Challenges in Evaluating Large Language Models

5th School on Automated Machine Learning
Special focus: Foundation Models

David Salinas. June 2025.

AutoML School 2025

MENU DU JOUR

ENTRÉES (10 MIN)

LLM Evaluations - Introduction

PLAT PRINCIPAL (30 MIN)

LLM Evaluations - Method Overview

Static Evaluations

Dynamic Evaluations

LLM as Judge

DESSERT (~45 MIN)

LLM & AutoML Perspectives

Routing LLMs

Tuning LLM Pipelines

A Case Study: Tuning LLM Judges

Complete Session

~85 min

AutoML School 2025

MENU DU JOUR

ENTRÉES (10 MIN)

LLM Evaluations - Introduction

PLAT PRINCIPAL (30 MIN)

LLM Evaluations - Method Overview

Static Evaluations

Dynamic Evaluations

LLM as Judge

DESSERT (~45 MIN)

LLM & AutoML Perspectives

Routing LLMs

Tuning LLM Pipelines

A Case Study: Tuning LLM Judges

Complete Session

~85 min

Goals of this lecture

AutoML School 2025

MENU DU JOUR

ENTRÉES (10 MIN)

LLM Evaluations - Introduction

PLAT PRINCIPAL (30 MIN)

LLM Evaluations - Method Overview

Static Evaluations

Dynamic Evaluations

LLM as Judge

DESSERT (~45 MIN)

LLM & AutoML Perspectives

Routing LLMs

Tuning LLM Pipelines

A Case Study: Tuning LLM Judges

Complete Session
~85 min

Goals of this lecture

- Know the main families of methods used to evaluate LLM

AutoML School 2025

MENU DU JOUR

ENTRÉES (10 MIN)

LLM Evaluations - Introduction

PLAT PRINCIPAL (30 MIN)

LLM Evaluations - Method Overview

Static Evaluations

Dynamic Evaluations

LLM as Judge

DESSERT (~45 MIN)

LLM & AutoML Perspectives

Routing LLMs

Tuning LLM Pipelines

A Case Study: Tuning LLM Judges

Complete Session
~85 min

Goals of this lecture

- Know the main families of methods used to evaluate LLM
- Understand the key challenges faced when evaluating LLMs

AutoML School 2025

MENU DU JOUR

ENTRÉES (10 MIN)

LLM Evaluations - Introduction

PLAT PRINCIPAL (30 MIN)

LLM Evaluations - Method Overview

Static Evaluations

Dynamic Evaluations

LLM as Judge

DESSERT (~45 MIN)

LLM & AutoML Perspectives

Routing LLMs

Tuning LLM Pipelines

A Case Study: Tuning LLM Judges

Complete Session

~85 min

Goals of this lecture

- Know the main families of methods used to evaluate LLM
- Understand the key challenges faced when evaluating LLMs
- Get ideas about research ideas / low-hanging fruits combining AutoML & LLMs evaluations

LLM

Lifecycle

LLM

Lifecycle

Huber web, book corpus

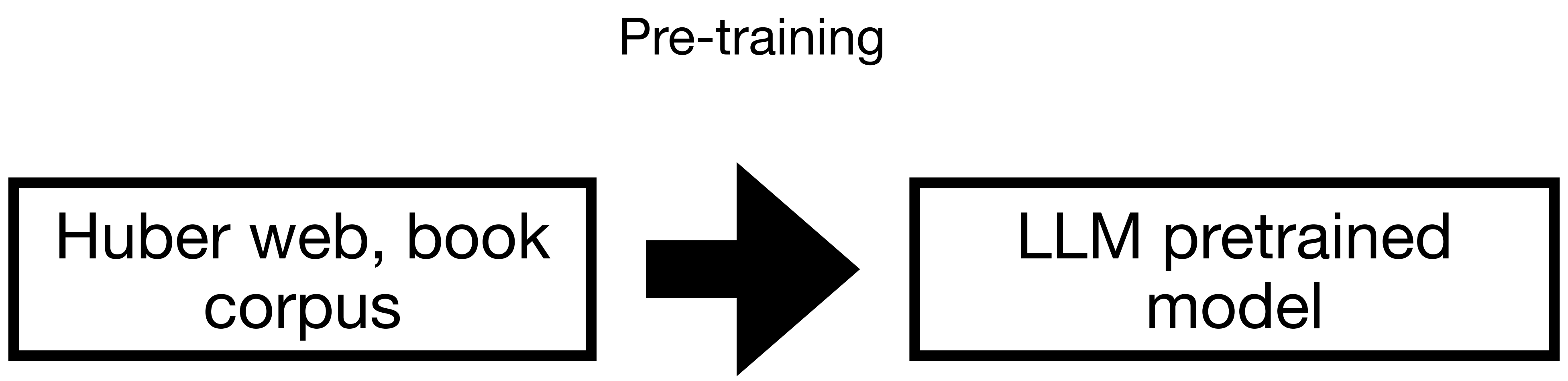
LLM
Lifecycle

Pre-training

Huber web, book corpus

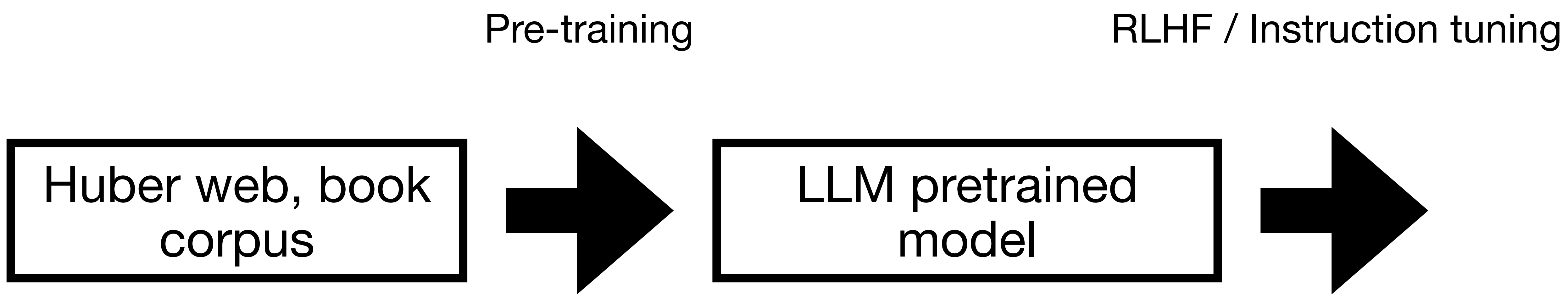
LLM

Lifecycle



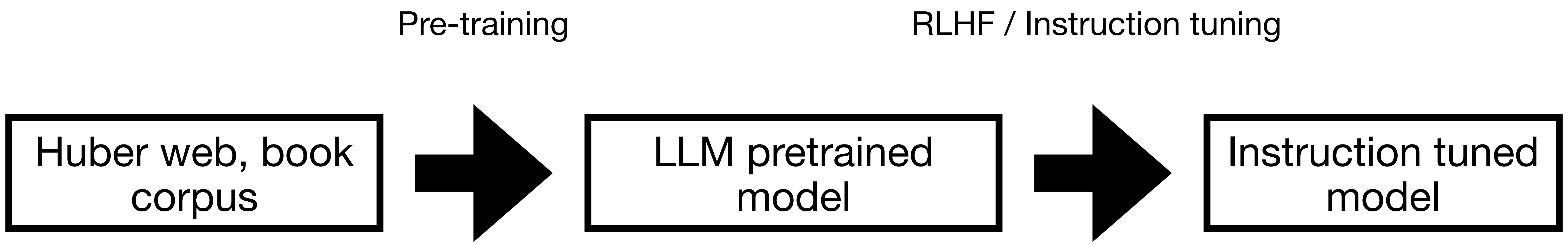
LLM

Lifecycle



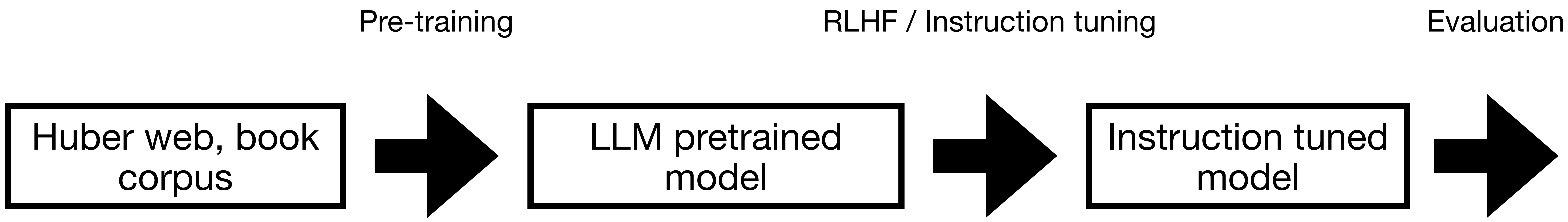
LLM

Lifecycle



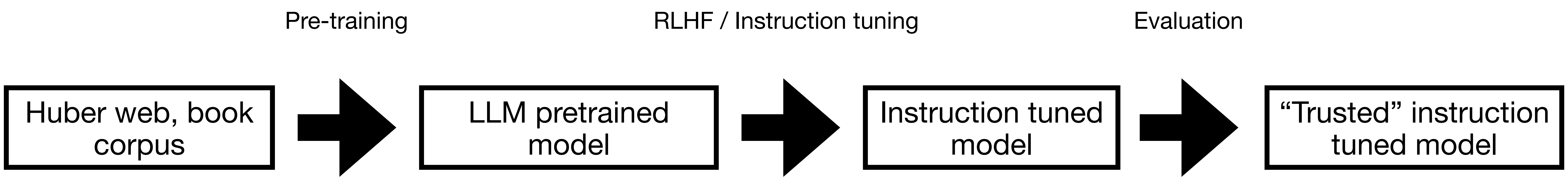
LLM

Lifecycle



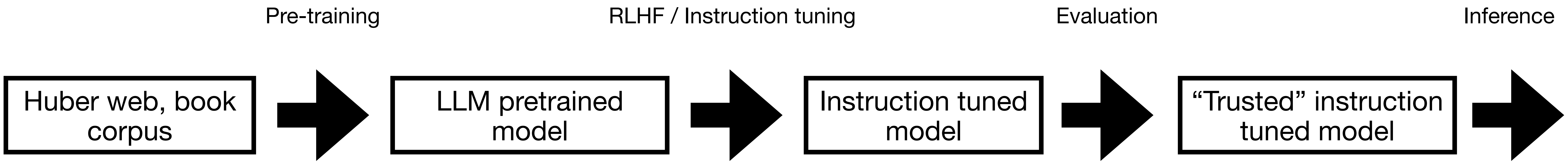
LLM

Lifecycle



LLM

Lifecycle



LLM

Lifecycle

LLM
Lifecycle

LLM Evaluation

Factual Knowledge
What is the capital of Australia?
Who wrote "Pride and Prejudice" and when was it published?
What are the main components of photosynthesis?

Language and Pattern Recognition
How many R's in strawberry?
What rhymes with "orange" in English?
Can you identify the grammatical error in this sentence: "Me and him went to th
store"?

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New Yo
3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

LLM Evaluation

- How do you select a LLM given the wide list (500+) of available options?

Factual Knowledge
What is the capital of Australia?
Who wrote "Pride and Prejudice" and when was it published?
What are the main components of photosynthesis?

Language and Pattern Recognition
How many R's in strawberry?
What rhymes with "orange" in English?
Can you identify the grammatical error in this sentence: "Me and him went to th
store"?

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New Yo
3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

LLM Evaluation

- How do you select a LLM given the wide list (500+) of available options?
- How would you evaluate a model that generate open-ended answer? 🤔

Factual Knowledge
What is the capital of Australia?
Who wrote "Pride and Prejudice" and when was it published?
What are the main components of photosynthesis?

Language and Pattern Recognition
How many R's in strawberry?
What rhymes with "orange" in English?
Can you identify the grammatical error in this sentence: "Me and him went to th store"?

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New Yo 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

LLM Evaluation

- How do you select a LLM given the wide list (500+) of available options?
- How would you evaluate a model that generate open-ended answer? 🤔

Factual Knowledge
What is the capital of Australia?
Who wrote "Pride and Prejudice" and when was it published?
What are the main components of photosynthesis?

Language and Pattern Recognition
How many R's in strawberry?
What rhymes with "orange" in English?
Can you identify the grammatical error in this sentence: "Me and him went to th store"?

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New Yo 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

LLM evaluations

Challenges

- Evaluating a generative model that produces open ended text is **hard**
- Many languages
- Many objectives
- Evaluating a single model with human annotations can costs thousands of dollars

Multilingual evaluations

Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks

Multilingual evaluations

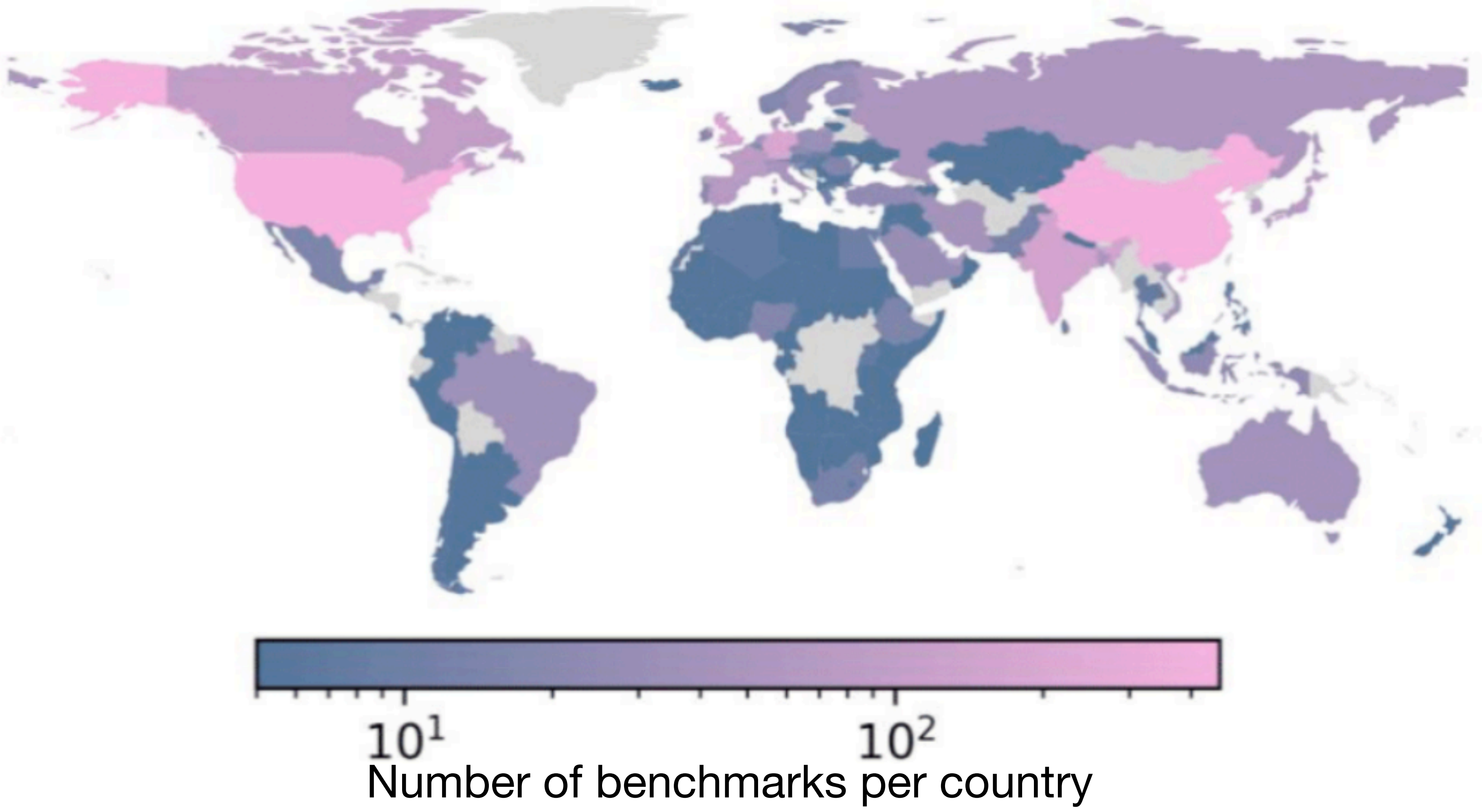
- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔

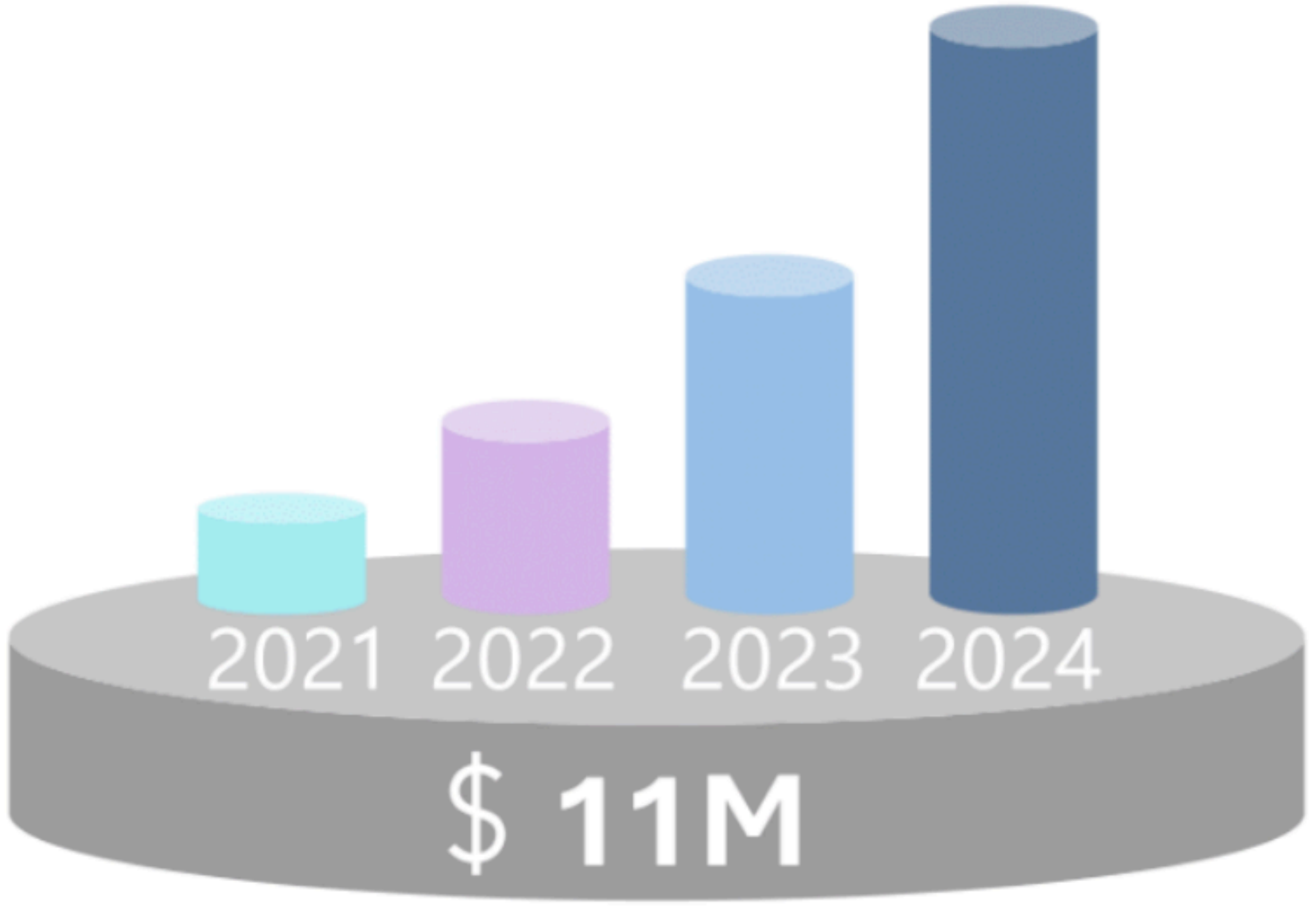
Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔
- Issues of language covering

Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔
- Issues of language covering

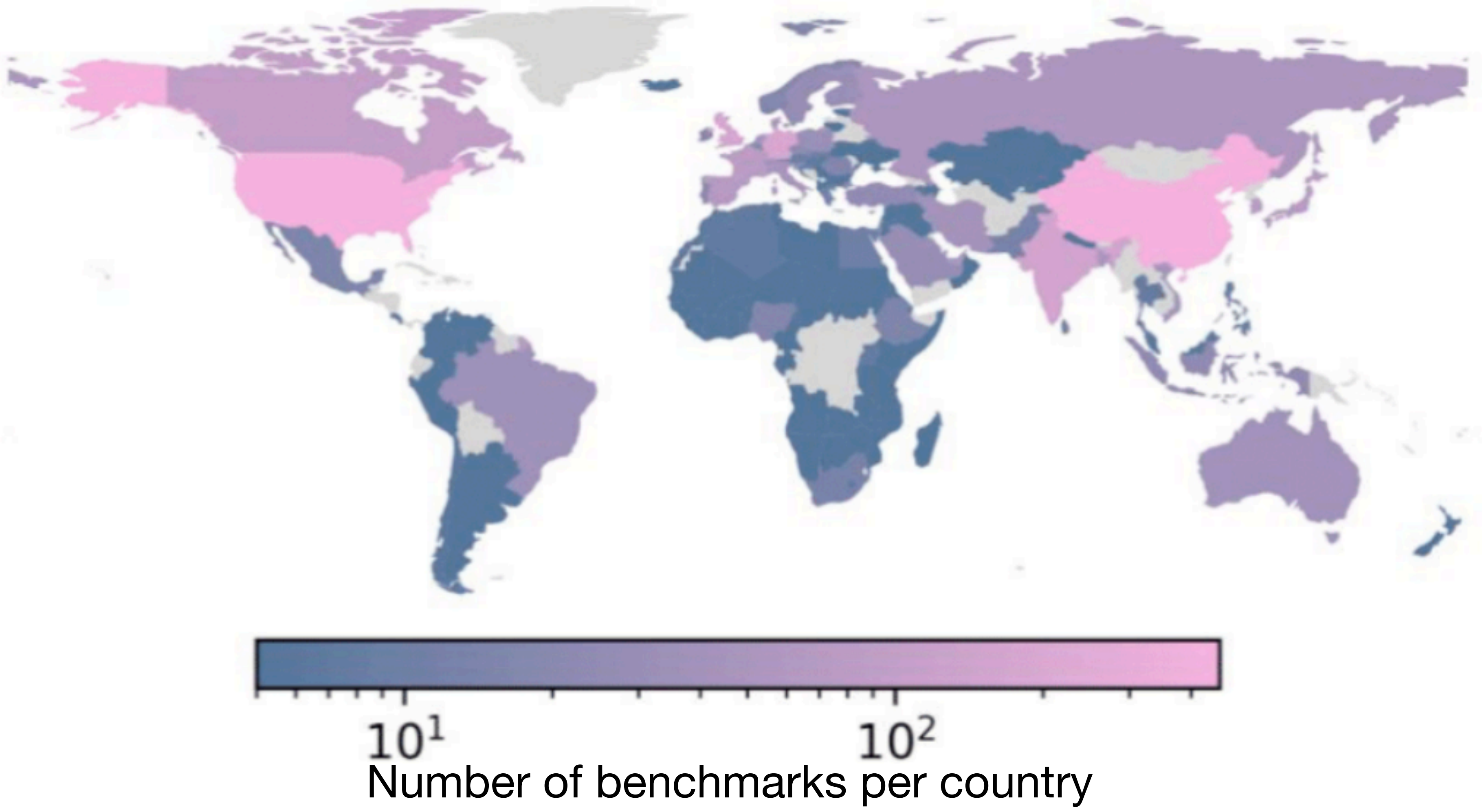


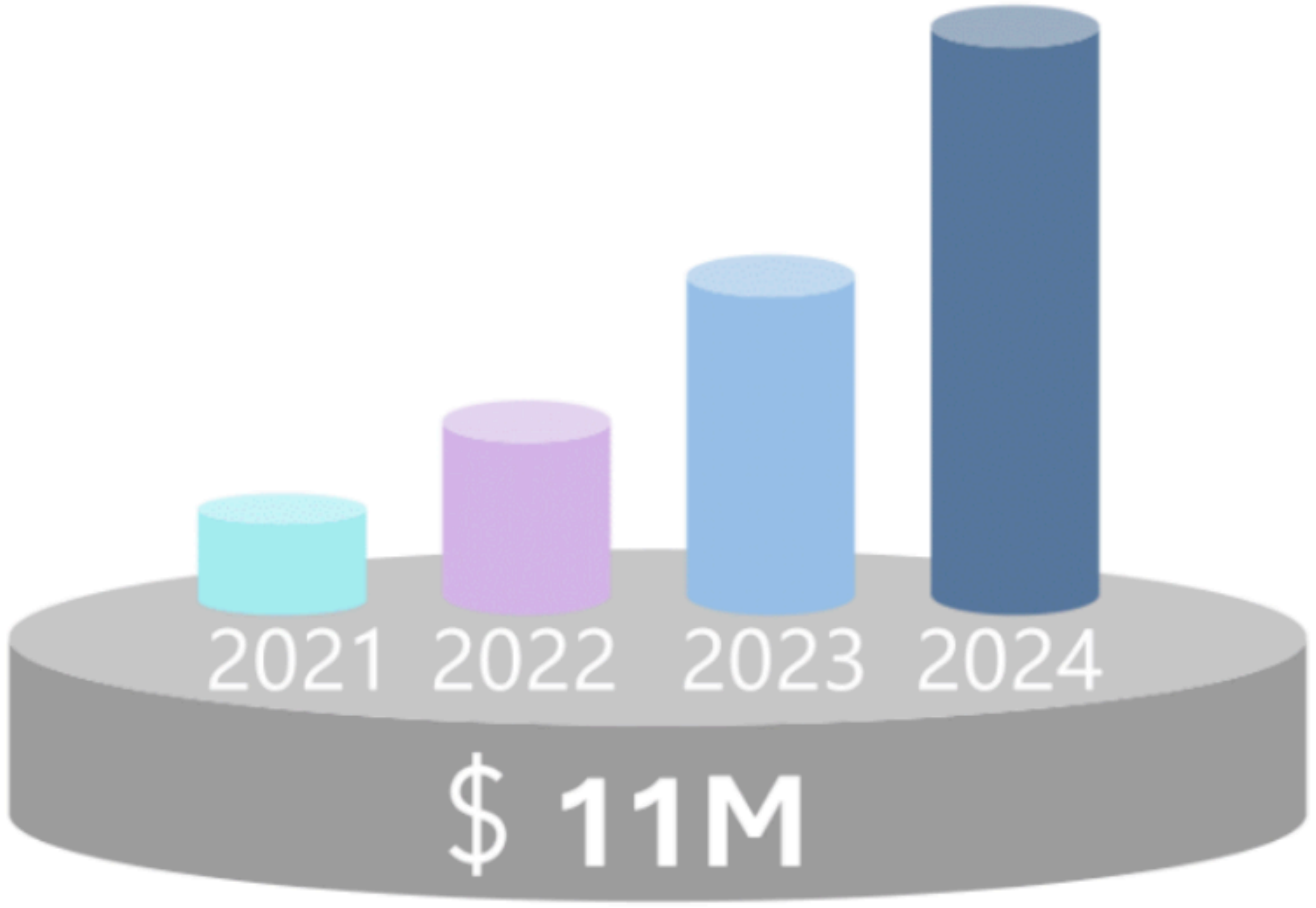


The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. Arxiv 2025.

Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔
- Issues of language covering
- Issues of automatic translation

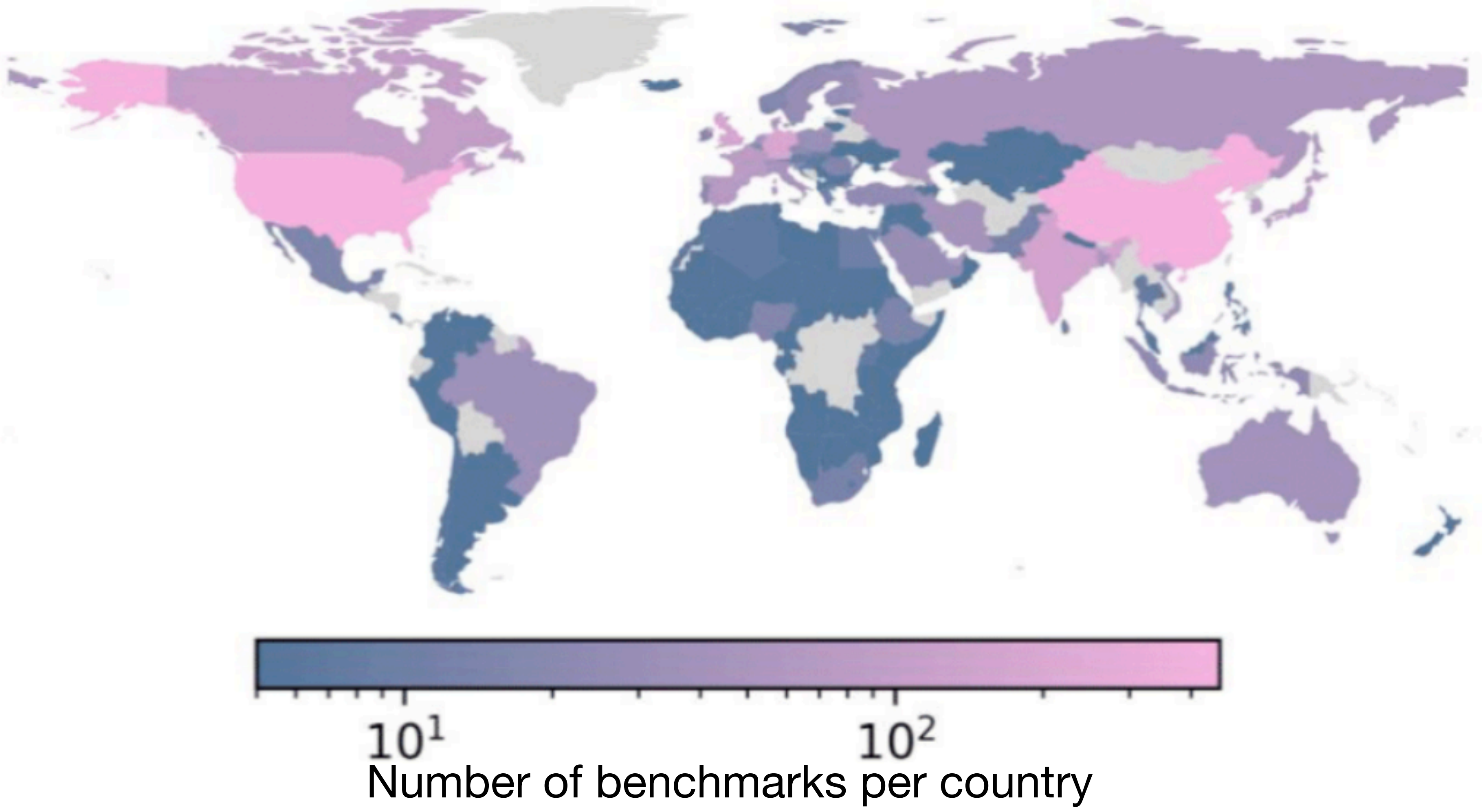


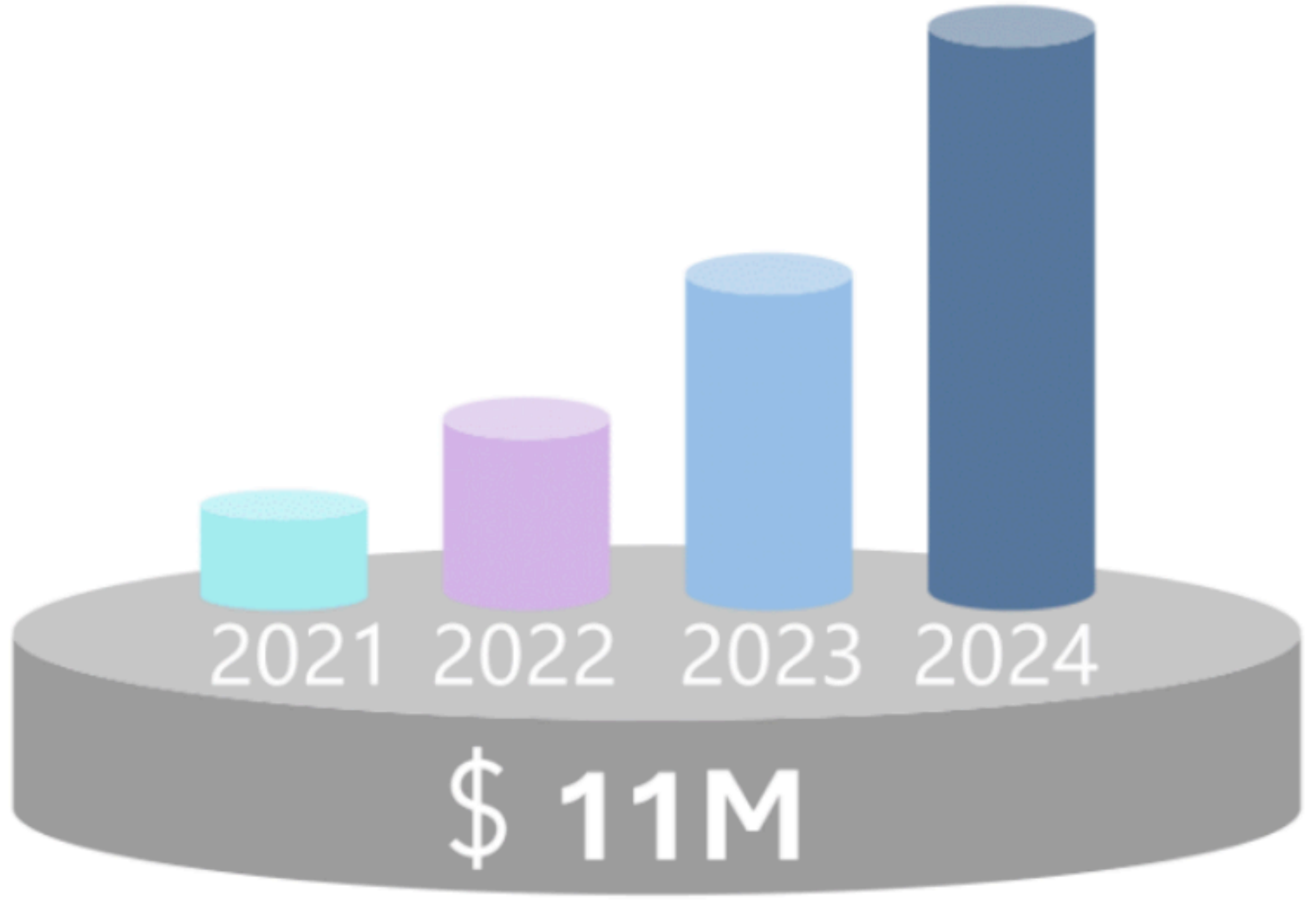


The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. Arxiv 2025.

Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔
- Issues of language covering
- Issues of automatic translation
  - Quick to get started but much worse correlation with human judgement

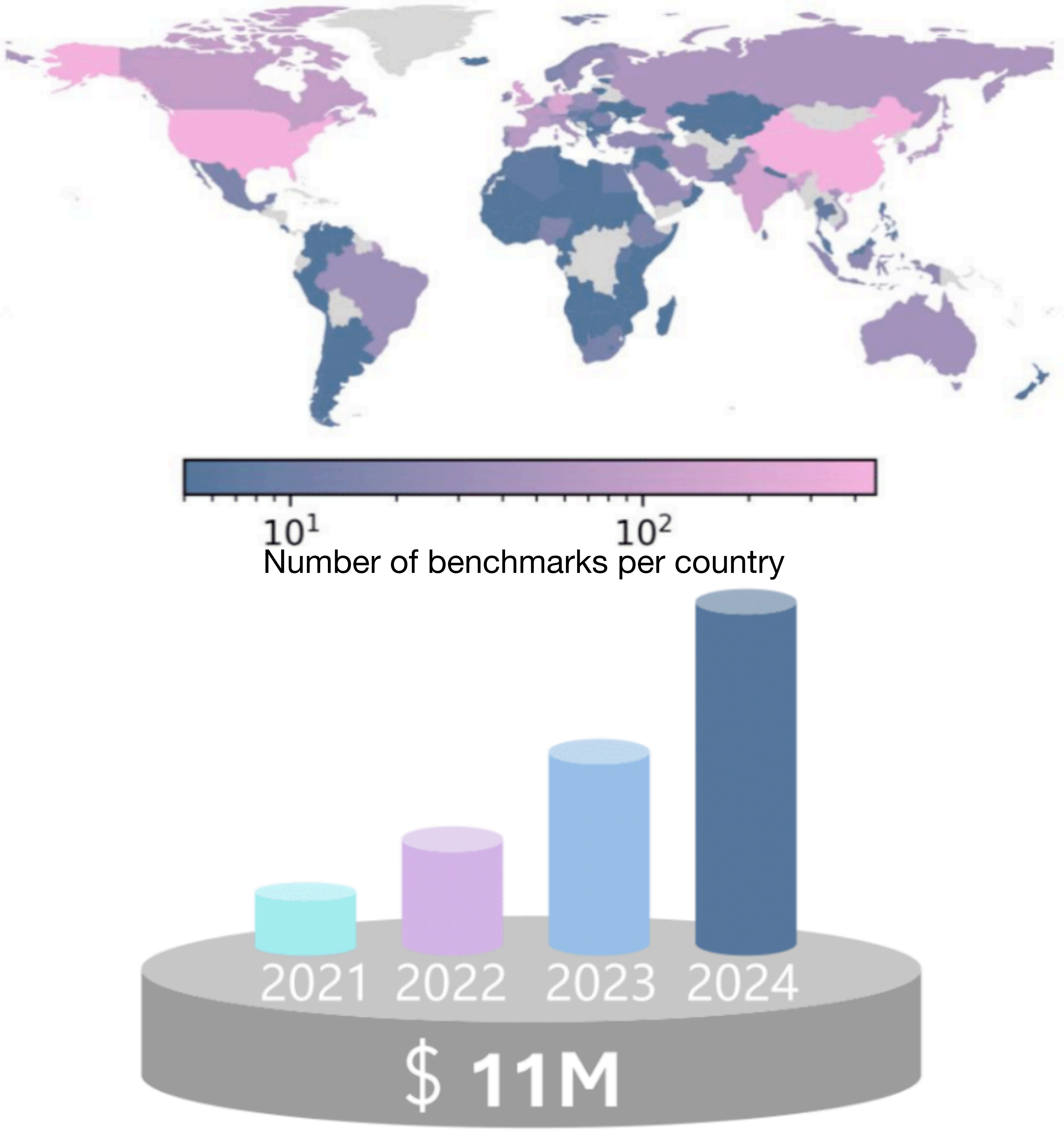




The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. Arxiv 2025.

Multilingual evaluations

- To evaluate mutilingual models, we need multilingual benchmarks
- How would you do it? 🤔
- Issues of language covering
- Issues of automatic translation
  - Quick to get started but much worse correlation with human judgement
- Cultural & bias of US/Western centric benchmark



The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. Arxiv 2025.

Evaluations goals

- Many objectives

Correctness/helpfulness

Fairness

Reasoning

Copyright

Safety

Style

Efficiency

…

Evaluations goals

- Many objectives

Correctness/helpfulness

Fairness

Reasoning

Copyright

Safety

Style

Efficiency

…

Each objective may require specific benchmarks

Evaluations goals

- Many objectives

Correctness/helpfulness

Fairness

Reasoning

Copyright

Safety

Style

Efficiency

…

Each objective may require specific benchmarks

How can all the objectives be *combined*?

LLM Evaluations - methods overview

Static Evaluations

Static zero and few-shot benchmarks

Static Evaluations

Static zero and few-shot benchmarks

- Prompt an LLM to answer questions either without examples (zero-shot) or with some examples (few-shots)

Static Evaluations

Static zero and few-shot benchmarks

- Prompt an LLM to answer questions either without examples (zero-shot) or with some examples (few-shots)
- Cheapest benchmark (just need to do inference)

Static Evaluations

Static zero and few-shot benchmarks

- Prompt an LLM to answer questions either without examples (zero-shot) or with some examples (few-shots)
- Cheapest benchmark (just need to do inference)
- An LLM can have a great score at static benchmarks if it has good knowledge…

Static Evaluations

Static zero and few-shot benchmarks

- Prompt an LLM to answer questions either without examples (zero-shot) or with some examples (few-shots)
- Cheapest benchmark (just need to do inference)
- An LLM can have a great score at static benchmarks if it has good knowledge…
- … but low utility if it cannot handle conversation, is too toxic or refuses to answer

Static Evaluations

Static zero and few-shot benchmarks

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
    - Prompt an LLM to answer questions
        - Without examples = zero-shot

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
    - Prompt an LLM to answer questions
        - Without examples = zero-shot
        - With some examples = few-shots

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5”
B "6x^2 + 4x + 6”
C “0"
D "x^2 + 1"
```

One example from MMLU

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5"
B "6x^2 + 4x + 6"
C "0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5”
B "6x^2 + 4x + 6”
C “0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5"
B "6x^2 + 4x + 6"
C "0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5”
B "6x^2 + 4x + 6”
C “0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😲 73%

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5”
B "6x^2 + 4x + 6”
C “0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😮 73%
- 8B models are quite good already…

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
    - Prompt an LLM to answer questions
        - Without examples = zero-shot
        - With some examples = few-shots
- Frameworks:

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5"
B "6x^2 + 4x + 6"
C "0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😯 73%
- 8B models are quite good already…

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots
- Frameworks:
  - Harness (Eleuther AI)

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5”
B "6x^2 + 4x + 6”
C “0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😯 73%
- 8B models are quite good already…

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots
- Frameworks:
  - Harness (Eleuther AI)
  - light-eval (Hugging-Face)

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5"
B "6x^2 + 4x + 6"
C "0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😯 73%
- 8B models are quite good already…

Static Evaluations

Static zero and few-shot benchmarks

- Static evaluations (MMLU, HellaSWAG and others)
  - Prompt an LLM to answer questions
    - Without examples = zero-shot
    - With some examples = few-shots
- Frameworks:
  - Harness (Eleuther AI)
  - light-eval (Hugging-Face)
  - HELM (Stanford)

```
Find the product of the given polynomials in
the given polynomial ring.
f(x) = 4x - 5, g(x) = 2x^2 - 4x + 2 in Z_8[x].

Choices:
A "2x^2 + 5"
B "6x^2 + 4x + 6"
C "0"
D "x^2 + 1"
```

One example from MMLU

Two questions for you!

🤔 Can you answer this MMLU example 👆?

🤔 What accuracy do you think LLama3.1-8B gets on average on MMLU?

- 😮 73%
- 8B models are quite good already…

Static Evaluations

How to use Harness

```
git clone https://github.com/EleutherAI/lm-evaluation-harness
cd lm-evaluation-harness
pip install -e .
```

```
lm_eval --model hf \
    --model_args pretrained=EleutherAI/gpt-j-6B \
    --tasks hellaswag \
    --device cuda:0 \
    --batch_size 8
```

Static Evaluations

How to use Harness

```
git clone https://github.com/EleutherAI/lm-evaluation-harness
cd lm-evaluation-harness
pip install -e .
```

