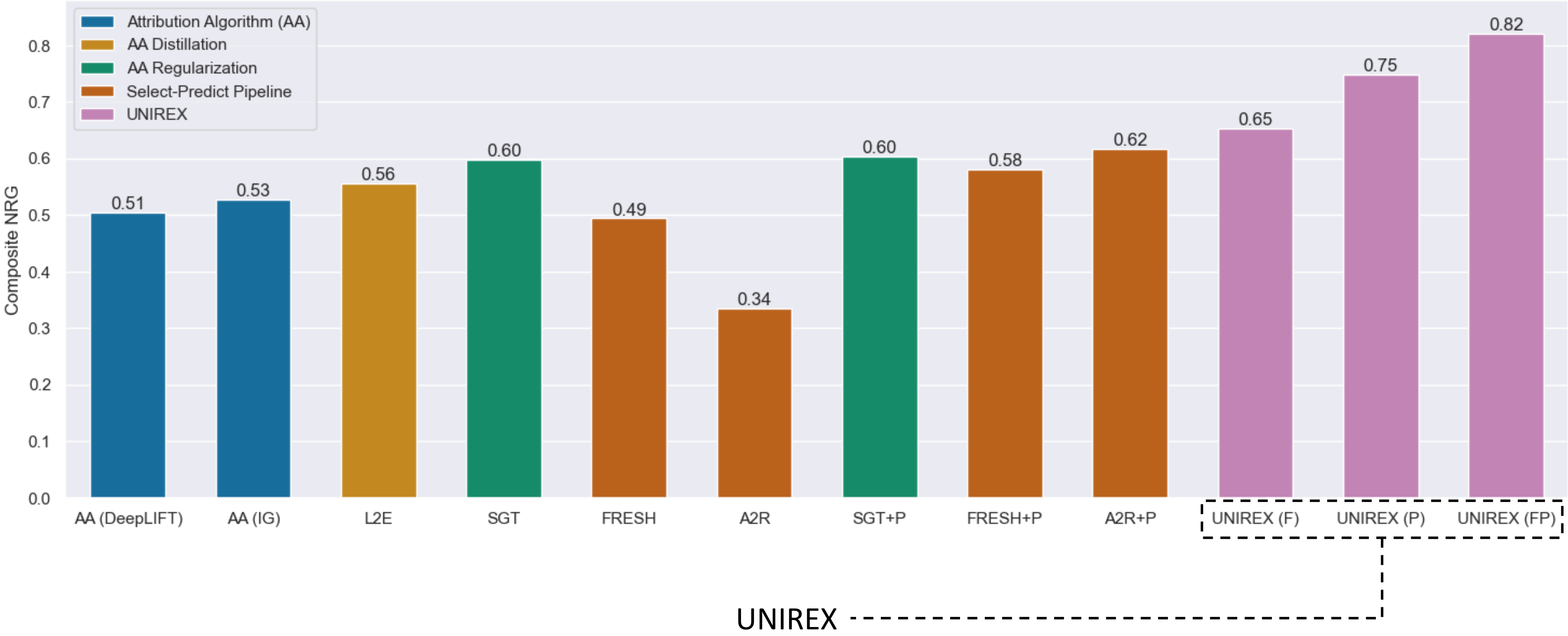
Attribution Algorithm Based

Experiments: Main Results



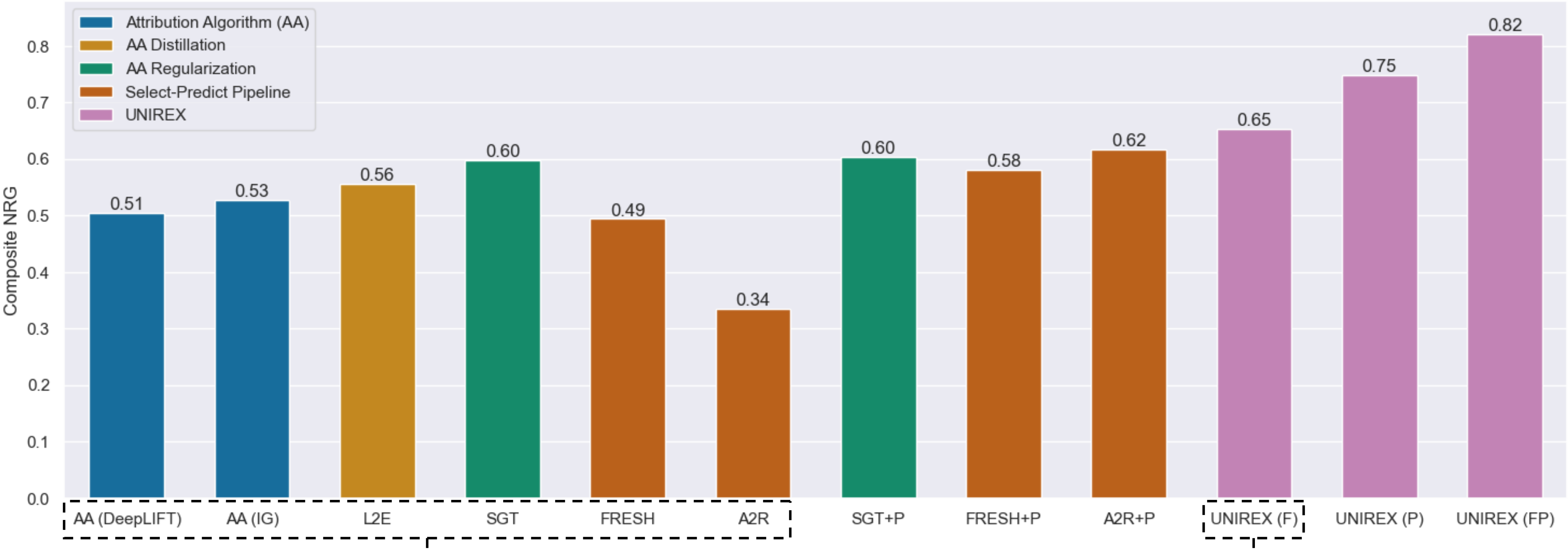
