Human-like compositional generalization through meta-learning

Brenden M. Lake
New York University
Meta AI

Human-like compositional generalization through meta-learning

Brenden M. Lake
New York University
Meta AI

Collaborators

Marco Baroni

Tal Linzen

Cognition, 28 (1988) 3–71

Connectionism and cognitive architecture: A critical analysis*

JERRY A. FODOR
CUNY Graduate Center

ZENON W. PYLYSHYN
University of Western Ontario

Abstract

This paper explores differences between Connectionist proposals for cognitive architecture and the sorts of models that have traditionally been assumed in cognitive science. We claim that the major distinction is that, while both Connectionist and Classical architectures postulate representational mental states, the latter but not the former are committed to a symbol-level of representation, or to a 'language of thought': i.e., to representational states that have combinatorial syntactic and semantic structure. Several arguments for combinatorial structure in mental representations are then reviewed. These include arguments based on the 'systematicity' of mental representation: i.e., on the fact that cognitive capacities always exhibit certain symmetries, so that the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents. We claim that such arguments make a powerful case that mind/brain architecture is not Connectionist at the cognitive level. We then consider the possibility that Connectionism may provide an account of the neural (or 'abstract neurological') structures in which Classical cognitive architecture is implemented. We survey a number of the standard arguments that have been offered in favor of Connectionism, and conclude that they are coherent only on this interpretation.

*This paper is based on a chapter from a forthcoming book. Authors' names are listed alphabetically. We wish to thank the Alfred P. Sloan Foundation for their generous support of this research. The preparation of this paper was also aided by a Killam Research Fellowship and a Senior Fellowship from the Canadian Institute for Advanced Research to ZWP. We also gratefully acknowledge comments and criticisms of earlier drafts by: Professors Noam Chomsky, William Demopoulos, Lila Gleitman, Russ Greiner, Norbert Hornstein, Keith Humphrey, Sandy Pentland, Steven Pinker, David Rosenthal, and Edward Stabler. Reprints may be obtained by writing to either author: Jerry Fodor, CUNY Graduate Center, 33 West 42 Street, New York, NY 10036, U.S.A.: Zenon Pylyshyn, Centre for Cognitive Science, University of Western Ontario, London, Ontario, Canada N6A 5C2.

0010-0277/88/$21.20

1. Introduction

Connectionist or *PDP* models are catching on. There are conferences and new books nearly every day, and the popular science press hails this new wave of theorizing as a breakthrough in understanding the mind (a typical example is the article in the May issue of *Science 86*, called "How we think: A new theory"). There are also, inevitably, descriptions of the emergence of Connectionism as a Kuhnian "paradigm shift". (See Schneider, 1987, for an example of this and for further evidence of the tendency to view Connectionism as the "new wave" of Cognitive Science.)

The fan club includes the most unlikely collection of people. Connectionism gives solace both to philosophers who think that relying on the pseudoscientific intentional or semantic notions of folk psychology (like goals and beliefs) mislead psychologists into taking the computational approach (e.g., P.M. Churchland, 1981; P.S. Churchland, 1986; Dennett, 1986); and to those with nearly the opposite perspective, who think that computational psychology is bankrupt because it doesn't address issues of intentionality or meaning (e.g., Dreyfus & Dreyfus, in press). On the computer science side, Connectionism appeals to theorists who think that serial machines are too weak and must be replaced by radically new parallel machines (Fahlman & Hinton, 1986), while on the biological side it appeals to those who believe that cognition can only be understood if we study it as neuroscience (e.g., Arbib, 1975; Sejnowski, 1981). It is also attractive to psychologists who think that much of the mind (including the part involved in using imagery) is not discrete (e.g., Kosslyn & Hatfield, 1984), or who think that cognitive science has not paid enough attention to stochastic mechanisms or to "holistic" mechanisms (e.g., Lakoff, 1986), and so on and on. It also appeals to many young cognitive scientists who view the approach as not only anti-establishment (and therefore desirable) but also rigorous and mathematical (see, however, footnote 2). Almost everyone who is discontent with contemporary cognitive psychology and current "information processing" models of the mind has rushed to embrace "the Connectionist alternative".

When taken as a way of modeling *cognitive architecture*, Connectionism really does represent an approach that is quite different from that of the Classical cognitive science that it seeks to replace. Classical models of the mind were derived from the structure of Turing and Von Neumann machines. They are not, of course, committed to the details of these machines as exemplified in Turing's original formulation or in typical commercial computers; only to the basic idea that the kind of computing that is relevant to understanding cognition involves operations on symbols (see Fodor 1976, 1987; Newell, 1980, 1982; Pylyshyn, 1980, 1984a, b). In contrast, Connec-

Systematic compositionality

The algebraic capacity to understand and produce novel combinations from known components

Systematic compositionality

The algebraic capacity to understand and produce novel combinations from known components

One-shot learning:
"This is how you dax"

Can you then:
"Dax twice?"
"Dax while jumping?"
"Dax wildly around the room?"

Reevaluating F&P's arguments in the age of deep learning

Recent benchmarks for compositional generalization

- SCAN (Lake & Baroni, 2018)
- CLOSURE (Bahdanau et al., 2019)
- DBCA (Keyers et al., 2019)
- Comparisons (Dasgupta et al., 2019)
- COGS (Kim & Linzen, 2020)
- gSCAN (Ruis et al., 2020)
- PCFG SET (Hupkes et al., 2020)
- NMT Challenge (Dankers et al., 2022)

What do we find?