```
lm_eval --model hf \
    --model_args pretrained=EleutherAI/gpt-j-6B \
    --tasks hellaswag \
    --device cuda:0 \
    --batch_size 8
```

Straightforward to use

Dynamic Evaluations

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ →

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:
  - golden metric but ? 🤔

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:
  - golden metric but ? 🤔
  - very expensive (💸💸💸)

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:
  - golden metric but ? 🤔
  - very expensive (💸💸💸)
  - slow, asynchronous

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:
  - golden metric but ? 🤔
  - very expensive (💸💸💸)
  - slow, asynchronous
- Ask an LLM to judge/rank models: less accurate but less expensive (💸)

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️ → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

Dynamic Evaluations

- How to evaluate open ended generated text? 🤔
- Ask iid humans to rank models:
  - golden metric but ? 🤔
  - very expensive (💸💸💸)
  - slow, asynchronous
- Ask an LLM to judge/rank models: less accurate but less expensive (💸)

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Human 🧑‍⚖️

LLM judge 🤖 → **Model 2**

Creative Tasks
Write a haiku about artificial intelligence.
Create a short story that begins with "The last library on Earth closed today."
Generate three marketing slogans for a sustainable clothing brand.

Analysis and Interpretation
What are the main themes in George Orwell's "1984"?
Compare and contrast renewable vs. non-renewable energy sources.
Analyze the potential economic impacts of universal basic income.

Planning and Recommendations
Can you recommend a two day trip to Hawaii?
What's a good study schedule for learning Spanish in 3 months?
Help me plan a vegetarian dinner party for 8 people.

Mathematical and Logical Reasoning
Can you prove the halting theorem?
If a train leaves Chicago at 2 PM traveling 60 mph, and another leaves New York at 3 PM traveling 80 mph, when do they meet?
Explain the prisoner's dilemma and its implications.

Technical Problem-Solving
Debug this Python code that's supposed to sort a list but isn't working properly.
Explain how blockchain technology works in simple terms.

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations

- One standard is ChatBot Arena which uses crowd annotation

- Getting high on this leaderboard is a **big deal**

- Elo-like ratings for selected models that mostly agrees with intuitive performance

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3	2	ChatGPT-4o-latest (2025-01-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9	10	o3-mini-high	1329

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3	[illegible]	[illegible]01-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9	10	o3-mini-high	1329

Best open weight model is #6

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3	[illegible]	[illegible]1-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9	10	o3-mini-high	1329
46	51	Llama-3.1-Tulu-3-70B	1243

Best open weight model is #6

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3		1-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9			1329
46	51	Llama-3.1-Tulu-3-70B	1243

Best open weight model is #6

… Best open weight model with open post-training receipt is #46

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3		1-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9			1329
46	51	Llama-3.1-Tulu-3-70B	1243
153	161	OLMo-7B-instruct	1015

Best open weight model is #6

… Best open weight model with open post-training receipt is #46

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3		01-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9			1329
46	51	Llama-3.1-Tulu-3-70B	1243
153	161	OLMo-7B-instruct	1015

Best open weight model is #6

… Best open weight model with open post-training receipt is #46

… Best fully open model is #153

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Very expensive ~3.5K$ per model of human annotation salary

Can we use something *cheaper*?

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3		1-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9			1329
46	51	Llama-3.1-Tulu-3-70B	1243
153	161	OLMo-7B-instruct	1015

Best open weight model is #6

… Best open weight model with open post-training receipt is #46

… Best fully open model is #153

LLM evaluations

Dynamic evaluations

- What about chat capability? There we clearly need dynamic evaluations
- One standard is ChatBot Arena which uses crowd annotation
- Getting high on this leaderboard is a **big deal**
- Elo-like ratings for selected models that mostly agrees with intuitive performance

Very expensive ~3.5K$ per model of human annotation salary

Can we use something *cheaper*?

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	2	Grok-3-Preview-02-24	1412
1	1	GPT-4.5-Preview	1411
3	5	Gemini-2.0-Flash-Thinking-Exp-01-21	1384
3	3	Gemini-2.0-Pro-Exp-02-05	1380
3		1-29)	1377
6	3	DeepSeek-R1	1363
6	10	Gemini-2.0-Flash-001	1357
7	3	o1-2024-12-17	1352
9	10	Qwen2.5-Max	1336
9	7	o1-preview	1335
9			1329
46	51	Llama-3.1-Tulu-3-70B	1243
153	161	OLMo-7B-instruct	1015

Best open weight model is #6

… Best open weight model with open post-training receipt is #46

… Best fully open model is #153

NB: Those results are from May 2025. The exact ranking may have changed but the trend remain.

Evaluation

Evaluating… at which cost?

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
    - trade off accuracy (spearman correlation with human judgement)…

Evaluation

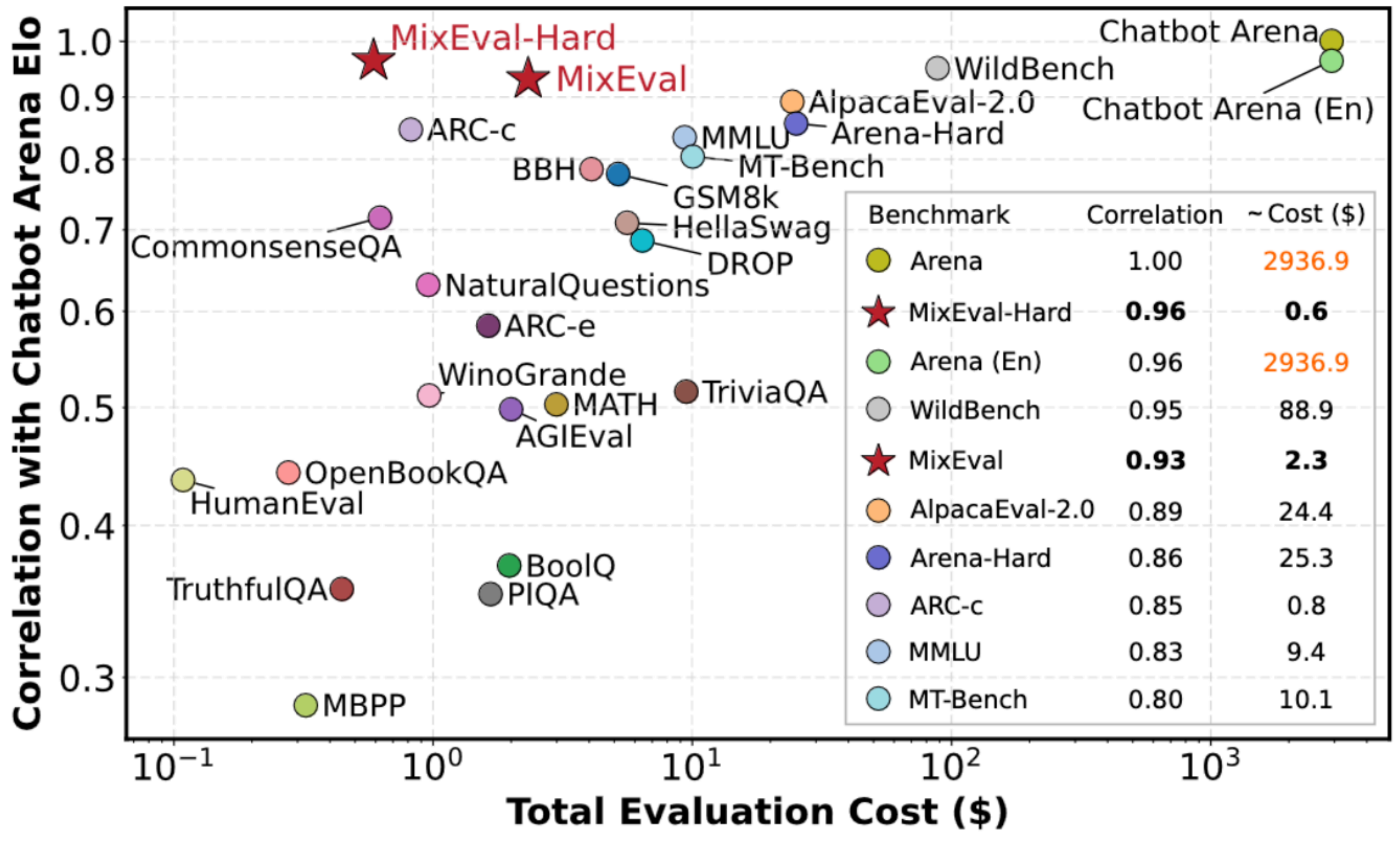
Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
  - trade off accuracy (spearman correlation with human judgement)…
  - … for cost

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
    - trade off accuracy (spearman correlation with human judgement)…
    - … for cost



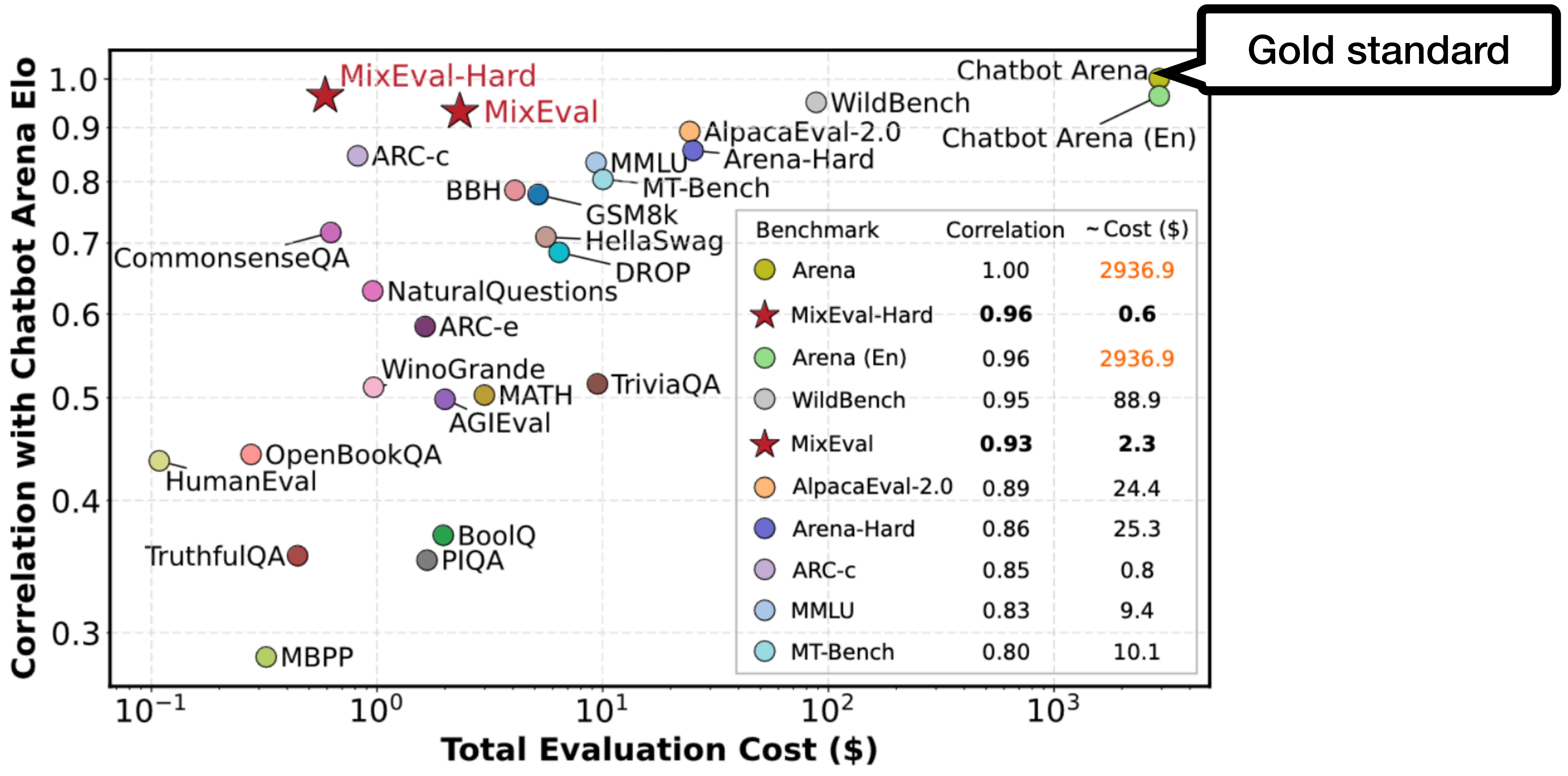
LLM evaluation cost and correlation with Elo ratings

Source: MixEval [Ni 2024]

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
    - trade off accuracy (spearman correlation with human judgement)…
    - … for cost



Benchmark	Correlation	~Cost ($)
Arena	1.00	2936.9
MixEval-Hard	**0.96**	**0.6**
Arena (En)	0.96	2936.9
WildBench	0.95	88.9
MixEval	**0.93**	**2.3**
AlpacaEval-2.0	0.89	24.4
Arena-Hard	0.86	25.3
ARC-c	0.85	0.8
MMLU	0.83	9.4
MT-Bench	0.80	10.1

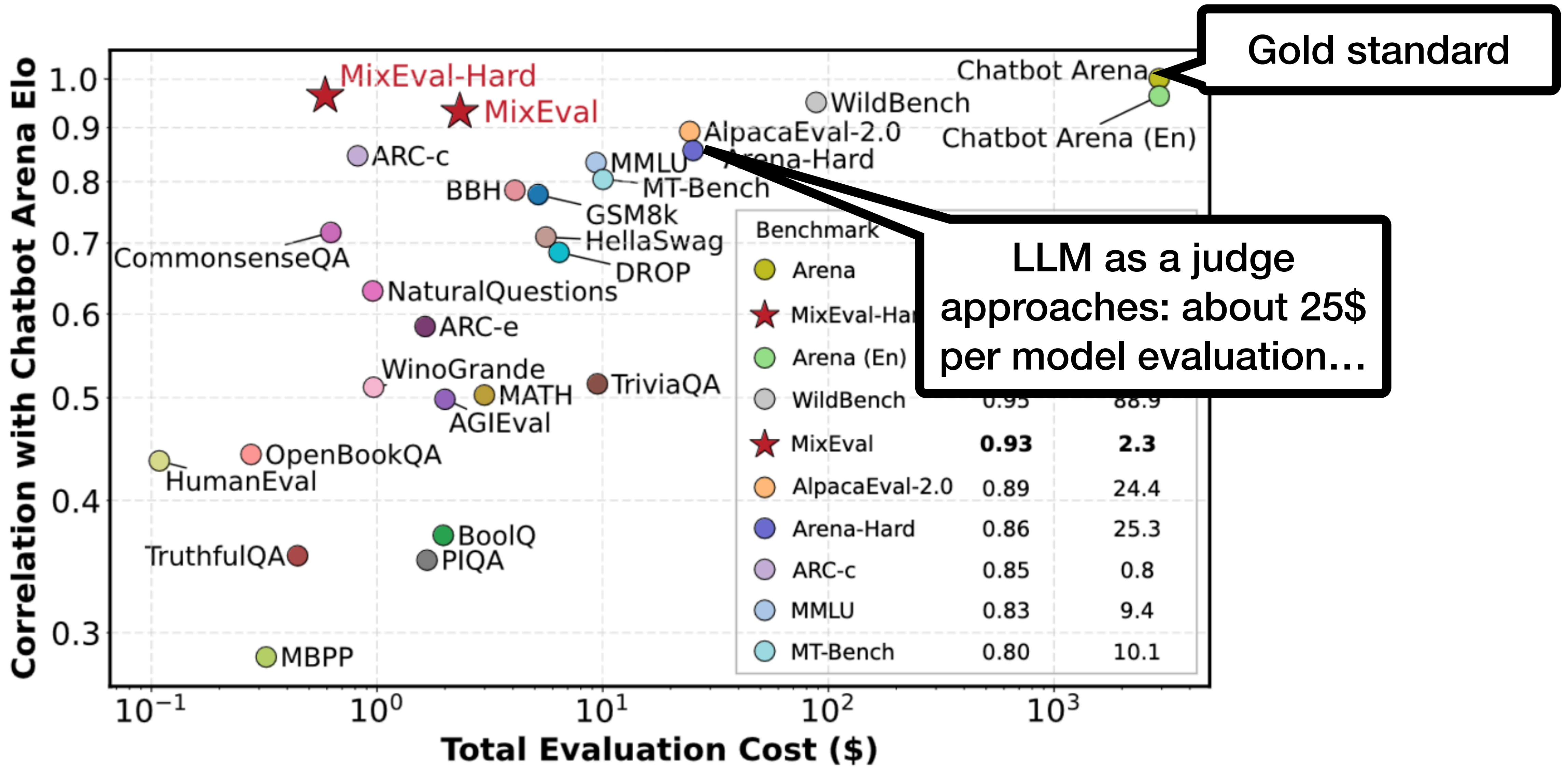
LLM evaluation cost and correlation with Elo ratings

Source: MixEval [Ni 2024]

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
    - trade off accuracy (spearman correlation with human judgement)…
    - … for cost



