Errudite: Scalable, Reproducible, and Testable Error Analysis

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Marco Tulio Ribeiro
Microsoft Research

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington





Motivation & Contributions

Error analysis is important for…

Uncovering bugs

Improving the state-of-art

Safeguarding deployments

Where We Are

We performed an error analysis on a sample of 100 questions

Fader et tal.
ACL'13

Chen et al.
ACL'16

We randomly select 50 incorrect questions and categorize them into 6 classes.

We sample 100 incorrect predictions and try to find common error categories.

Wadhwa et al.
ACL'18

Where We Are

We performed an error analysis on a sample of 100 questions

Fader et tal.
ACL'13

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Chen et al.
ACL'16

We randomly select incorrect questions and categorize them into 6 classes.

We sample 100 incorrect predictions and try to find common error categories.

Wadhwa et al.
ACL'18

Where We Are

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Where We Are & Our Contribution

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Principles & Errudite

Precise & reproducible hypotheses

+

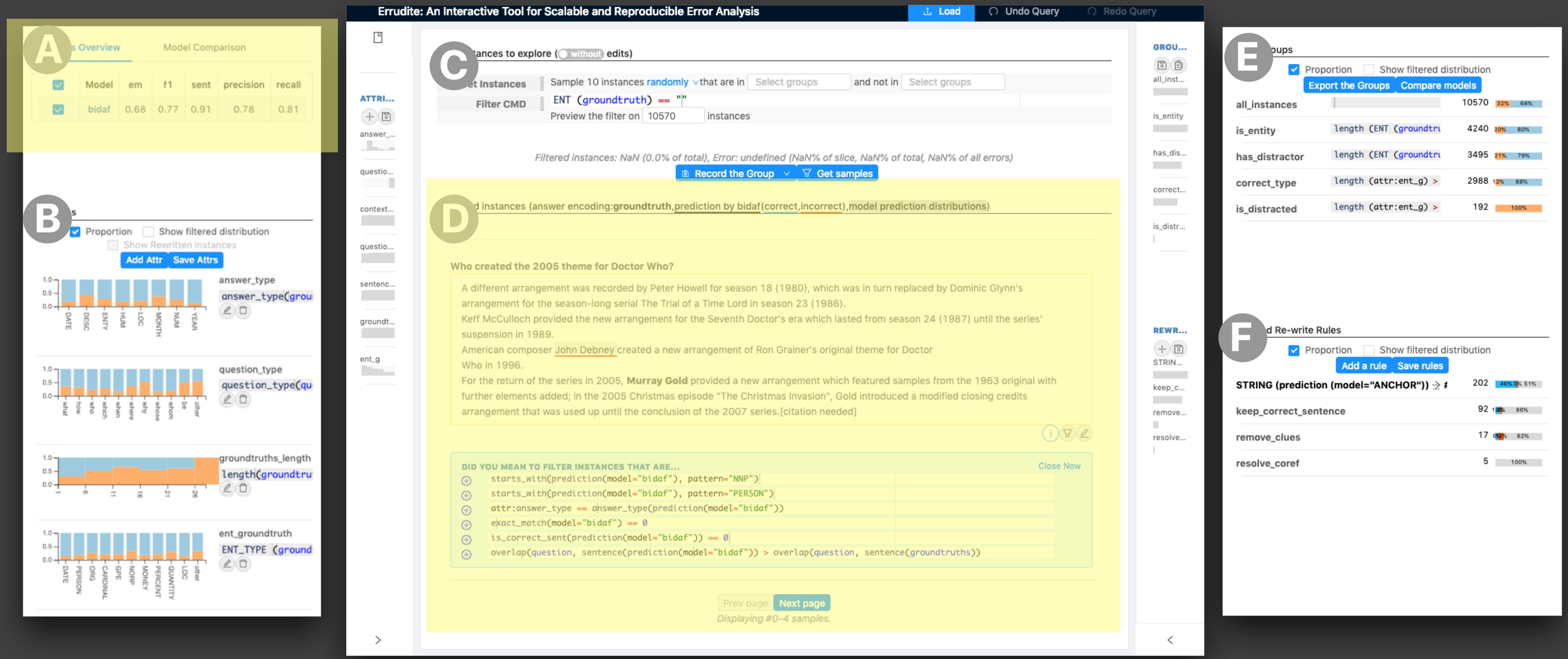
Scale up to the entire dev set

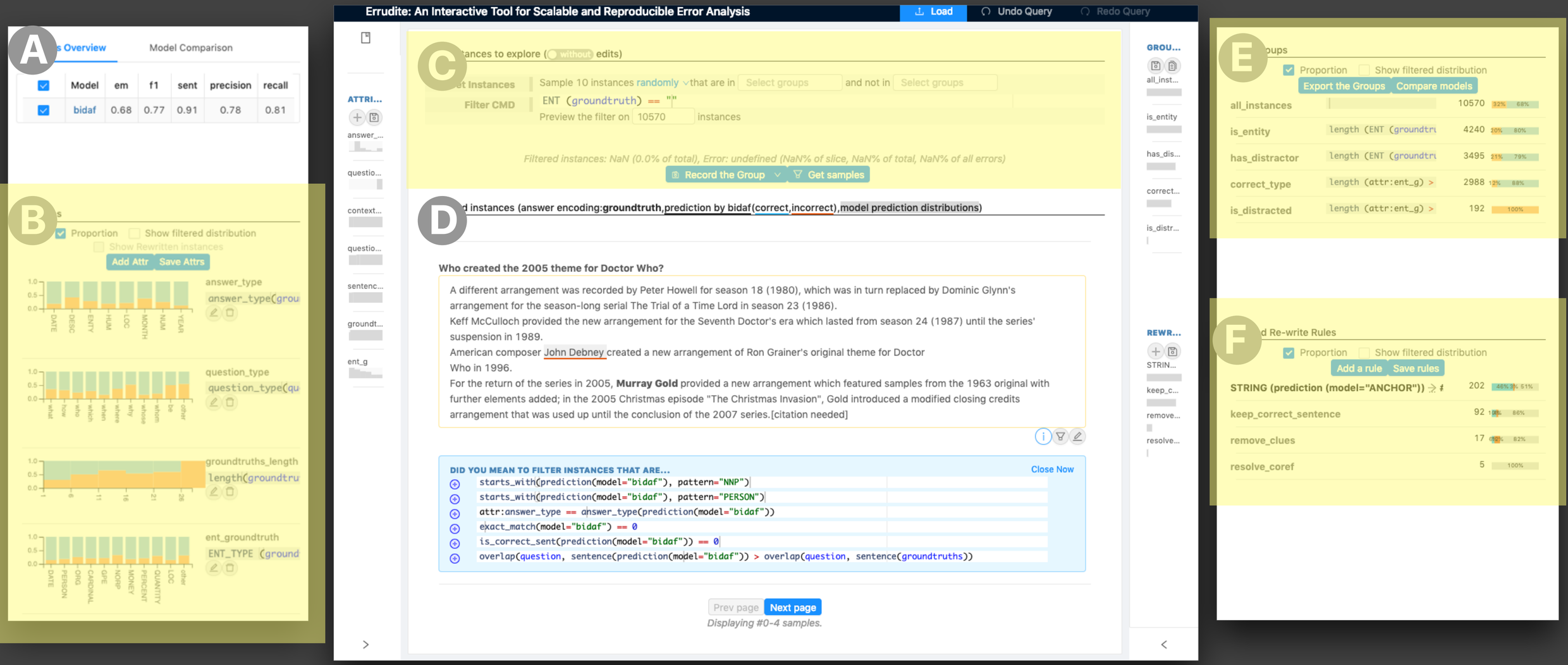
+

Cover errors & correct instances

+

Test via counterfactual analysis


A
Overview
Model Comparison
Model em f1 sent precision recall
bidaf 0.68 0.77 0.91 0.78 0.81
B
Proportion
Show filtered distribution
Show Rewritten instances
Add Attr
Save Attrs
answer_type
question_type
groundtruths_length
ent_groundtruth
Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis
Load
Undo Query
Redo Query
C
without
edits)
Sample 10 instances randomly that are in
Select groups
and not in
Select groups
Filter CMD
ENT (groundtruth) == ""
Preview the filter on
10570
instances
Filtered instances: NaN (0.0% of total), Error: undefined (NaN% of slice, NaN% of total, NaN% of all errors)
Record the Group
Get samples
D
instances (answer encoding:groundtruth,prediction by bidaf(correct,incorrect),model prediction distributions)
Who created the 2005 theme for Doctor Who?
A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, Murray Gold provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]
DID YOU MEAN TO FILTER INSTANCES THAT ARE...
Close Now
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf")) > overlap(question, sentence(groundtruths))
Prev page
Next page
Displaying #0-4 samples.
E
Proportion
Show filtered distribution
Export the Groups
Compare models
all_instances 10570
is_entity length (ENT (groundtru 4240
has_distractor length (ENT (groundtru 3495
correct_type length (attr:ent_g) > 2988
is_distracted length (attr:ent_g) > 192
F
Re-write Rules
Proportion
Show filtered distribution
Add a rule
Save rules
STRING (prediction (model="ANCHOR")) 202
keep_correct_sentence 92
remove_clues 17
resolve_coref 5