Somewhat surprisingly, neural networks still struggle on tests of systematicity

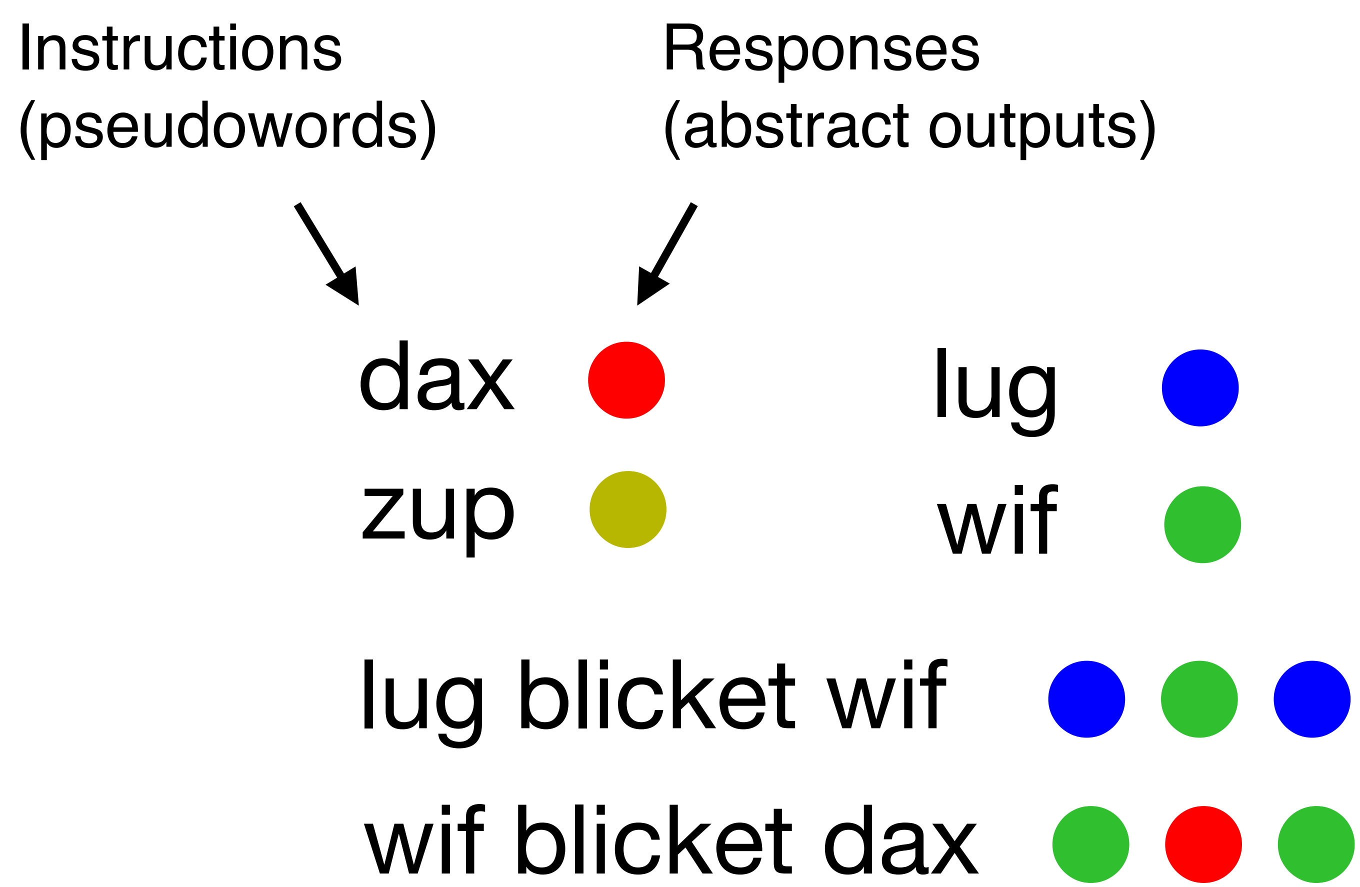
Goals of this work

1. Behavioral studies to compare humans and machines side-by-side on the same tests of systematicity

2. An approach to building neural networks that can achieve human-like systematic generalization, through an optimization procedure that encourages systematicity

A test of systematicity for humans and machines

Support



Query

dax blicket zup?

A test of systematicity for humans and machines

Support

dax

lug

zup

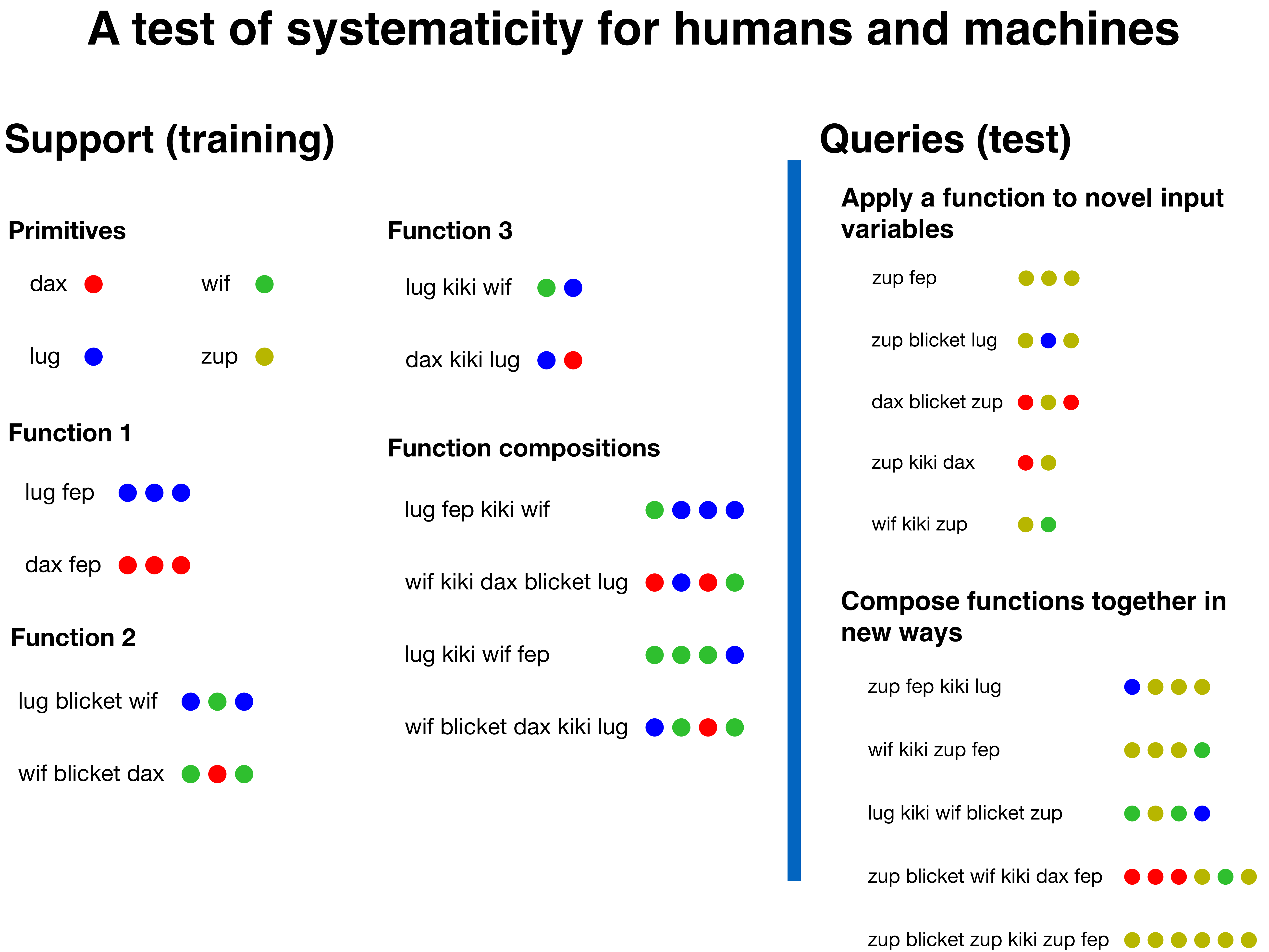
wif

lug blicket wif

wif blicket dax

Query

dax blicket zup?