Select-Predict Pipeline

Experiments: Main Results



Experiments: Main Results

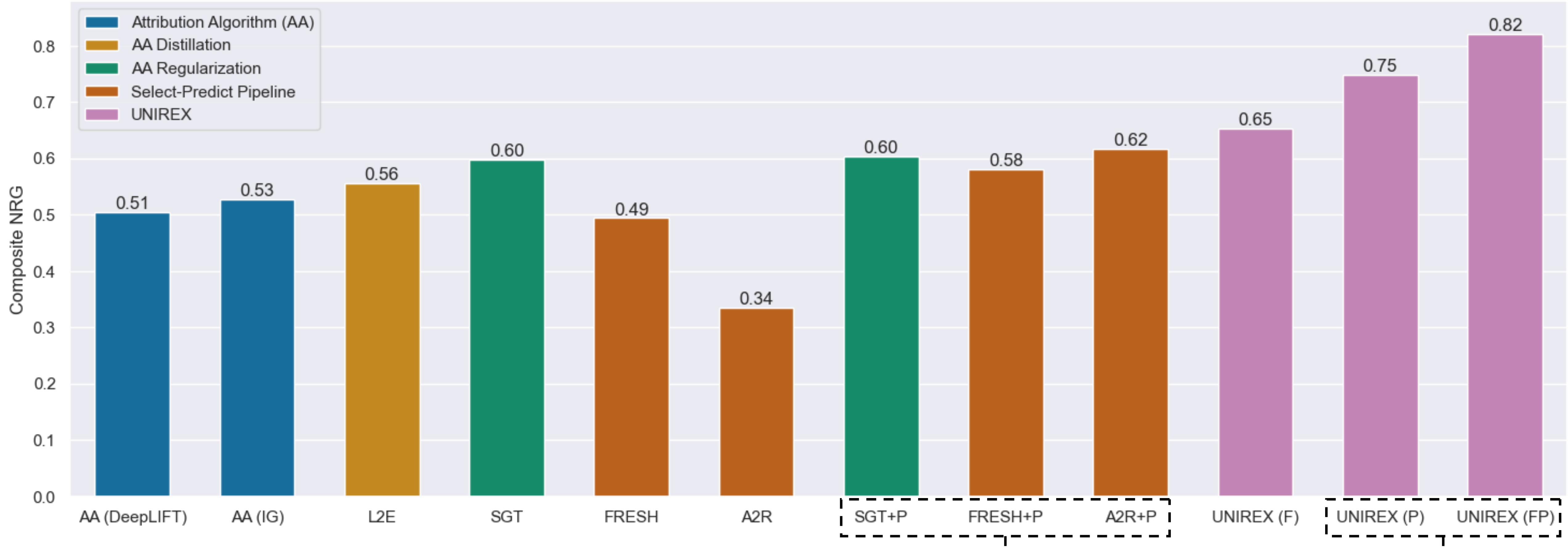
UNIREX (F) beats all baselines!