Video demo: https://tinyurl.com/errudite-video

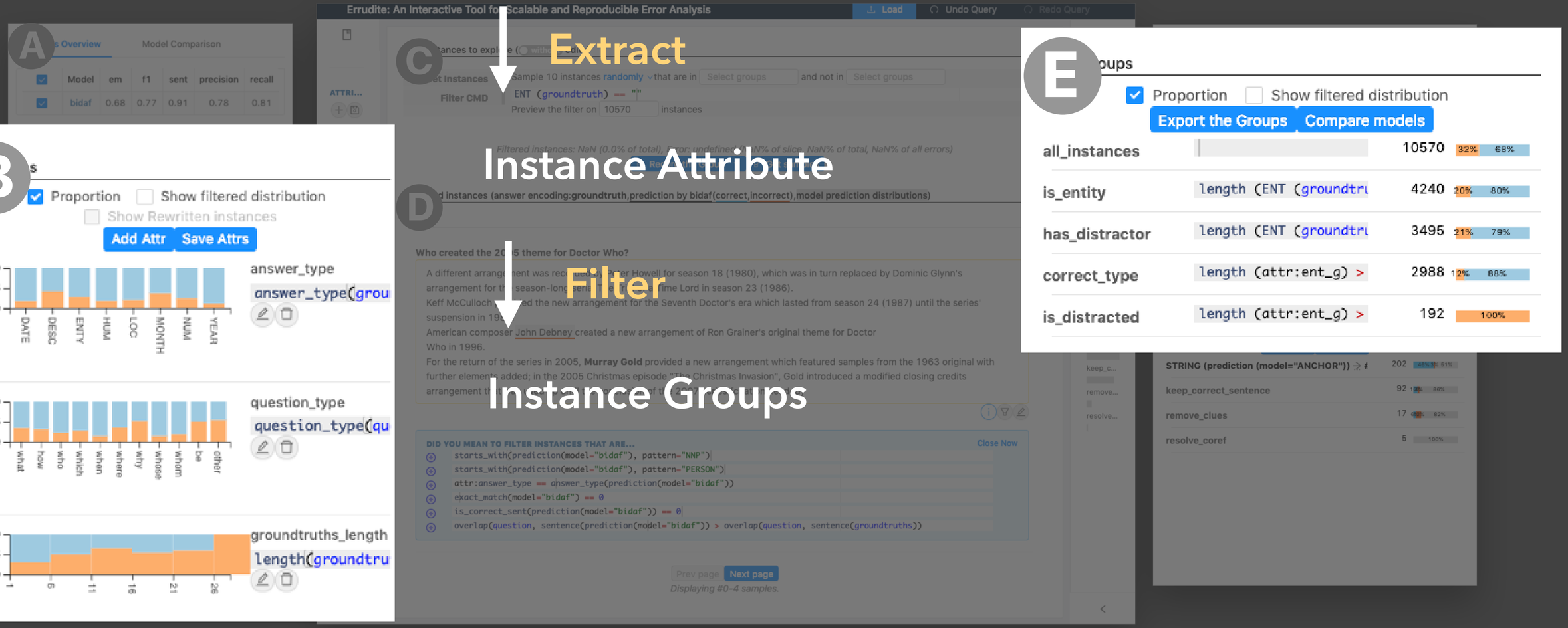
Core Design

Precise & Reproducible Domain Specific Language

Precise DSL (Domain Specific Language)

DSL = Attribute Extractor + Target + Operators

```
length(q) > 20
```



"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

Too ambiguous to reproduce

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

User study: What is imprecise answer boundaries?

"The model is making predictions with missing or additional words…?"

D1 No exact match, but high overlap

```
exact_match(p(m)) == 0
  and f1(p(m)) > 0.7
```

D2 Off by at most 2 tokens both on the left and right

```
exact_match(p(m)) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

User study: What is imprecise answer boundaries?

"The model is making predictions with missing or additional words…?"

D1 No exact match, but high overlap

```
exact_match(p(m)) == 0
  and f1(p(m)) > 0.7
```

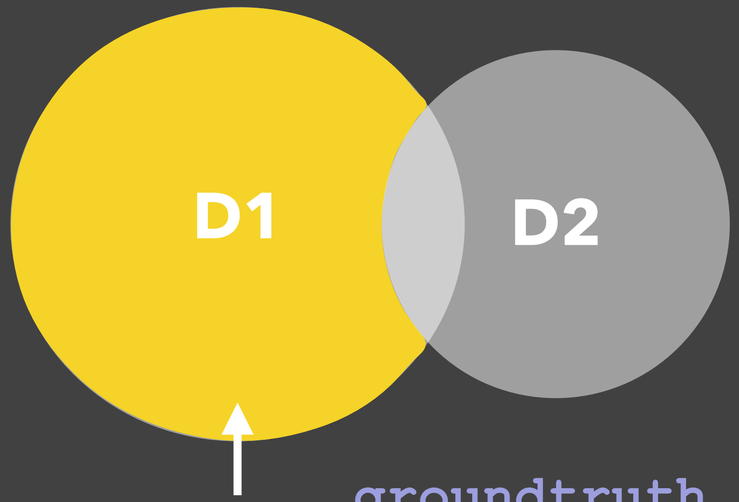
D2 Off by at most 2 tokens both on the left and right

```
exact_match(p(m)) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

User study: What is imprecise answer boundaries?

D1 No exact match, but high overlap

D2 Off by at most 2 tokens both on the left and right



groundtruth

…the **polynomial time hierarchy** collapses.

…believed that the polynomial hierarchy does..

prediction

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible hypotheses

Quantify instances with a domain specific language

+

Cover errors & correct instances

+

Test via counterfactual analysis

Design & Use Scenario

Examine the distractor hypothesis

on BiDAF (Seo et al., 2016), with SQuAD (10570 instances; Rajpurkar et al., 2016)

Independently tested by **4** (out of 10) participants in the user study

Scenario: distractor hypothesis

Common belief: BiDAF…

- **Matches entity types**
  Knows to find a *PERSON*
- **Finds the exact answer spans**
  Distracted by other *PERSON* spans

Who created the 2005 theme for Doctor Who?

…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement… featured sampled from the 1963 original.