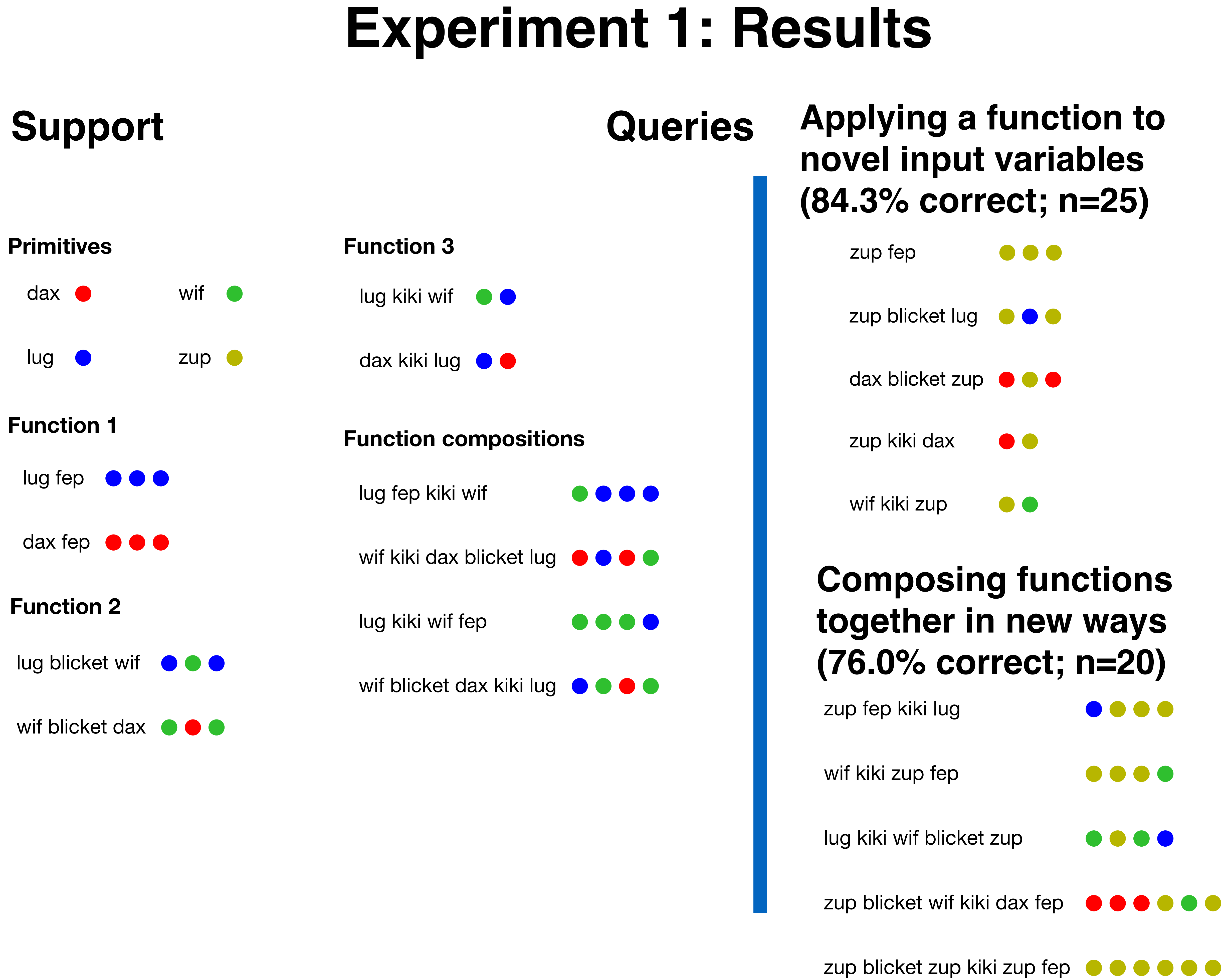

A test of systematicity for humans and machines
Support (training)
Primitives
dax
wif
lug
zup
Function 1
lug fep
dax fep
Function 2
lug blicket wif
wif blicket dax
Function 3
lug kiki wif
dax kiki lug
Function compositions
lug fep kiki wif
wif kiki dax blicket lug
lug kiki wif fep
wif blicket dax kiki lug
Queries (test)
Apply a function to novel input variables
zup fep
zup blicket lug
dax blicket zup
zup kiki dax
wif kiki zup
Compose functions together in new ways
zup fep kiki lug
wif kiki zup fep
lug kiki wif blicket zup
zup blicket wif kiki dax fep
zup blicket zup kiki zup fep

Behavioral experiment 1: Design

- Four primitive instructions:
  - dax ➔ 🔴
  - lug ➔ 🔵
  - wif ➔ 🟢
  - zup ➔ 🟡
- Modifier ("fep"-thrice):
  - dax fep ➔ 🔴 🔴 🔴
  - lug fep ➔ 🔵 🔵 🔵
- Conjunctions ("blicket"-surround, "kiki"-after):
  - wif blicket dax ➔ 🟢 🔴 🟢
  - lug blicket wif ➔ 🔵 🟢 🔵
  - dax kiki lug ➔ 🔵 🔴
- Simplifications:
  - No scope ambiguity ("lug kiki [wif fep]" ➔ 🟢 🟢 🟢 🔵)

- Instructions: "learn a set of commands and their corresponding outputs"
- Outputs produced by dragging symbols from a pool of options
- Curriculum learning
- Support set remained visible during query phase
- Participants recruited on AMT

🟡 🟢 🔵 🔴 ...


Experiment 1: Results
Support
Queries
Primitives
dax
wif
lug
zup
Function 1
lug fep
dax fep
Function 2
lug blicket wif
wif blicket dax
Function 3
lug kiki wif
dax kiki lug
Function compositions
lug fep kiki wif
wif kiki dax blicket lug
lug kiki wif fep
wif blicket dax kiki lug
Applying a function to novel input variables (84.3% correct; n=25)
zup fep
zup blicket lug
dax blicket zup
zup kiki dax
wif kiki zup
Composing functions together in new ways (76.0% correct; n=20)
zup fep kiki lug
wif kiki zup fep
lug kiki wif blicket zup
zup blicket wif kiki dax fep
zup blicket zup kiki zup fep

Experiment 1: Results

Support

dax

zup

lug

wif

lug blicket wif

wif blicket dax

Query

Correct answer

dax blicket zup ?

85% participants

Representative mistake

dax blicket zup ?

“1-to-1” bias?

Experiment 1: Results

Support

dax

wif

lug fep

lug

zup

dax fep

lug kiki wif

lug fep kiki wif

dax kiki lug

wif kiki dax blicket lug

lug blicket wif

lug kiki wif fep

wif blicket dax

wif blicket dax kiki lug

Query

Correct answer

wif kiki zup fep ?

85% participants

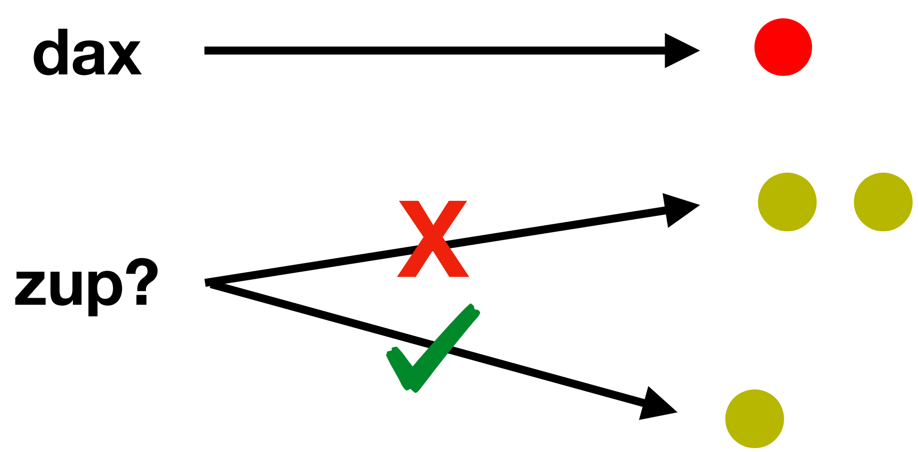
Representative mistake

wif kiki zup fep ?

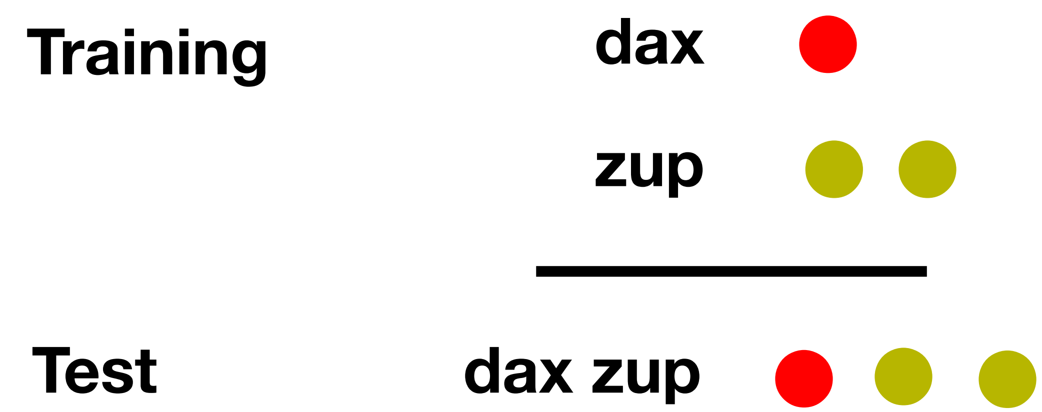
“iconic concatenation” bias?

Candidate inductive biases

1-to-1:
each input symbol corresponds
to exactly one output symbol

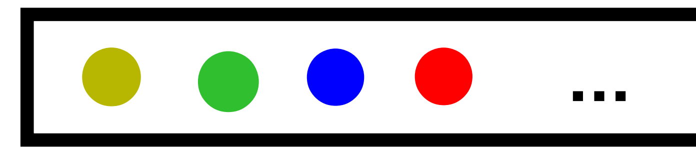


iconic concatenation (IC): first in first out



Experiment 2: Examining inductive biases - perform a seq2seq task with NO training examples!

pool (6 items)



zup?

zup zup?

dax zup?

zup tufa?

zup wif zup?

zup wif blicket?

