Evolution and compression in LLMs:

on the emergence of human-aligned categorization

Nathaniel Imel and Noga Zaslavsky

CogInterp @ NeurIPS

7 Dec 2025

infoCogLab

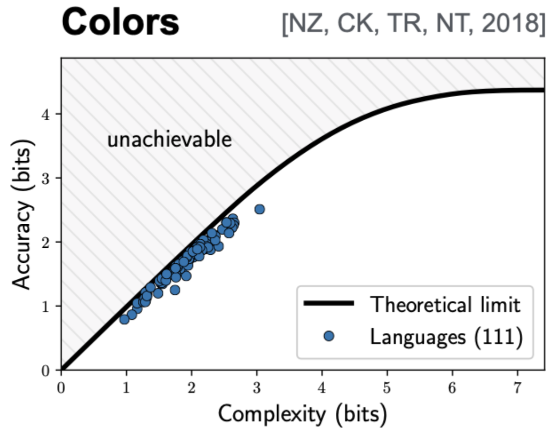
information cognition language

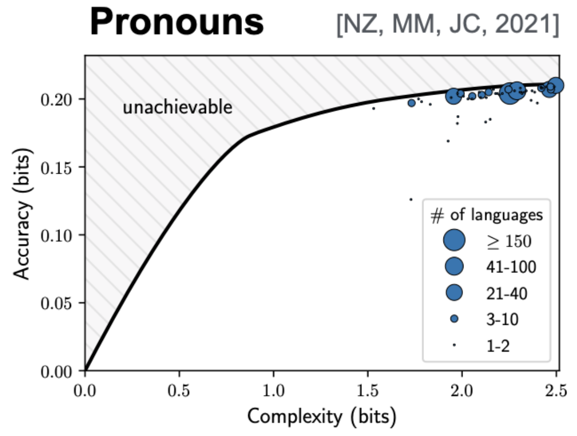
preprint!

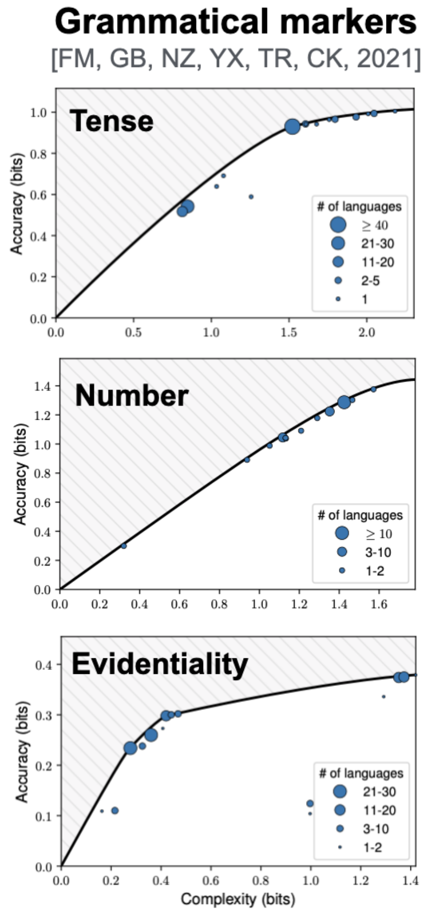
On human-aligned semantic categorization in LLMs

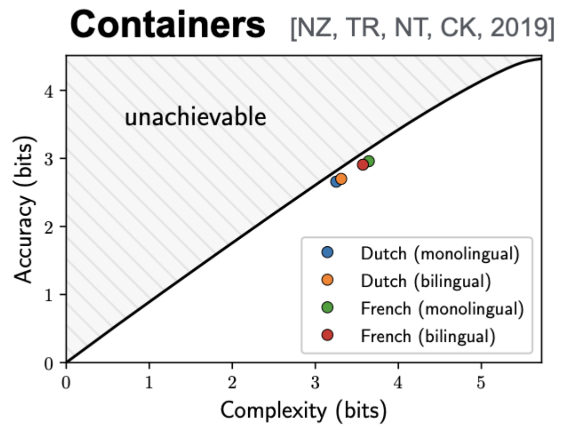
On human-aligned semantic categorization in LLMs

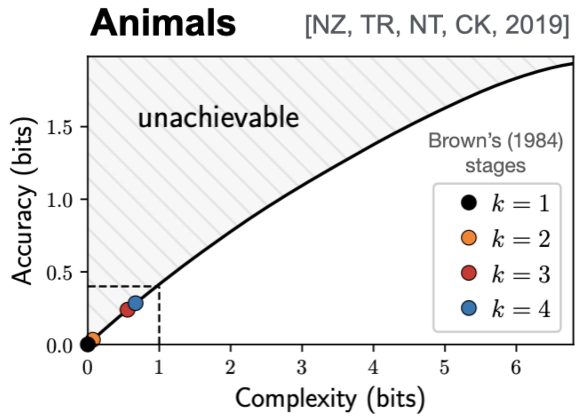
- Systems of semantic categories in human language are optimized for **efficiency** via the **Information Bottleneck** (IB) complexity-accuracy trade-off [1-5]







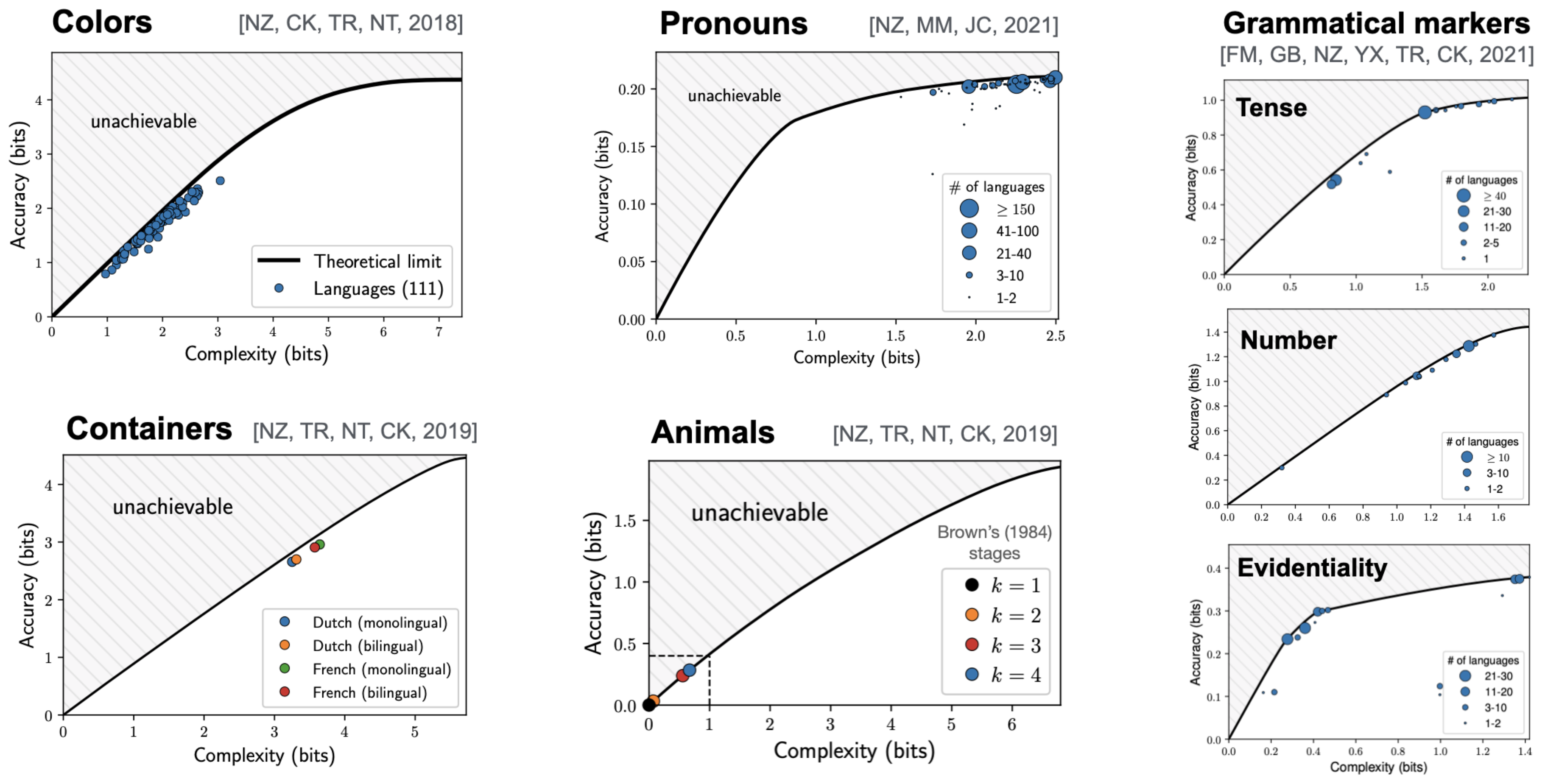




[1] Tishby et al. (1999) [2] Zaslavsky et al. (2018); [3] Zaslavsky et al. (2021); [4] Zaslavsky et al. (2019); [5] Mollica et al. (2021)

On human-aligned semantic categorization in LLMs

- Systems of semantic categories in human language are optimized for **efficiency** via the **Information Bottleneck** (IB) complexity-accuracy trade-off [1-5]

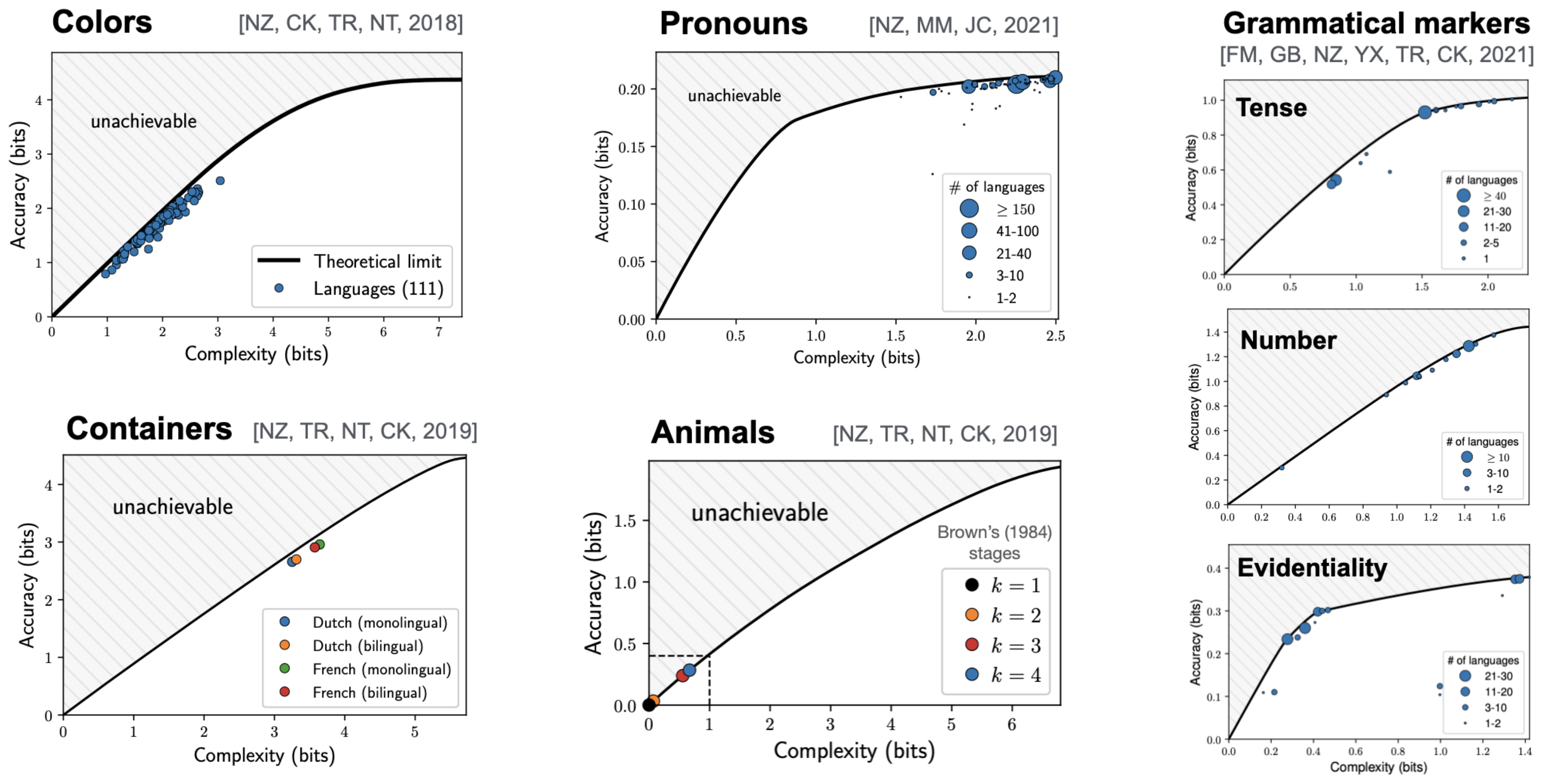


- Large language models (LLMs) are not trained for the IB objective, which raises the question:

[1] Tishby et al. (1999) [2] Zaslavsky et al. (2018); [3] Zaslavsky et al. (2021); [4] Zaslavsky et al. (2019); [5] Mollica et al. (2021)

On human-aligned semantic categorization in LLMs

- Systems of semantic categories in human language are optimized for **efficiency** via the **Information Bottleneck** (IB) complexity-accuracy trade-off [1-5]



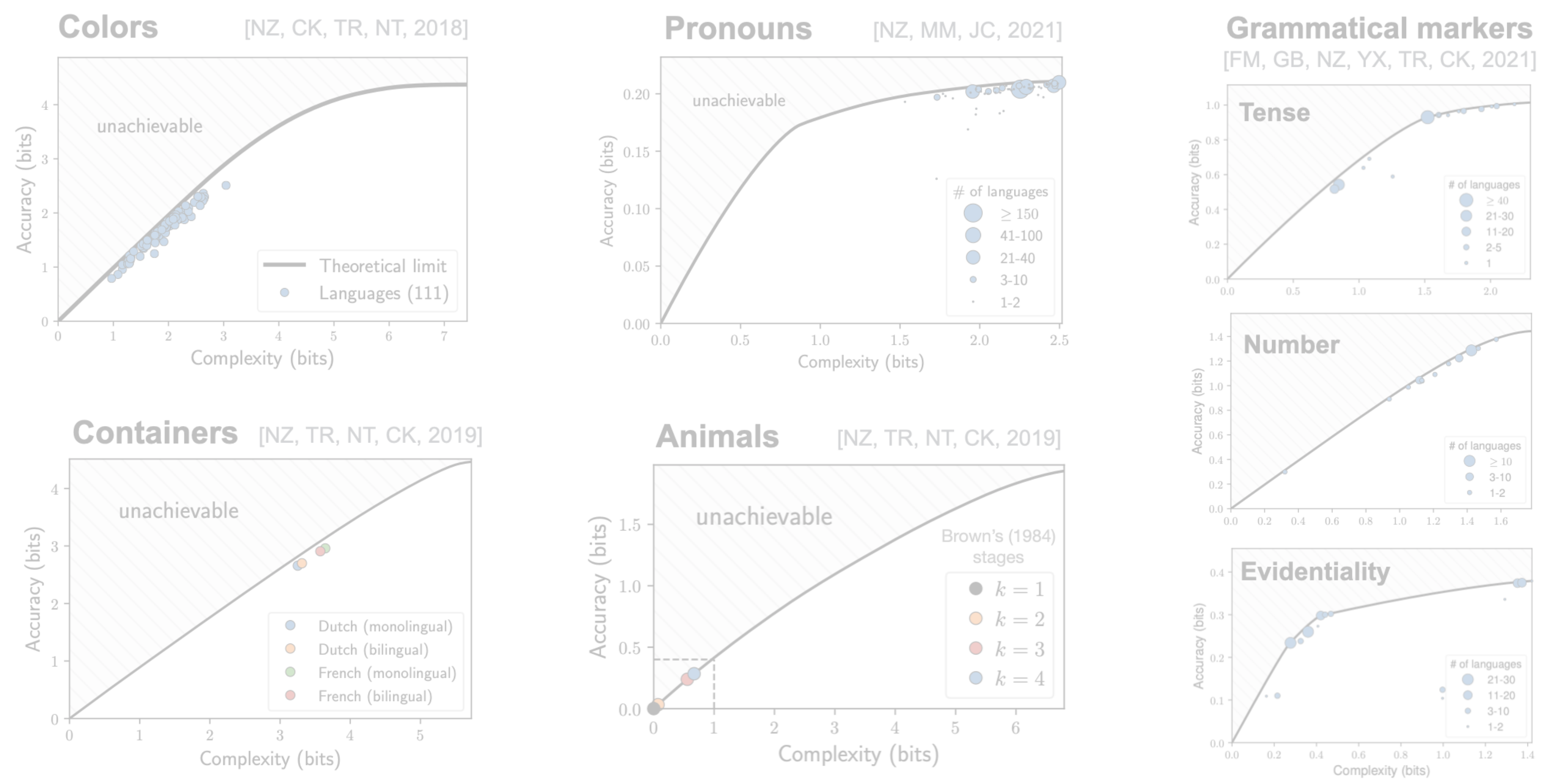
- Large language models (LLMs) are not trained for the IB objective, which raises the question:

Do LLMs share with humans a bias

to maintain semantic efficiency?

[1] Tishby et al. (1999) [2] Zaslavsky et al. (2018); [3] Zaslavsky et al. (2021); [4] Zaslavsky et al. (2019); [5] Mollica et al. (2021)

On human-aligned semantic categorization in LLMs

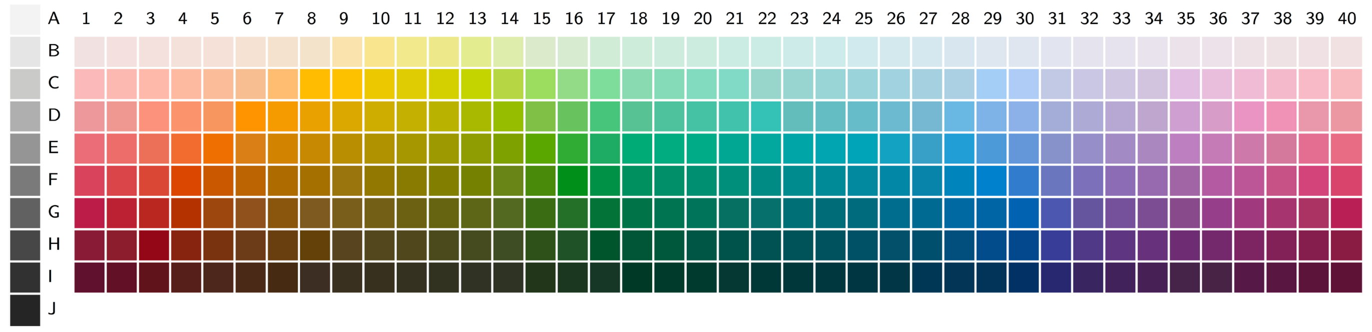


Do LLMs share with humans a bias

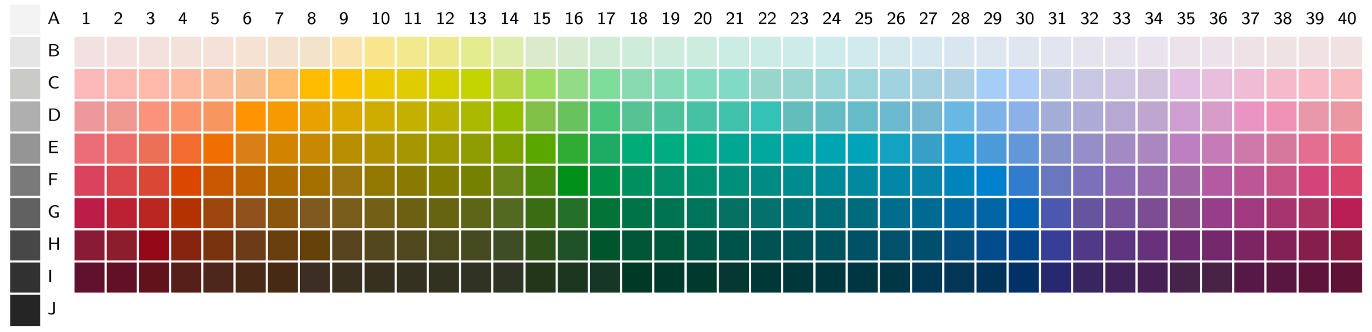
to maintain semantic efficiency?

- To address this question, we perform an in-depth analysis of LLM **color naming**

Why talk about color?

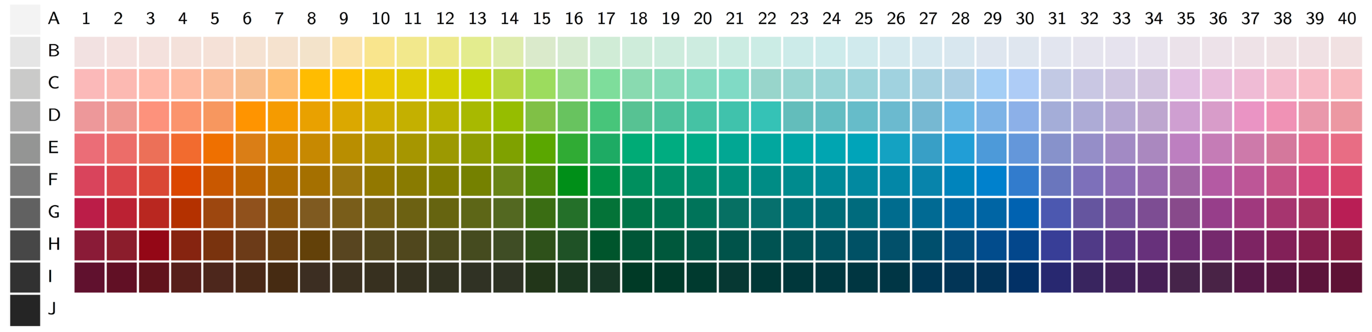


Why talk about color?



- Practical implications for AI

Why talk about color?

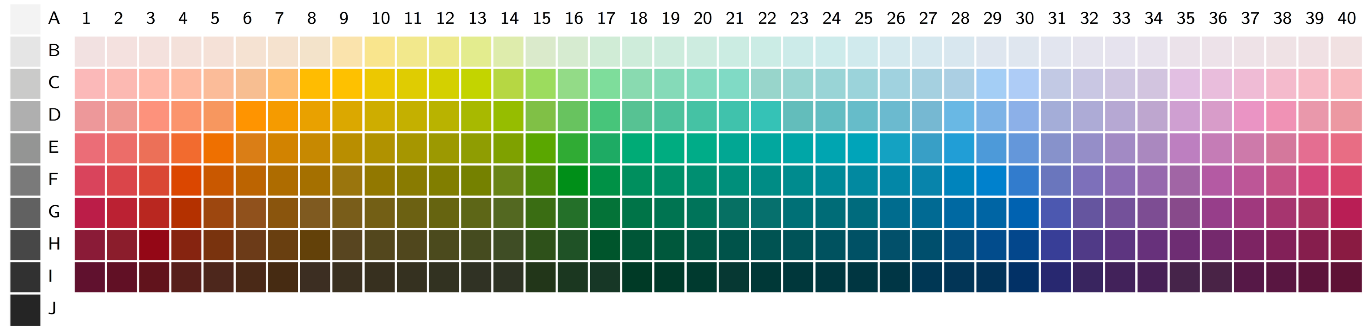


- Practical implications for AI

- rarely available human behavioral data [1-5]

[1] Berlin & Kay (1969) [2] Cook et al. (2005); [3] Lindsey & Brown (2014); [4] Xu et al. (2013); [5] Imel et al. (2025)

Why talk about color?



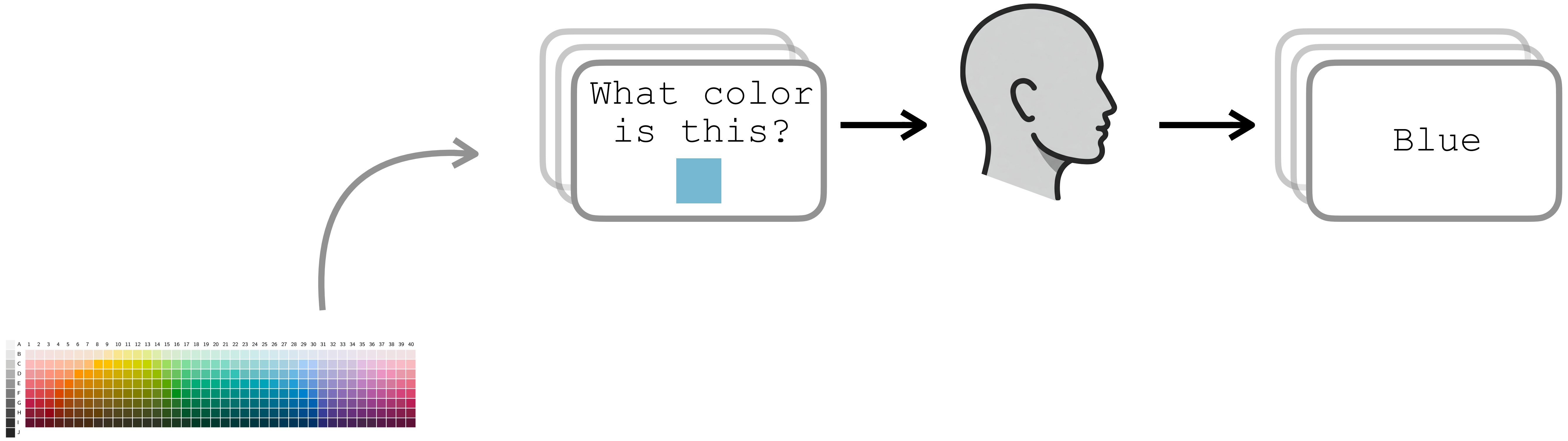
- Practical implications for AI
- rarely available human behavioral data [1-5]
- well established IB naming model [6,7]

[1] Berlin & Kay (1969) [2] Cook et al. (2005); [3] Lindsey & Brown (2014); [4] Xu et al. (2013); [5] Imel et al. (2025); [6] Zaslavsky et al. (2018); [7] Zaslavsky et al. (2022)

Study 1: English color naming in LLMs

Study 1: English color naming in LLMs

Empirical comparison: color naming with English speakers [1]



What color is this?

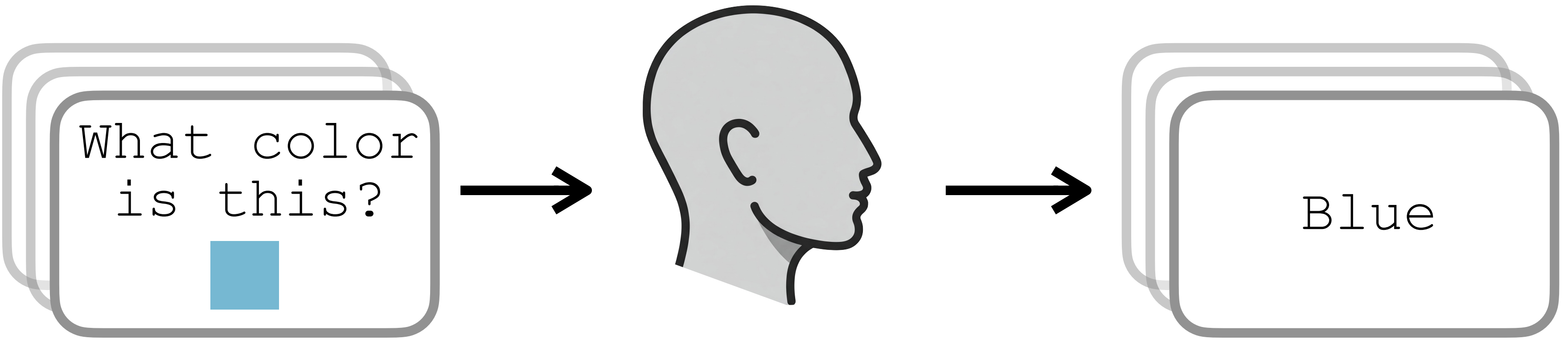
Blue

WCS stimulus grid

[1] Lindsey & Brown (2014)

Study 1: English color naming in LLMs

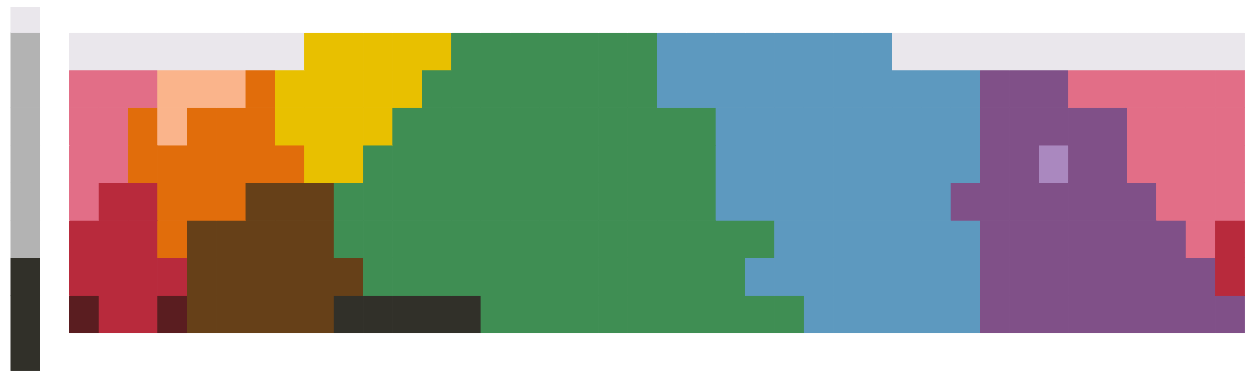
Empirical comparison: color naming with English speakers [1]



What color is this?

Blue

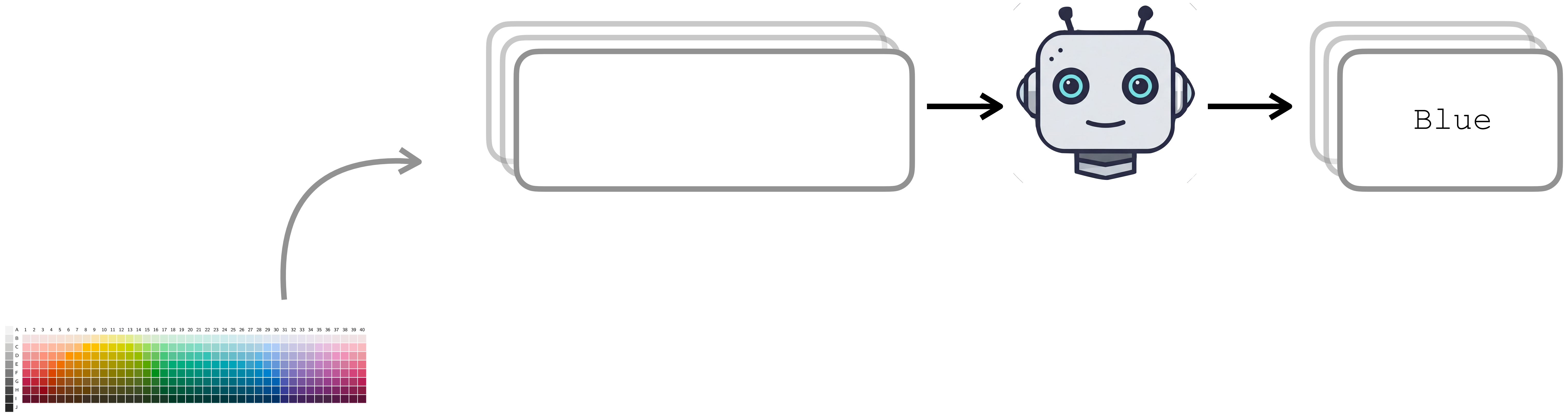
English color naming system



[1] Lindsey & Brown (2014)

Study 1: English color naming in LLMs

English color naming with **LLMs**

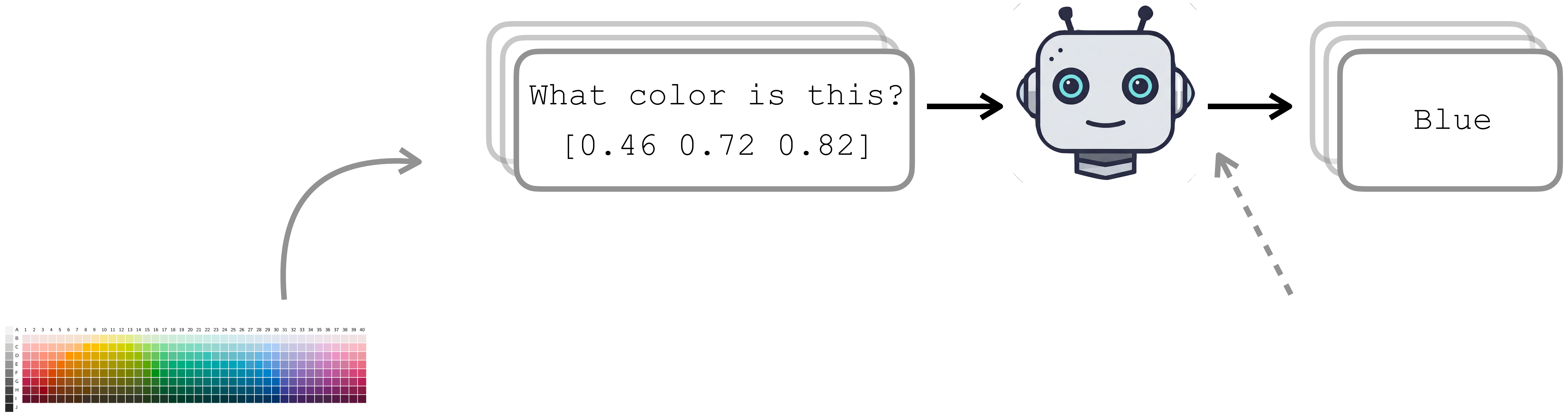


WCS stimulus grid

see also Marjieh et al. (2024)

Study 1: English color naming in LLMs

English color naming with **LLMs**



```
What color is this?
[0.46 0.72 0.82]
```

```
Blue
```

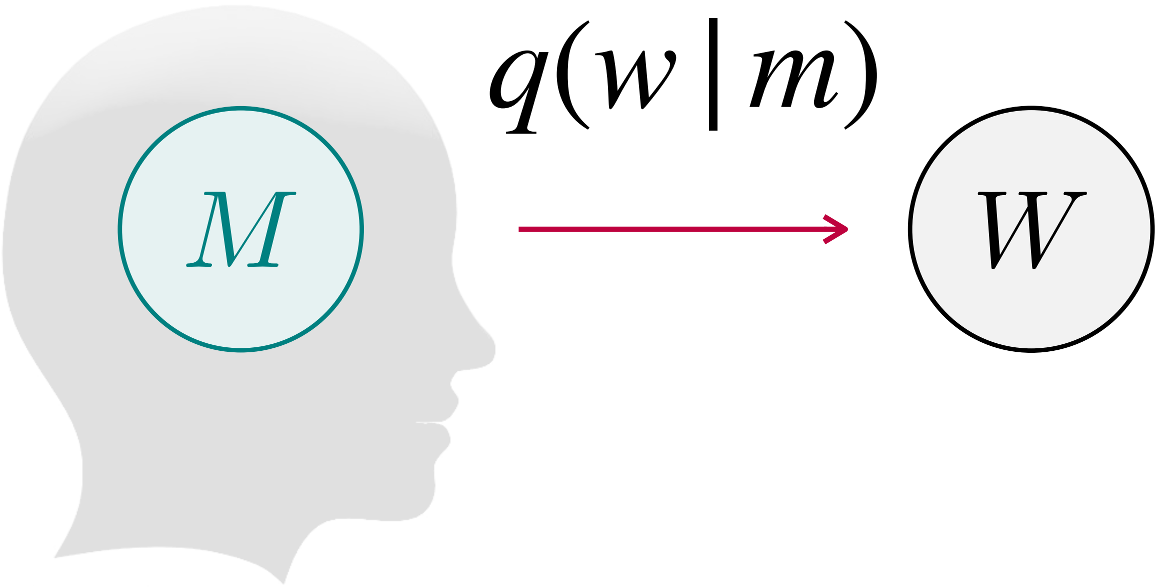
WCS stimulus grid

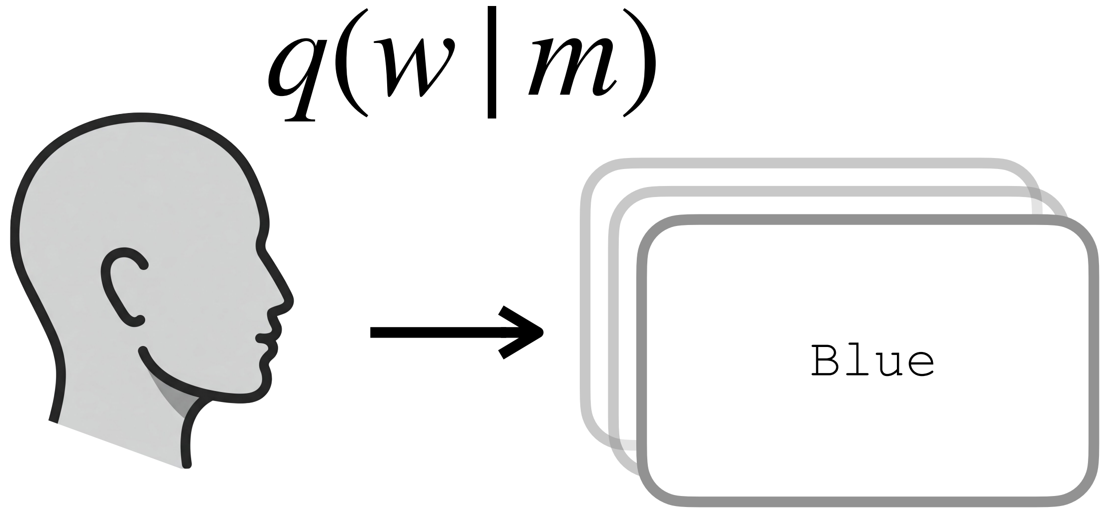
* ask me about multimodal models in Q&A!

