Human-like Category Learning by Injecting Ecological Priors from Large Language Models into Neural Networks

Akshay K Jagadish, Julian Coda-Forno, Mirko Thalmann,
Eric Schulz*, & Marcel Binz*

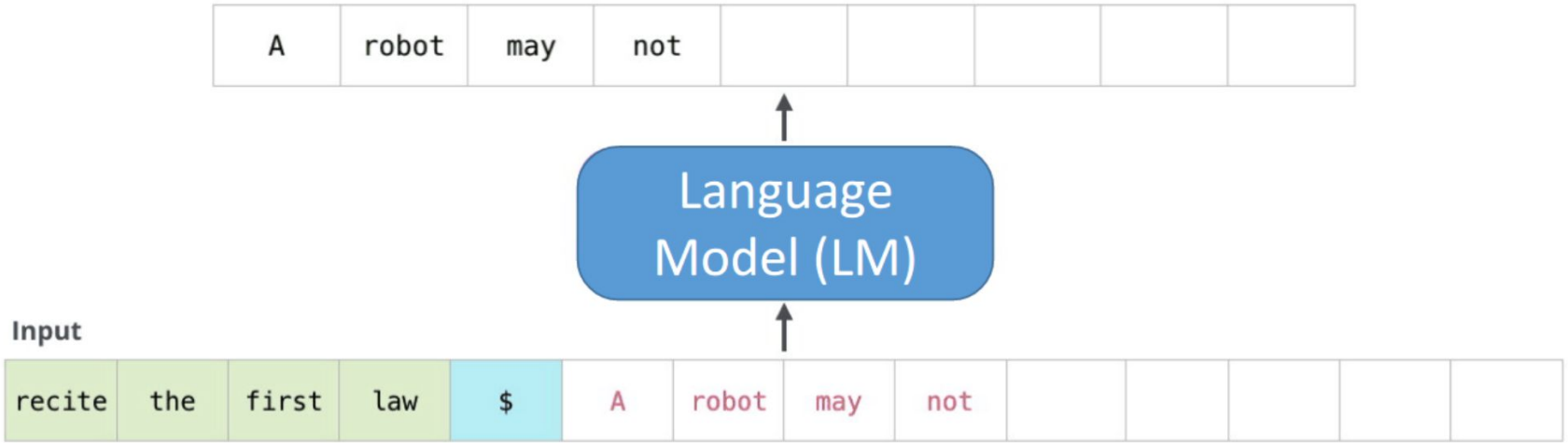
*shared authorship

Ecological rationality

Simon et al. 1953, 1990.; Goldstein and Gigerenzer, 2002

how do we define ecologically valid tasks?

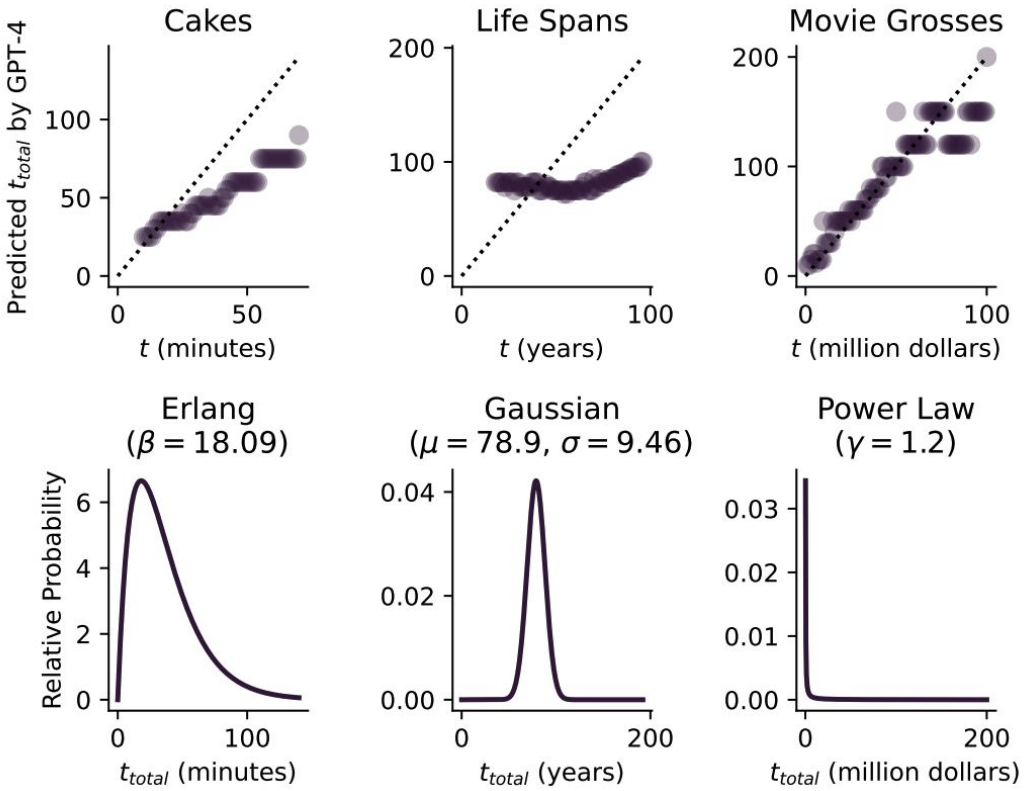
Large Language Models (LLMs)



“LLMs [..] are powerful new cultural technologies, analogous to earlier technologies like writing, print, libraries, internet search and even language itself.”

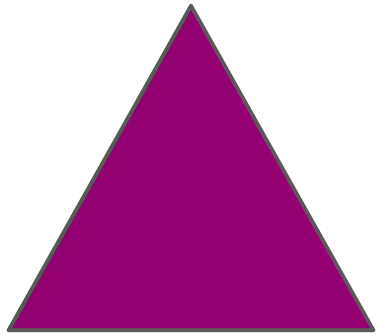
Yiu,...,Gopnik, 2023

LLMs internalize everyday distributions similar to humans



Griffiths et al. 2023 Griffiths and Tenenbaum 2006

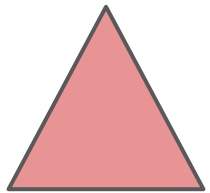
Can LLMs be used to generate ecologically valid tasks?



Choose A or B?

Choice: A

Correct Choice!

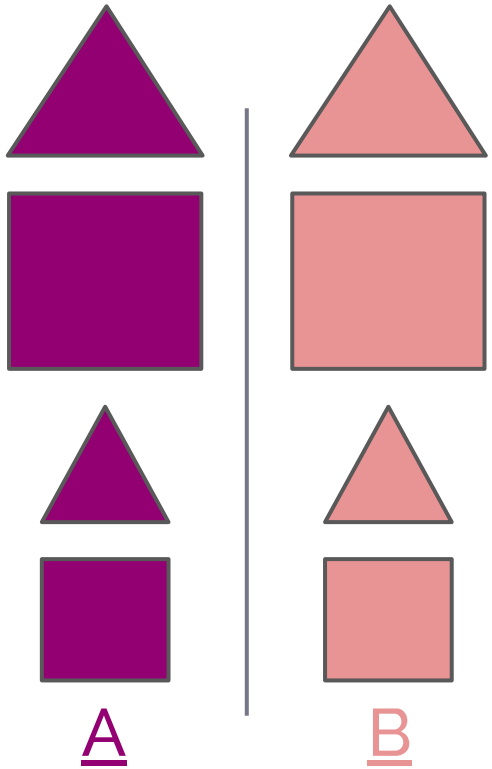


Choose A or B?

Choice: A

Wrong Choice!

Category learning



Canonical task in psychology (Shepard et al. 1961)

Effects replicated successfully (Nosofsky et al. 1986, Medin et al. 1978, Badham et al. 2017, Devraj et al. 2022)

Many models of human categorisation exist (Smith et al. 1998, Anderson et al. 1991, Sandborn et al. 2010, Griffiths et al. 2017)

Shepherd et al. 1961; Sandborn et al. 2010; Nosofsky 1984; Levering 2019; Nosofsky 1988;

How can we generate category learning tasks using LLMs?

Two step approach

Step 1: Synthesize feature names and category labels

Step 2: Generate data for category learning tasks

Quick side note on LLMs considered

Achieved varying degrees of success with

- LLaMA
- GPT-3.5
- GPT-4

Results reported for **Claude-2**

- process up to 100k tokens
- instruction tuned
- good performance on preliminary tests
- temperature set to 1.