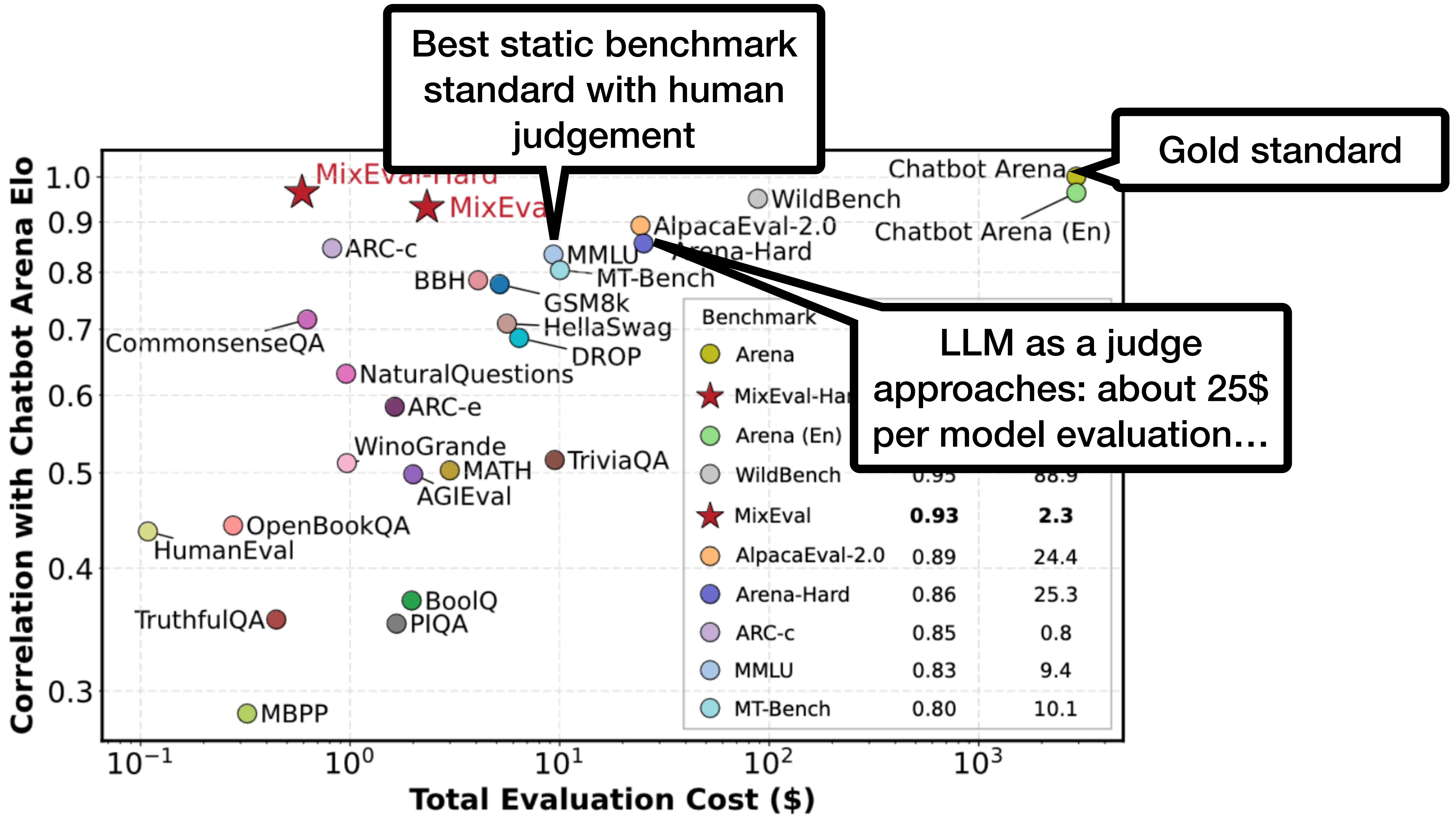
LLM evaluation cost and correlation with Elo ratings

Source: MixEval [Ni 2024]

Evaluation

Evaluating… at which cost?

- But Chatbot Arena is expensive (~3.5K$ if annotators would be paid minimum US wage)
- Can we do better?
- Multiobjective problem:
  - trade off accuracy (spearman correlation with human judgement)…
  - … for cost



LLM evaluation cost and correlation with Elo ratings

Source: MixEval [Ni 2024]

LLM as a judge

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Model 1: Llama3-70B, Model 2 : GPT4-o

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions
- 3. Ask an LLM judge (GPT4-o) which model is better on each instruction

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Model 1: Llama3-70B, Model 2 : GPT4-o

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions
- 3. Ask an LLM judge (GPT4-o) which model is better on each instruction

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Model 1: Llama3-70B, Model 2 : GPT4-o

LLM judge

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions
- 3. Ask an LLM judge (GPT4-o) which model is better on each instruction

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

Model 1: Llama3-70B, Model 2 : GPT4-o

LLM judge → **Model 2**

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions
- 3. Ask an LLM judge (GPT4-o) which model is better on each instruction
- 4. Return the winrate on all instructions for the model against the baseline

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

→ **Model 2**

LLM judge

Model 1: Llama3-70B, Model 2 : GPT4-o

LLM as a judge

- Idea 💡: ask LLM to tell which LLM output is better
- 1. Ask the model you want to evaluate (say LLama3-8B) to generate outputs on a set of fixed instructions (805 for Alpaca-Eval, 500 for Arena-Hard)
- 2. Ask a baseline (GPT4-o) to generate outputs on all instructions
- 3. Ask an LLM judge (GPT4-o) which model is better on each instruction
- 4. Return the winrate on all instructions for the model against the baseline

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues, and popular music. Here are some cool music styles and artists from the 1920s:**Jazz** Louis Armstrong: Considered one of the most influential jazz musicians of all time…

Model 2: The 1920s, often referred to as the \"Roaring Twenties,\" was a period that saw a significant explosion in the popularity of jazz and the beginning of widespread commercial recording…

LLM judge → **Model 2**

Model 1: Llama3-70B, Model 2 : GPT4-o

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13)	57.5%	51.3%
2	GPT-4 Turbo (04/09)	55.0%	46.1%
3	Yi-Large Preview	51.9%	57.5%
4	GPT-4 Preview (11/06)	50.0%	50.0%
5	Claude 3 Opus (02/29)	40.5%	29.1%
6	GPT-4	38.1%	23.6%
7	Qwen1.5 72B Chat	36.6%	26.5%
8	GPT-4 (03/14)	35.3%	22.1%
9	Claude 3 Sonnet (02/29)	34.9%	25.6%
10	Llama 3 70B Instruct	34.4%	33.2%

Leaderboard: winrate against GPT4-turbo

LLM as a judge

Current challenges

LLM as a judge

Current challenges

- “Smallish" issues:

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)
- Bigger issues:

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)
- Bigger issues:
  - Lack of proper cross validation (same set of models used to select hyperparameters and report results)

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)
- Bigger issues:
  - Lack of proper cross validation (same set of models used to select hyperparameters and report results)
  - Bias of the judge model (GPT4 tends to prefer GPT4)

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)
- Bigger issues:
  - Lack of proper cross validation (same set of models used to select hyperparameters and report results)
  - Bias of the judge model (GPT4 tends to prefer GPT4)
  - Current SOTA use close model (GPT4) as the judge

LLM as a judge

Current challenges

- “Smallish" issues:
  - Preference for verbose outputs
  - Bias of the position: judges have small preference for first position
  - Data contamination (small because one can update the benchmarked prompts over time)
- Bigger issues:
  - Lack of proper cross validation (same set of models used to select hyperparameters and report results)
  - Bias of the judge model (GPT4 tends to prefer GPT4)
  - Current SOTA use close model (GPT4) as the judge
  - Cost of evaluation (still 20-25$ for Alpaca Eval and Arena hard)

A story of leaderboards

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
    - Static:
        - OpenLLM leaderboard
        - Helm
    - Dynamic

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm
  - Dynamic
    - ChatbotArena

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
    - Static:
        - OpenLLM leaderboard
        - Helm
    - Dynamic
        - ChatbotArena
        - Openrouter

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm
  - Dynamic
    - ChatbotArena
    - Openrouter
    - LiveBench LLM

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
    - Static:
        - OpenLLM leaderboard
        - Helm
    - Dynamic
        - ChatbotArena
        - Openrouter
        - LiveBench LLM
    - Dynamic (synthetic)

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
    - Static:
        - OpenLLM leaderboard
        - Helm
    - Dynamic
        - ChatbotArena
        - Openrouter
        - LiveBench LLM
    - Dynamic (synthetic)
        - Alpaca-Eval

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
    - Static:
        - OpenLLM leaderboard
        - Helm
    - Dynamic
        - ChatbotArena
        - Openrouter
        - LiveBench LLM
    - Dynamic (synthetic)
        - Alpaca-Eval
        - Arena-Hard

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm
  - Dynamic
    - ChatbotArena
    - Openrouter
    - LiveBench LLM
  - Dynamic (synthetic)
    - Alpaca-Eval
    - Arena-Hard
    - Translated versions

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm
  - Dynamic
    - ChatbotArena
    - Openrouter
    - LiveBench LLM
  - Dynamic (synthetic)
    - Alpaca-Eval
    - Arena-Hard
    - Translated versions

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - OpenLLM leaderboard
    - Helm
  - Dynamic
    - ChatbotArena
    - Openrouter
    - LiveBench LLM
  - Dynamic (synthetic)
    - Alpaca-Eval
    - Arena-Hard
    - Translated versions

A story of leaderboards

- Leaderboards are important as they help standardize the evaluation process
- Can help the community to focus on a single direction (Imagenet)
- Leaderboards for LLMs:
  - Static:
    - $\boxed{\text{OpenLLM leaderboard}}$
    - Helm
  - Dynamic
    - $\boxed{\text{ChatbotArena}}$
    - Openrouter
    - LiveBench LLM
  - Dynamic (synthetic)
    - Alpaca-Eval
    - Arena-Hard
