Sequence-to-Sequence Networks Learn the Meaning of Reflexive Anaphora

CRAC 2020

Robert Frank and Jackson Petty, Yale University

Primitive Computations in Language Processing

Primitive Computations in Language Processing

- Systematic Mapping:

  Kamala introduced Joe → INTRODUCED(KAMALA, JOE)

Primitive Computations in Language Processing

- Systematic Mapping:

  Kamala introduced Joe → INTRODUCED(KAMALA, JOE)

- Context-sensitive Mapping:

  Kamala introduced herself → INTRODUCED(KAMALA, KAMALA)

  Stacey introduced herself → INTRODUCED(STACEY, STACEY)

Algebraic abstraction

Marcus (1998, 2001)

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:
  - The set-up:

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:
  - The set-up:
    - training: ‘a rose is a rose’, ‘a pig is a pig’, ‘a book is a book’

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:
  - The set-up:
    - training: ‘a rose is a rose’, ‘a pig is a pig’, ‘a book is a book’
    - test: ‘a house is a ______’

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:
  - The set-up:
    - training: ‘a rose is a rose’, ‘a pig is a pig’, ‘a book is a book’
    - test: ‘a house is a ______’
  - The result:

Algebraic abstraction

Marcus (1998, 2001)

- Human cognition exploits algebraic rules, which allow the expression of variable identity.
- Explores whether SRN language models (Elman 1990) can learn such rules:
    - The set-up:
        - training: ‘a rose is a rose’, ‘a pig is a pig’, ‘a book is a book’
        - test: ‘a house is a ______’
    - The result:
        - Failure to predict the correct word!

Algebraic abstraction and anaphora

Frank, Mathis and Badecker (2013)

Algebraic abstraction and anaphora

Frank, Mathis and Badecker (2013)

- Reflexive interpretation requires an algebraic rule:

$$[\![\text{herself}]\!] = \lambda P.\lambda x.P(x, x)$$

Algebraic abstraction and anaphora

Frank, Mathis and Badecker (2013)

- Reflexive interpretation requires an algebraic rule:

$$[\![\text{herself}]\!] = \lambda P.\lambda x.P(x, x)$$

- Can SRNs learn to interpret reflexive anaphora?

ALICE	SEE	ALICE
↑	↑	↑
Alice	sees	herself

Algebraic abstraction and anaphora

Frank, Mathis and Badecker (2013)

- Reflexive interpretation requires an algebraic rule:

$$[\![\text{herself}]\!] = \lambda P.\lambda x.P(x,x)$$

- Can SRNs learn to interpret reflexive anaphora?

ALICE	SEE	ALICE
↑	↑	↑
Alice	sees	herself

- Result: No generalization for names that were not included as reflexive antecedents in the training data!