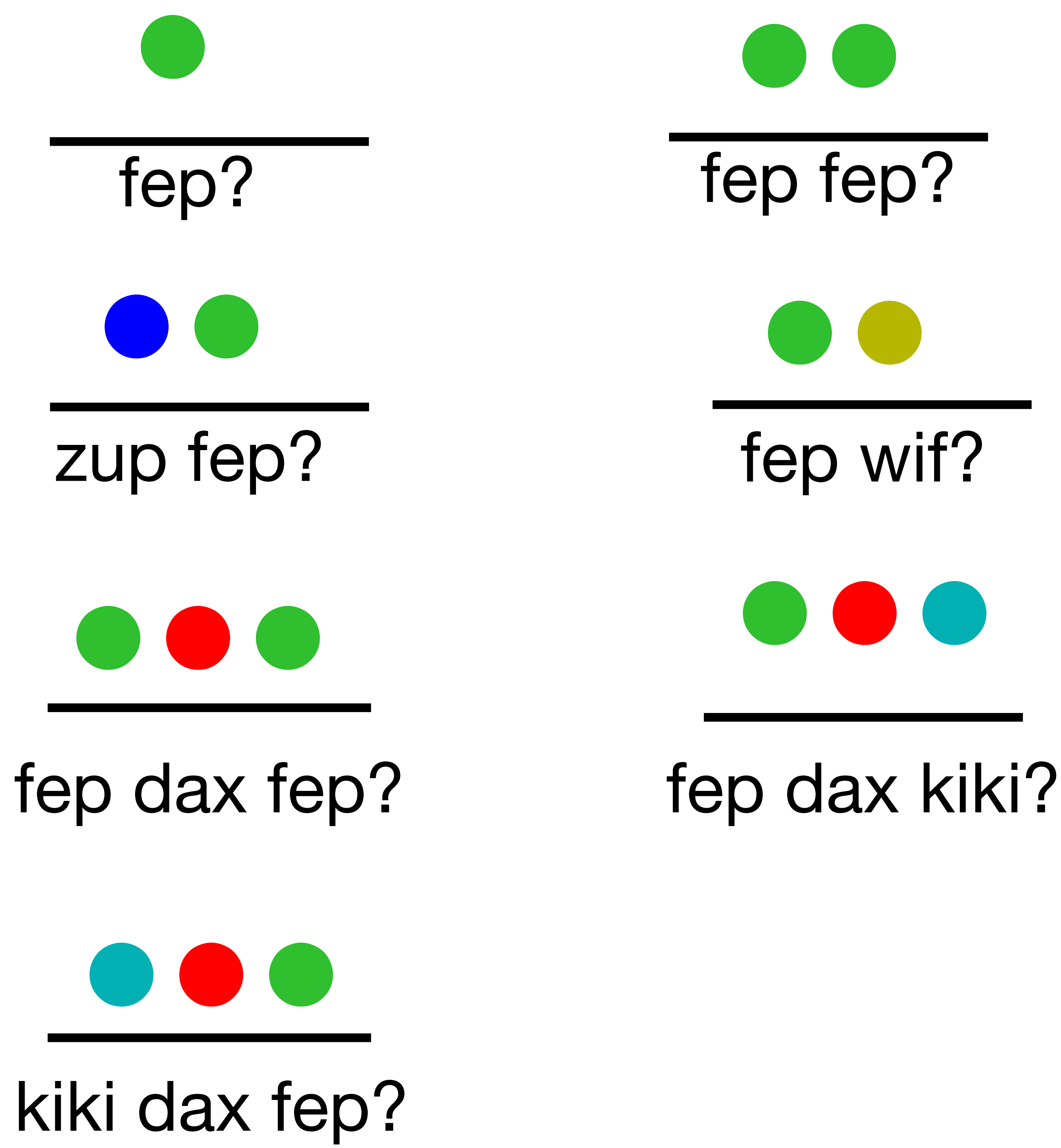
blicket wif zup?

Experiment 2: Results

Representative response on open-ended task

(59% responded this way, shows 3 inductive biases:

1-to-1; Iconic Concatenation (IC); Mutual Exclusivity (ME))

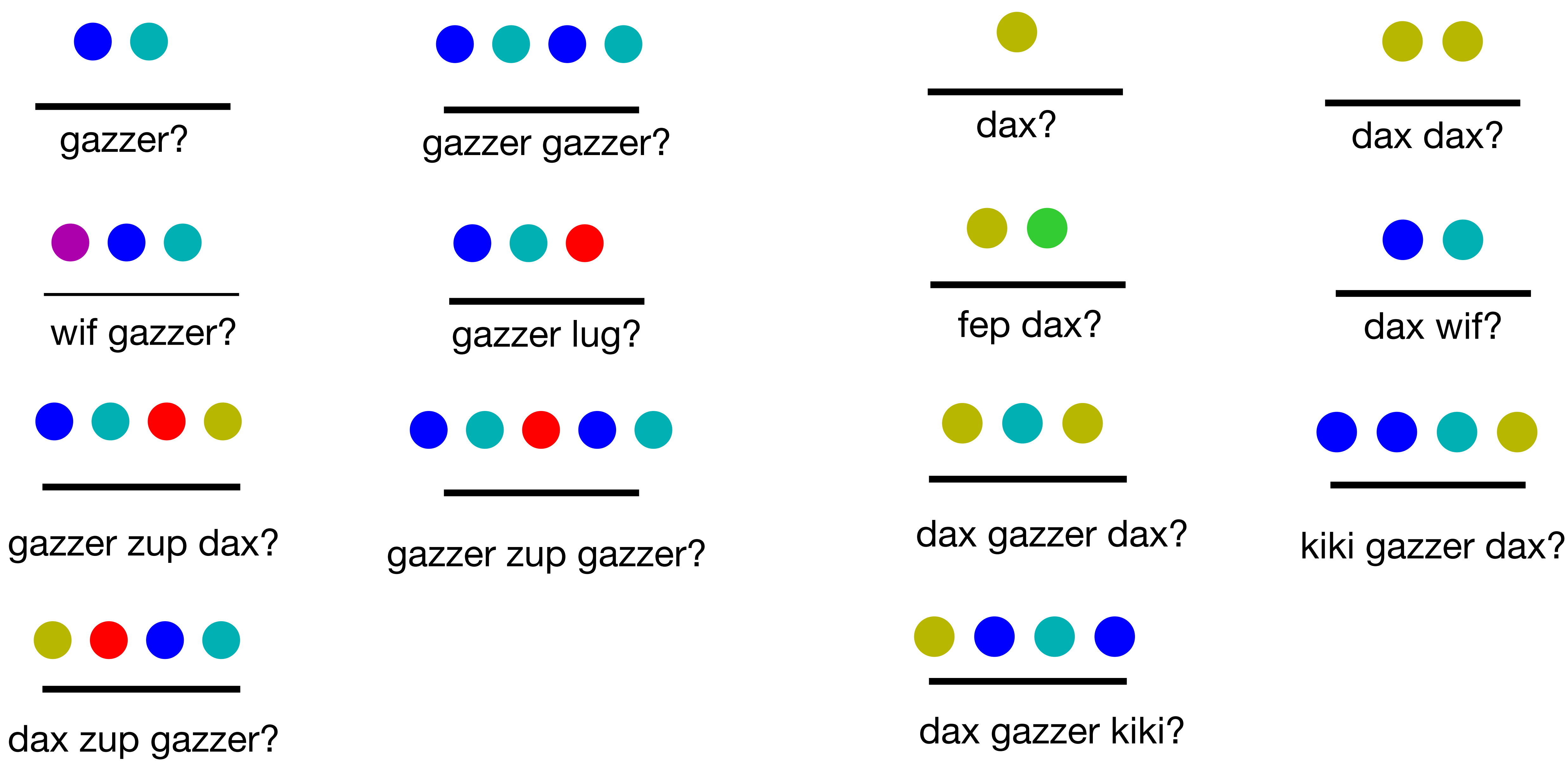


Experiment 2: Results

Alternative response 1

(follows IC, ME)

Alternative response 2



Goals of this work

1. Behavioral studies to compare humans and machines side-by-side on the same tests of systematicity

2. An approach to building neural networks that can achieve human-like systematic generalization, through an optimization procedure that encourages systematicity