    - Translated versions

We will focus on those

OpenLLM Leaderboard

OpenLLM Leaderboard

- Created in 2022 by HuggingFace

OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)

OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏

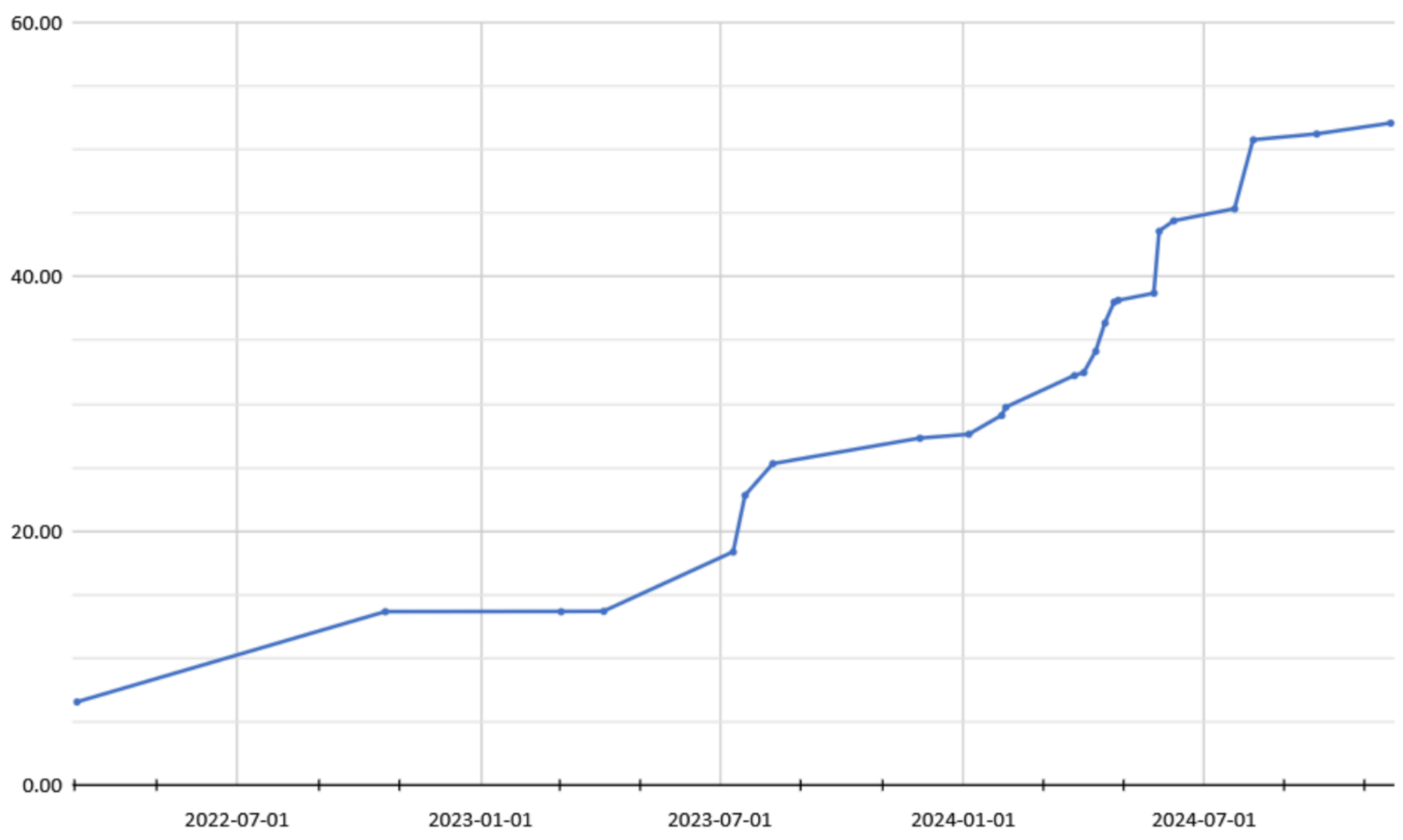
OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025

OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025

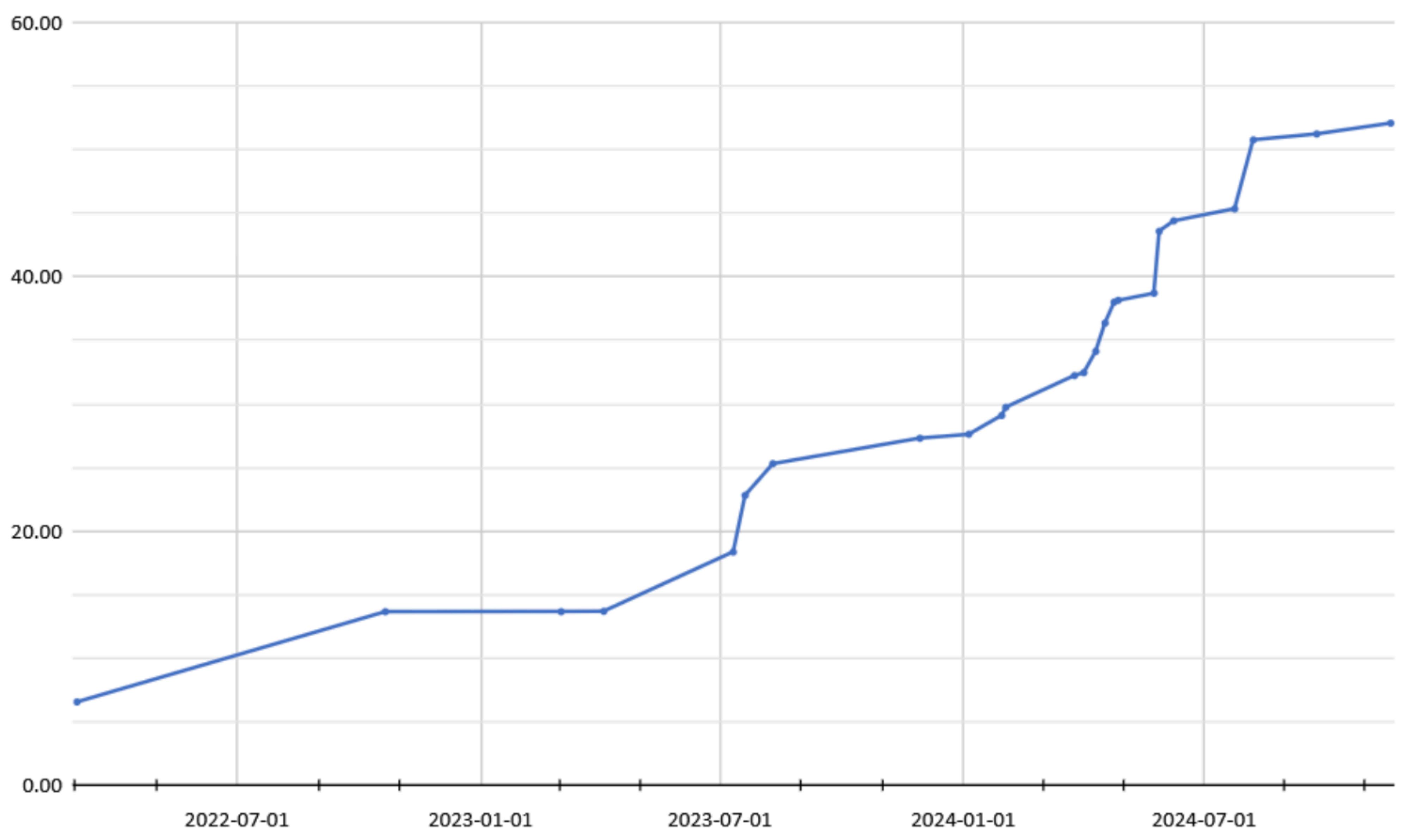
Average accuracy over time



OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025
  - No automatic check for contamination

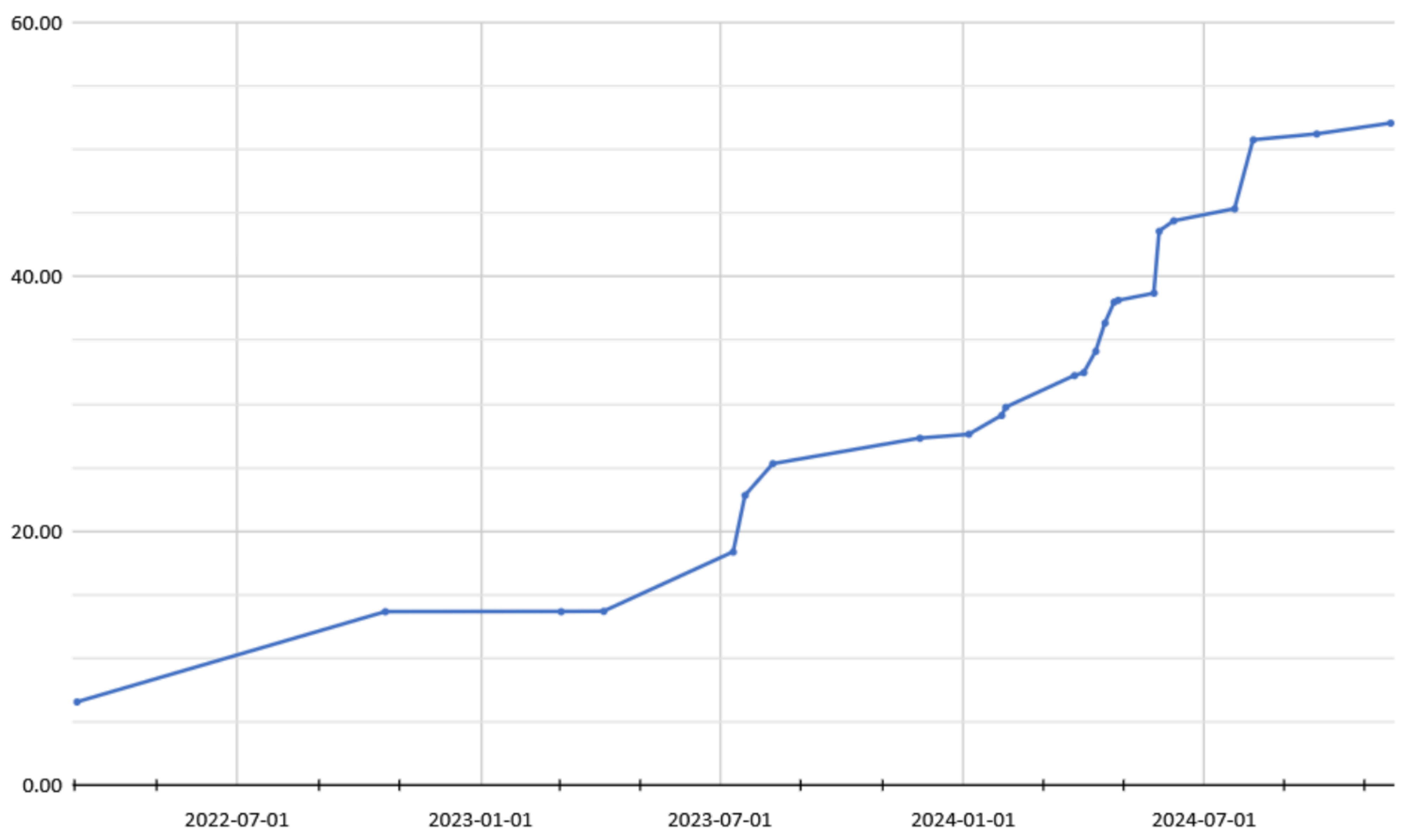
Average accuracy over time



OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025
  - No automatic check for contamination
  - Top solution could fine-tune arbitrarily on the test set

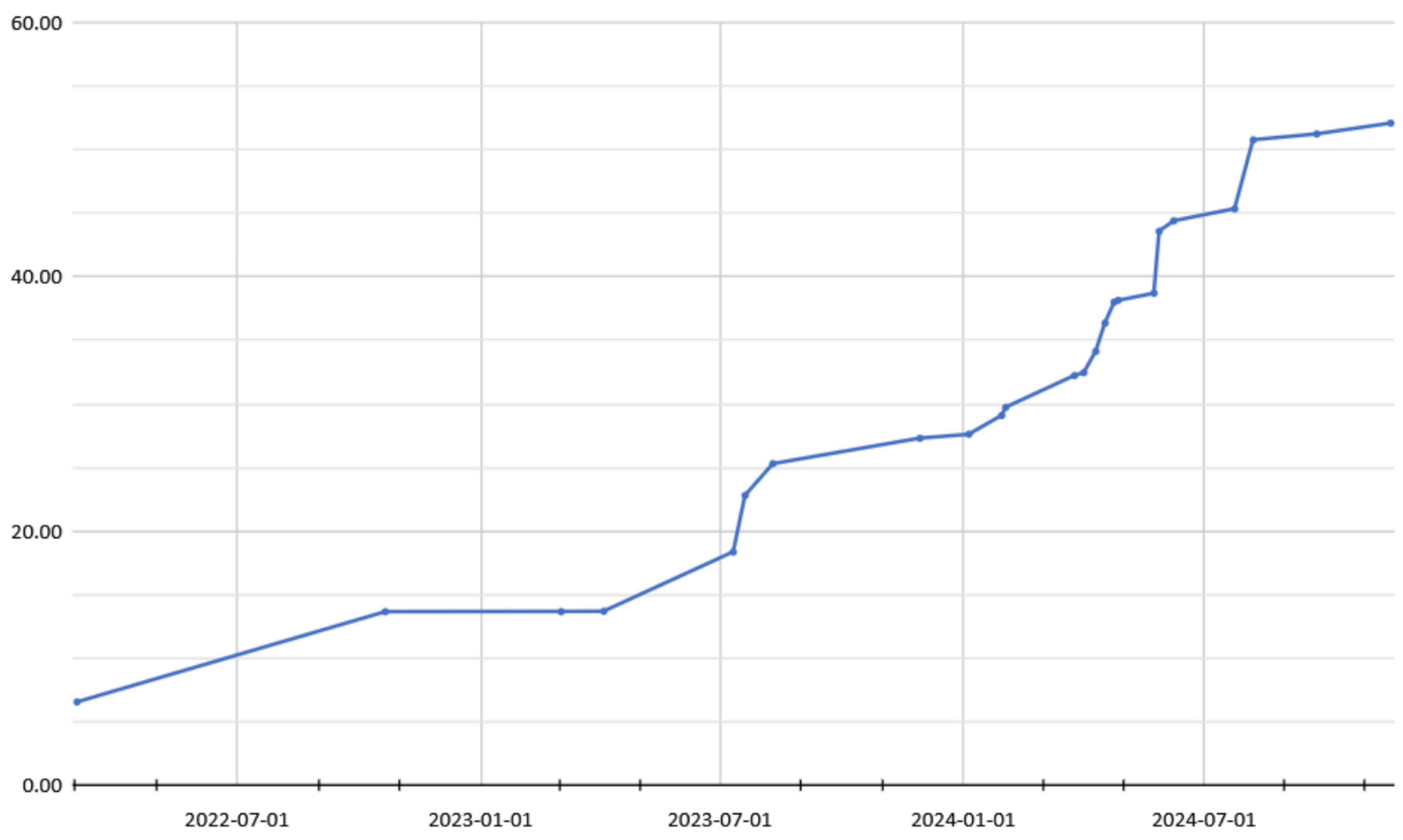
Average accuracy over time



OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025
    - No automatic check for contamination
    - Top solution could fine-tune arbitrarily on the test set
    - Try to fix and then abandon it

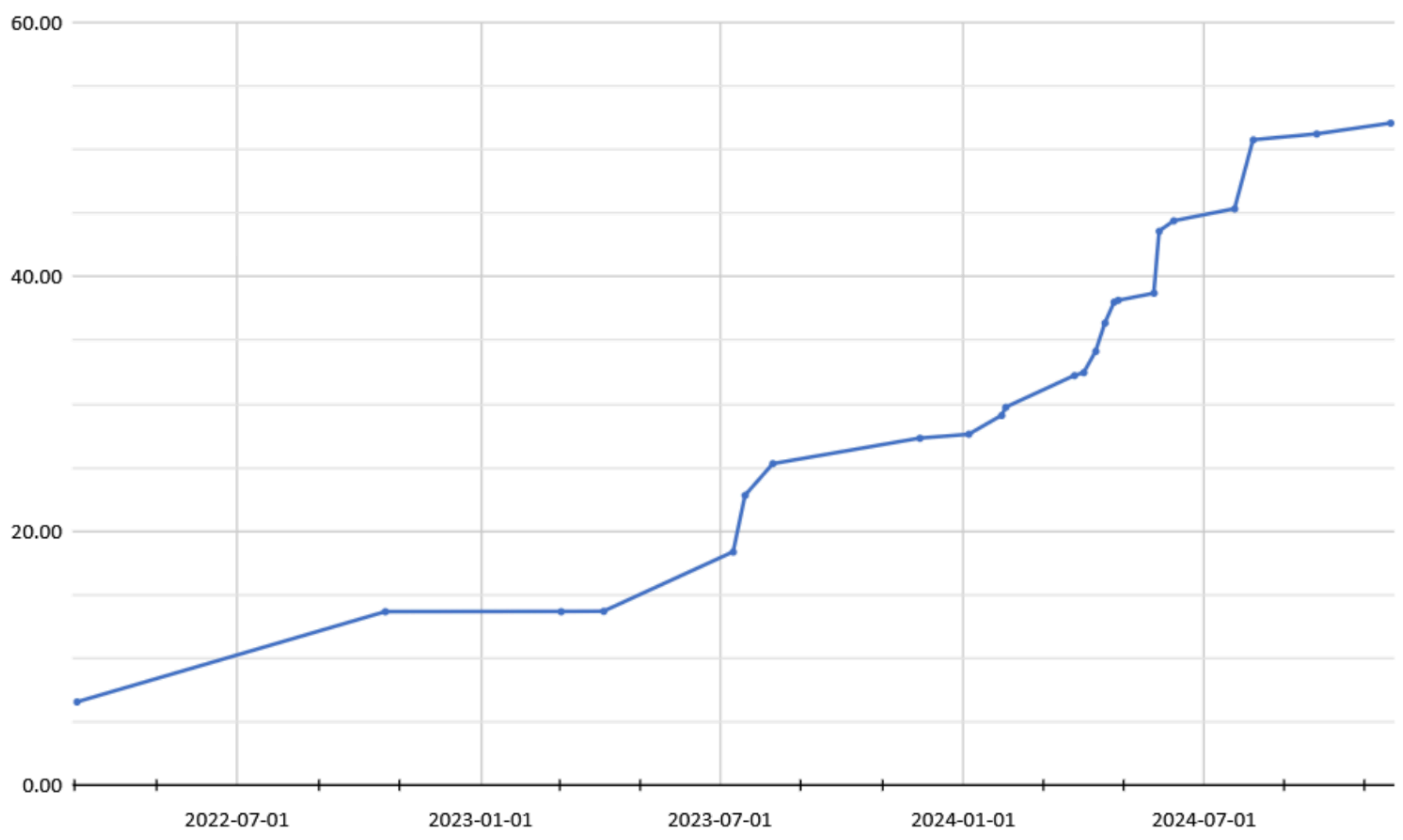
Average accuracy over time



OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025
  - No automatic check for contamination
  - Top solution could fine-tune arbitrarily on the test set
  - Try to fix and then abandon it
- Still a massive compute gain to the community 13K models were evaluated for free for ~3 million GPU hours

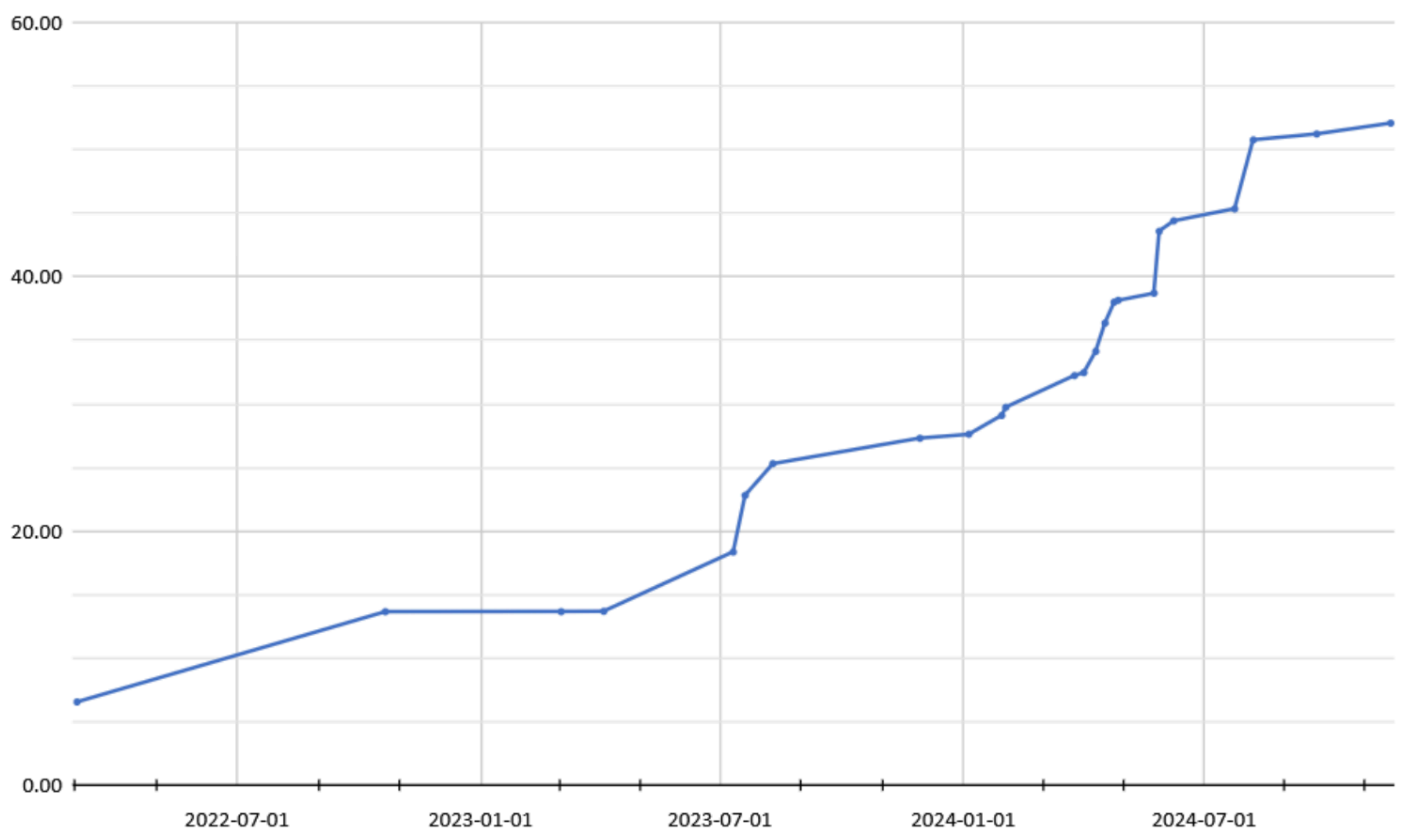
Average accuracy over time



OpenLLM Leaderboard

- Created in 2022 by HuggingFace
- Average scores on a list of static benchmarks (list got updated 2 times)
- Models evaluated **for free** on Hugging Face cluster 🙏
- Widely popular for some time but ended in April 2025
    - No automatic check for contamination
    - Top solution could fine-tune arbitrarily on the test set
    - Try to fix and then abandon it
- Still a massive compute gain to the community 13K models were evaluated for free for ~3 million GPU hours
- All data publicly available

Average accuracy over time



Chatbot Arena leaderboard

Chatbot Arena leaderboard

- Chatbot Arena was introduced in 2024 to evaluate LLMs

Chatbot Arena leaderboard

- Chatbot Arena was introduced in 2024 to evaluate LLMs
- Dynamic benchmark: pair of completions are shown to the user who pick the best

Chatbot Arena leaderboard

- Chatbot Arena was introduced in 2024 to evaluate LLMs
- Dynamic benchmark: pair of completions are shown to the user who pick the best
- Ratings computed for participating models (Bradley-Terry coefficients) that mostly agrees with intuitive performance

Chatbot Arena leaderboard

- Chatbot Arena was introduced in 2024 to evaluate LLMs
- Dynamic benchmark: pair of completions are shown to the user who pick the best
- Ratings computed for participating models (Bradley-Terry coefficients) that mostly agrees with intuitive performance
- Importance sampling used to select most uncertain pair models

Chatbot Arena leaderboard

- Chatbot Arena was introduced in 2024 to evaluate LLMs
- Dynamic benchmark: pair of completions are shown to the user who pick the best
- Ratings computed for participating models (Bradley-Terry coefficients) that mostly agrees with intuitive performance
- Importance sampling used to select most uncertain pair models

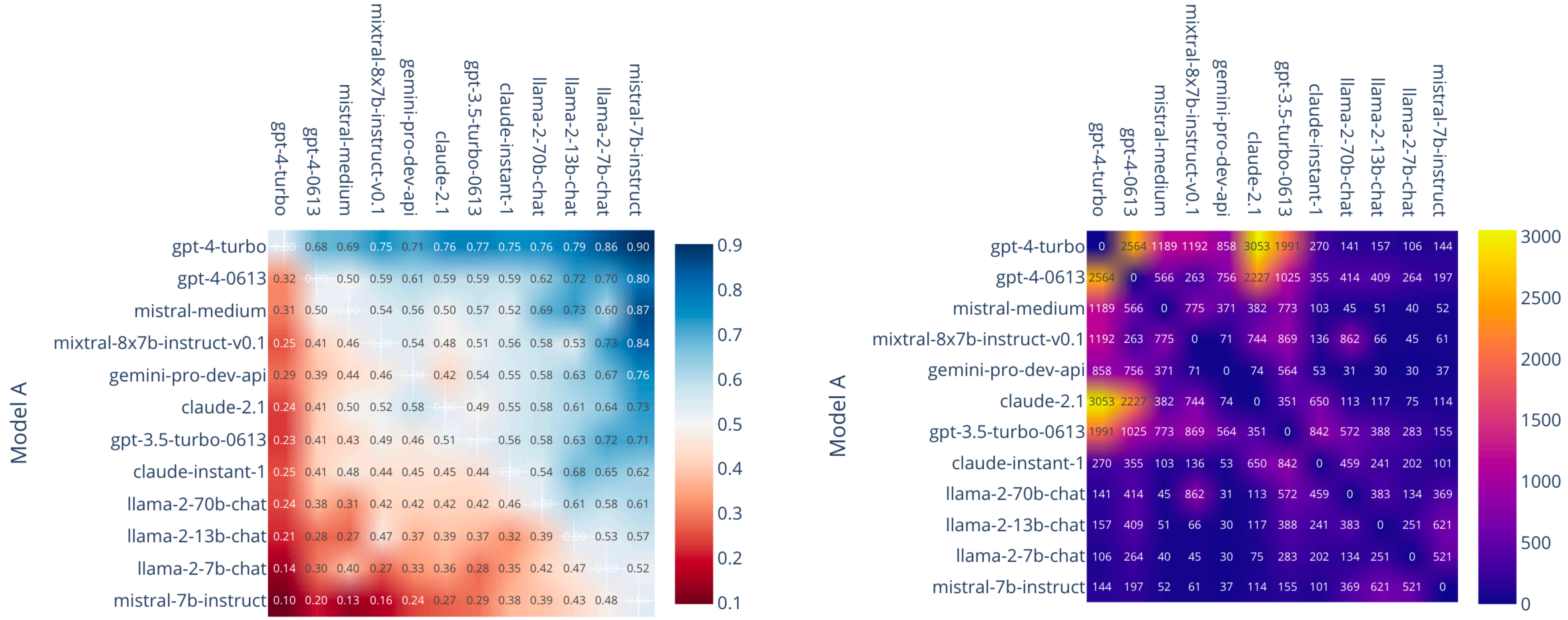


Figure 2. Win-rate (left) and battle count (right) between a subset of models in Chatbot Arena.

Chatbot Arena leaderboard

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:
    - Avoid contamination / overfitting to specific ChatBot Arena users

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:
  - Avoid contamination / overfitting to specific ChatBot Arena users
  - But private company sees **a lot of** data…

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:
  - Avoid contamination / overfitting to specific ChatBot Arena users
  - But private company sees **a lot of** data…
- Need for a transparent leaderboard

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:
  - Avoid contamination / overfitting to specific ChatBot Arena users
  - But private company sees **a lot of** data…
- Need for a transparent leaderboard

The Leaderboard Illusion

Shivalika Singh[*1], Yiyang Nan[1], Alex Wang[2], Daniel D'souza[1], Sayash Kapoor[3], Ahmet Üstün[1], Sanmi Koyejo[4], Yuntian Deng[5], Shayne Longpre[6], Noah A. Smith[7,8], Beyza Ermis[1], Marzieh Fadaee[♦1], and Sara Hooker[♦1]

[1]Cohere Labs, [2]Cohere, [3]Princeton University, [4]Stanford University, [5]University of Waterloo, [6]Massachusetts Institute of Technology, [7]Allen Institute for Artificial Intelligence, [8]University of Washington

Chatbot Arena leaderboard

- **Small** fraction of the data has been publicly released (3% of battles)
- One benefit of not releasing the dataset:
  - Avoid contamination / overfitting to specific ChatBot Arena users
  - But private company sees **a lot of** data…
- Need for a transparent leaderboard

The Leaderboard Illusion

Shivalika Singh[*1], Yiyang Nan[1], Alex Wang[2], Daniel D'souza[1], Sayash Kapoor[3], Ahmet Üstün[1], Sanmi Koyejo[4], Yuntian Deng[5], Shayne Longpre[6], Noah A. Smith[7,8], Beyza Ermis[1], Marzieh Fadaee[♦1], and Sara Hooker[♦1]

[1]Cohere Labs, [2]Cohere, [3]Princeton University, [4]Stanford University, [5]University of Waterloo, [6]Massachusetts Institute of Technology, [7]Allen Institute for Artificial Intelligence, [8]University of Washington

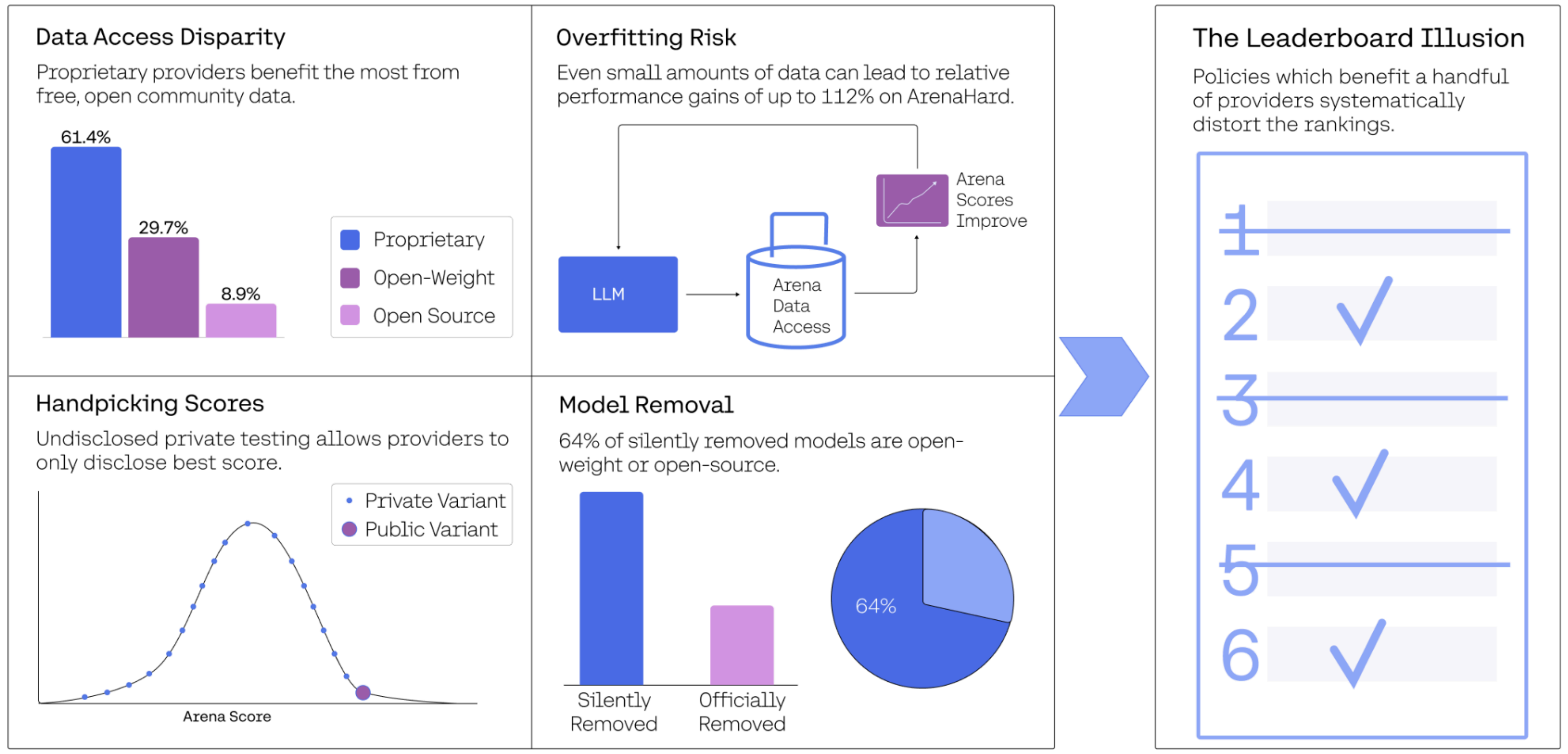


Figure 1: **Overview of key insights**. We investigate the prevalence of **undisclosed private testing and selective score reporting** on the Arena (Section 3), and highlight significant **data access disparities** between proprietary and open-source providers (Section 4.1). These disparities enable **overfitting to the Arena** (Section 4.2). Furthermore, **model deprecation practices** lack transparency, with many models silently deprecated without any notification to providers. We demonstrate how these deprecations contribute to unreliable rankings on the leaderboard (Section 5).

TabArena

Tabular leaderboard

- Live leaderboard for tabular methods: http://tabarena.ai
- Joint work with Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, Frank Hutter
- Very easy to host your leaderboard on Gradio/Hugging Face:
  - check out https://pypi.org/project/gradio-leaderboard/

#	Type	Model	Elo [⬆️]	Normalized Score [⬆️]	Rank [⬇️]	Median Train Time (s/1K) [⬇️]	Median Predict Time (s/
0	📊	AutoGluon 1.3 (4h)	1588	0.682	8.37	1408.78	3.333
1	🧠🔁	RealMLP (tuned + ensemble)	1566	0.638	9	6566.62	10.264
2	🌳	LightGBM (tuned + ensemble)	1529	0.583	10.43	417.05	2.639