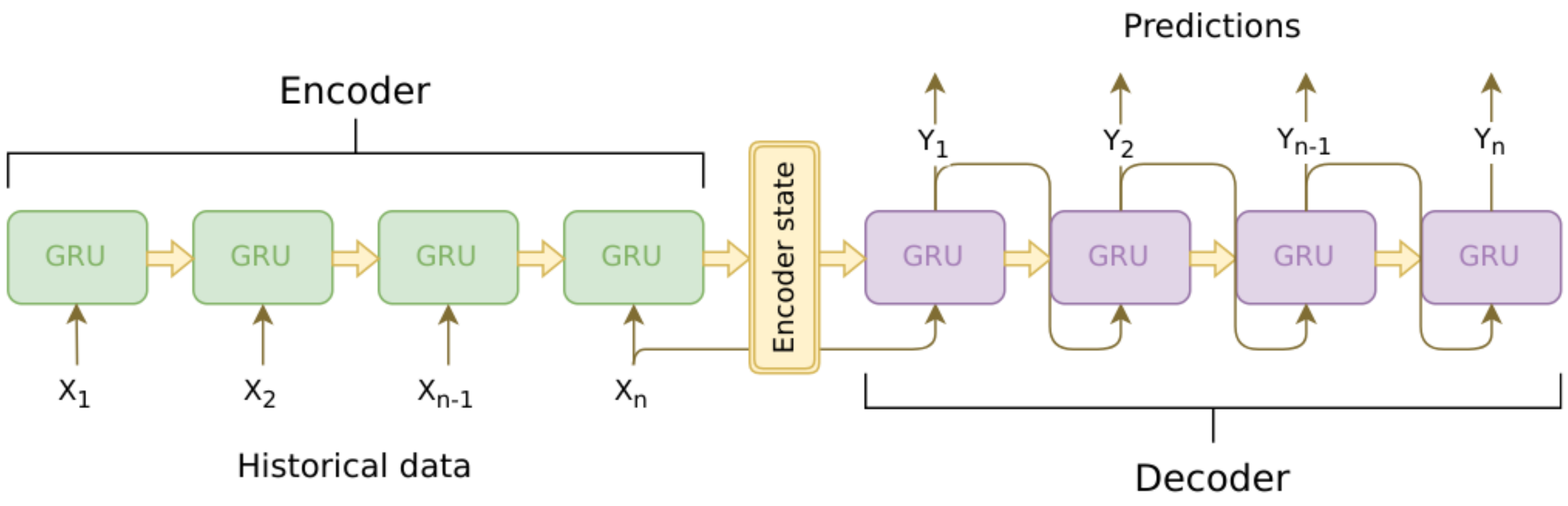
Modern Neural Network Models

Modern Neural Network Models

- **Recurrent units**: LSTMs and GRUs more robustly encode state

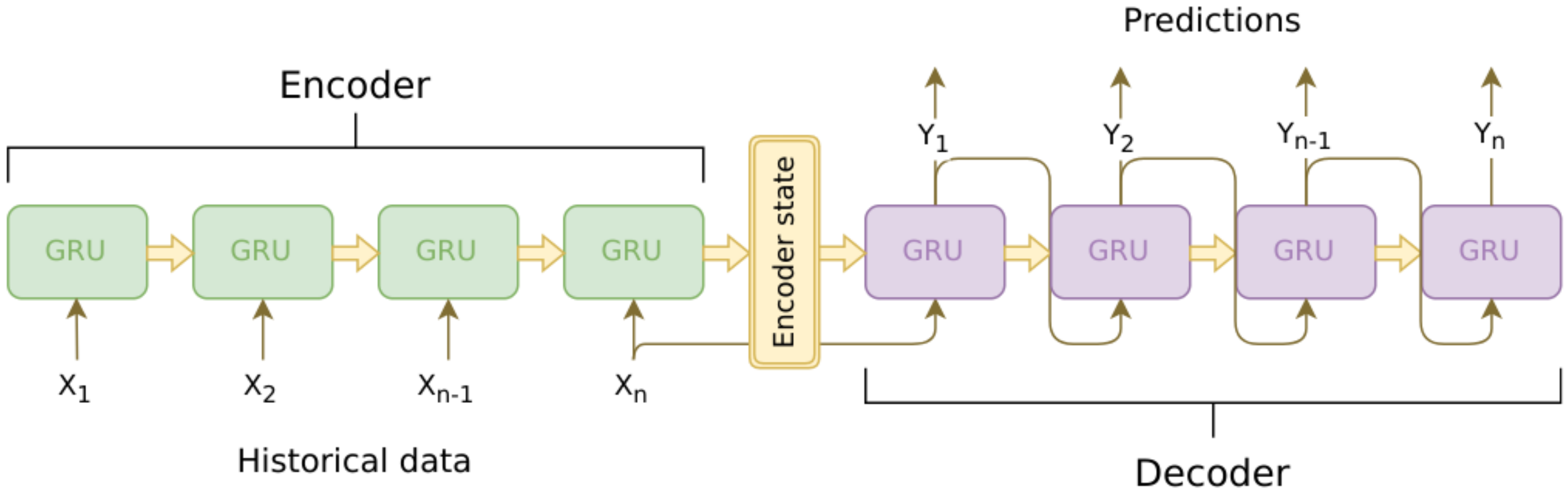
Modern Neural Network Models

- **Recurrent units**: LSTMs and GRUs more robustly encode state
- **Sequence-to-Sequence Architectures**: more flexibility in mapping between input and output



Modern Neural Network Models

- **Recurrent units**: LSTMs and GRUs more robustly encode state
- **Sequence-to-Sequence Architectures**: more flexibility in mapping between input and output



- **Attention mechanisms**: context-sensitive access to encodings

Questions

Questions

1. Are modern neural networks capable of algebraic generalization in reflexive anaphora? Can they learn to interpret a reflexive with a novel antecedent?

Questions

1. Are modern neural networks capable of algebraic generalization in reflexive anaphora? Can they learn to interpret a reflexive with a novel antecedent?
2. What effect does lexical support have? Does the variety of antecedents for a reflexive in the training data impact the network's ability to generalize?

Questions

1. Are modern neural networks capable of algebraic generalization in reflexive anaphora? Can they learn to interpret a reflexive with a novel antecedent?
2. What effect does lexical support have? Does the variety of antecedents for a reflexive in the training data impact the network's ability to generalize?
3. What effect does structural support have? Does the presence of an antecedent in certain structural positions during training affect how well networks learn to generalize to that antecedent?

The Dataset

The Dataset

- Synthetic, pairs of simple sentences and predicate calculus expressions representing their meanings

The Dataset

- Synthetic, pairs of simple sentences and predicate calculus expressions representing their meanings

Intransitive Sentences

Alice swims → SWIM(ALICE)

Bob runs → RUN(BOB)

Claire eats → EAT(CLAIRE)

The Dataset

- Synthetic, pairs of simple sentences and predicate calculus expressions representing their meanings

Intransitive Sentences	**Transitive Sentences**
Alice swims → SWIM(ALICE)	John sees Claire → SEE(JOHN, CLAIRE)
Bob runs → RUN(BOB)	Alice hears Bob → HEAR(ALICE, BOB)
Claire eats → EAT(CLAIRE)	Claire knows Claire → KNOW(CLAIRE, CLAIRE)

The Dataset

- Synthetic, pairs of simple sentences and predicate calculus expressions representing their meanings

Intransitive Sentences

Alice swims → SWIM(ALICE)

Bob runs → RUN(BOB)

Claire eats → EAT(CLAIRE)

Transitive Sentences

John sees Claire → SEE(JOHN, CLAIRE)

Alice hears Bob → HEAR(ALICE, BOB)

Claire knows Claire → KNOW(CLAIRE, CLAIRE)

Reflexive Sentences

John sees himself → SEE(JOHN, JOHN)

Claire hears herself → SEE(CLAIRE, CLAIRE)

The Dataset

- Synthetic, pairs of simple sentences and predicate calculus expressions representing their meanings

Intransitive Sentences

Alice swims → SWIM(ALICE)

Bob runs → RUN(BOB)

Claire eats → EAT(CLAIRE)

Transitive Sentences

John sees Claire → SEE(JOHN, CLAIRE)

Alice hears Bob → HEAR(ALICE, BOB)

Claire knows Claire → KNOW(CLAIRE, CLAIRE)

Reflexive Sentences

John sees himself → SEE(JOHN, JOHN)

Claire hears herself → SEE(CLAIRE, CLAIRE)

- Intended to test whether networks can learn a context-dependent interpretation of reflexives.

Experimental Procedure

Experimental Procedure

- Follow *Poverty of the Stimulus* paradigm: each of our experiments involves systematically removing some class of sentences from the dataset, which form the ***generalization set.***

Experimental Procedure

- Follow *Poverty of the Stimulus* paradigm: each of our experiments involves systematically removing some class of sentences from the dataset, which form the ***generalization set.***

- Remaining data is divided into training, validation and ***test set***.

Experimental Procedure

- Follow *Poverty of the Stimulus* paradigm: each of our experiments involves systematically removing some class of sentences from the dataset, which form the ***generalization set.***
- Remaining data is divided into training, validation and ***test set***.
- We report ***full sentence accuracy*** for both generalization and test sets.

Experimental Procedure

- Follow *Poverty of the Stimulus* paradigm: each of our experiments involves systematically removing some class of sentences from the dataset, which form the ***generalization set.***

- Remaining data is divided into training, validation and ***test set***.

- We report ***full sentence accuracy*** for both generalization and test sets.

- For each experiment, we trained 5 randomly initialized sequence-to-sequence networks with each recurrent unit type (SRN, LSTM, GRU) with and without multiplicative attention (Luong et al., 2015).

Experimental Procedure

- Follow *Poverty of the Stimulus* paradigm: each of our experiments involves systematically removing some class of sentences from the dataset, which form the ***generalization set.***
- Remaining data is divided into training, validation and ***test set***.
- We report ***full sentence accuracy*** for both generalization and test sets.
- For each experiment, we trained 5 randomly initialized sequence-to-sequence networks with each recurrent unit type (SRN, LSTM, GRU) with and without multiplicative attention (Luong et al., 2015).
- Trained with SGD for maximum of 100 epochs (with early stopping).

Q1: Can networks generalize reflexive meanings?