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

100 << 2000+ errors in total

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

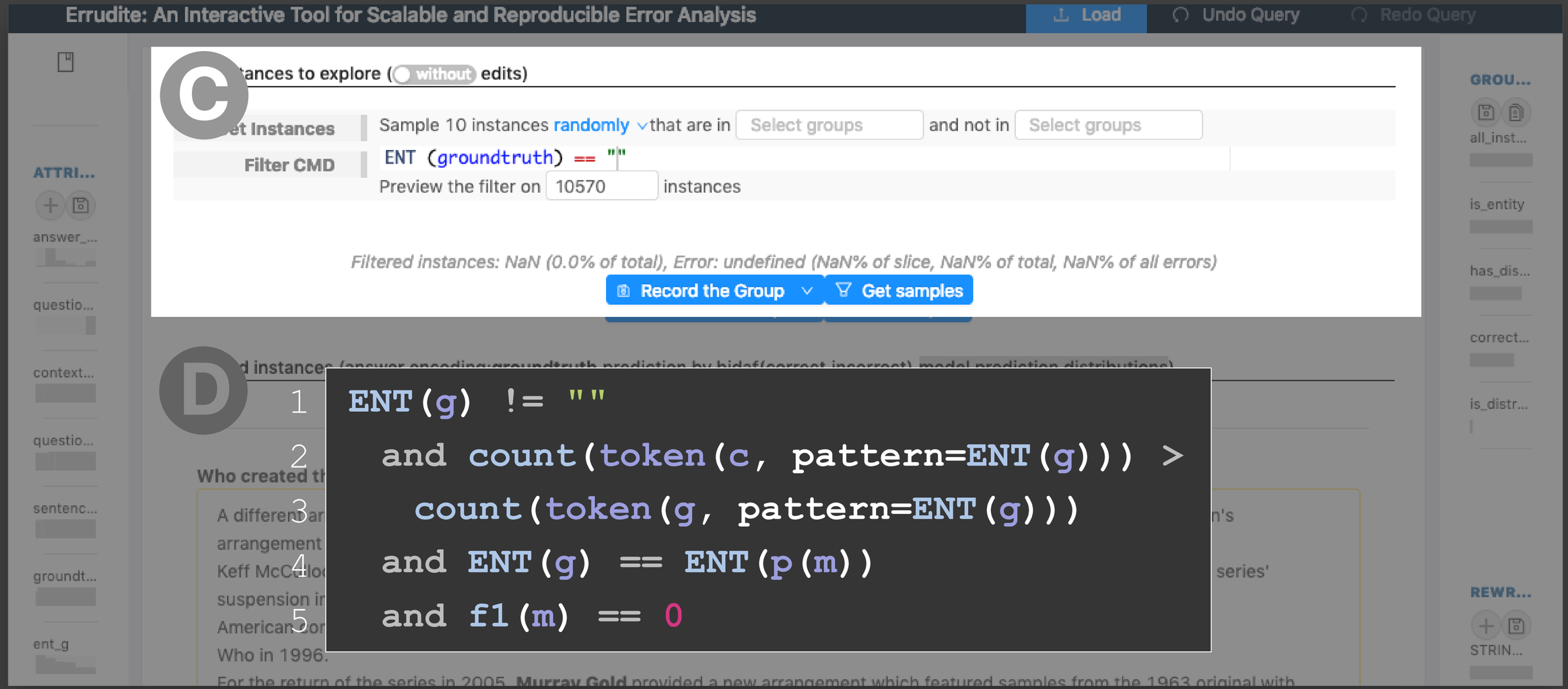
+

Cover errors & correct instances

+

Test via counterfactual analysis

Build distractor groups with DSL



Build distractor groups with DSL

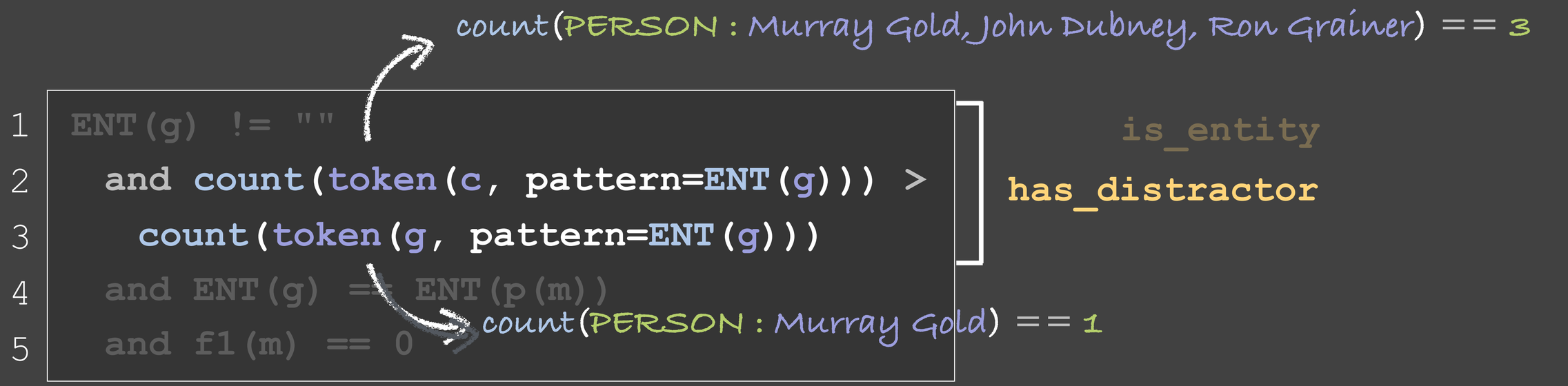
ENT(Murray Gold) == PERSON

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

is_entity

*"The **g**roundtruth is an **ENT**ity."*

Build distractor groups with DSL



*"There are more tokens matching the ground truth entity type (**ENT(g)**) in the whole **c**ontext than in the **g**roundtruth."*

Build distractor groups with DSL

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

ENT(John Debney) == PERSON

is_entity

has_distractor

correct_type

*"The **m**odel **p**rediction **ENT**ity type matches the **g**roundtruth **ENT**ity type."*

Build distractor groups with DSL

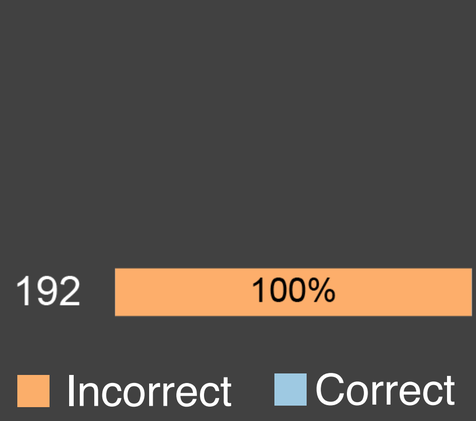
```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

is_entity
has_distractor
correct_type
is_distracted

"The model prediction is incorrect."

Build distractor groups with DSL

```
ENT(g) != ""
  and count(token(c, pattern=ENT(g))) >
    count(token(g, pattern=ENT(g)))
  and ENT(g) == ENT(p(m))
  and f1(m) == 0
```



5.7% of all BiDAF errors:

The distractor hypothesis seems correct!

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

Wrongly prioritize groups that are well-handled in average.

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

Wrongly prioritize groups that are well-handled in average.

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

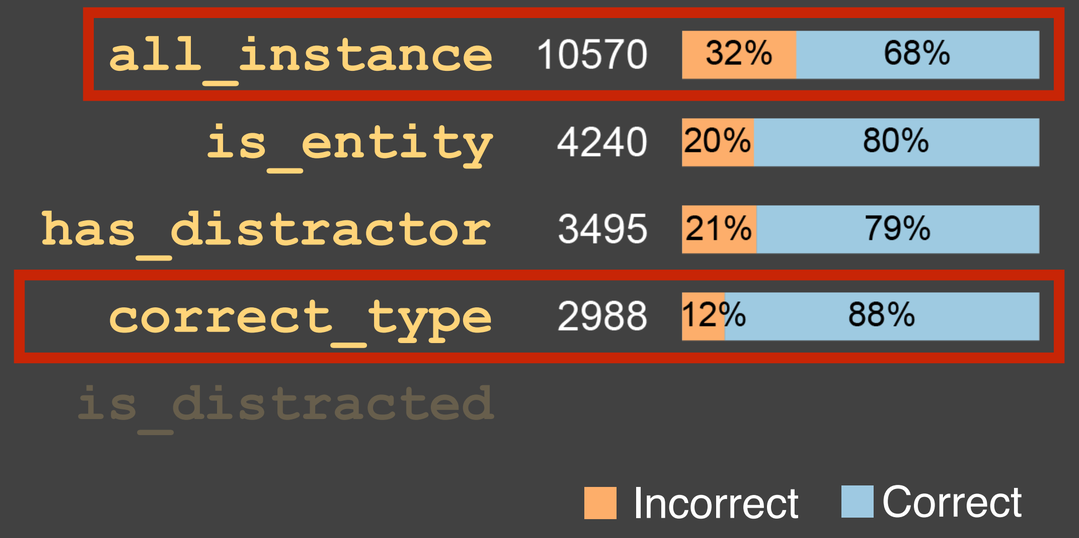
Cover errors & correct instances

+

Test via counterfactual analysis

Build distractor groups with DSL

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```



88% EM > 68% EM:

BiDAF performs better when have distractors & entity type is matched, than overall.

Reject / revise the hypothesis!

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

`is_distracted` 192 100%

HAS distractor prediction

!=

IS WRONG due to distractor prediction

Who created the 2005 theme for Doctor Who?

…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement… featured sampled from the 1963 original.

Distractor entity?

Multi-sentence reasoning?

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

HAS distractor prediction !=
IS WRONG due to distractor prediction

Errudite

Precise & reproducible hypotheses

+

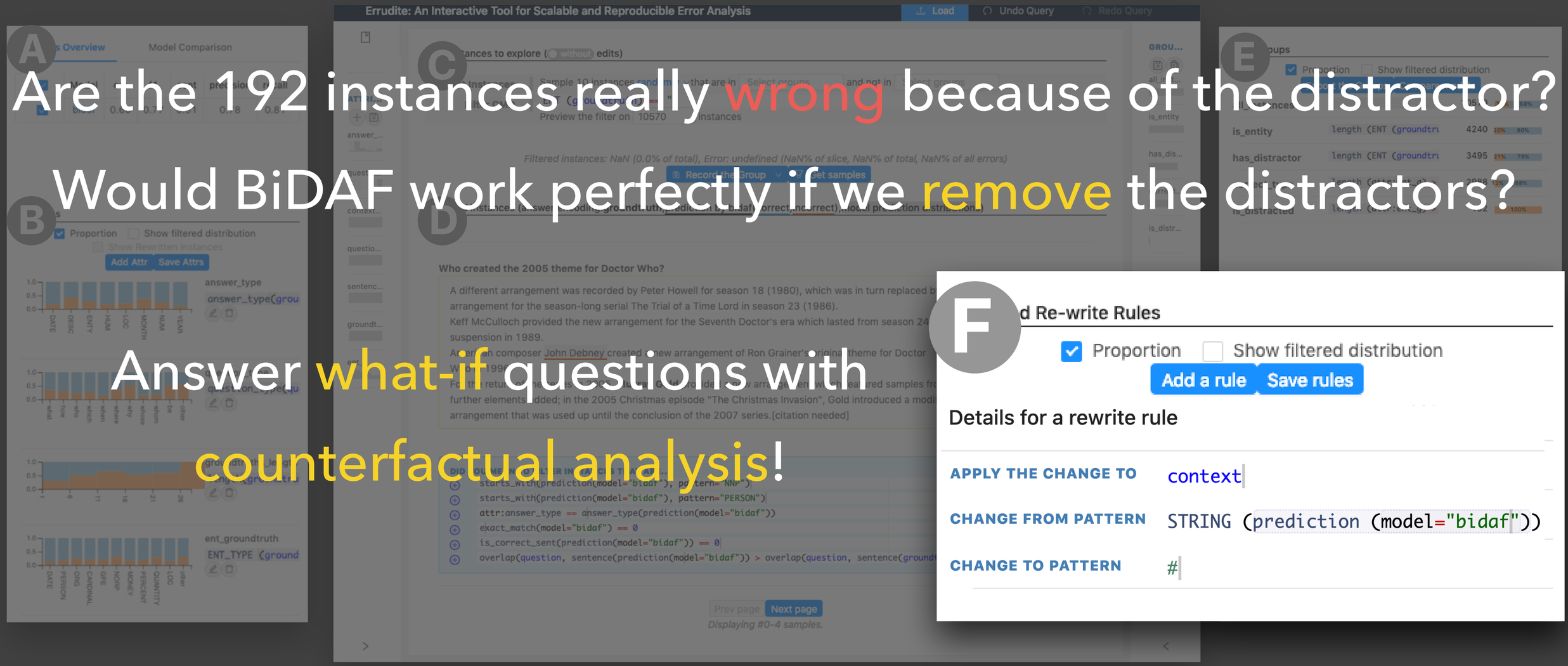
Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis



Are the 192 instances really wrong because of the distractor?

Would BiDAF work perfectly if we remove the distractors?

Answer what-if questions with counterfactual analysis!

Counterfactual Analysis with Rewrite Rules

`rewrite(`**`target , from → to`**`)`

Re-write the **target** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

`rewrite(target , from → to )`

Re-write the **target** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

`rewrite( c , from → to )`

Re-write the **context** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

`rewrite( c , string(p(m)) → to )`

Re-write the **context** part of an instance

by replacing the **model predicted distractor string**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