Two step approach

Step 1: Synthesize feature names and category labels

Step 2: Generate data for category learning tasks

Prompt

I am a psychologist who wants to run a category learning experiment. In a category learning experiment, there are many different {three}-dimensional stimuli, each of which belongs to one of two possible real-world categories.

Please generate names for {three} stimulus feature dimensions and {two} corresponding categories for {250} different category learning experiments:
- feature dimension 1, feature dimension 2, ..., feature dimension {3}, category label 1, category label 2

Feature 1	Feature 2	Feature 3	Category 1	Category 2
sodium	fat	protein	healthy	unhealthy
price	comfort	mileage	economy car	luxury car
rhythm	melody	harmony	jazz	classical

Two step approach

Step 1: Synthesize feature names and category labels

Step 2: Generate data for category learning tasks

Prompt

I am a psychologist who wants to run a category learning experiment. For a category learning experiment, I need a list of stimuli and their category labels. Each stimulus is characterized by {three} distinct features: {sodium}, {fat}, and {protein}. These features can take only numerical values. The category label can take the values {healthy} or {unhealthy} and should be predictable from the feature values of the stimulus.

Please generate a list of {400} stimuli with their feature values and their corresponding category labels using the following template for each row:

- feature value 1, feature value 2,..., feature value {3}, category label

An example task

Sodium	Fat	Protein	Category
250	15	20	healthy
220	17	22	unhealthy
320	26	31	unhealthy
145	11	20	healthy

Is there an emerging theme in the LLM generated tasks?


mammals
kitchenknives
vodka
electric
english
scotch
nonmetals
classical
dramas
electricguitars
suvs
cabernetsauvignon
laptops
roses
conifers
reptiles
trucks
ales
lagers
lions
movies
tropical
comedy
sedans
pinotnoir
insects
birds
bourbon
europe
vegetables
diamonds
stouts
metals
runningshoes
ipas
sportscars
rye
lattes
fish
chardonnay
acoustic
rock
desktops
jazz
pop
drama
tigers
fruits
comedies
acousticguitars

Do these tasks capture real-world statistics?

Ecological-validity of LLM-generated tasks

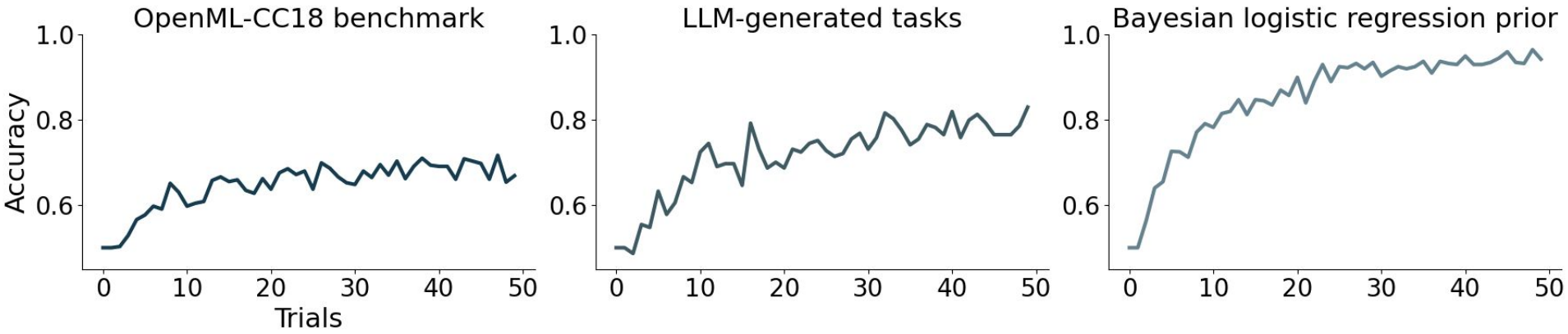
Comparison:

- OpenML-CC18 classification benchmark (Bischl et al. 2019)
- Bayesian logistic regression prior (Binz et al. 2022, Speekenbrink et al. 2008, 2010)
- Bayesian neural network prior (Müller et al. 2022, Levering et al. 2020)

Statistics:

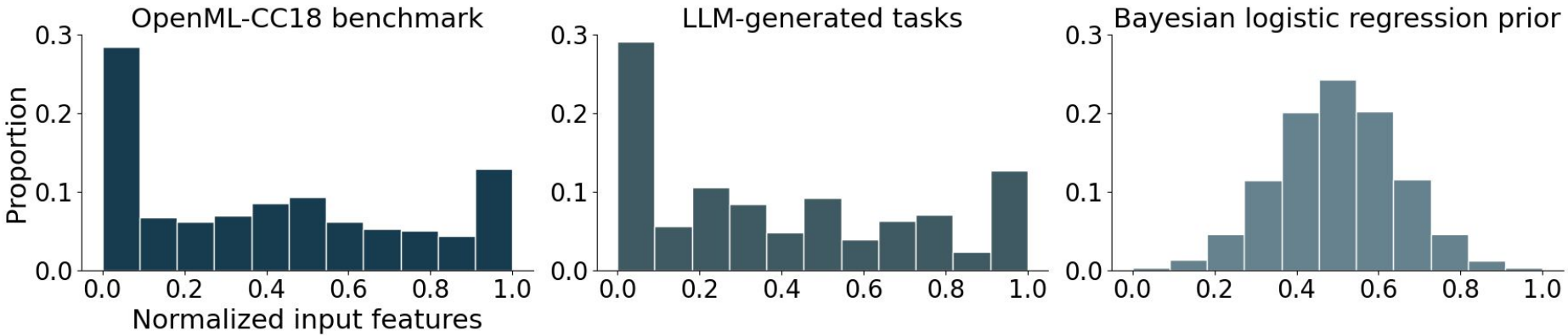
1. Classification performance
2. Normalized input features
3. Pair-wise input feature correlations
4. Sparsity
5. Linearity