Experiment 1

Generalization

Alice *verbs* herself

Training

Claire *verbs* herself
Eliza *verbs* herself
Bob *verbs* himself
…

Bob *verbs*
Alice *verbs*
Claire *verbs*
…

Bob *verbs* Alice
Alice *verbs* Claire
Alice *verbs* Alice
…

Experiment 1

Mean test set accuracy (over random seeds)

Experiment 1

Mean test set accuracy (over random seeds)

Experiment 1

Mean test set accuracy (over random seeds)

	No Attention	Attention

Experiment 1

Mean test set accuracy (over random seeds)

	No Attention	Attention
SRN	100	100

Experiment 1

Mean test set accuracy (over random seeds)

	No Attention	Attention
SRN	100	100
GRU	100	100

Experiment 1

Mean test set accuracy (over random seeds)

	No Attention	Attention
SRN	100	100
GRU	100	100
LSTM	100	100

Experiment 1

Mean generalization accuracy

Experiment 1

Mean generalization accuracy

Experiment 1

Mean generalization accuracy

	No Attention	Attention

Experiment 1

Mean generalization accuracy

	No Attention	Attention
SRN	100	100

Experiment 1

Mean generalization accuracy

	No Attention	Attention
SRN	100	100
GRU	100	100

Experiment 1

Mean generalization accuracy

	No Attention	Attention
SRN	100	100
GRU	100	100
LSTM	100	100

Q1: Can networks generalize reflexive meanings?

Experiment 2

Generalization

Alice *verbs* herself

Alice *verbs* Alice

→ VERB(ALICE, ALICE)

Training

Claire *verbs* herself
Eliza *verbs* herself
Bob *verbs* himself
…

Bob *verbs*
Alice *verbs*
Claire *verbs*
…

Bob *verbs* Alice
Alice *verbs* Claire
…

Experiment 2

Mean test set accuracy

Experiment 2

Mean test set accuracy

Experiment 2

Mean test set accuracy

	No Attention	Attention

Experiment 2

Mean test set accuracy

	No Attention	Attention
SRN	100	100

Experiment 2

Mean test set accuracy

	No Attention	Attention
SRN	100	100
GRU	100	100

Experiment 2

Mean test set accuracy

	No Attention	Attention
SRN	100	100
GRU	100	100
LSTM	100	100

Experiment 2

Mean test set accuracy

	No Attention	Attention
SRN	100	100
GRU	100	100
LSTM	100	100

Test set accuracy for **all** of our experiments is at ceiling, so we do not report it in subsequent discussion!

Experiment 2

Mean generalization accuracy

Experiment 2

Mean generalization accuracy

Experiment 2

Mean generalization accuracy

	No Attention	Attention

Experiment 2
Mean generalization accuracy

	No Attention	Attention
SRN	86	100

Experiment 2

Mean generalization accuracy

	No Attention	Attention
SRN	86	100
GRU	100	100

Experiment 2

Mean generalization accuracy

	No Attention	Attention
SRN	86	100
GRU	100	100
LSTM	100	100

Key Point, contra (Marcus 1991 and Frank et al. 2013):

Seq2Seq models ARE capable of learning an abstract reflexive meaning that generalizes to a novel antecedent!

Q2: How does lexical support affect generalization?

Experiment 3: vary number of held-out antecedents

Generalization

Alice *verbs* herself/Alice

Claire *verbs* herself/Claire

Eliza *verbs herself*/Eliza

…

Training

Bob *verbs* himself
Zelda *verbs* herself
…

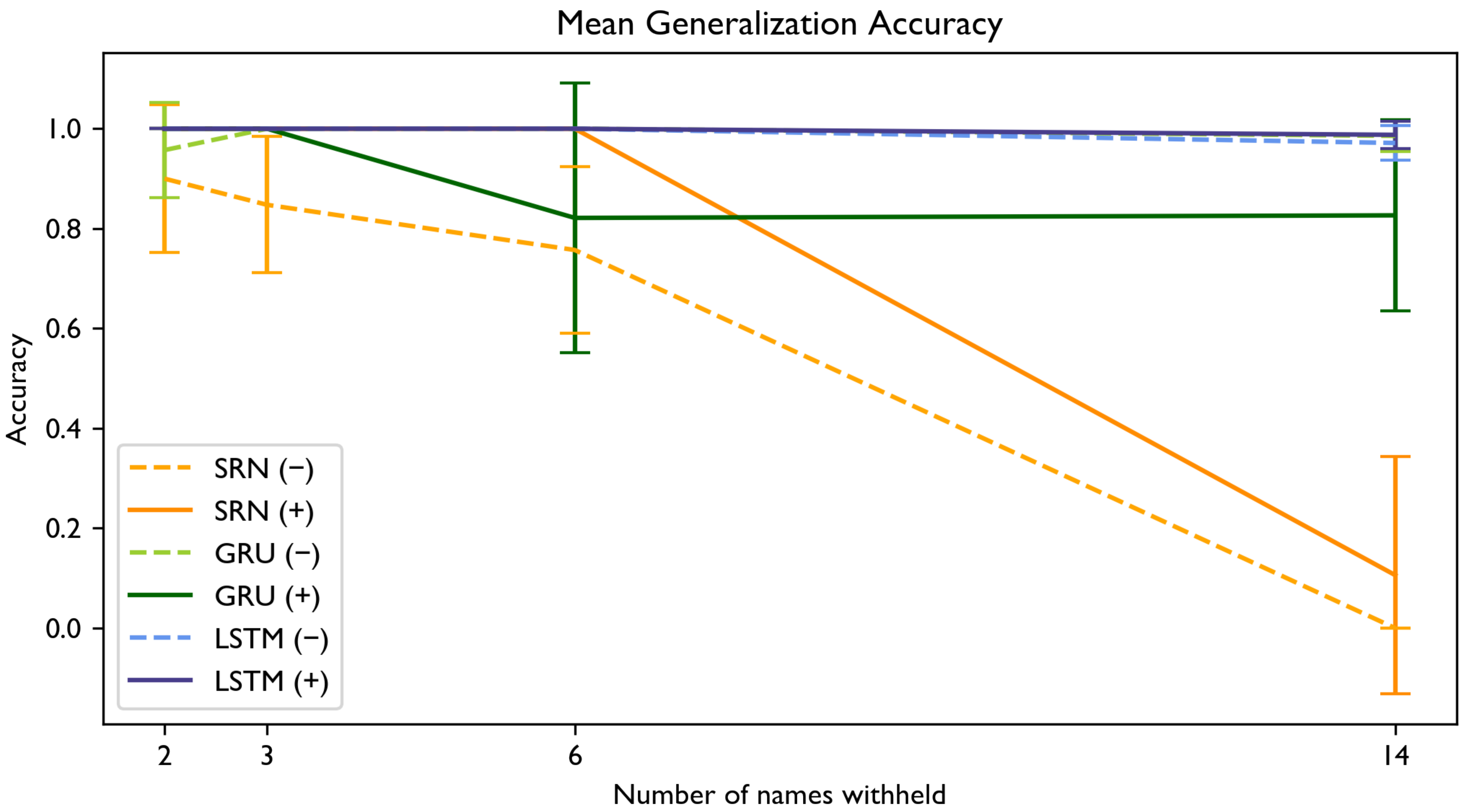
Bob *verbs*
Alice *verbs*
Claire *verbs*
…