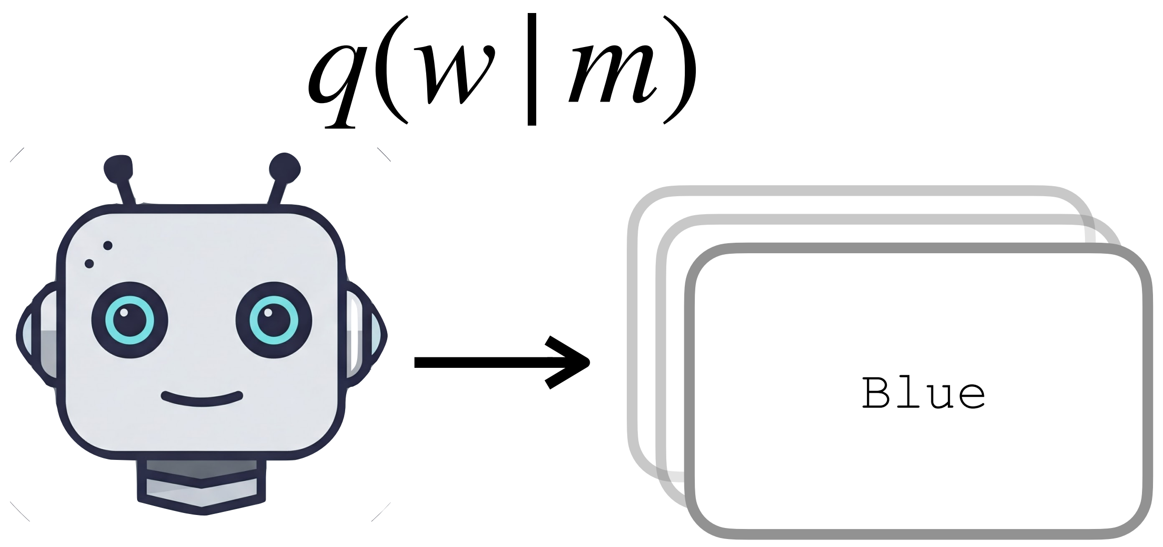
see also Marjieh et al. (2024)

Study 1: English color naming in LLMs — IB model

each color naming system: a stochastic encoder $q(w\,|\,m)$

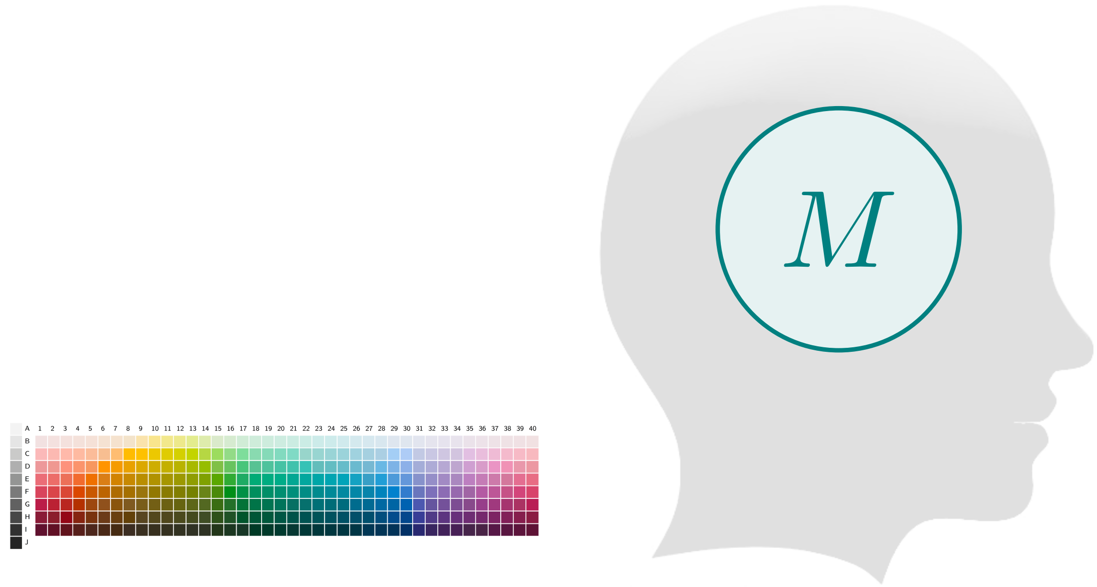






Study 1: English color naming in LLMs — IB model

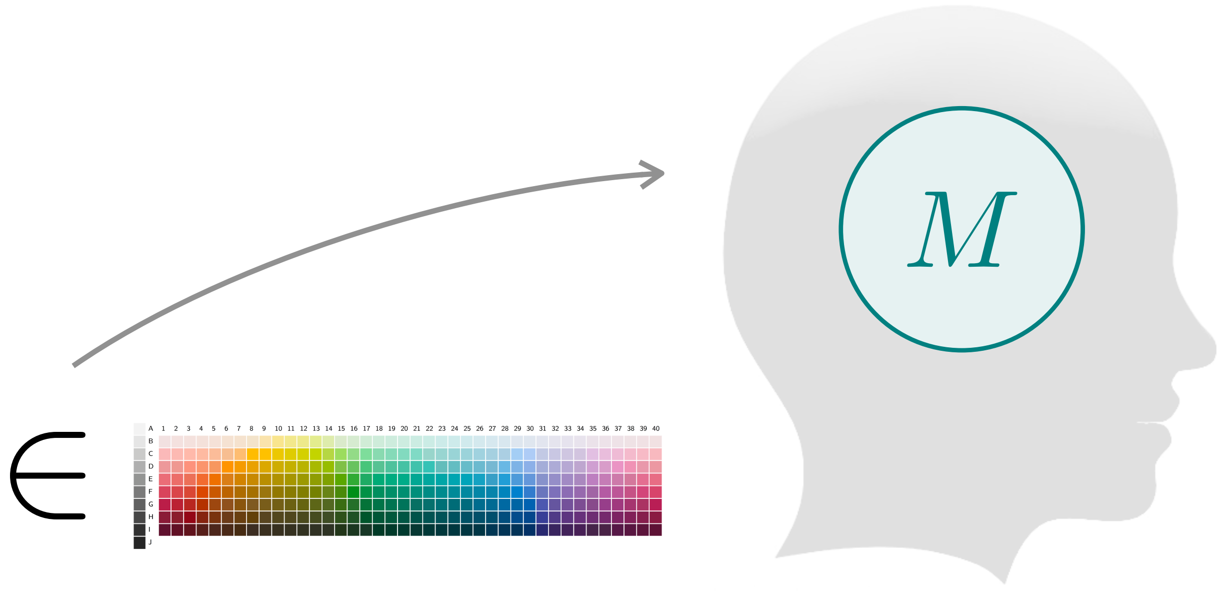
$u_t \in$



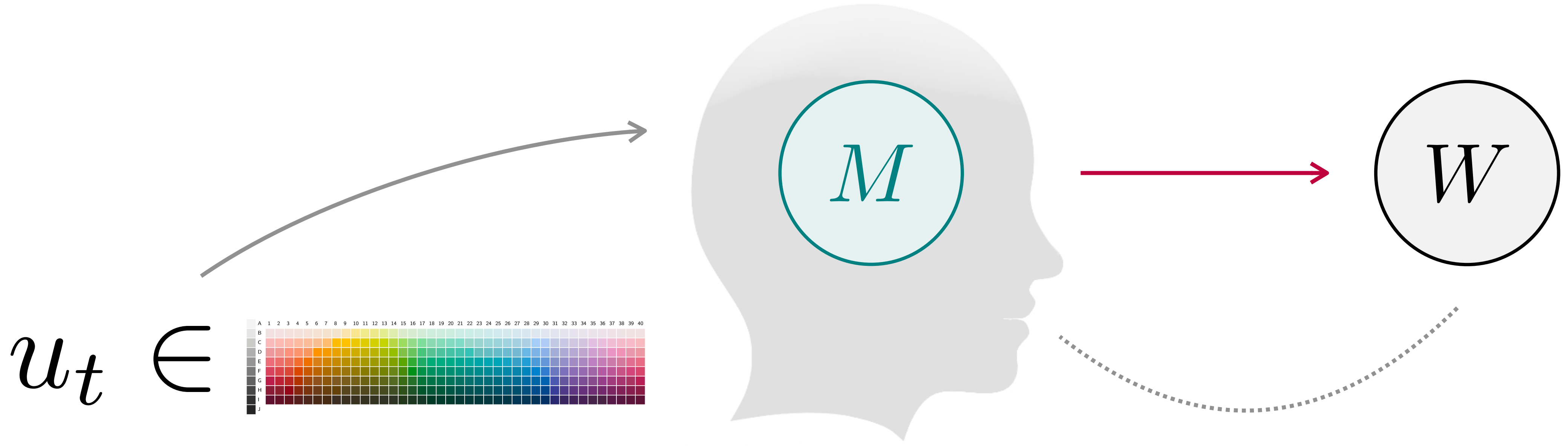
Study 1: English color naming in LLMs — IB model

$$m_t(u) \propto \exp(\gamma\ \mathbf{Sim}(u_t, u))$$

$u_t \in$



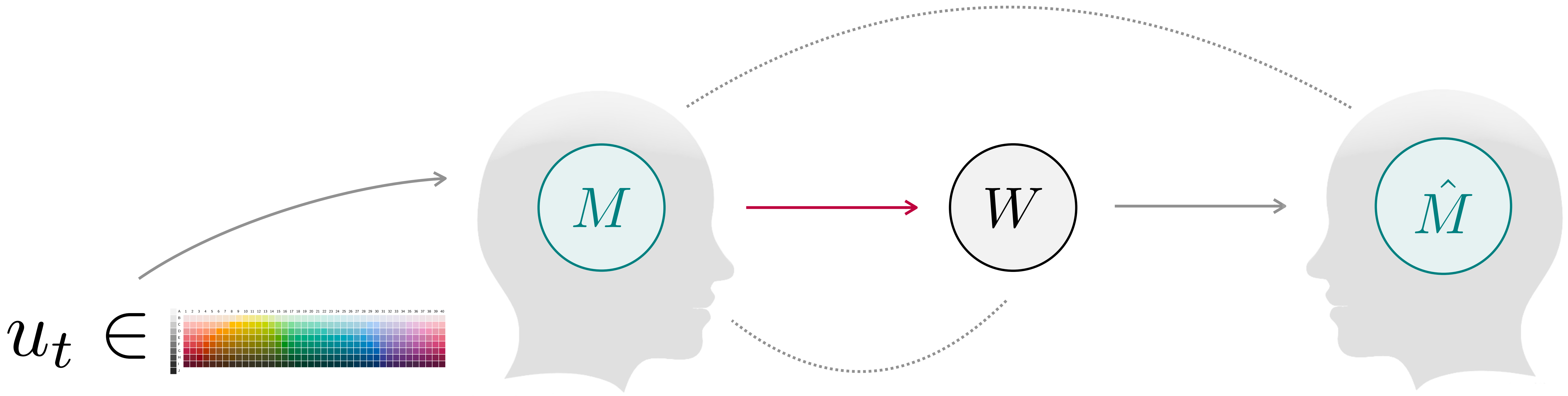
Study 1: English color naming in LLMs — IB model



complexity $= I_q(M; W)$

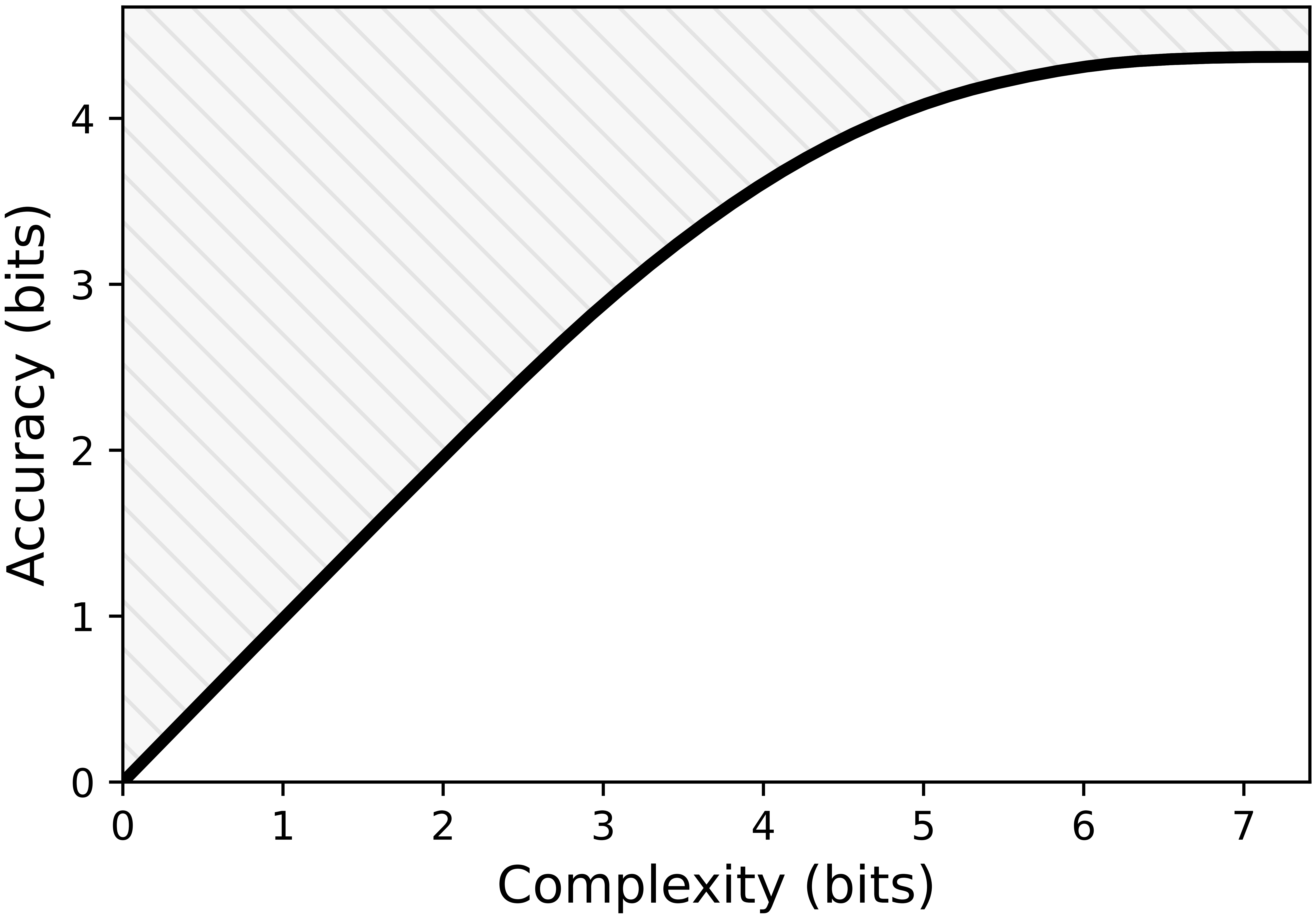
Study 1: English color naming in LLMs — IB model

accuracy $= \text{const} - \mathbb{E}_q[\ D_{KL}[M \parallel \hat{M}]\ ]$

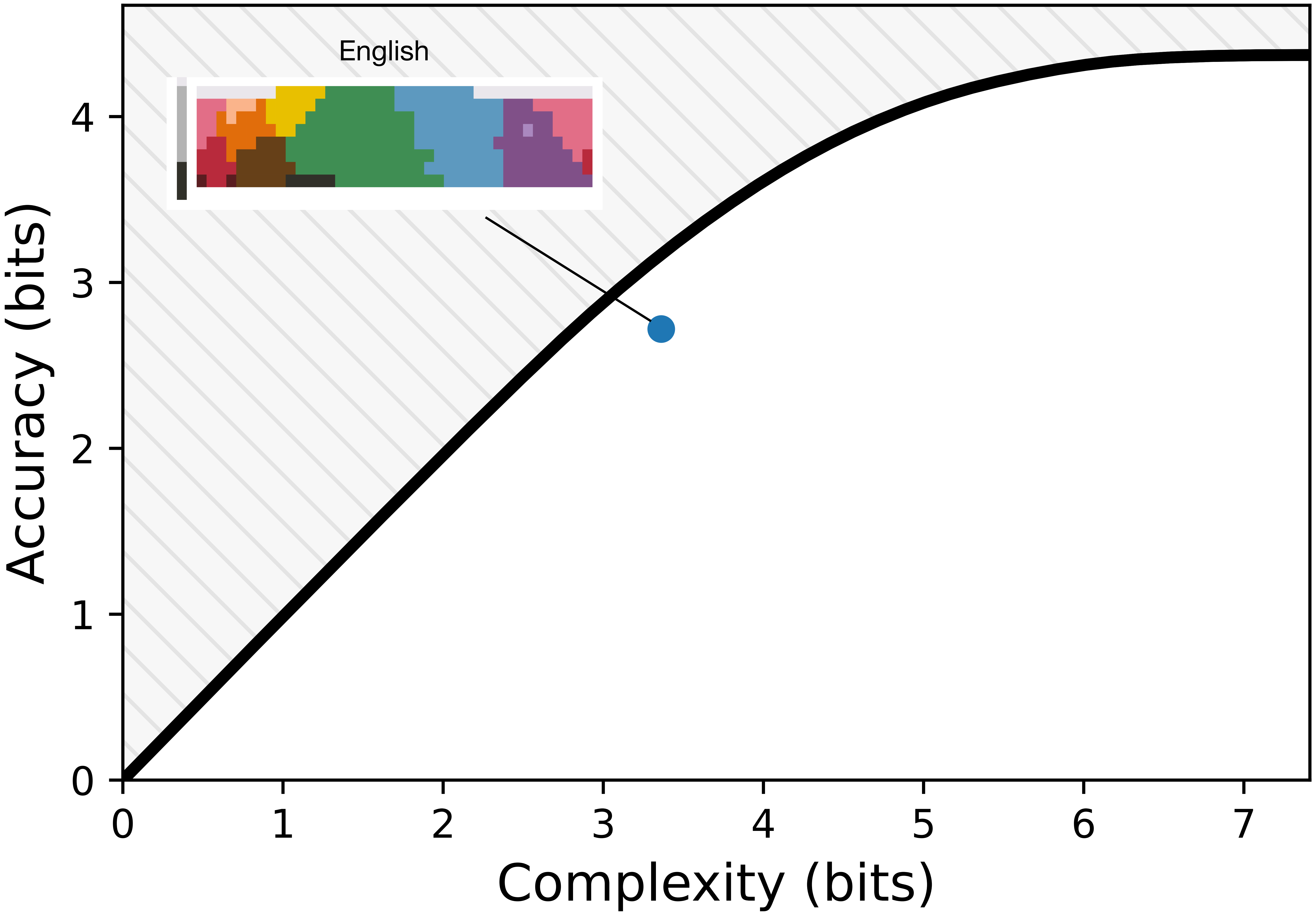


complexity $= I_q(M; W)$

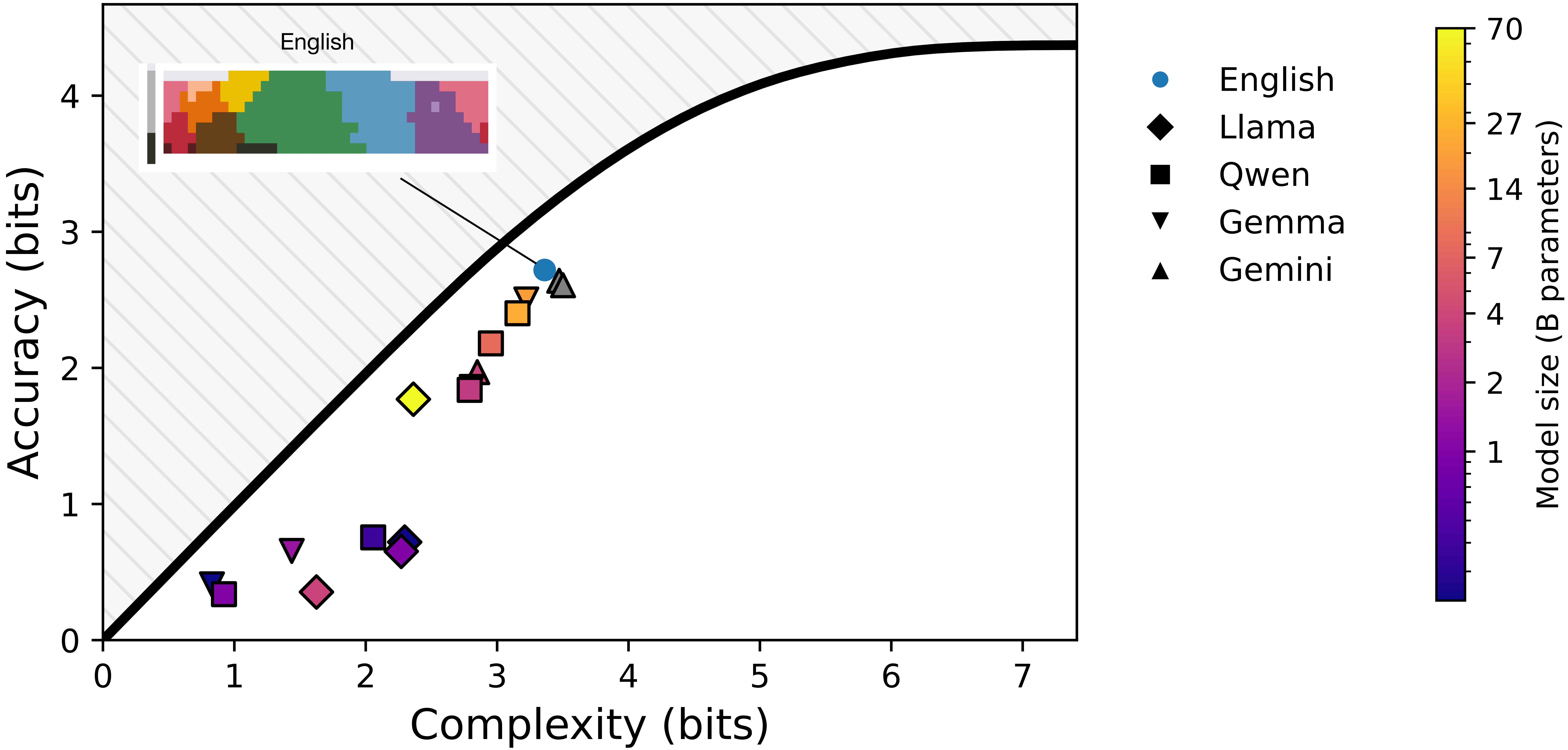
Study 1: English color naming in LLMs — IB model



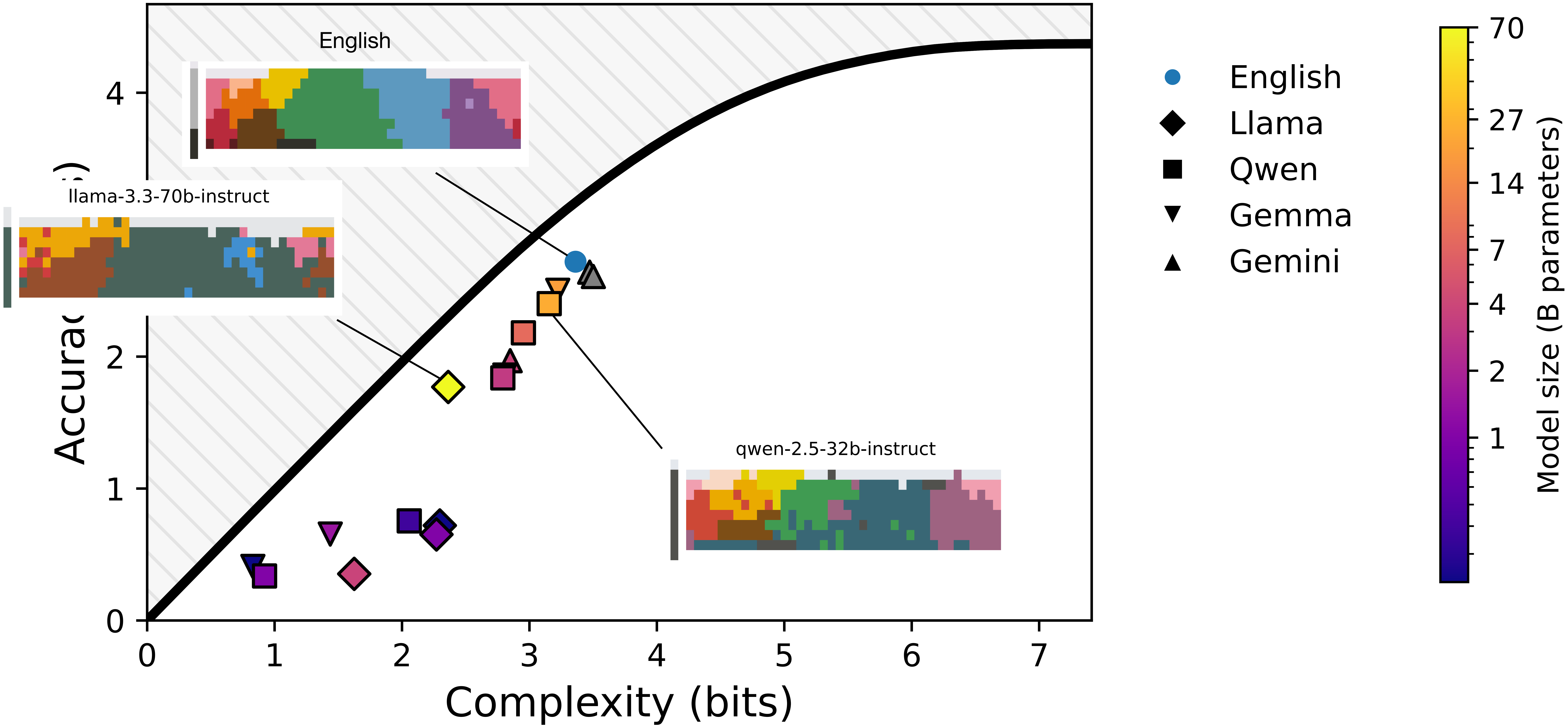
Study 1: English color naming in LLMs — IB model



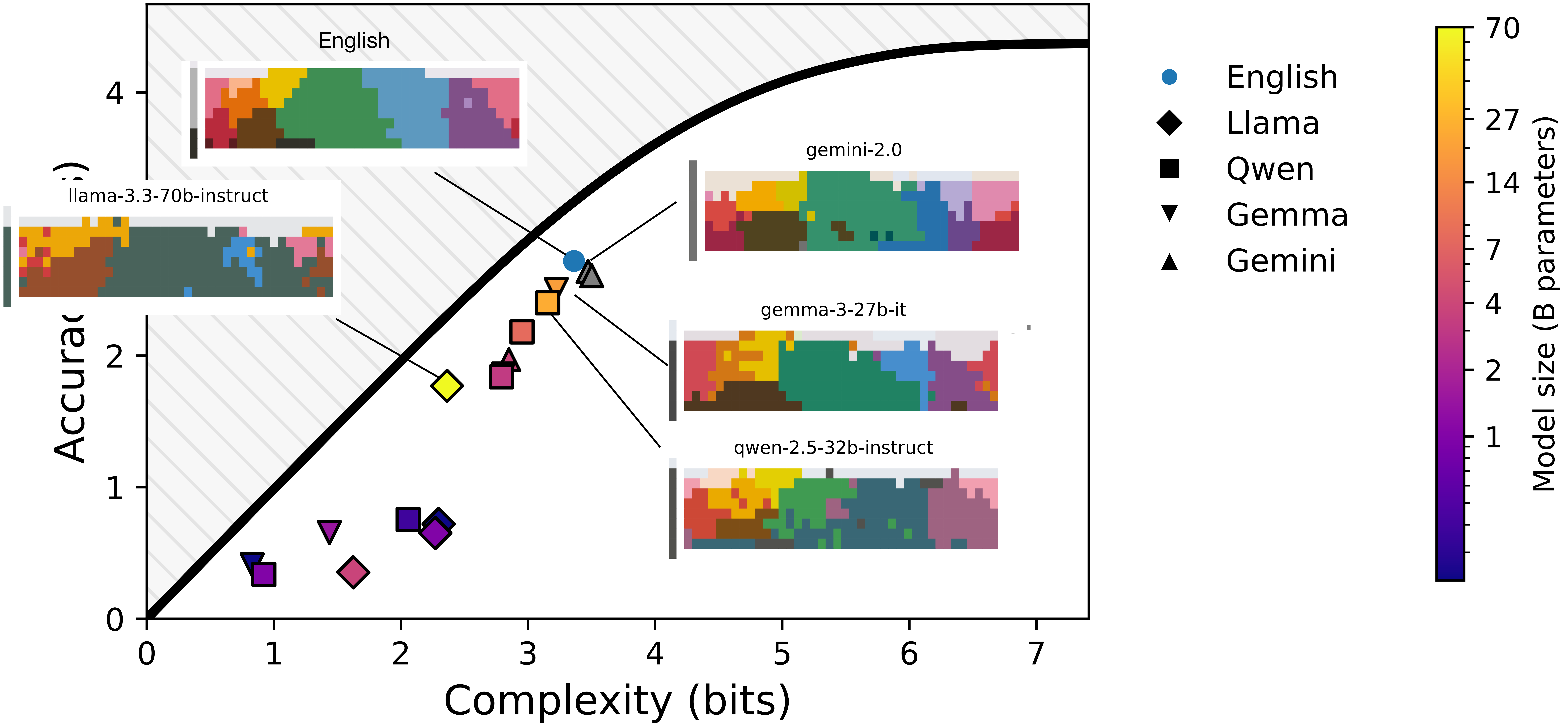
Study 1 Results: LLM systems' efficiency tradeoffs



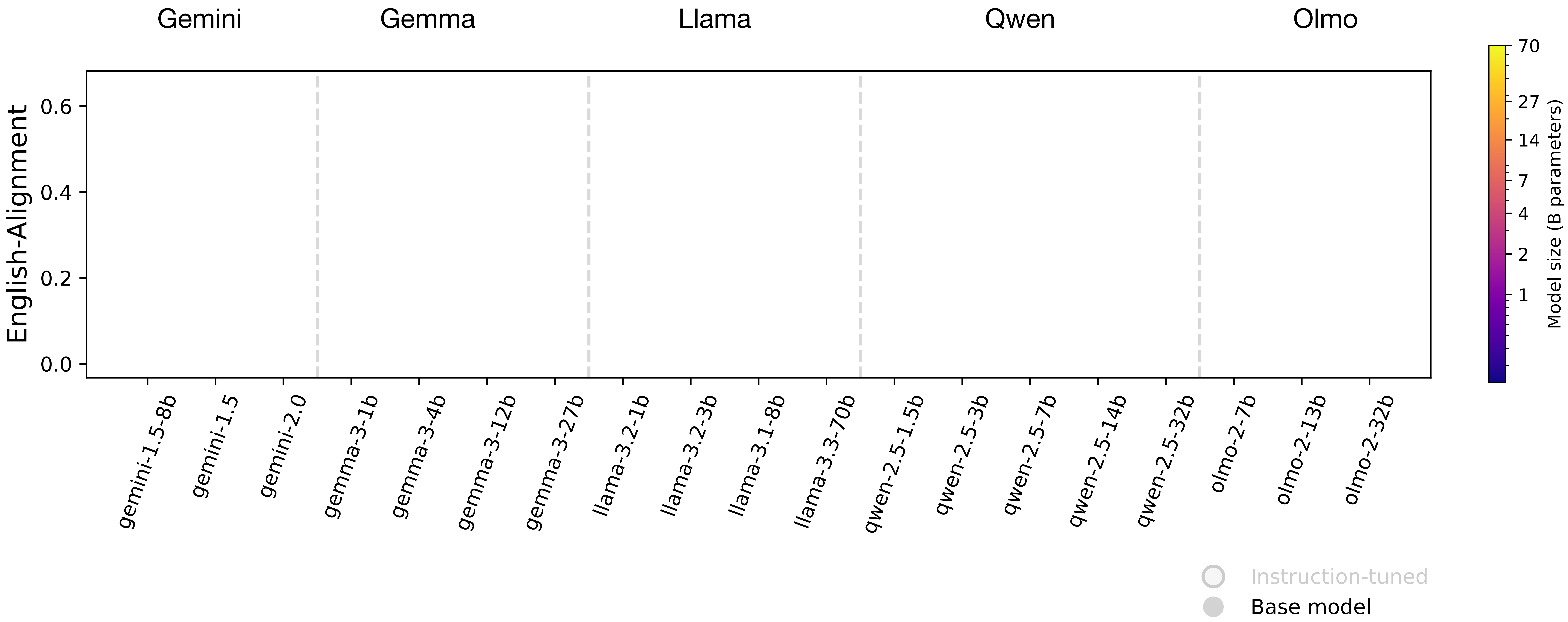
Study 1 Results: LLM systems' efficiency tradeoffs



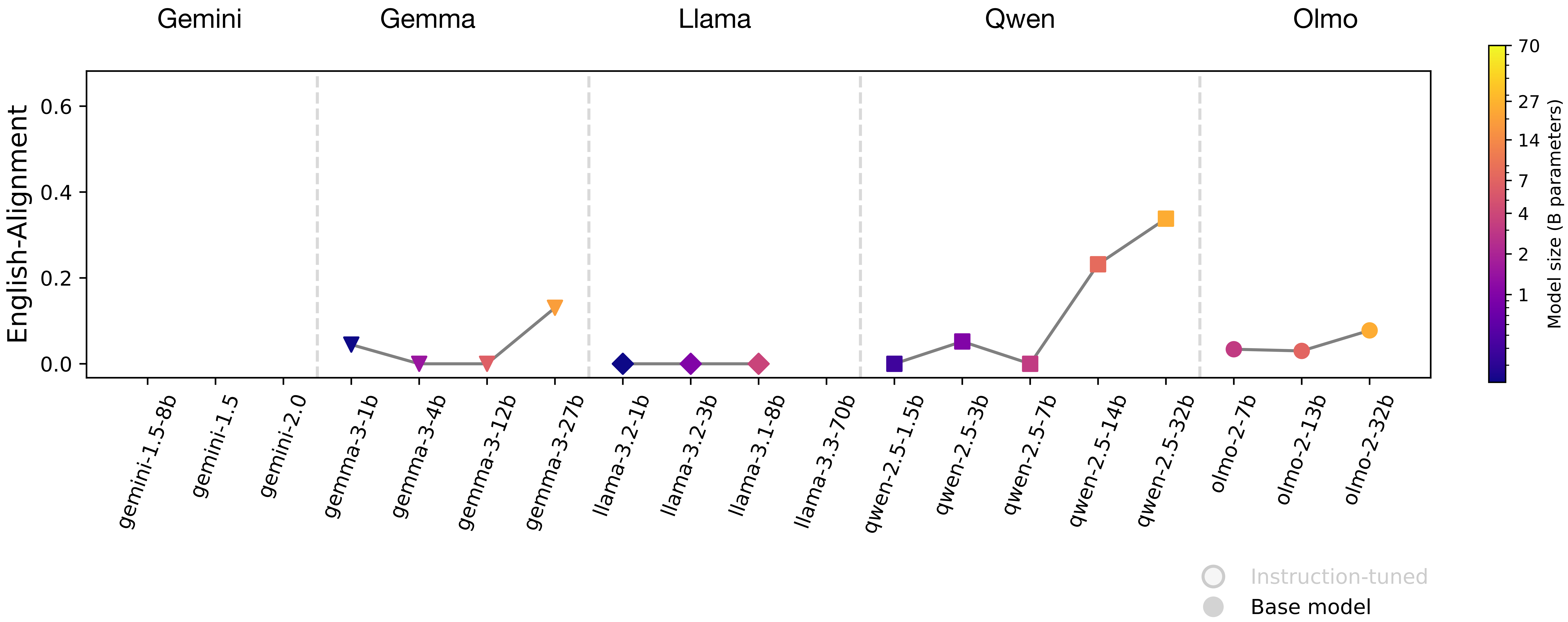
Study 1 Results: LLM systems' efficiency tradeoffs



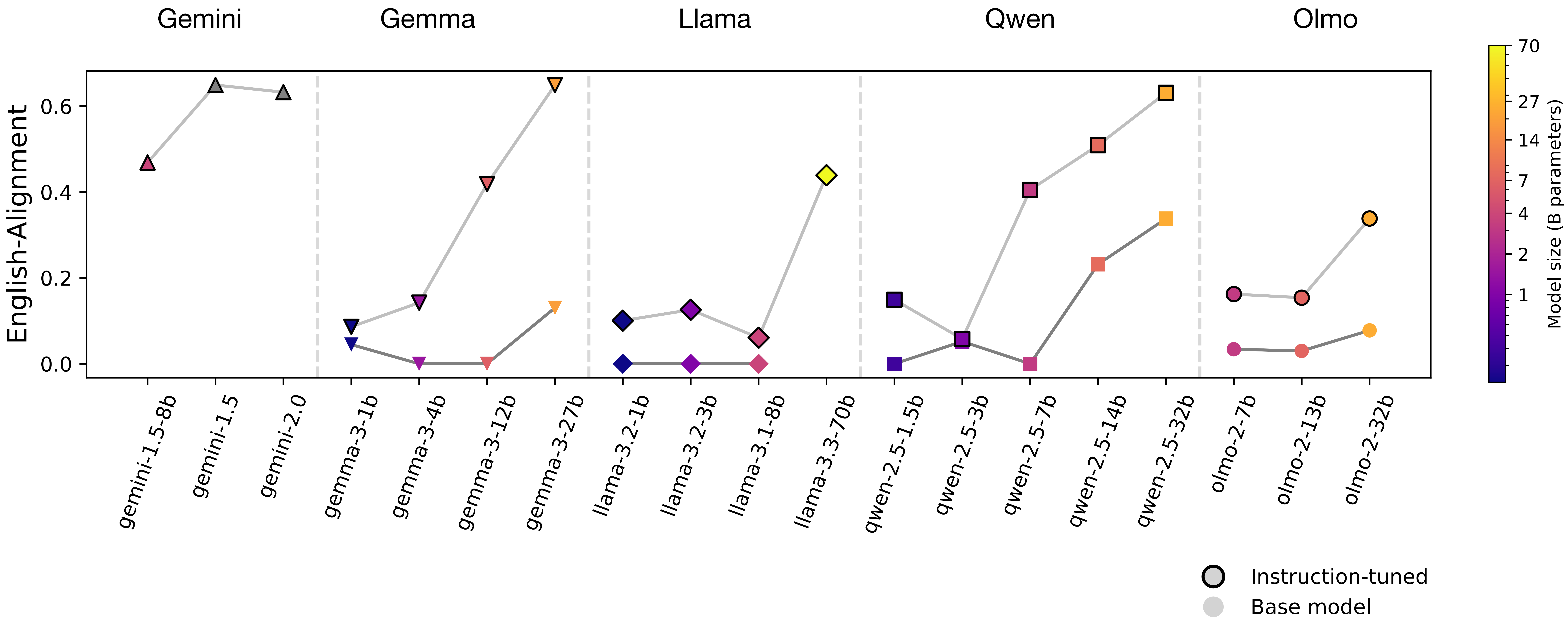
Study 1 Results: LLM alignment to human color naming



Study 1 Results: LLM alignment to human color naming



Study 1 Results: LLM alignment to human color naming



Study 1: Summary

- Capturing English color naming patterns is **non-trivial** for LLMs,

Study 1: Summary

- Capturing English color naming patterns is **non-trivial** for LLMs,
- larger, instruction-tuned models often achieve better **alignment and efficiency.**

Are better models simply mimicking patterns in (English) training data?