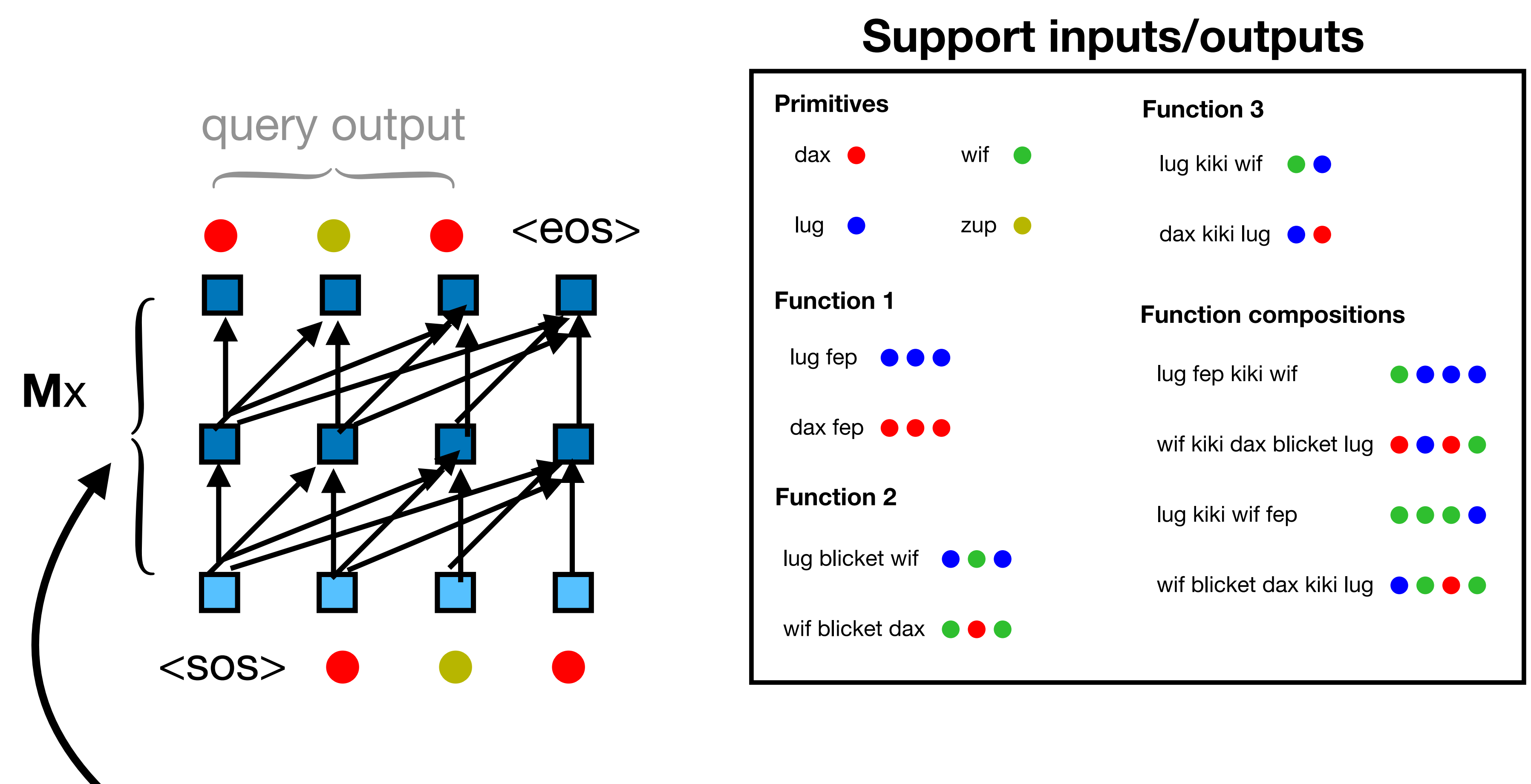
Goals for a computational framework

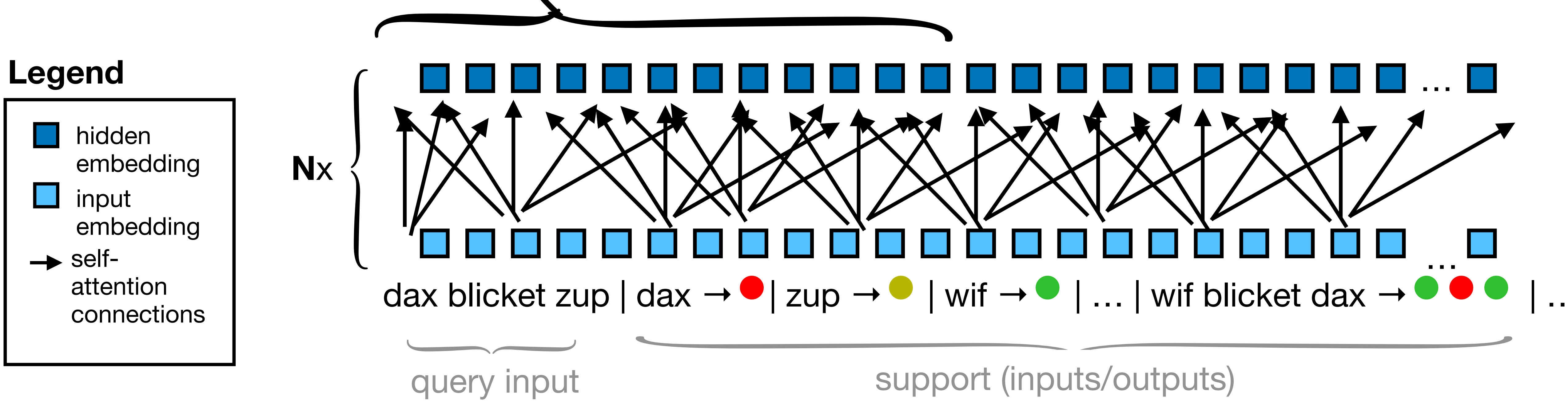
We would like neural network models that can do

- Few-shot induction of primitives and functions, and compose them flexibly and algebraically
- Prefer hypotheses that capture certain input/output regularities in meaning (1-to-1, IC, and ME)
- Model adult compositional skills (in this case, through meta-learning)
  - *Importantly*, we do not intend to model the process by which people acquire these skills

Behaviorally-Informed Meta-Learning (BIML)

- Specify desired behavior with high-level guidance and/or direct human examples
- Guides a neural network to parameter values that, when faced with a novel task, produce human-like generalizations and overcome challenges of systematicity