Bob *verbs* Alice
Alice *verbs* Claire
Alice *verbs* Alice
…

Experiment 3

Generalization accuracy (on "*P verbs* herself" sentences)

- LSTMs (-) and (+) and GRUs (-) do excellently!
- GRUs (+) drop off somewhat beyond 3 antecedents withheld
- SRNs drop off significantly after 6 antecedents withheld

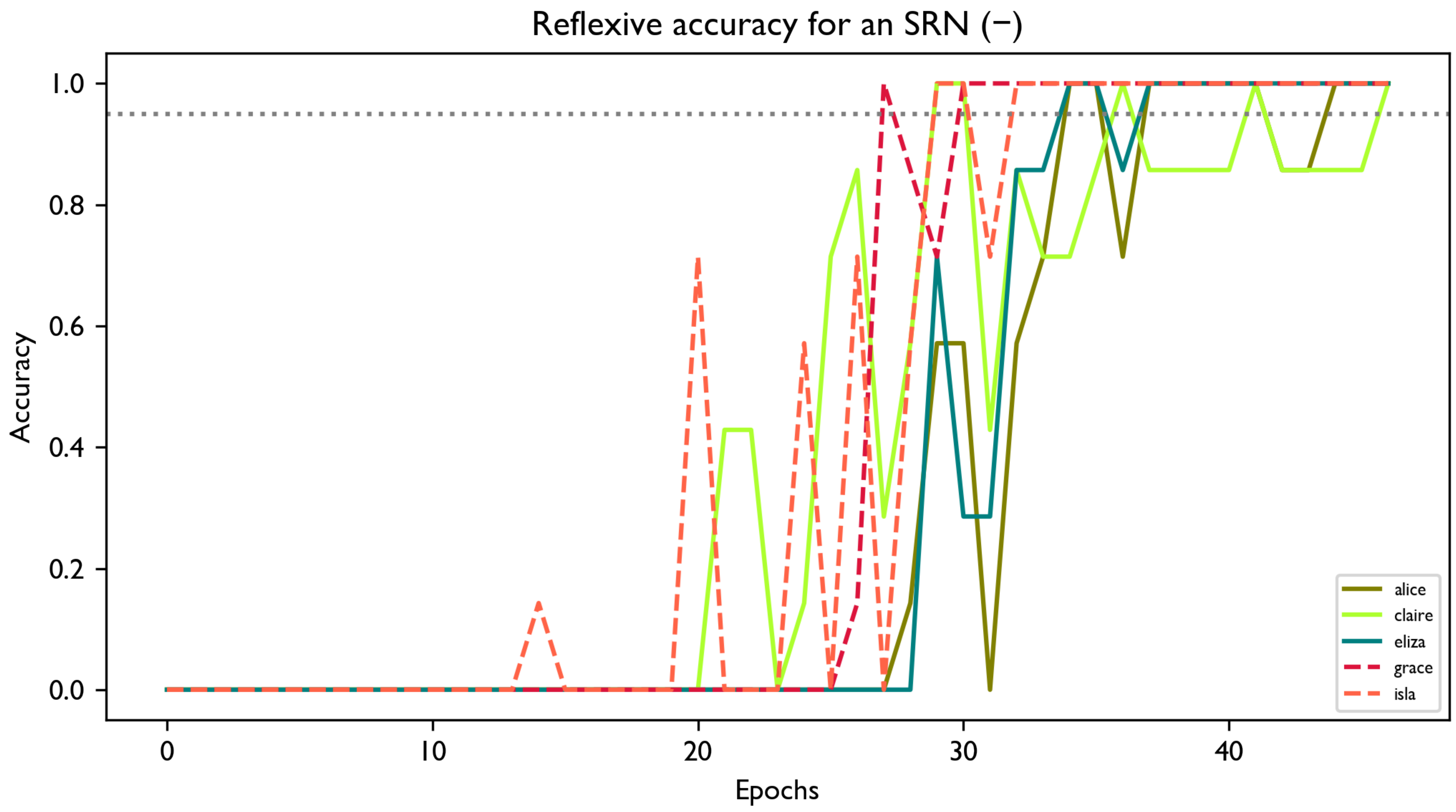


Experiment 3

Learning curve for "*P verbs* herself" sentences

- Networks learn how to interpret reflexives in a piecemeal fashion, even if they do generalize!

- In-sample antecedents (dashed lines) are typically learned before out-of sample antecedents.



Q3: How does structural support affect generalization?

Q3: How does structural support affect generalization?

- The structure of both the form and meaning provides support

Q3: How does structural support affect generalization?

- The structure of both the form and meaning provides support
  - Impact of seeing “Alice” in subject vs object position?

Q3: How does structural support affect generalization?

- The structure of both the form and meaning provides support
  - Impact of seeing “Alice” in subject vs object position?
  - Impact of seeing ALICE as the first vs second argument to a predicate?

Q3: How does structural support affect generalization?

- The structure of both the form and meaning provides support
  - Impact of seeing “Alice” in subject vs object position?
  - Impact of seeing ALICE as the first vs second argument to a predicate?
- Intransitive sentences provide an interesting point of note

Q3: How does structural support affect generalization?

- The structure of both the form and meaning provides support
  - Impact of seeing “Alice” in subject vs object position?
  - Impact of seeing ALICE as the first vs second argument to a predicate?
- Intransitive sentences provide an interesting point of note
  - Unary predicates ⇒ ambiguity in role of single argument

Q3: How does structural support affect generalization?

Experiment 4a: Alice is no transitive subject!