```
rewrite(   c     , string(p(m))   →   "#" )
```

Re-write the **context** part of an instance

by replacing the **model predicted distractor string**

with **a placeholder token** “#”

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

```
rewrite(   c    , string(p(m))   →   "#" )
```

Incorrect
Incorrect

Q: Who created the 2005 theme for Doctor Who?

C: …~~John Dobney~~ # created a new arrangement of Ron Grainer's …
Murray Gold provided a new arrangement…

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we remove the distractors?

```
rewrite(   c    , string(p(m))   →   "#" )
```

Incorrect
Incorrect

Another distractor is still confusing the model!

Counterfactual Analysis with Rewrite Rules

```
rewrite(   c    , string(p(m))   →   "#" )
```

`p(m)` for the 192 rewritten `is_distracted` instances are…

Incorrect
Incorrect

Another distractor is still confusing the model!

Counterfactual Analysis with Rewrite Rules

`rewrite( c , string(p(m)) → "#" )`

`p(m)` for the 192 rewritten `is_distracted` instances are…

Incorrect → Incorrect — 29% — ✓ Another distractor is still confusing the model!

Incorrect → Correct — 48% — ✓ The distractor was fooling the model!

Unchanged — 23% — ✗ Other factors are at play!

age of 18, ~~10.5%~~ # from 18 to 24…

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…	**Errudite**
Subjectively defined hypotheses	Precise & reproducible hypotheses
+	+
Small samples	Scale up to the entire dev set
+	+
Focus exclusively on errors	Cover errors & correct instances
+	+
No Test on true cause	Test via counterfactual analysis

Deliveries: Precise + Reproducible + Re-applicable

Attribute

```
ENT(g)
```

Groups

```
all_instance
is_entity
has_distractor
correct_type
is_distracted
```

Rewrite rule