Learning curves of LLMs match real-world data



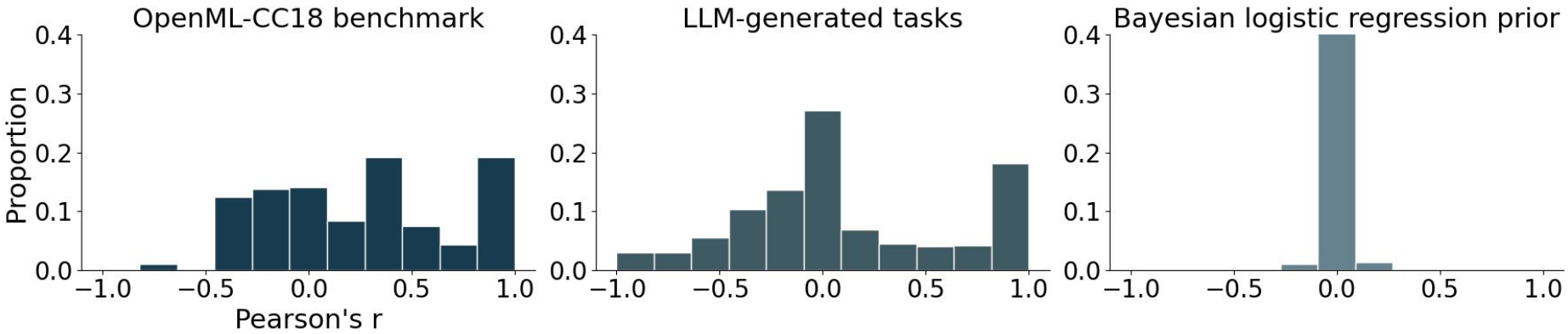
Jagadish et al. 2024

Distribution of normalized input features match real-world data



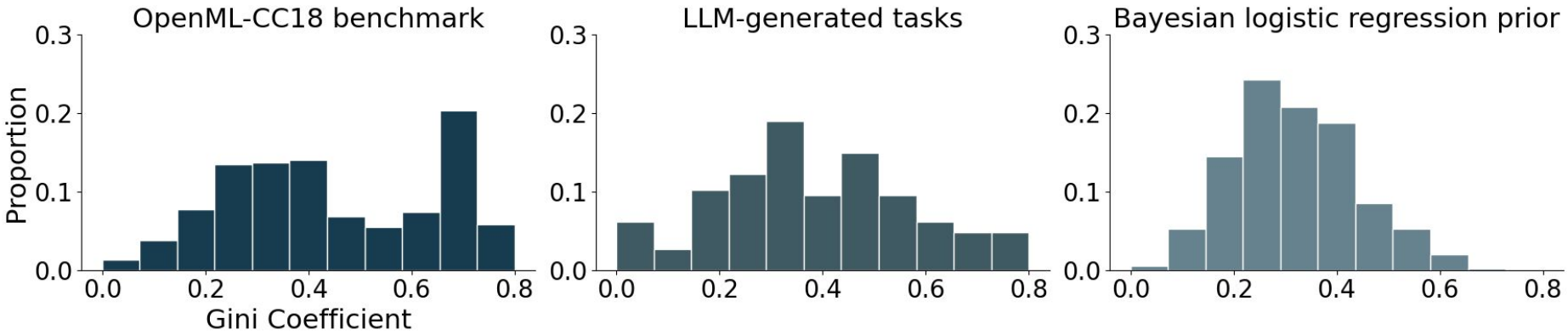
Jagadish et al. 2024

Distribution of input feature correlations match real-world data



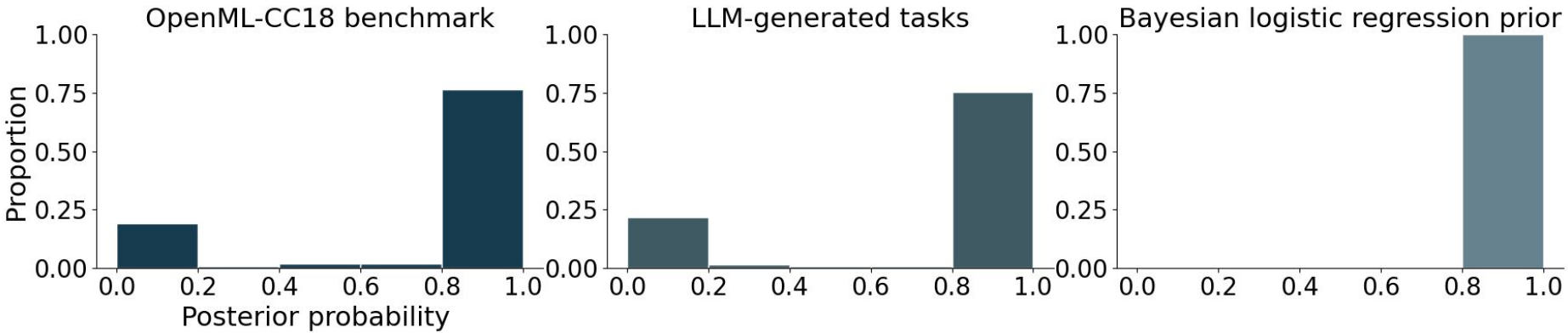
Jagadish et al. 2024

Distribution of sparsity match real-world data



Jagadish et al. 2024

Proportion of linear tasks match real-world data



Jagadish et al. 2024

Interim summary

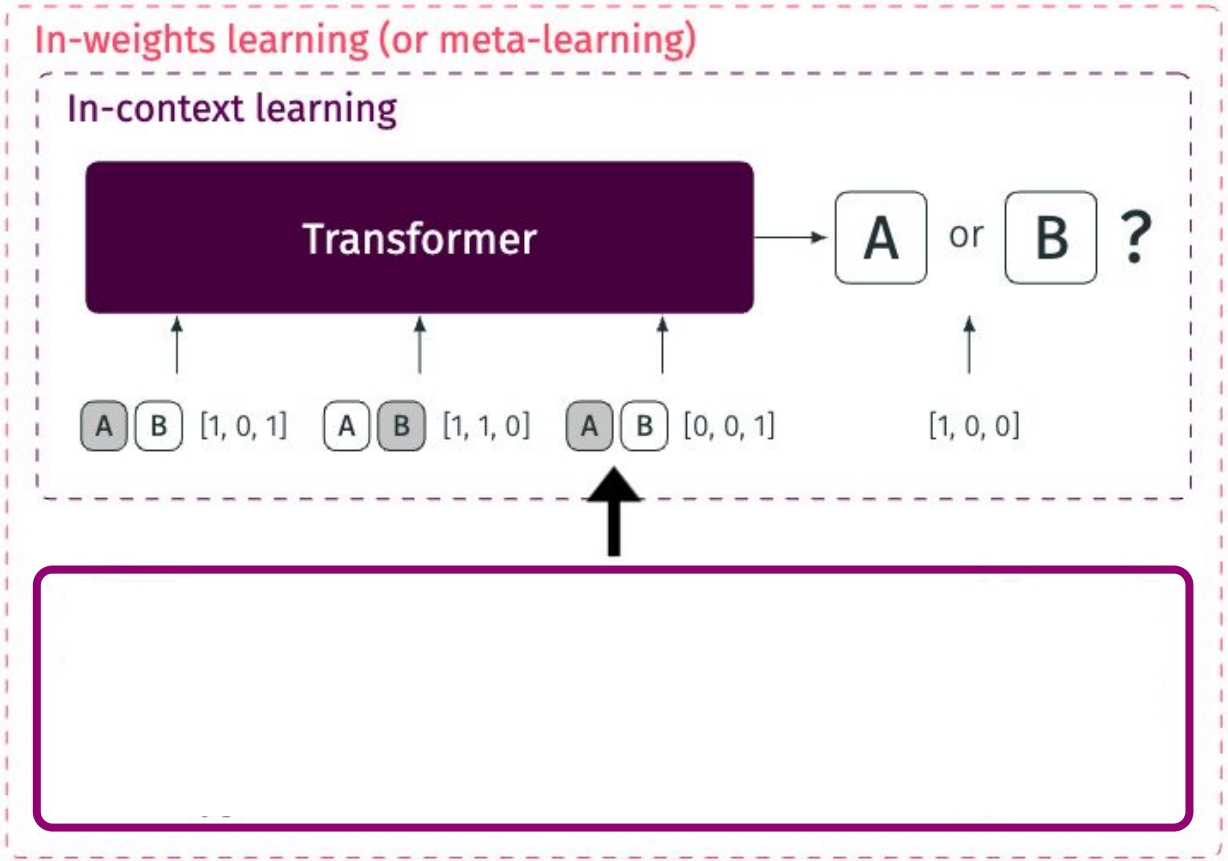
1. Two step approach to generate category learning tasks from LLMs

2. Statistics of tasks generated by LLMs match real-world classification tasks

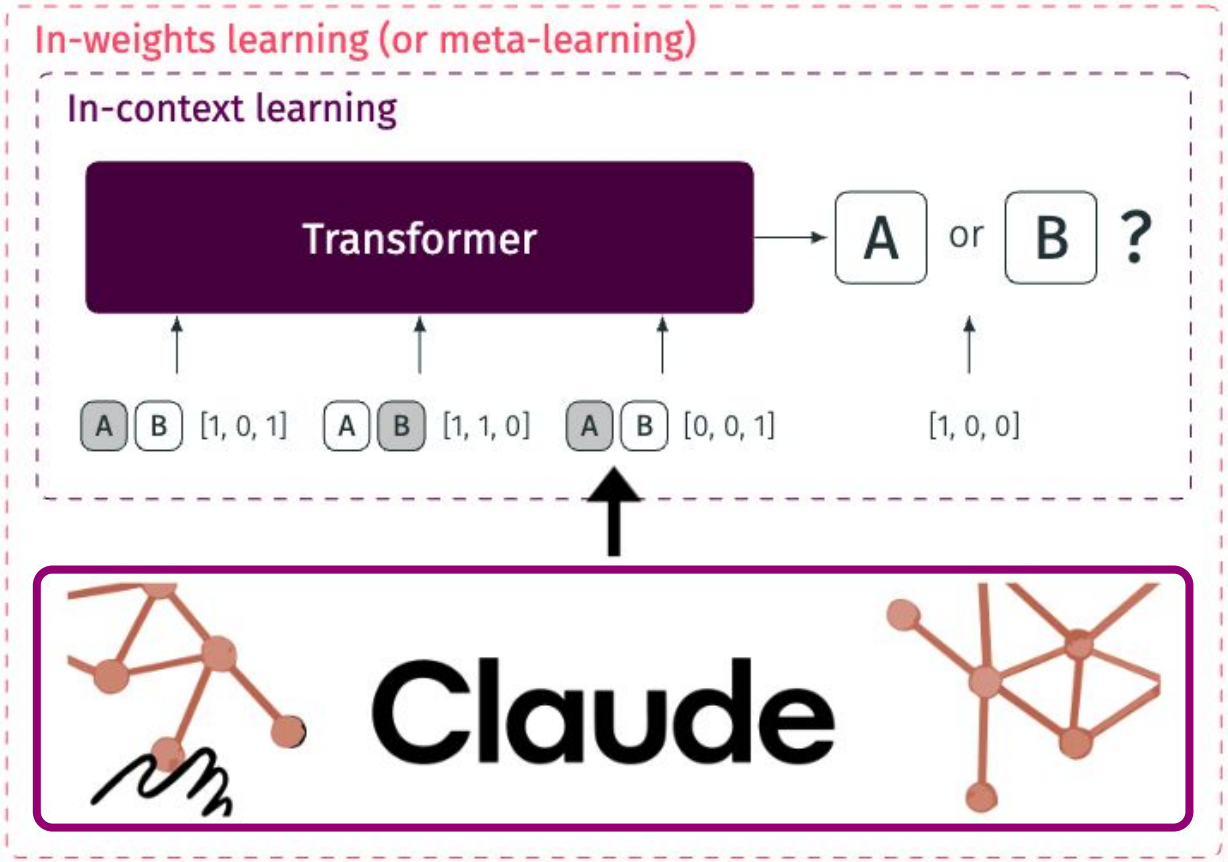
Interim summary

Simon et al. 1953, 1990.