Generalization

Alice *verbs* herself

Alice *verbs* Alice

Alice *verbs* Bob

Training

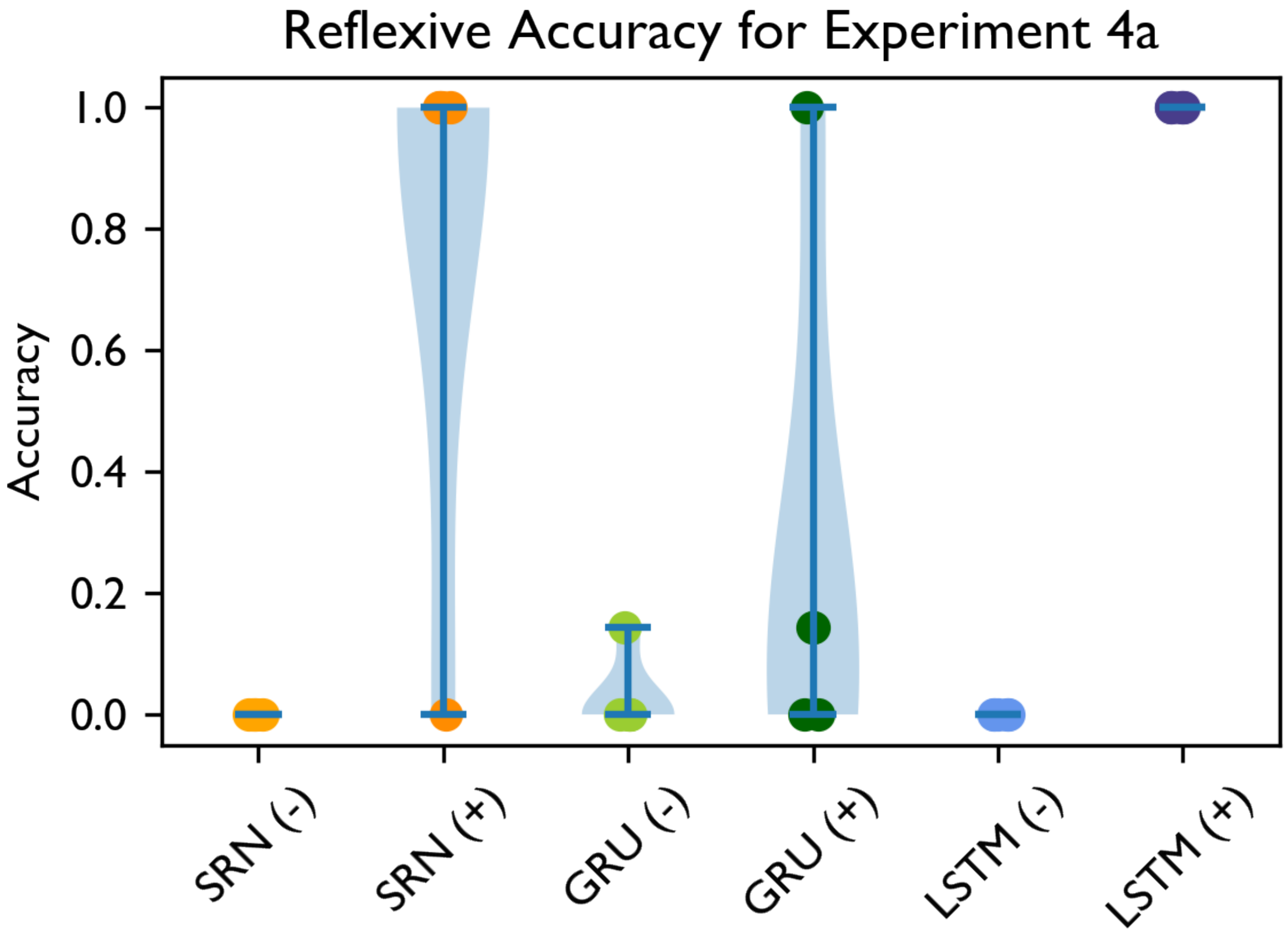
Bob *verbs* himself
Claire *verbs* herself
Daniel *verbs* himself
...

Bob *verbs*
Alice *verbs*
Claire *verbs*
...

Claire *verbs* Alice
Daniel *verbs* Bob
John *verbs* John
...

Experiment 4a

Generalization accuracy (on "Alice *verbs* herself" sentences)



SRN (-)	SRN (+)	GRU (-)	GRU (+)	LSTM (-)	LSTM (+)
0.0	**0.8**	**0.03**	**0.26**	**0.0**	**1.0**

- LSTMs (+) do excellently!
- SRNs (+) outperform GRUs (+)
- Attention definitely matters

Q3: How does structural support affect generalization?

Experiment 4b: Alice is no transitive or intransitive subject!