3	🧠🔁	TabM (tuned + ensemble)	1527	0.592	10.44	38348.6	18.194
4	🌳	CatBoost (tuned + ensemble)	1485	0.555	12.29	1658.43	0.653
5	🌳	CatBoost (tuned)	1470	0.545	12.86	1658.43	0.081
6	🌳	LightGBM (tuned)	1450	0.519	13.76	417.05	0.334
7	🌳	XGBoost (tuned + ensemble)	1436	0.502	14.27	693.49	1.689
8	🧠🔁	TabM (tuned)	1432	0.501	14.48	38348.6	2.038
9	🌳	CatBoost (default)	1429	0.508	14.7	6.83	0.08
10	🧠🔁	ModernNCA (tuned + ensemble)	1425	0.519	14.86	20604.6	62.202

Evaluation

Wrapping up

Evaluation

Wrapping up

- Evaluating LLM is hard:

Evaluation

Wrapping up

- Evaluating LLM is hard:
    - Many languages

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives
  - Cost can be high

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives
  - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives
  - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy
- Still early days, lot of low hanging fruits

Evaluation

Wrapping up

- Evaluating LLM is hard:
    - Many languages
    - Many objectives
    - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy
- Still early days, lot of low hanging fruits
    - How to achieve good tradeoffs between objectives

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives
  - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy
- Still early days, lot of low hanging fruits
  - How to achieve good tradeoffs between objectives
  - Optimal sampling strategies

Evaluation

Wrapping up

- Evaluating LLM is hard:
    - Many languages
    - Many objectives
    - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy
- Still early days, lot of low hanging fruits
    - How to achieve good tradeoffs between objectives
    - Optimal sampling strategies
    - Better ways to improve low-resource languages & provide less culturally biased benchmarks

Evaluation

Wrapping up

- Evaluating LLM is hard:
  - Many languages
  - Many objectives
  - Cost can be high
- Tradeoff must be made, for instance cost vs accuracy
- Still early days, lot of low hanging fruits
  - How to achieve good tradeoffs between objectives
  - Optimal sampling strategies
  - Better ways to improve low-resource languages & provide less culturally biased benchmarks
- Dire need of trusted high-quality leaderboards

LLM & AutoML perspectives

Automatic Model Selection

Automatic Model Selection

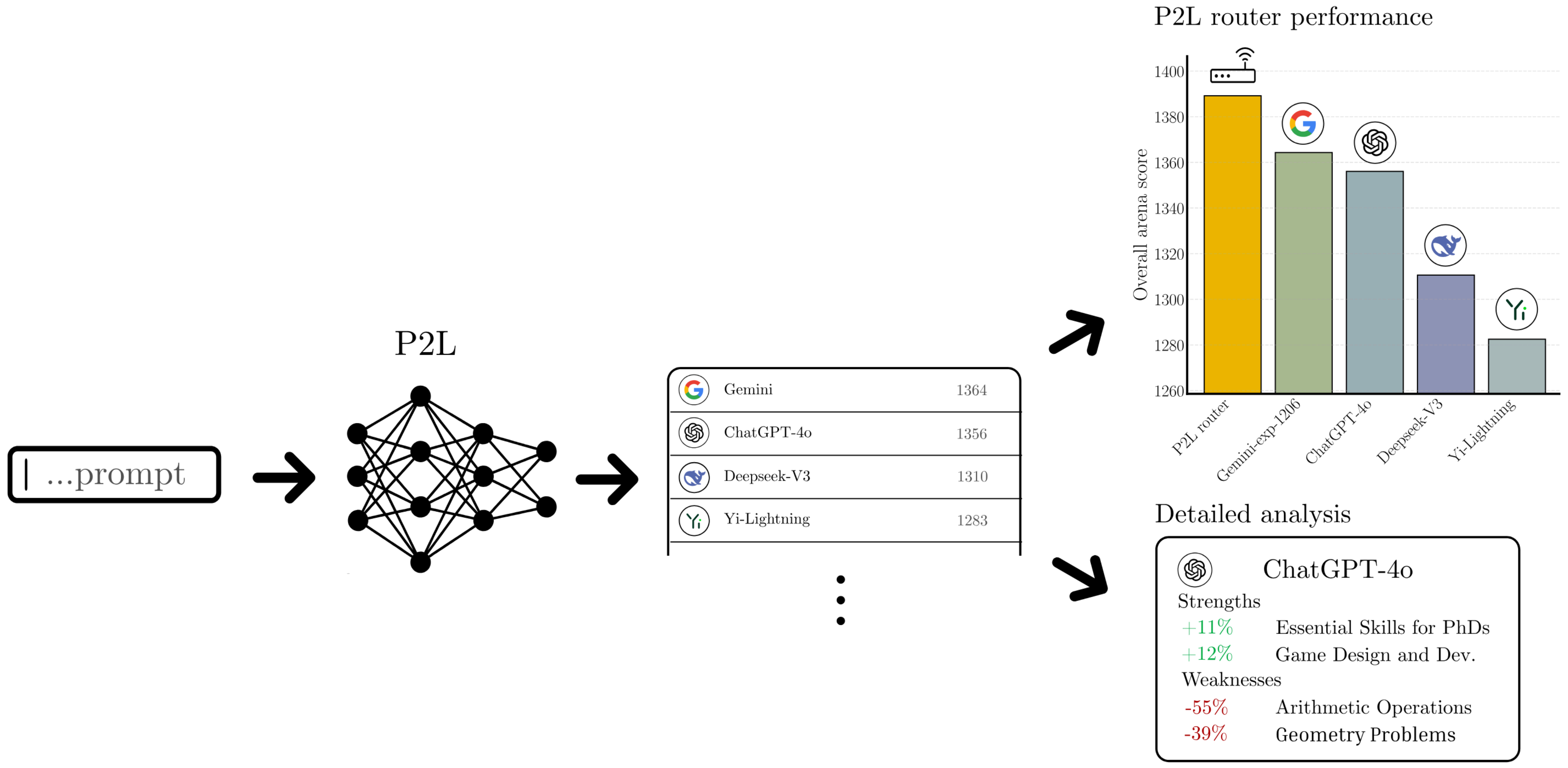
- Can we select among the 500+ models based on the prompt?

Automatic Model Selection

- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used

Automatic Model Selection

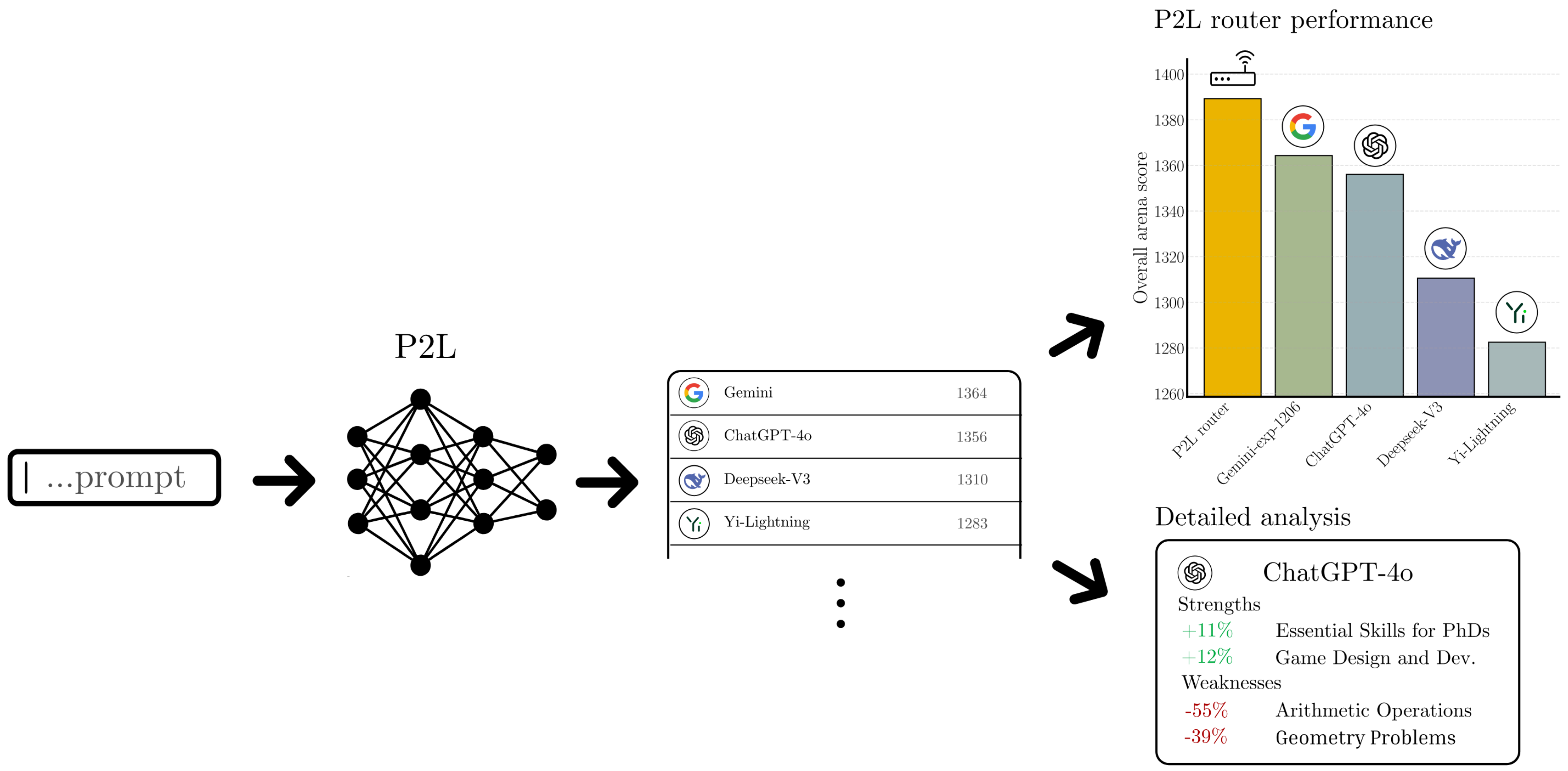
- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used



Prompt-to-Leaderboard. Arxiv 2025.

Automatic Model Selection

- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used

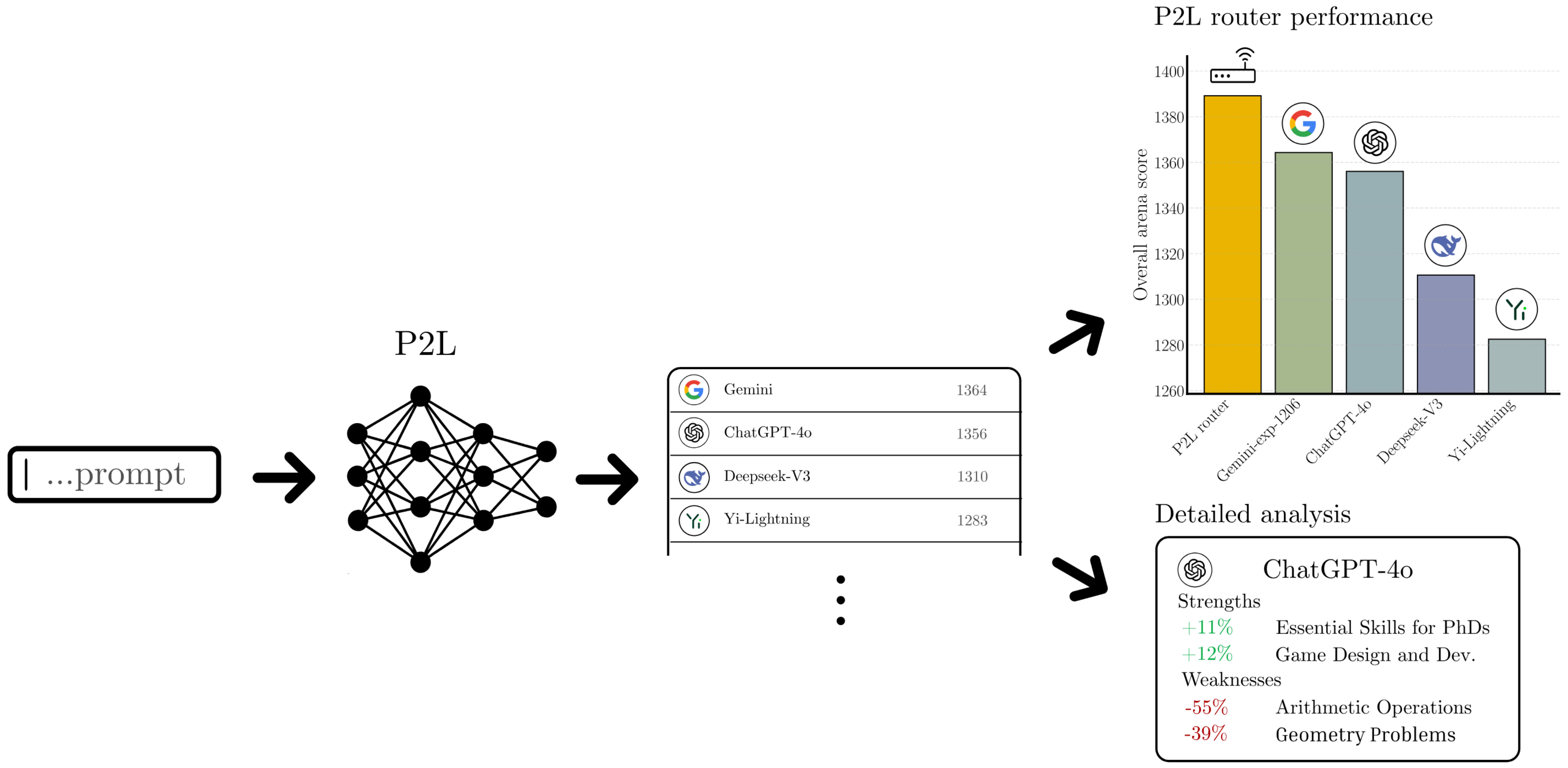


Prompt-to-Leaderboard. Arxiv 2025.

- Predict model ranks based on prompt features
- Got top model on CB arena
- Private dataset 😢

Automatic Model Selection

- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used
- Companies/services: Not Diamond startup + Open Router

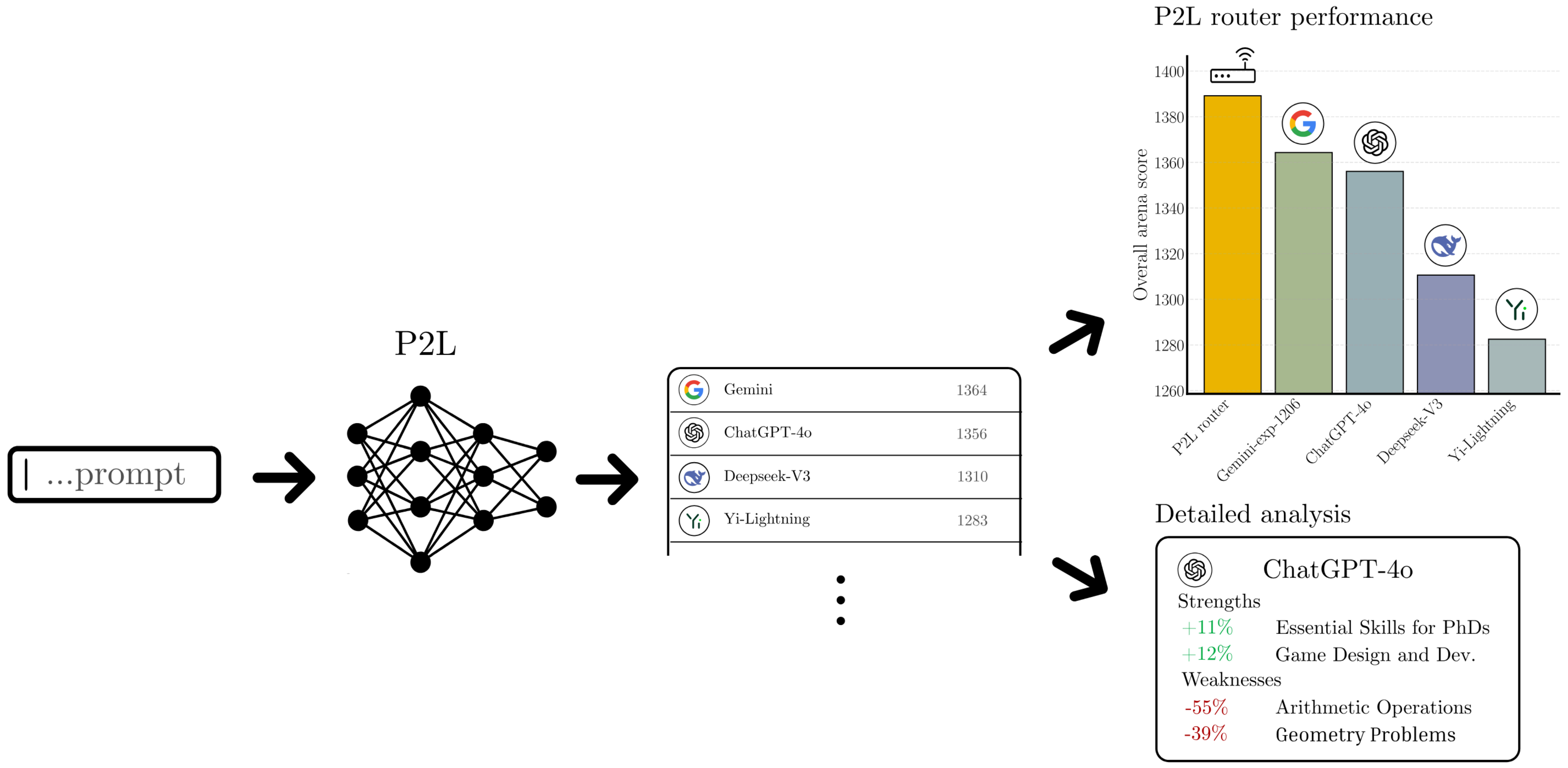


Prompt-to-Leaderboard. Arxiv 2025.

- Predict model ranks based on prompt features
- Got top model on CB arena
- Private dataset 😢

Automatic Model Selection

- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used
- Companies/services: Not Diamond startup + Open Router
- Easy connection with AutoML: transfer-learning/meta-learning/portfolio

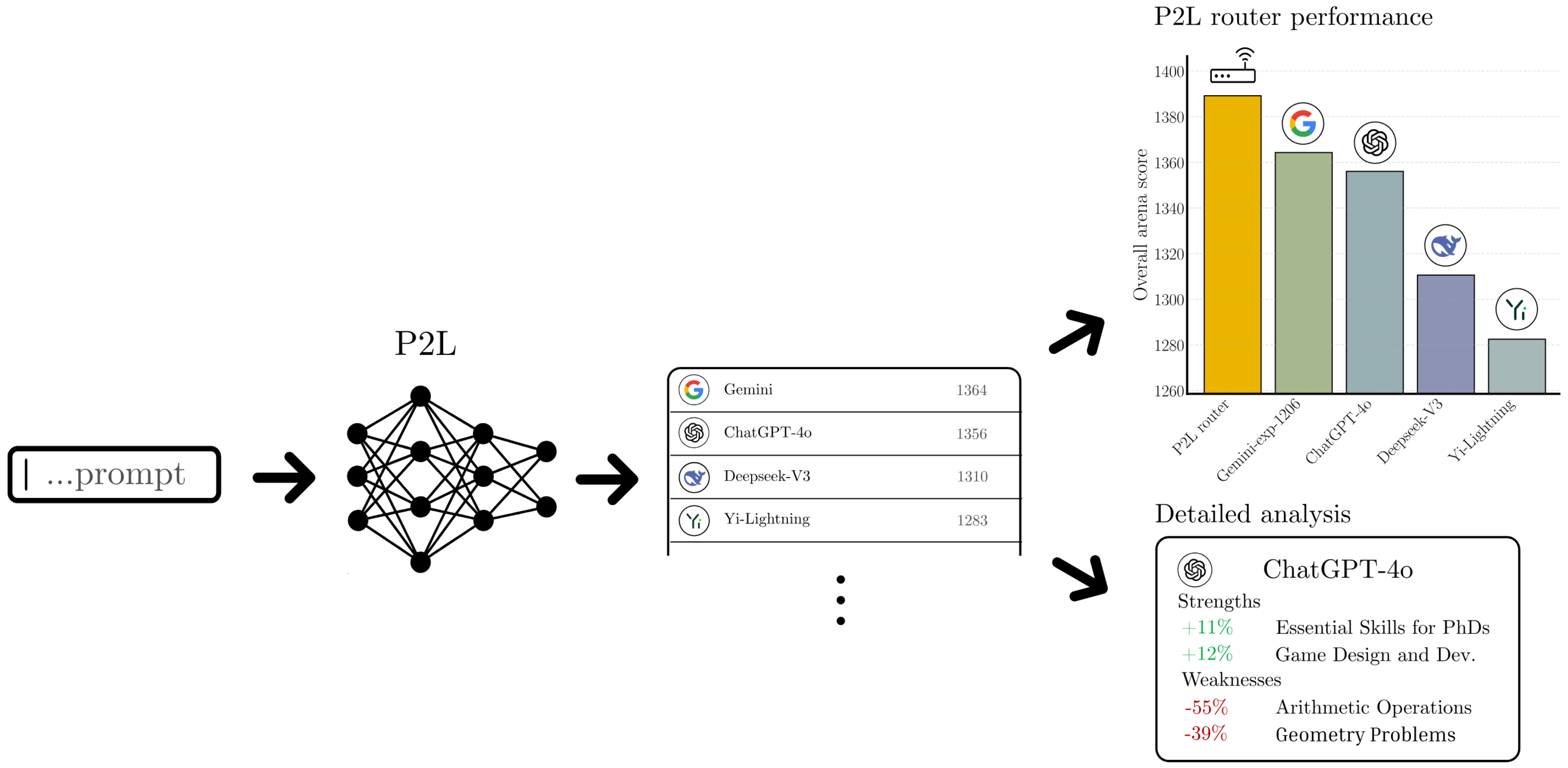


Prompt-to-Leaderboard. Arxiv 2025.

- Predict model ranks based on prompt features
- Got top model on CB arena
- Private dataset 😢

Automatic Model Selection

- Can we select among the 500+ models based on the prompt?
- Some model may be better on coding, for simple prompt (2+2?) simple/cheaper models can be used
- Companies/services: Not Diamond startup + Open Router
- Easy connection with AutoML: transfer-learning/meta-learning/portfolio
  - Low hanging fruit! 🍇



Prompt-to-Leaderboard. Arxiv 2025.

- Predict model ranks based on prompt features
- Got top model on CB arena
- Private dataset 😢

Instruction tuning quick recap

Instruction tuning quick recap

- SFT (self-supervised-training) step:

Instruction tuning quick recap

- SFT (self-supervised-training) step:
    - Humans annotate instructions and chat assistant possible answers

Instruction tuning quick recap

- SFT (self-supervised-training) step:
    - Humans annotate instructions and chat assistant possible answers
    - Model is trained on next token prediction on the generated dataset

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
  - Generate two different completions for each instruction

Instruction tuning quick recap

- SFT (self-supervised-training) step:
    - Humans annotate instructions and chat assistant possible answers
    - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
    - Generate two different completions for each instruction
    - Ask human to say which generation is better

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
  - Generate two different completions for each instruction
  - Ask human to say which generation is better

Instruction: write me a poem about Jazz.

Completion 1:

Jazz sucks, listen to rap, yo.

Completion 2:

In twilight's hush, where city streets

Do softly whisper secrets sweet,

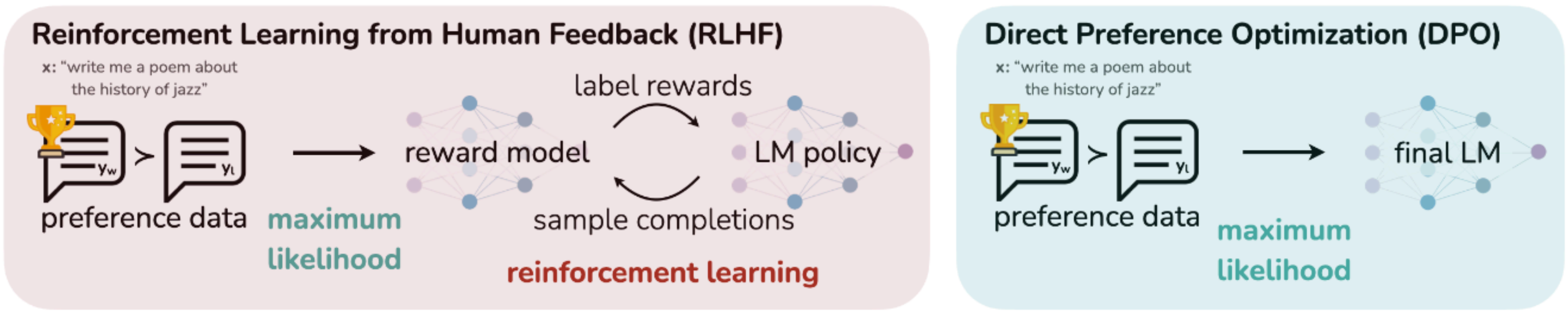
The jazz notes dance upon the air,

A improvisational flair

Human preference: 2

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
  - Generate two different completions for each instruction
  - Ask human to say which generation is better



DPO [Rafailov 2023]

Instruction: write me a poem about Jazz.

Completion 1:

Jazz sucks, listen to rap, yo.

Completion 2:

In twilight's hush, where city streets

Do softly whisper secrets sweet,

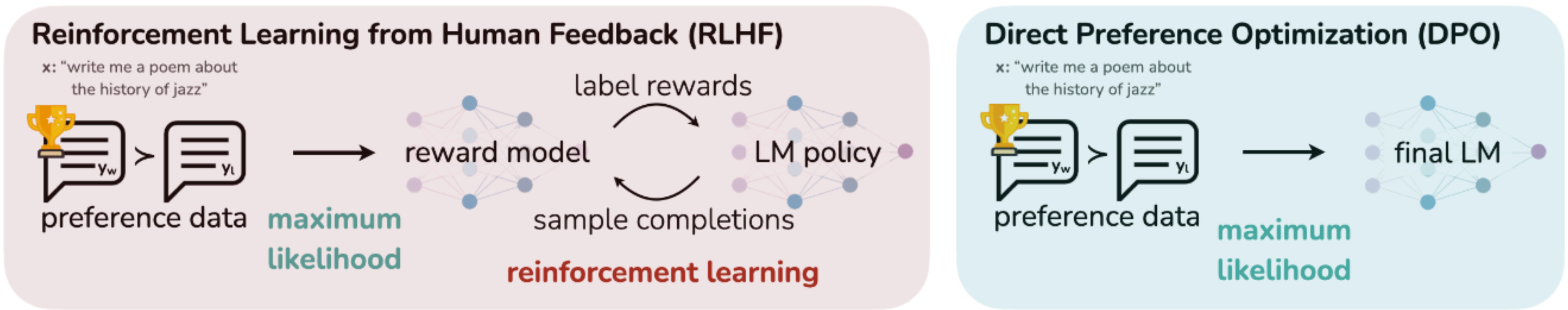
The jazz notes dance upon the air,

A improvisational flair

Human preference: 2

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
  - Generate two different completions for each instruction
  - Ask human to say which generation is better



DPO [Rafailov 2023]

- Online RL (PPO)
  - Learn reward model to predict user preference
  - Optimize model with PPO or other RL algorithm with the reward model
  - Regularize toward initial models

Instruction: write me a poem about Jazz.

Completion 1:

Jazz sucks, listen to rap, yo.

Completion 2:

In twilight's hush, where city streets

Do softly whisper secrets sweet,

The jazz notes dance upon the air,

A improvisational flair

Human preference: 2

Instruction tuning quick recap

- SFT (self-supervised-training) step:
    - Humans annotate instructions and chat assistant possible answers
    - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
    - Generate two different completions for each instruction
    - Ask human to say which generation is better

Instruction: write me a poem about Jazz.

Completion 1:

Jazz sucks, listen to rap, yo.

Completion 2:

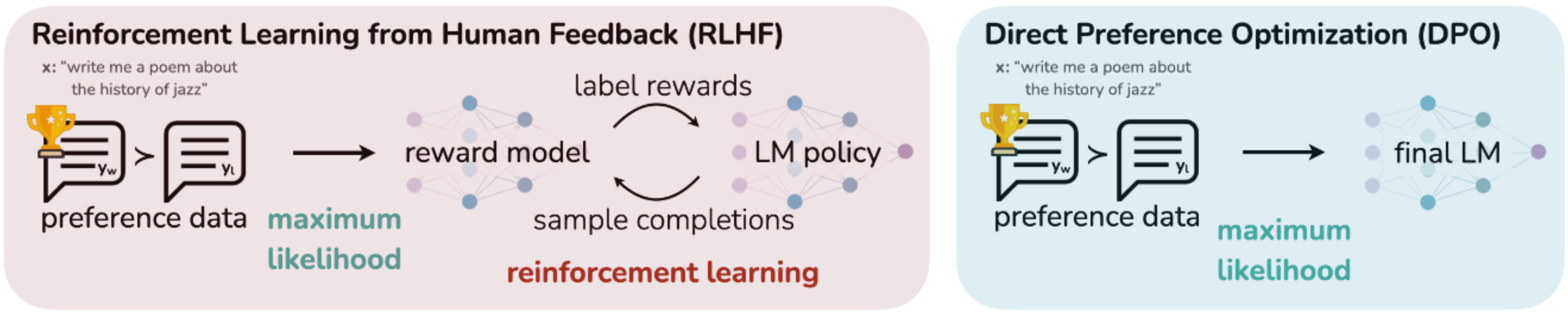
In twilight's hush, where city streets

Do softly whisper secrets sweet,

The jazz notes dance upon the air,

A improvisational flair

Human preference: 2



DPO [Rafailov 2023]

- Online RL (PPO)
    - Learn reward model to predict user preference
    - Optimize model with PPO or other RL algorithm with the reward model
    - Regularize toward initial models

- Offline RL (DPO)
    - Optimize model directly to predict human preferences
    - Also regularize toward initial model

Instruction tuning quick recap

- SFT (self-supervised-training) step:
  - Humans annotate instructions and chat assistant possible answers
  - Model is trained on next token prediction on the generated dataset
- Instruction tuning step:
  - Generate two different completions for each instruction
  - Ask human to say which generation is better

Instruction: write me a poem about Jazz.

Completion 1:

Jazz sucks, listen to rap, yo.

Completion 2:

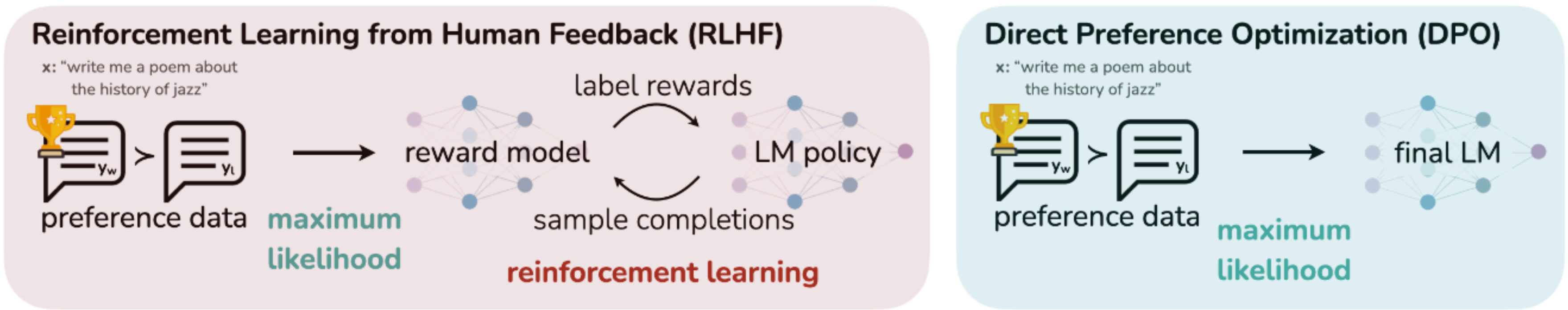
In twilight's hush, where city streets

Do softly whisper secrets sweet,

The jazz notes dance upon the air,

A improvisational flair

Human preference: 2



DPO [Rafailov 2023]

- Online RL (PPO)
  - Learn reward model to predict user preference
  - Optimize model with PPO or other RL algorithm with the reward model
  - Regularize toward initial models
- Offline RL (DPO)
  - Optimize model directly to predict human preferences
  - Also regularize toward initial model
- Others
  - Sample N completion and pick the one with highest reward from a reward models
  - SFT on best answer from annotators
  - …

Automatic Instruction Tuning

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
    - Base model to be used

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
    - Base model to be used
    - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
    - Base model to be used
    - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
    - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
    - (Q)Lora hyperparameters

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
    - Base model to be used
    - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
    - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
    - (Q)Lora hyperparameters
    - Dataset to be used

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
    - Base model to be used
    - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
    - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
    - (Q)Lora hyperparameters
    - Dataset to be used
- Fine tuning is the secret sauce

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena
- Best available open-source pipeline: Tulu3

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena
- Best available open-source pipeline: Tulu3
  - Data: Automatic data generation with synthetic users

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena
- Best available open-source pipeline: Tulu3
  - Data: Automatic data generation with synthetic users
  - LLM judges used to fine-tune model and discriminate good/bad answers

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena
- Best available open-source pipeline: Tulu3
  - Data: Automatic data generation with synthetic users
  - LLM judges used to fine-tune model and discriminate good/bad answers
  - Close model used for LLM judges

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

Automatic Instruction Tuning

- Instruction Tuning Pipelines have a **lot** of hyperparameters
  - Base model to be used
  - Optimiser usual suspects (optimizer algorithm, learning-rate, batch-size, …)
  - RLHF hyperparameters (RLHF method, regularisation wrt initial model)
  - (Q)Lora hyperparameters
  - Dataset to be used
- Fine tuning is the secret sauce
  - Best open-weight model: #6 in ChatBot Arena
  - Best open-weight model with open post-training pipeline: #46 in ChatBot Arena
  - Best fully open model: #153 in ChatBot Arena
- Key difficulties:
  - Best data is not public: Meta spend tens of millions $ on instruction-tuned annotated data for Llama
  - Human Evaluation is very expensive: $3500 to evaluate a single model in Chatbot Arena
- Best available open-source pipeline: Tulu3
  - Data: Automatic data generation with synthetic users
  - LLM judges used to fine-tune model and discriminate good/bad answers
  - Close model used for LLM judges
- Can we tune the LLM judges to be cheaper from open weights models??

🤔 Question for you: why cant we just apply AutoML to instruction tuning?

A case study: tuning LLM judges

Multiobjective and multi fidelity optimization

Preamble

Multiobjective & multifidelity optimization

Preamble

Multiobjective & multifidelity optimization

- Who is familiar with

Preamble

Multiobjective & multifidelity optimization

- Who is familiar with
  - multiobjective optimization? 🤔

Preamble

Multiobjective & multifidelity optimization

- Who is familiar with
  - multiobjective optimization? 🤔
  - multifidelity optimization? 🤔

Multiobjective optimization

Multiobjective optimization

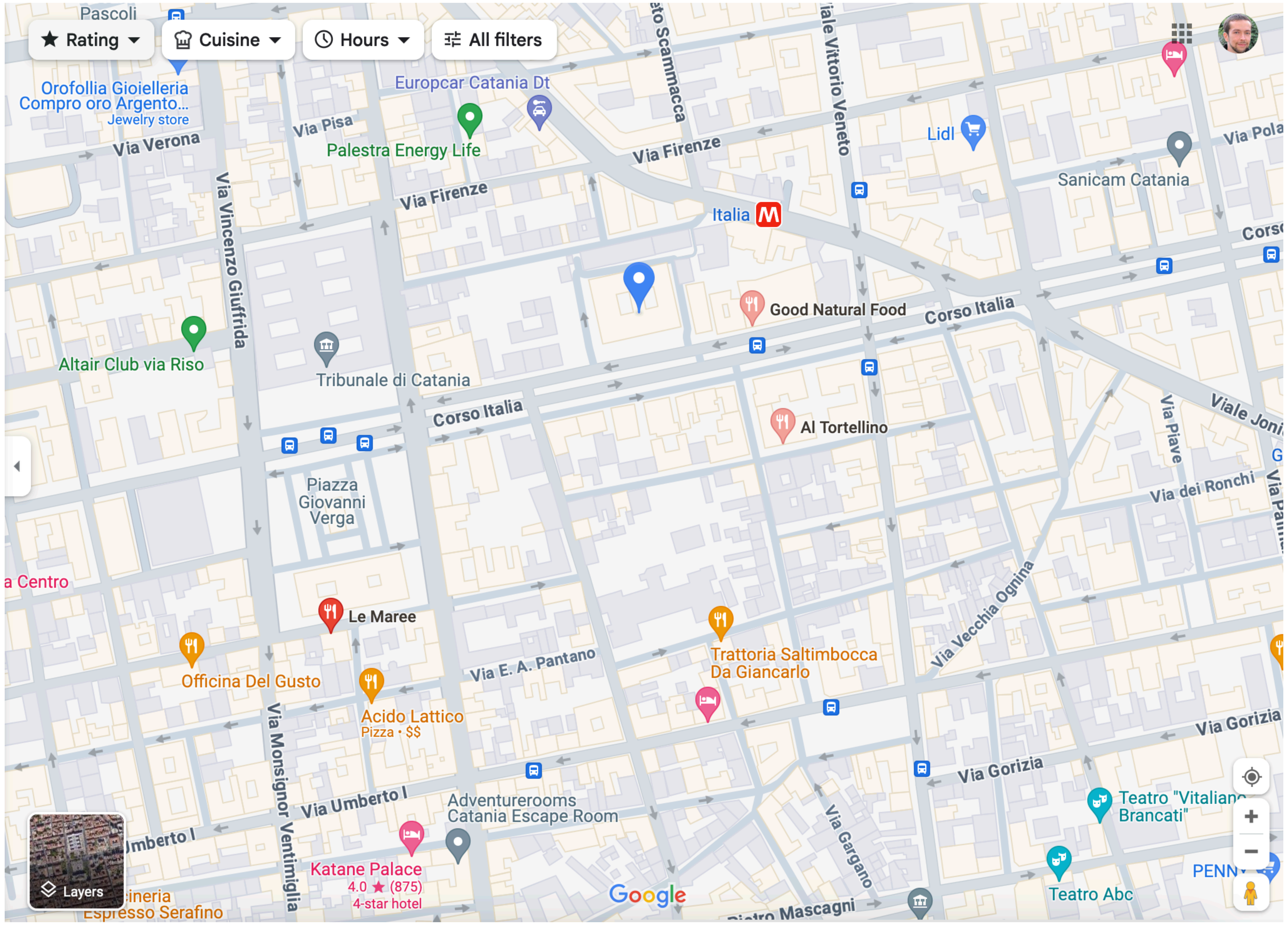
- Optimization is often about optimising a single objective but what if we have many?

Multiobjective optimization

- Optimization is often about optimising a single objective but what if we have many?
- Assume you want to pick a restaurant close to the Summer school…

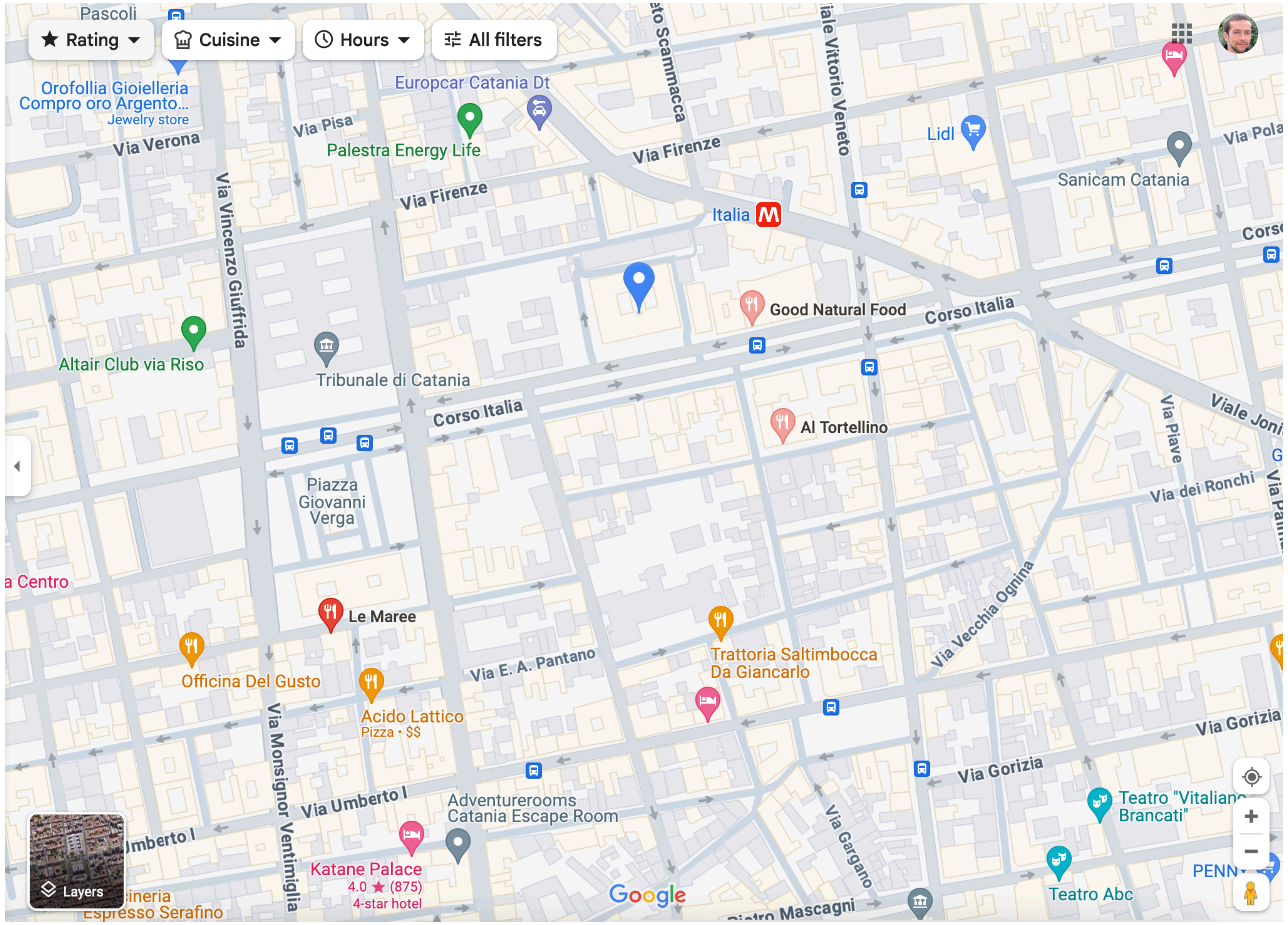
Multiobjective optimization

- Optimization is often about optimising a single objective but what if we have many?
- Assume you want to pick a restaurant close to the Summer school…



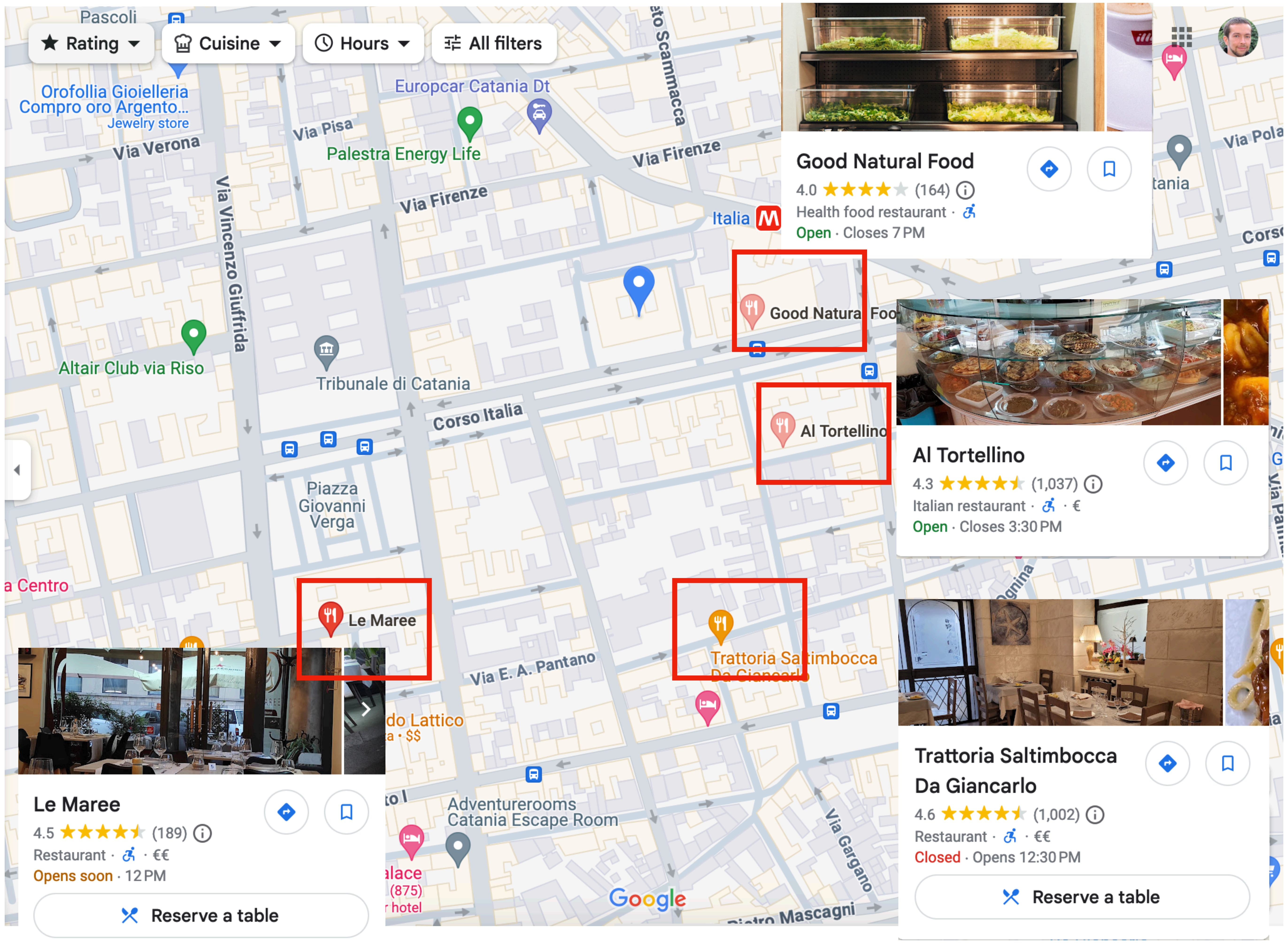
Multiobjective optimization

- Optimization is often about optimising a single objective but what if we have many?
- Assume you want to pick a restaurant close to the Summer school…
- And you only care about rating **and** distance



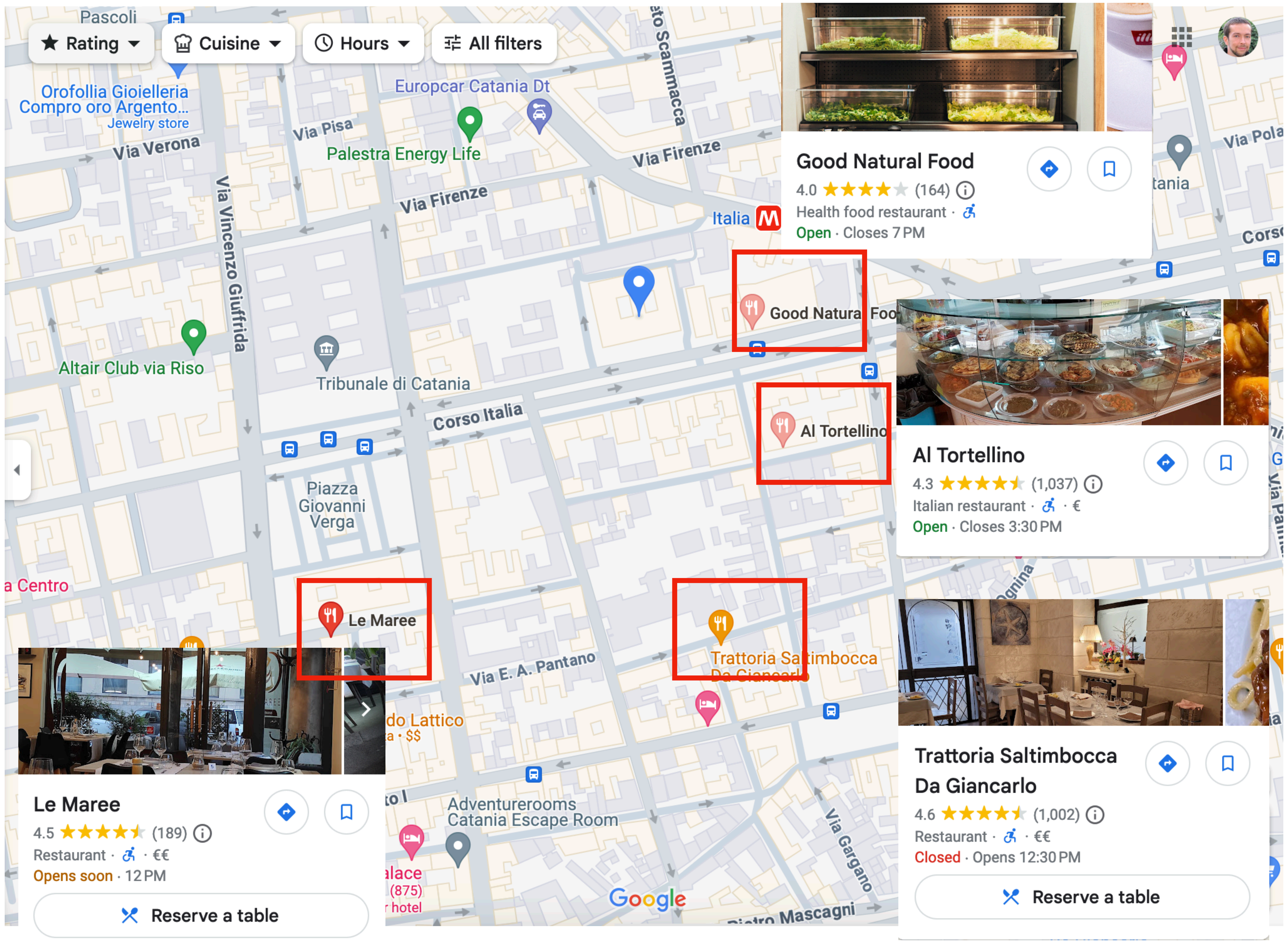
Multiobjective optimization

- Optimization is often about optimising a single objective but what if we have many?
- Assume you want to pick a restaurant close to the Summer school…
- And you only care about rating **and** distance



Multiobjective optimization

- Optimization is often about optimising a single objective but what if we have many?
- Assume you want to pick a restaurant close to the Summer school…