Derving an ecologically rational model



Jagadish et al. 2024



Jagadish et al. 2024

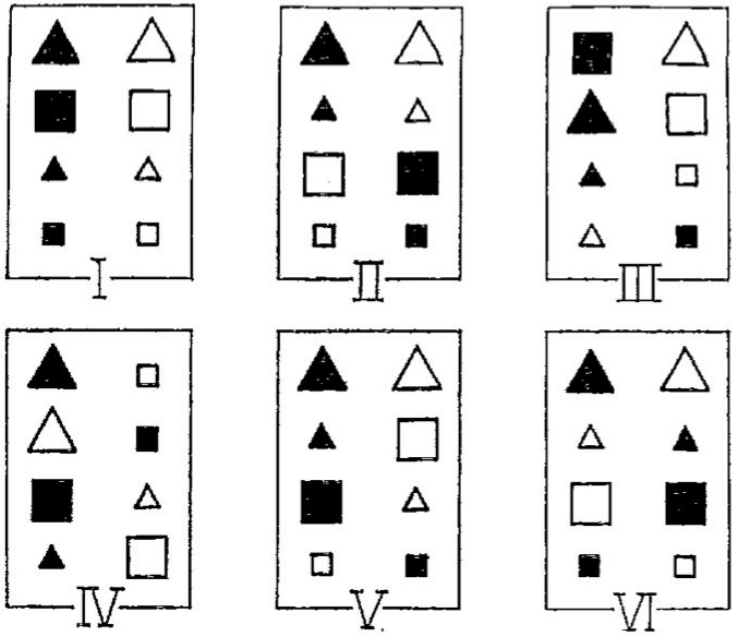
How much of human behavior does ERMI explain?

We looked at three different effects in human category learning:

1. Learning difficulties
2. Learning strategy
3. Generalisation

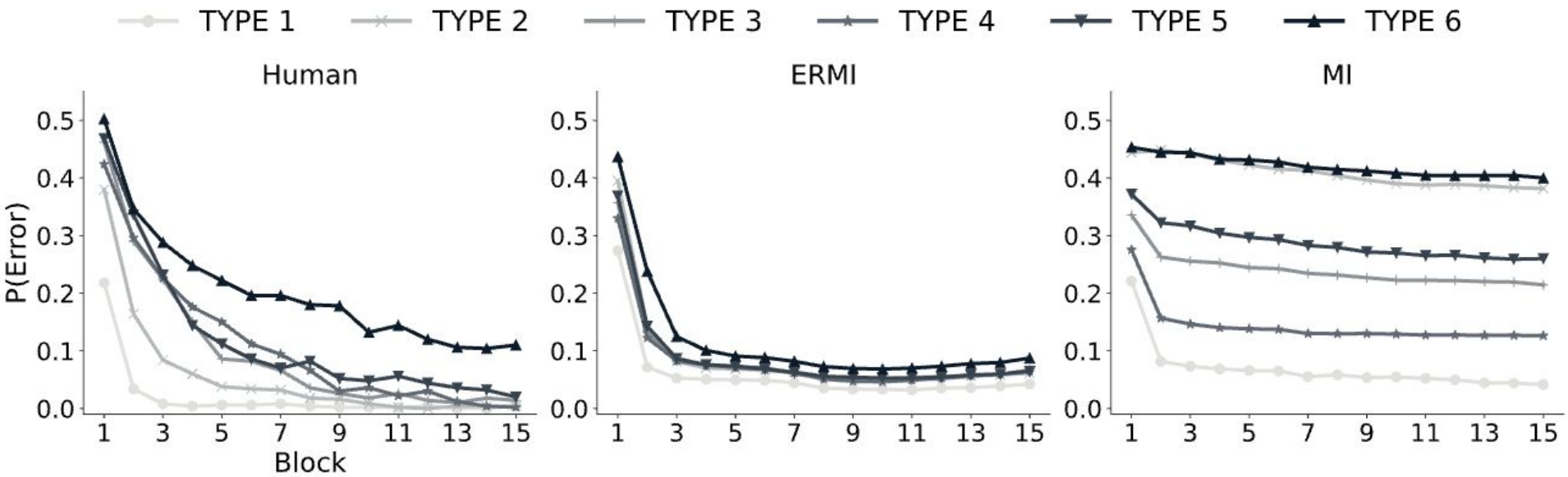
Can ERMI explain human learning difficulties?

Task difficulty increases from Type I to Type VI



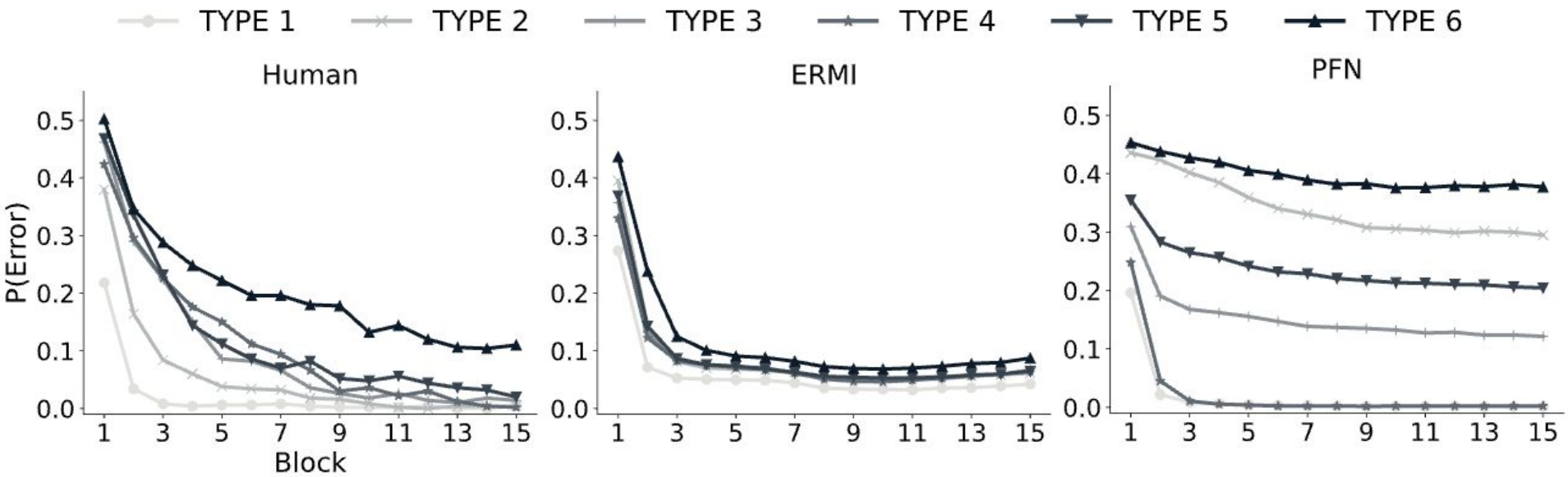
Sherpard et al. 1961; Nosofsky et al. 1988