Generalization

Alice *verbs* herself

Alice *verbs* Alice

Alice *verbs*

Alice *verbs* Bob

Training

Bob *verbs* himself
Claire *verbs* herself
Daniel *verbs* himself
…

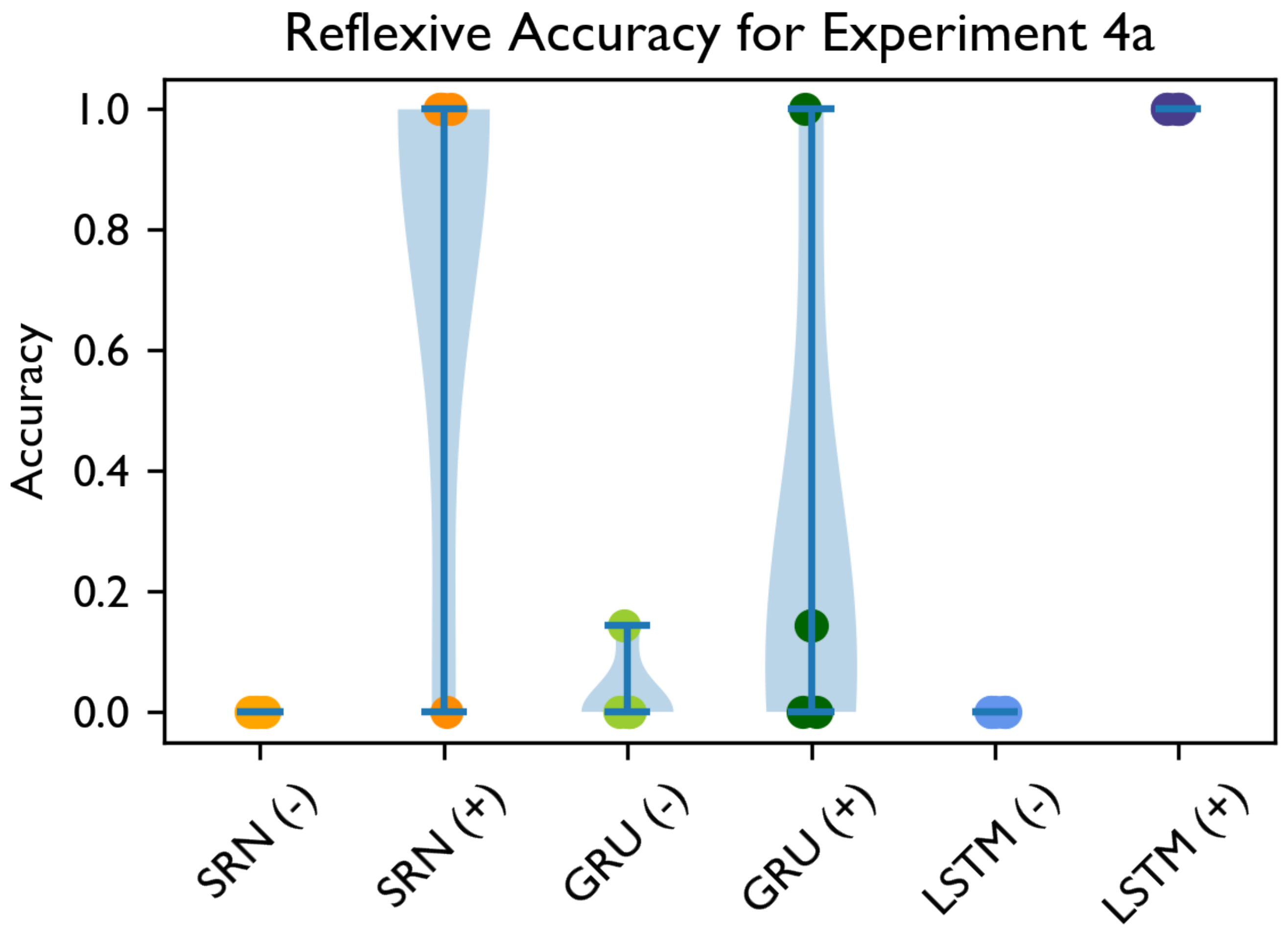
Bob *verbs*
Claire *verbs*
…

Claire *verbs* Alice
Daniel *verbs* Bob
John *verbs* John
…

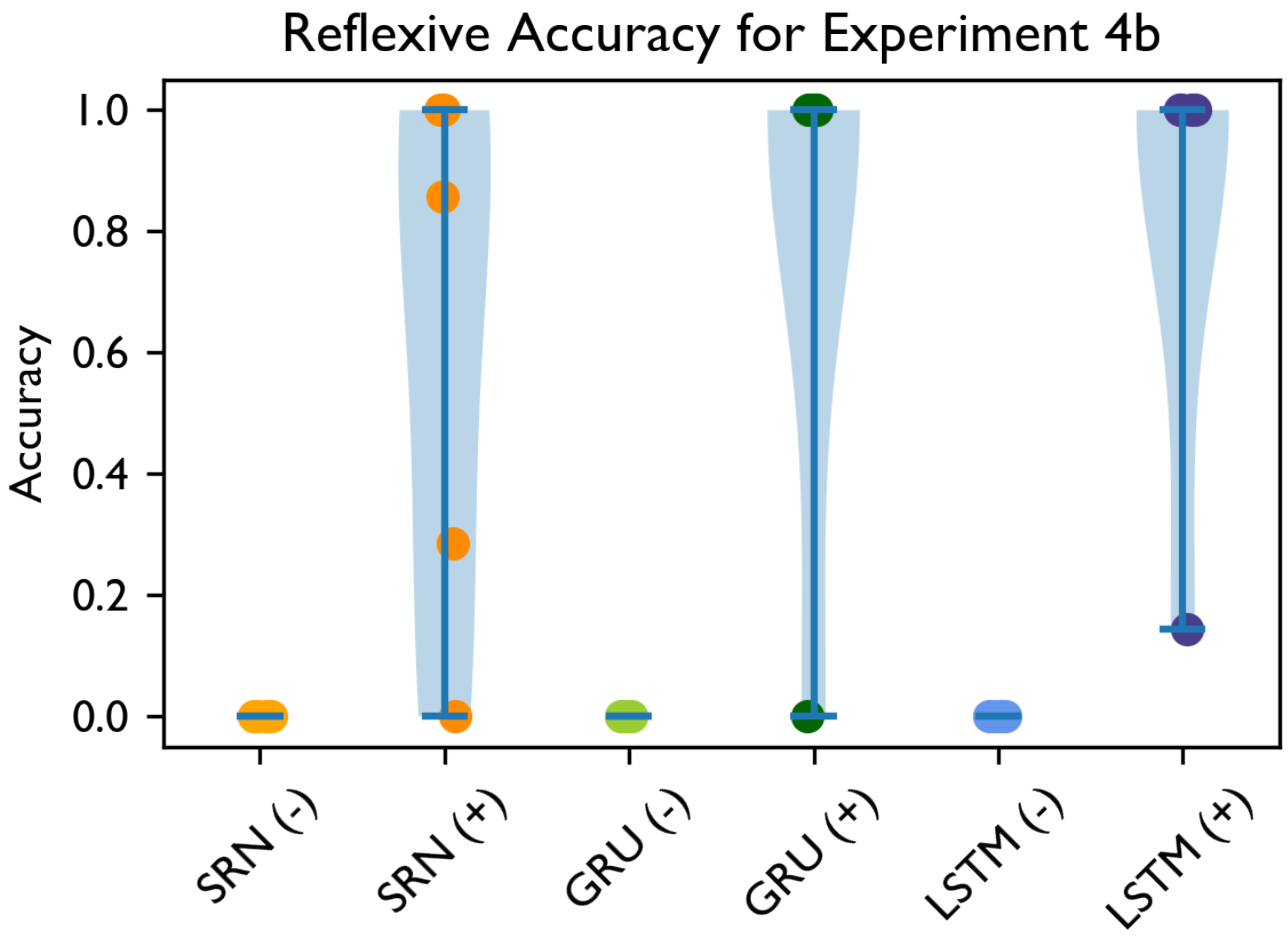
Experiment 4b

Generalization accuracy (on “Alice *verbs* herself” sentences)

- Attention still matters
- LSTM and SRN performance dropped
- **GRU performance went up**



0.0 0.8 0.03 0.26 0.0 1.0



0.0 0.63 0.0 0.8 0.0 0.83

Q3: How does structural support affect generalization?

Experiment 5a: Alice is no object!