Are better models simply mimicking patterns in (English) training data?

Or have LLMs **acquired an inductive bias** towards human-like IB-efficiency?

Are better models simply mimicking patterns in (English) training data?

Or have LLMs **acquired an inductive bias** towards human-like IB-efficiency?

- To address this, we simulate cultural transmission of color naming systems in LLMs

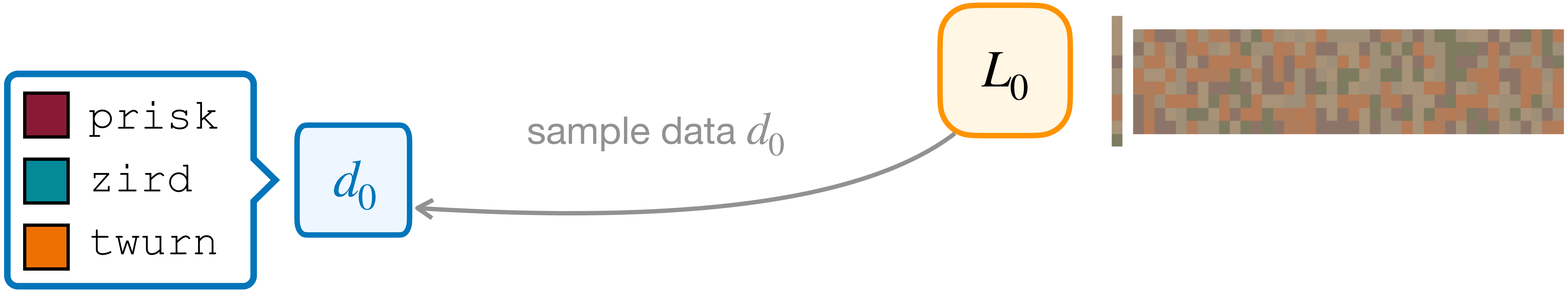
Background: human iterated learning of color systems [1]

[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]

[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]



🟥 prisk

🟦 zird

🟧 twurn

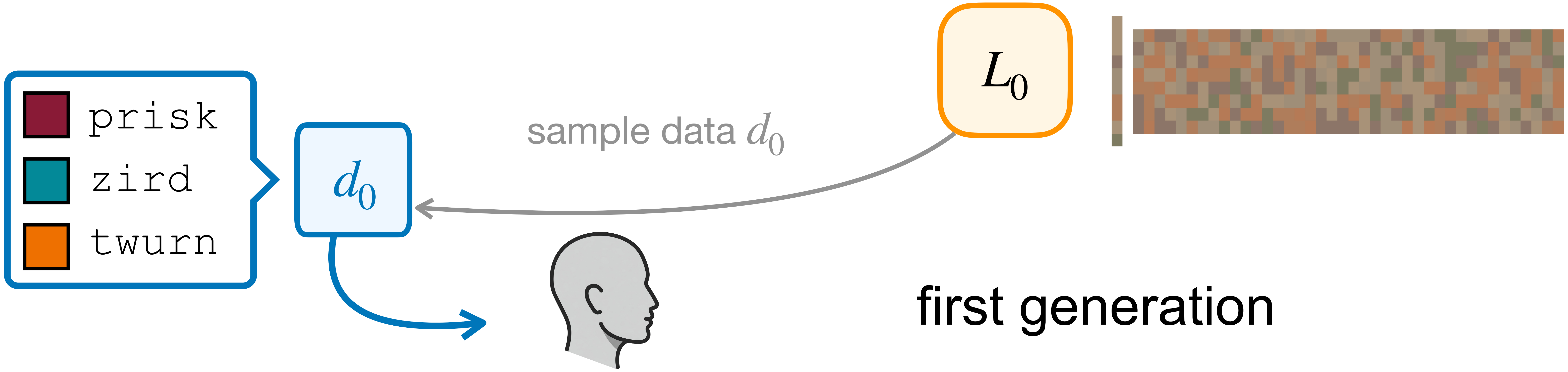
d_0

sample data d_0

L_0

[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]



🟥 prisk

🟦 zird

🟧 twurn

d_0

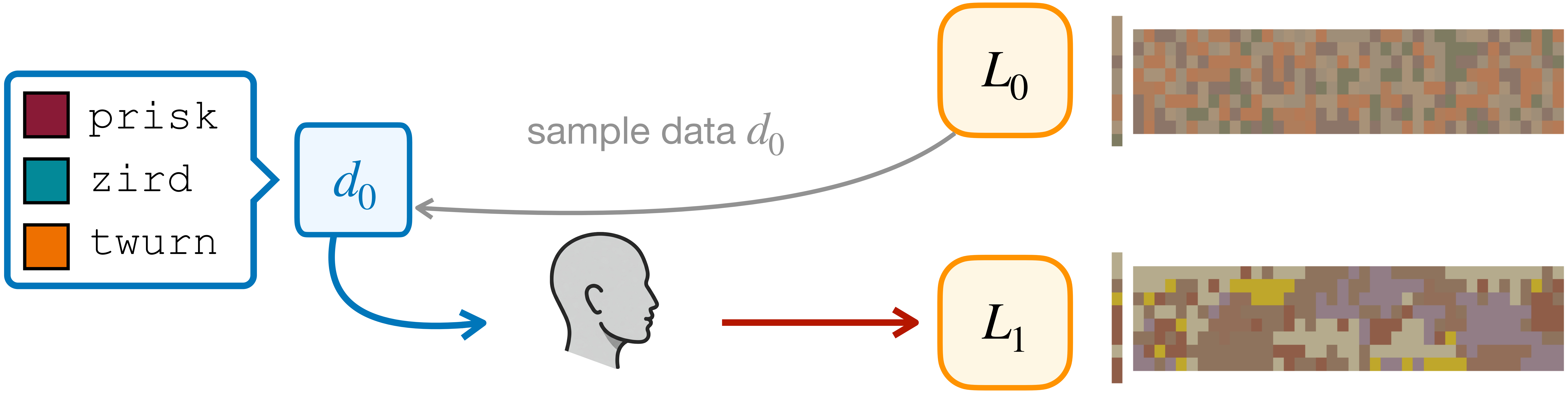
sample data d_0

L_0

first generation

[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]



🟥 prisk

🟦 zird

🟧 twurn

d_0

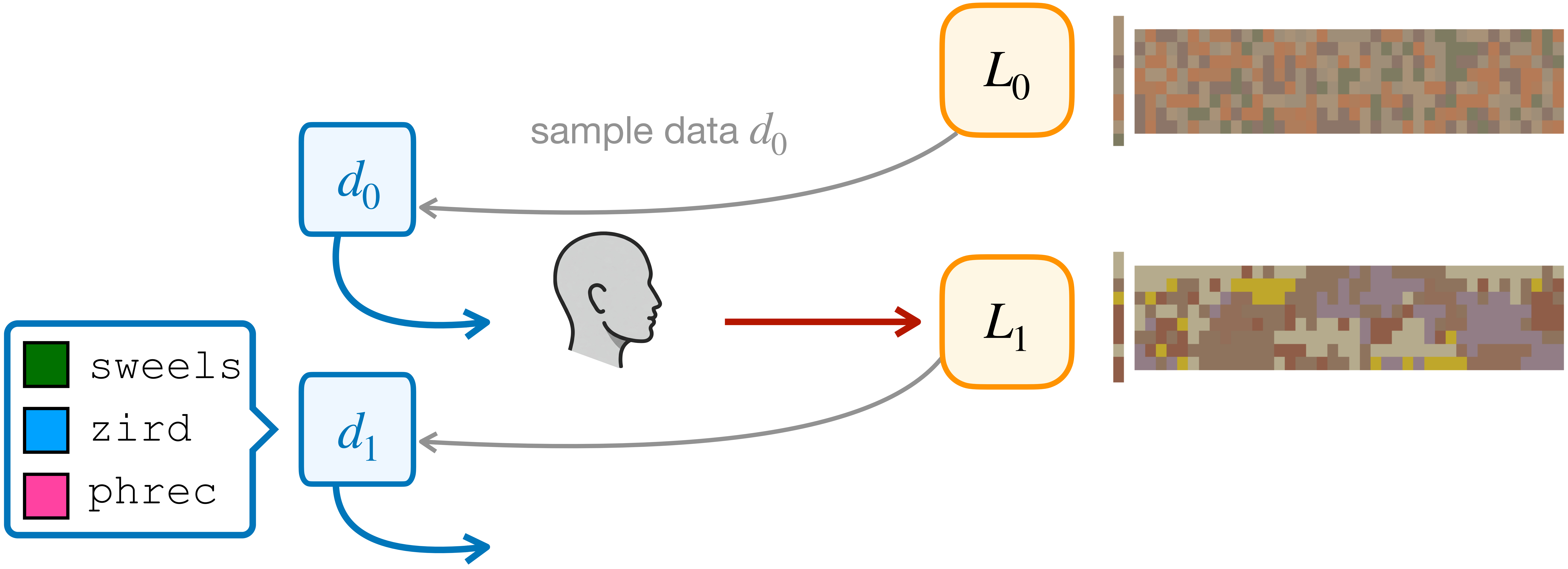
sample data d_0

L_0

L_1

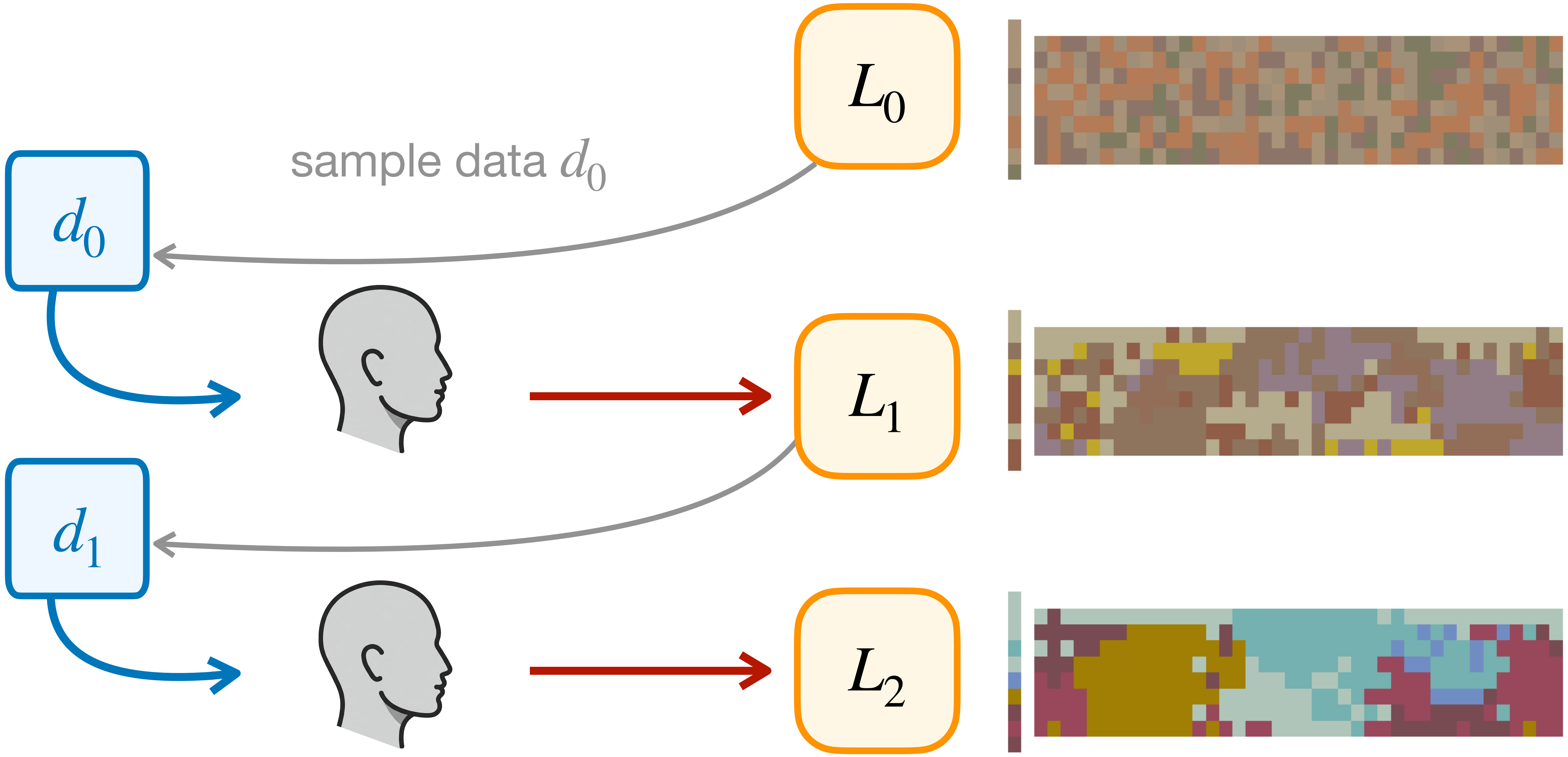
[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]



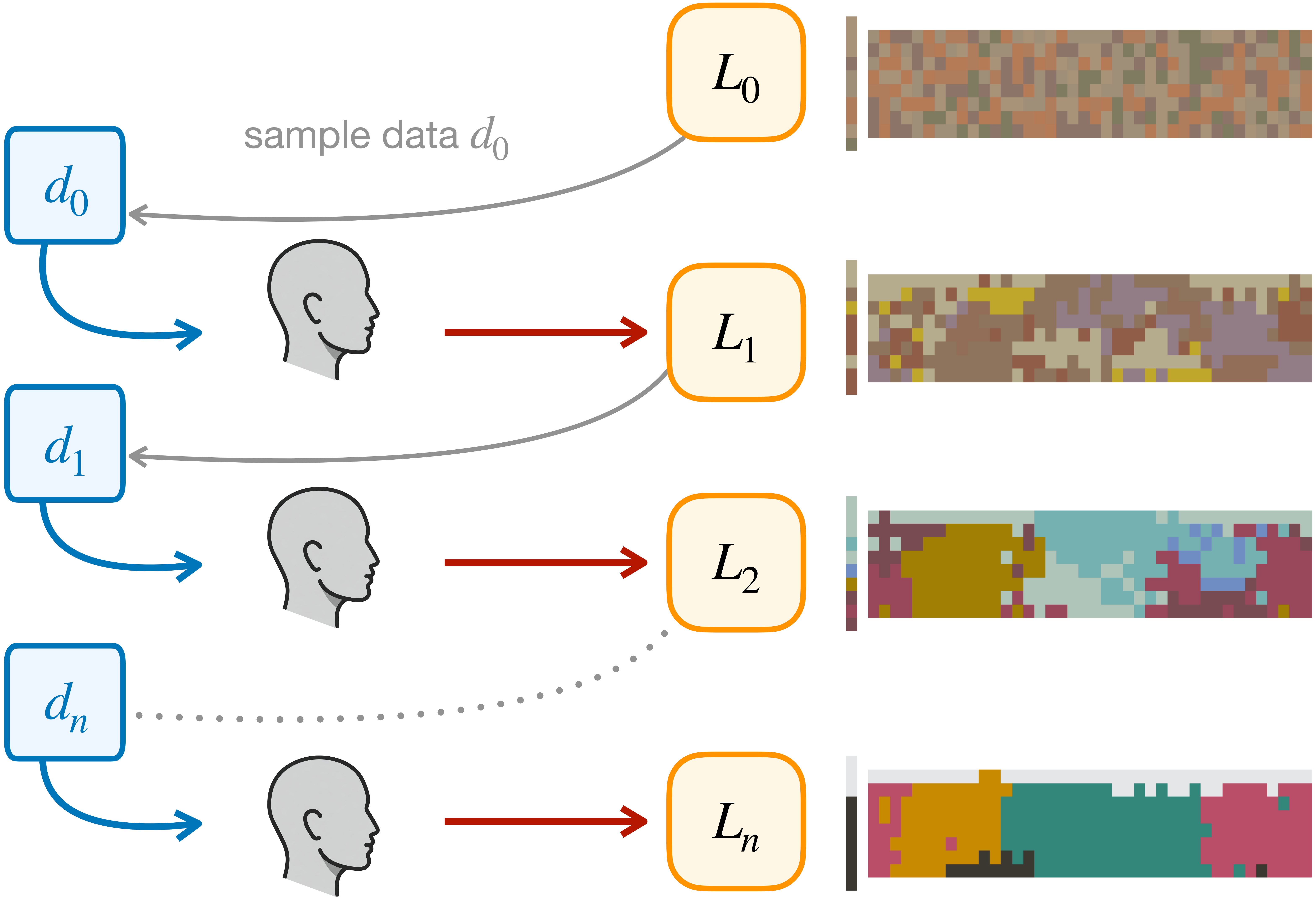
[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]



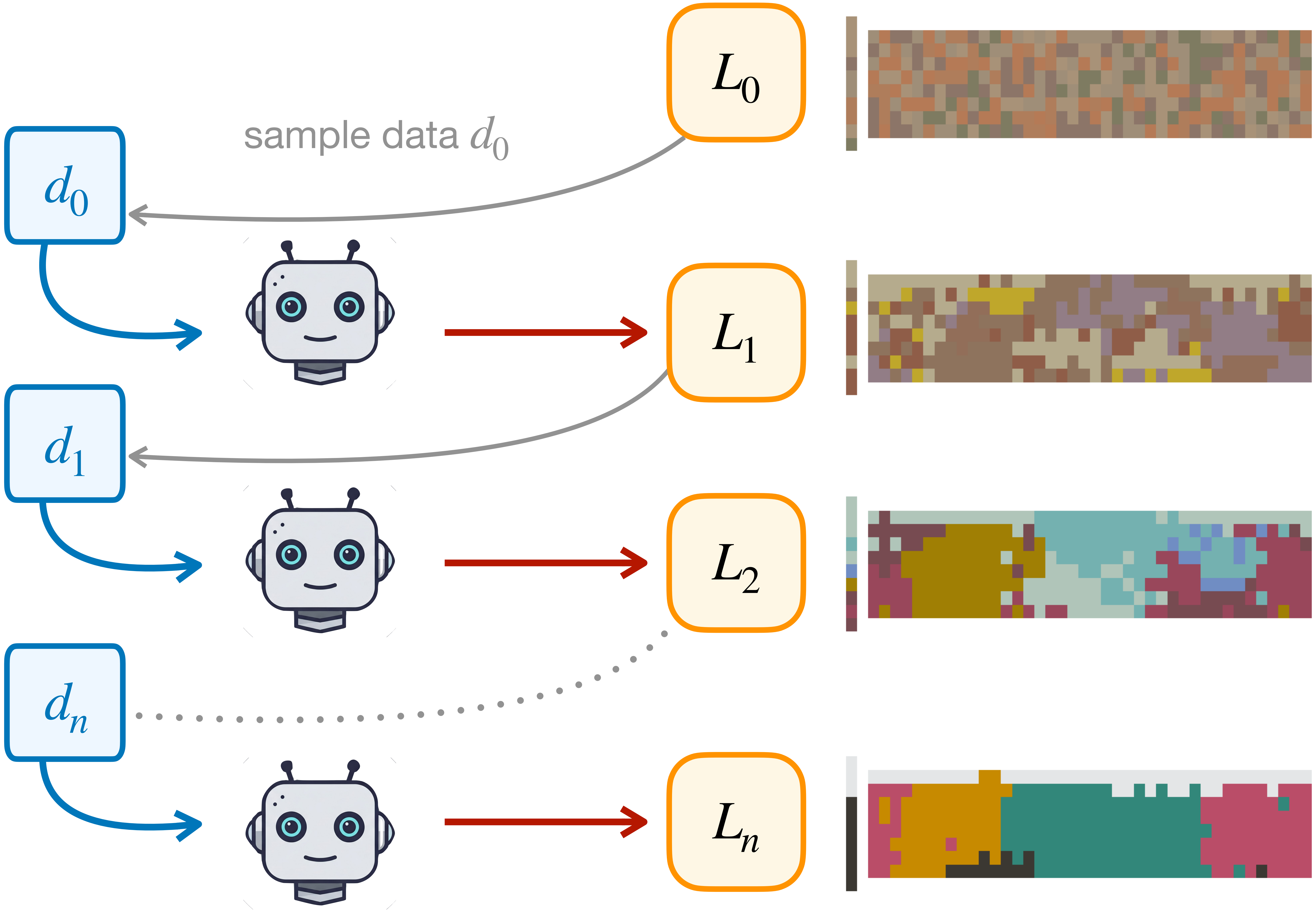
[1] Xu et al. (2013)

Background: human iterated learning of color systems [1]

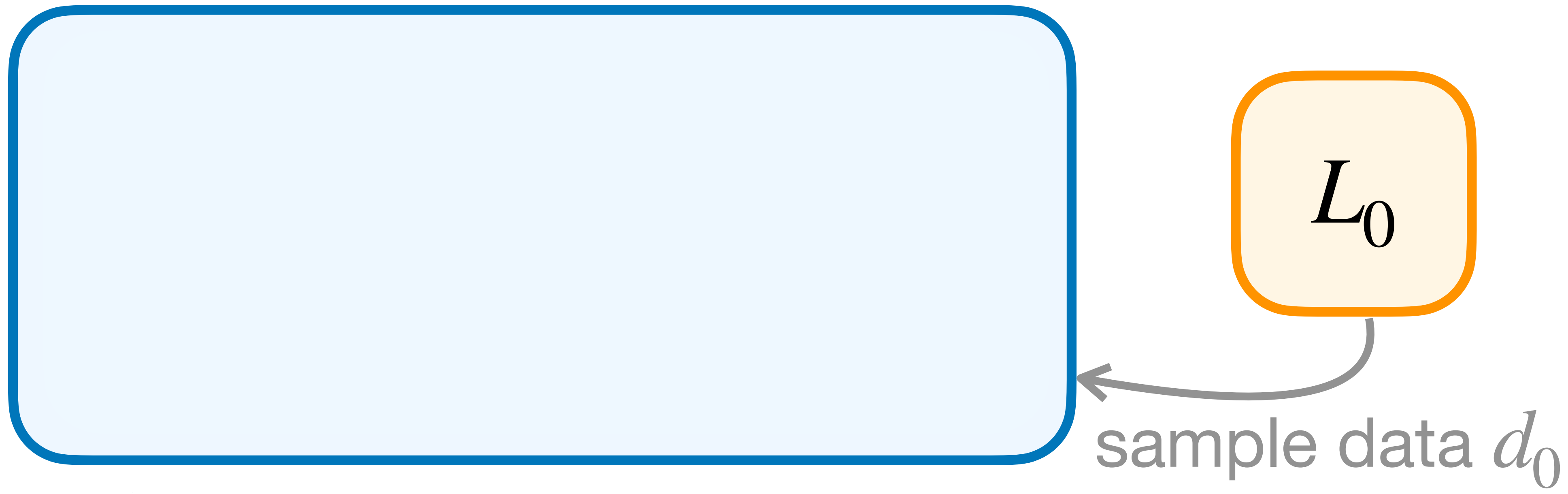


[1] Xu et al. (2013)

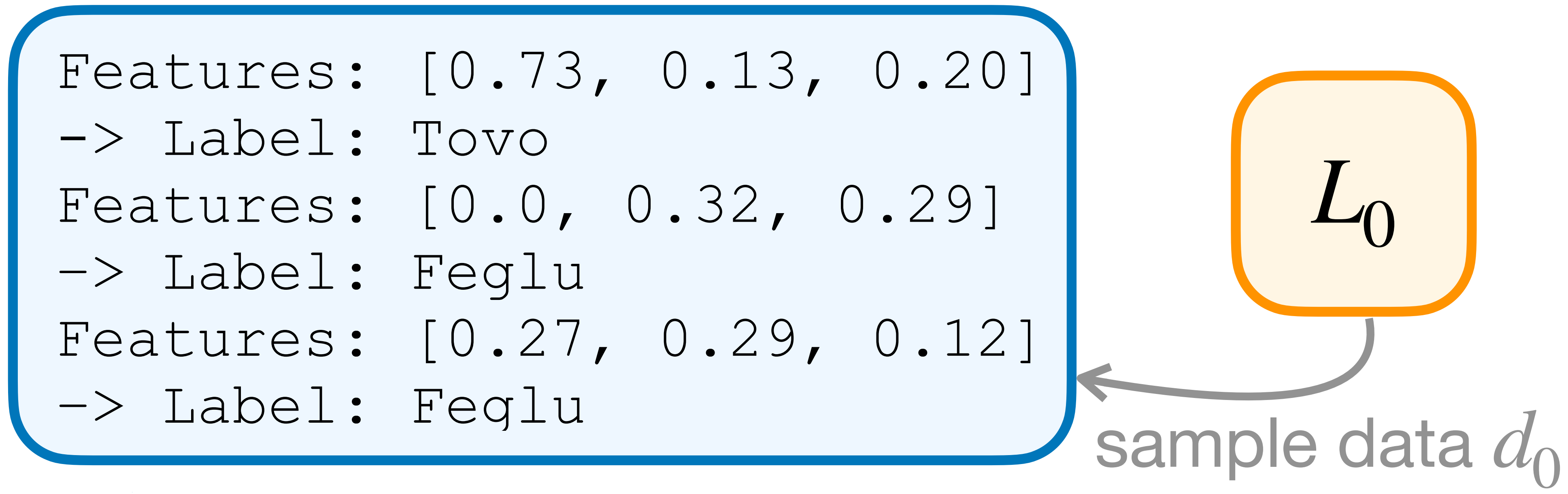
here: iterated language learning in LLMs



Adapt to LLMs with **iterated in-context language learning**



Adapt to LLMs with **iterated in-context language learning**

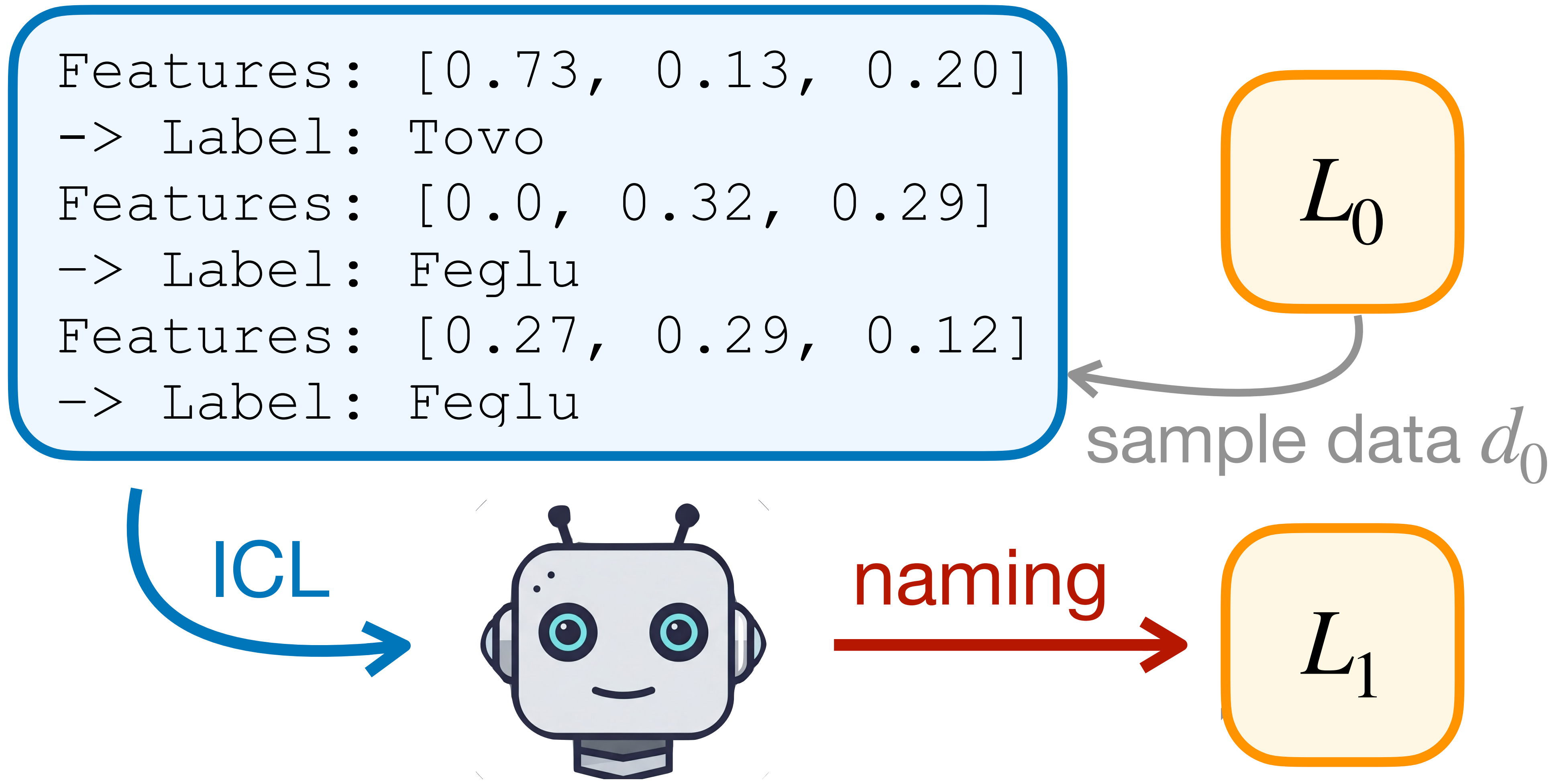


```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```

L_0

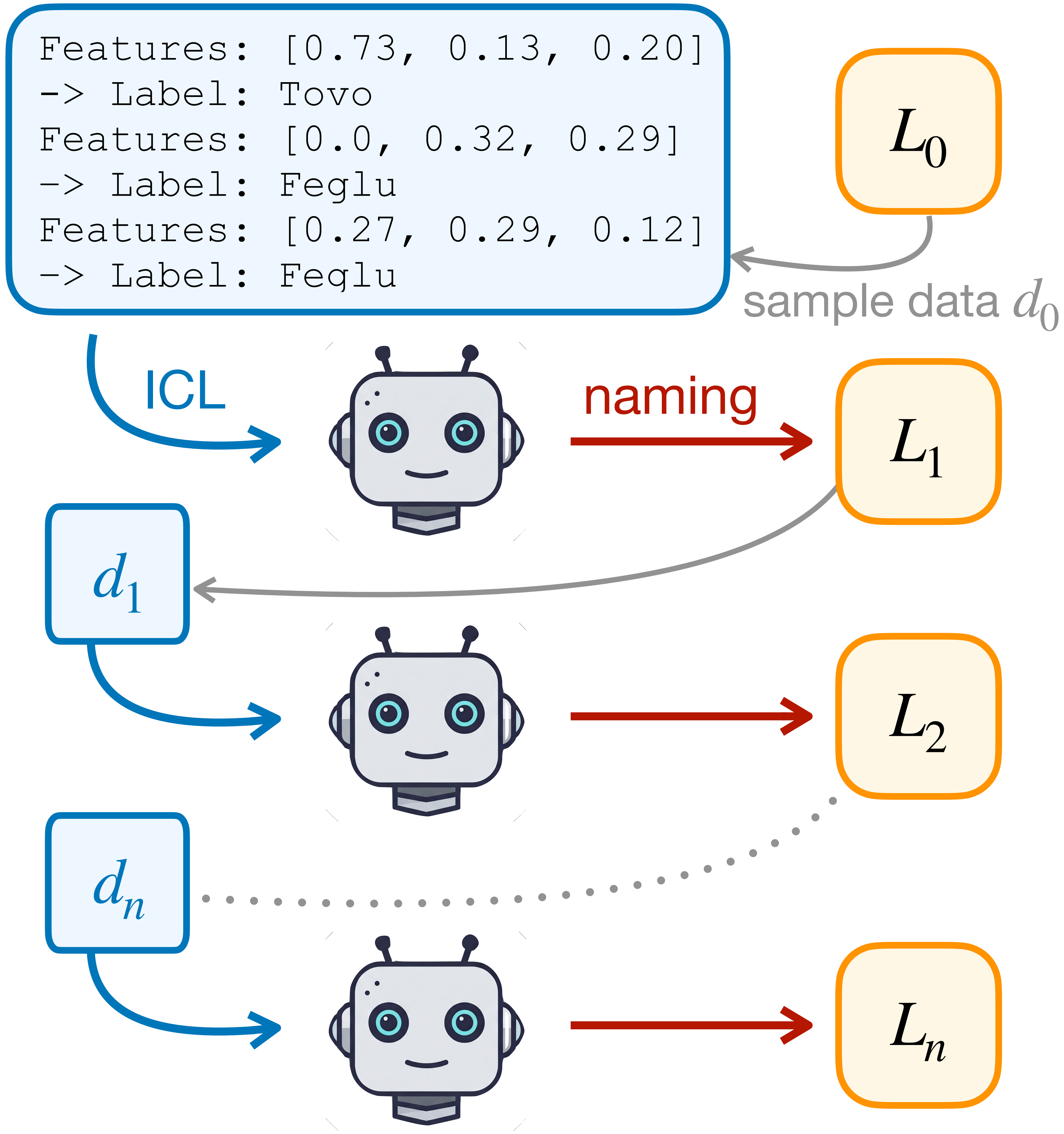
sample data d_0

Adapt to LLMs with **iterated in-context language learning**