ERMI shows human-like learning difficulties



Sherpard et al. 1961; Nosofsky et al. 1988;; **Jagadish** et al. 2024

ERMI shows human-like learning difficulties

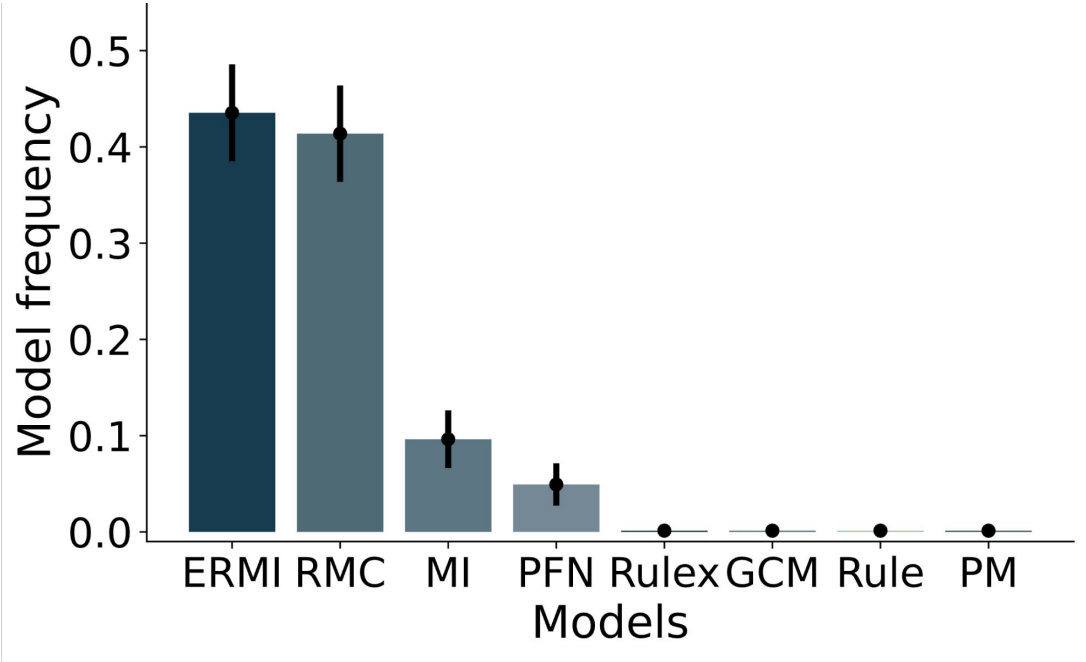


Sherpard et al. 1961; Nosofsky et al. 1988; **Jagadish** et al. 2024

ERMI explains human data from Badham et al. 2017 better than cognitive models

1. **MI:** Meta-learned Inference on Bayesian logistic regression prior (Binz et al. 2022, Speekenbrink et al. 2008, 2010)
2. **PFN:** Meta-learned inference on Bayesian neural network prior (Müller et al. 2022, Levering et al. 2020)
3. **RMC:** Rational model of categorisation (Anderson et al. 1991)
4. **GCM:** Generalized context model (Nosofsky, 1986)
5. **PM:** Prototype Model (Homa and Cultice, 1984)
6. **Rule:** Rule-based model (Ashby and Townsend, 1986)
7. **Rulex:** Rule plus exception model (Nosofsky, 1994)

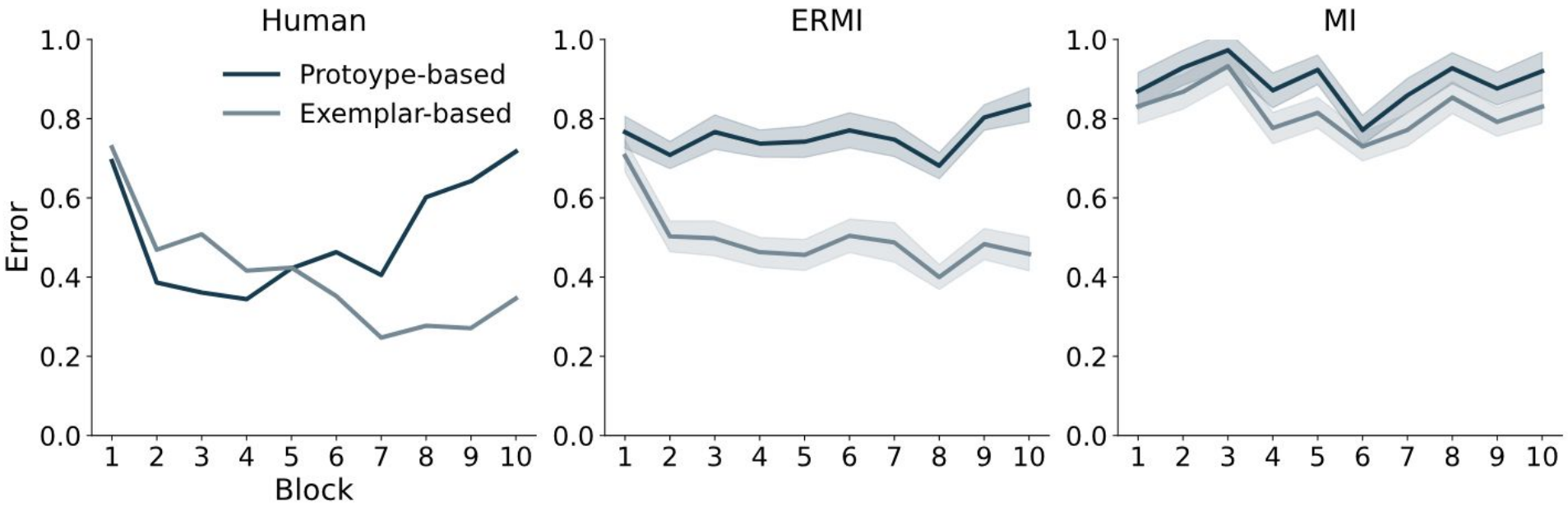
ERMI explains human data from Badham et al. 2017 better than cognitive models



Rigoux et al. 2014; Badham et al. 2017; **Jagadish** et al. 2024

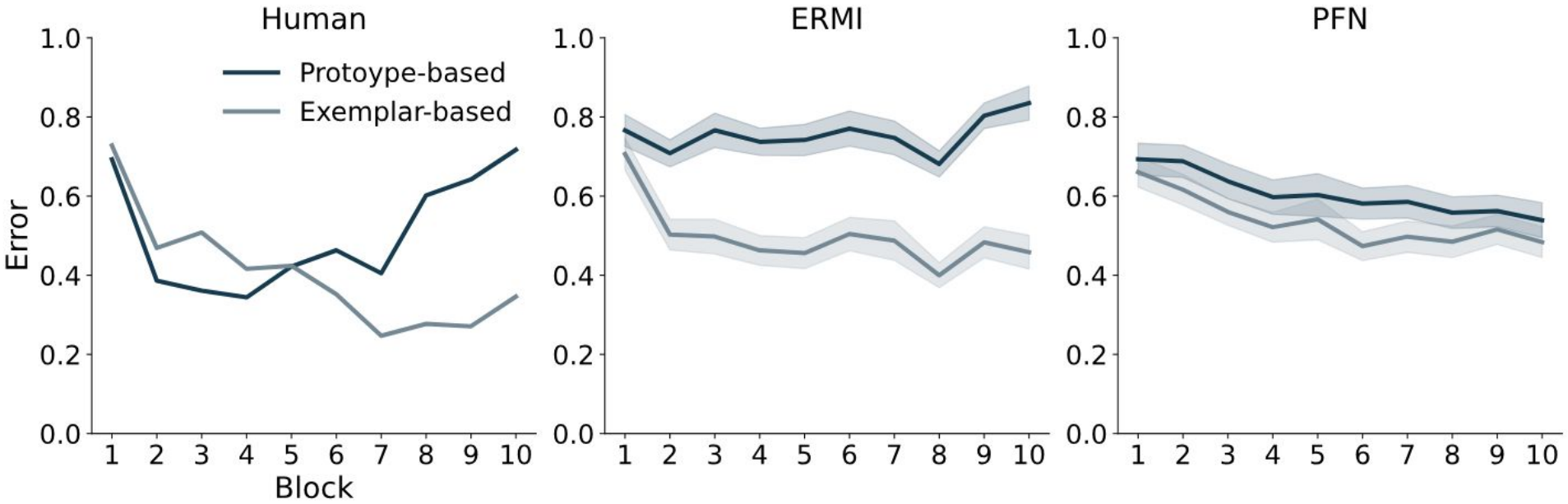
Does ERMI shift to exemplar-based learning strategy?