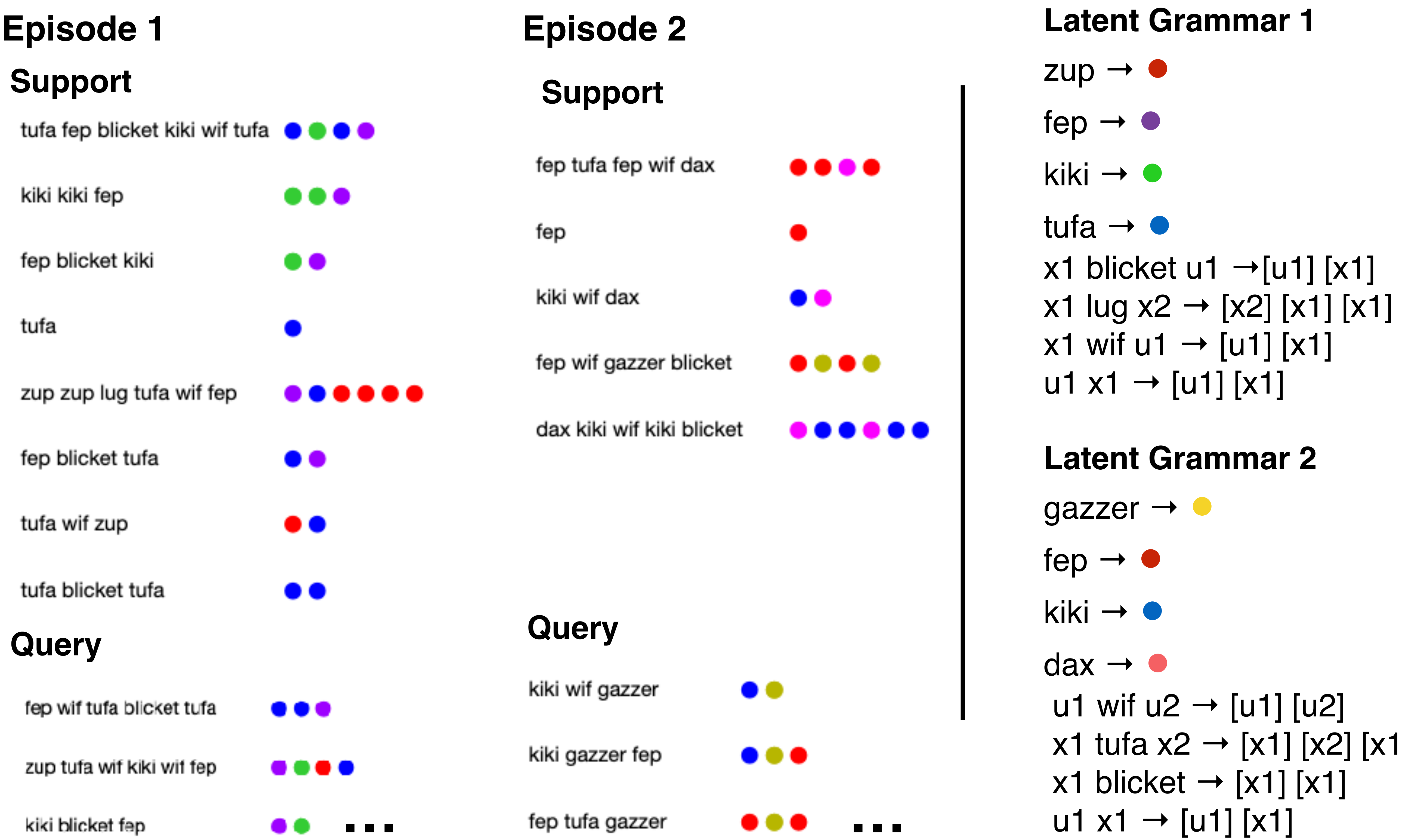




Behaviorally-Informed Meta-Learning (BIML)

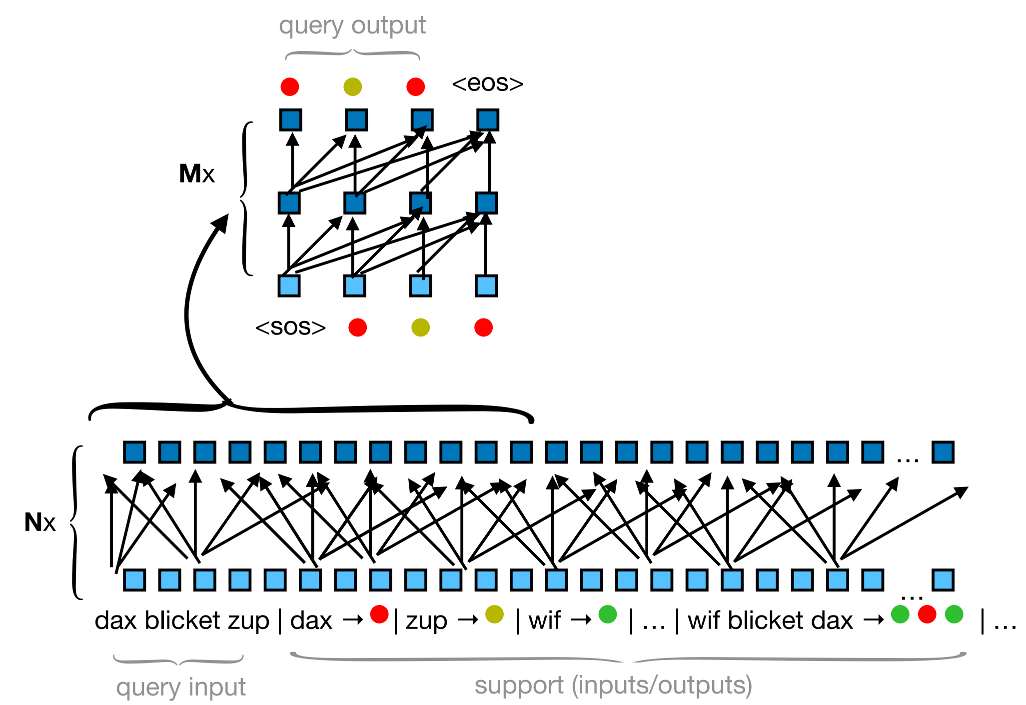
Optimization over a series of dynamically changing seq2seq tasks (episodes) that encourage systematic generalization (Lake, 2019, *NeurIPS*).

- Each episode samples a latent grammar, with 4 primitive and 3 compositional functions
- Queries paired with both grammar-based (algebraic) and biased-based outputs



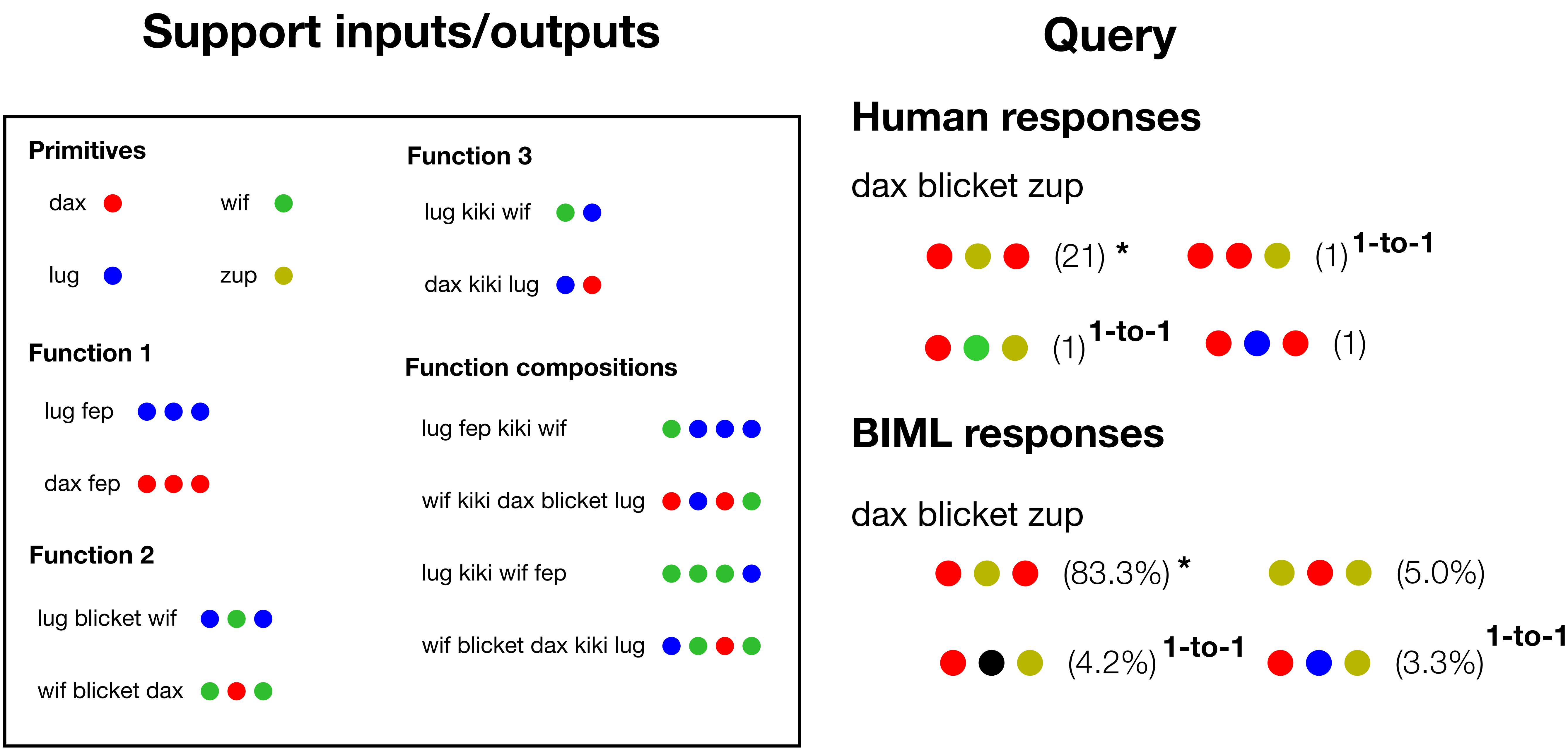
Comparing people and BIML on few-shot instruction learning

- After optimization, BIML's most likely outputs are perfectly systematic (100% consistent with grammar)
- When sampling over possible outputs, BIML accuracy (83%) is closer to human performance
- For predicting human responses (algebraic and bias-based)…