```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```

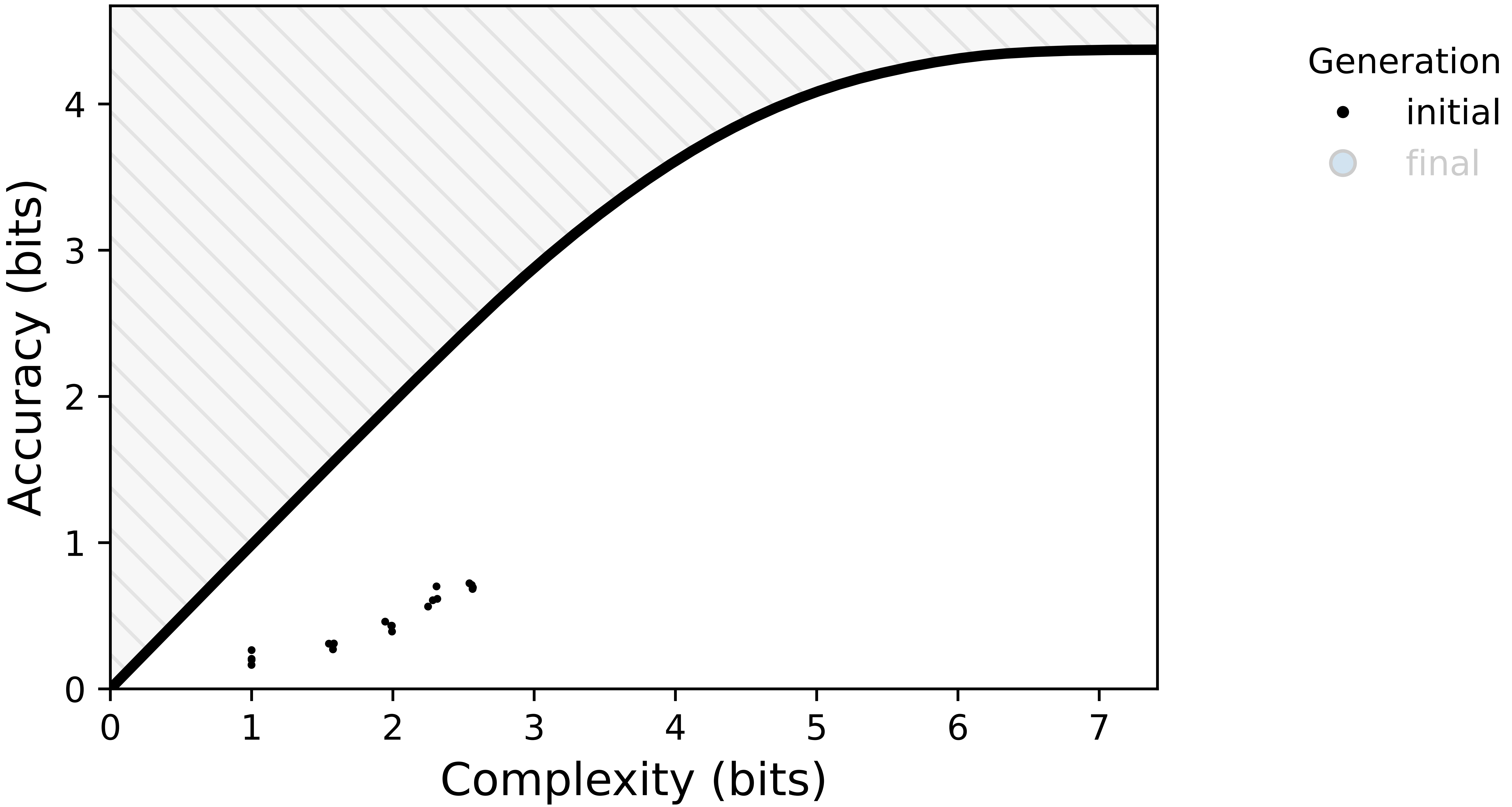
Adapt to LLMs with **iterated in-context language learning**



```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```

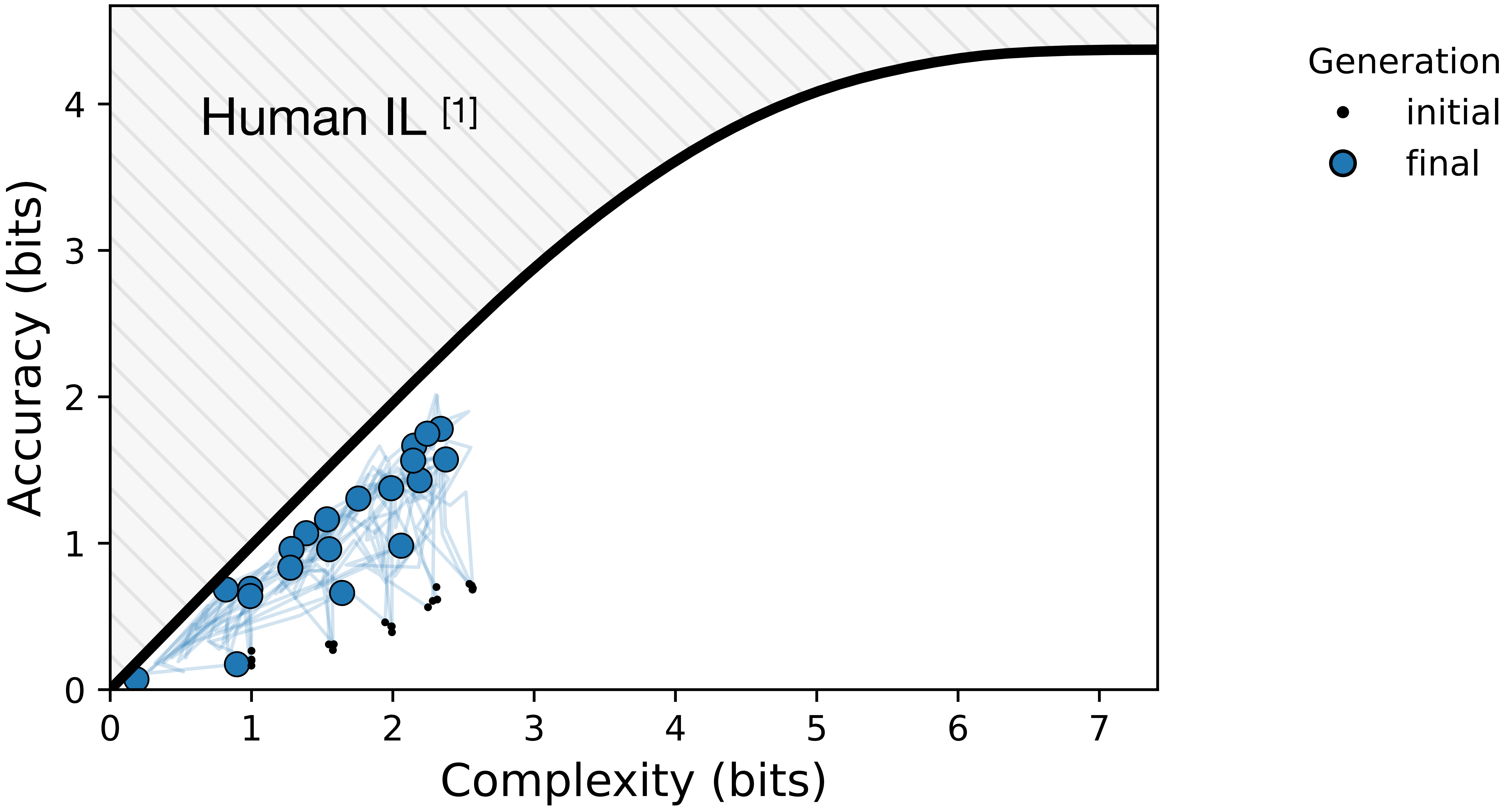
Study 2 (IICLL) Results

Study 2 (IICLL) Results



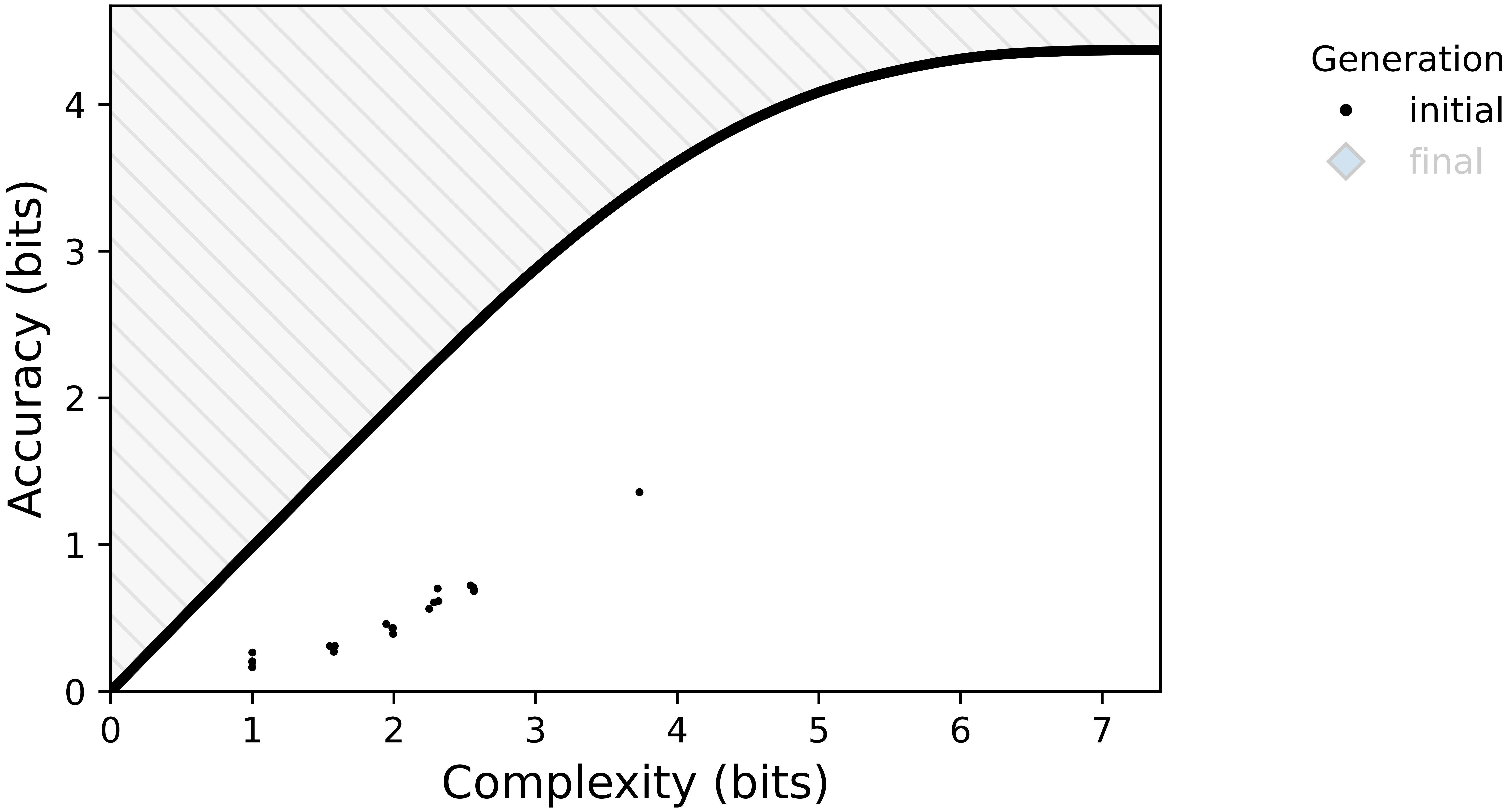
[1] Imel et al. (2025)

Study 2 (IICLL) Results

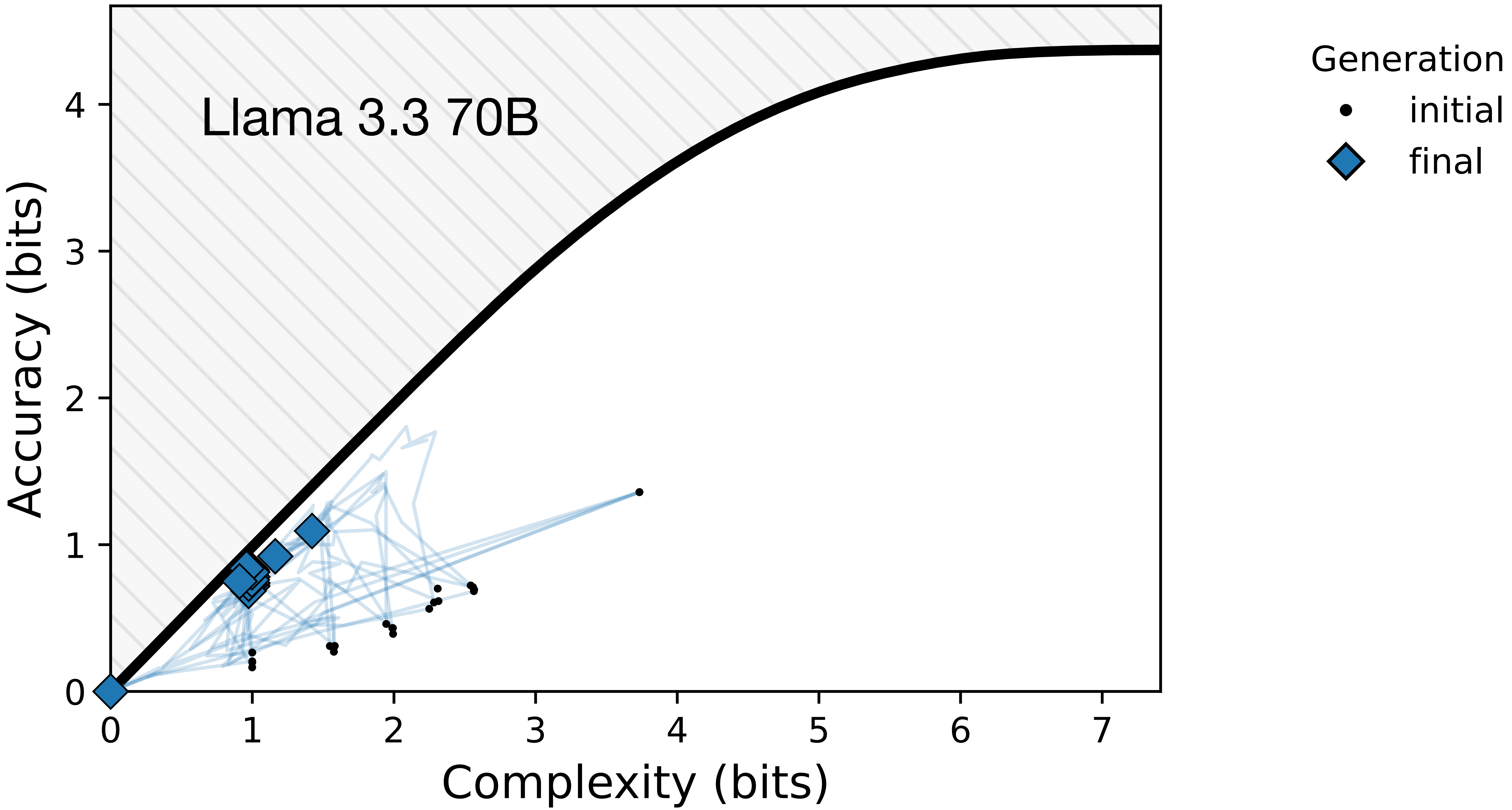


[1] Imel et al. (2025)

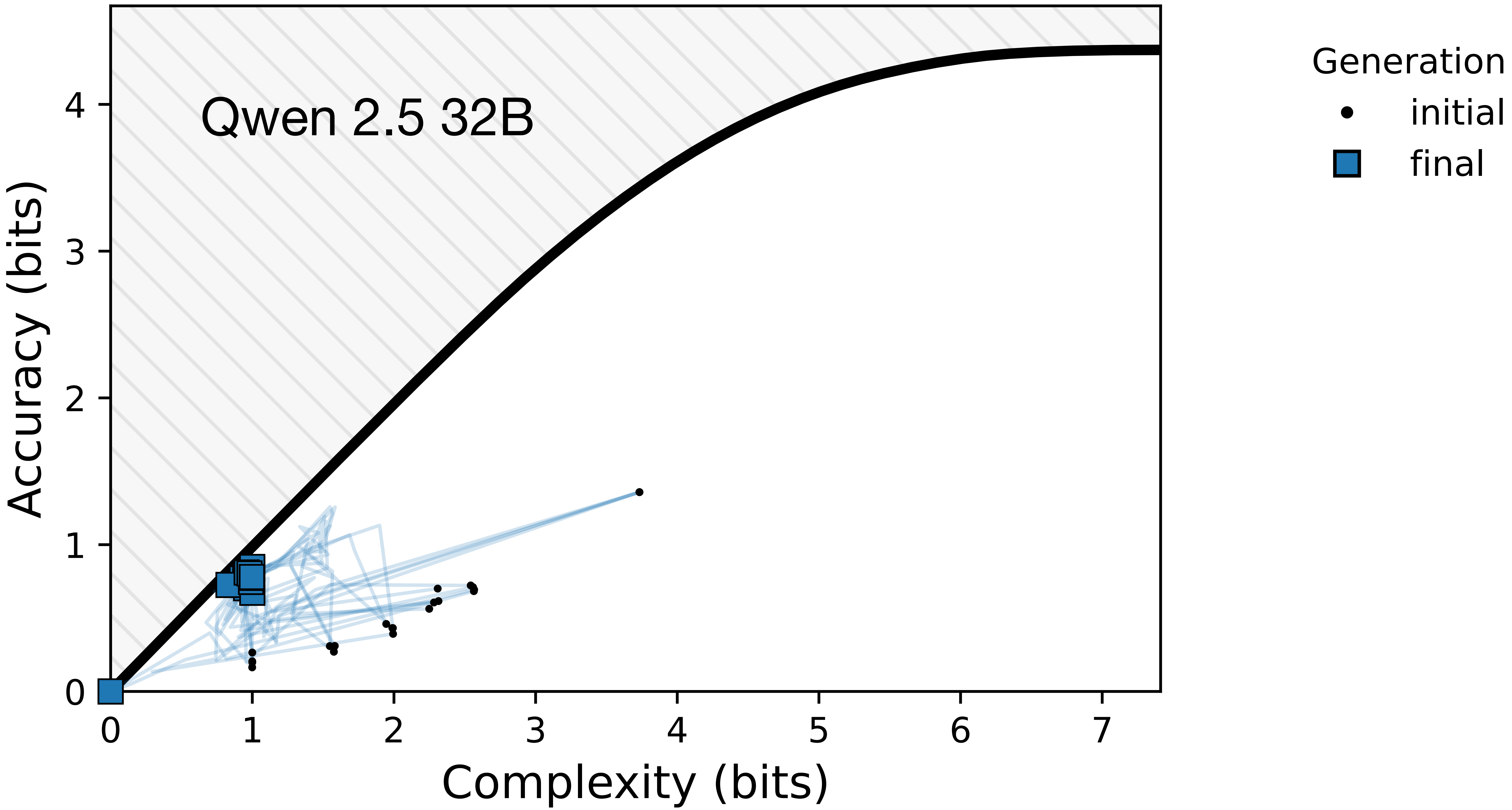
Study 2 (IICLL) Results



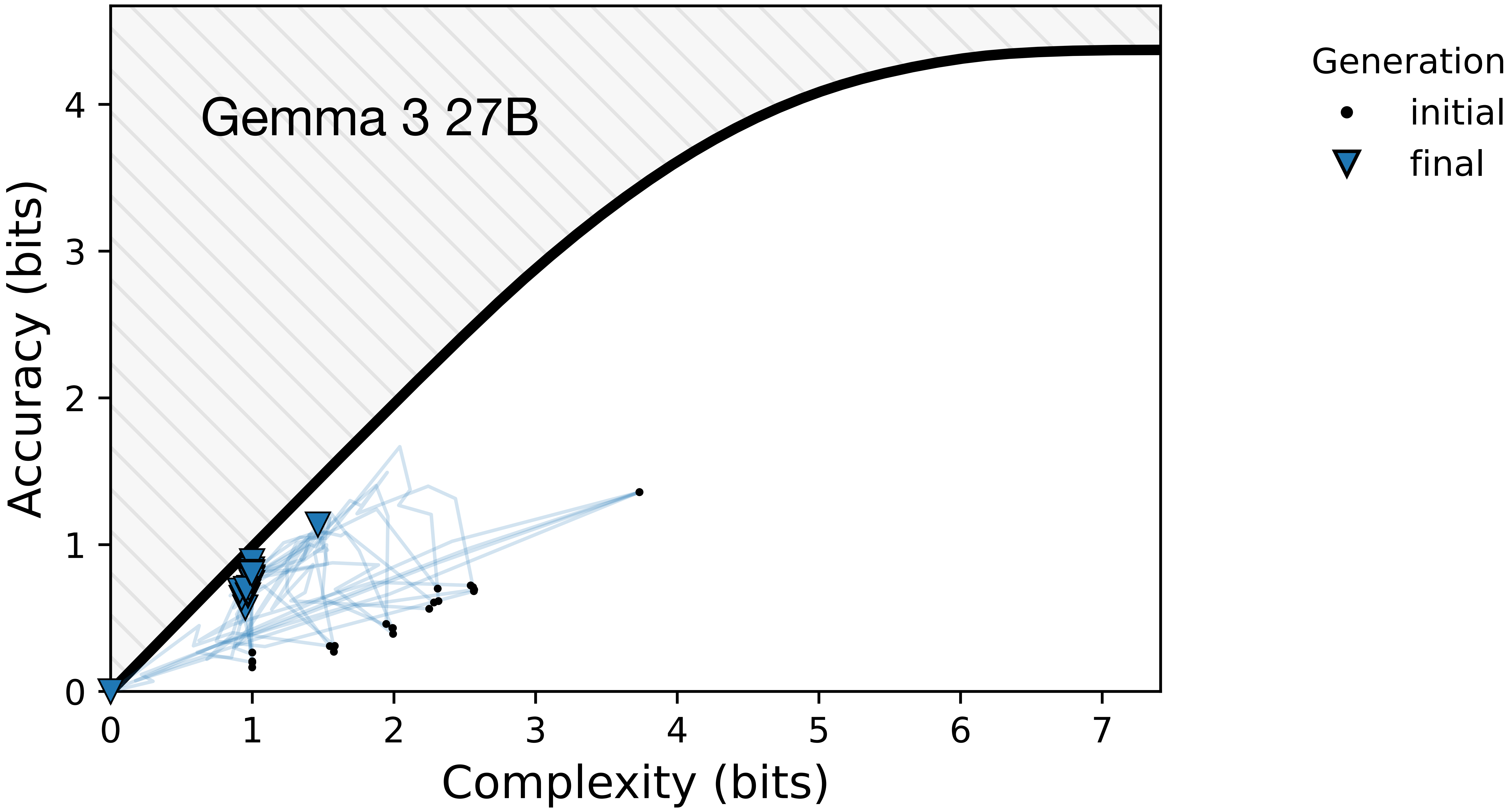
Study 2 (IICLL) Results: Llama



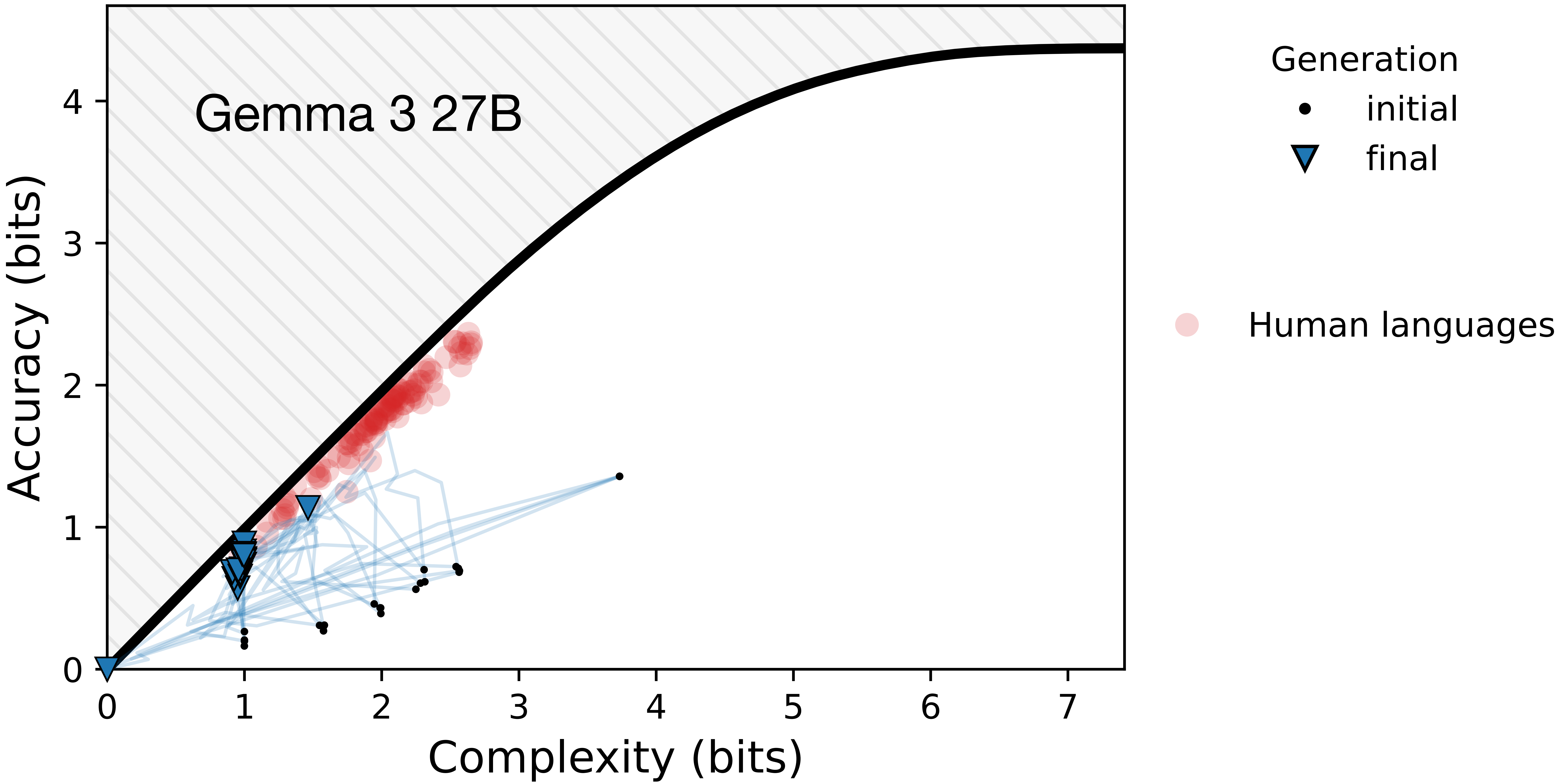
Study 2 (IICLL) Results: Qwen



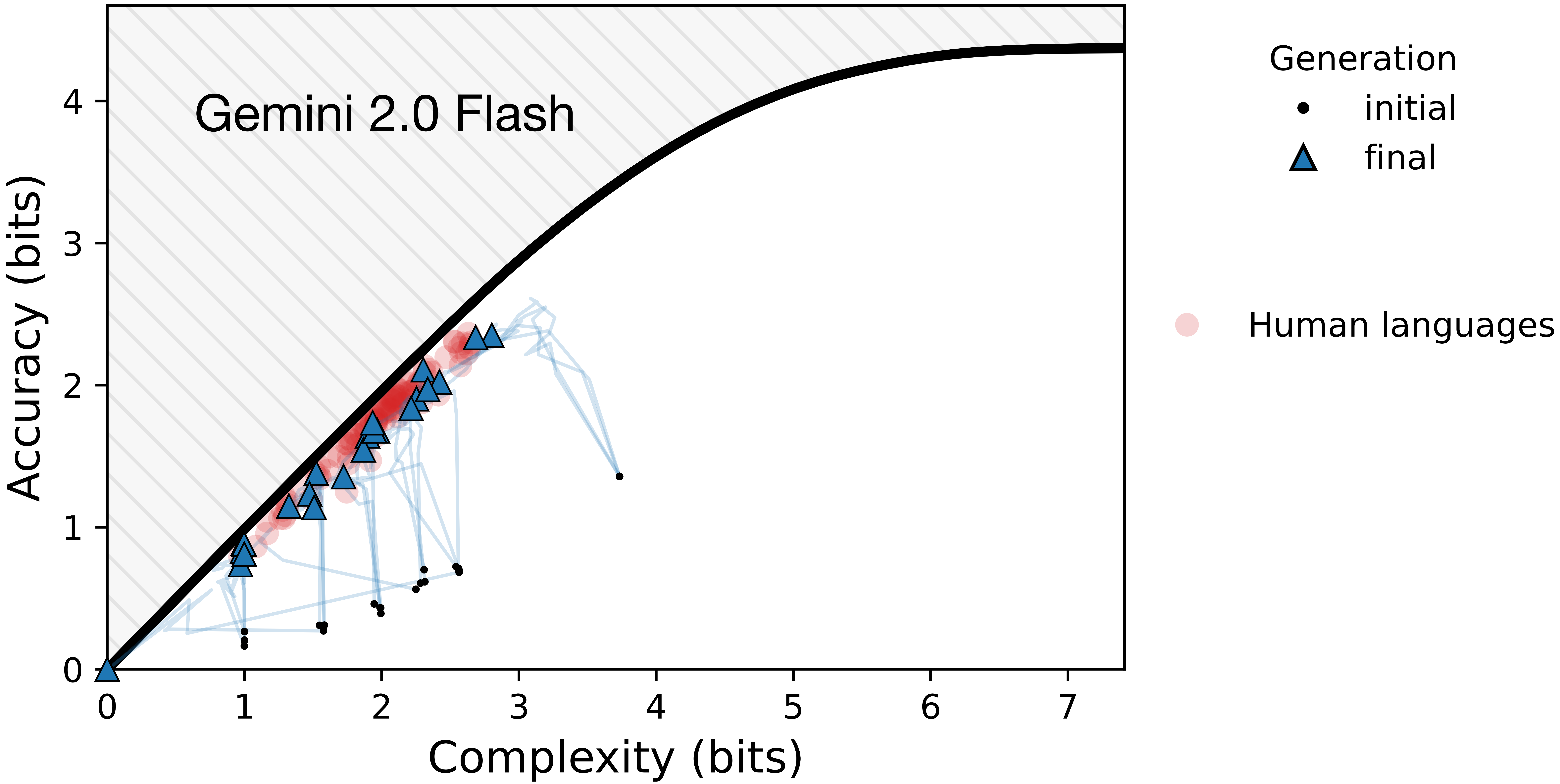
Study 2 (IICLL) Results: Gemma



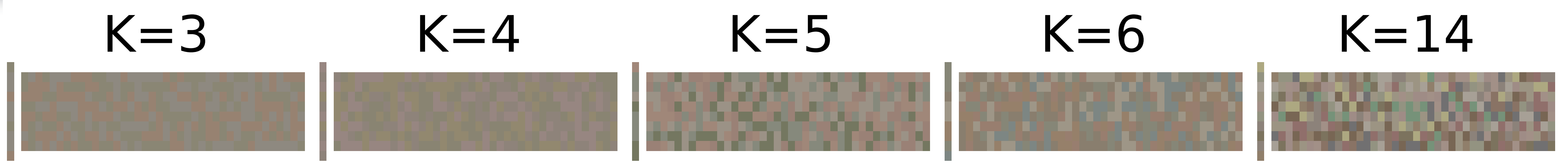
Study 2 (IICLL) Results: Gemma



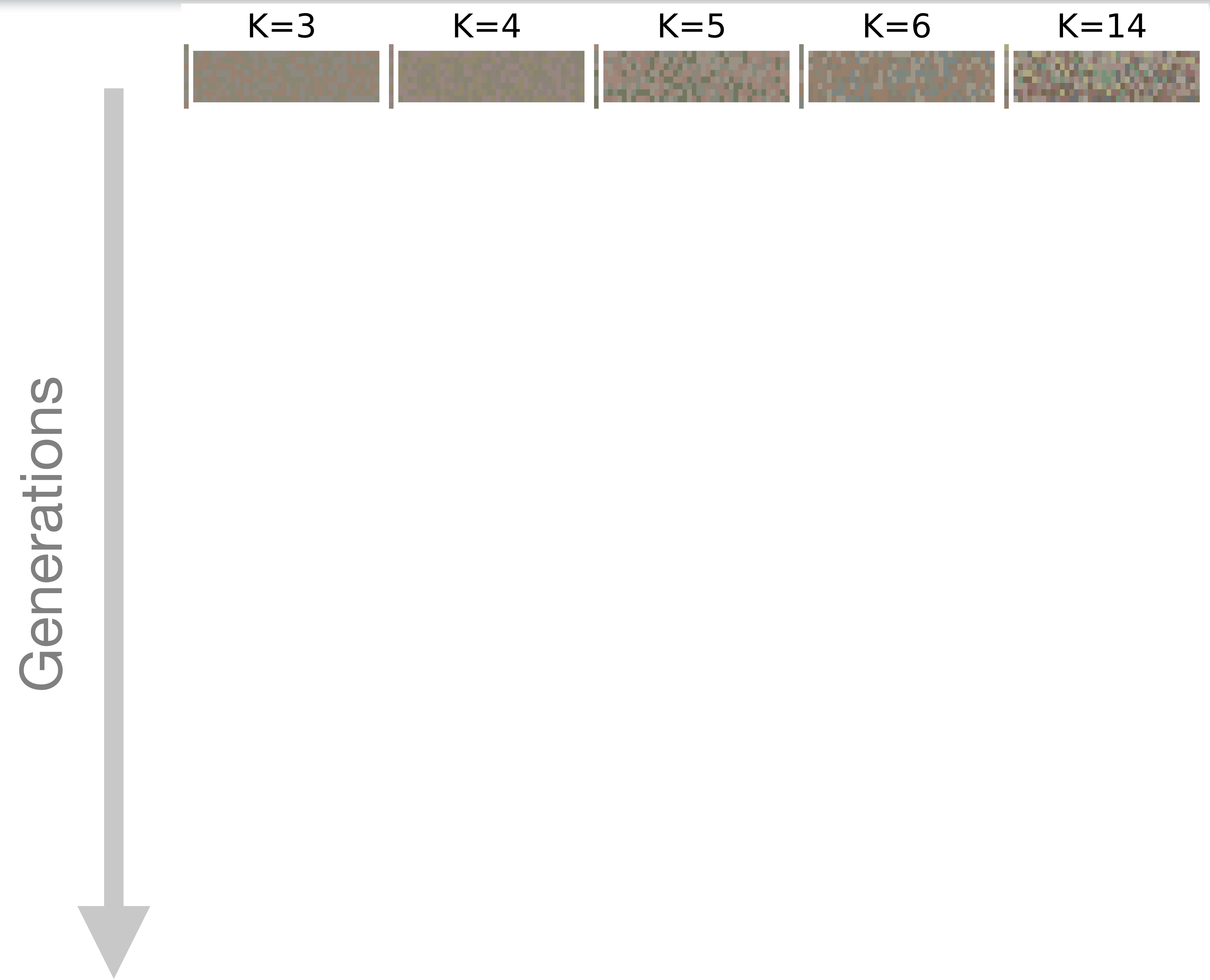
Study 2 (IICLL) Results: Gemini



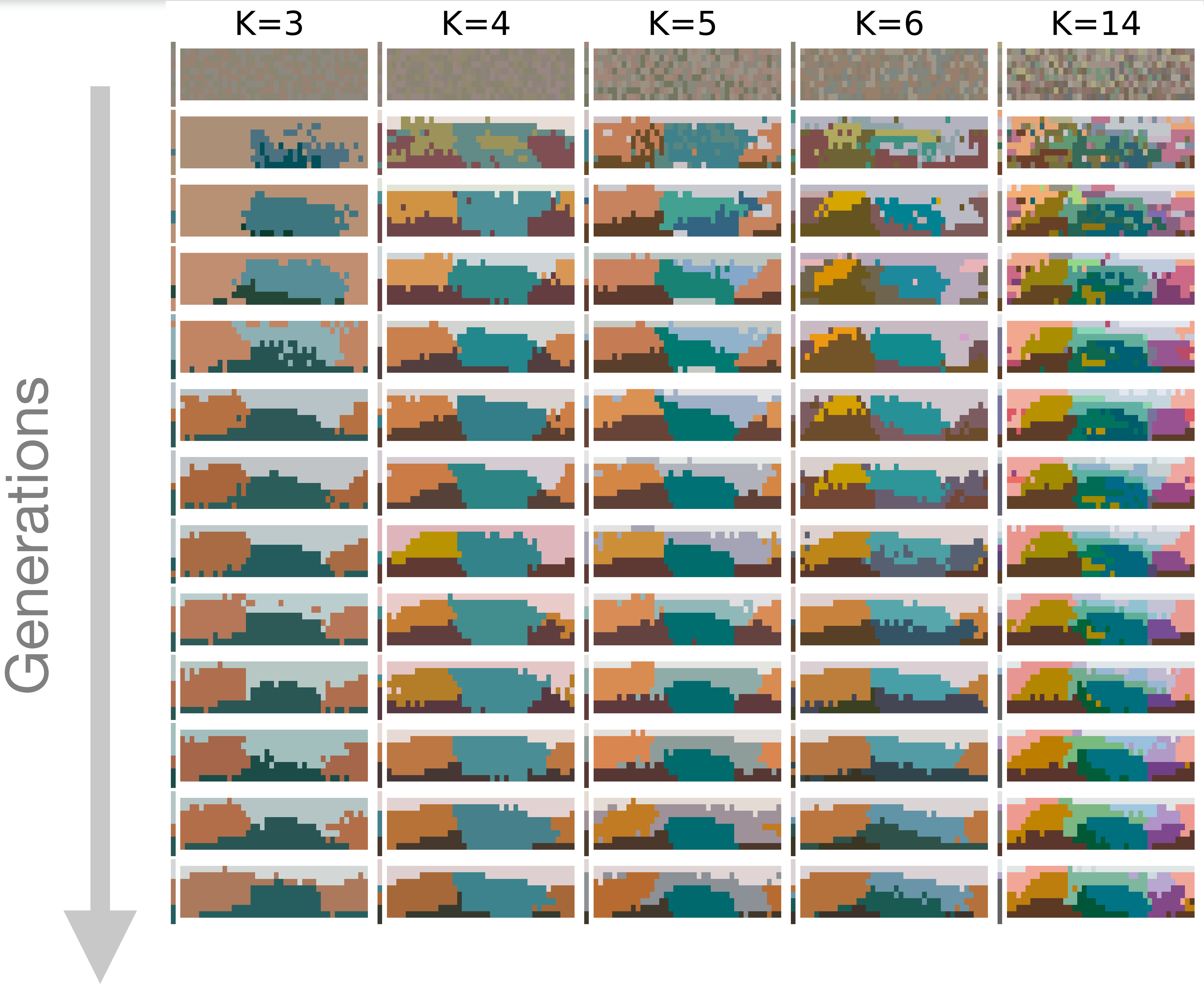
Study 2 (IICLL) Results: Gemini color systems over time



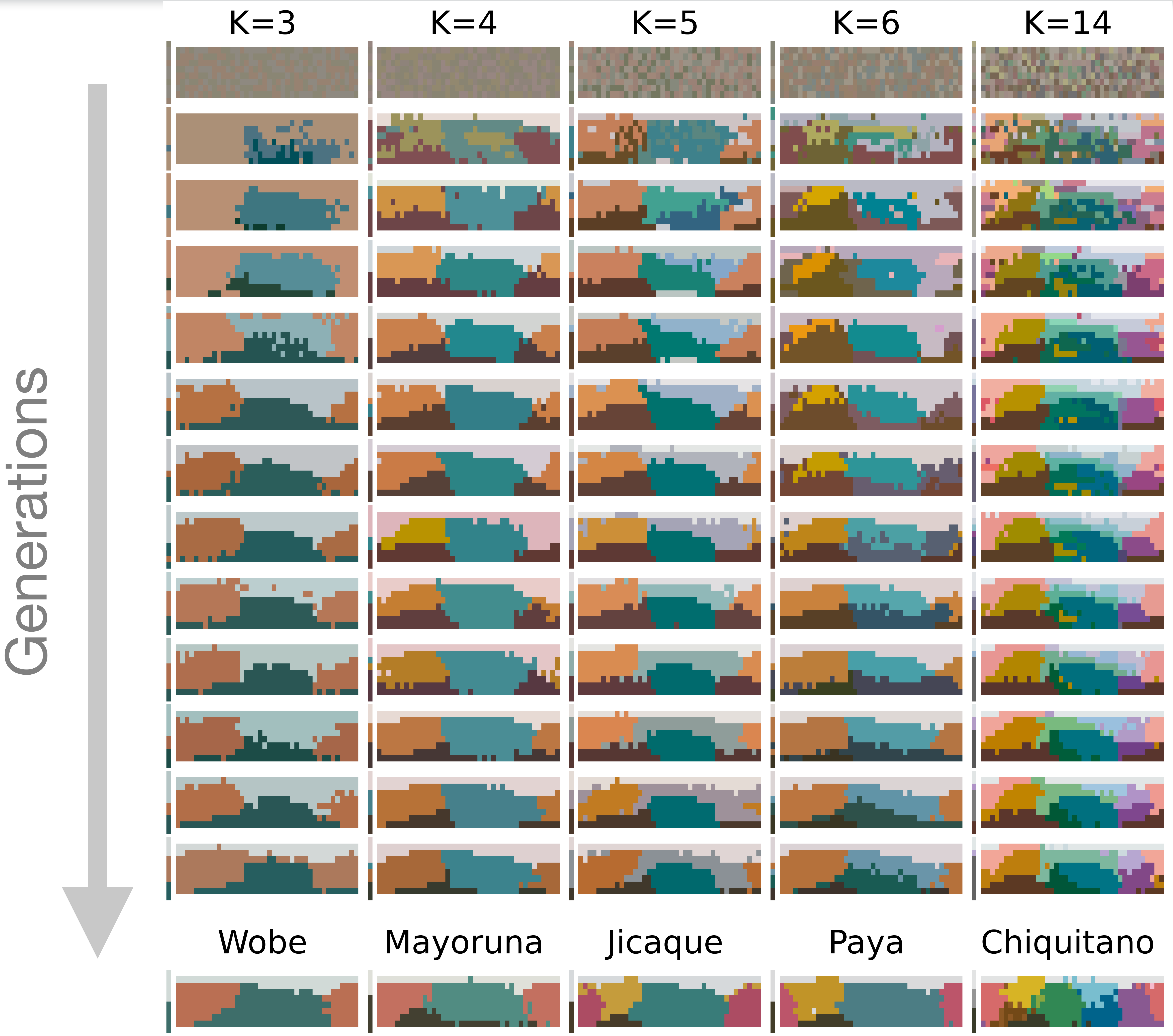
Study 2 (IICLL) Results: Gemini color systems over time



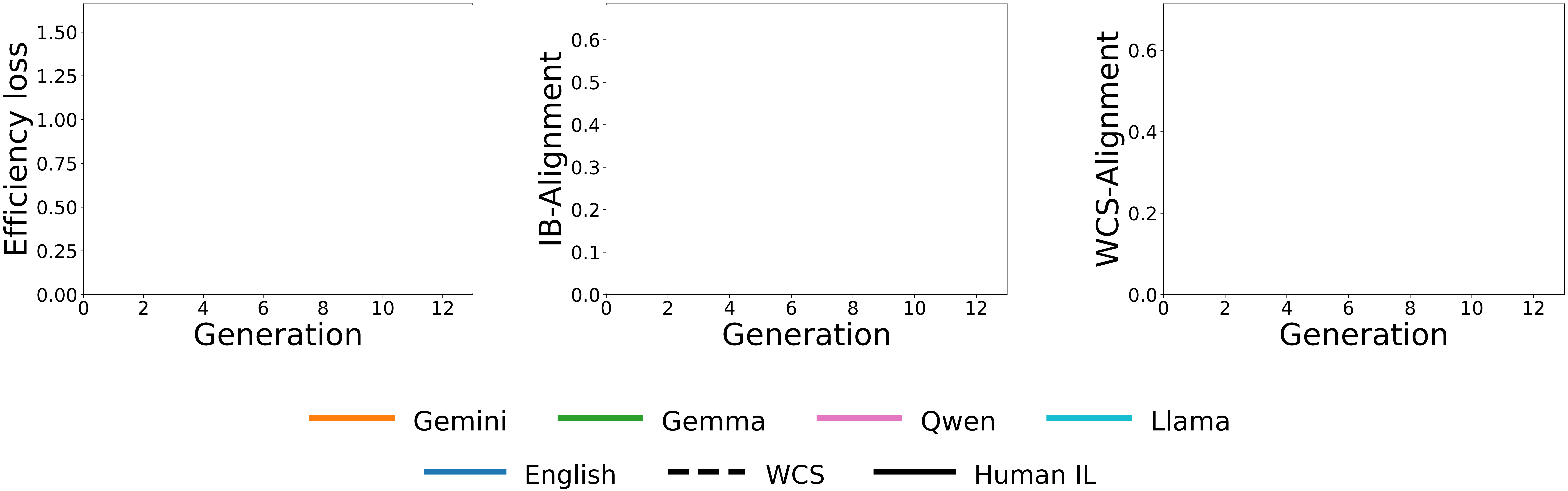
Study 2 (IICLL) Results: Gemini color systems over time



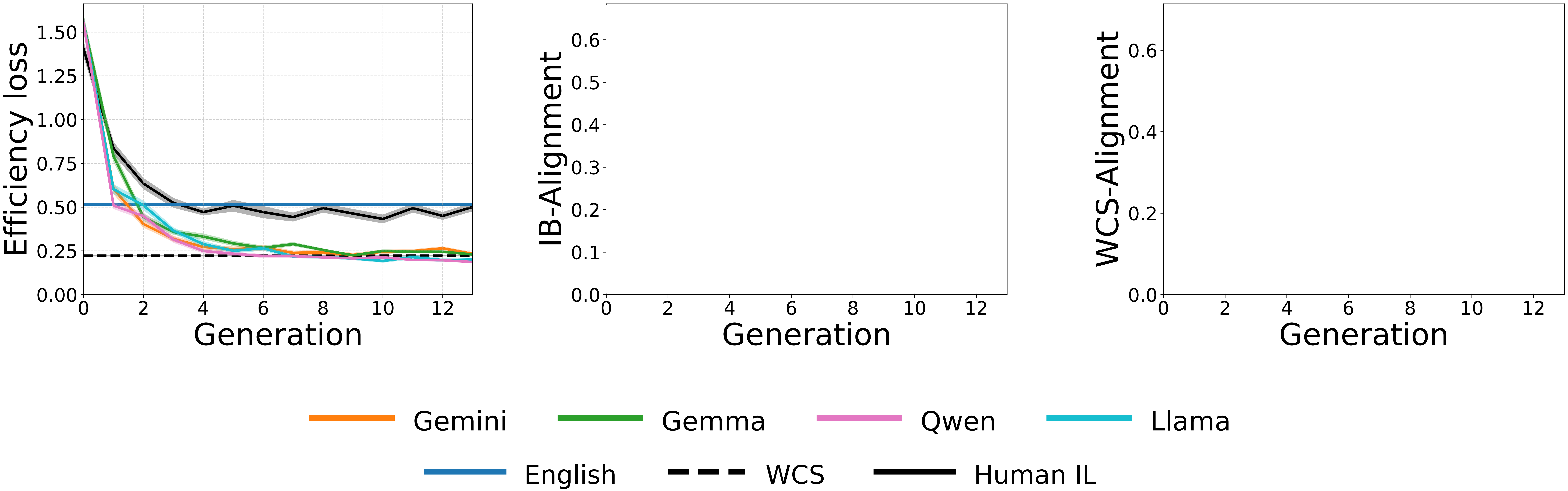
Study 2 (IICLL) Results: Gemini color systems over time



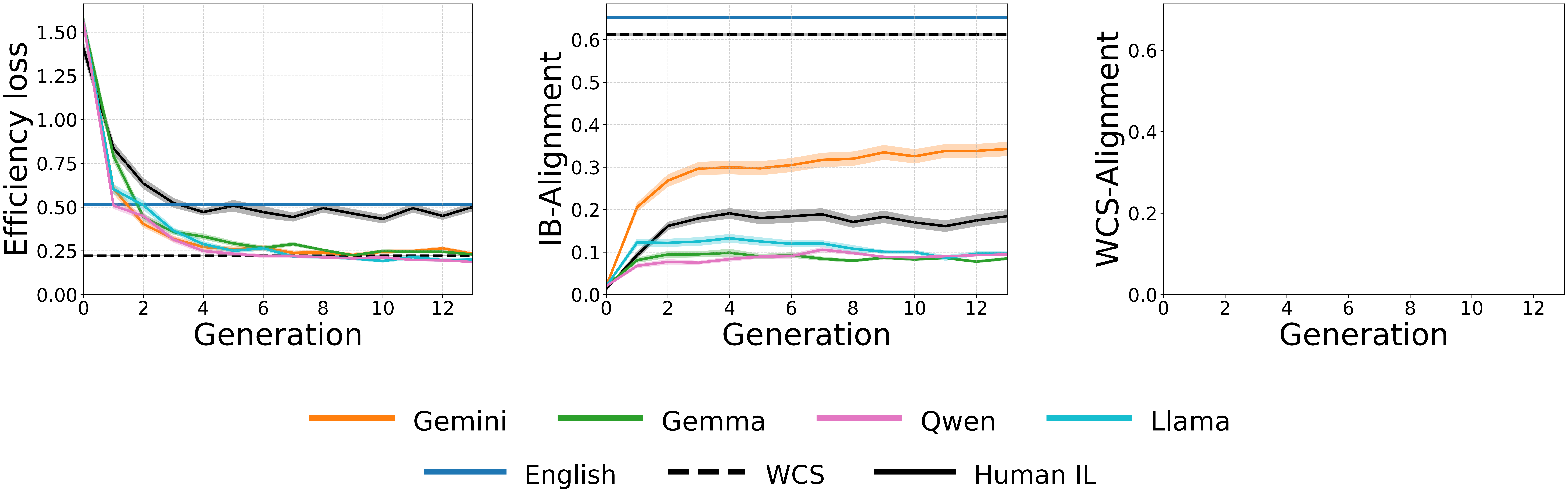
Study 2 (IICLL) Results: efficiency and alignment over time



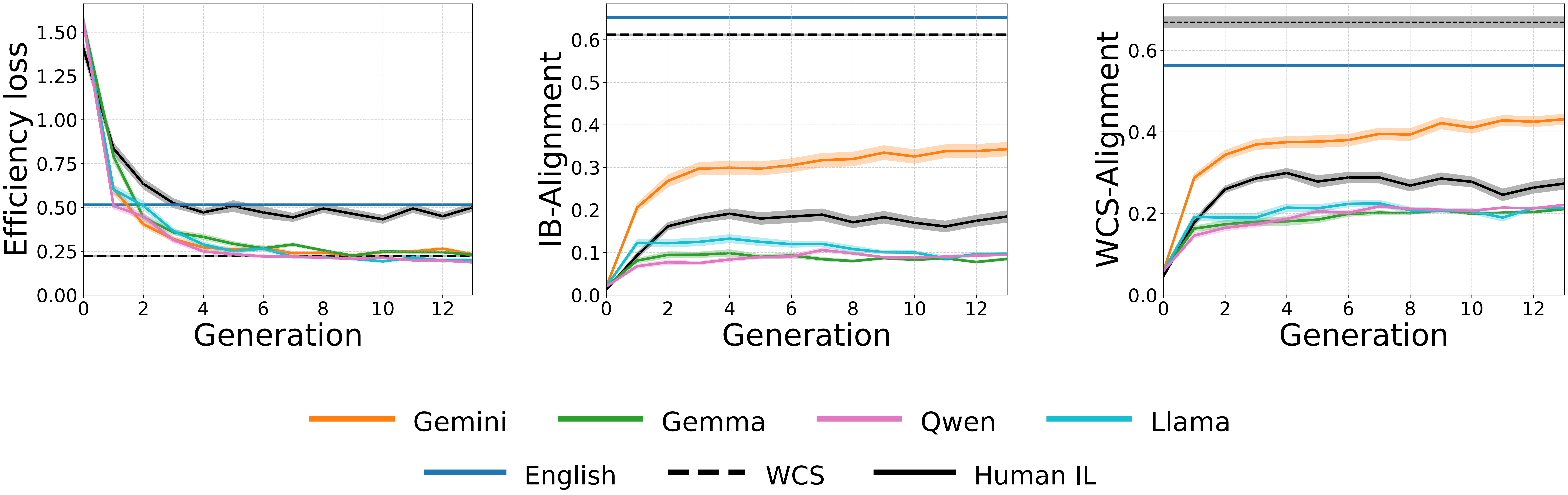
Study 2 (IICLL) Results: efficiency and alignment over time



Study 2 (IICLL) Results: efficiency and alignment over time



Study 2 (IICLL) Results: efficiency and alignment over time



Conclusions

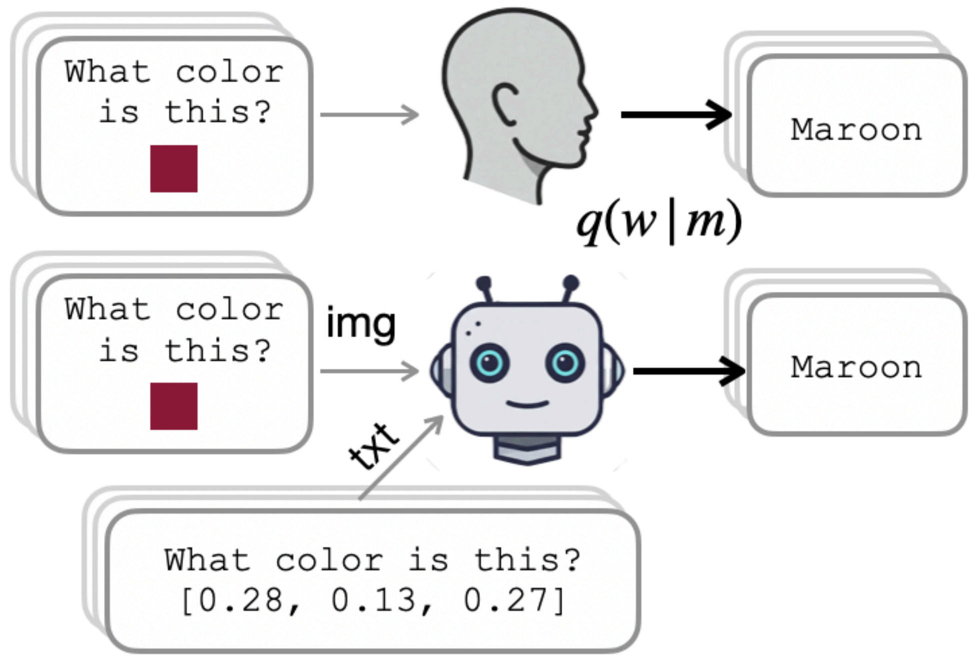
Conclusions

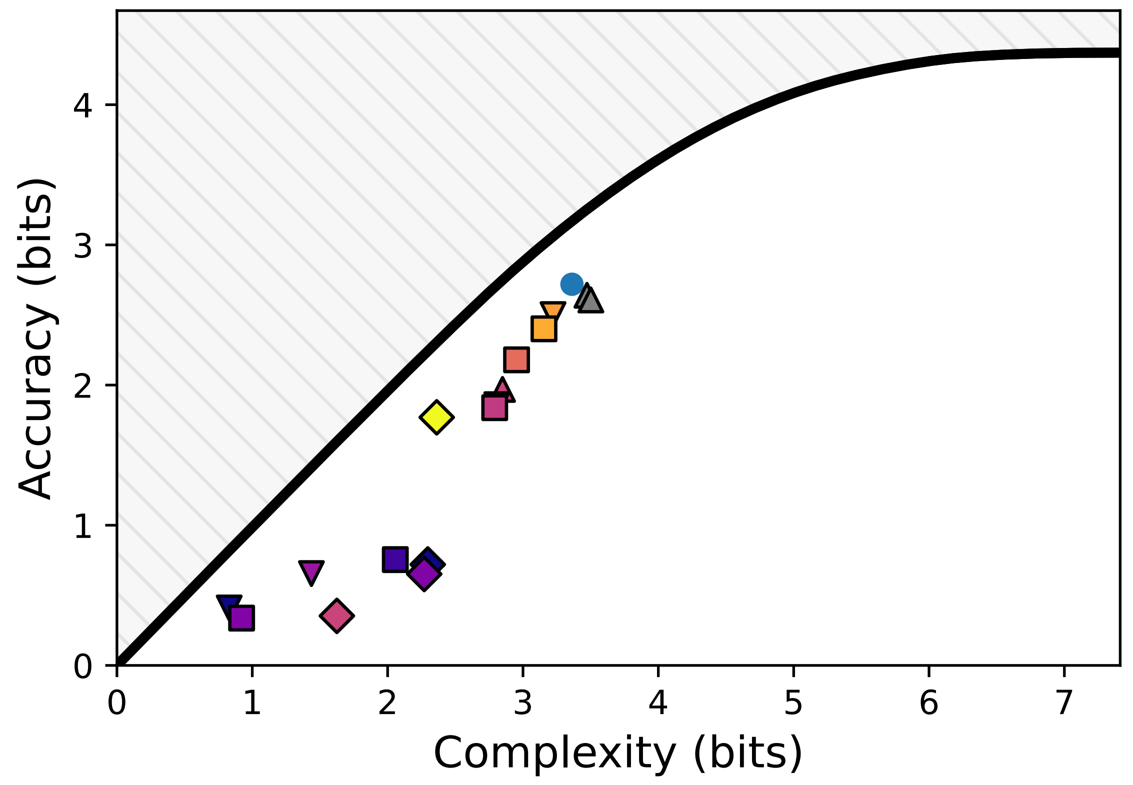
- **human-aligned semantic categories can emerge in LLMs** via the same fundamental principle that underlies semantic efficiency in humans.

Conclusions

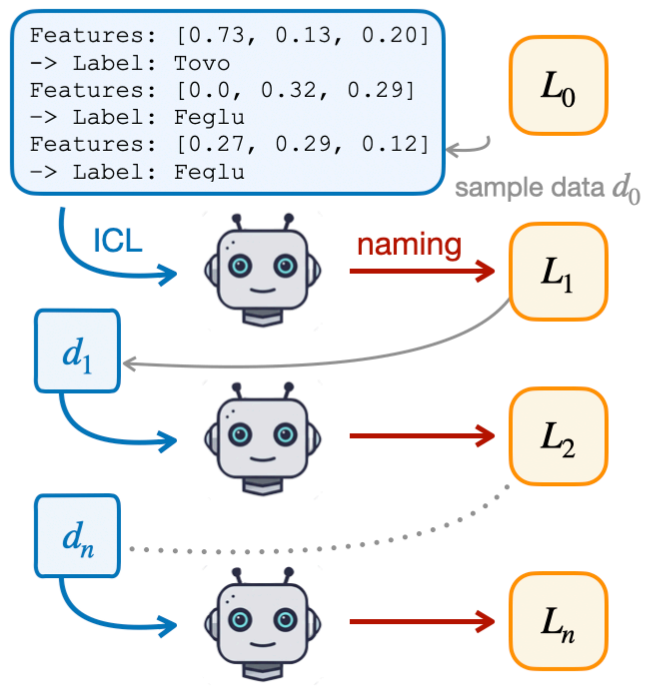
- **human-aligned semantic categories can emerge in LLMs** via the same fundamental principle that underlies semantic efficiency in humans.

English color naming

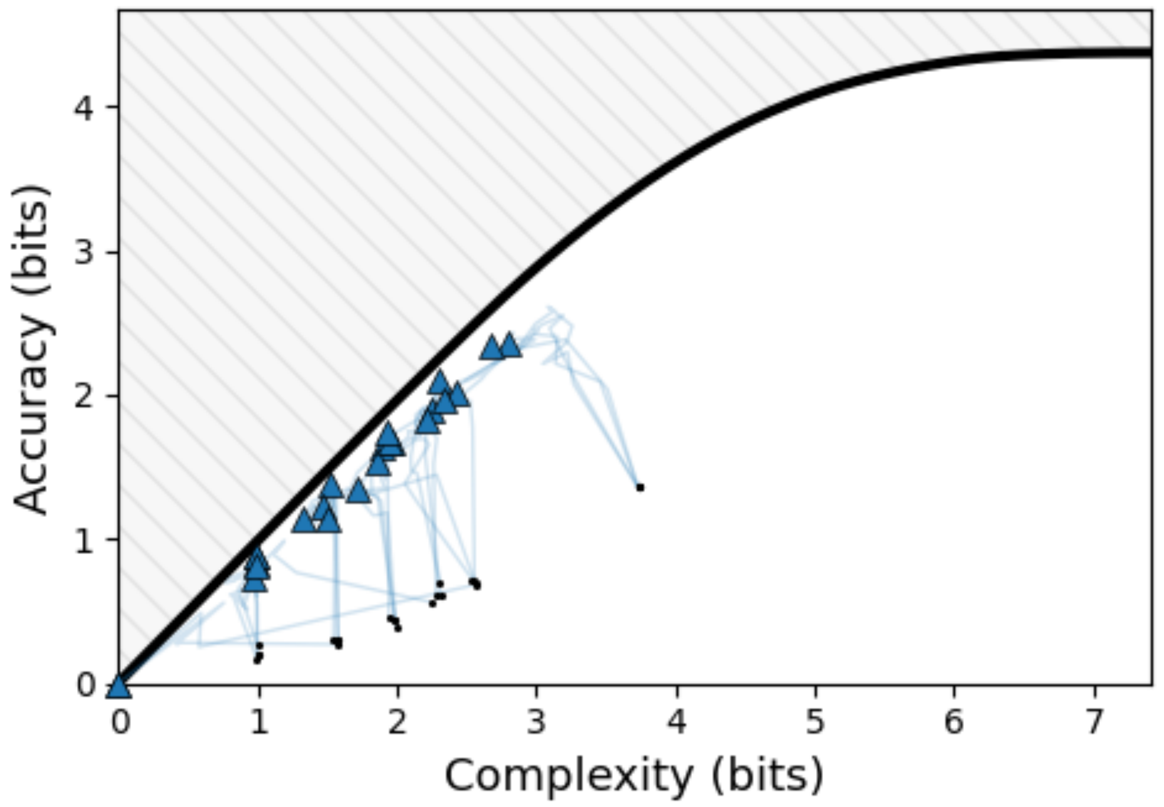




Cultural transmission via IICLL



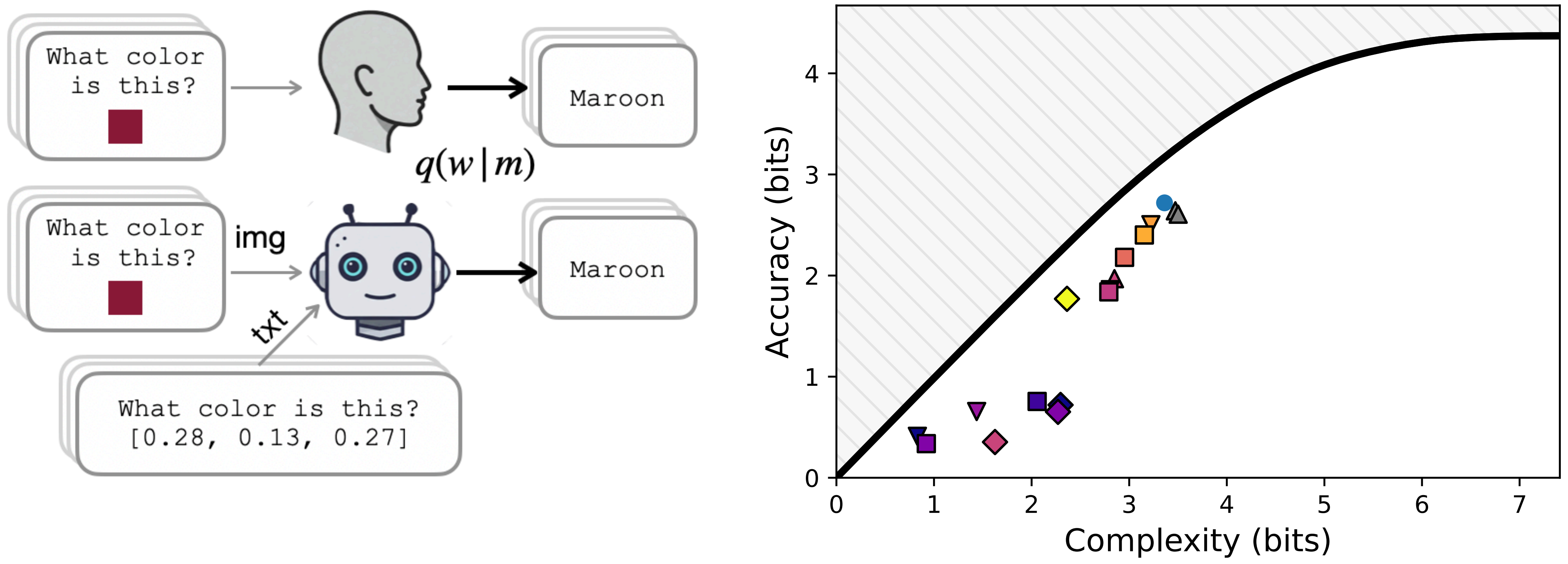
```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```