```
rewrite(
   c,
   string(p(m))→"#")
```

Deliveries: Precise + Reproducible + Re-applicable

Attribute + Groups + Rewrite rule



applied to…

BiDAF is …

not particularly bad at distractors.

Seemingly distractor errors can be due to other factors.

Deliveries: Precise + Reproducible + Re-applicable

Attribute + Groups + Rewrite rule



Re-applied to…

Other datasets & Other models…

? at handling distractor.

User Study

10 participants = NLP graduate students + QA engineers from industry

Examine BiDAF (Seo et al., 2016) on SQuAD (Rajpurkar et al., 2016)

One hour section: Replicate prior error analysis + Freely explore the model

User Feedback: Did they like Errudite?

Enhanced their error analysis experience.

Systematically scaled up the analysis.

Precise and inspiring more confidence.

Much faster exploration.

Biased conclusion due to…

Subjectively defined hypotheses

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Principles & Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

Errudite improves error analysis…

Precise

Reproducible

Scalable

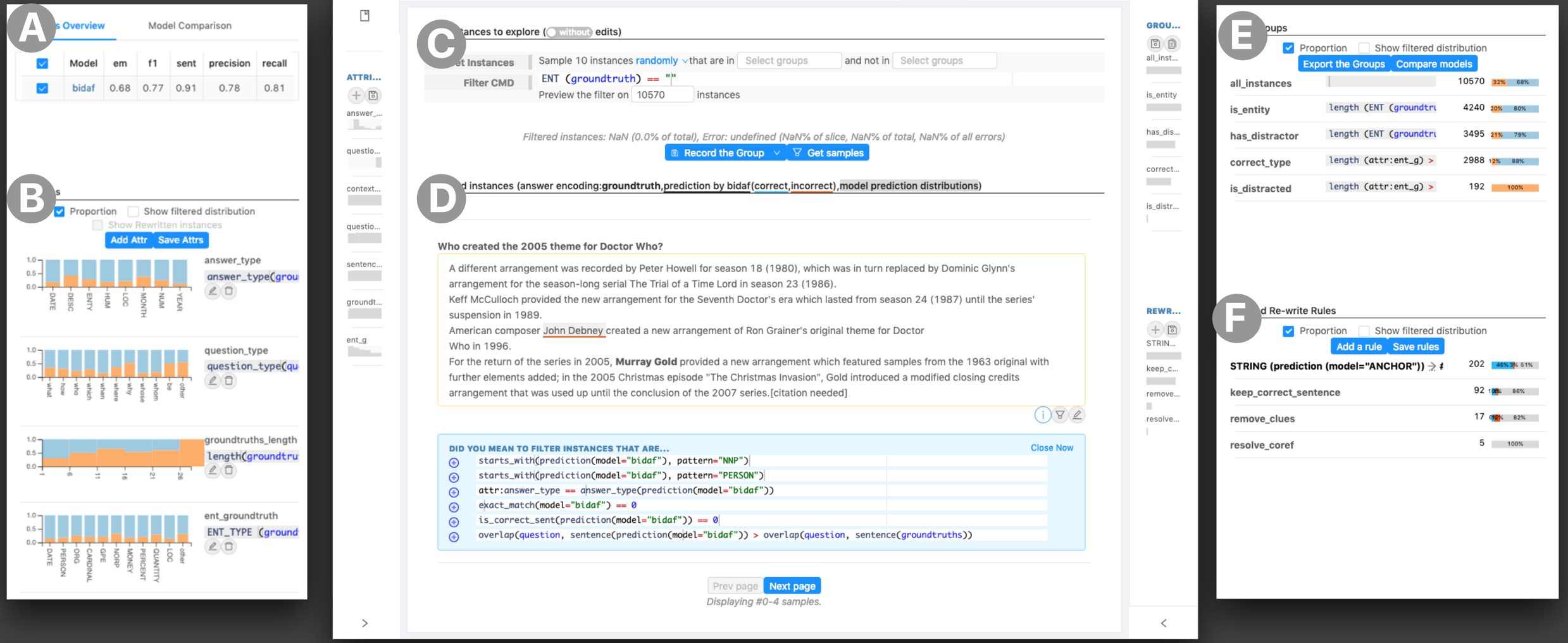
Testable

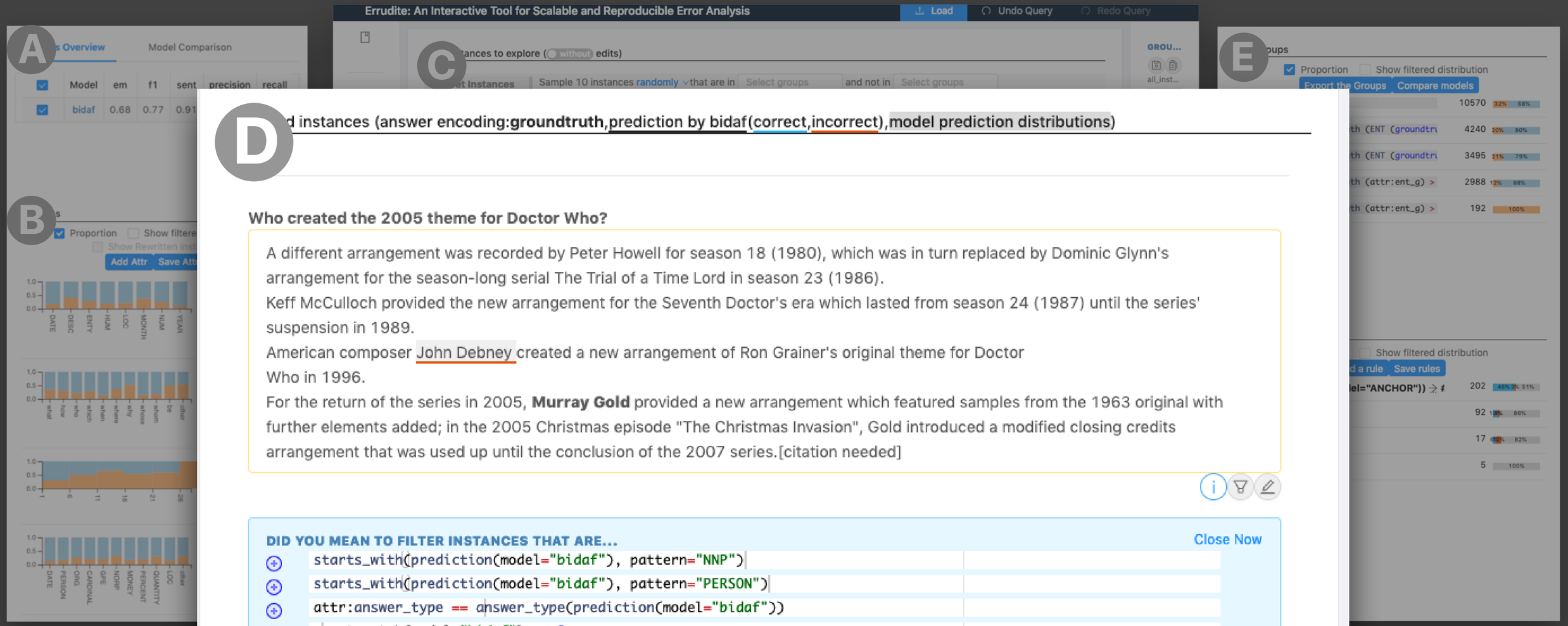
Better error analysis will then…

Uncover bugs

Improve the state-of-art

Safeguard deployments


A
Overview
Model Comparison
Model em f1 sent precision recall
bidaf 0.68 0.77 0.91 0.78 0.81
B
Proportion
Show filtered distribution
Show Rewritten instances
Add Attr
Save Attrs
answer_type
answer_type(grou
DATE DESC ENTY HUM LOC MONTH NUM YEAR
question_type
question_type(qu
what how who which when where why whose whom be other
groundtruths_length
length(groundtru
ent_groundtruth
ENT_TYPE (ground
DATE PERSON ORG CARDINAL GPE NORP MONEY PERCENT QUANTITY LOC other
Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis
Load
Undo Query
Redo Query
ATTRI...
answer_...
questio...
context...
questio...
sentenc...
groundt...
ent_g
C
Instances to explore (without edits)
Sample 10 instances randomly that are in
Select groups
and not in
Select groups
Filter CMD
ENT (groundtruth) == ""
Preview the filter on 10570 instances
Filtered instances: NaN (0.0% of total), Error: undefined (NaN% of slice, NaN% of total, NaN% of all errors)
Record the Group
Get samples
D
instances (answer encoding:groundtruth,prediction by bidaf(correct,incorrect),model prediction distributions)
Who created the 2005 theme for Doctor Who?
A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, Murray Gold provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]
DID YOU MEAN TO FILTER INSTANCES THAT ARE...
Close Now
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf")) > overlap(question, sentence(groundtruths))
Prev page
Next page
Displaying #0-4 samples.
GROU...
all_inst...
is_entity
has_dis...
correct...
is_distr...
REWR...
STRIN...
keep_c...
remove...
resolve...
E
oups
Proportion
Show filtered distribution
Export the Groups
Compare models
all_instances 10570 32% 68%
is_entity length (ENT (groundtru 4240 20% 80%
has_distractor length (ENT (groundtru 3495 21% 79%
correct_type length (attr:ent_g) > 2988 12% 88%
is_distracted length (attr:ent_g) > 192 100%
F
d Re-write Rules
Proportion
Show filtered distribution
Add a rule
Save rules
STRING (prediction (model="ANCHOR")) ≠ 202 46% 3% 51%
keep_correct_sentence 92 14% 86%
remove_clues 17 18% 82%
resolve_coref 5 100%


A
Overview
Model Comparison
Model em f1 sent precision recall
bidaf 0.68 0.77 0.91
B
Proportion
Add Attr
Save Att
C
Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis
Load
Undo Query
Redo Query
Sample 10 instances randomly that are in Select groups and not in Select groups
D
d instances (answer encoding:groundtruth,prediction by bidaf(correct,incorrect),model prediction distributions)
Who created the 2005 theme for Doctor Who?
A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, Murray Gold provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]
DID YOU MEAN TO FILTER INSTANCES THAT ARE...
Close Now
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf")) > overlap(question, sentence(groundtruths))
E
Proportion
Show filtered distribution
Export the Groups
Compare models
10570
4240
3495
2988
192
Save rules
202
92
17
5



How many brownish peaks are there?

groundtruth:**2** *(* 10)*

saaa:4 | vqacounting:5

DID YOU MEAN TO FILTER INSTANCES THAT ARE... Close Now

A

```
starts_with(question, pattern="how many ADJ")
starts_with(question, pattern="ADV ADJ ADJ")
attr:question_type == "how many"
```

See more general suggestions?

Future Work

More sophisticated suggestions

Collaborative error analysis

Model comparison

Extend to more domains

Thank you!

https://tinyurl.com/errudite-video

Come talk to me if you want a (small) **live** demo!!






Backup

Qualitative Description

The model is bad on long questions

questions with more than N tokens



Refer to

Qualitative Description

The model is bad on long questions

questions with more than 20 tokens

Precise **DSL** (Domain Specific Language)

Qualitative Description The model is bad on long questions

questions with more than 20 tokens

Quantitative Description

```
length(q) > 20
```



Translate with **DSL**

Precise DSL (Domain Specific Language)

Qualitative Description The model is bad on long questions
questions with more than 20 tokens

Quantitative Description `length(q) > 20`



Attribute Extractor	Target	Operators
length	question	>
question_type	context	!=
answer_type	groundtruth	in
	prediction (model)	has_any
	token	
	sentence	

DSL (Domain Specific Language) **Attribute Extractor**

Basic Attributes	`Length`
General purpose linguistic features	`LEMMA,POS,ENT`
Standard prediction performance metrics	`f1,accuracy`
Between-target relations	`overlap(t1, t2)`
Domain-specific attribute	`answer_type,question_type`

DSL (Domain Specific Language) **Attribute Extractor**

Who created the 2005 theme for Doctor Who?

…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement... featured sampled from the 1963 original.

Function Name	Definition	DSL Code and resulting Output Values
`sentence`, `token`	Extractors for desired spans from targets — sentences or sub-phrases.	`sentence(g)`→`For the return of...` `token(c,pattern="PERSON")`→`[John,Debney,...]`
`exact_match`, `f1`, `is_correct_sent`	Performance functions that measure different levels of correctness.	`f1(m) == 0`, `exact_match(m) == 0` `is_correct_sent(m) == False`
`length`	Length of the target.	`length(q)== 9`, `length(g)== 2`
`POS`, `ENT`, `LEMMA`	Tokens in the target that have certain patterns of POS tags, named entity, etc.	`ENT(g,get_root=True)=="PERSON"`
`has_pattern`, `starts_with`, `ends_with`	To check whether the target contains certain pattern. `pattern` automatically detects queries on POS tags and entity types.	`starts_with(q,pattern="who VBZ")==True` `has_pattern(g,pattern="PERSON")==True`
`overlap(t1,t2)`	The ratio of tokens in `t1` that also occur in `t2`.	`overlap(q, sentence(g)) == 0.25`

Table 1: Definitions for a subset of attribute extractors, including sample values from the example in Fig. 2

Suggestions via programming-by-demonstration

Who created the 2005 theme for Doctor Who?

…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement... featured sampled from the 1963 original.