Generalization

Alice *verbs* herself

Alice *verbs* Alice

Bob *verbs* Alice

Training

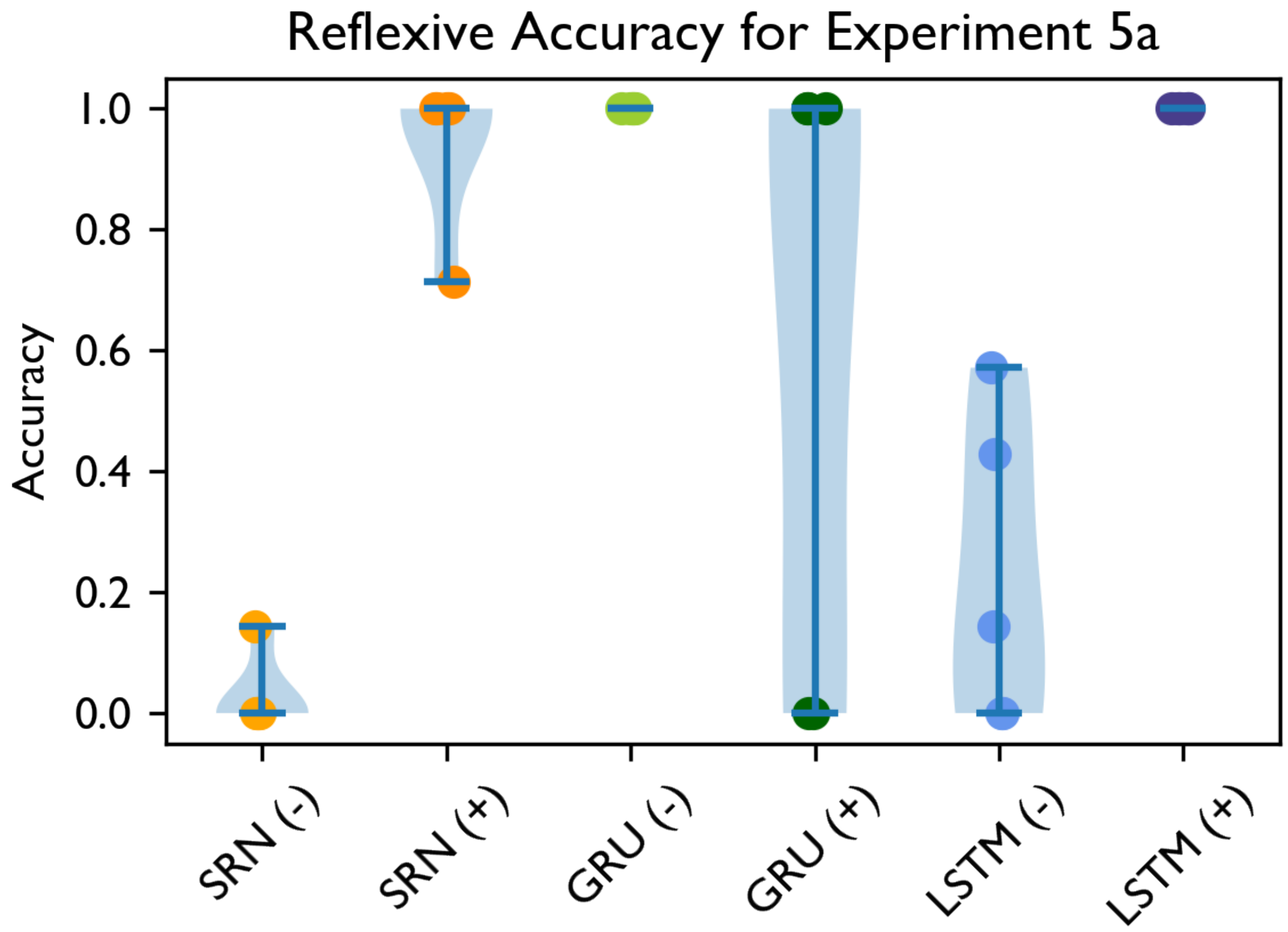
Bob *verbs* himself
Claire *verbs* herself
Daniel *verbs* himself
…

Bob *verbs*
Alice *verbs*
Claire *verbs*
…

Alice *verbs* Claire
Daniel *verbs* Bob
John *verbs* John
…

Experiment 5a

Generalization accuracy (on “Alice *verbs* herself” sentences)



0.03 0.94 0.98 0.60 0.23 1.0

- LSTMs (+) do excellently (again)!
- SRNs (+) outperform GRUs (+) (again)!
- GRUs (-) perform excellently!?

Q3: How does structural support affect generalization?

Experiment 5b: Alice is no object (or intransitive subject)!