ERMI also becomes more exemplar-based with learning



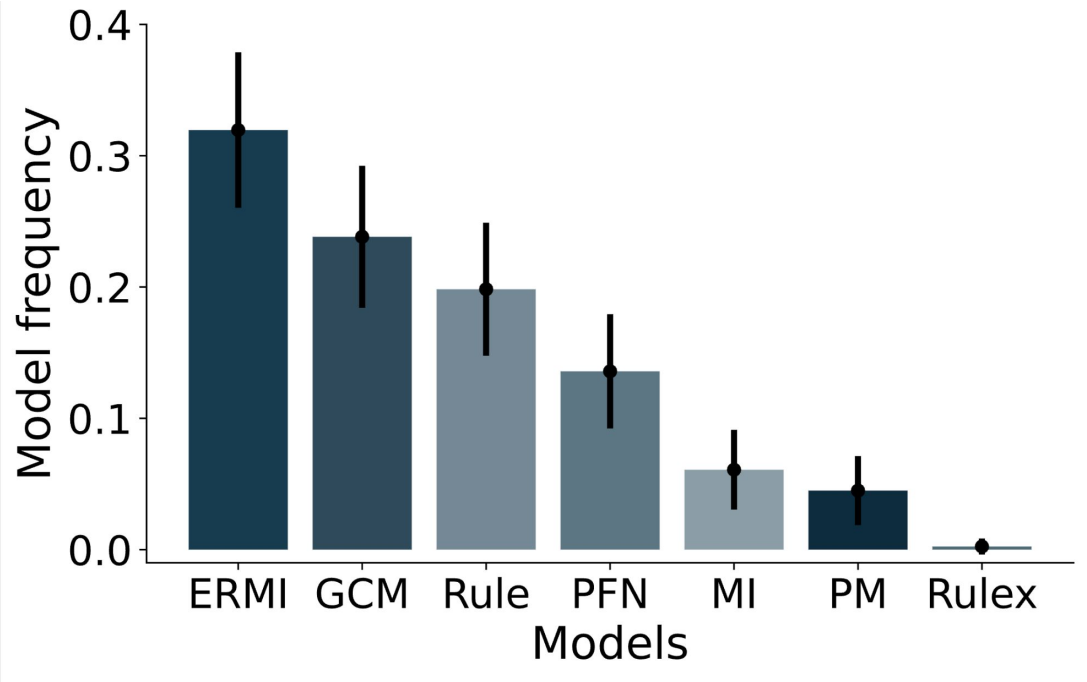
Smith et al. 1998; **Jagadish** et al. 2024

ERMI also becomes more exemplar-based with learning



50

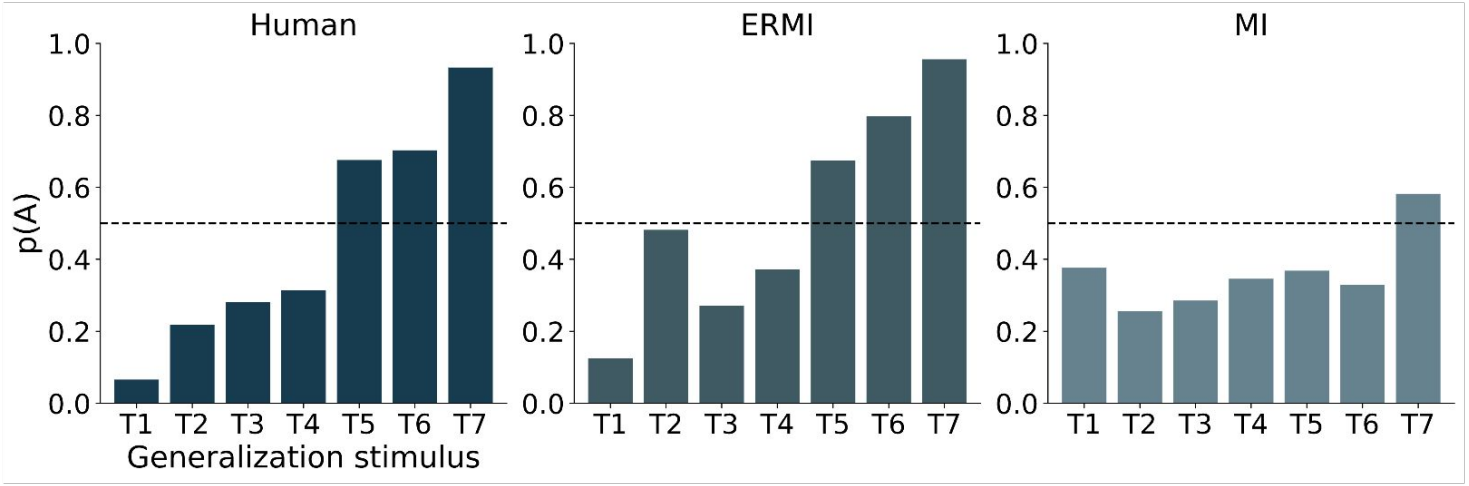
ERMI explains human data from Devraj et al. 2022 better than cognitive models



Devraj et al. 2022; **Jagadish** et al. 2024

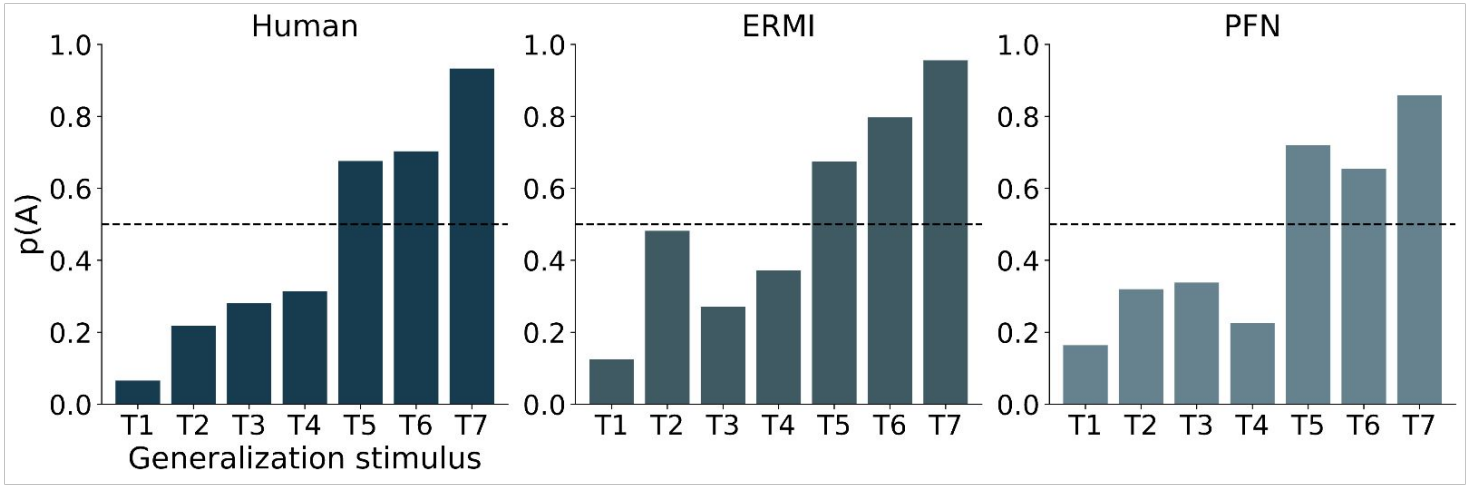
Does ERMI generalize to unseen stimuli like humans?

Johannsen et al. 2002



T1: [1 0 1 1]
T2: [1 0 1 0]
T3: [0 1 1 1]
T4: [1 1 0 1]
T5: [1 1 0 0]
T6: [0 1 1 0]
T7: [0 0 0 0]