Not Optimized for Plausibility

Experiments: Main Results

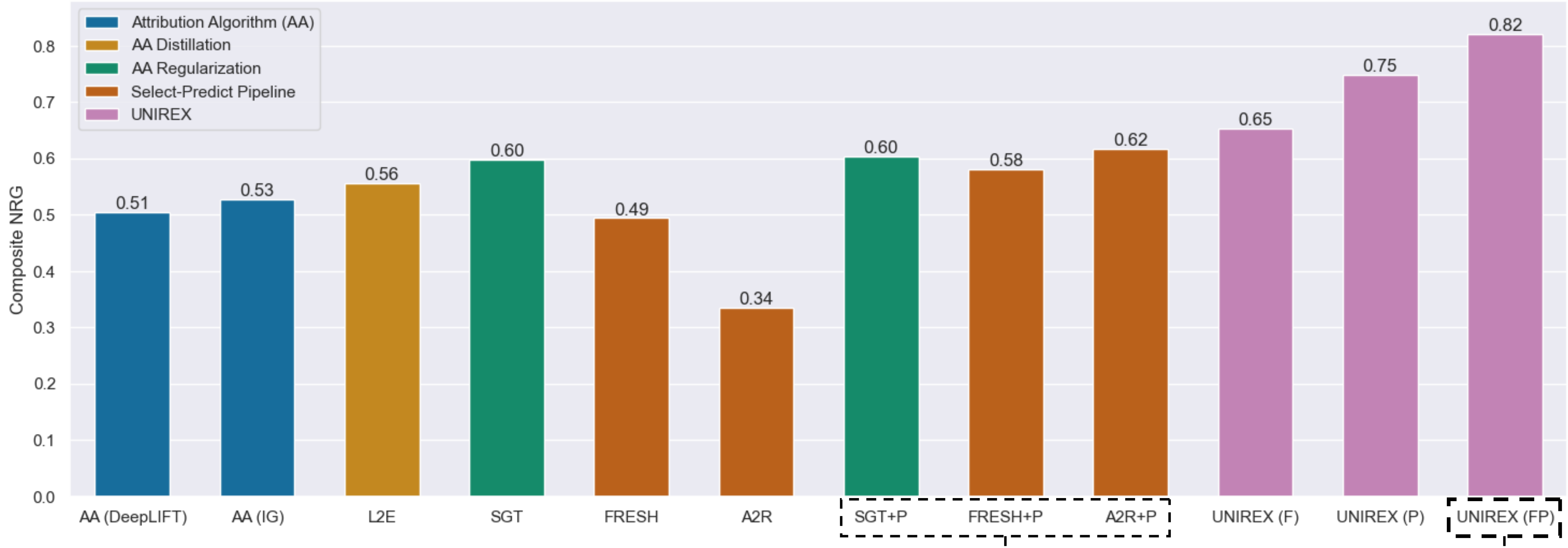
UNIREX (FP) beats all methods!



Optimized for Plausibility

Experiments: Main Results

UNIREX (FP) beats all methods!



Optimized for Plausibility

Faithfulness and plausibility are complementary objectives!

Thank You!