Generalization

Alice *verbs* herself

Alice *verbs* Alice

Alice *verbs*

Bob *verbs* Alice

Training

Bob *verbs* himself
Claire *verbs* herself
Daniel *verbs* himself
…

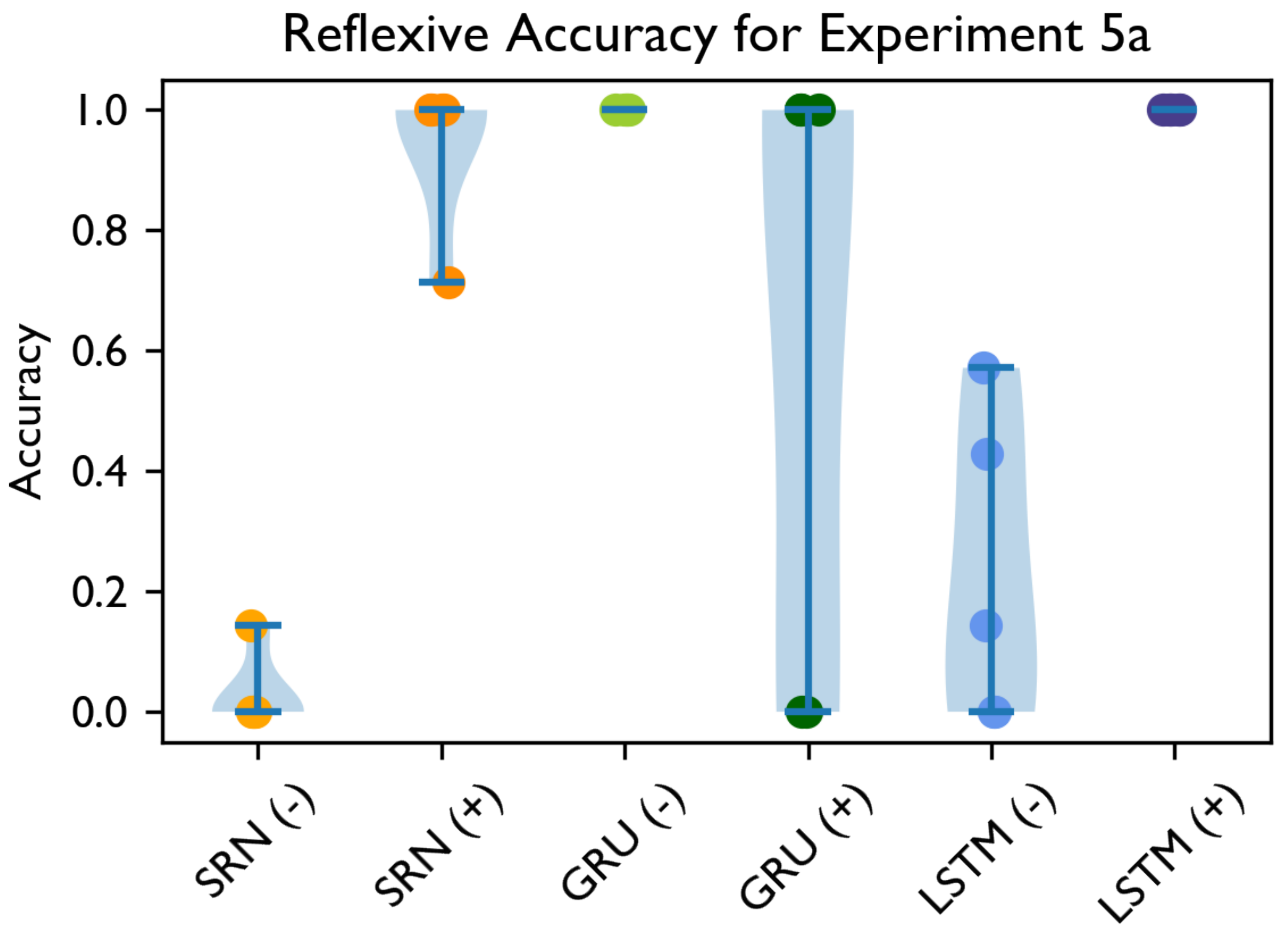
Bob *verbs*
Claire *verbs*
…

Alice *verbs* Claire
Daniel *verbs* Bob
John *verbs* John
…

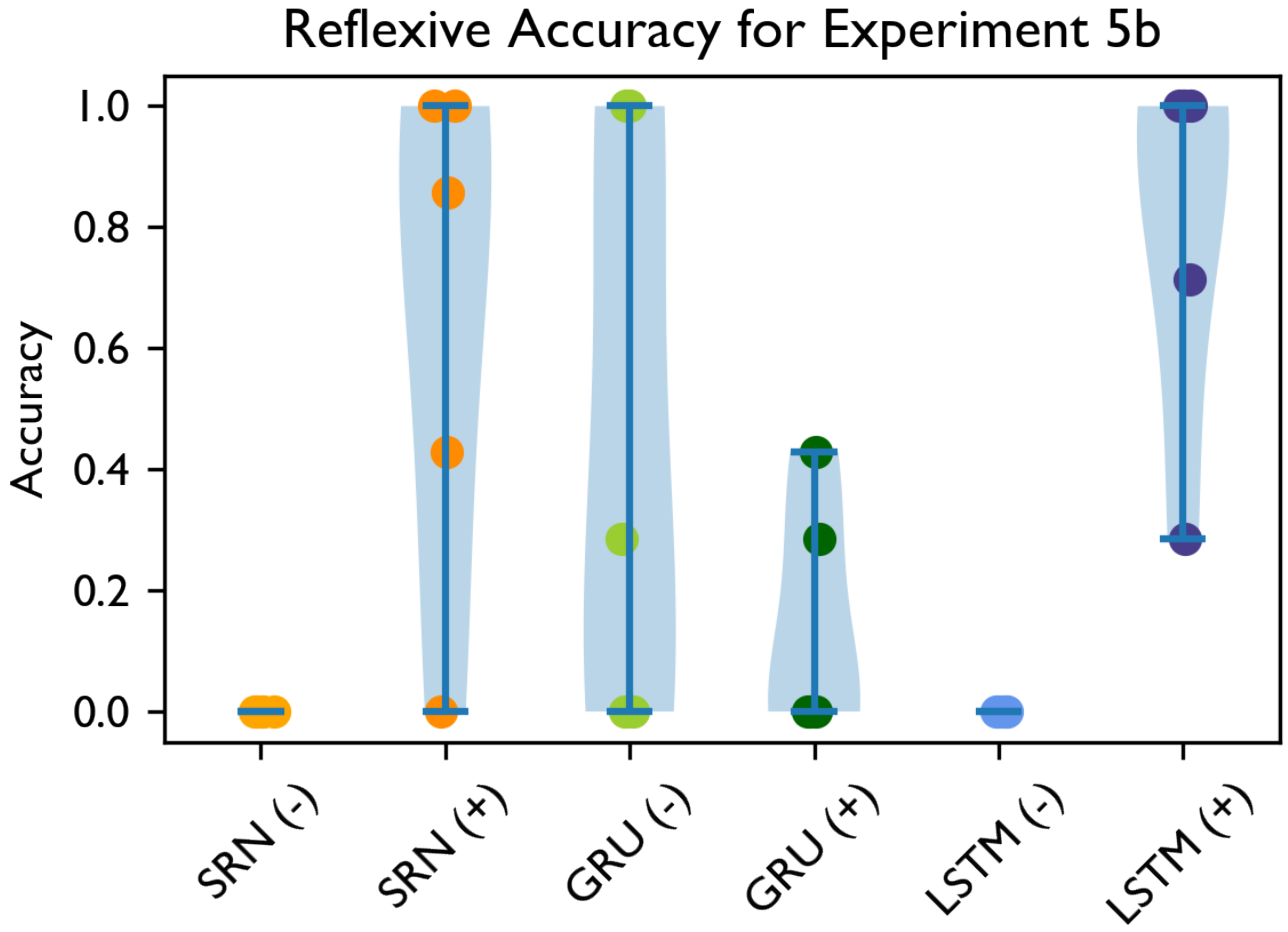
Experiment 5b

Generalization accuracy (on “Alice *verbs* herself” sentences)

- Performance degrades overall



0.03 0.94 0.98 0.60 0.23 1.0



0.0 0.65 0.45 0.14 0.00 0.80

Questions with which we began...

1. Are modern neural networks capable of algebraic generalization in reflexive anaphora? Can they learn to interpret a reflexive with a novel antecedent?

   Yes! Seq2Seq architectures with even the simplest recurrent unit (SRNs) and no attention can do it!

2. What effect does lexical support have? Does the variety of antecedents for a reflexive in the training data impact the network's ability to generalize?

   Generalization depends on the number of antecedents presented during training, though this varies by architecture. LSTMs and GRUs generalize in the presence of limited lexical support.

3. What effect does structural support have? Does the presence of an antecedent in certain structural positions affect how well networks learn to generalize to that antecedent?

   Presence of the antecedent as a subject or object in non-reflexive sentences in training does affect generalization.

Open questions

Open questions

- Structural dependence of anaphora

    The student near the teacher sees herself → SEE(STUDENT,STUDENT)

Open questions

- Structural dependence of anaphora

    The student near the teacher sees herself → SEE(STUDENT,STUDENT)

- Relation to systematicity and SCAN (and proposed solutions)

    jump twice → JUMP JUMP

Thank you for watching

Robert Frank and Jackson Petty, Yale University