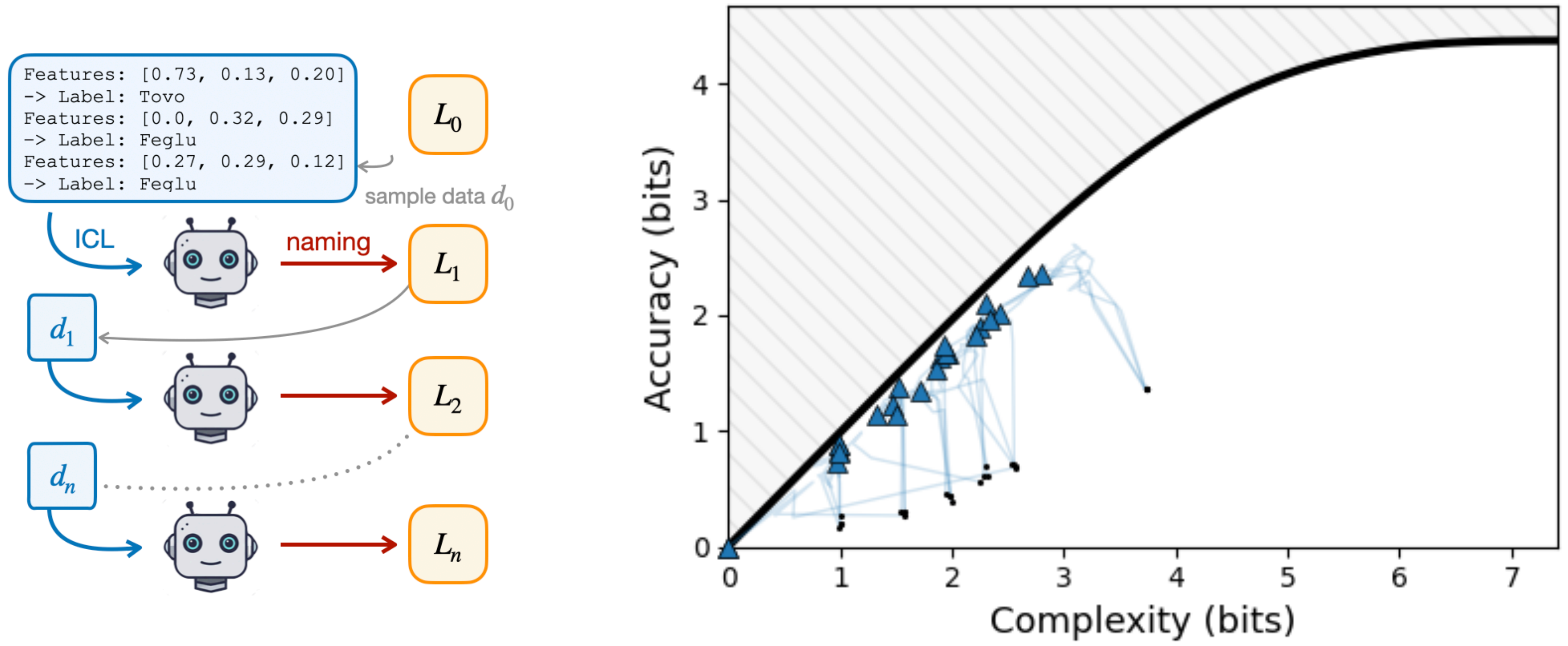
Conclusions

- **human-aligned semantic categories can emerge in LLMs** via the same fundamental principle that underlies semantic efficiency in humans.

English color naming



Cultural transmission via IICLL



```
Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu
```

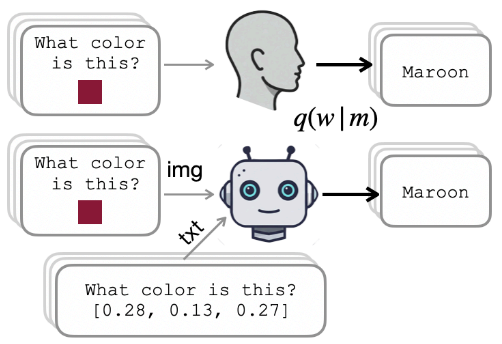
- Importantly, neither humans nor LLMs are **explicitly trained** for optimizing the IB objective, suggesting IB-efficiency may **emerge** to support intelligent behavior.

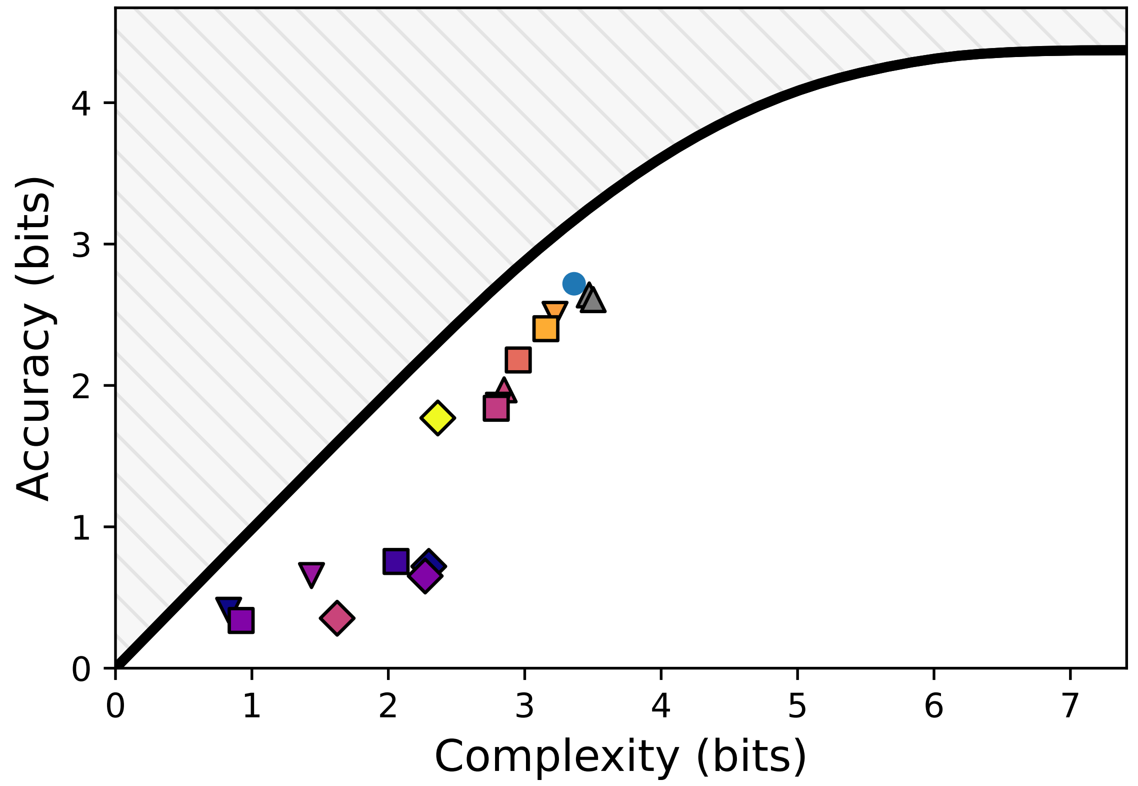
Thank you!