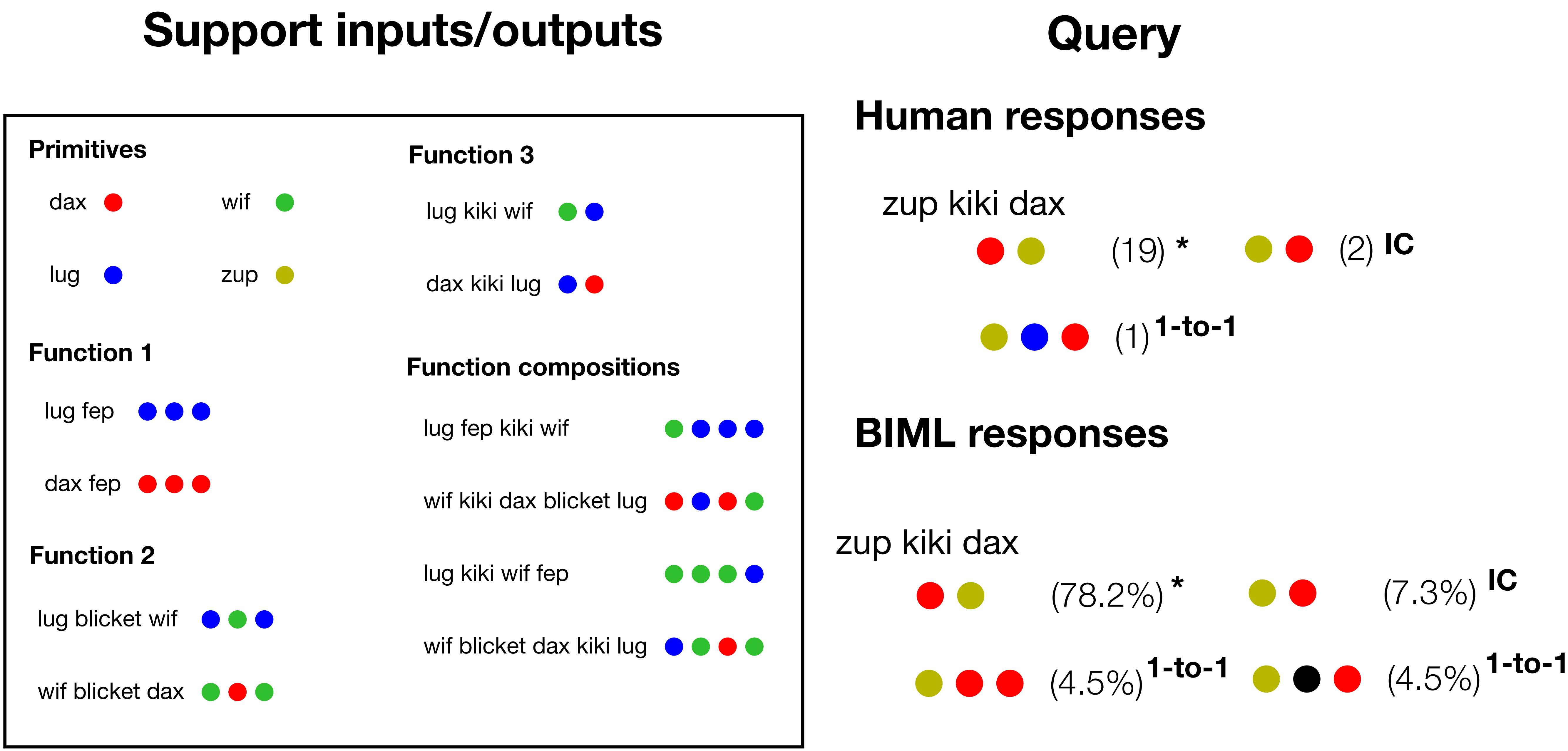


	Log-likelihood (larger is better)
Baseline	-1926.5
Symbolic (algebraic only)	-538.1
Symbolic (tuned)	-357.9
BIML (algebraic only)	-455.7
BIML	**-356.0**

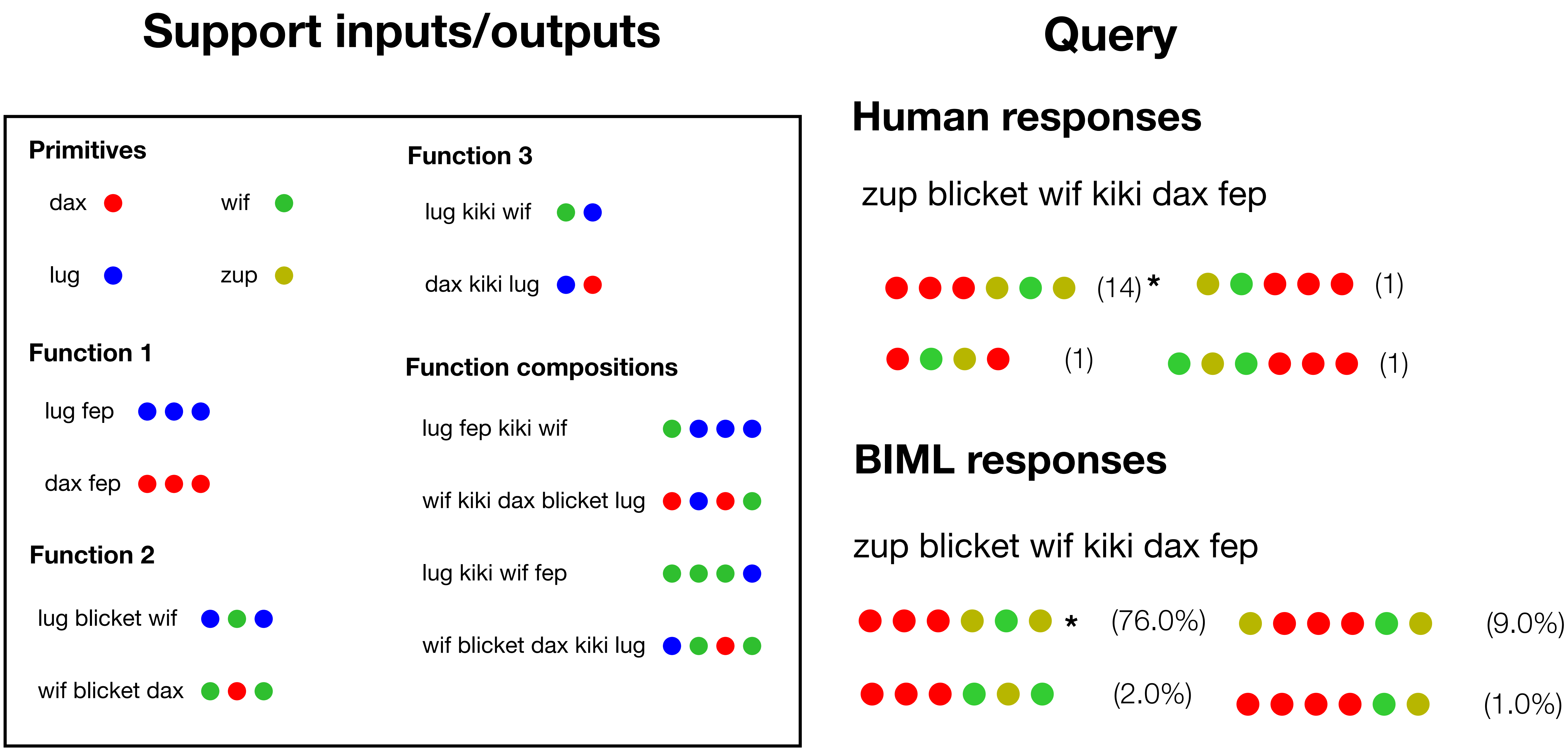
Comparing people and BIML on few-shot instruction learning