```
starts_with(p(m),pattern="NNP"))
starts_with(p(m),pattern="PERSON"))
answer_type(g) == answer_type(p(m))
exact_match(m) == 0
is_correct_sent(m) == False
overlap(q, sentence(p(m))) > overlap(q, sentence(g))
```

Suggestions via programming-by-demonstration

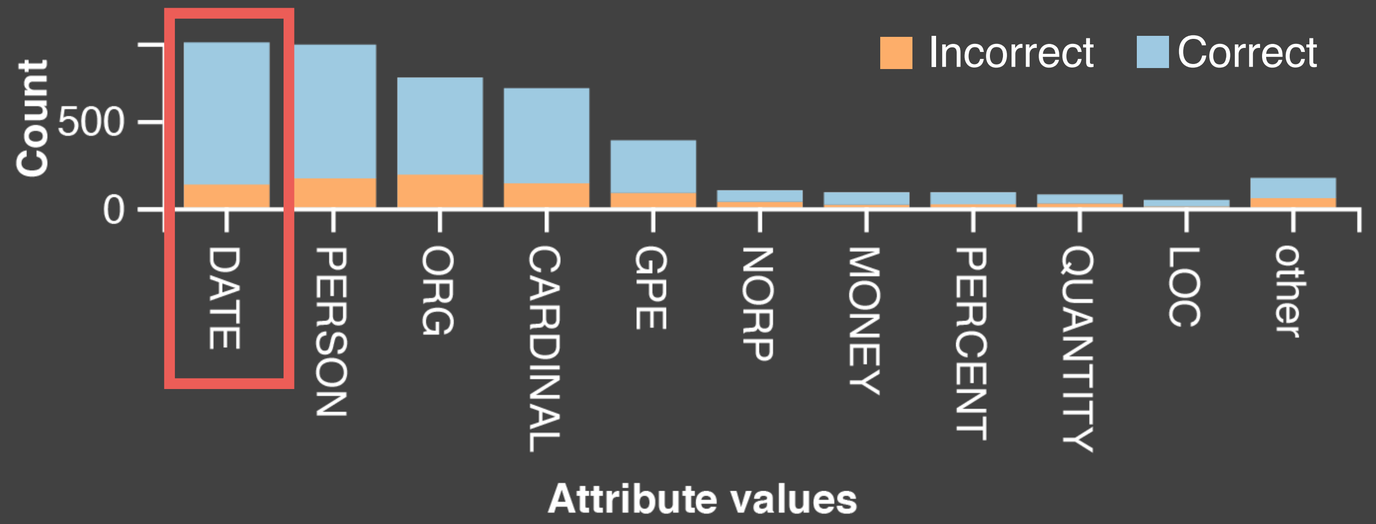
~~Who~~ **What person** created the 2005 theme for Doctor Who?

DID YOU WANT TO GENERALIZE TO...

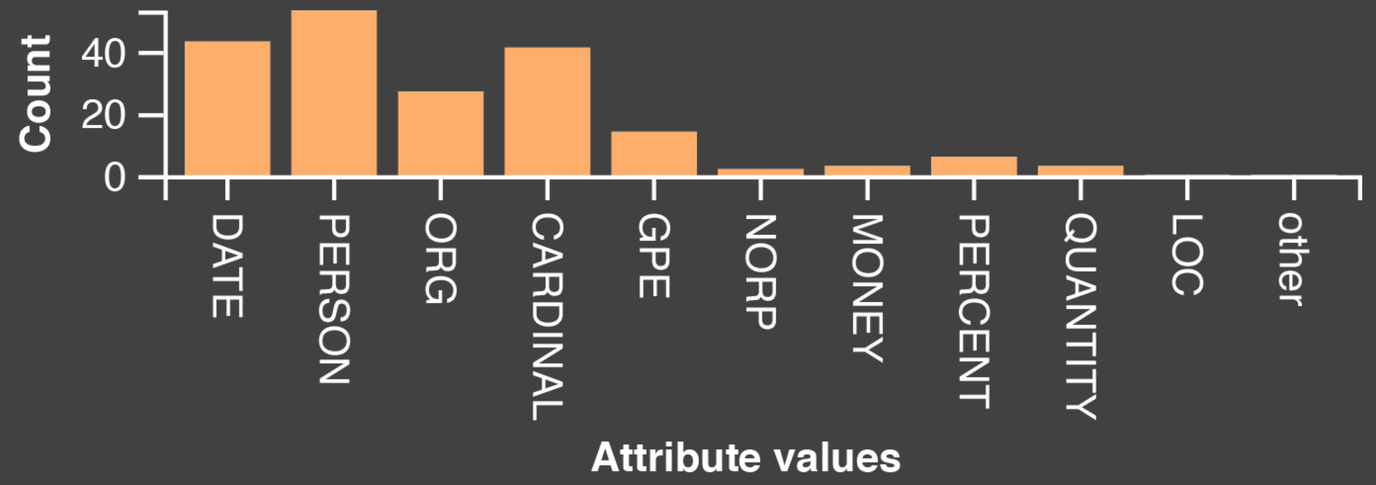
who → what person	☐ keep
who created → what person created	☐ keep
who VERB → what person VERB	☑ keep
who created the → what person created the	☐ keep
who VERB DET → what person VERB DET	☐ keep

Check attribute distribution `ENT(g)` in groups

`is_entity`

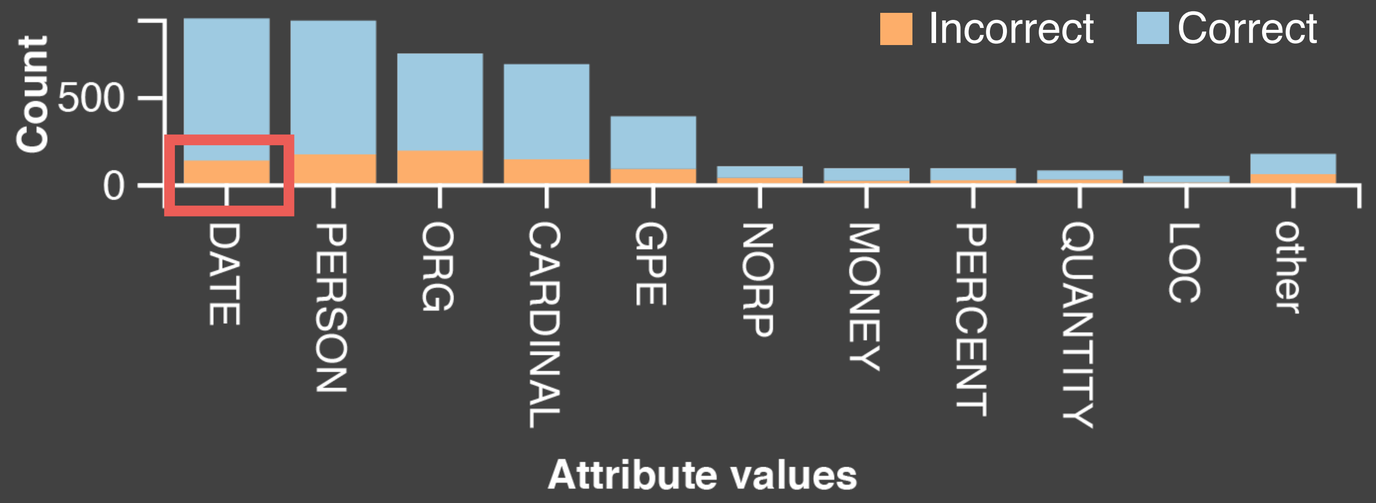


`is_distracted`

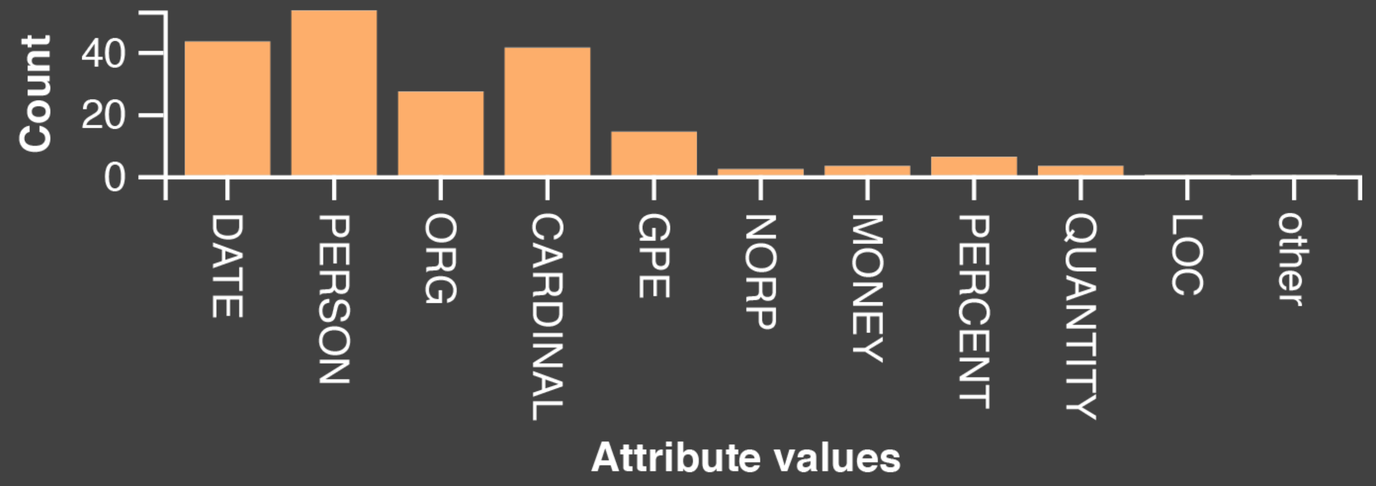


Check attribute distribution `ENT(g)` in groups

`is_entity`

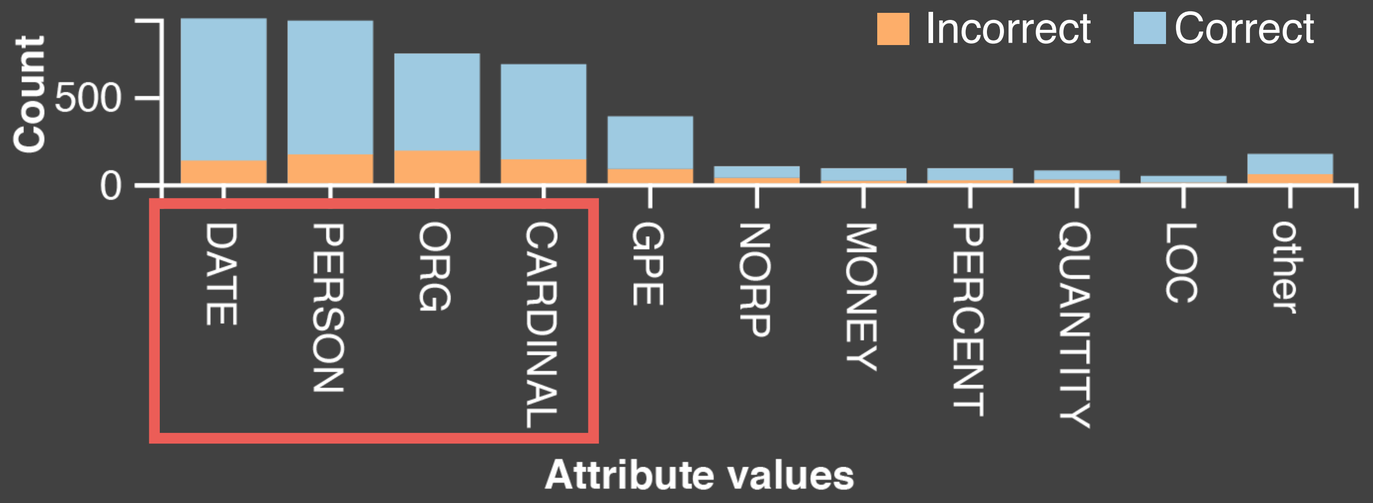


`is_distracted`

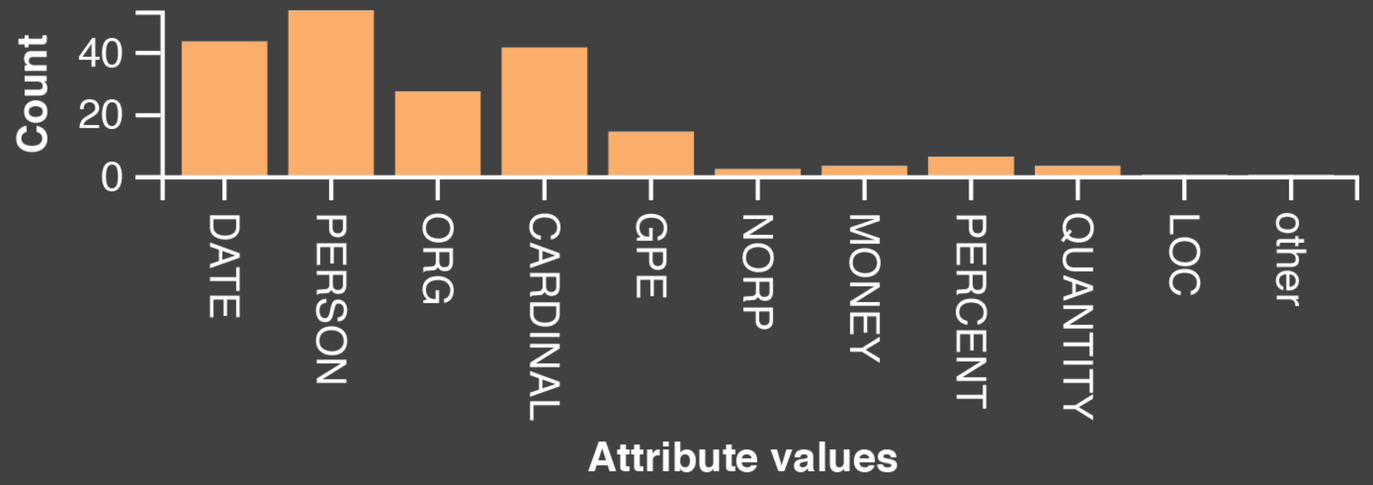


Check attribute distribution `ENT(g)` in groups

`is_entity`



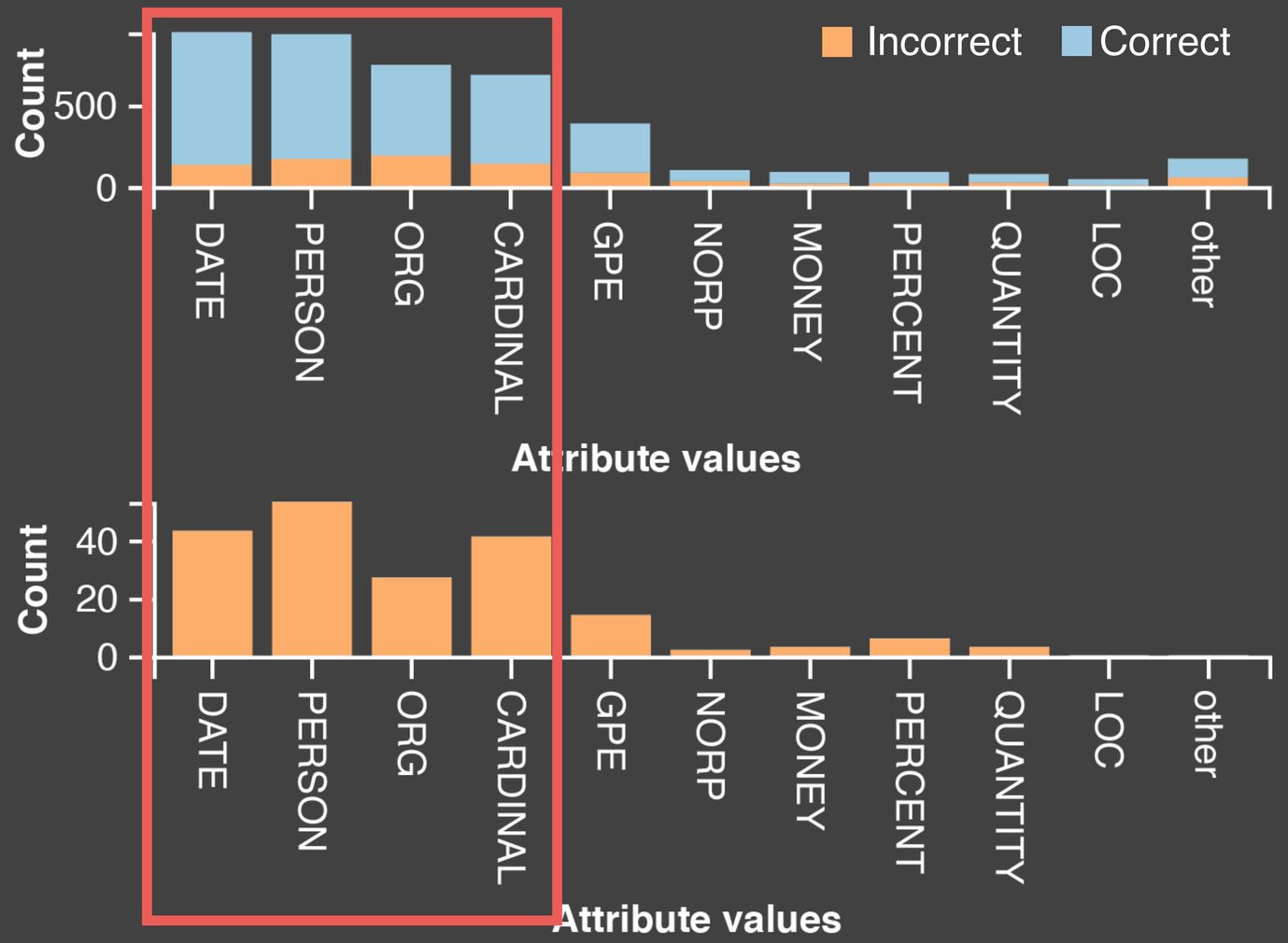
`is_distracted`



Check attribute distribution `ENT(g)` in groups

`is_entity`

`is_distracted`



User Study

10 participants = NLP graduate students + QA engineers from industry