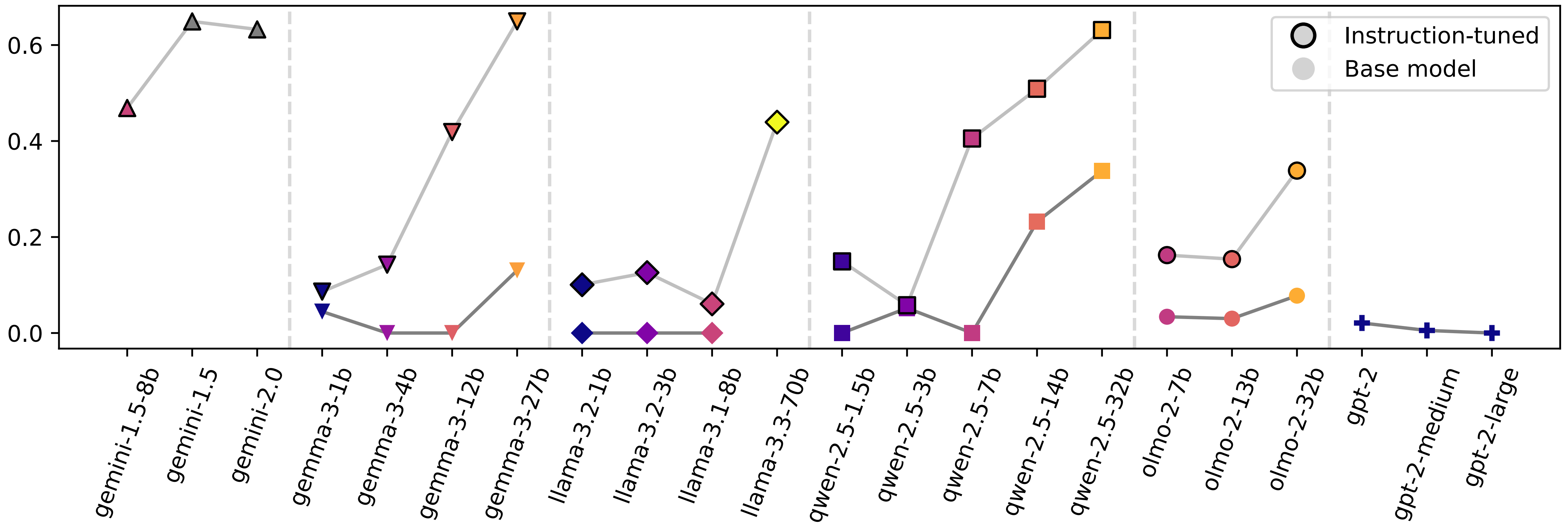
infoCogLab
information cognition language

preprint ->

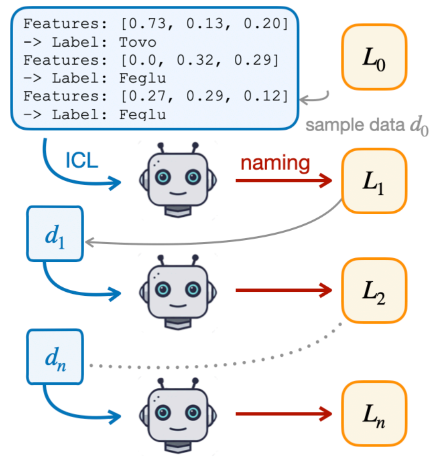
English color naming

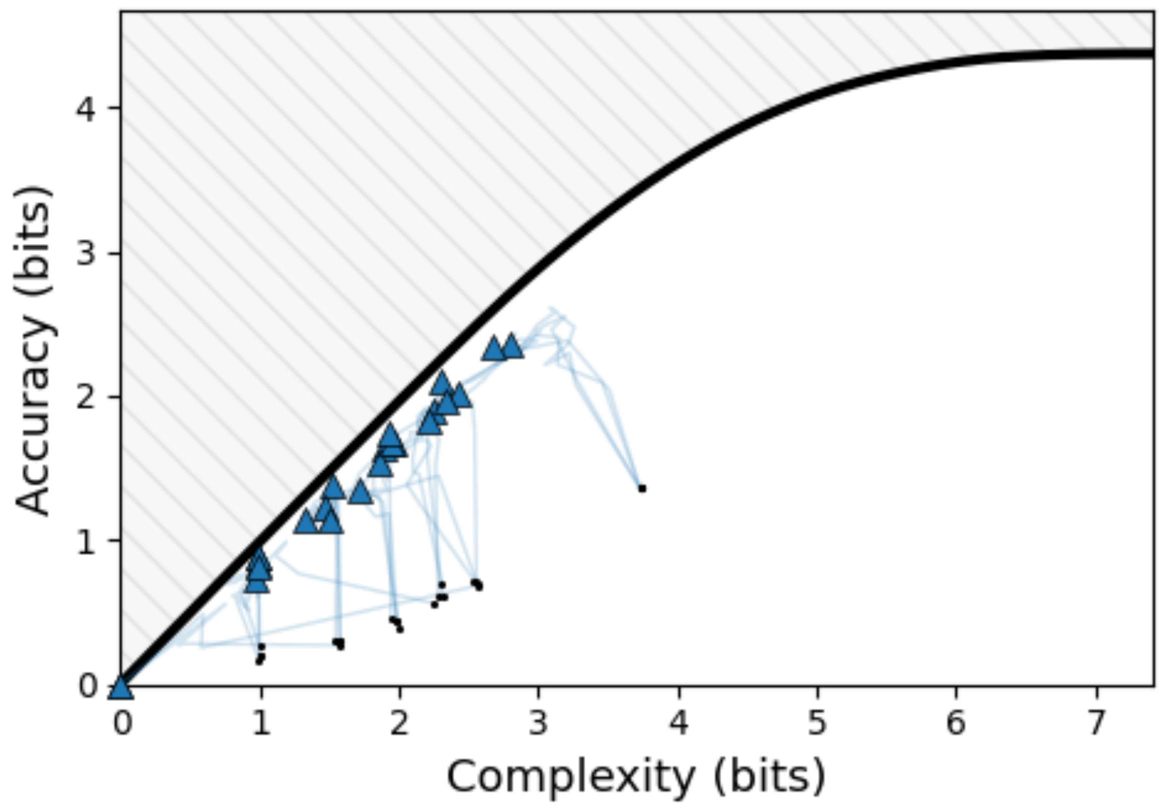


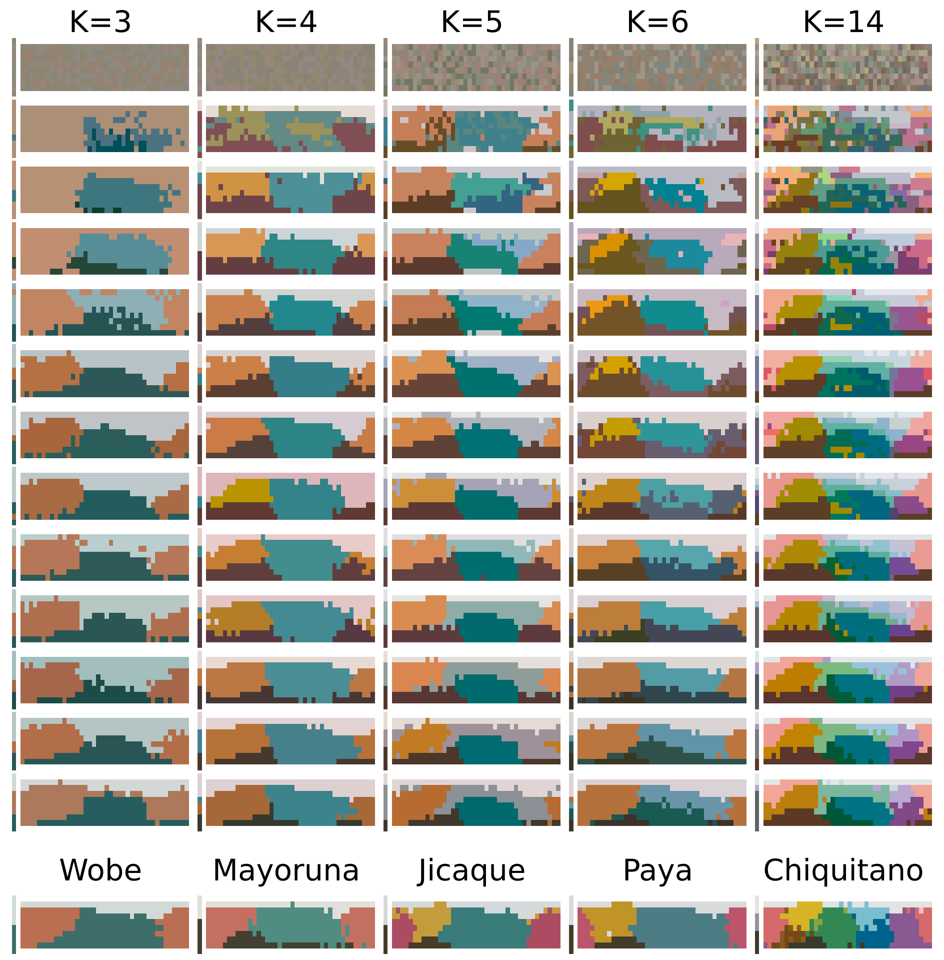




Cultural transmission (IICLL)


Features: [0.73, 0.13, 0.20]
-> Label: Tovo
Features: [0.0, 0.32, 0.29]
-> Label: Feglu
Features: [0.27, 0.29, 0.12]
-> Label: Feglu





Acknowledgements

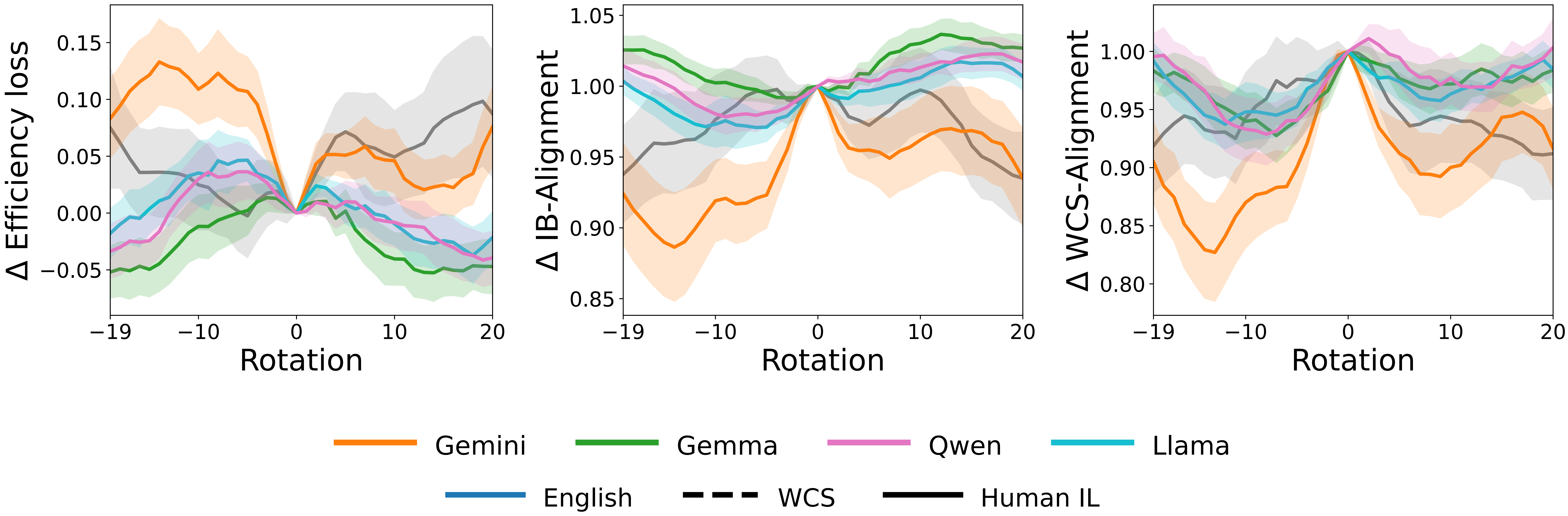
Noga Zaslavsky
(Advisor)

*This work was supported by a Google Cloud Platform Credit Award

Jing Xu
(Data)

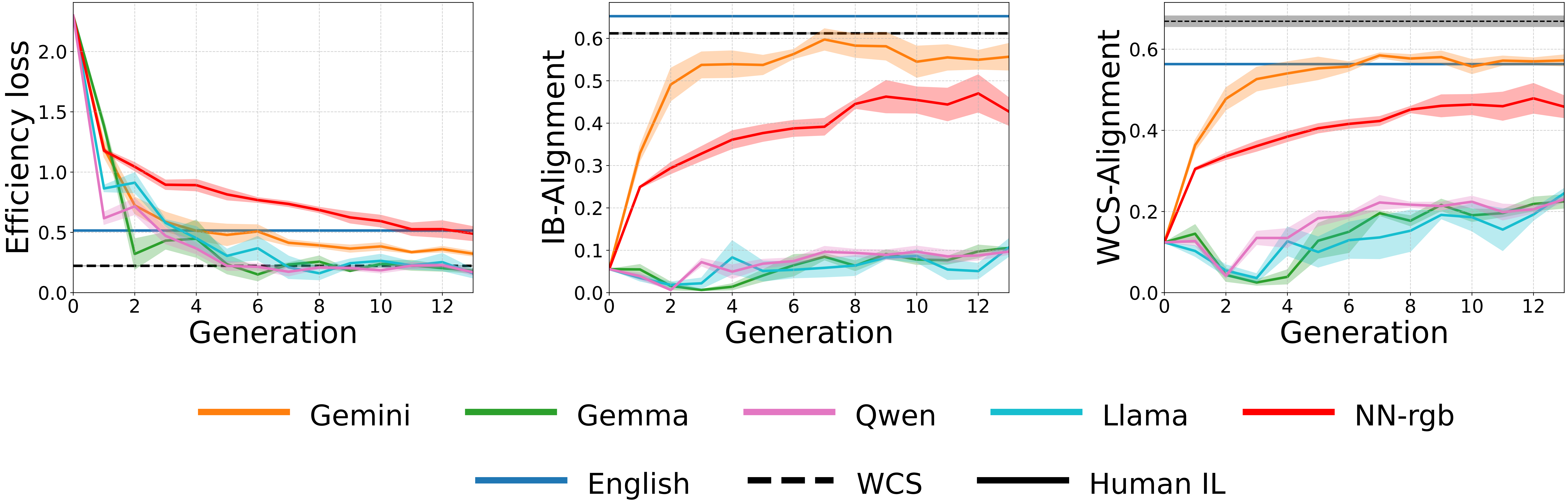
extra material

Study 2: rotation analysis of final generation systems



nearest neighbor baseline

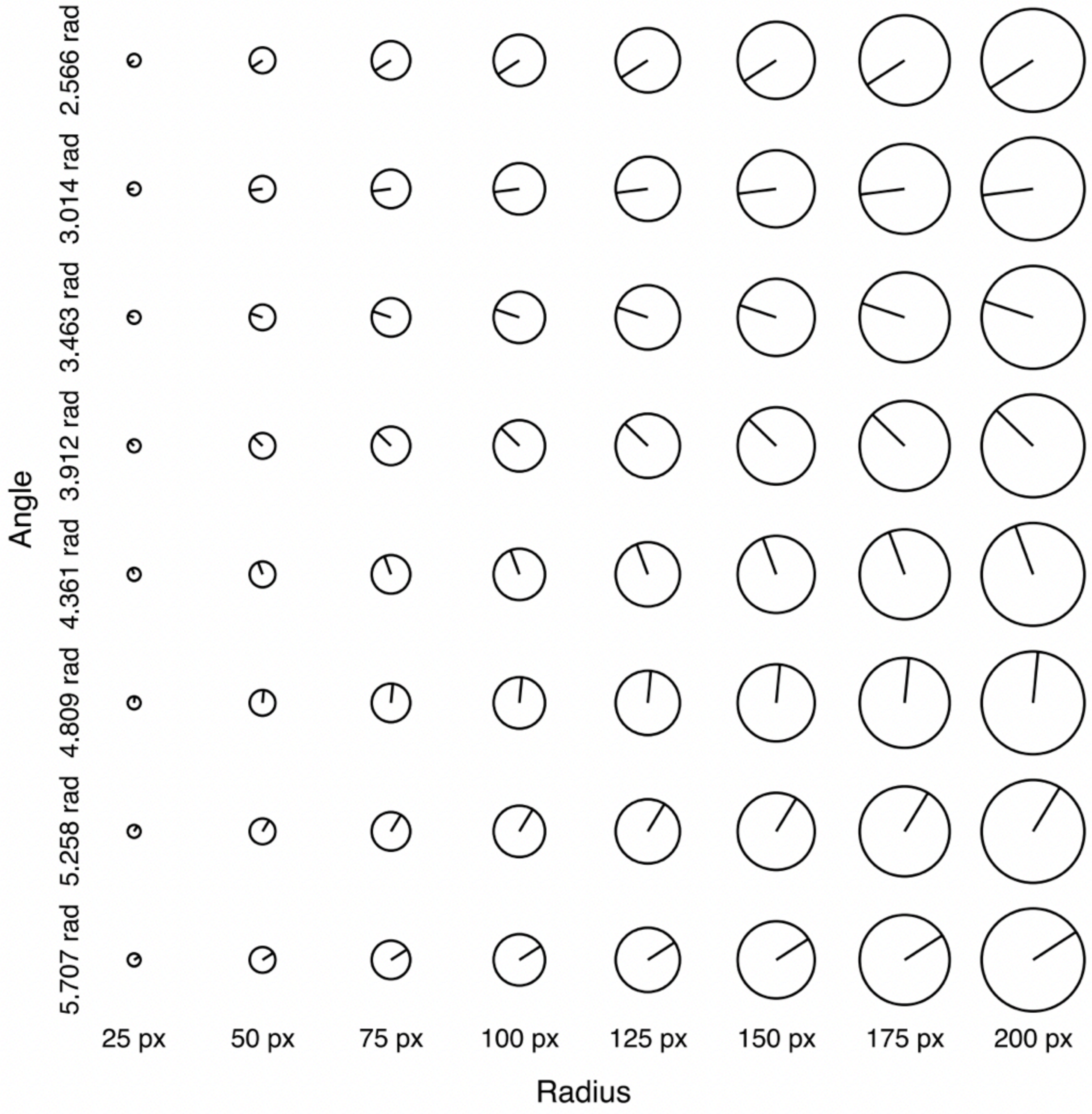
nearest neighbor baseline



$k = 14$ condition

Preliminary results: IICLL beyond the domain of color

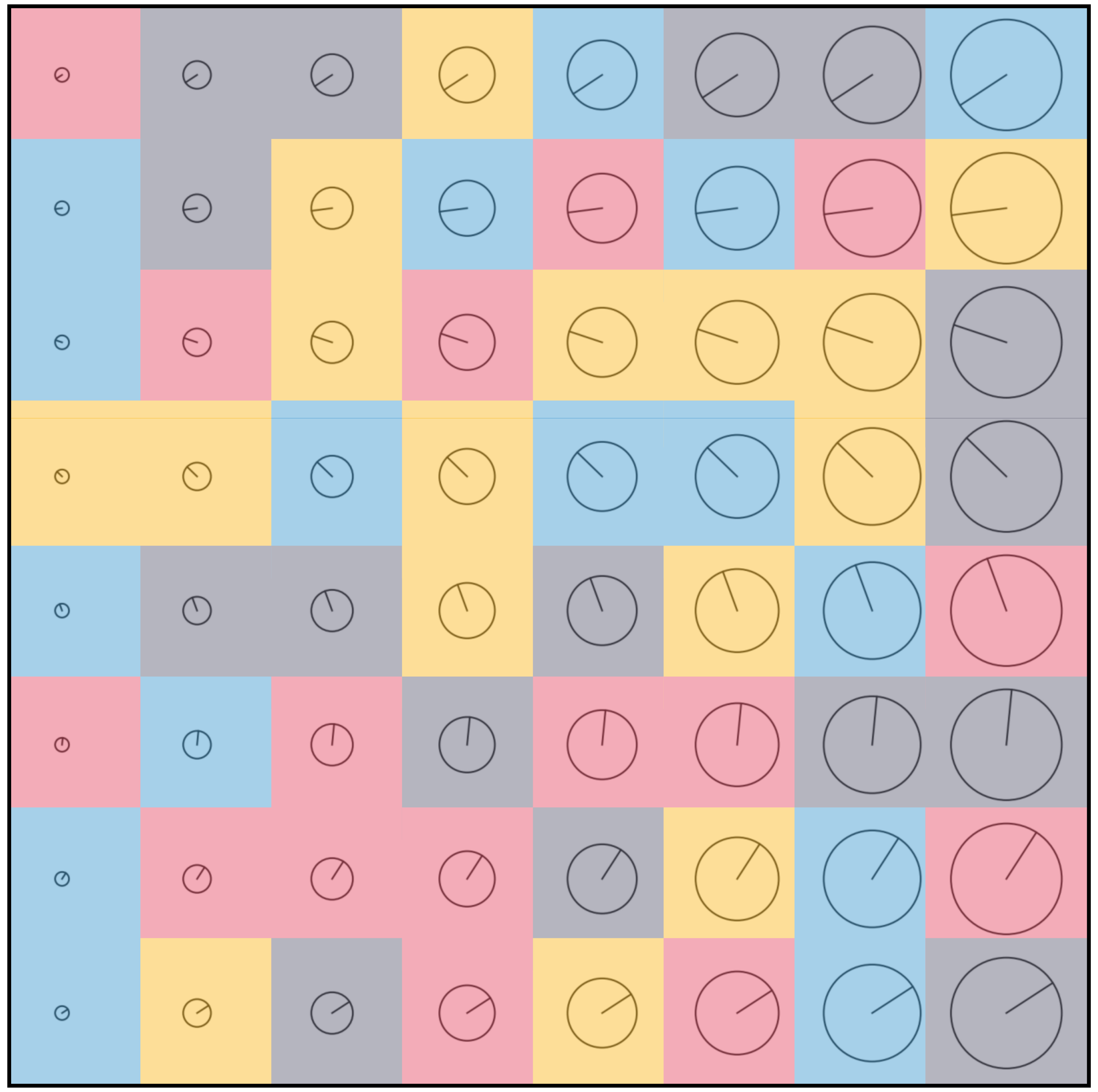
Shepard circle naming



[1] Shepard (1964)

Shepard circle naming

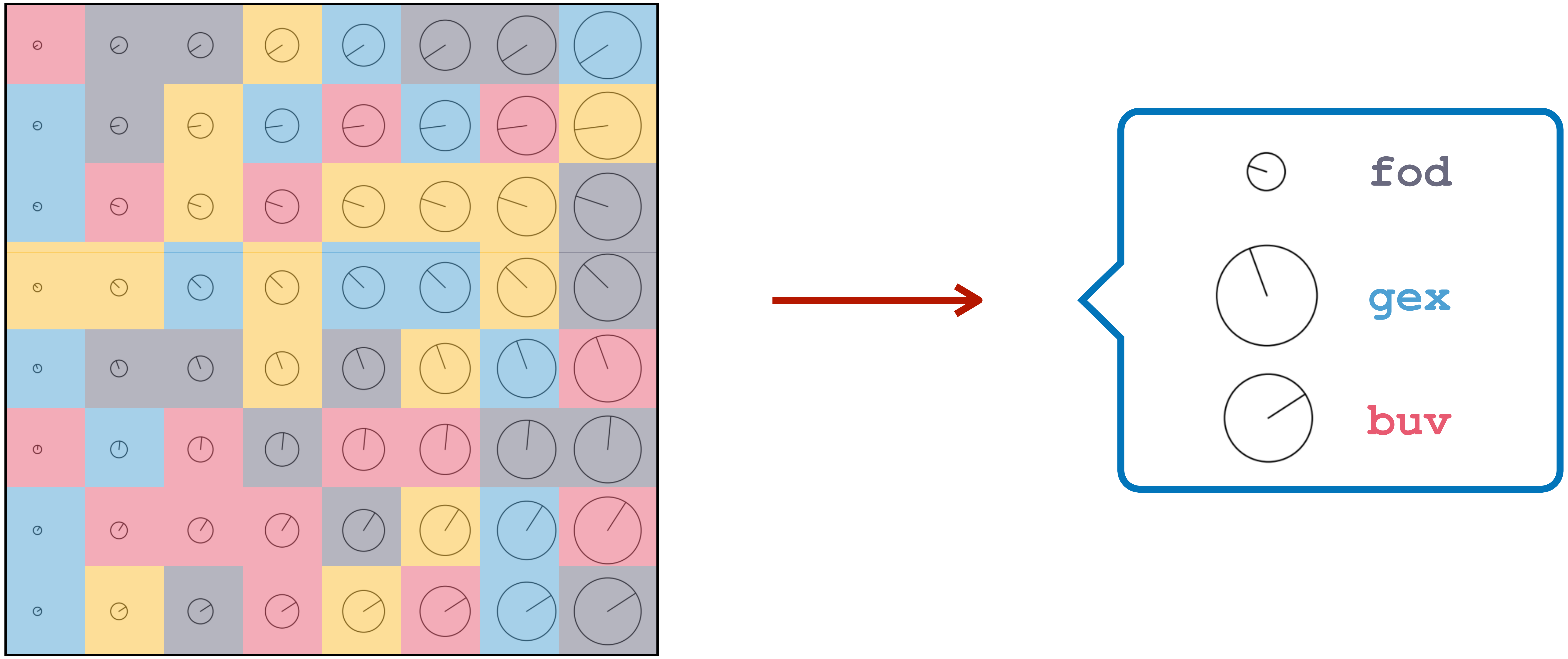
example random system



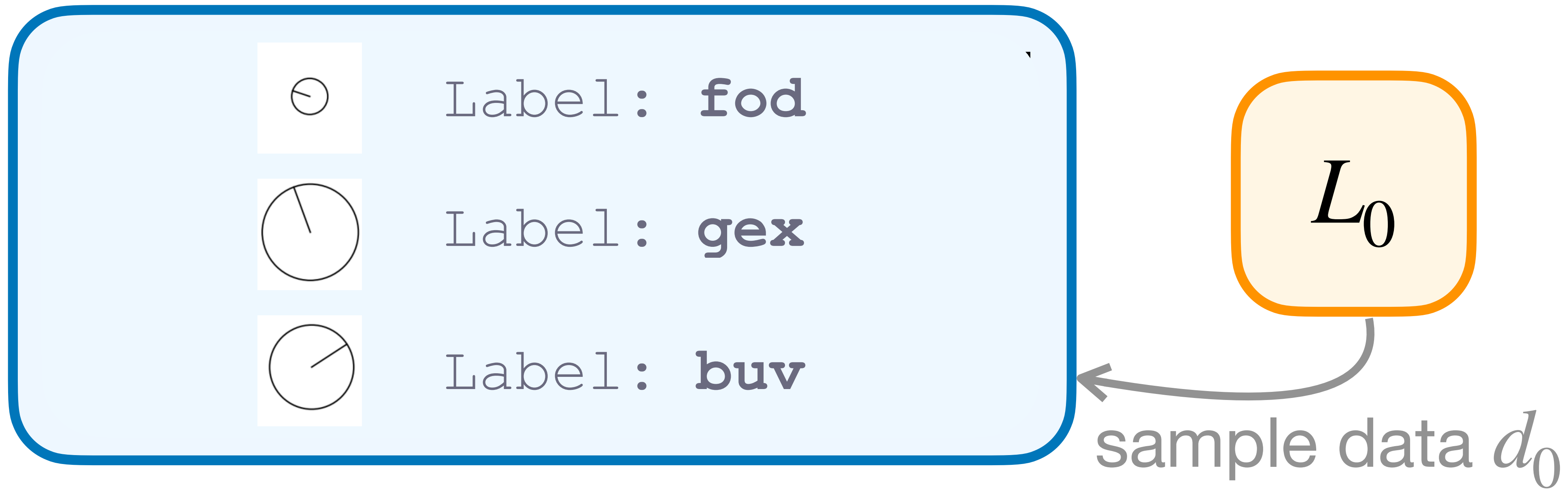
Shepard circle naming

example random system

example (meaning, term) pairs

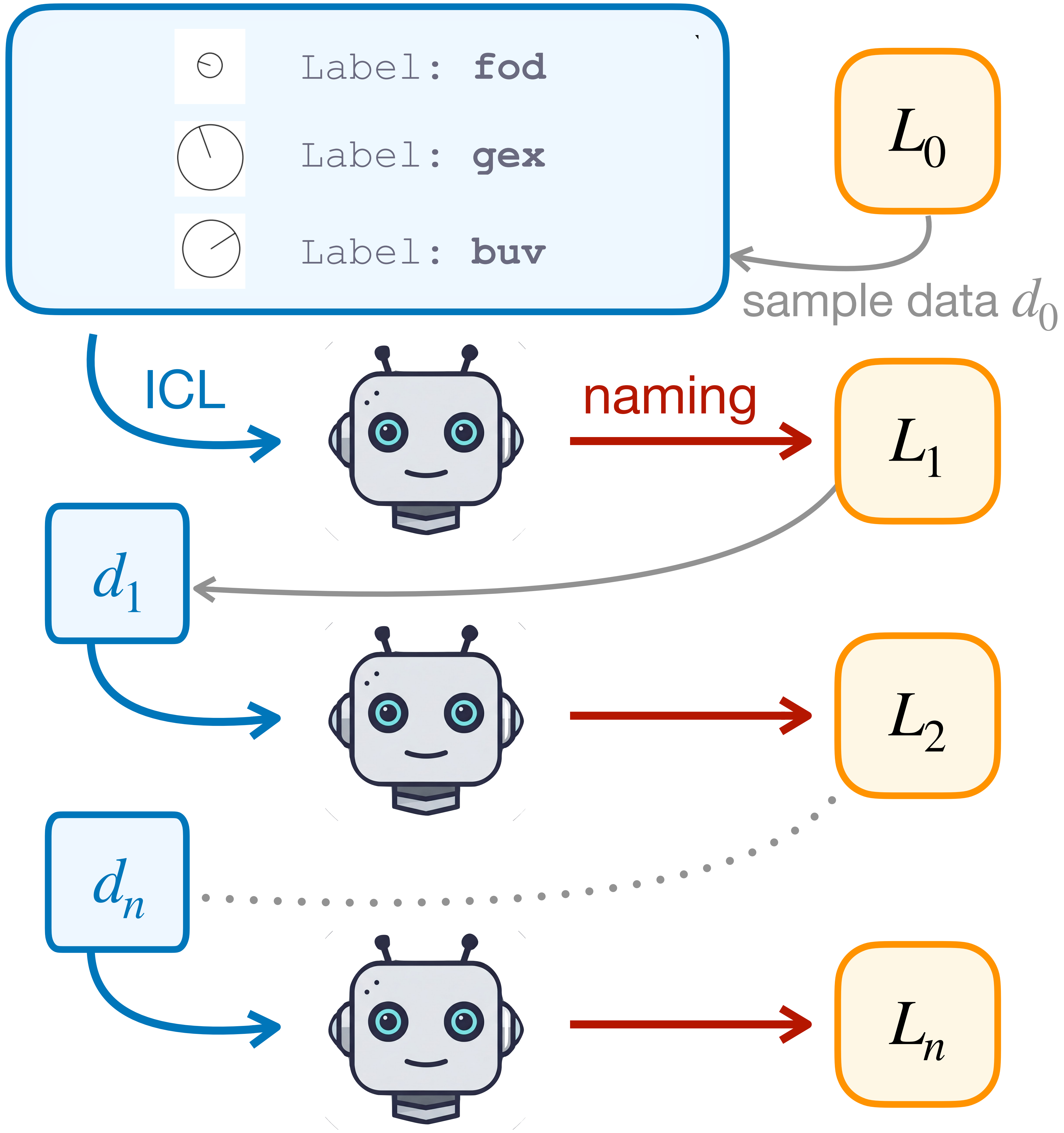


Shepard circle naming

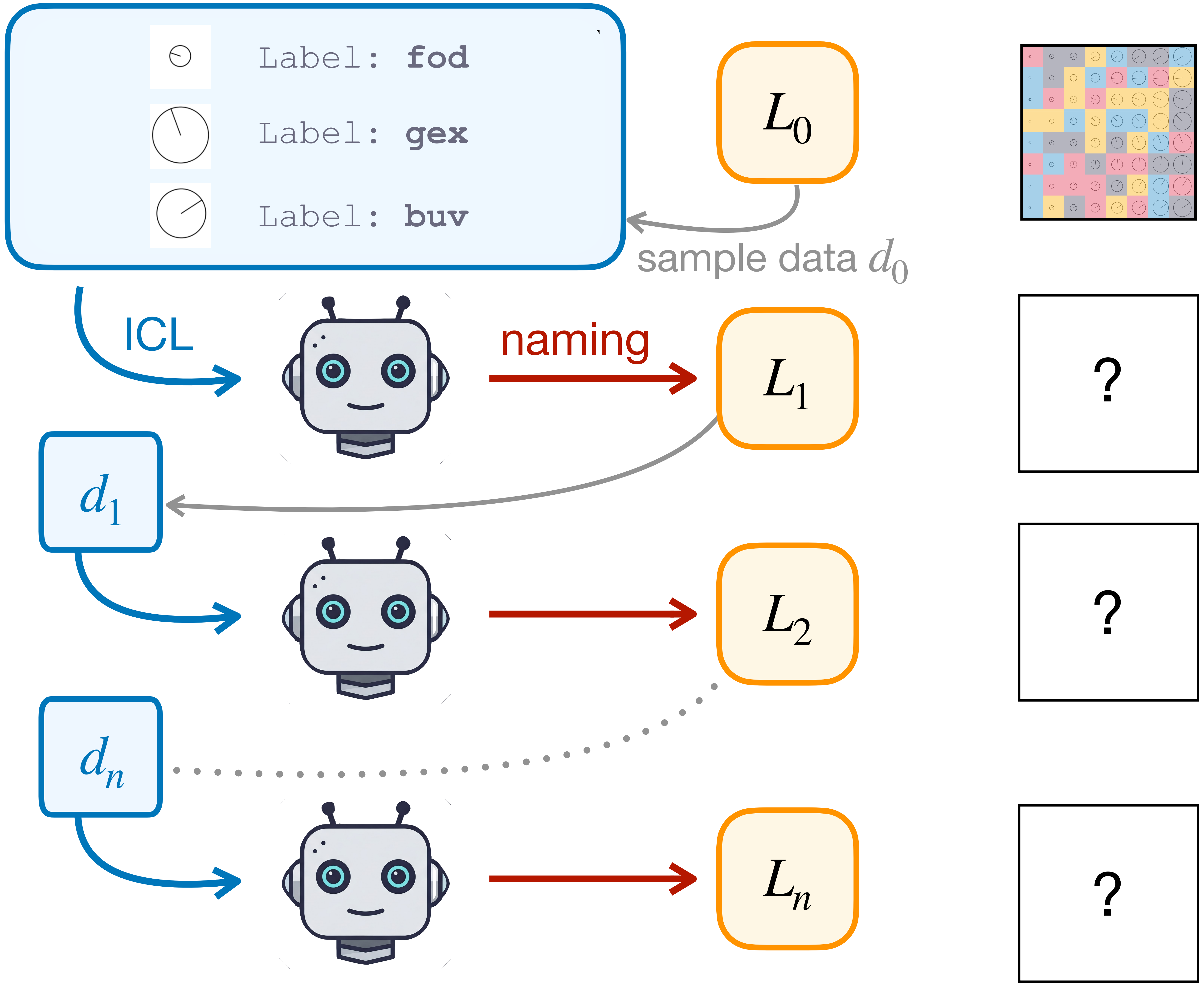


```
Label: fod
Label: gex
Label: buv
```

Shepard circle naming



Shepard circle naming



Label: **fod**

Label: **gex**

Label: **buv**

L_0

sample data d_0

ICL

naming

L_1

d_1

L_2

d_n

L_n

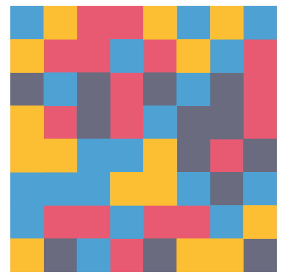
?

?

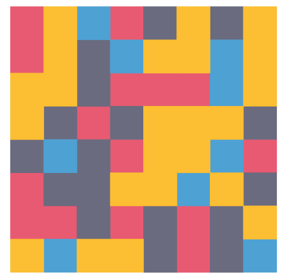
?