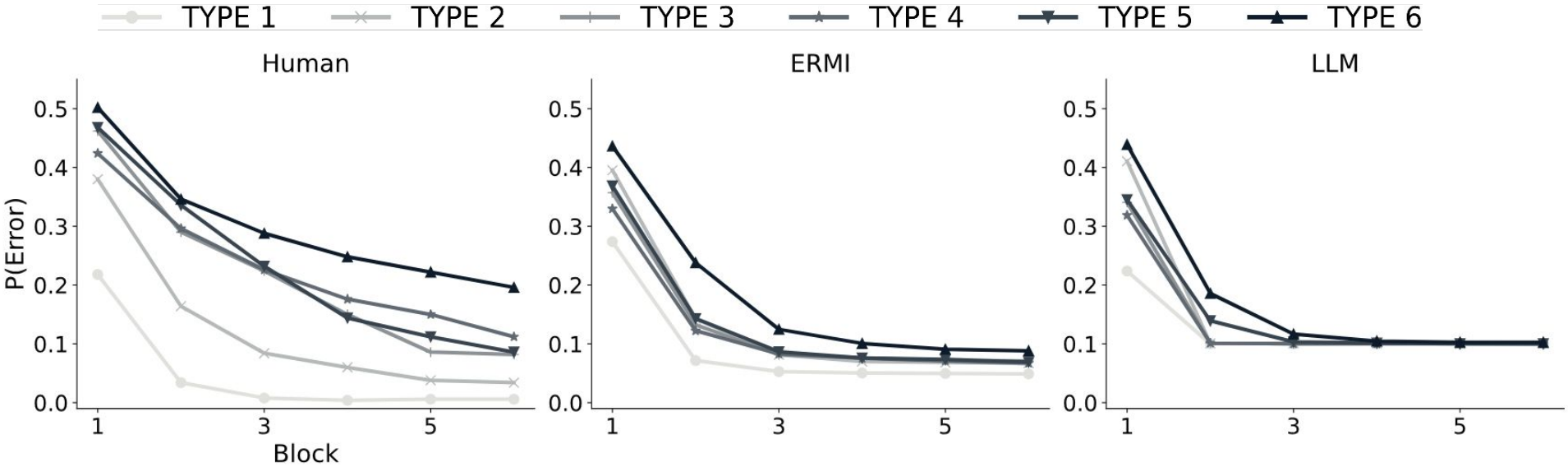
Johannsen et al. 2002



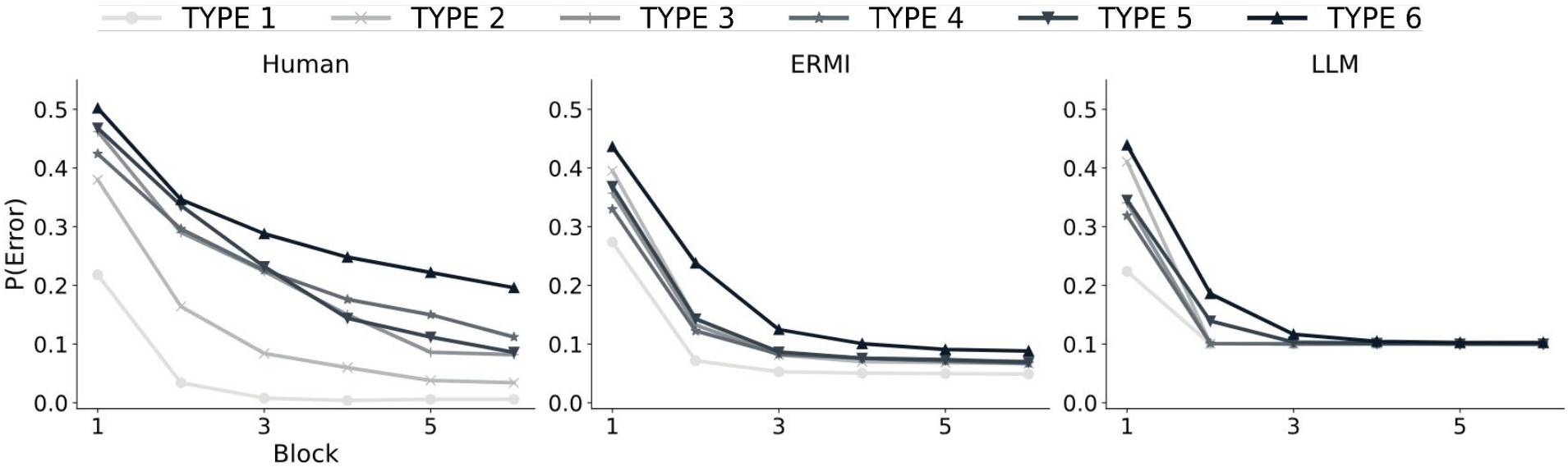
T1: [1 0 1 1]
T2: [1 0 1 0]
T3: [0 1 1 1]
T4: [1 1 0 1]
T5: [1 1 0 0]
T6: [0 1 1 0]
T7: [0 0 0 0]

What if we directly let an LLM perform a category learning task?

ERMI is better fit to humans both qualitatively and quantitatively in Shepard et al. 1961

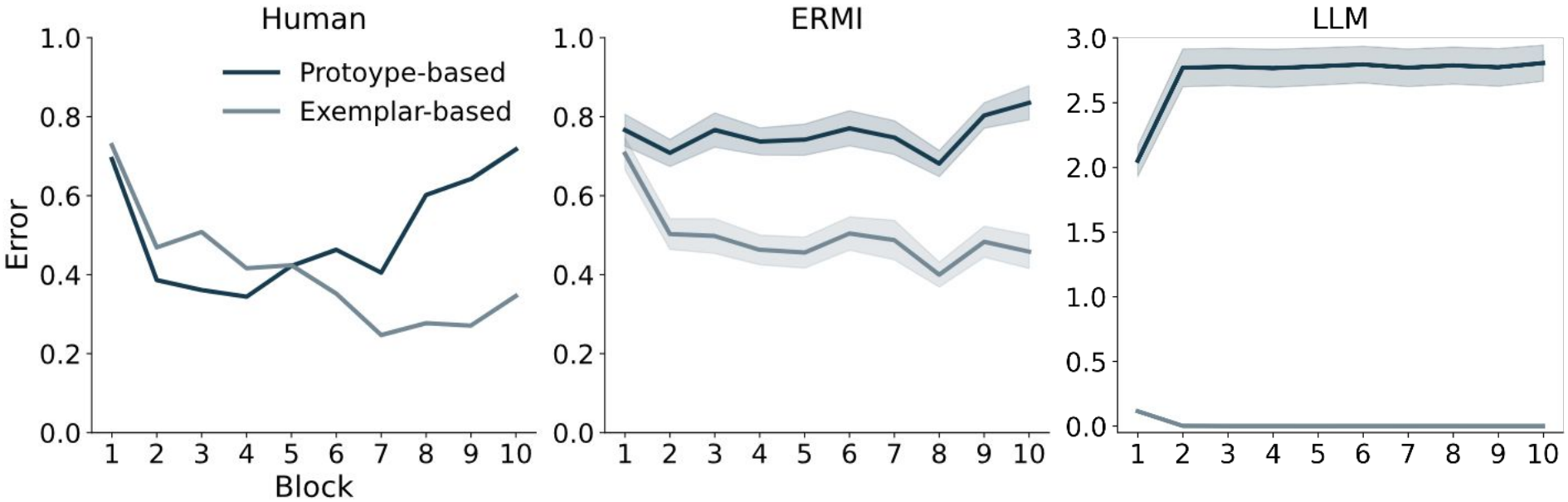


ERMI is better fit to humans both qualitatively and quantitatively in Shepard et al. 1961



ERMI also offers a better fit to humans than LLM in term of BIC

ERMI is better fit to humans both qualitatively and quantitatively to Smith and Minda, 1998



ERMI also offers a better fit to humans than LLM in term of BIC

Does ERMI perform well on ML tabular classification tasks?

ERMI achieves state-of-the-art performance on machine learning benchmarks

Tabular classification benchmark based on OpenML-CC18:

- 23 binary classification tasks (less than 100 features)
- 30 training points and 70 testing points (Müller et al. 2022)
- reduced the input dimensionality to 4 (based on highest ANOVA F-value to the target)

Table 1. Performance metrics on OpenML-CC18 benchmark.

MEAN	SVM	XGBOOST	TABPFN	ERMI
ACC.	69.29%	70.17%	70.51%	**70.95%**
RANK	2.76	2.61	2.85	**2.26**

Summary

- Large language models can generate ecologically valid data
- A class of models called ecologically rational meta-learned inference (ERMI)
- ERMI displays human-like category learning in three experiments
  - human-like learning difficulties
  - human-like learning strategies
  - human-like generalization
- ERMI achieves state-of-the-art performance on ML Benchmarks

Surprising how far we can go just by meta-learning on the right data!