Comparing people and BIML on few-shot instruction learning



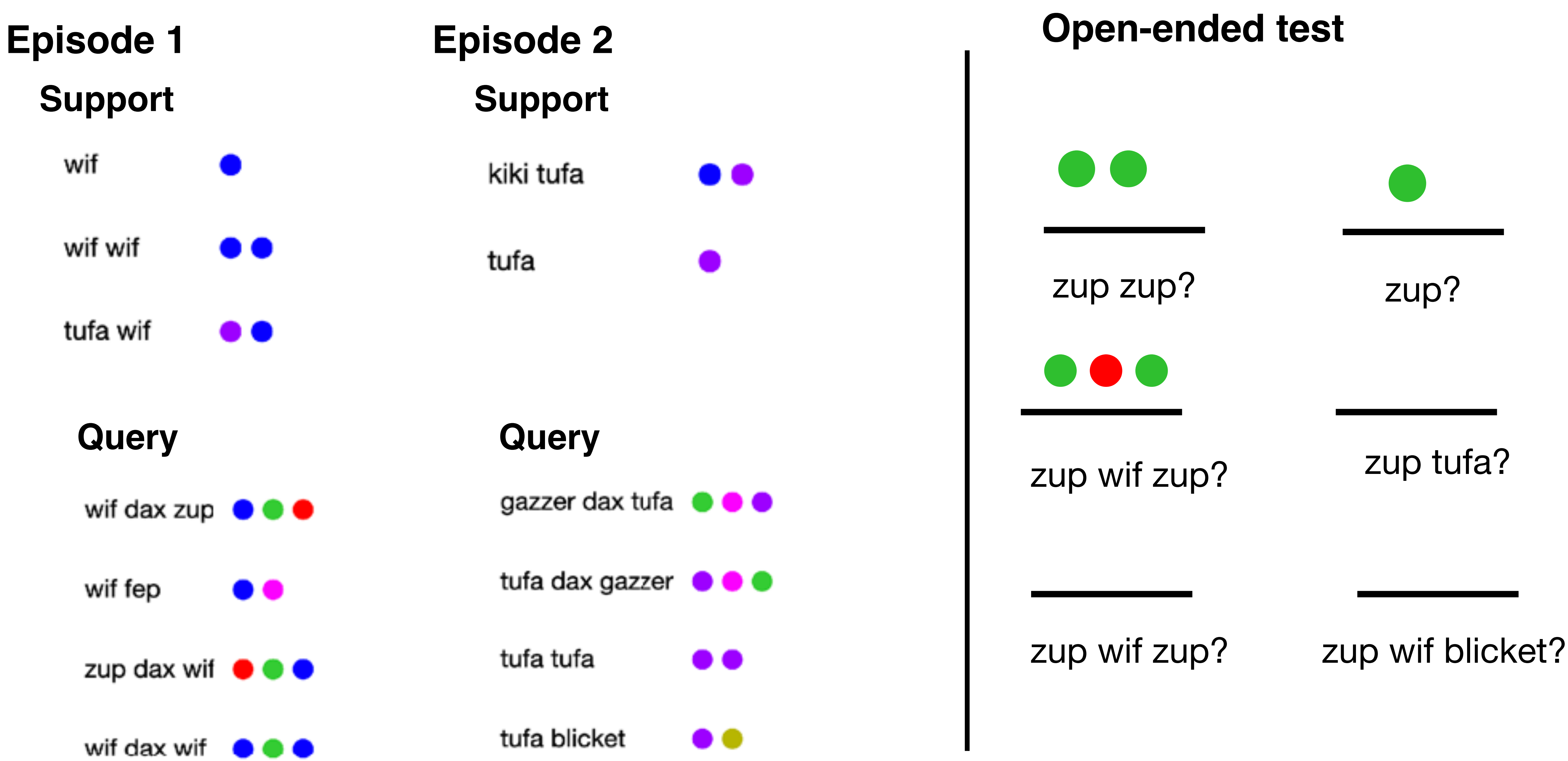
Comparing people and BIML on few-shot instruction learning



Comparing people and BIML on open-ended instruction task

Optimization over a series of dynamically changing seq2seq tasks (episodes).

- Episodes are based on augmented versions of human responses from Experiment 2
- Final model is evaluated on open-ended test task

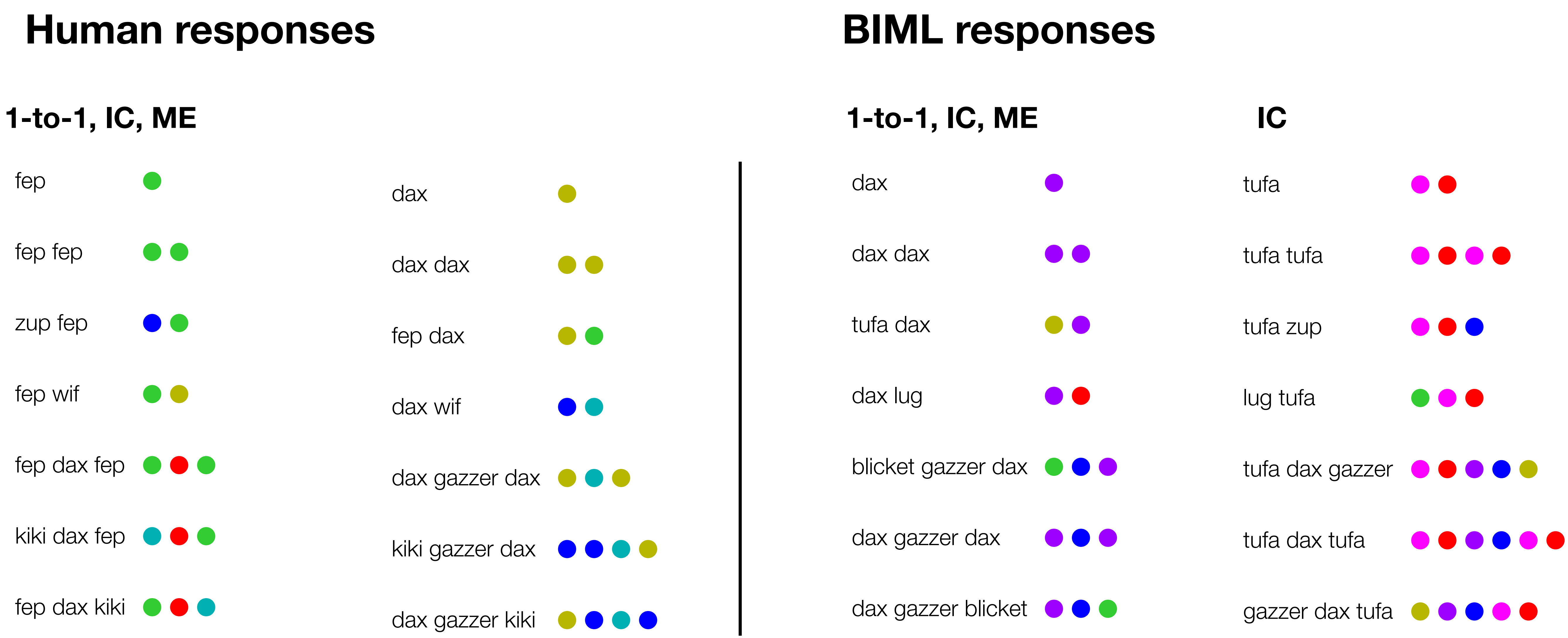


Comparing people and BIML on open-ended instruction task

- After optimization, 65% of BIML samples recreate the modal human response pattern (59% of people)
- For predicting human open-ended responses…

	Log-likelihood (larger is better)
Baseline	-173.2
Symbolic (tuned)	-92.6
BIML (algebraic only)	-150.1
BIML	**-64.2**

Comparing people and BIML on open-ended instruction task



Limitations and open questions

We would like neural network models that can do

✓ • Few-shot induction of primitives and functions, and compose them flexibly and algebraically
✓ • Prefer hypotheses that capture certain input/output regularities in meaning (1-to-1, IC, and ME)
✓ • Model adult compositional skills (in this case, through meta-learning)

Limitations and open questions

- How can a model learn entirely new primitives, rather than simply new primitive mappings?
- How do these abilities develop? How do people come to this rich starting point?

Conclusions

1. Despite remarkable progress in deep learning, F&P's (1988) article is still being debated today

2. Here, we used behavioral studies to compare humans and machines side-by-side on the same tests of systematicity
   - most common response is algebraic
   - People also rely on inductive biases that are good heuristics but can also lead people astray (1-to-1, IC, ME)

3. BIML shows how neural nets can achieve human-like systematic generalization, through an optimization procedure that encourages systematicity.

4. Hopefully informs engineering efforts to build more capable and more human-like AI systems