Shepard circle naming: initial results from Gemini

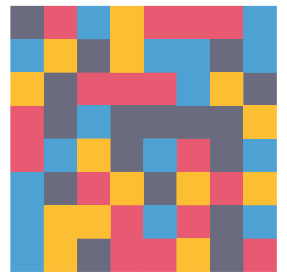
A



B

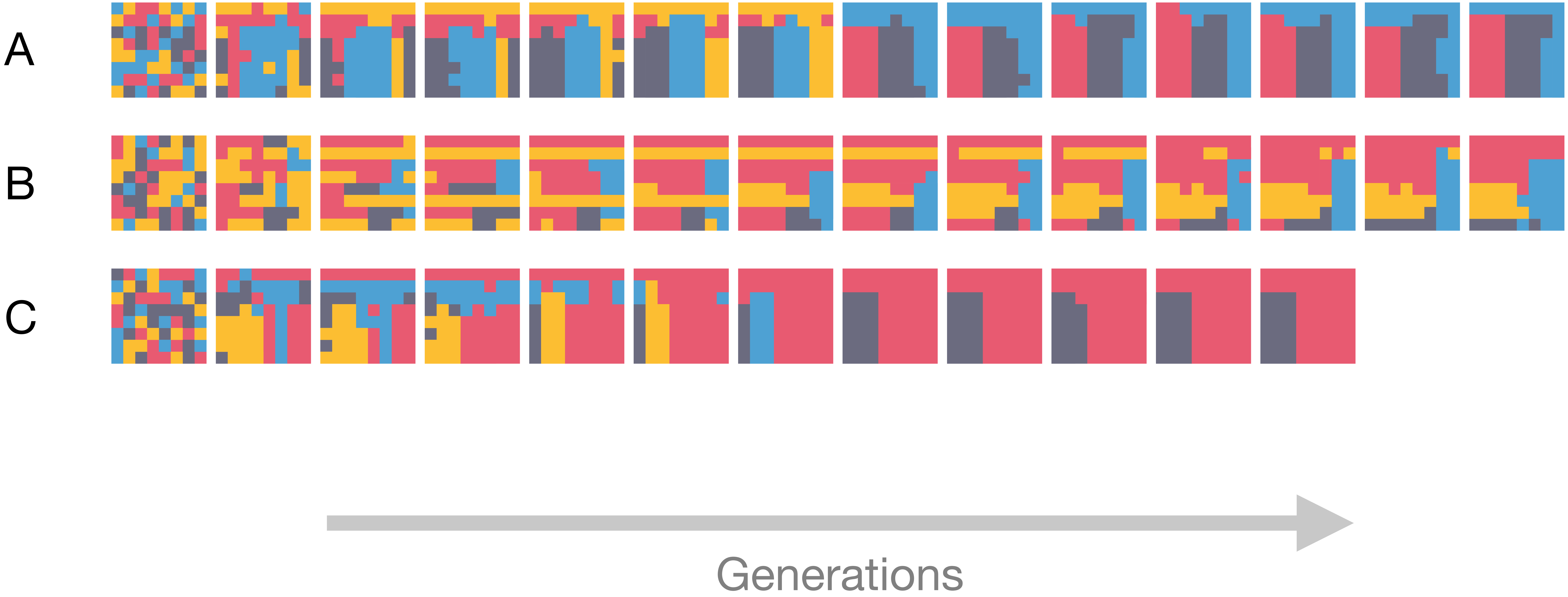


C



Generations

Shepard circle naming: initial results from Gemini



IICLL: example prompt + decoding

IICLL: example prompt + decoding

```
Features: [0.73579176, 0.13100809, 0.20245084] -> Label:
Tovo


Features: [0.0, 0.32875953, 0.29290289] -> Label: Feglu


  ...


Features: [0.0, 0.33817158, 0.21461567] -> Label: Mib
```

‘training’ examples
in context

IICLL: example prompt + decoding

```
Features: [0.73579176, 0.13100809, 0.20245084] -> Label:
Tovo


Features: [0.0, 0.32875953, 0.29290289] -> Label: Feglu


   ...


Features: [0.0, 0.33817158, 0.21461567] -> Label: Mib
```

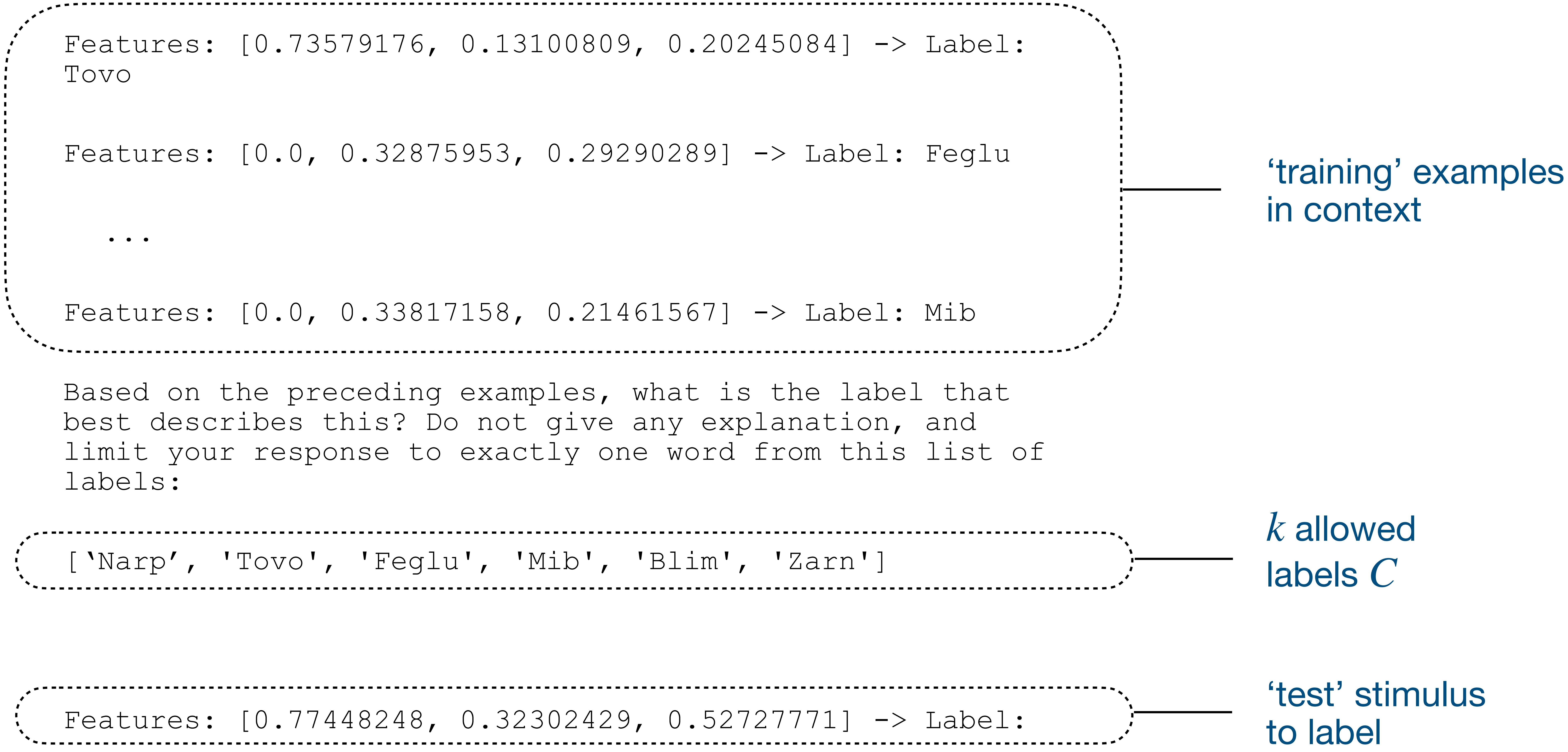
‘training’ examples
in context

```
Based on the preceding examples, what is the label that
best describes this? Do not give any explanation, and
limit your response to exactly one word from this list of
labels:

[‘Narp’, 'Tovo', 'Feglu', 'Mib', 'Blim', 'Zarn']
```

k allowed
labels C

IICLL: example prompt + decoding



```
Features: [0.73579176, 0.13100809, 0.20245084] -> Label:
Tovo


Features: [0.0, 0.32875953, 0.29290289] -> Label: Feglu


   ...


Features: [0.0, 0.33817158, 0.21461567] -> Label: Mib
```

‘training’ examples in context

```
Based on the preceding examples, what is the label that
best describes this? Do not give any explanation, and
limit your response to exactly one word from this list of
labels:

[‘Narp’, 'Tovo', 'Feglu', 'Mib', 'Blim', 'Zarn']
```

k allowed labels C

```
Features: [0.77448248, 0.32302429, 0.52727771] -> Label:
```

‘test’ stimulus to label

IICLL: example prompt + decoding

```
Features: [0.73579176, 0.13100809, 0.20245084] -> Label:
Tovo

Features: [0.0, 0.32875953, 0.29290289] -> Label: Feglu

   ...

Features: [0.0, 0.33817158, 0.21461567] -> Label: Mib
```

‘training’ examples in context

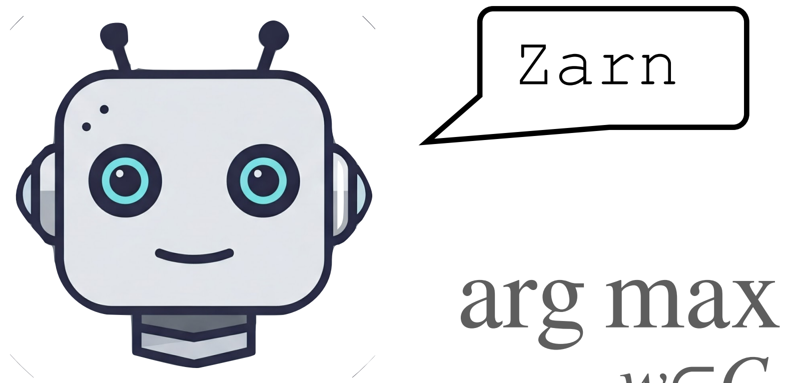
```
Based on the preceding examples, what is the label that
best describes this? Do not give any explanation, and
limit your response to exactly one word from this list of
labels:
```

```
[‘Narp’, 'Tovo', 'Feglu', 'Mib', 'Blim', 'Zarn']
```

k allowed labels C

```
Features: [0.77448248, 0.32302429, 0.52727771] -> Label:
```

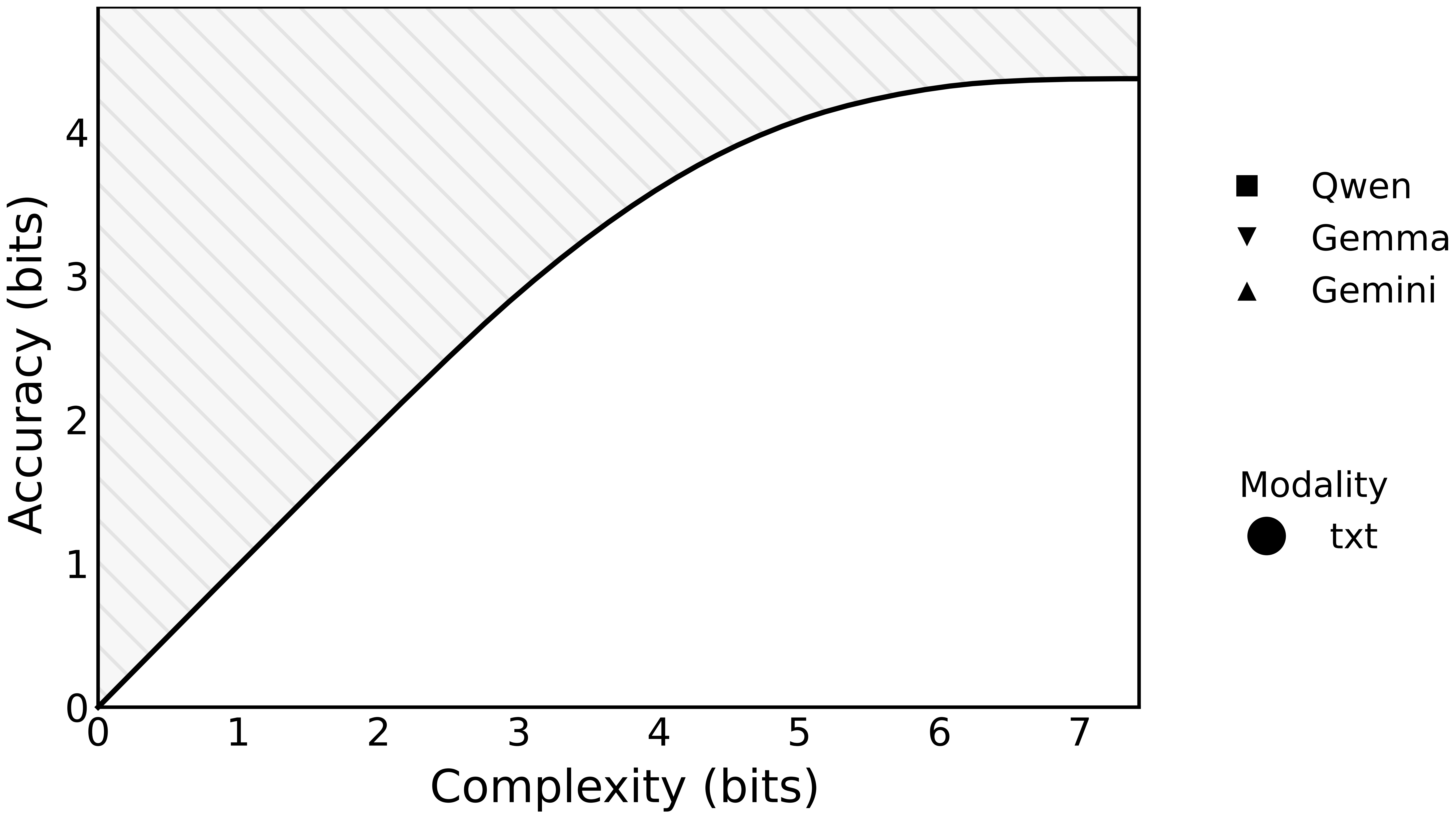
‘test’ stimulus to label



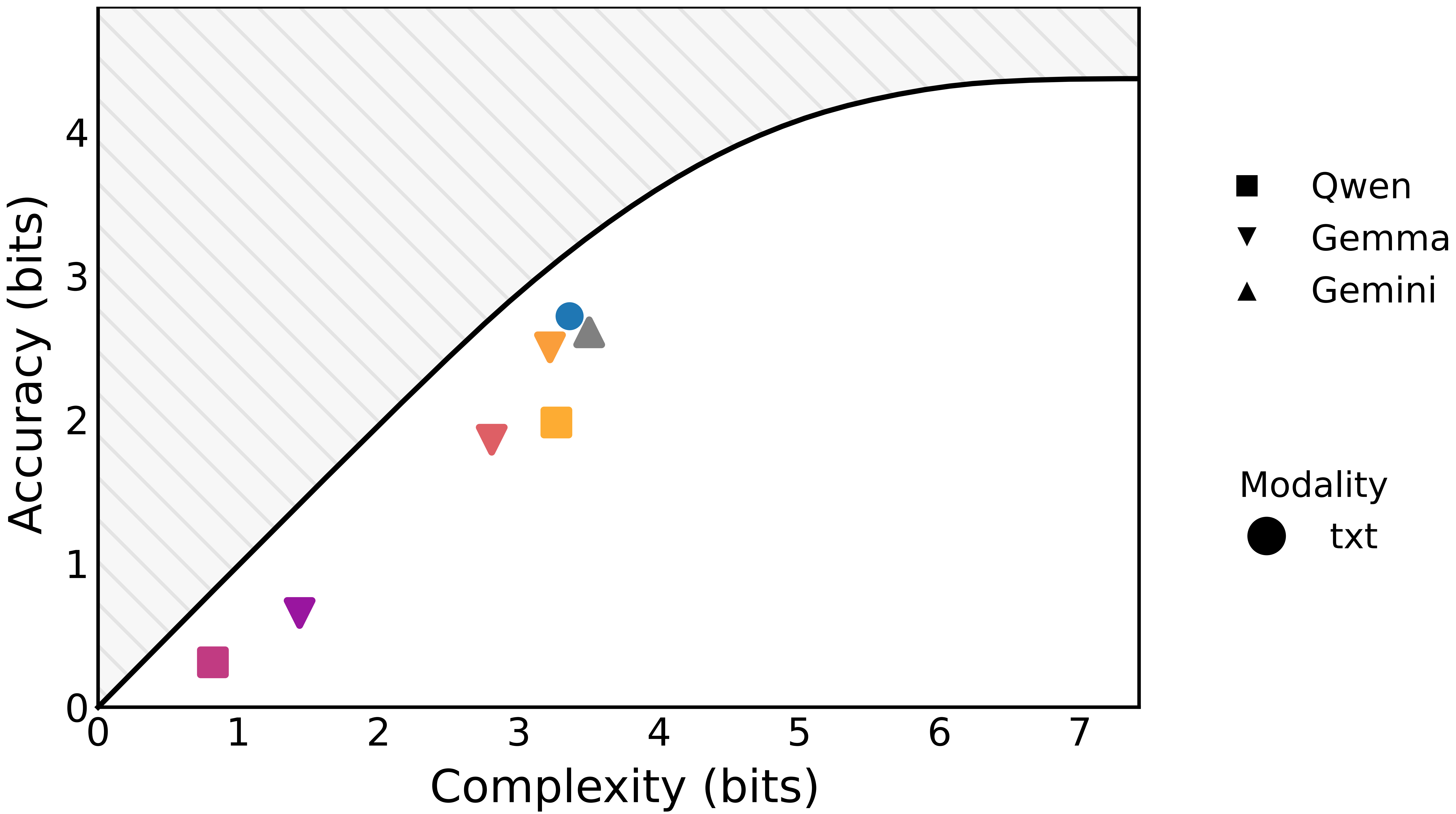
$$\arg\max_{w \in C} p_{\mathrm{LM}}(w_t \mid w_{<t})$$

greedy logprob-based decoding

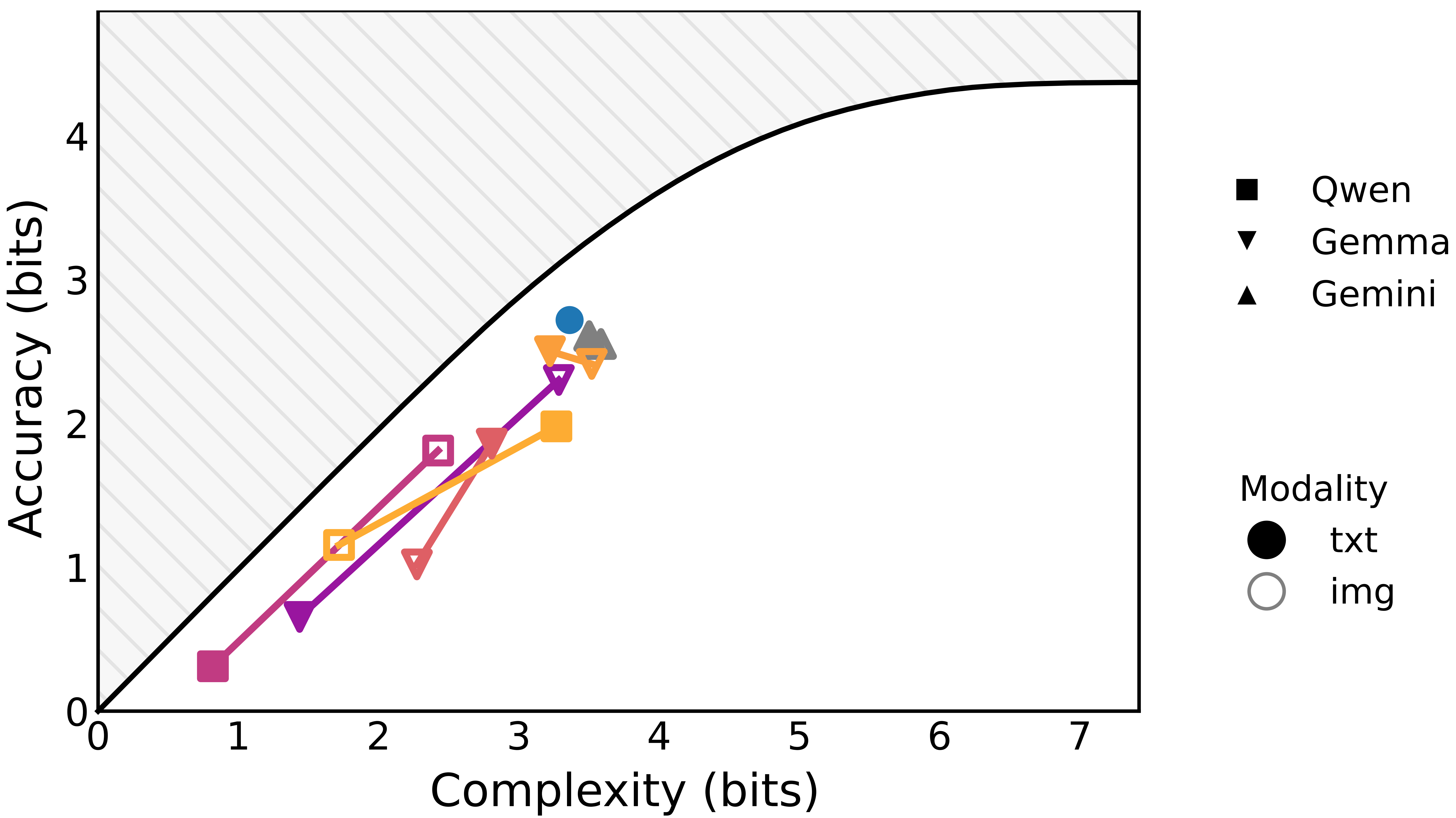
Study 1 Results: vLLM systems’ efficiency tradeoffs



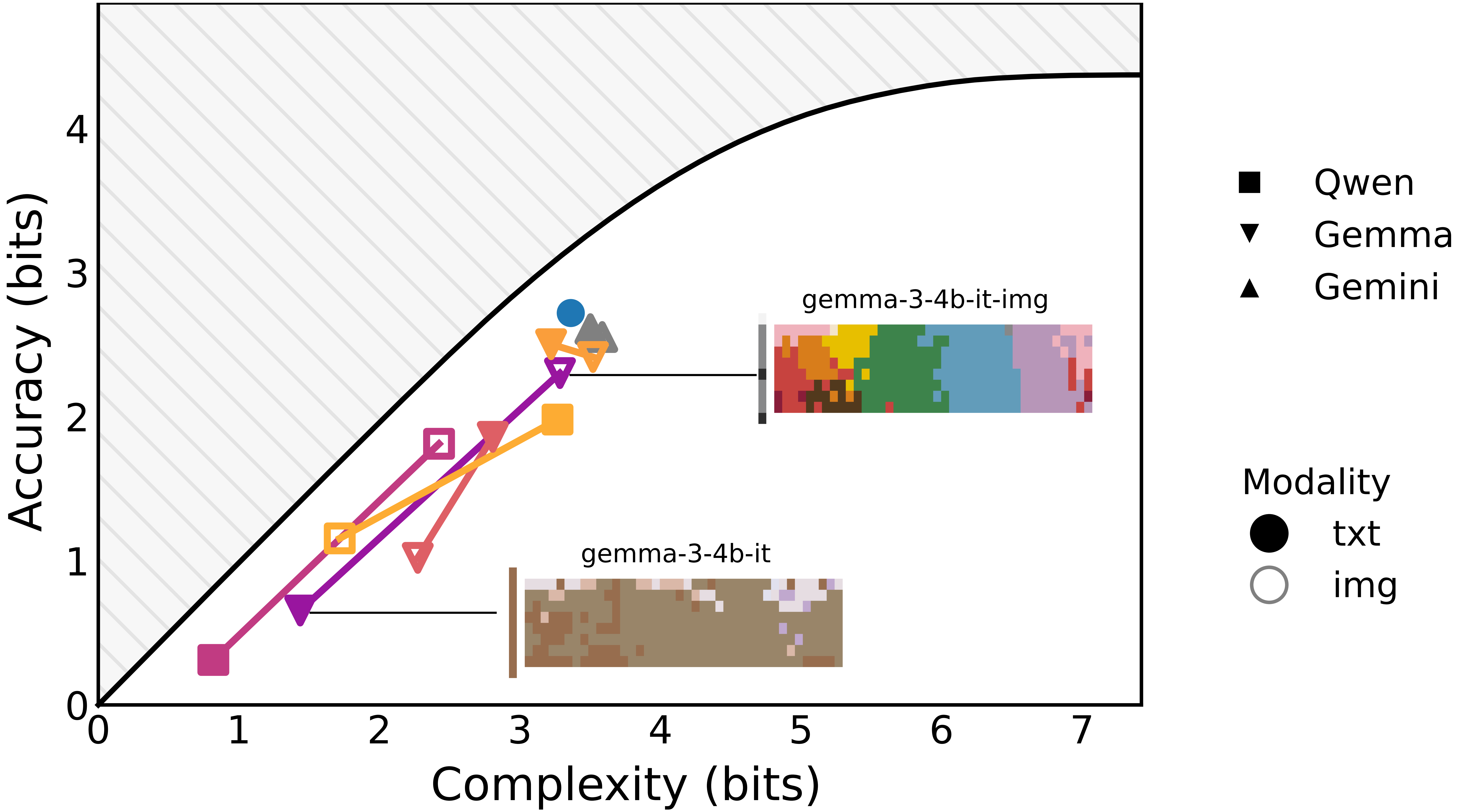
Study 1 Results: vLLM systems' efficiency tradeoffs



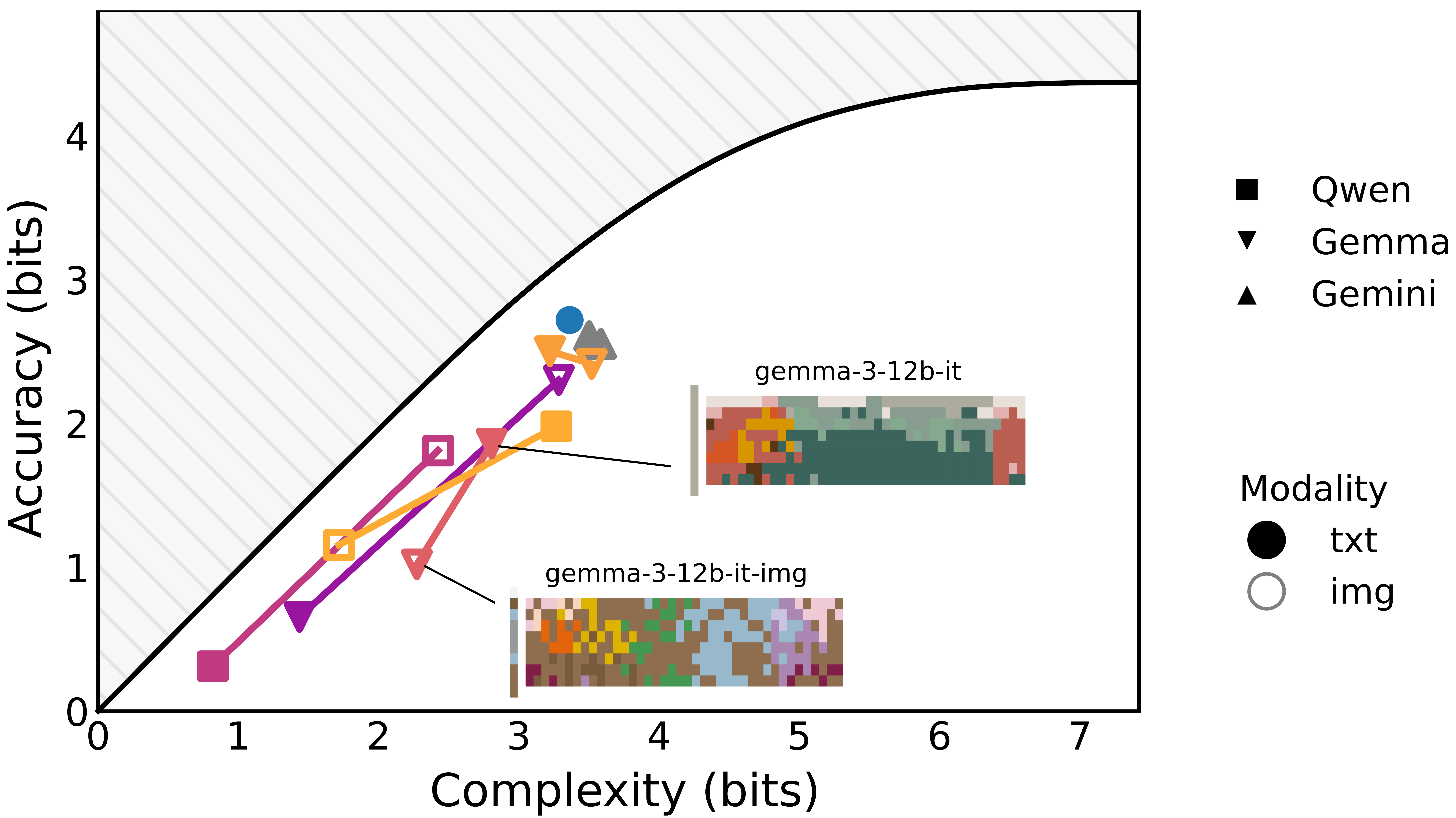
Study 1 Results: vLLM systems' efficiency tradeoffs



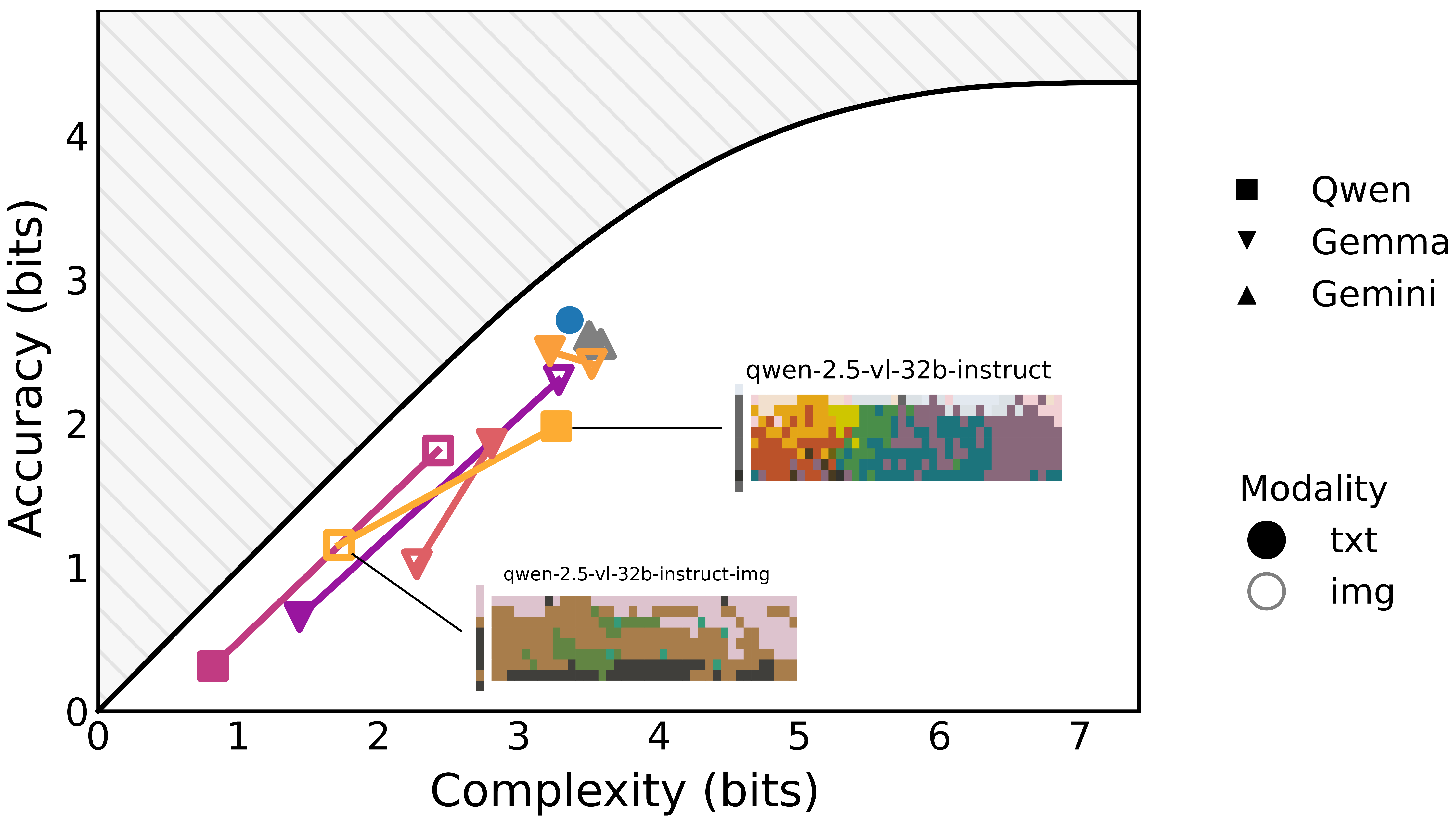
Study 1 Results: vLLM systems’ efficiency tradeoffs



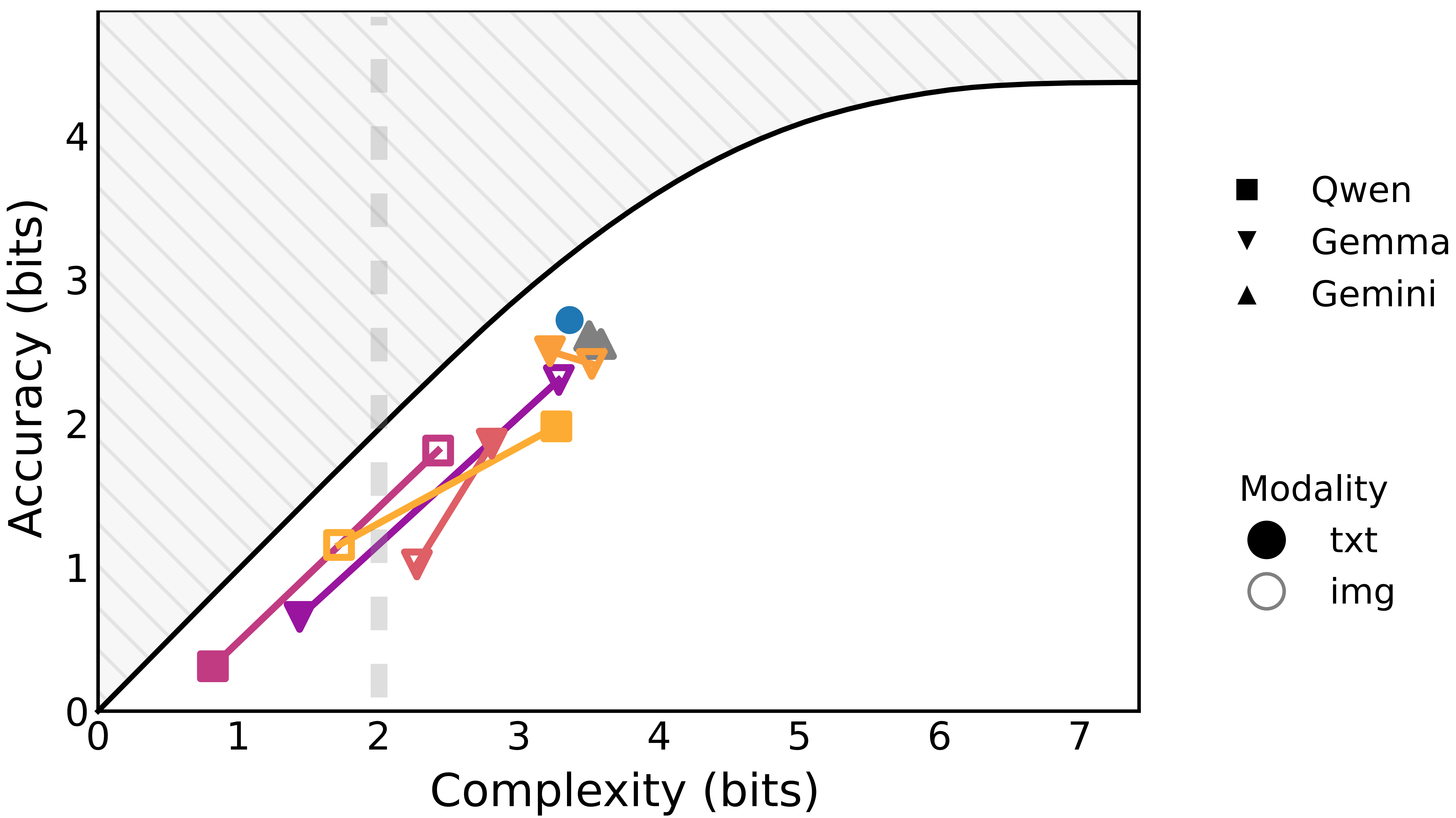
Study 1 Results: vLLM systems' efficiency tradeoffs