Examine BiDAF (Seo et al., 2016) on SQuAD (Rajpurkar et al., 2016)

One hour section: Replicate prior error analysis + Freely explore the model

Study #1 Replication: Errudite flexible enough?

Read BiDAF error analysis: 50 errors, hand-labeled into 6 classes

↓

Recreate 4 classes with Errudite on the entire dataset

↓

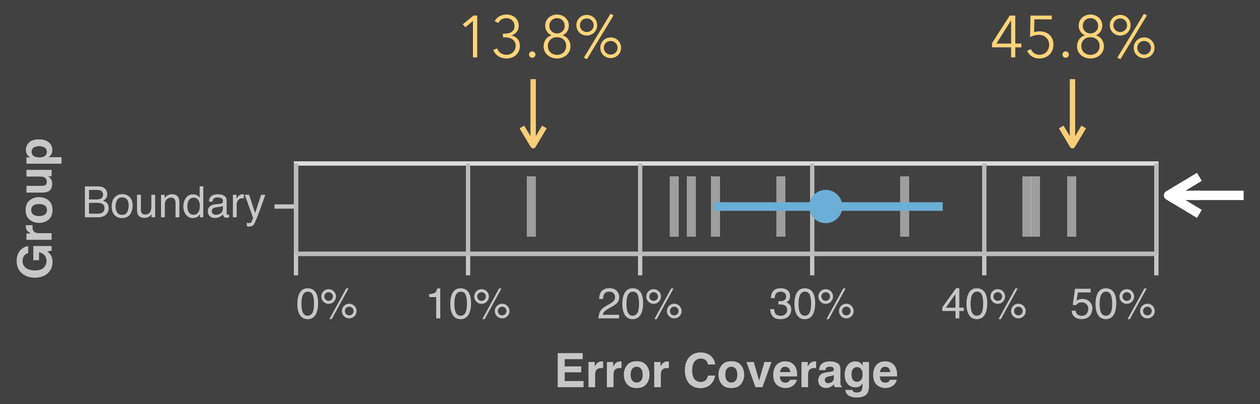
Rate closeness: Recreated groups == originals?

semantic

Users were able to express their intended groups well with Errudite.

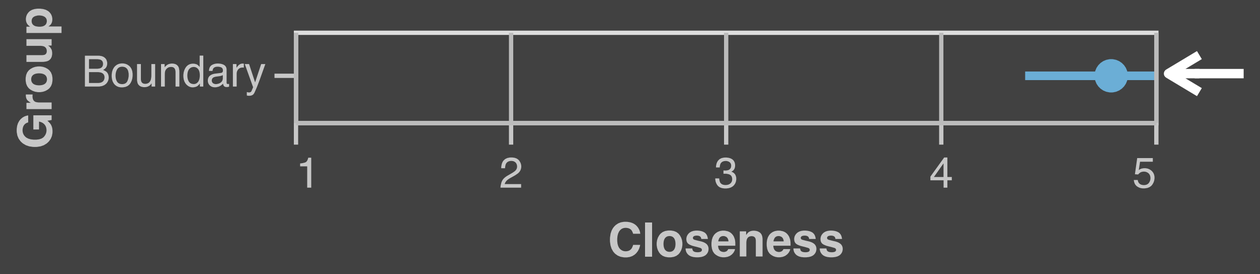
Study #1 Replication: Easy group ≠ reproducible!

How many errors are **covered** by user-built *Imprecise Error Boundary*?



← **Groups with low inter-agreement!**

How **close** does the approximation match the paper definition?



← **Most confident, an easy group**

Study #1 Replication: Easy group ≠ reproducible!

Error type	Ratio (%)	Example
Imprecise answer boundaries	50	**Context**: "The Free Movement of Workers Regulation articles 1 to 7 set out the main provisions on equal treatment of workers." **Question**: "Which articles of the Free Movement of Workers Regulation set out the primary provisions on equal treatment of workers?" **Prediction**: "1 to 7", **Answer**: "articles 1 to 7"

D1 Off by at most 2 tokens both on the left and right