Thank you :)

Collaborators and colleagues from the CPI Lab

Scraping real-world classification tasks

OpenML Curated Classification Benchmark (CC-18)

- real referenced datasets
- 500 <= #observations <=10000
- #classes >=2, each with at least two observations for each
- not predictable using single feature or simple decision tree

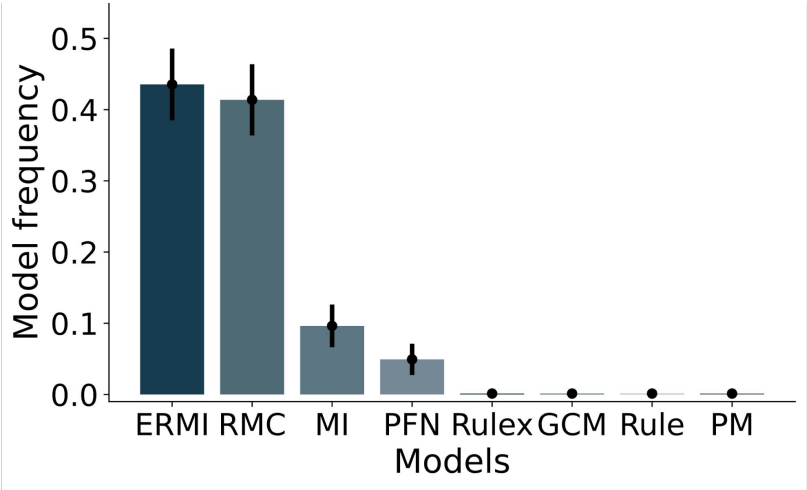
Filtering criteria

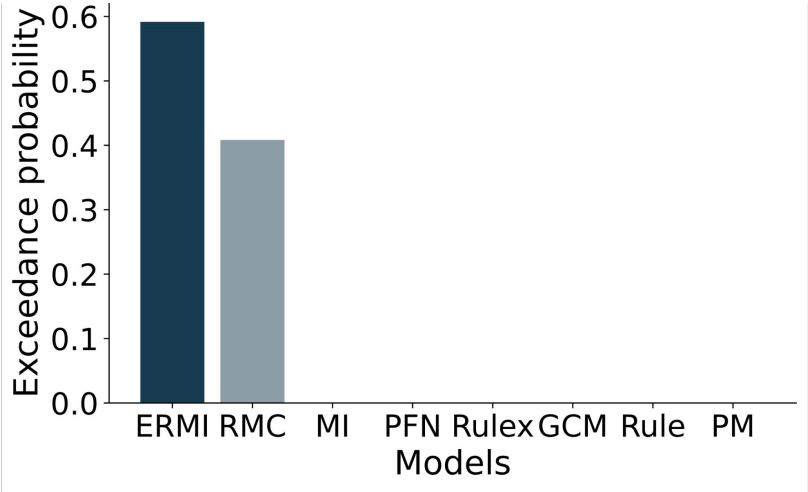
- binary classification
- not Nan
- #features < 100

Resulted in a total of 24 datasets

Bischl et al. 2019.

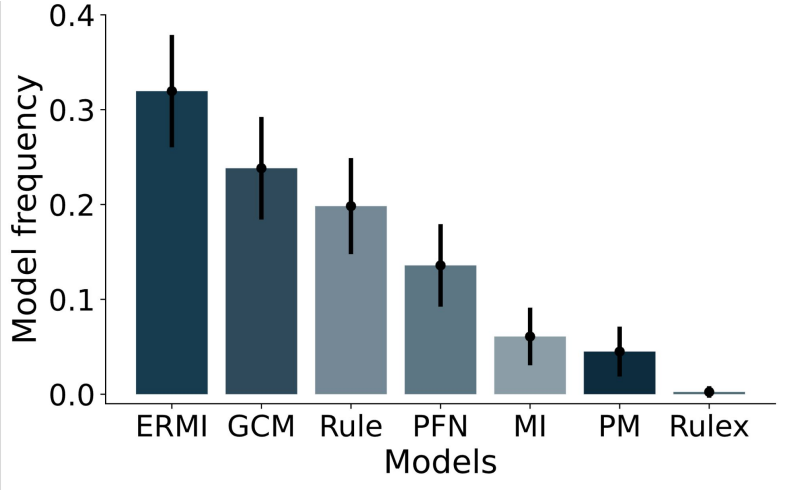
ERMI explains human data from Badham et al. 2017 better than cognitive models

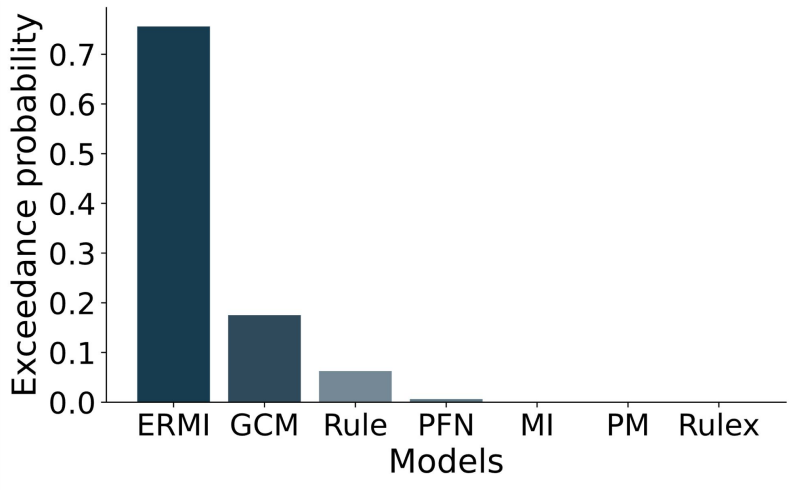




Rigoux et al. 2014; Badham et al. 2017; **Jagadish** et al. 2024

ERMI explains human data from Devraj et al. 2022 better than cognitive models





Devraj et al. 2022; **Jagadish** et al. 2024

Prompt for Badham et al. 2017 study

In this experiment, you will be shown examples of geometric objects. Each object has three different features: size, color, and shape. Your job is to learn a rule based on the object features that allows you to tell whether each example belongs in the {A} or {B} category. As you are shown each example, you will be asked to make a category judgment and then you will receive feedback. At first you will have to guess, but you will gain experience as you go along. Try your best to gain mastery of the {A} and {B} categories.

- In trial 1, you picked category {A} for Big Black Square and category {A} was correct.
- In trial 2, you picked category {A} for Small Black Triangle and category {B} was correct

Human: What category would a Small Black Triangle belong to? (Give the answer in the form "Category ⟨your answer⟩").
Assistant: Category

Prompt for Devraj et al. 2022 study

In this experiment, you will be shown examples of nonsense word stimuli. Look carefully at each word and decide if it belongs to group {U} or group {M}. Respond with {U} if you think it is a group {U} word and {M} if you think it is a group {M} word. You will recieve feedback about the correct group after each of your response. At first, the task will seem quite difficult, but with time and practice, you should be able to answer correctly.

- In trial 1, you picked group {M} for wafuzi and group {U} was correct.
- In trial 2, you picked group {M} for gyfuzi and group {U} was correct.

Human: What group would the word gyfuzi belong to? (Give the answer in the form "Group ⟨your answer⟩").
Assistant: Group

Data set	Log. Reg.	SVM	XGBoost	TabPFN	ERMI
kr-vs-kp classification	0.8257	0.8514	0.7986	**0.8664**	0.8450
credit-g classification	**0.6421**	0.6357	0.6350	0.6036	0.6150
diabetes classification	0.6771	**0.7079**	0.6786	0.6886	0.6950
spambase classification	0.5407	0.7664	0.7536	**0.7993**	0.7757
tic-tac-toe classification	0.5536	0.5950	**0.6071**	0.5914	**0.6071**
electricity classification	0.5543	0.6007	**0.7036**	0.6871	0.6436
pc4 Software defect prediction	0.7136	0.7521	**0.7886**	0.7707	0.7714
pc3 Software defect prediction	0.6514	0.7264	**0.7357**	0.7279	0.7107
kc2 Software defect prediction	0.5893	0.7314	**0.7257**	**0.7257**	**0.7257**
kc1 Software defect prediction	0.6271	0.6707	**0.6743**	0.6679	0.6521
pc1 Software defect prediction	0.5336	0.5964	0.6514	0.6064	0.6493
wdbc classification	0.9121	0.9207	0.9014	**0.9221**	0.9093
phoneme classification	0.5793	**0.7314**	0.6979	0.6921	0.7200
qsar-biodeg classification	0.5779	0.7014	0.6850	0.6921	**0.7064**
ilpd classification	0.5493	**0.6386**	0.6229	0.6121	0.6286
ozone-level-8hr classification	0.6614	0.6907	0.6707	0.6471	**0.6950**
banknote-authentication classification	0.7721	0.9229	0.8457	**0.9657**	0.9379
blood-transfution-service-center	0.4714	0.5493	0.5879	0.5671	**0.6186**
phishing websites classification	0.7929	0.8071	**0.8157**	**0.8157**	0.8129
bank-marketing classification	0.5829	0.5614	**0.7386**	0.7350	0.7171
wilt classification	0.5171	0.5736	0.6393	0.6371	**0.6507**
numerai28.6 classification	0.4857	0.4779	**0.5029**	0.4779	0.4986
churn classification	0.6321	0.7271	0.6800	0.7186	**0.7329**
Mean Acc.	62.80 ± 0.66	69.29 ± 0.62	$70.17\% \pm 0.52$	$70.51\% \pm 0.63$	**70.95%± 0.54**
Mean rank	4.52 ± 0.21	2.76 ± 0.26	2.61 ± 0.30	2.85 ± 0.27	**2.26 ± 0.22**