Study 1 Results: vLLM systems’ efficiency tradeoffs

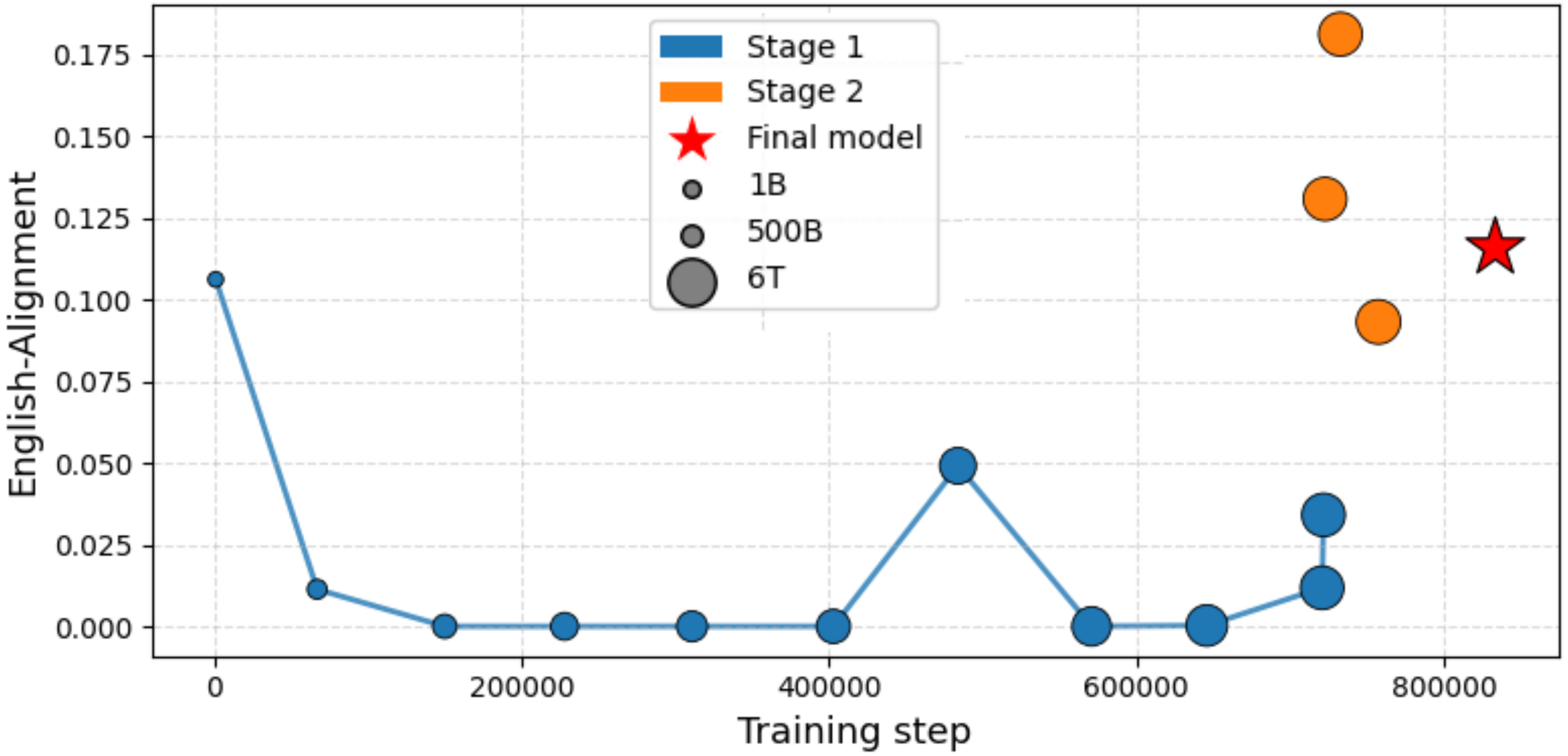


Study 1 Results: vLLM systems' efficiency tradeoffs

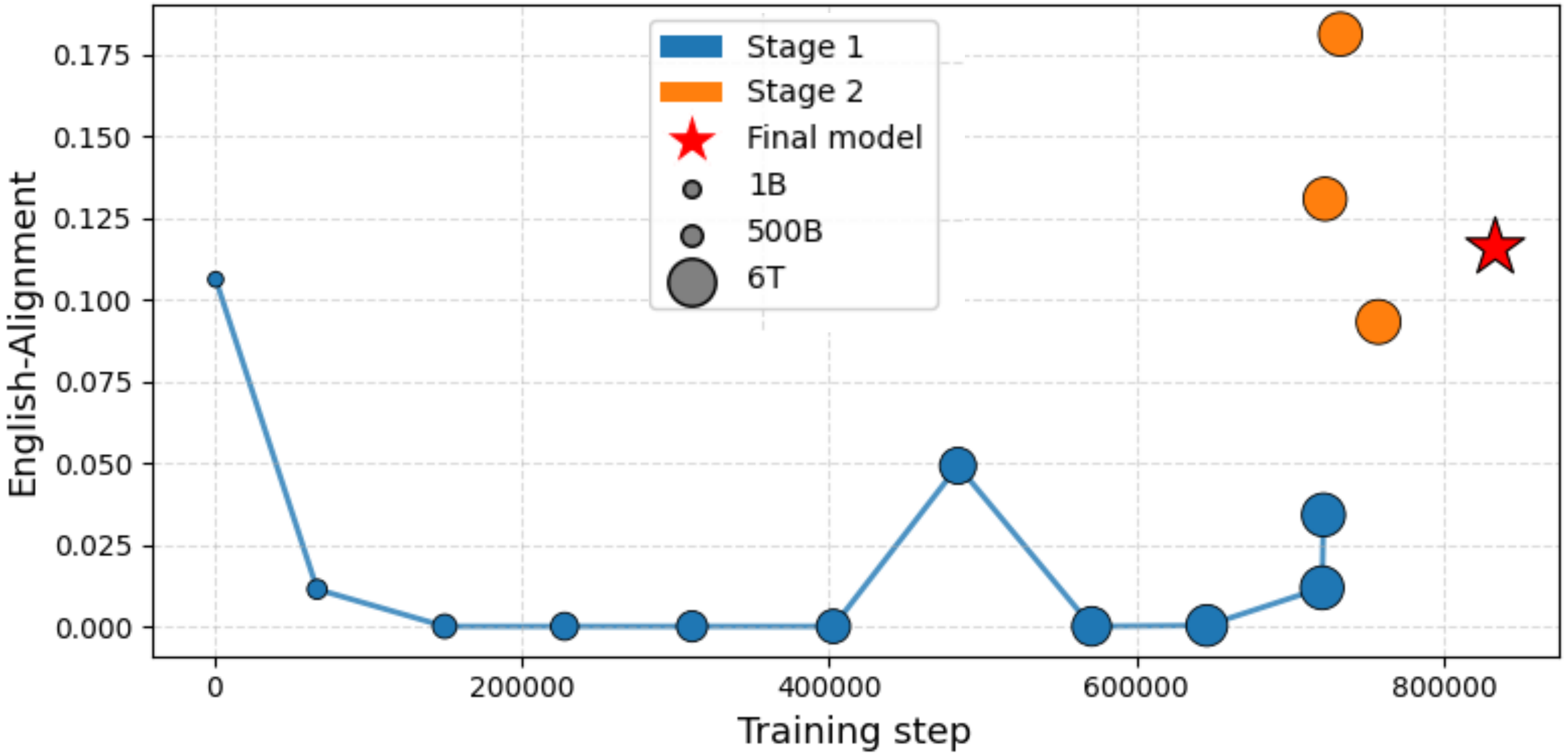


Study 1 Results: Olmo alignment over training

Study 1 Results: Olmo alignment over training

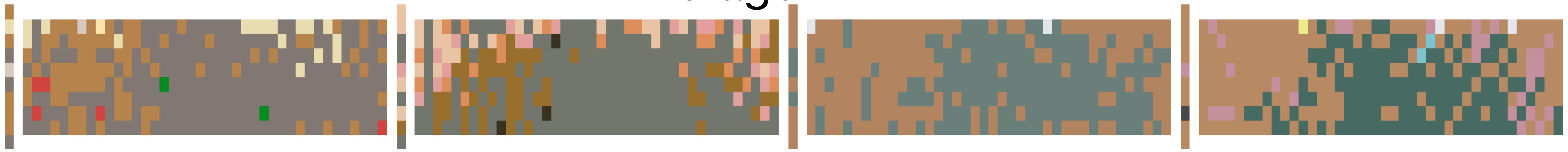


Study 1 Results: Olmo alignment over training



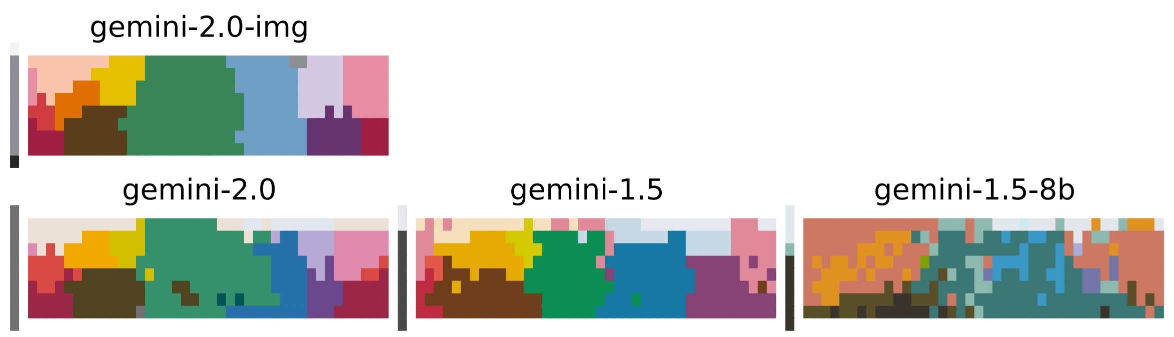
stage 1

stage 2

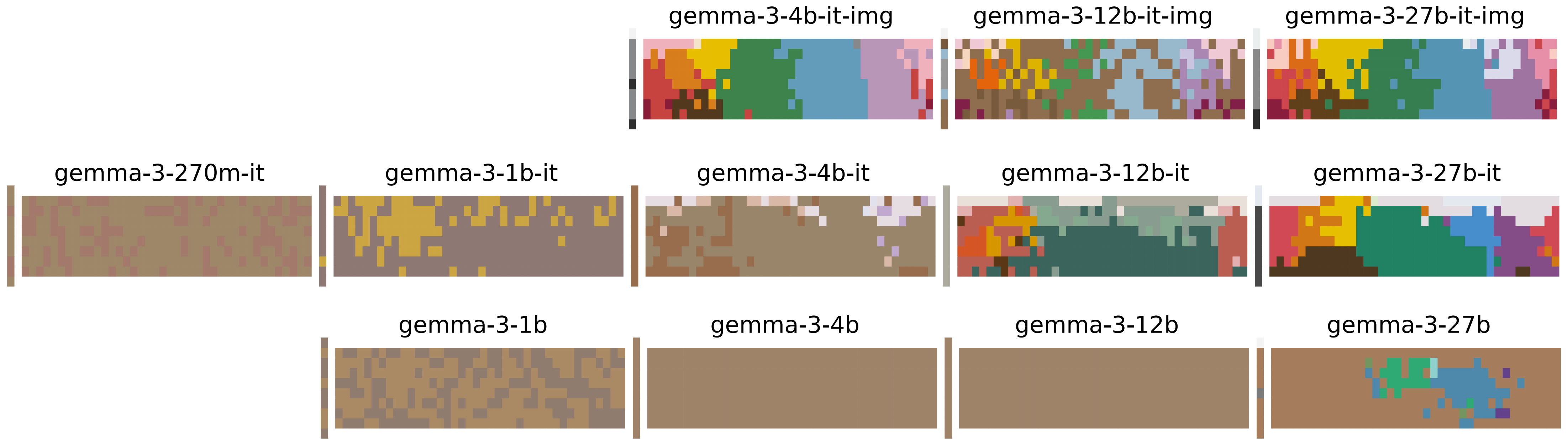


Study 1 Results: LLM systems (all)

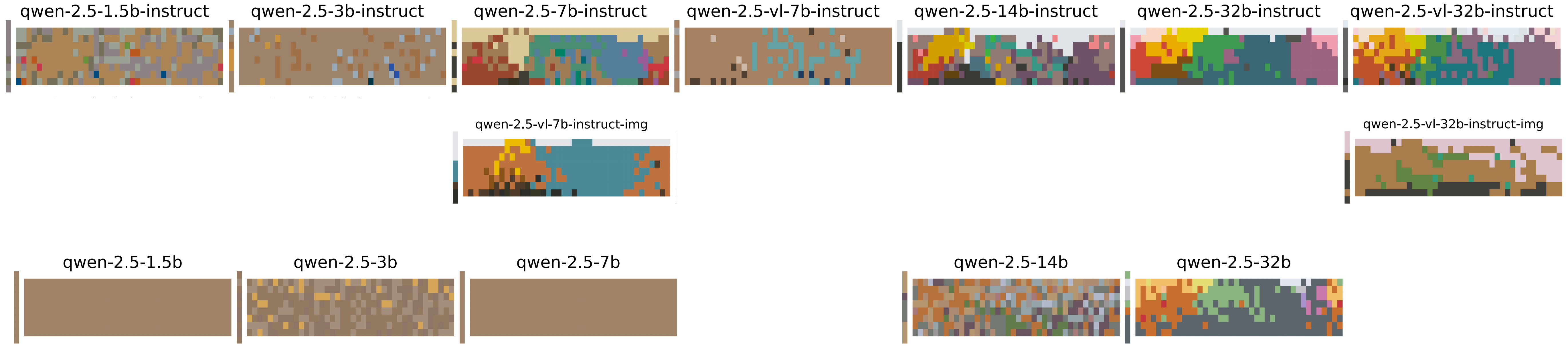
Gemini



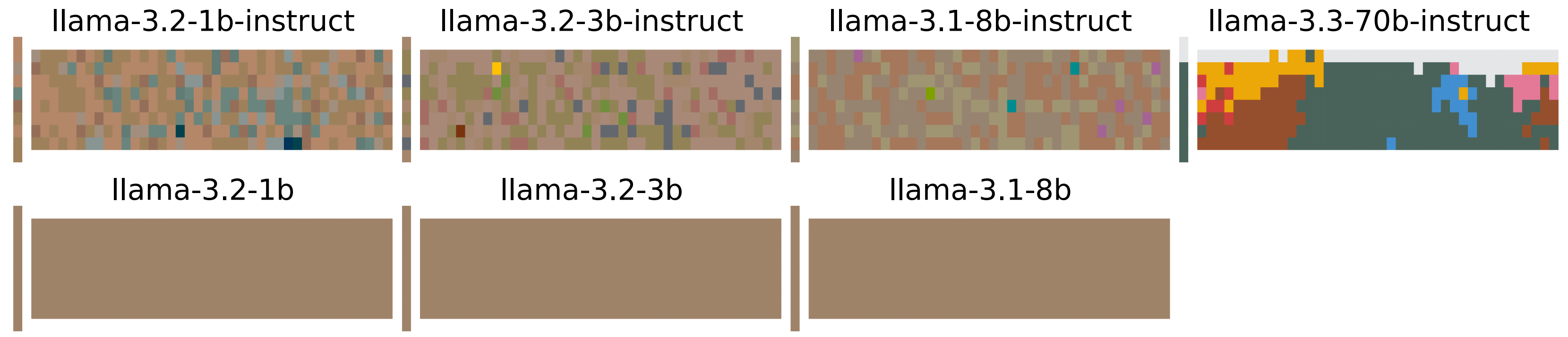
Gemma



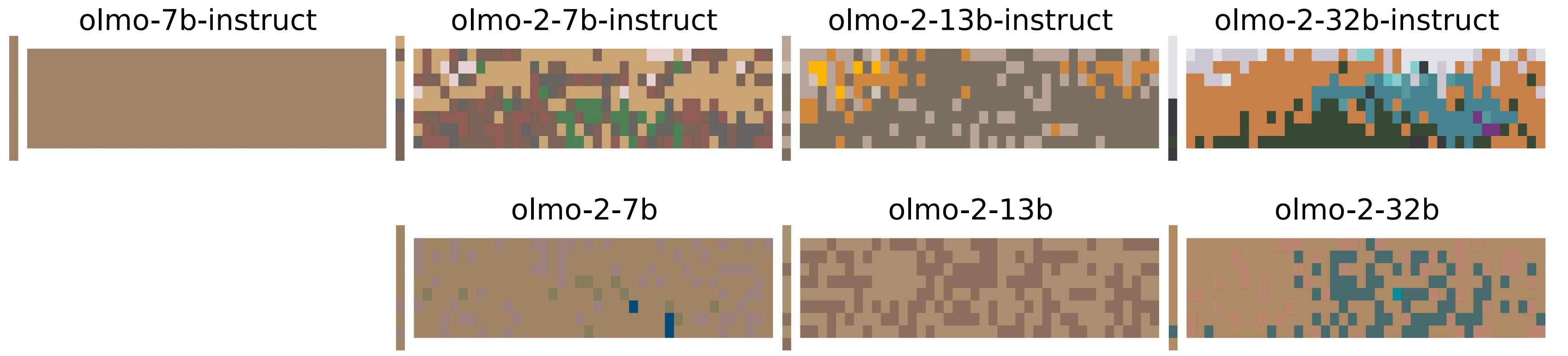
Qwen



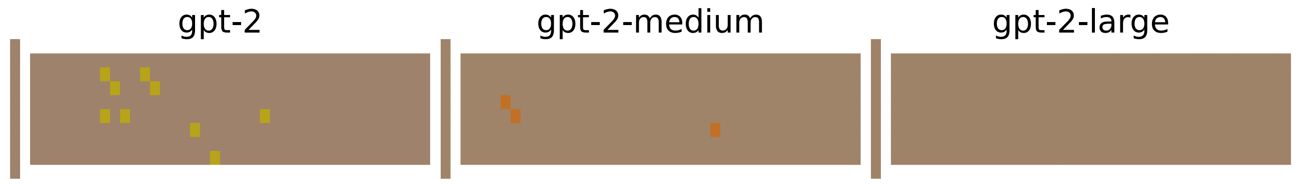
Llama



Olmo



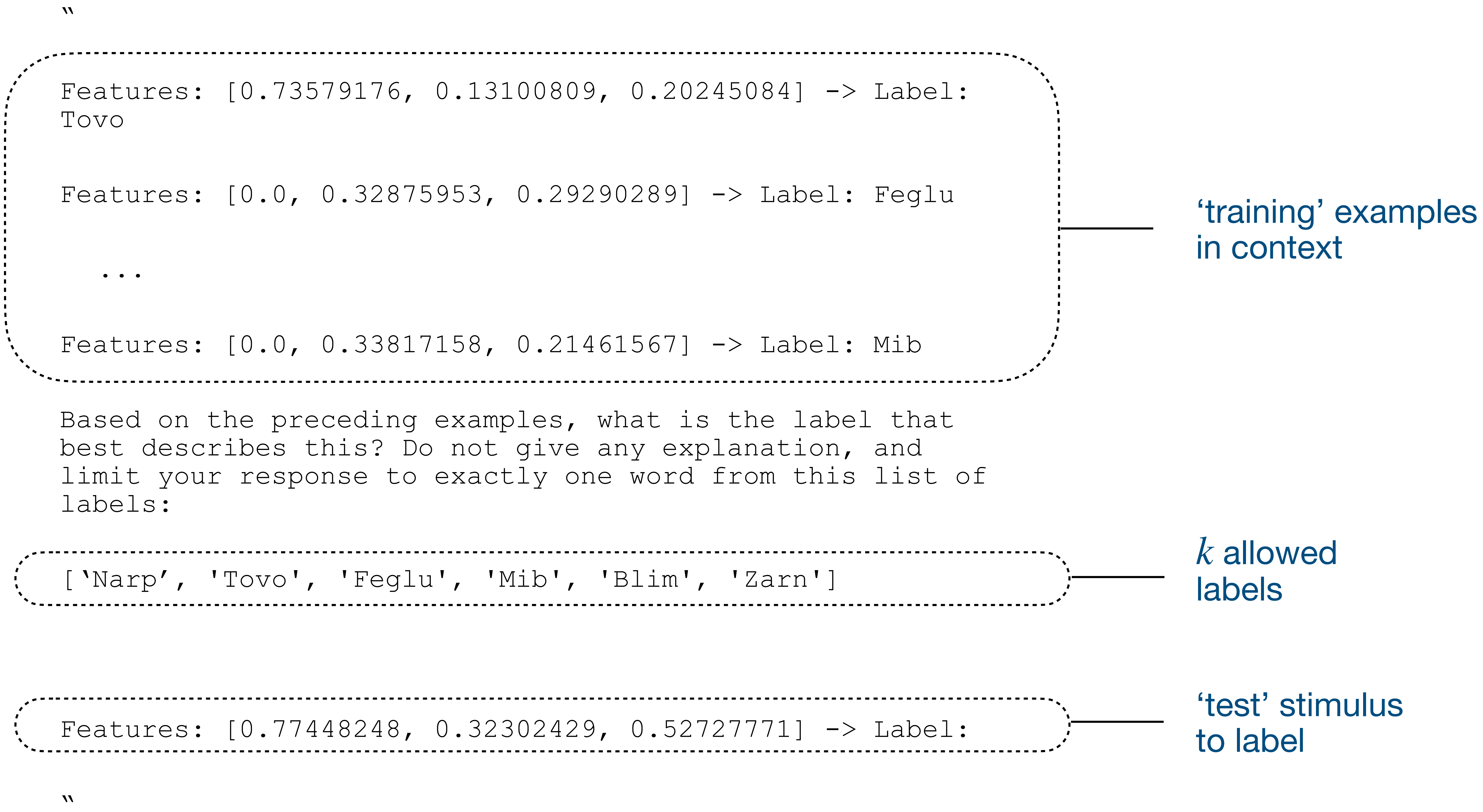
GPT-2



Study 1: Example prompt

```
“What color is this [81.35, 9.68, -11.26]? You may
only use one of the following allowed labels:
['Red', 'Blue', 'Yellow', 'Green', 'Orange',
'Purple', 'Pink', 'Brown', 'Black', 'White',
'Gray', 'Peach', 'Lavender', 'Maroon']. Please
provide only a single label from the list just
provided. Do not give any explanation."
```

Study 2: Example prompt



```
"

Features: [0.73579176, 0.13100809, 0.20245084] -> Label:
Tovo


Features: [0.0, 0.32875953, 0.29290289] -> Label: Feglu


   ...


Features: [0.0, 0.33817158, 0.21461567] -> Label: Mib

Based on the preceding examples, what is the label that
best describes this? Do not give any explanation, and
limit your response to exactly one word from this list of
labels:

['Narp', 'Tovo', 'Feglu', 'Mib', 'Blim', 'Zarn']


Features: [0.77448248, 0.32302429, 0.52727771] -> Label:

"
```

‘training’ examples in context

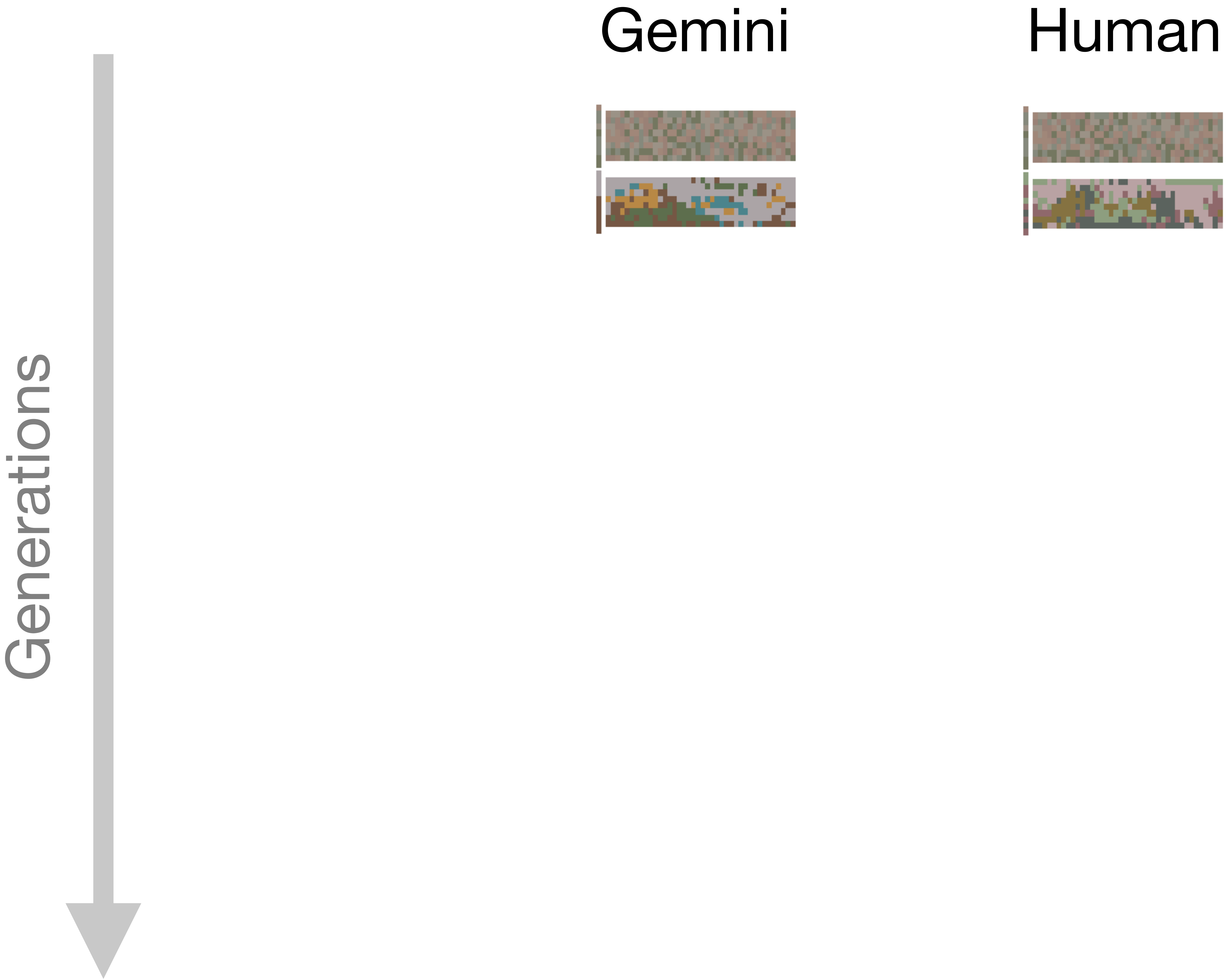
k allowed labels

‘test’ stimulus to label

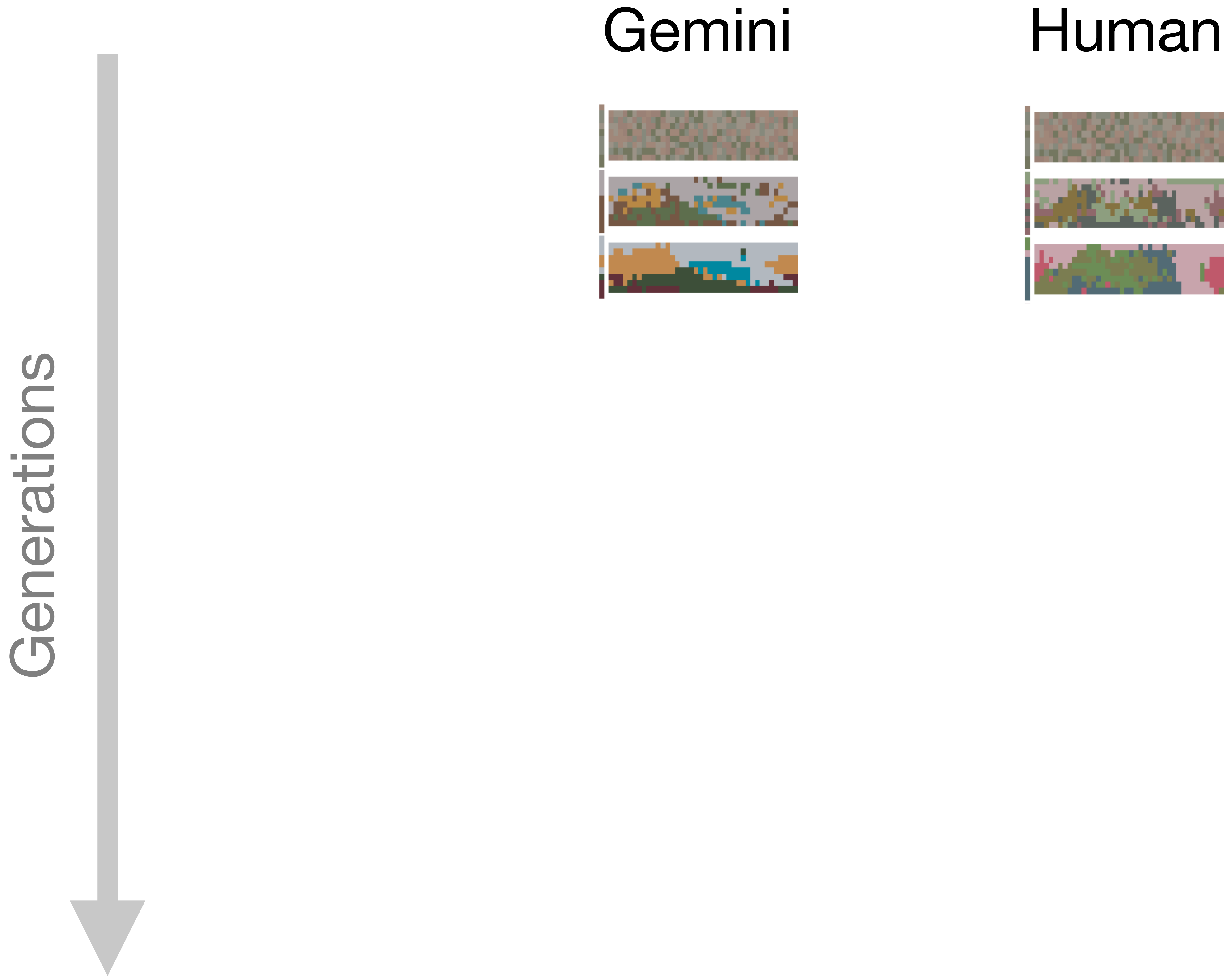
Study 2 (IICLL) Results: LLM color systems over time

Gemini
Human
Generations

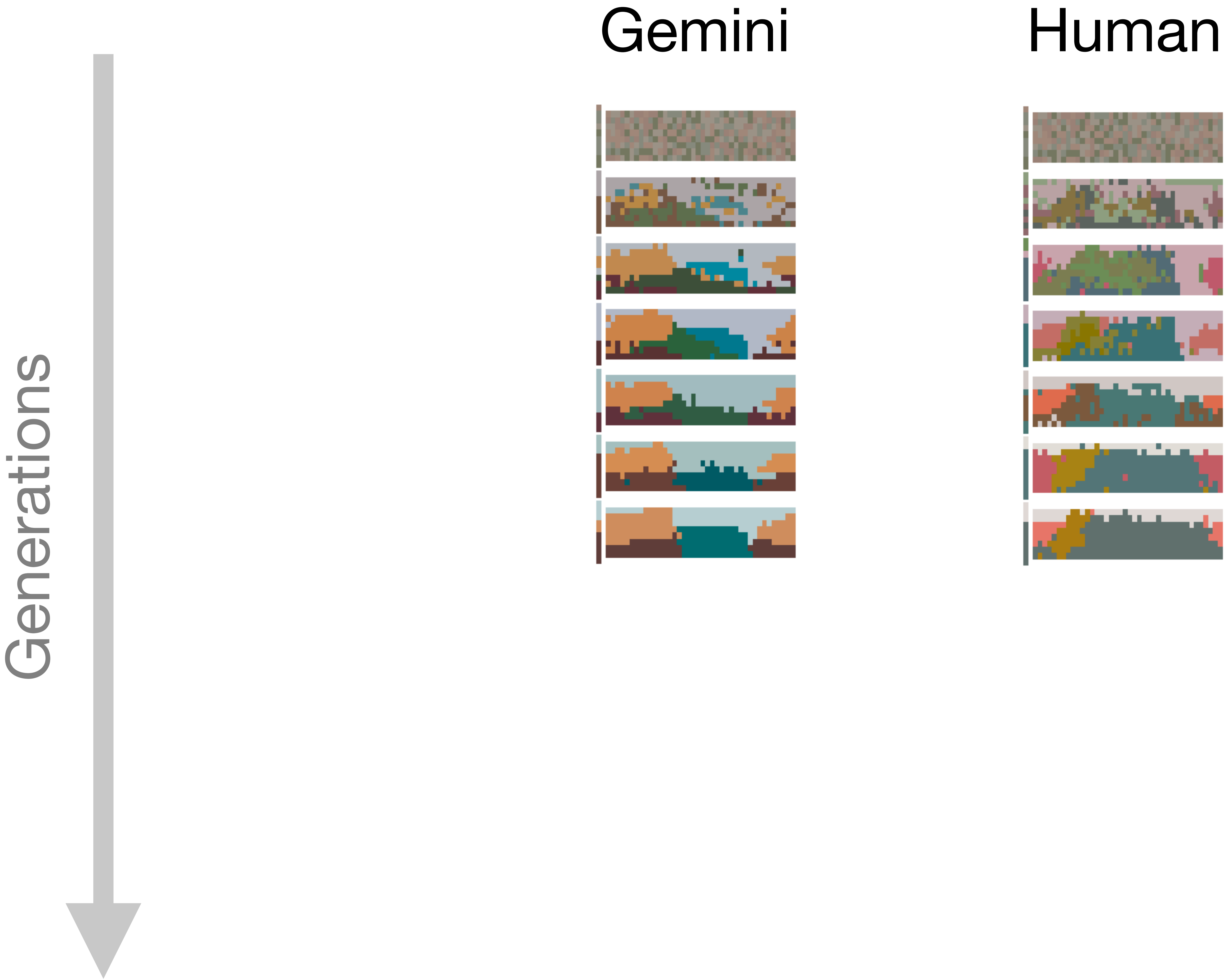
Study 2 (IICLL) Results: LLM color systems over time



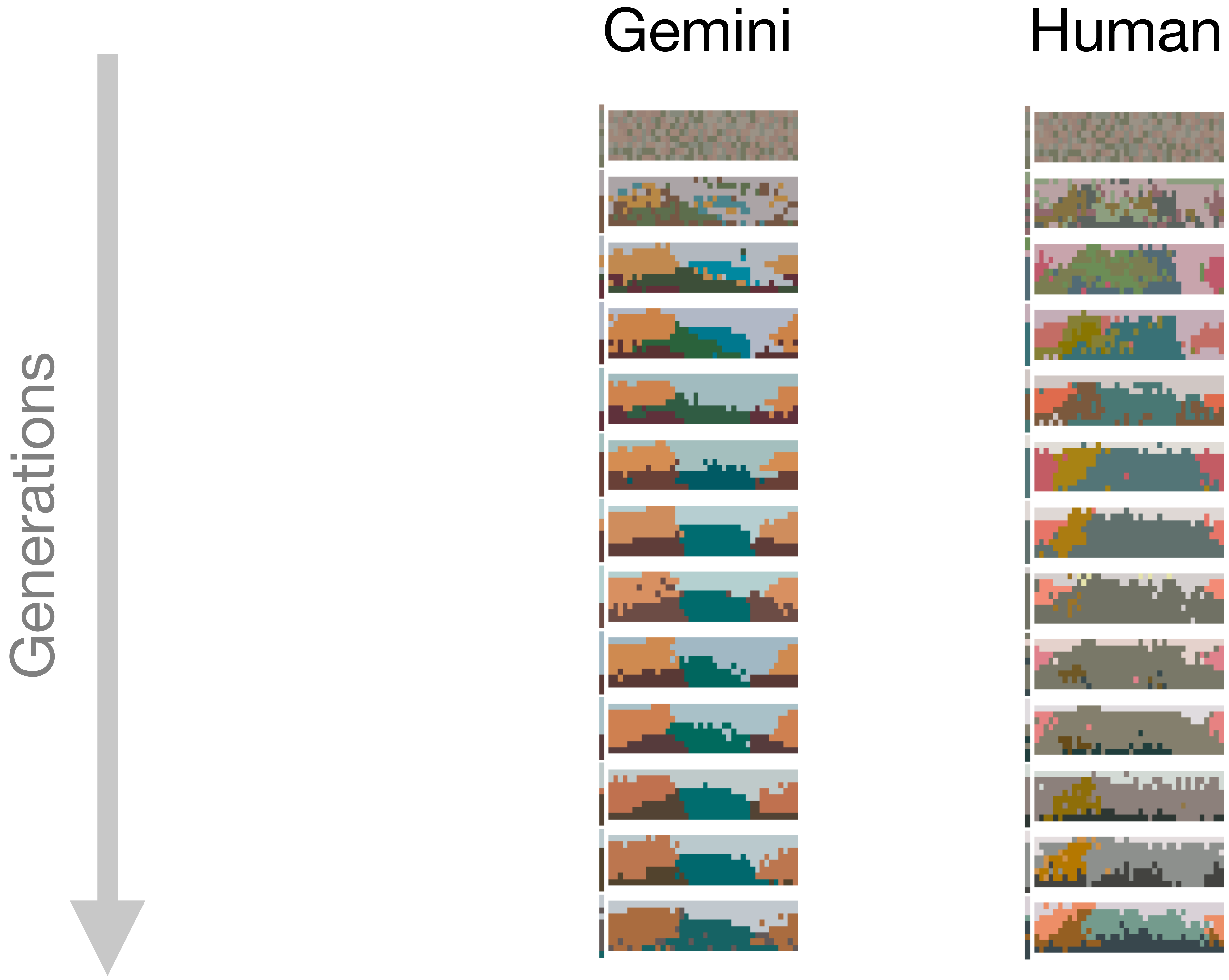
Study 2 (IICLL) Results: LLM color systems over time



Study 2 (IICLL) Results: LLM color systems over time



Study 2 (IICLL) Results: LLM color systems over time



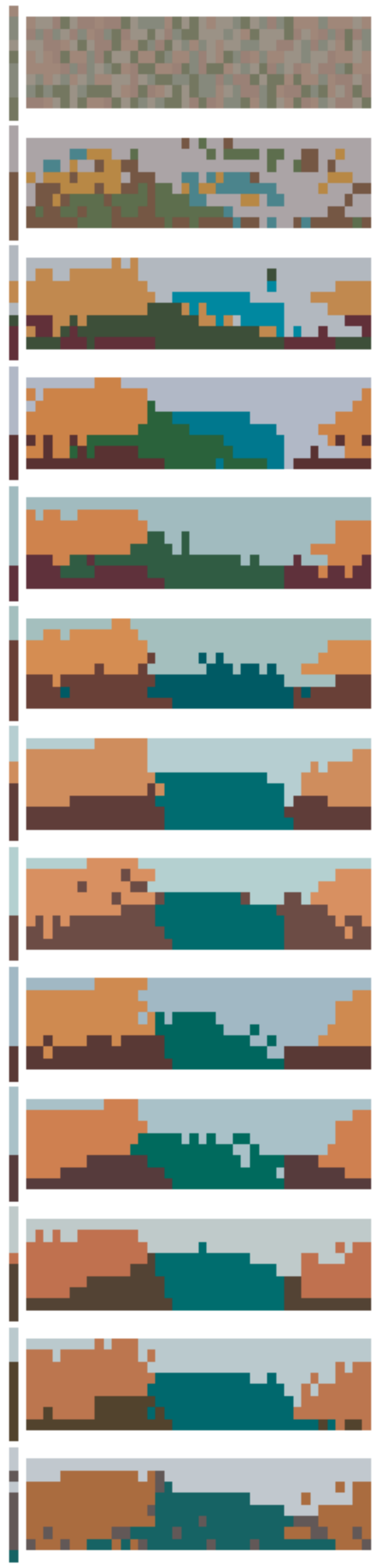
Gemini

Human

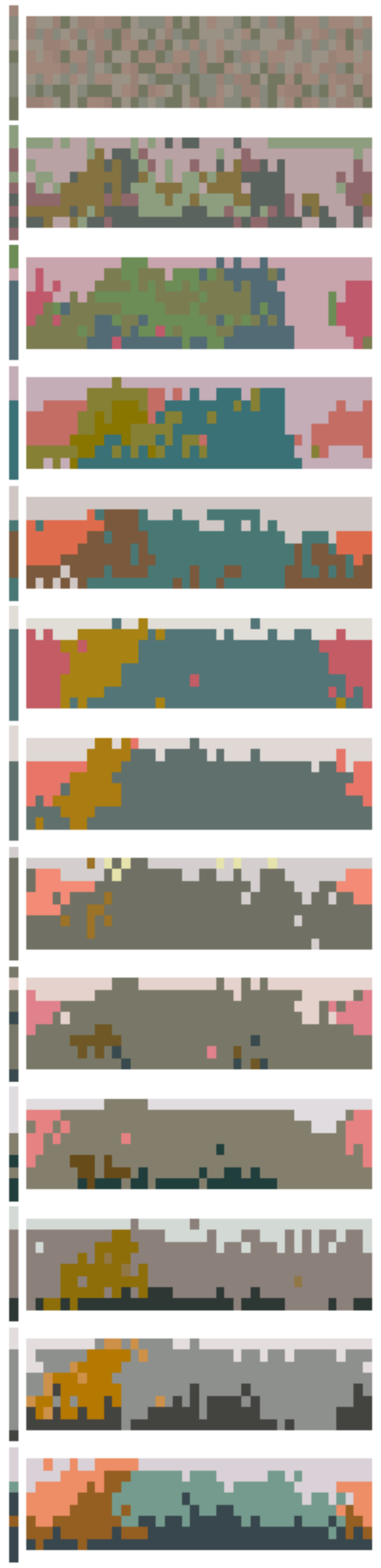
Generations

Study 2 (IICLL) Results: LLM color systems over time

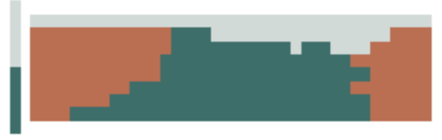
Gemini



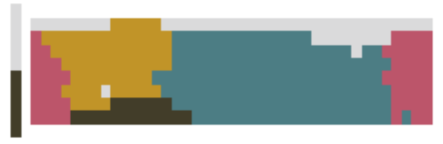
Human



Wobe



Paya



closest WCS languages