```
exact_match(m) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

Coverage = 22.1%

D2 No exact match, but high overlap

```
exact_match(m) == 0
  and f1(m) > 0.7
```

Coverage = 13.8%

Study #1 Replication: Easy group ≠ reproducible!

Error type	Ratio (%)	Example
Imprecise answer boundaries	50	**Context**: "The Free Movement of Workers Regulation articles 1 to 7 set out the main provisions on equal treatment of workers." **Question**: "Which articles of the Free Movement of Workers Regulation set out the primary provisions on equal treatment of workers?" **Prediction**: "1 to 7", **Answer**: "articles 1 to 7"

D1 Off by at most 2 tokens both on the left and right

```
exact_match(m) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

Coverage = 22.1%

D2 No exact match, but high overlap

```
exact_match(m) == 0
  and f1(m) > 0.7
```

Coverage = 13.8%

Study #1 Replication: Easy group ≠ reproducible!

D1 Off by at most 2 tokens both on the left and right

`Coverage = 22.1%`

D2 No exact match, but high overlap

`Coverage = 13.8%`

D1 D2 …**commercial, scientific, and cultural** growth…

D1 D2 …**from Karakorum** in Mongolia to Khanbaliq…

D1 D2 …the **polynomial time hierarchy** collapses.
…believed that the polynomial hierarchy does..

Users were able to express their intended groups well with Errudite.

Ambiguous manual labels prevents consistent replication, even when users thought they did!

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…

Subjectively defined group

+

Small samples

+

Focus exclusively on errors

+

No Test on true cause

Errudite

Precise & reproducible grouping

+

Scale up to the entire dev set

+

Cover errors & correct instances

+

Test via counterfactual analysis

Study #2 Exploration: Errudite Useful Enough?

Freely explore BiDAF with Errudite, think aloud

↓

Describe their observations / insights on BiDAF

↓

Rate insights on importance, confidence, relative easiness

Study #2 Exploration: Errudite Useful Enough?

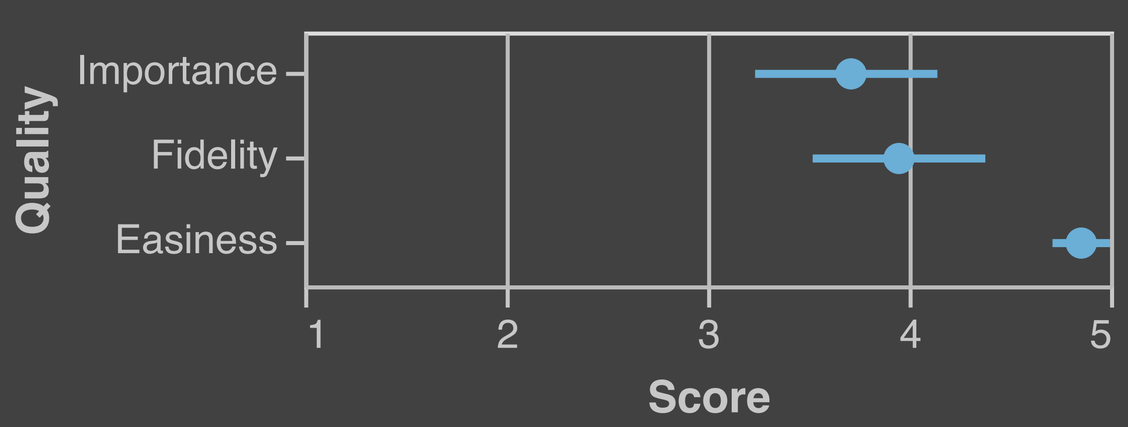
Users reported μ = 2.1, σ = 0.94 findings.

Confirmed prior hypotheses

Extended previous knowledge

Rejected prior hypotheses

Users thought their insights are…



Users learned more about the model (μ= 3.9,σ=0.94).

User Feedback: Did they like Errudite?

Enhanced their error analysis experience.

Systematically scaled up the analysis.

Precise and inspiring more confidence.

Much faster exploration.