- And you only care about rating **and** distance



🤔 How to select which restaurant to go?

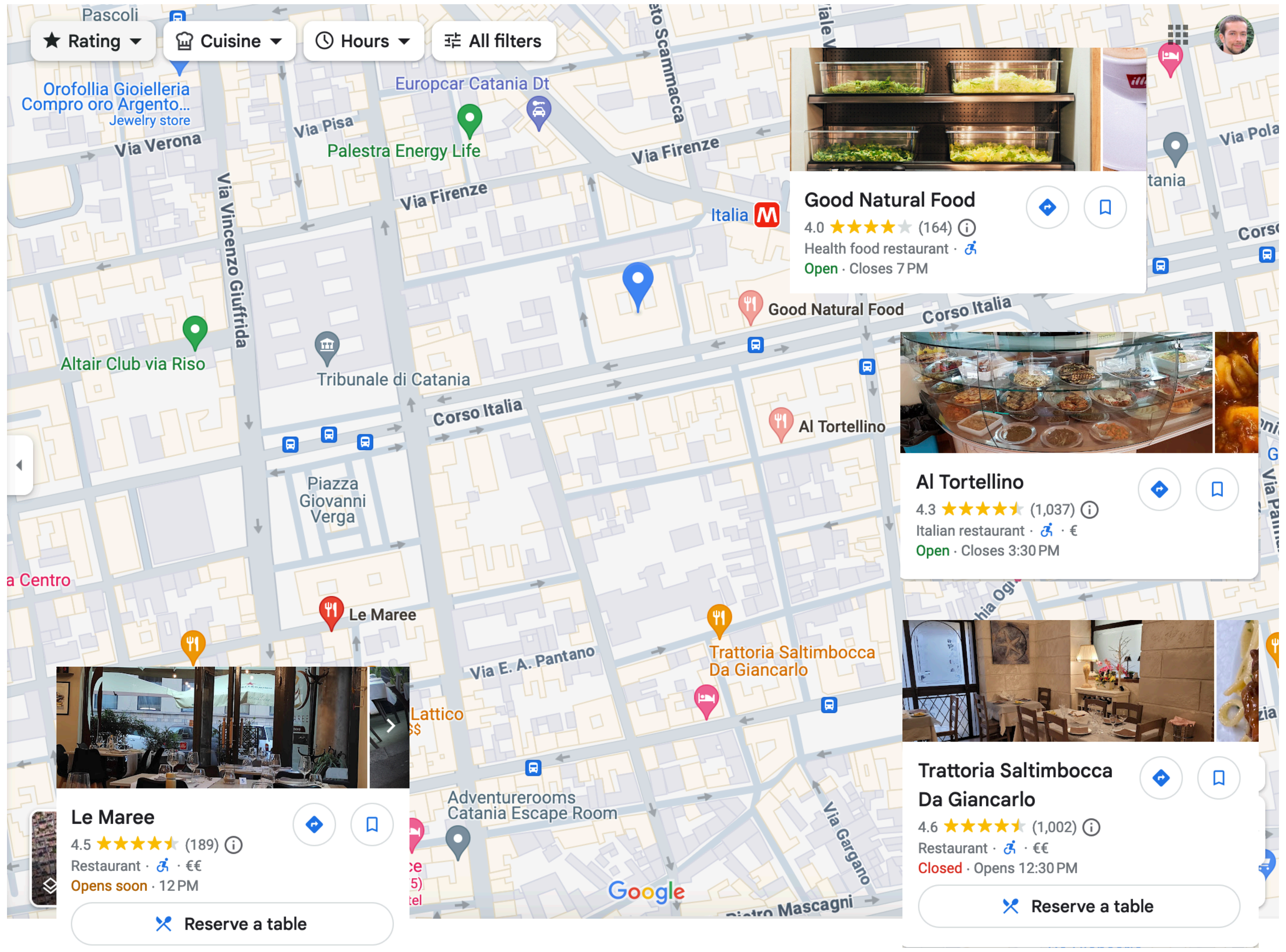
How to select which restaurant to go?

How to select which restaurant to go?

- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives

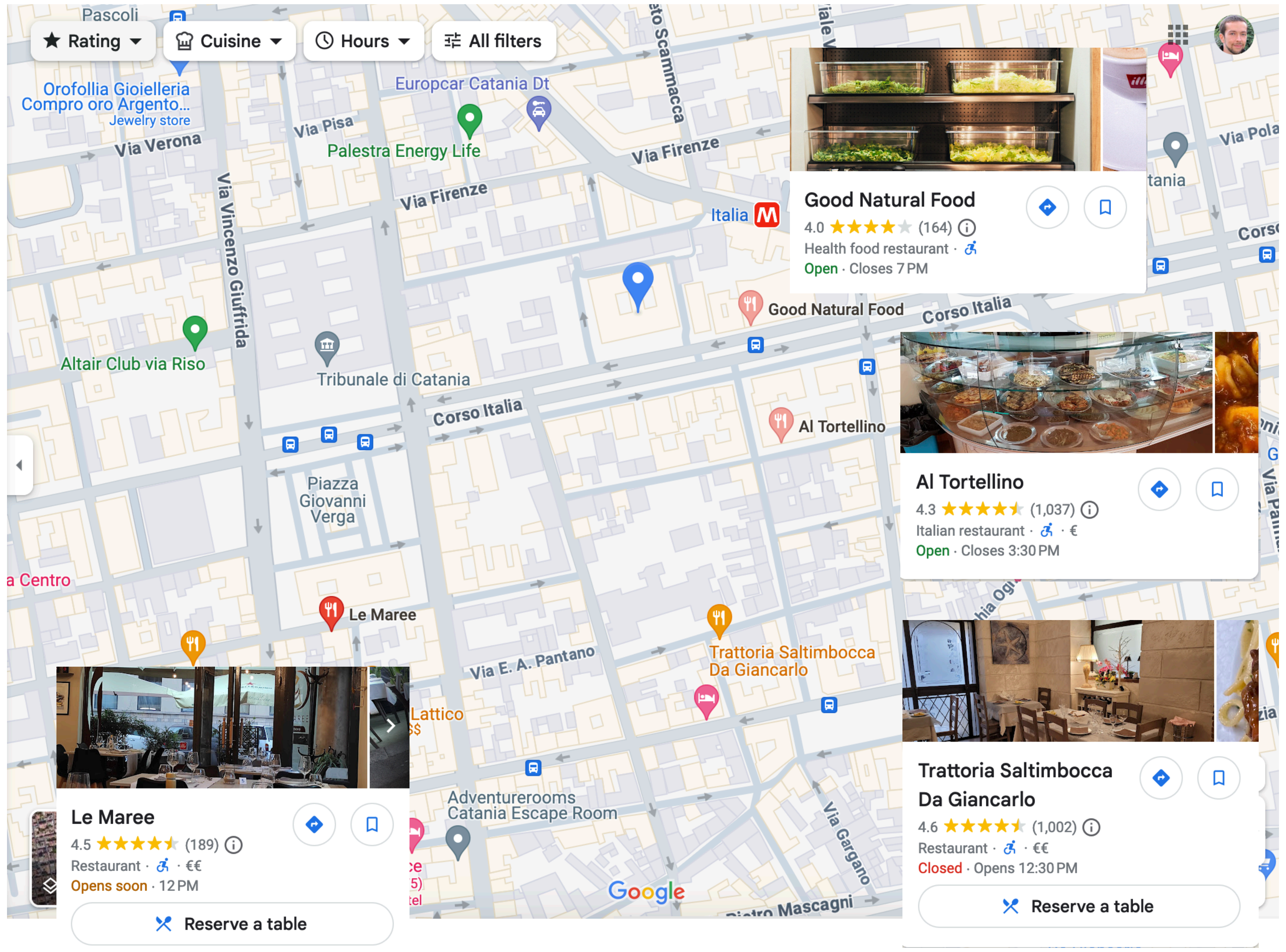
How to select which restaurant to go?

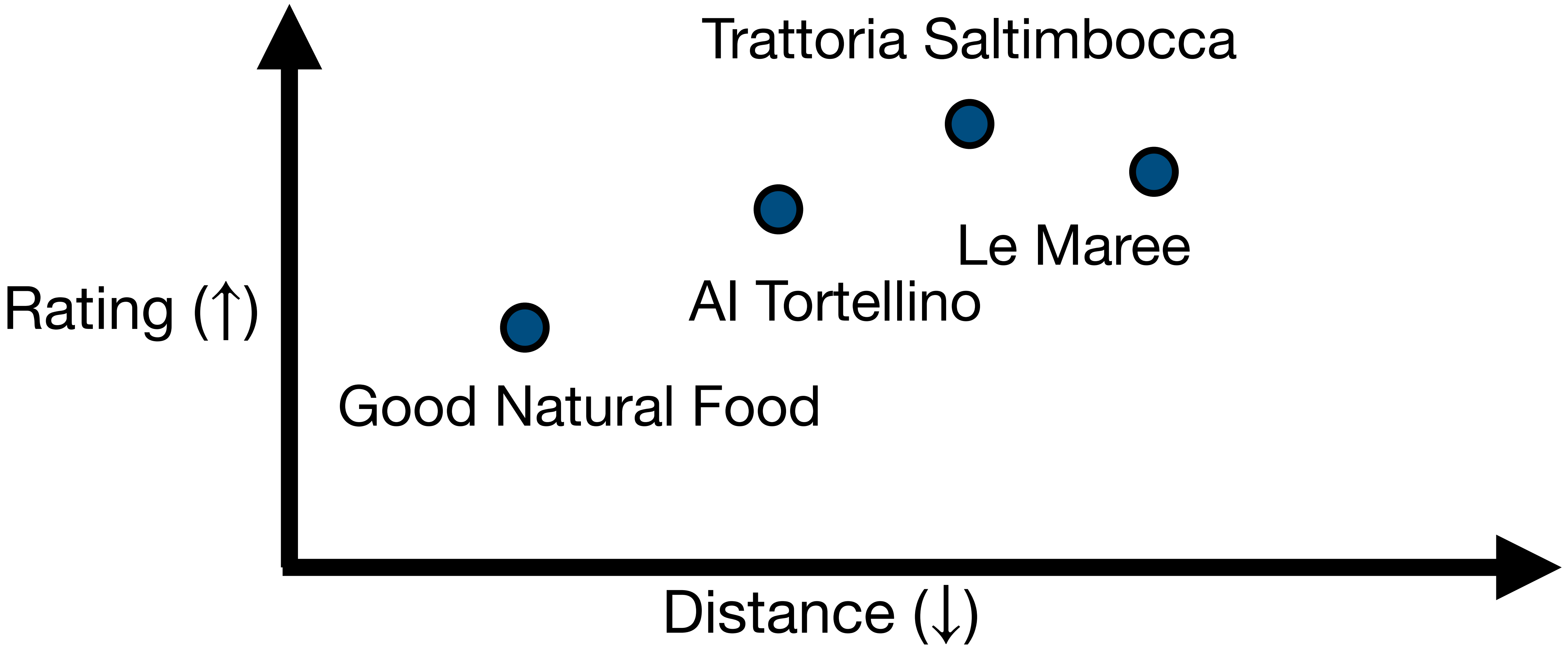
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives



How to select which restaurant to go?

- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives

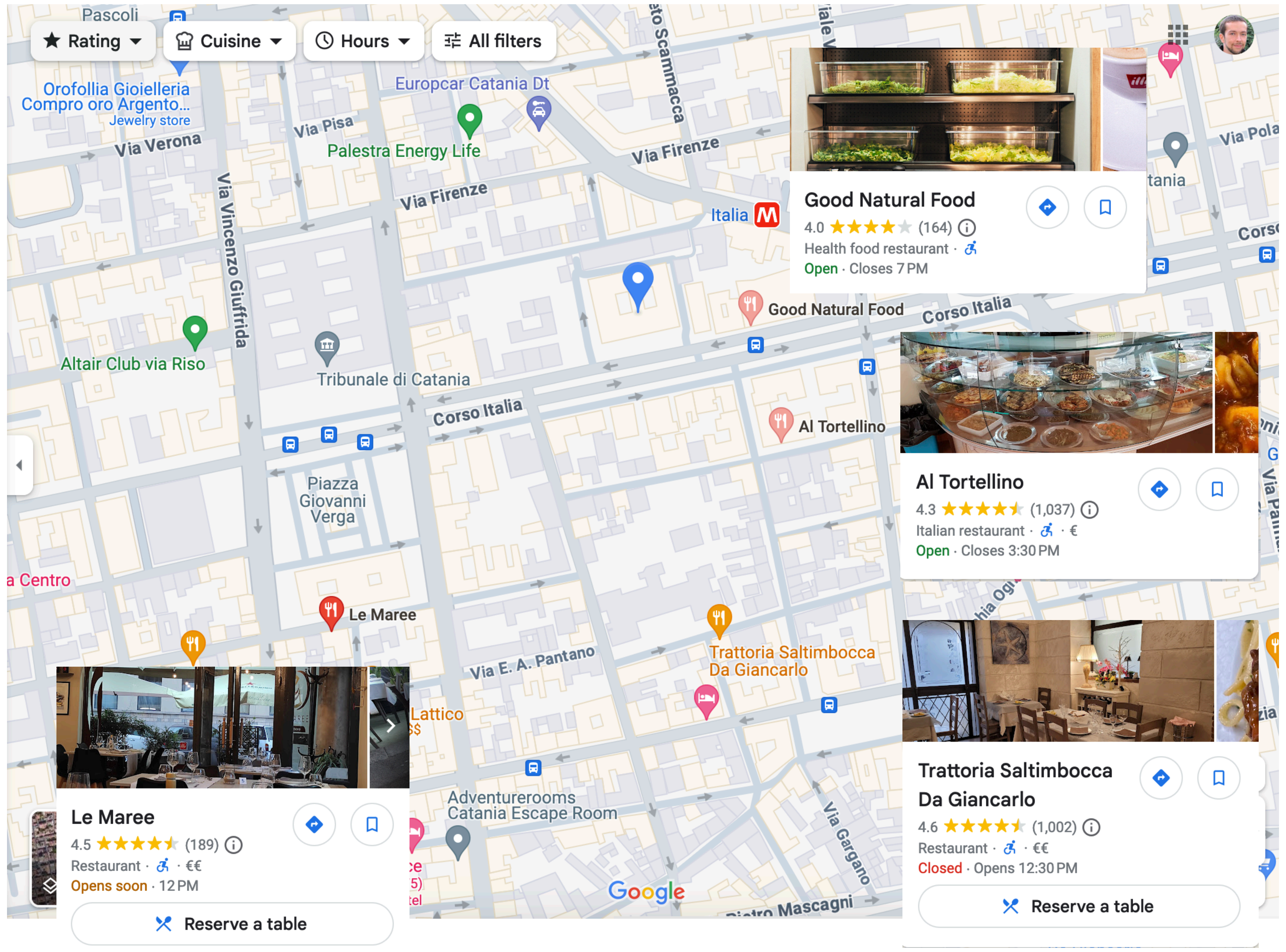


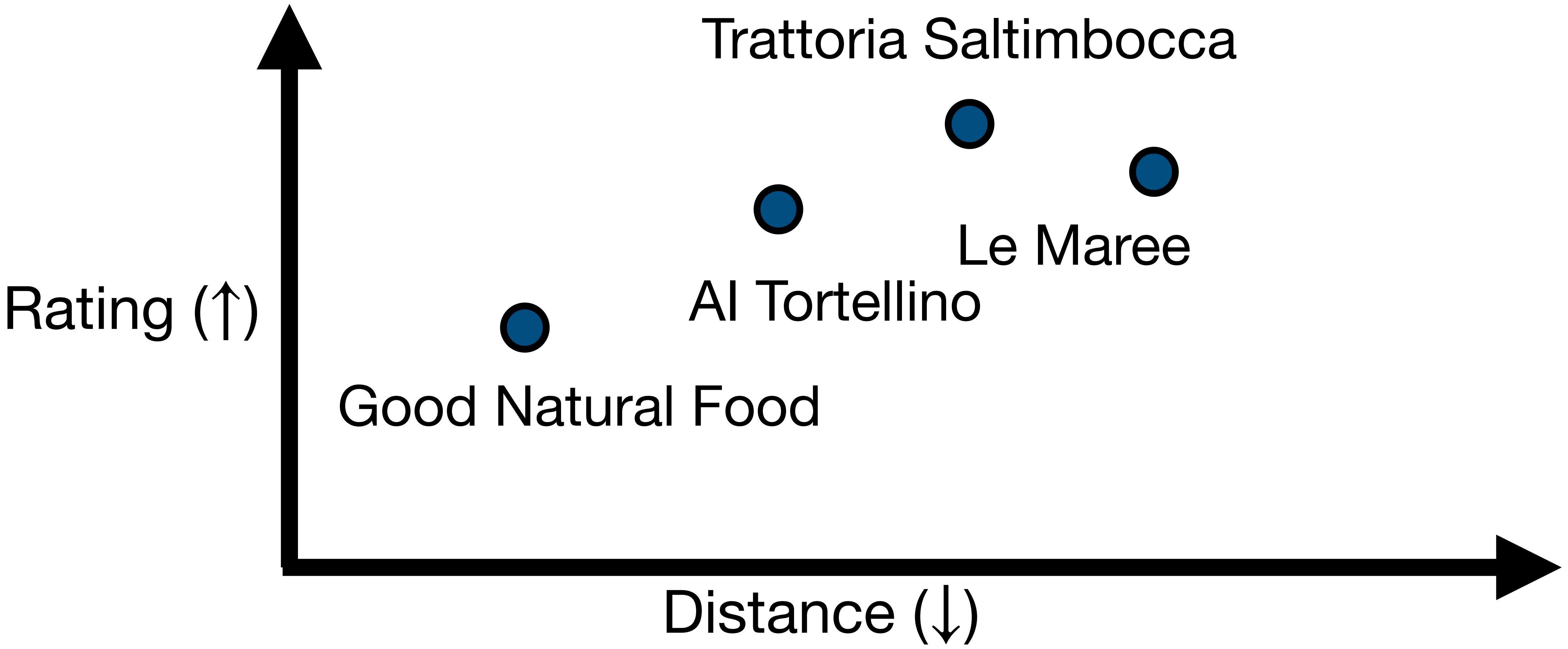


How to select which restaurant to go?

- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives

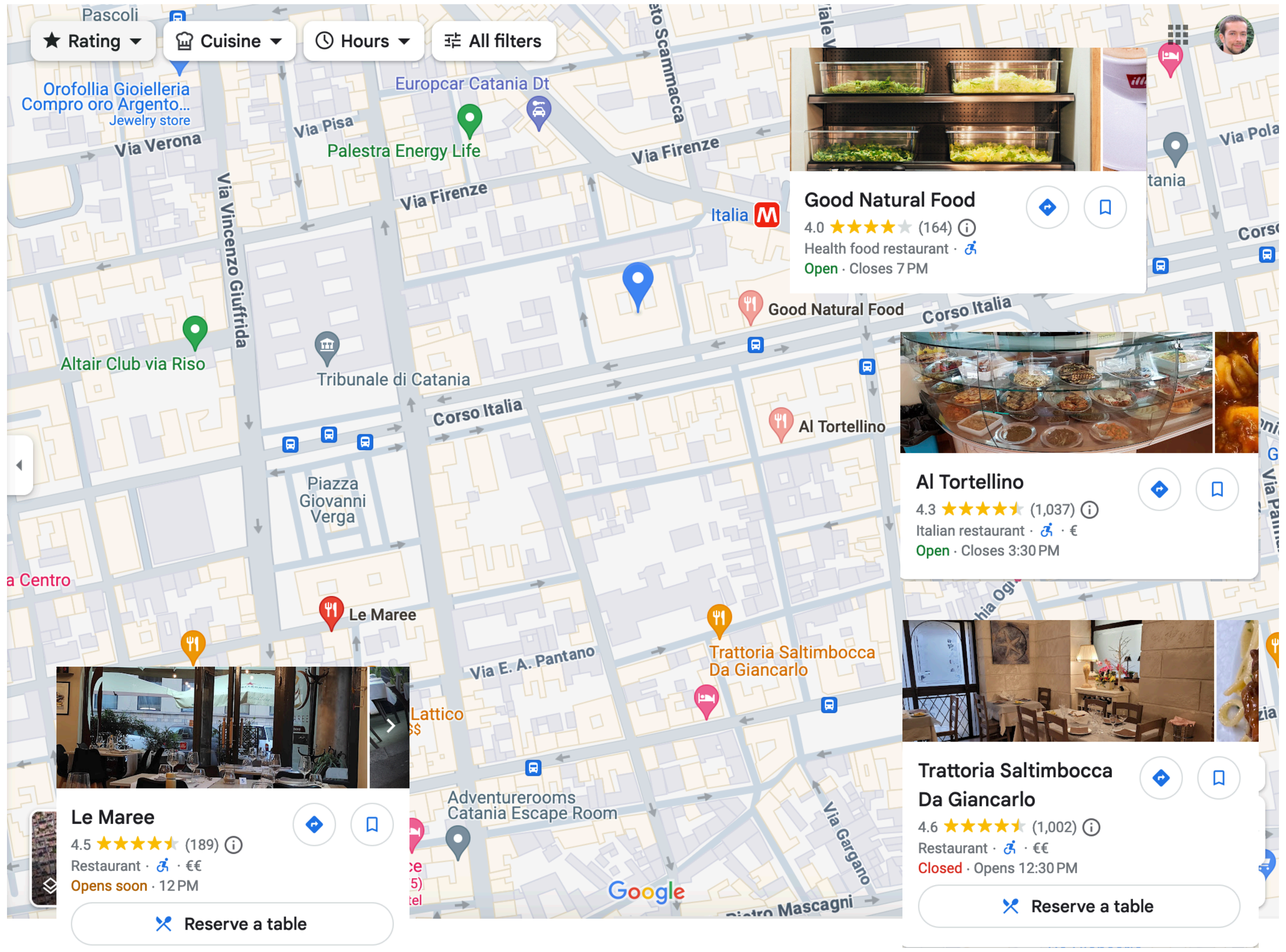
- No just a single best solution!

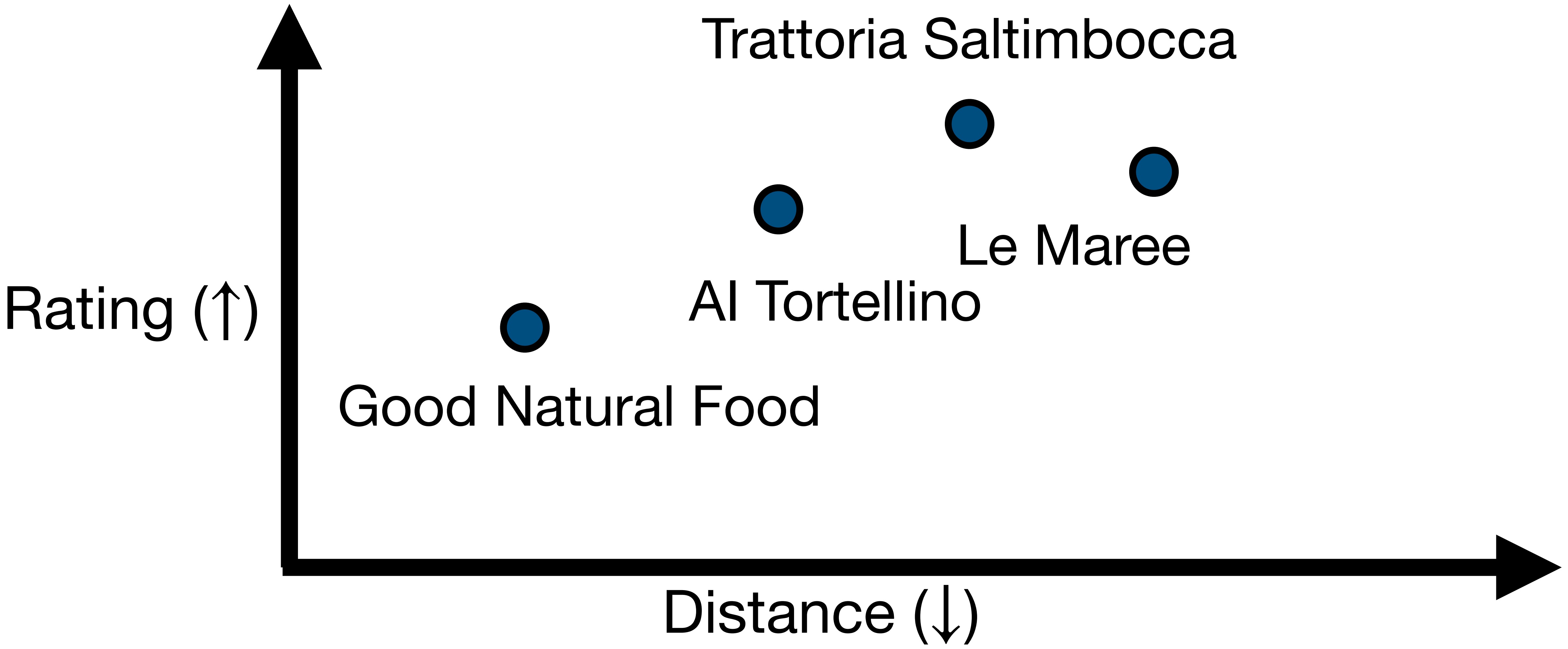




How to select which restaurant to go?

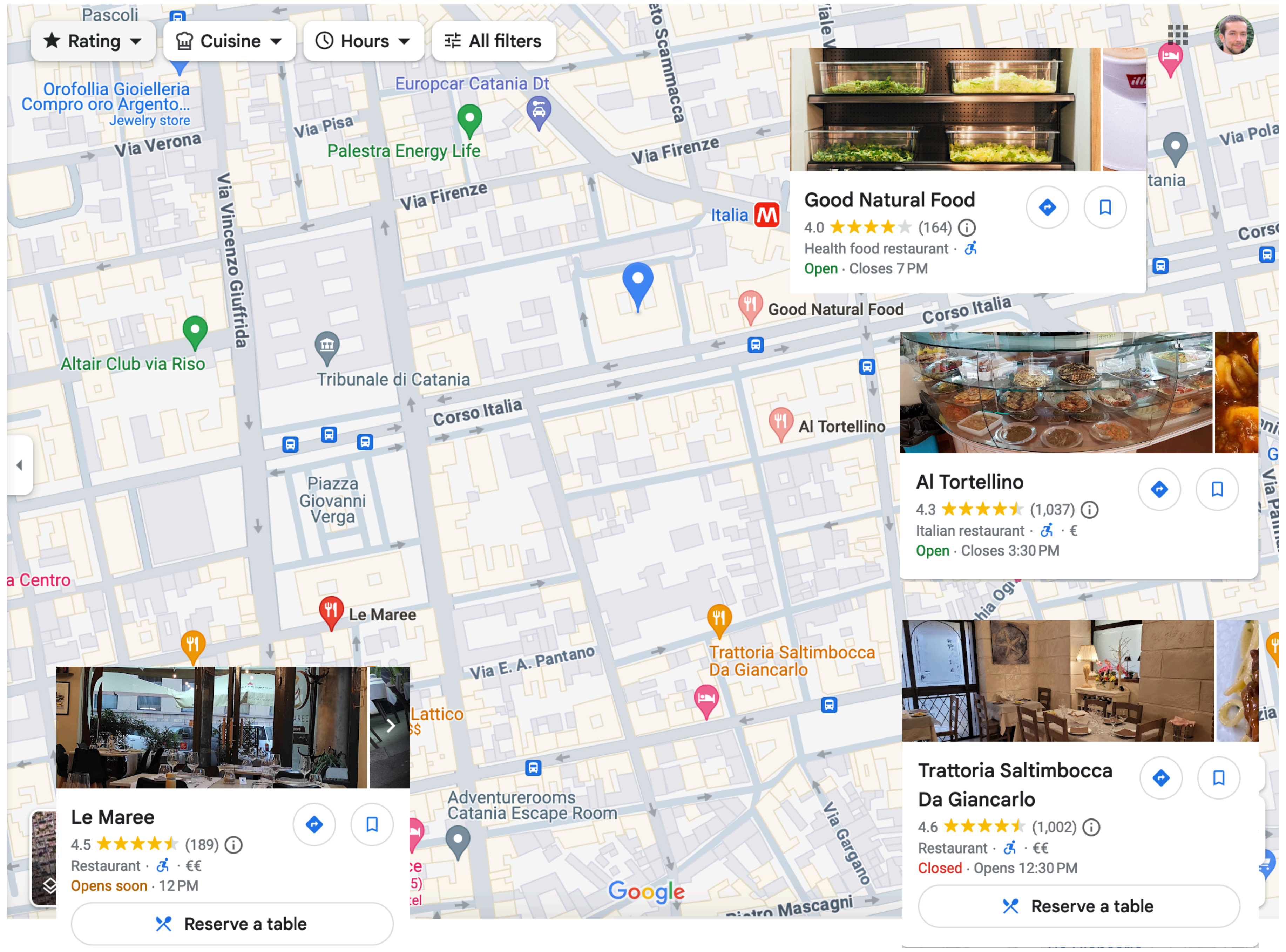
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?

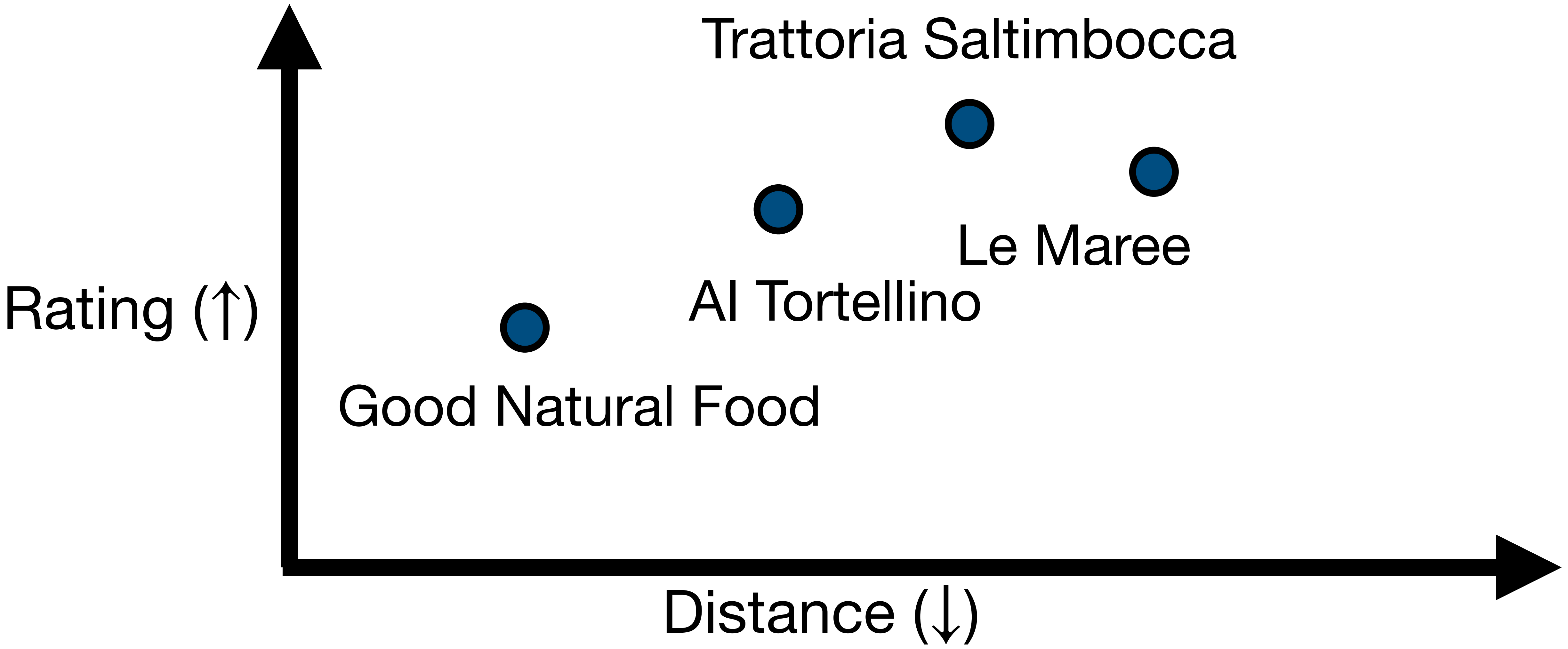




How to select which restaurant to go?

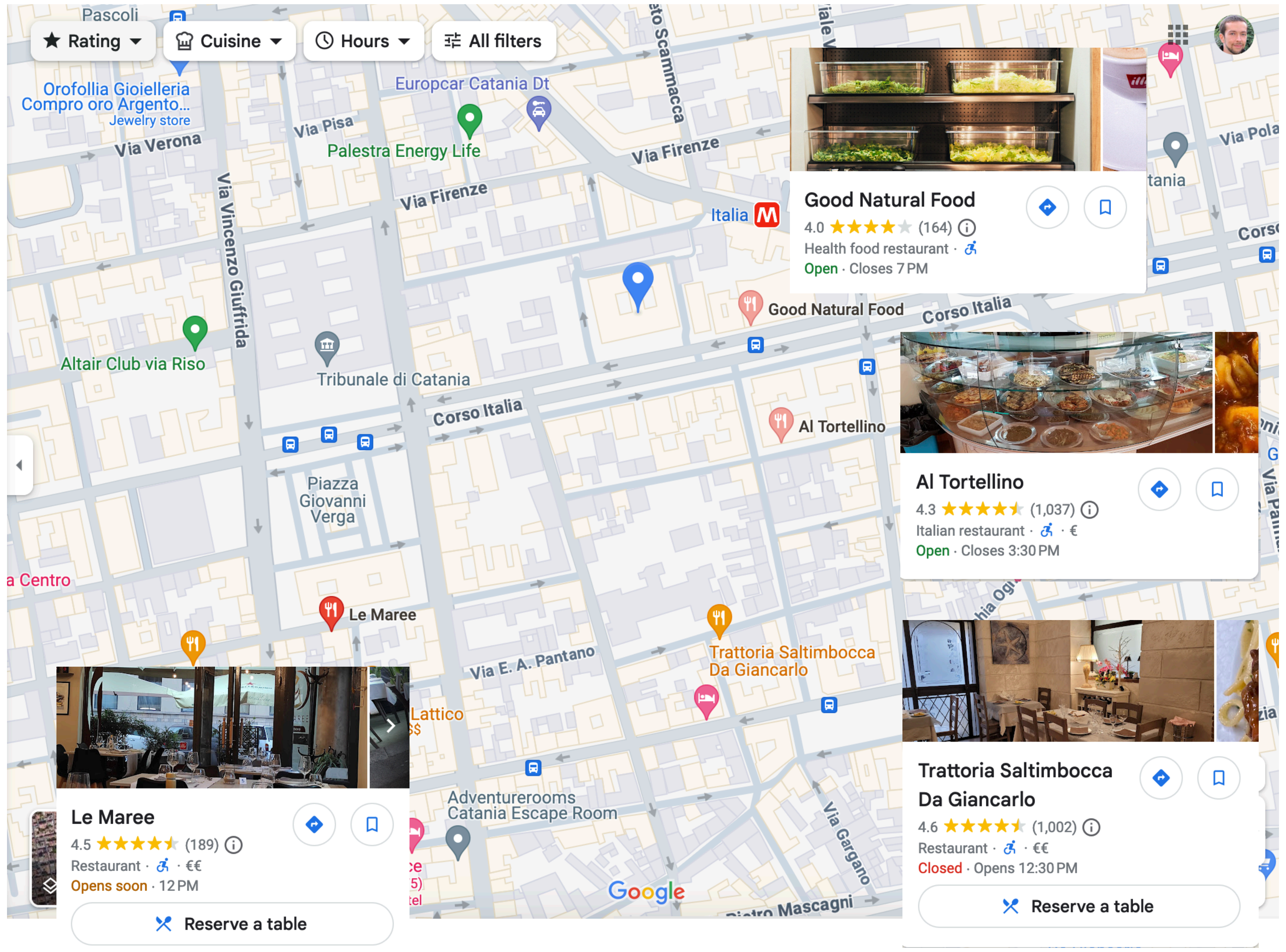
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal

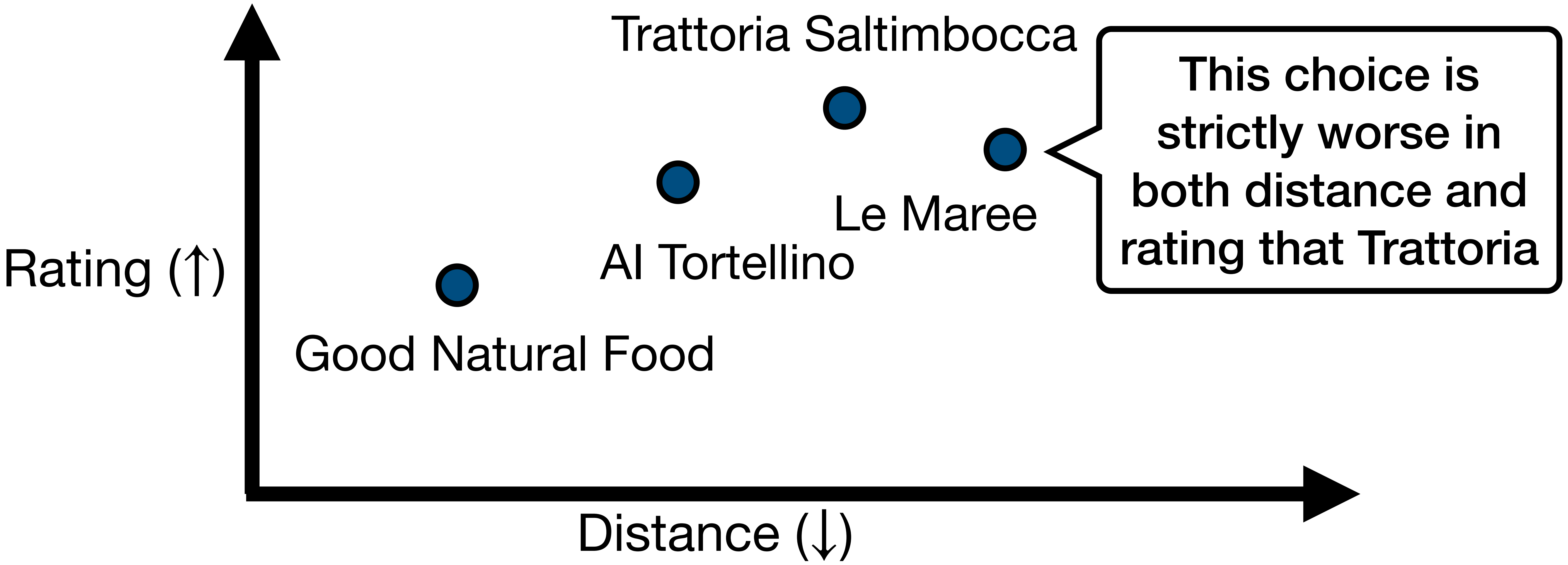




How to select which restaurant to go?

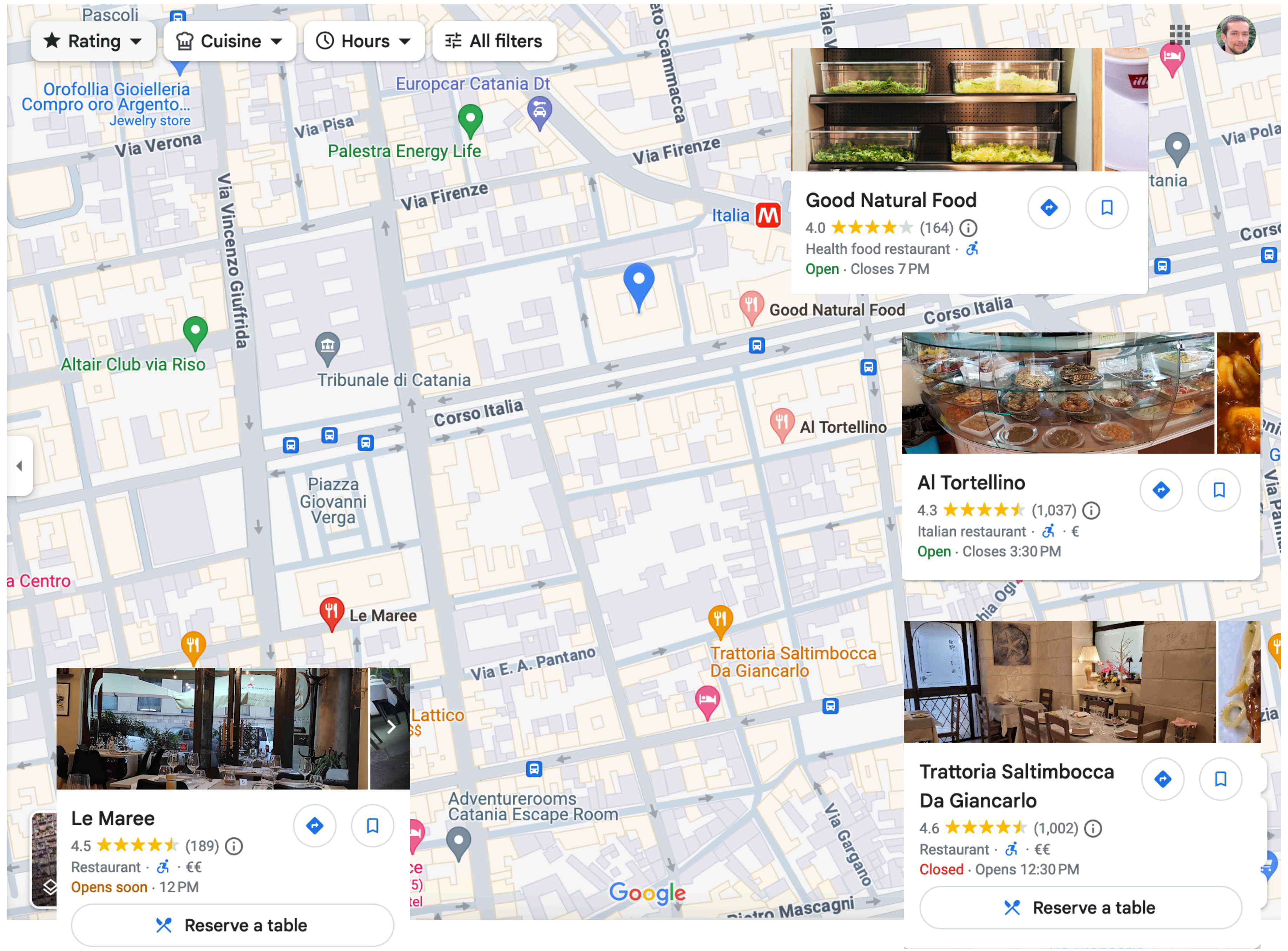
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal

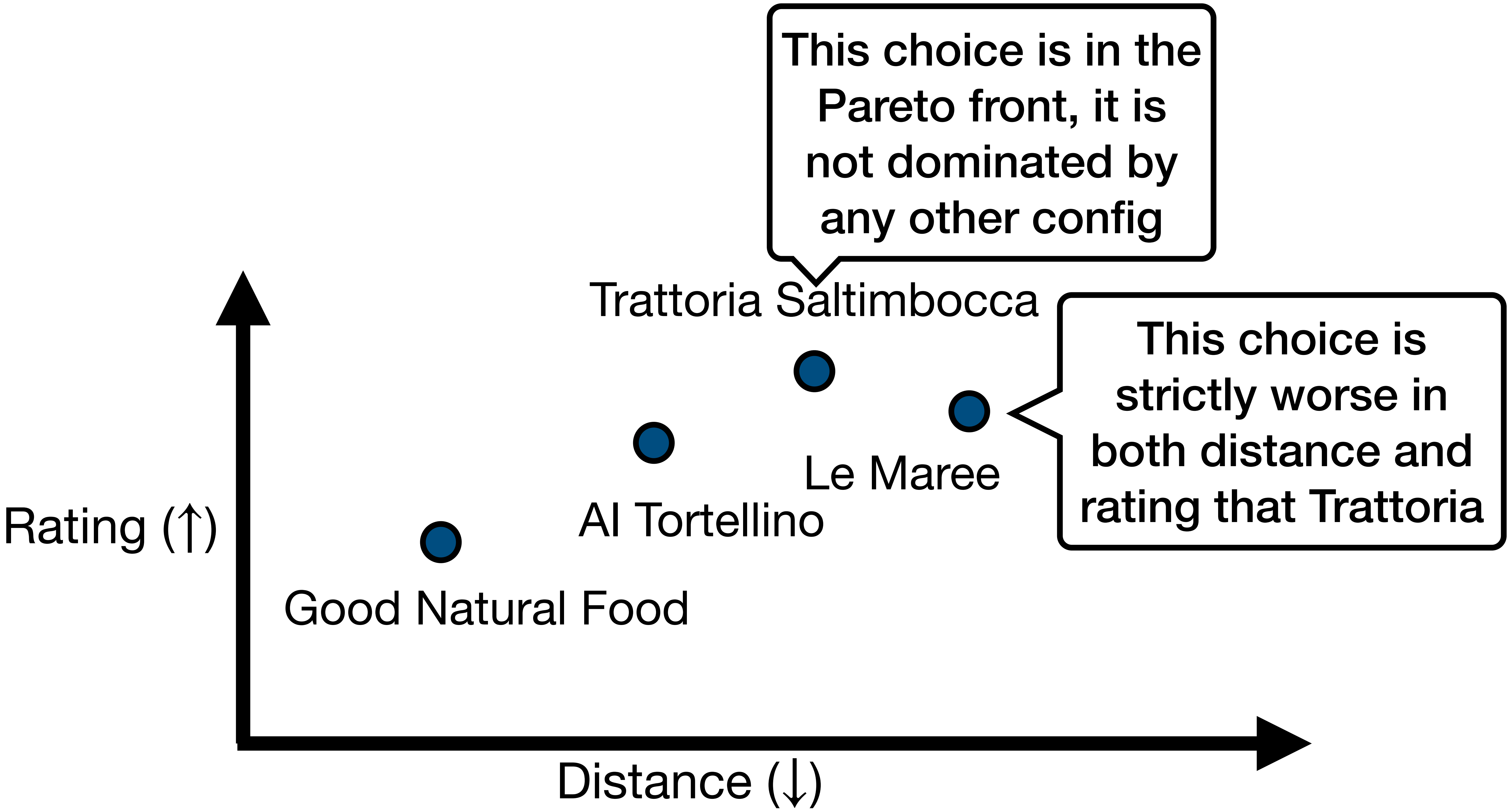




How to select which restaurant to go?

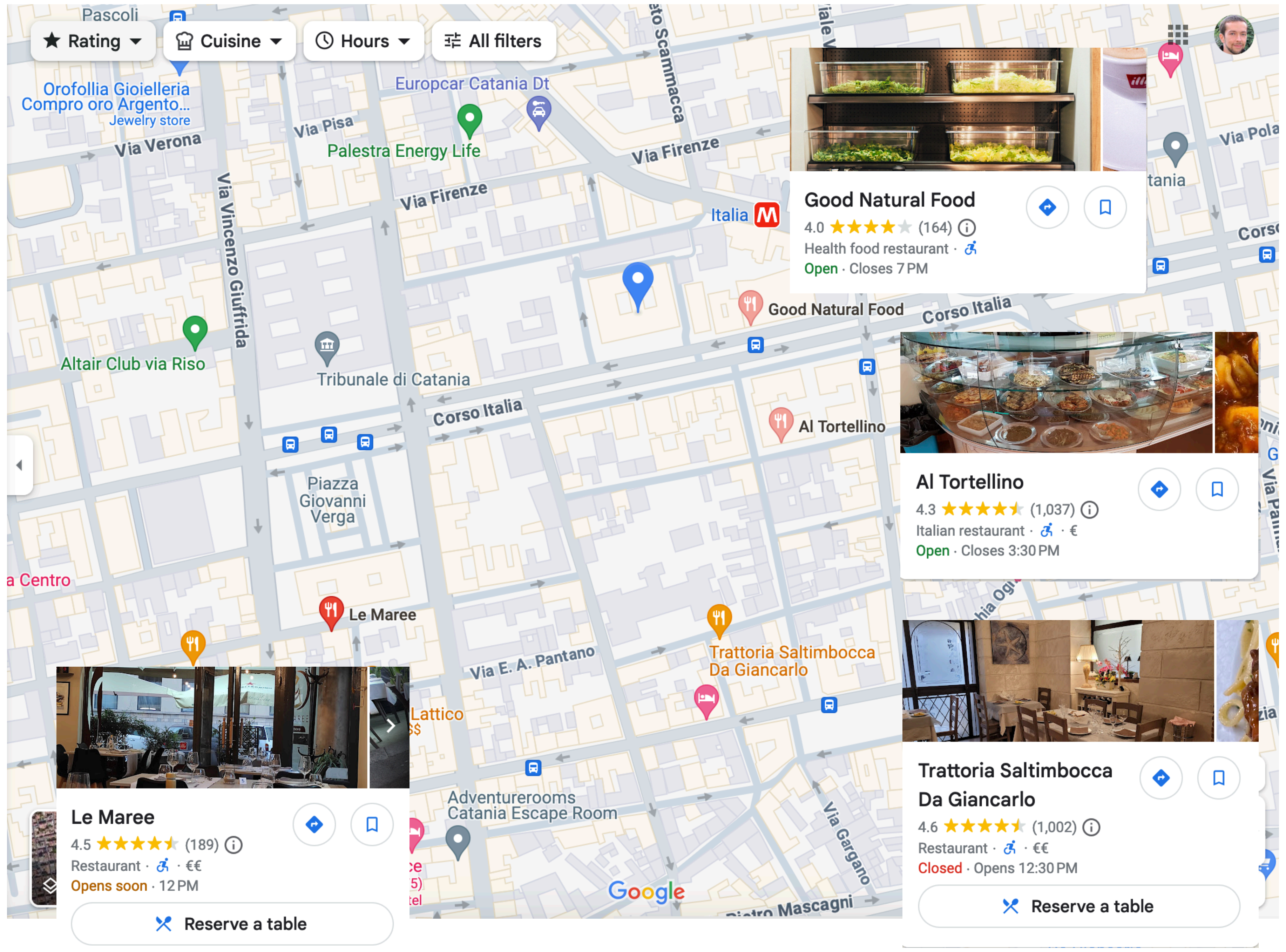
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal

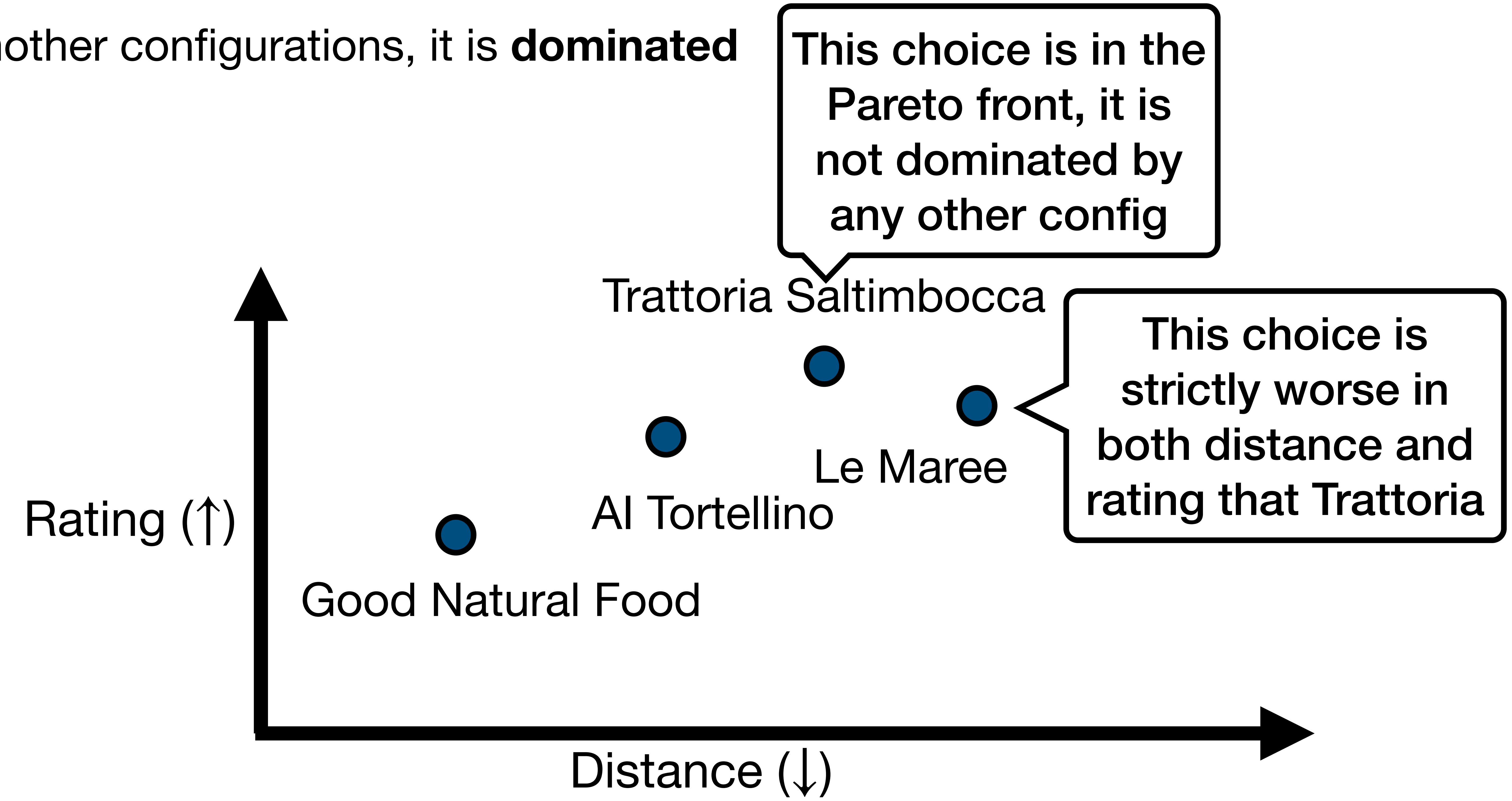




How to select which restaurant to go?

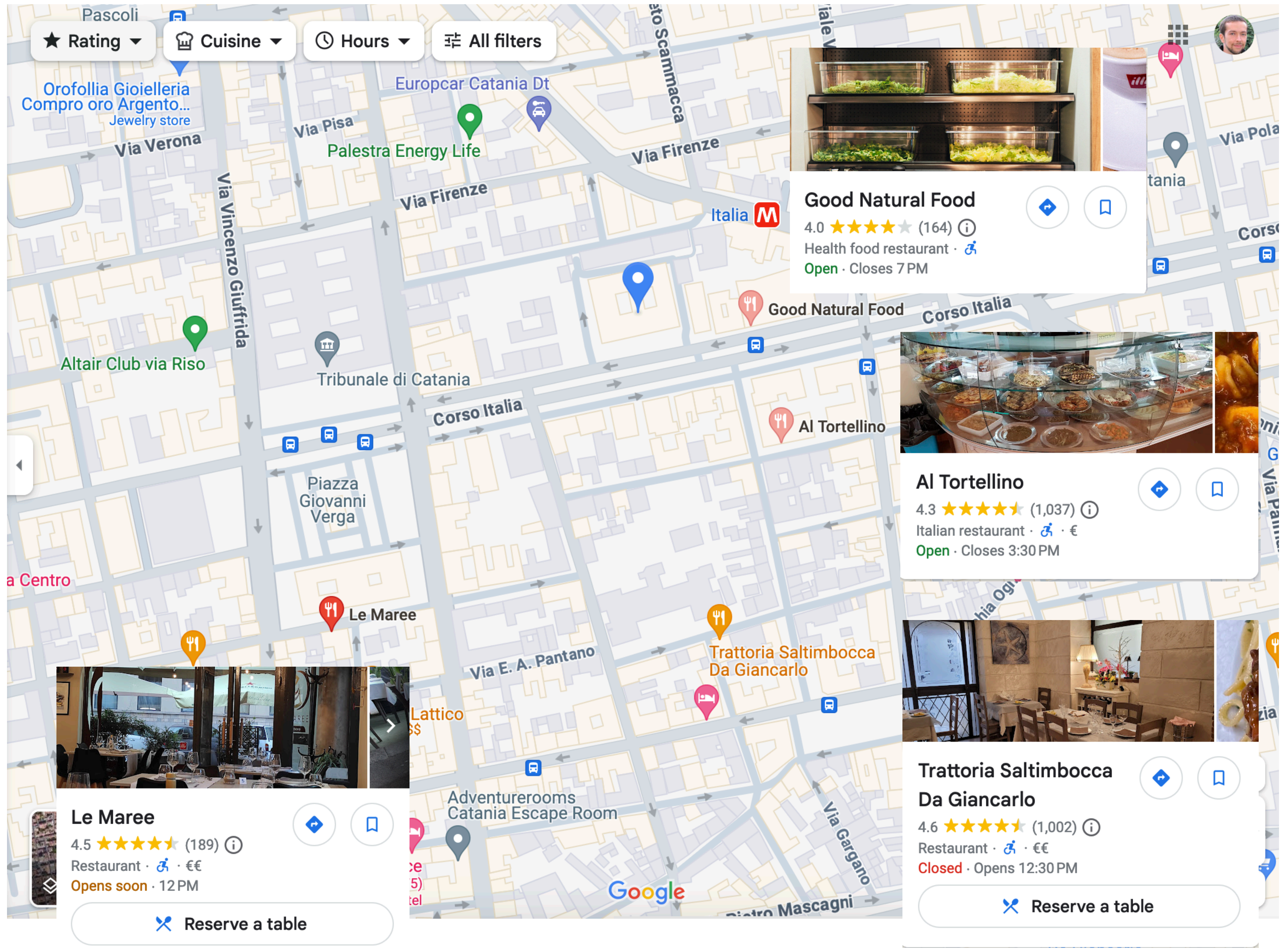
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal
- If one configuration is worse for all metrics than another configurations, it is **dominated**

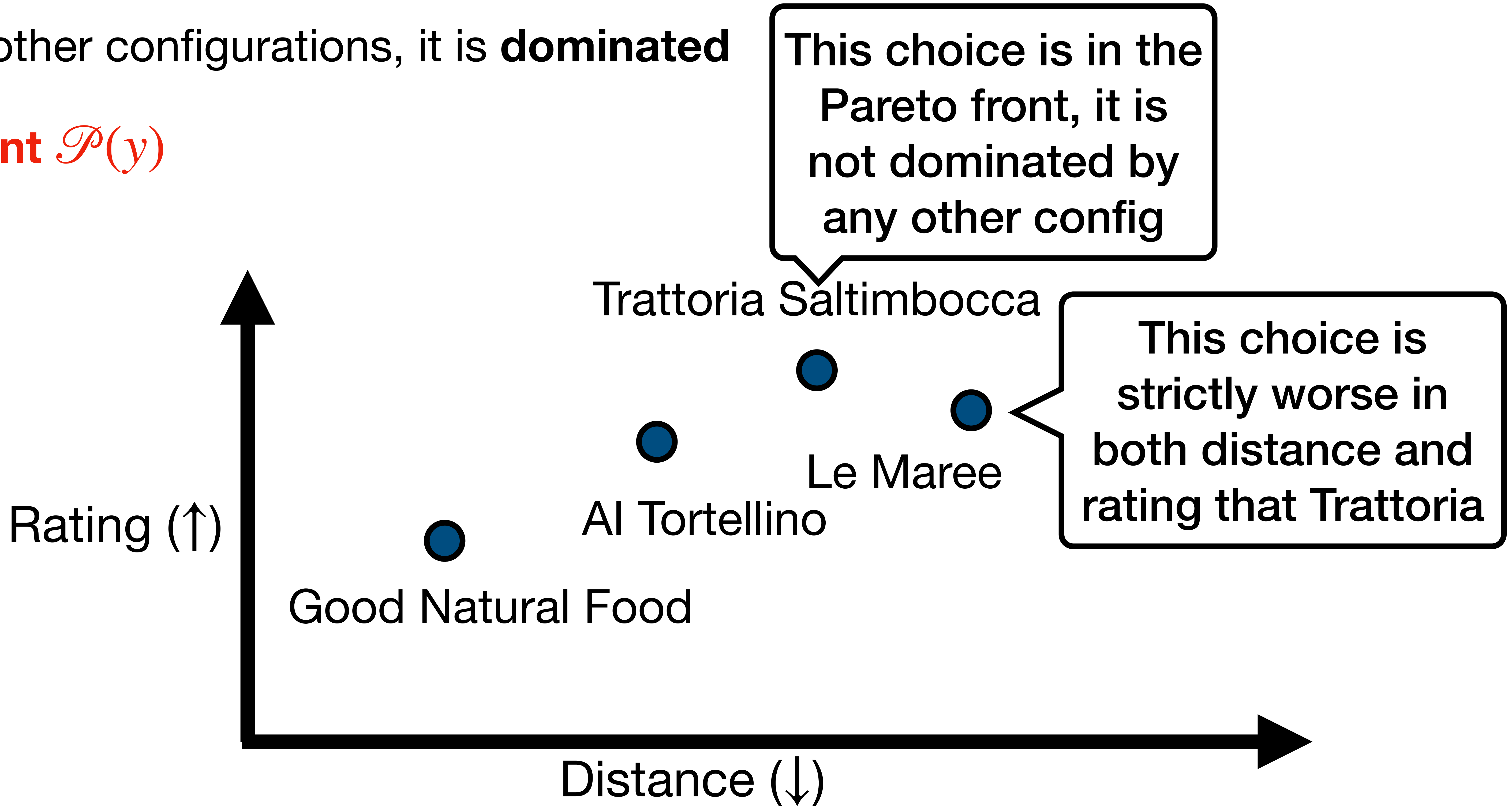




How to select which restaurant to go?

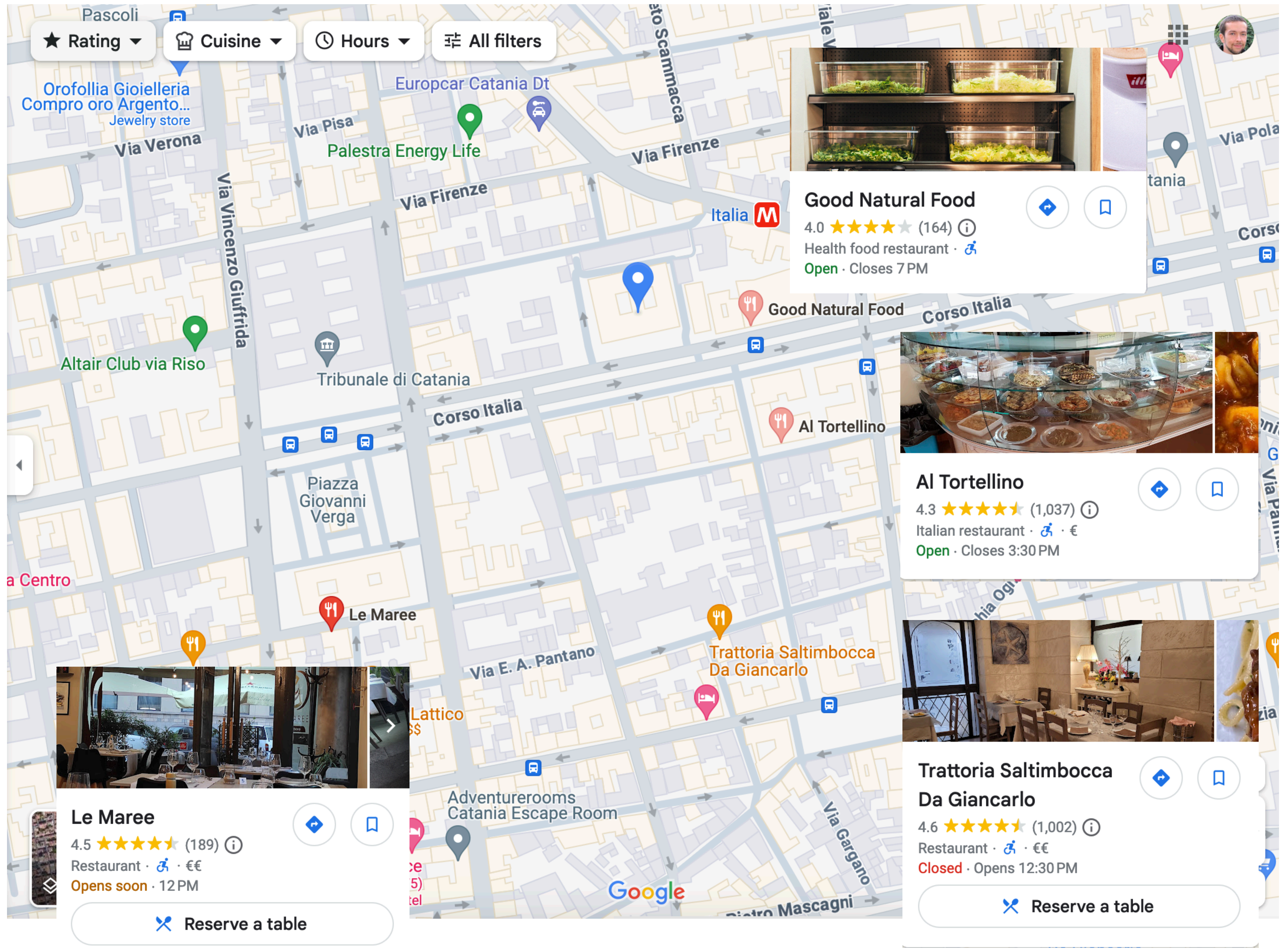
- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal
- If one configuration is worse for all metrics than another configurations, it is **dominated**
- The set of non dominated options is the **Pareto front** $\mathscr{P}(y)$

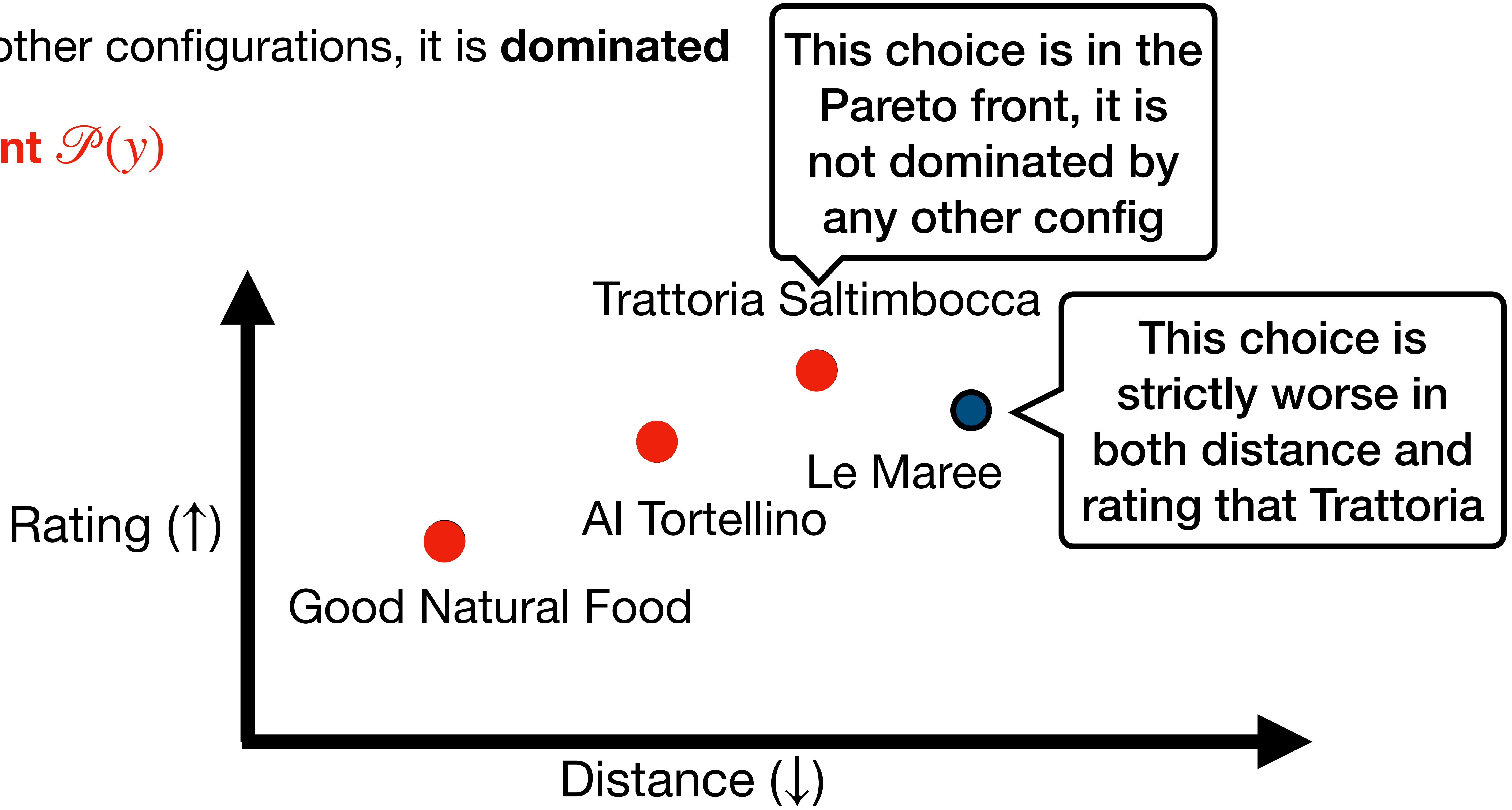




How to select which restaurant to go?

- Plot objective observations $y \in \mathbb{R}^{n \times d}$ where n is the number of configurations (restaurants) and d is the number of objectives
- No just a single best solution!
- 🤔 Which configuration to pick?
- Some choices are sub-optimal
- If one configuration is worse for all metrics than another configurations, it is **dominated**
- The set of non dominated options is the **Pareto front** $\mathscr{P}(y)$





Pareto front

Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized

Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.

Pareto front

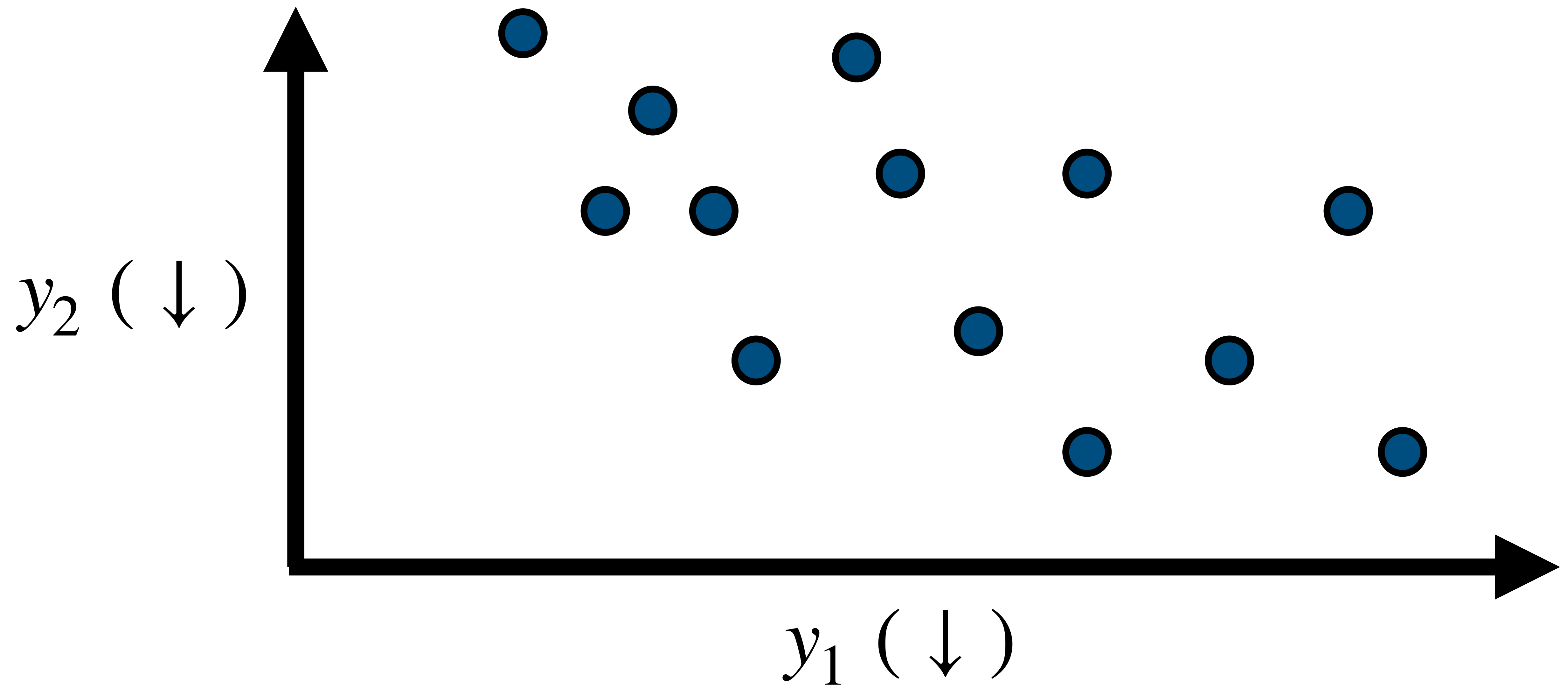
- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathscr{P}(y)$ is the set of non dominated options $\mathscr{P}(y) = \{y_i \,|\, \nexists y_j \prec y_i\}$

Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathscr{P}(y)$ is the set of non dominated options $\mathscr{P}(y) = \{y_i \mid \nexists y_j \prec y_i\}$
- 🤔 Which configurations are in the Pareto front $\mathscr{P}(y)$? 🤔

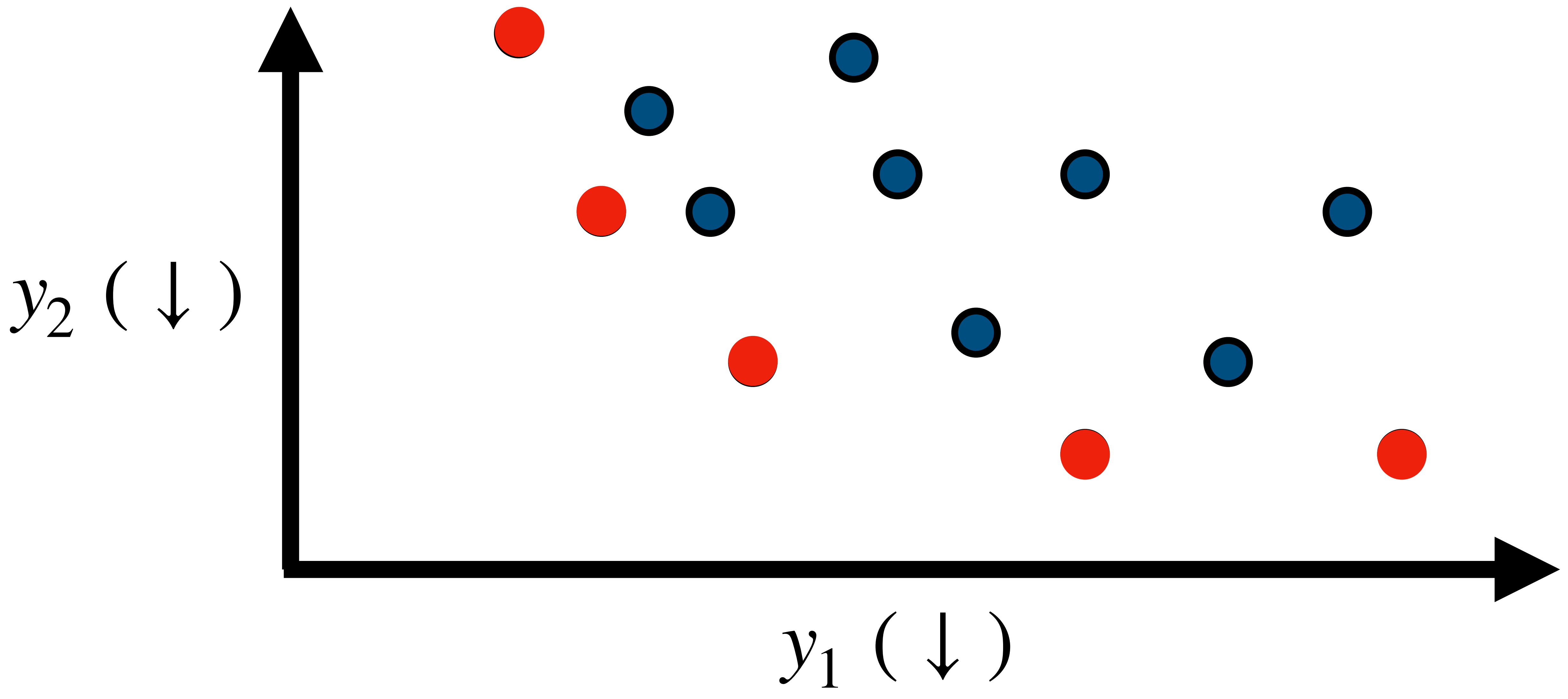
Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathscr{P}(y)$ is the set of non dominated options $\mathscr{P}(y) = \{y_i | \nexists y_j \prec y_i\}$
- 🤔 Which configurations are in the Pareto front $\mathscr{P}(y)$? 🤔



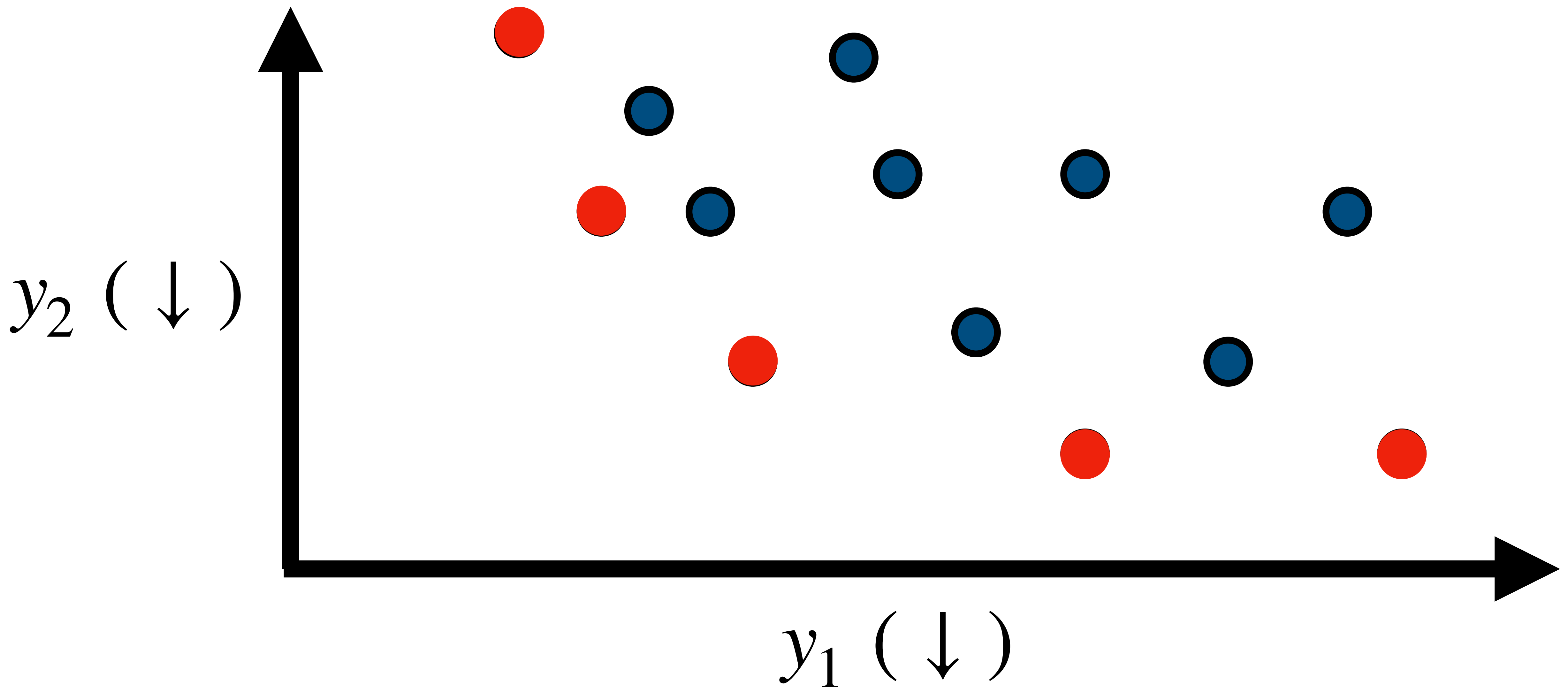
Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathscr{P}(y)$ is the set of non dominated options $\mathscr{P}(y) = \{y_i \,|\, \nexists y_j \prec y_i\}$
- 🤔 Which configurations are in the Pareto front $\mathscr{P}(y)$? 🤔



Pareto front

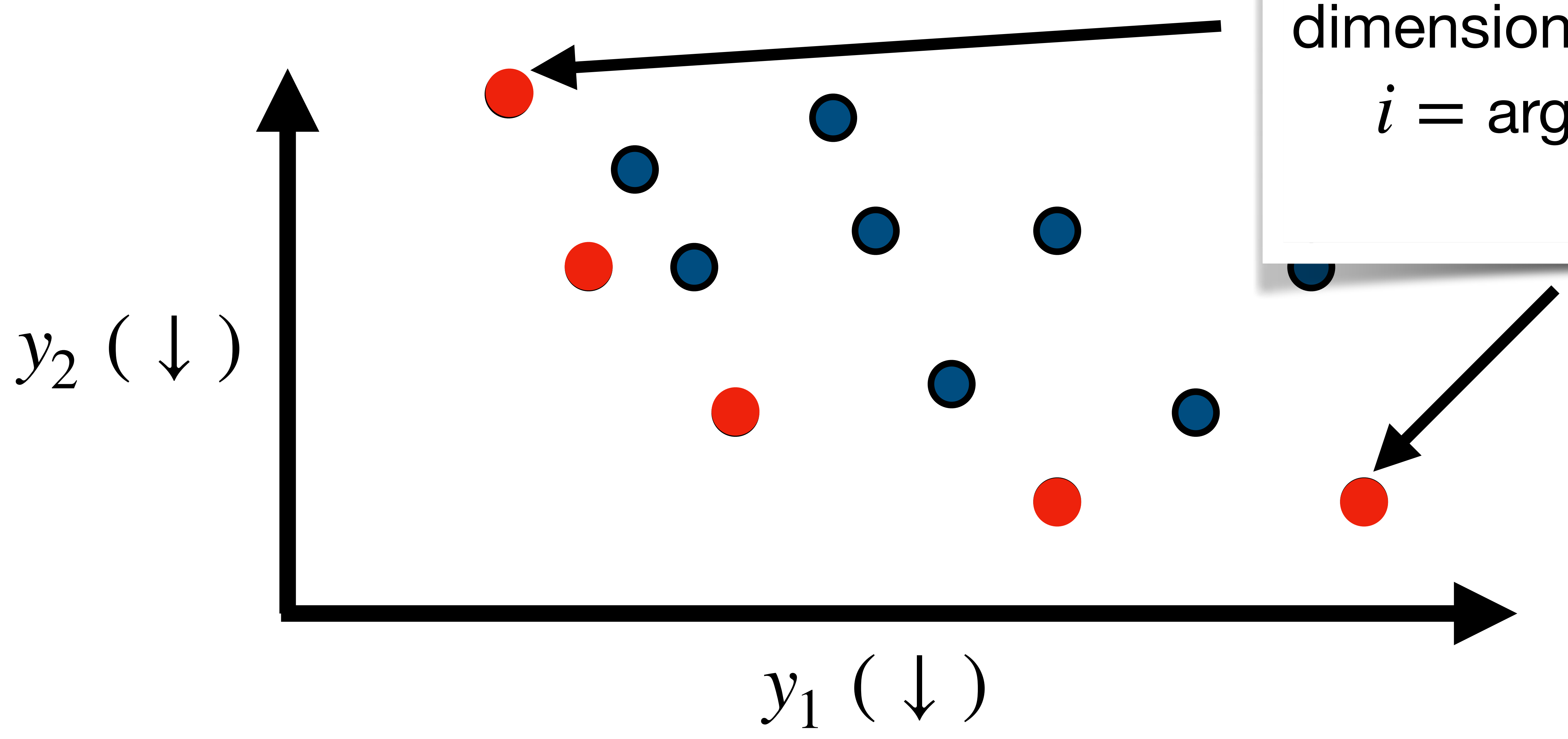
- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathcal{P}(y)$ is the set of non dominated options $\mathcal{P}(y) = \{y_i \mid \nexists y_j \prec y_i\}$
- 🤔 Which configurations are in the Pareto front $\mathcal{P}(y)$? 🤔
- 🤔 Given objectives $y \in \mathbb{R}^{n \times d}$, can you give a procedure that finds d configurations that are in the Pareto front $\mathcal{P}(y)$ in $\mathcal{O}(d)$?



Pareto front

- Metrics observations $y \in \mathbb{R}^{n \times d}$, assume all metrics are to be minimized
- We say that y_i dominate y_j (which we denote $y_i \prec y_j$) iff $\forall k, y_{ik} \leq y_{jk}$ and $\exists k, y_{ik} < y_{jk}$.
- The Pareto front $\mathscr{P}(y)$ is the set of non dominated options $\mathscr{P}(y) = \{y_i \,|\, \nexists y_j \prec y_i\}$
- 🤔 Which configurations are in the Pareto front $\mathscr{P}(y)$? 🤔
- 🤔 Given objectives $y \in \mathbb{R}^{n \times d}$, can you give a procedure that finds d configurations that are in the Pareto front $\mathscr{P}(y)$ in $\mathcal{O}(d)$?

The configuration minimizing the j-th dimension is always on the Pareto front! $i = \text{argmin}_{i \in [n]} y_{ij}$ is always on the Pareto front



Multifidelity

Asynchronous Succesful halving (ASHA)

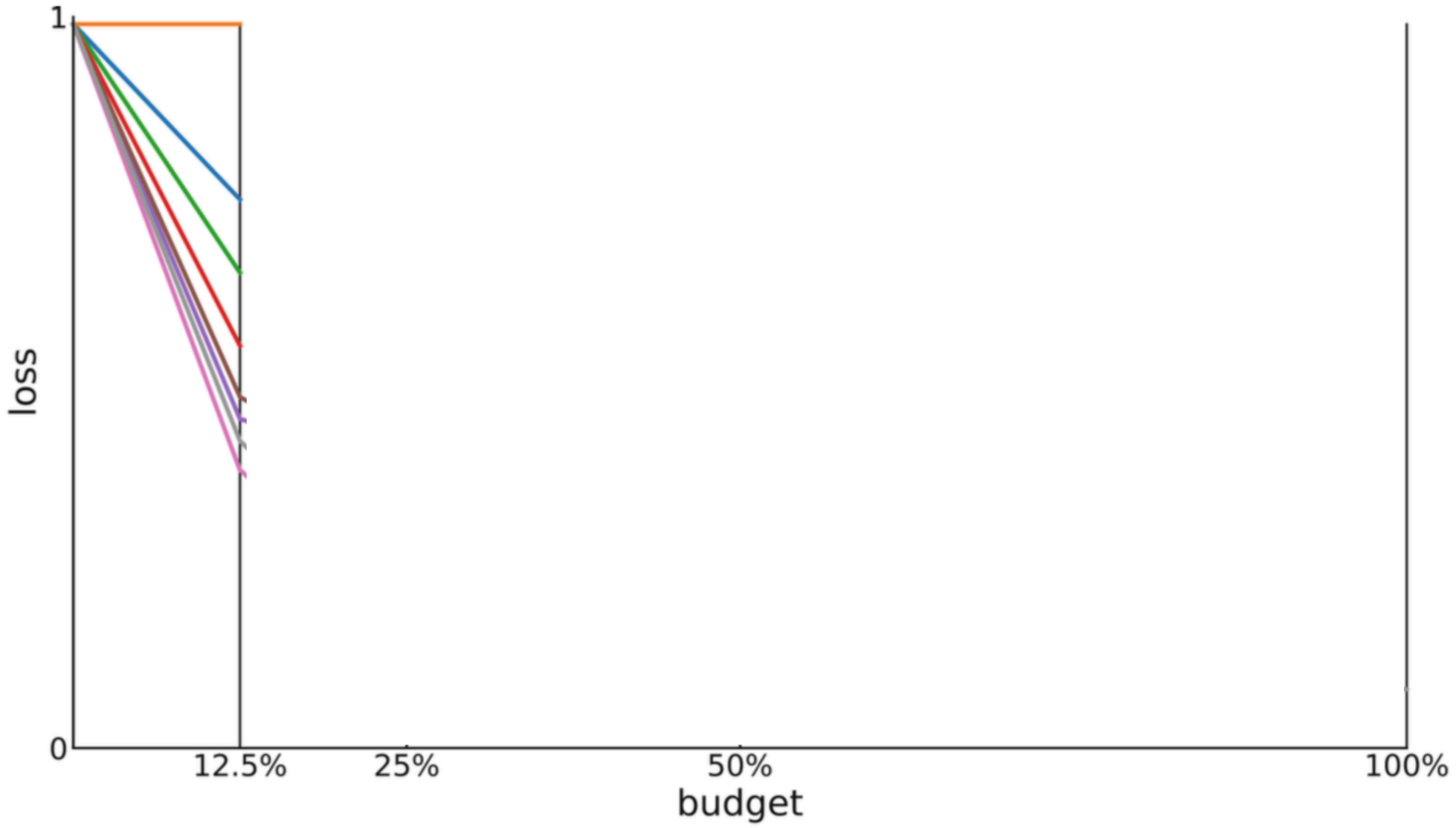


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?

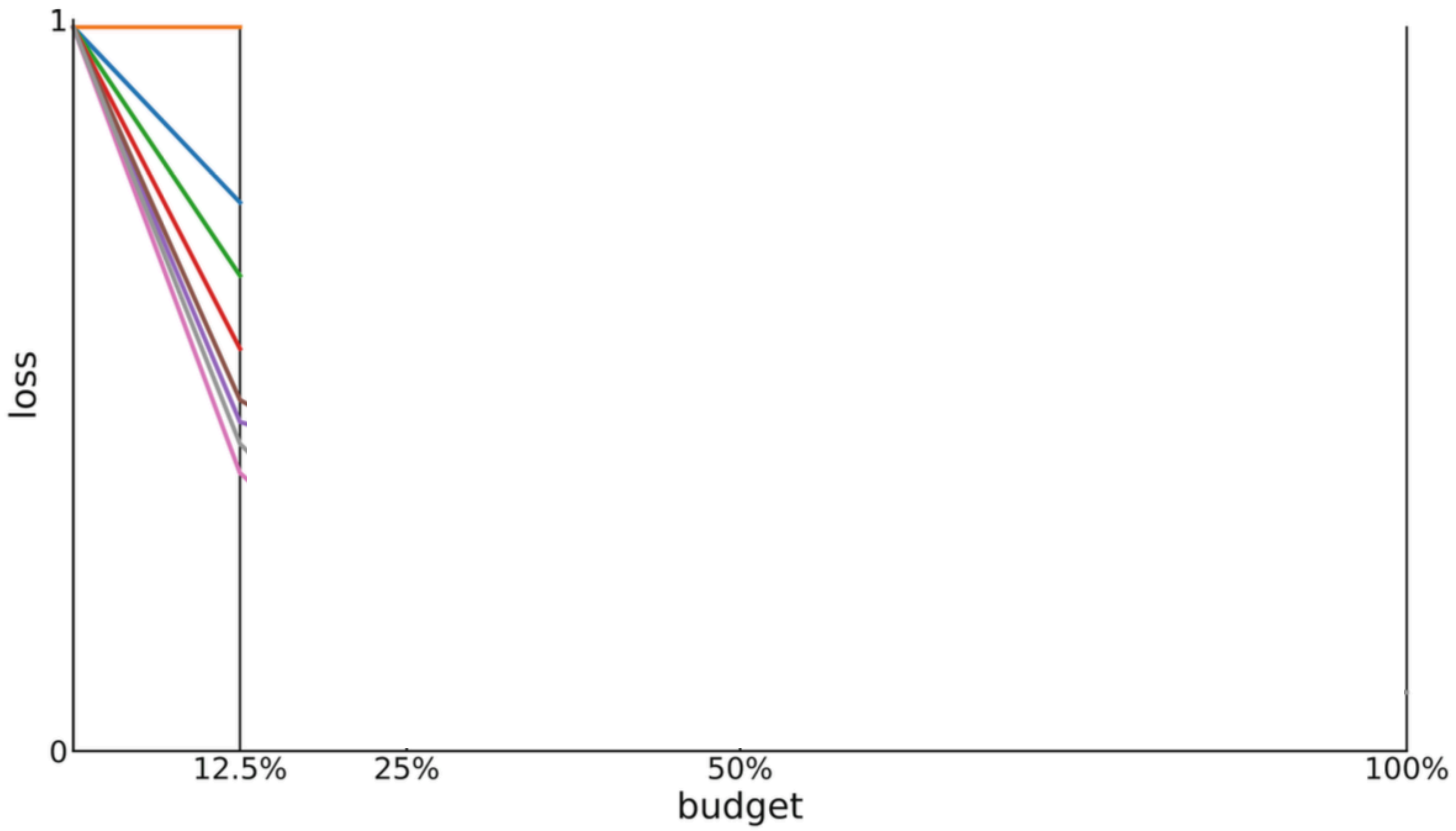


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

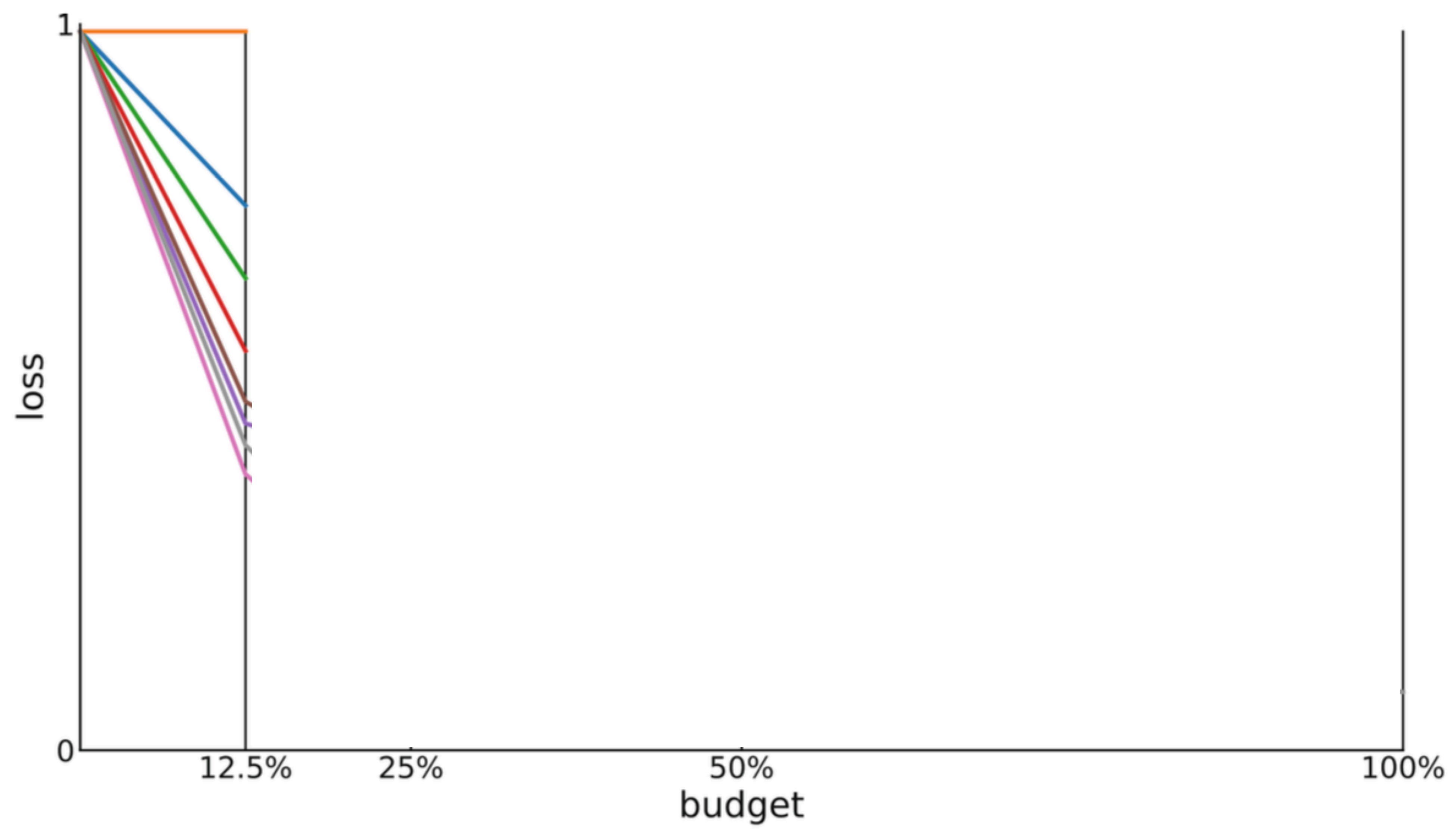


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

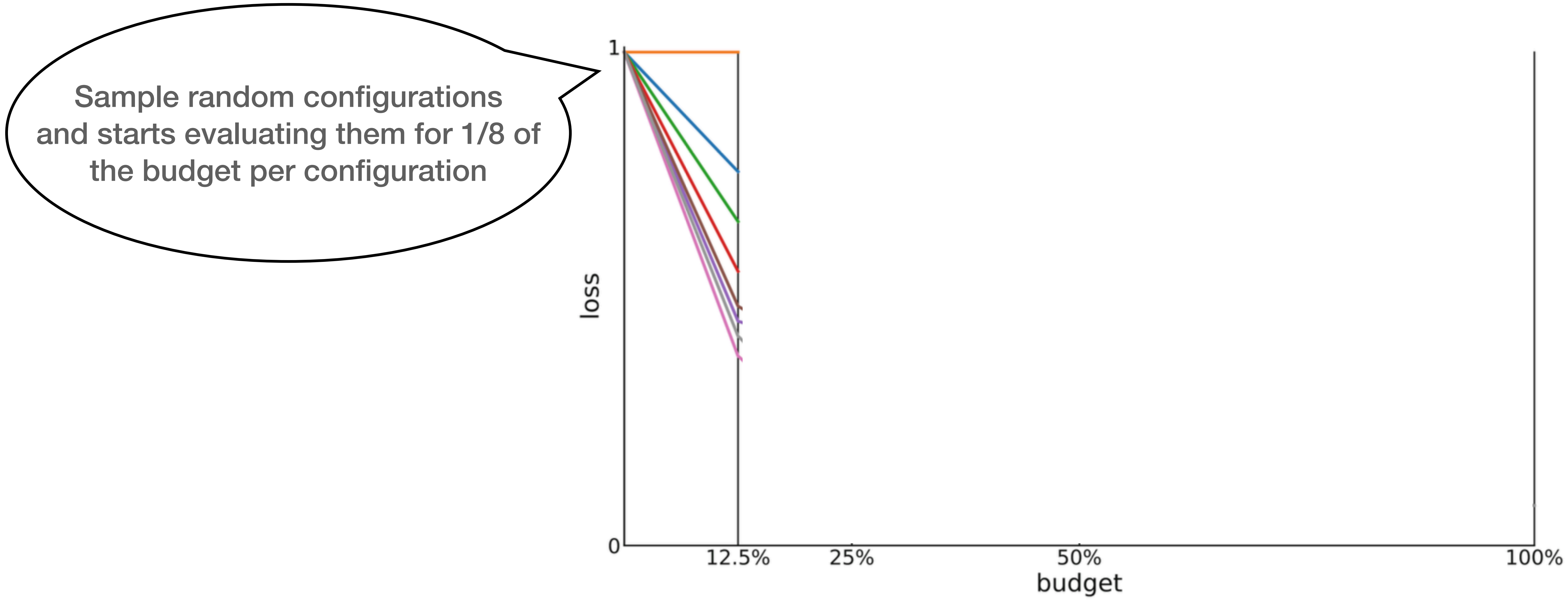


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

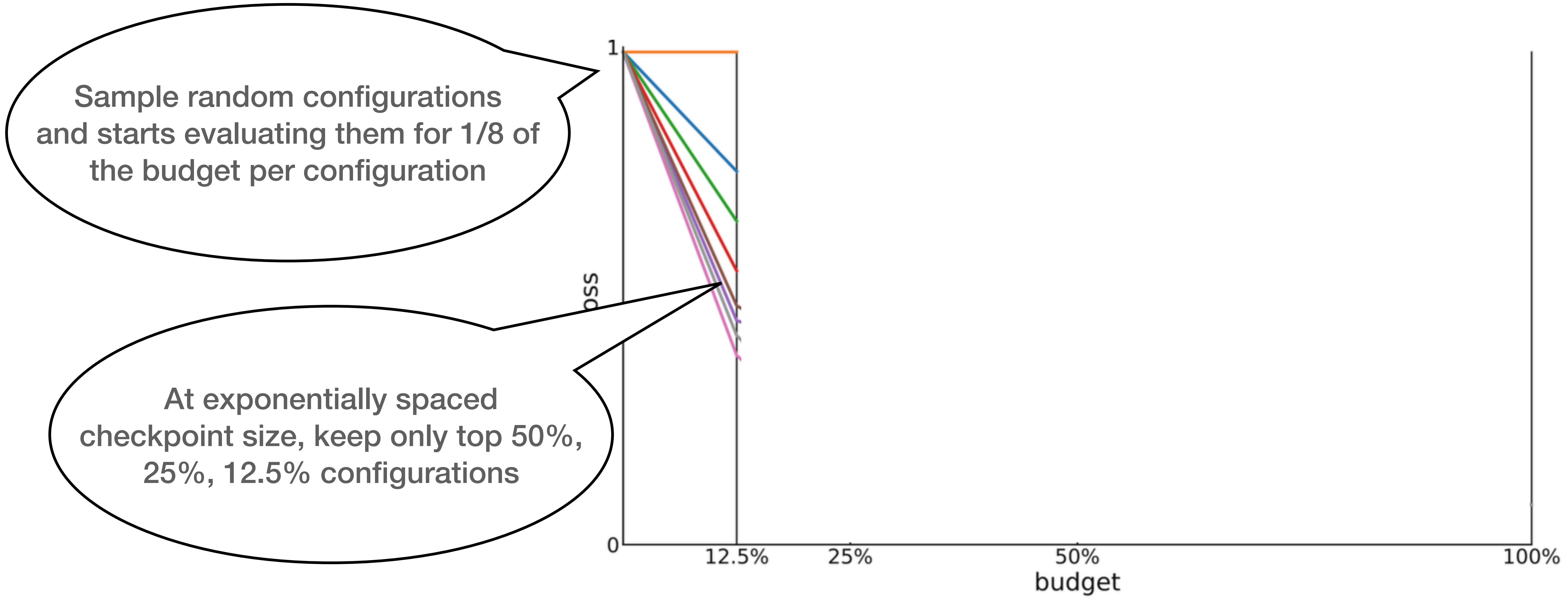


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

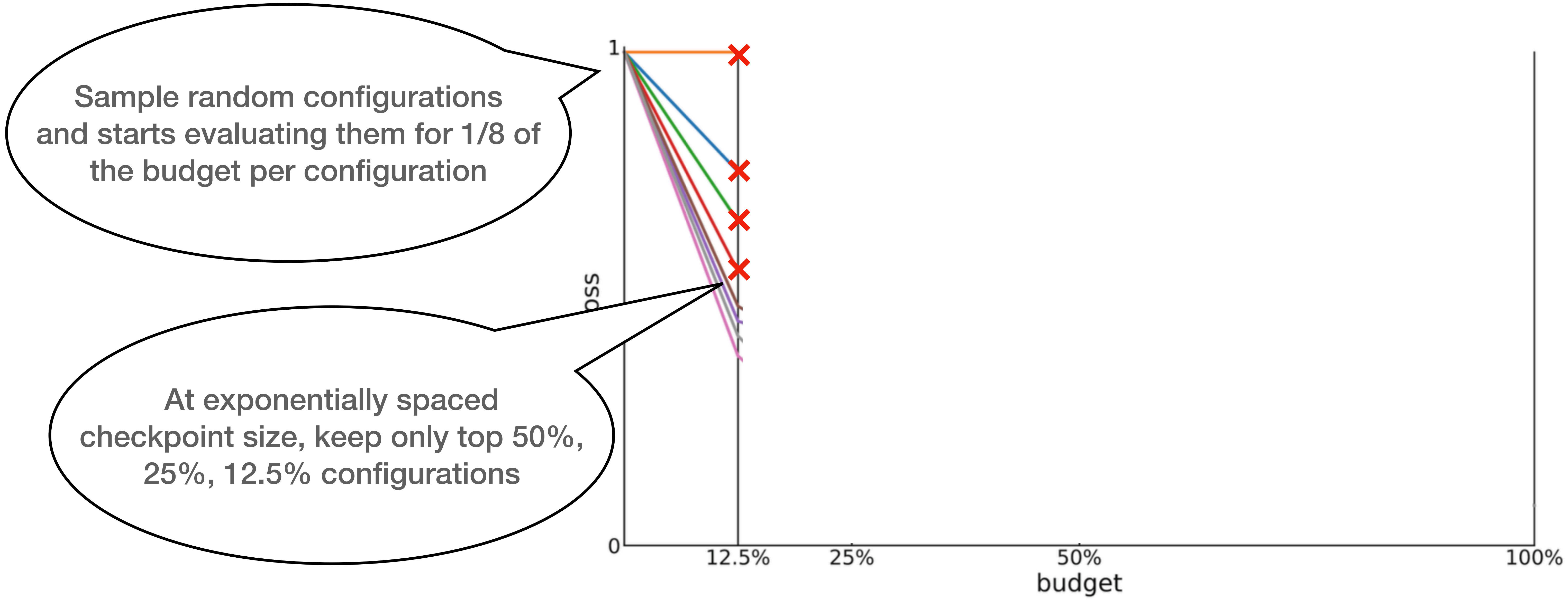


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

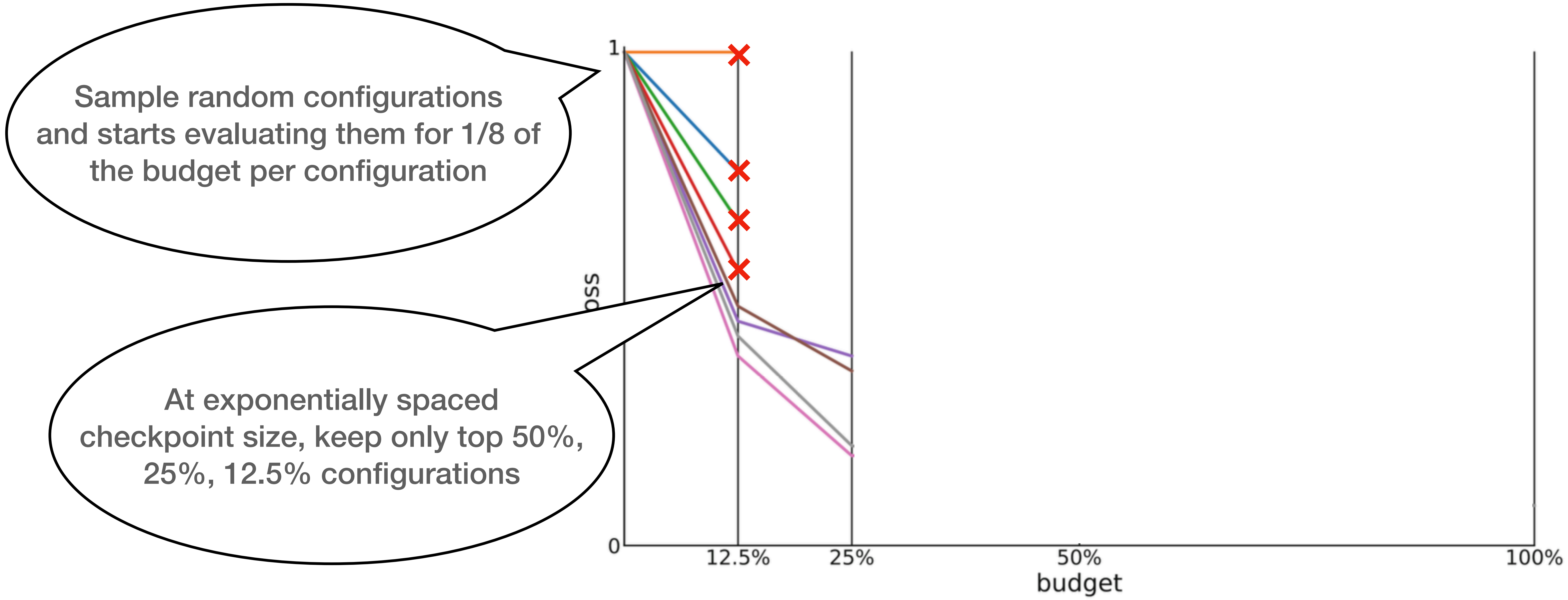


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

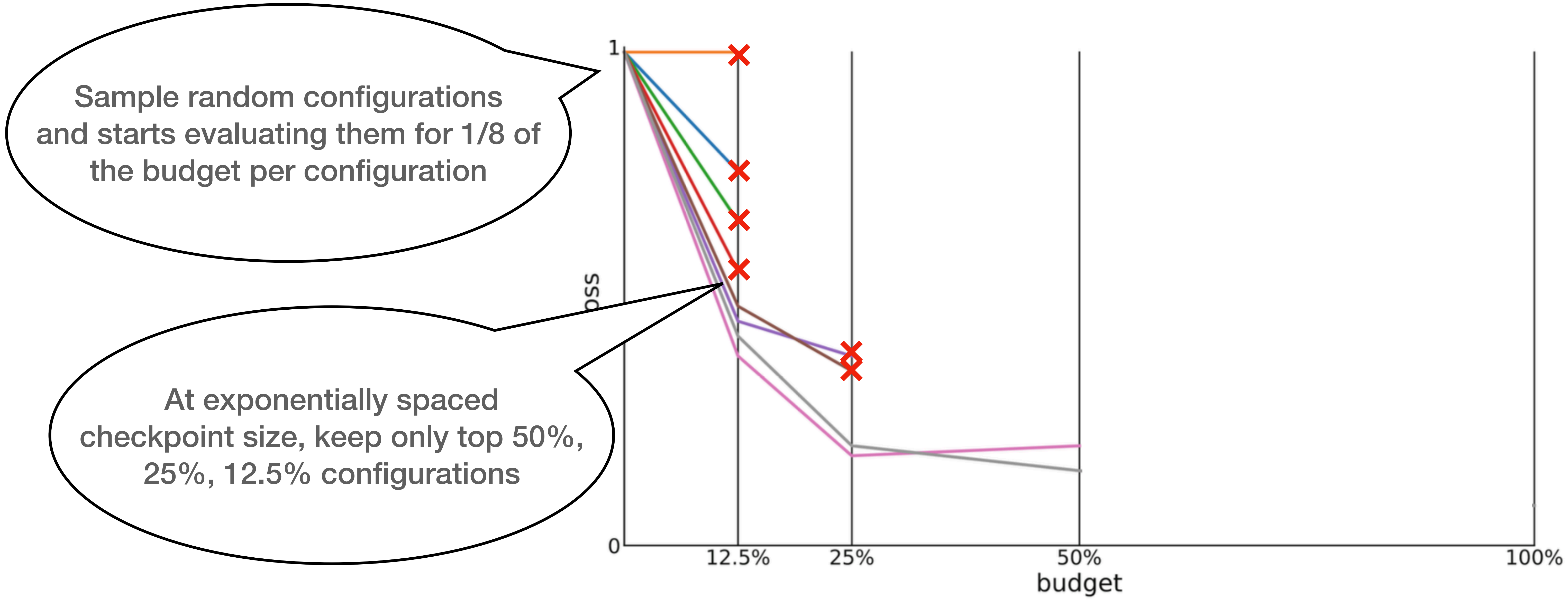


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

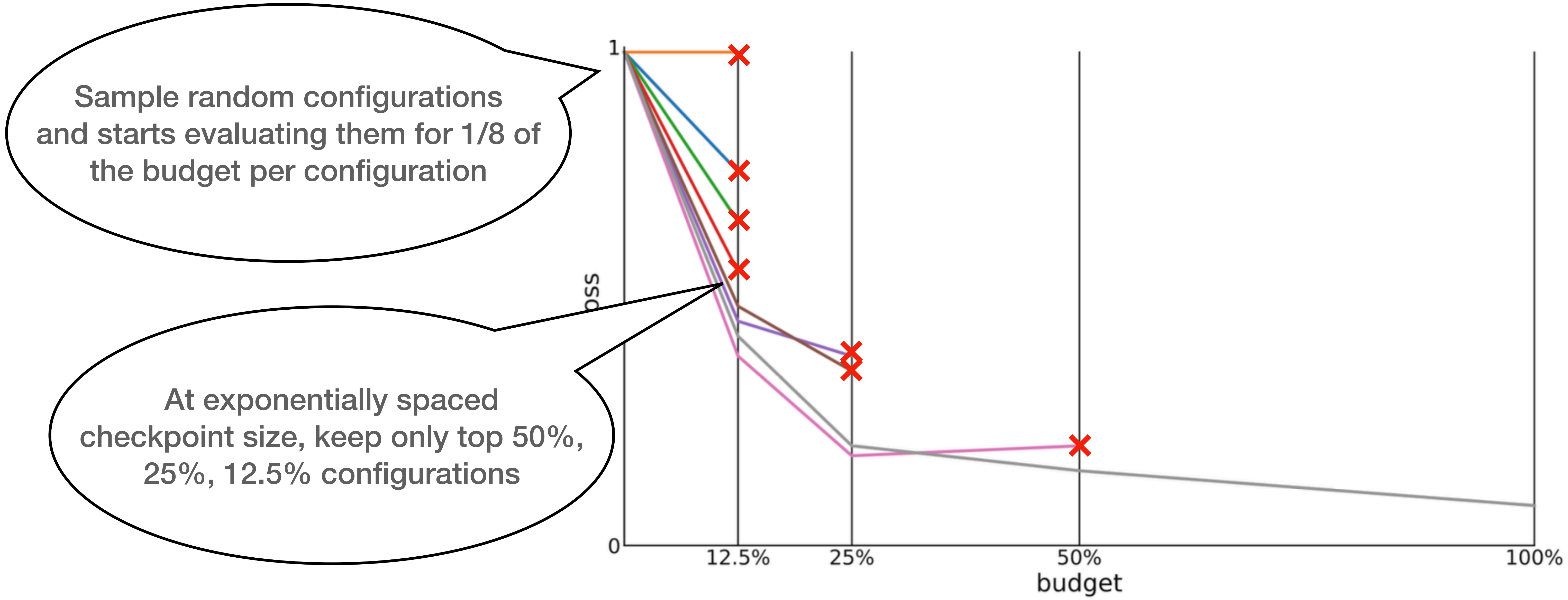


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

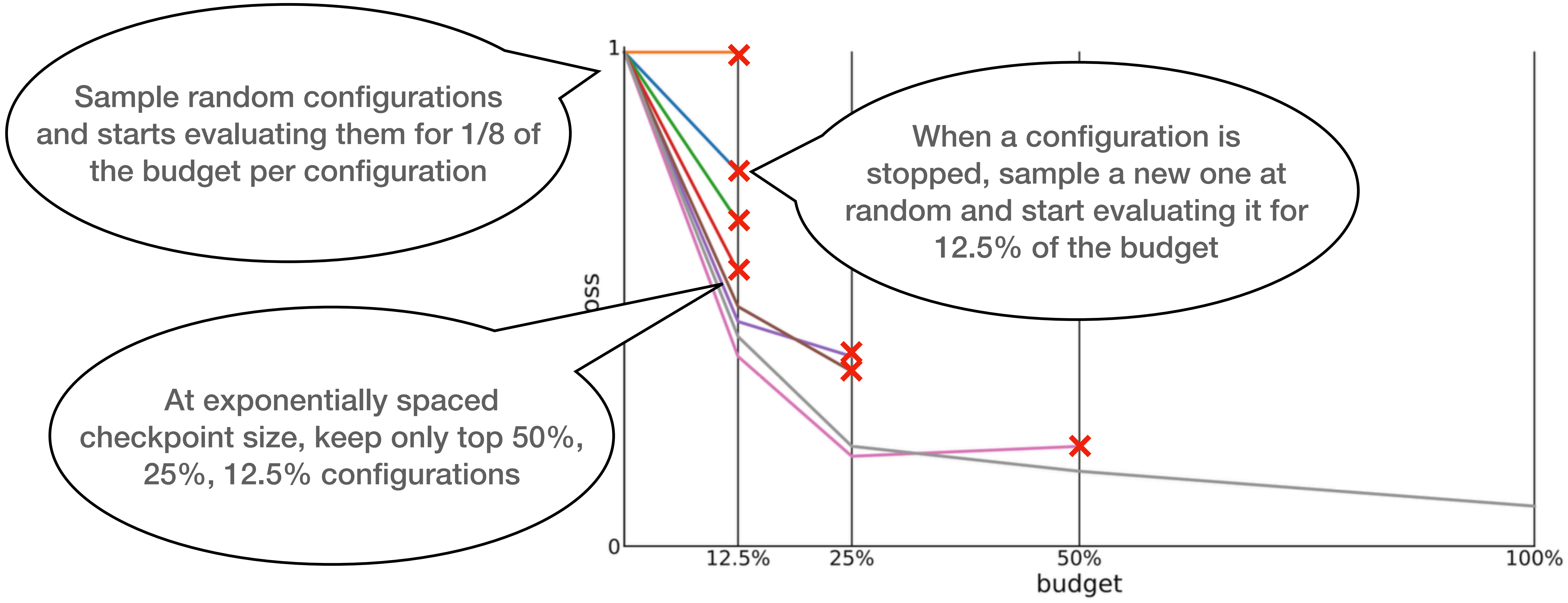


Image credit: Matthias Feurer.

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

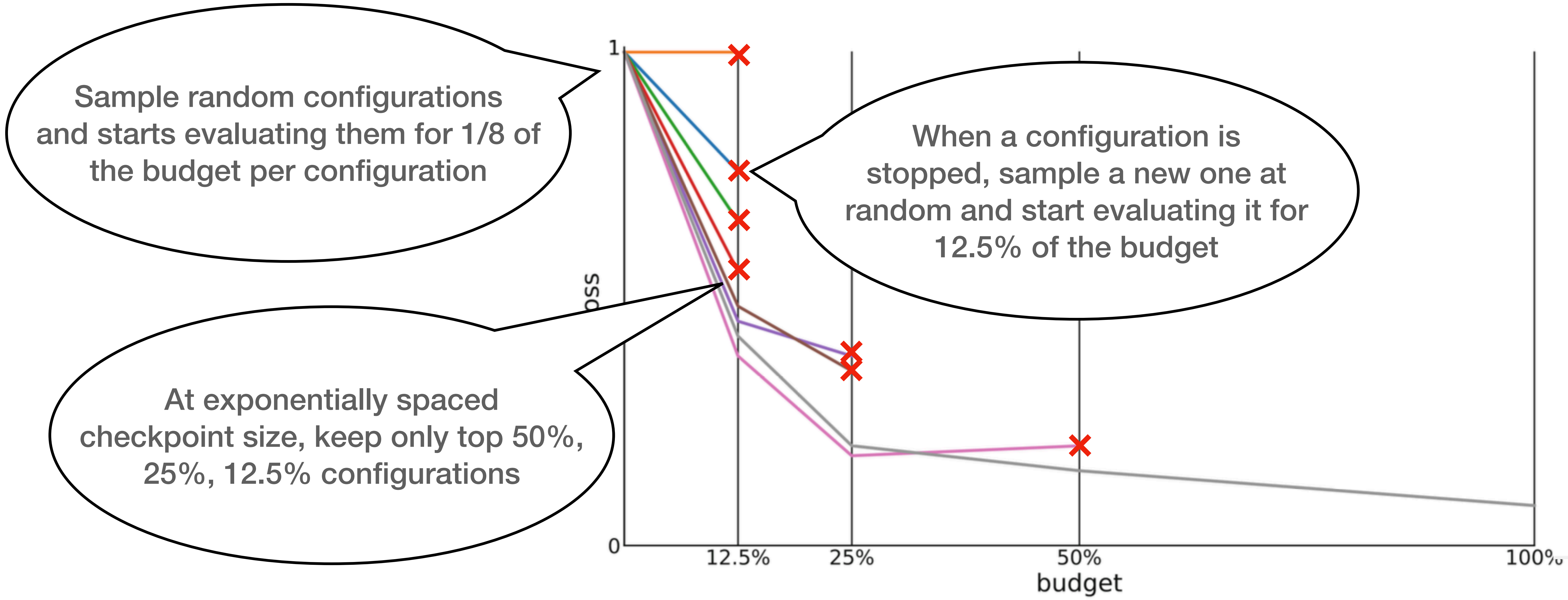
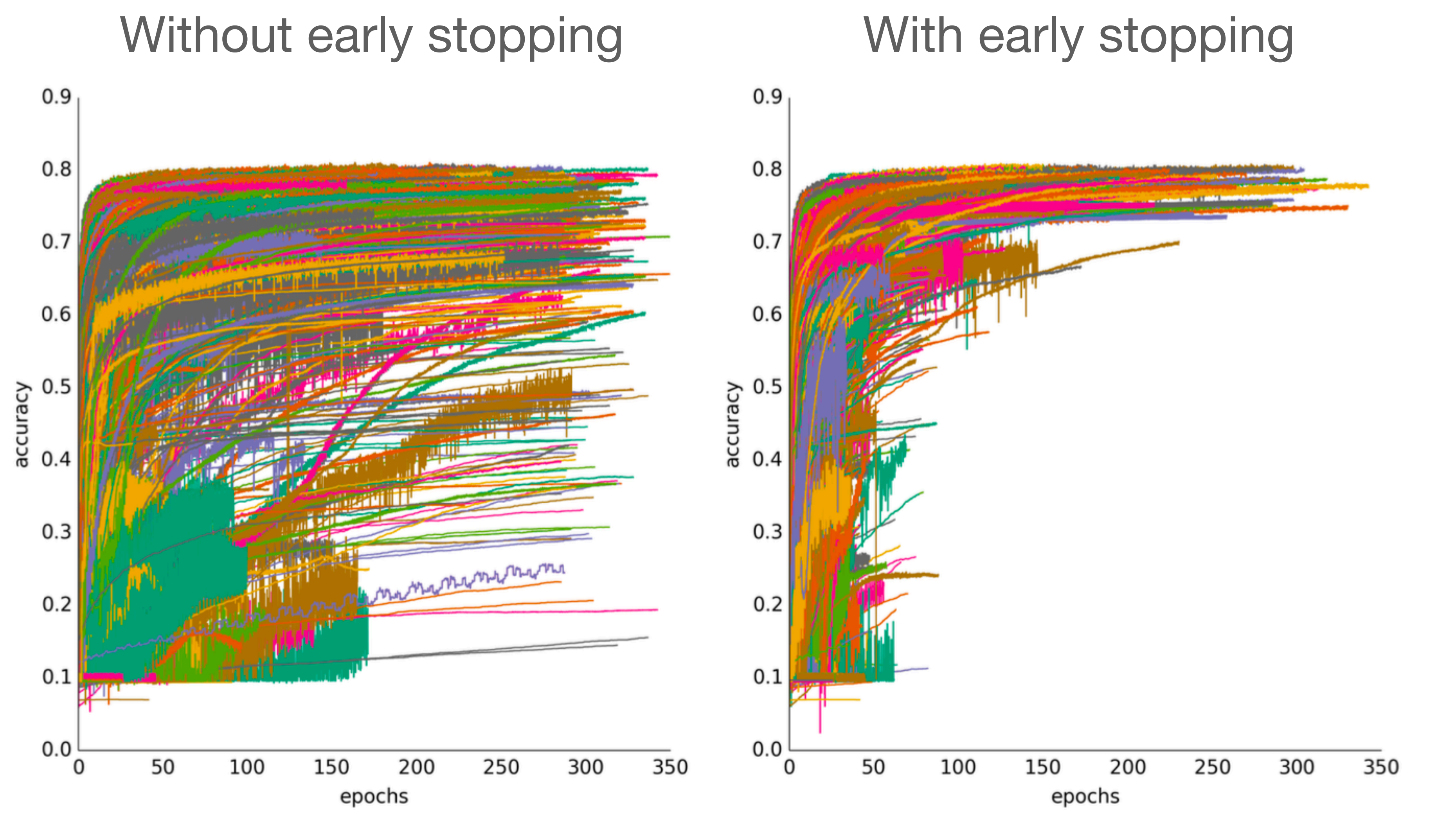


Image credit: Matthias Feurer.



Wistuba and Grabocka. Meta-Learning for Hyperparameter Optimization 2023

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

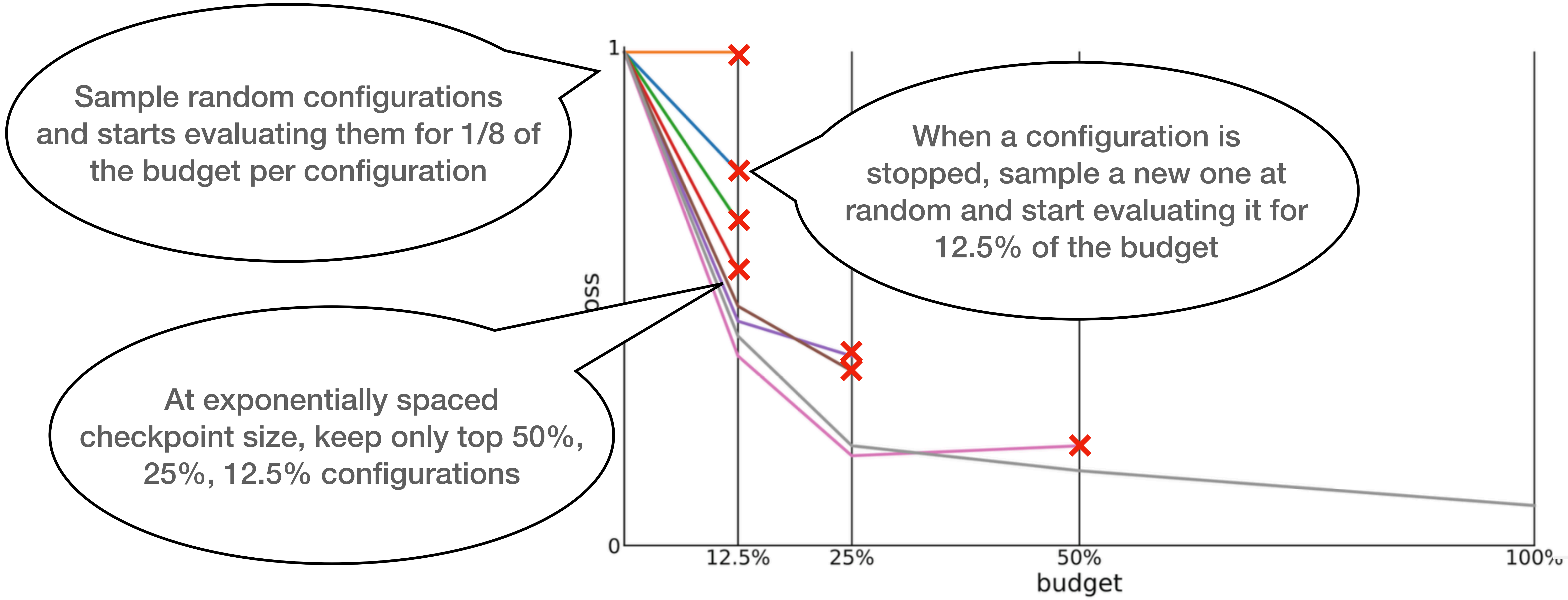
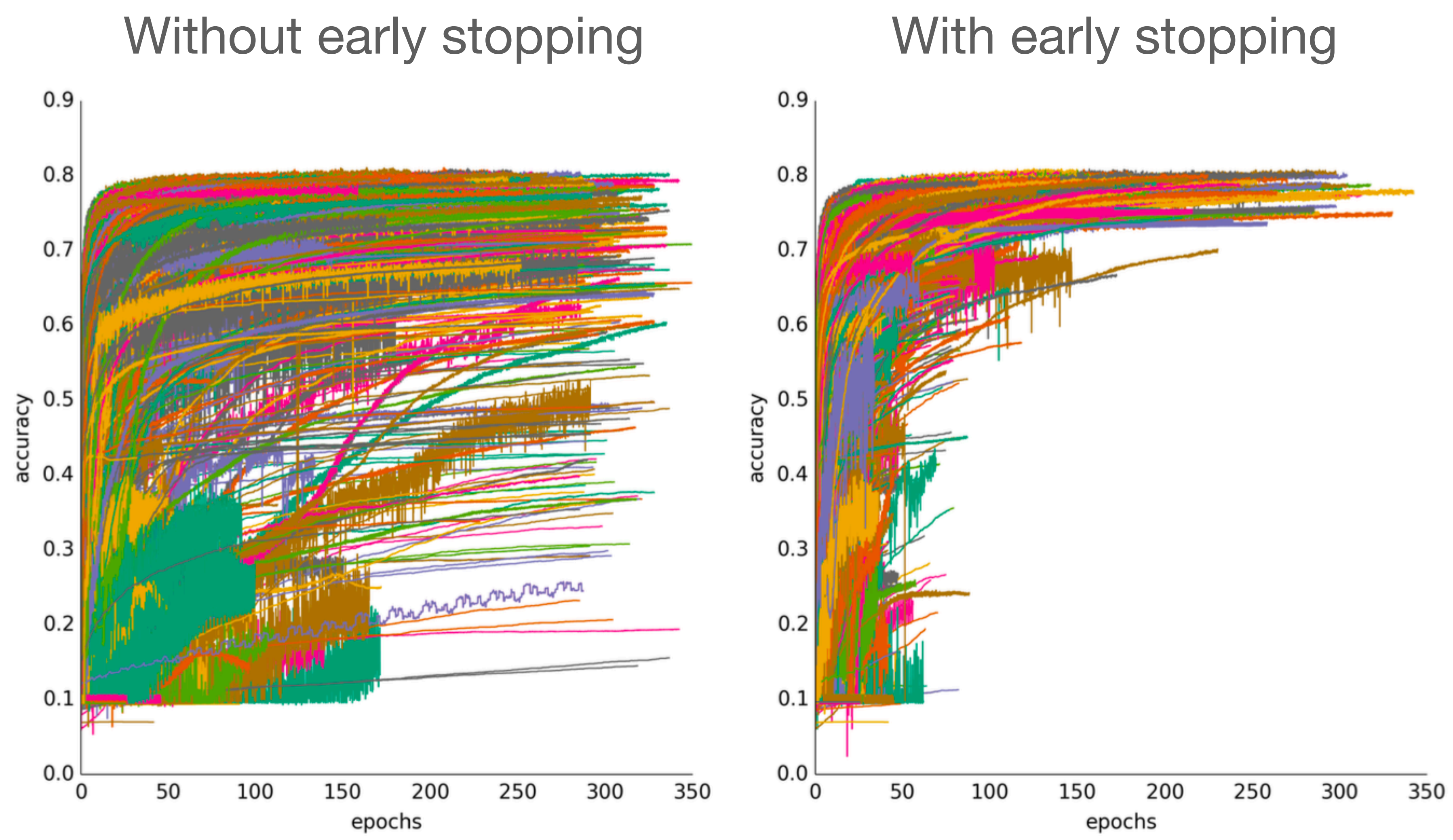


Image credit: Matthias Feurer.



Wistuba and Grabocka. Meta-Learning for Hyperparameter Optimization 2023

- 👍 **Great performance** in practice and one can use multiple workers

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

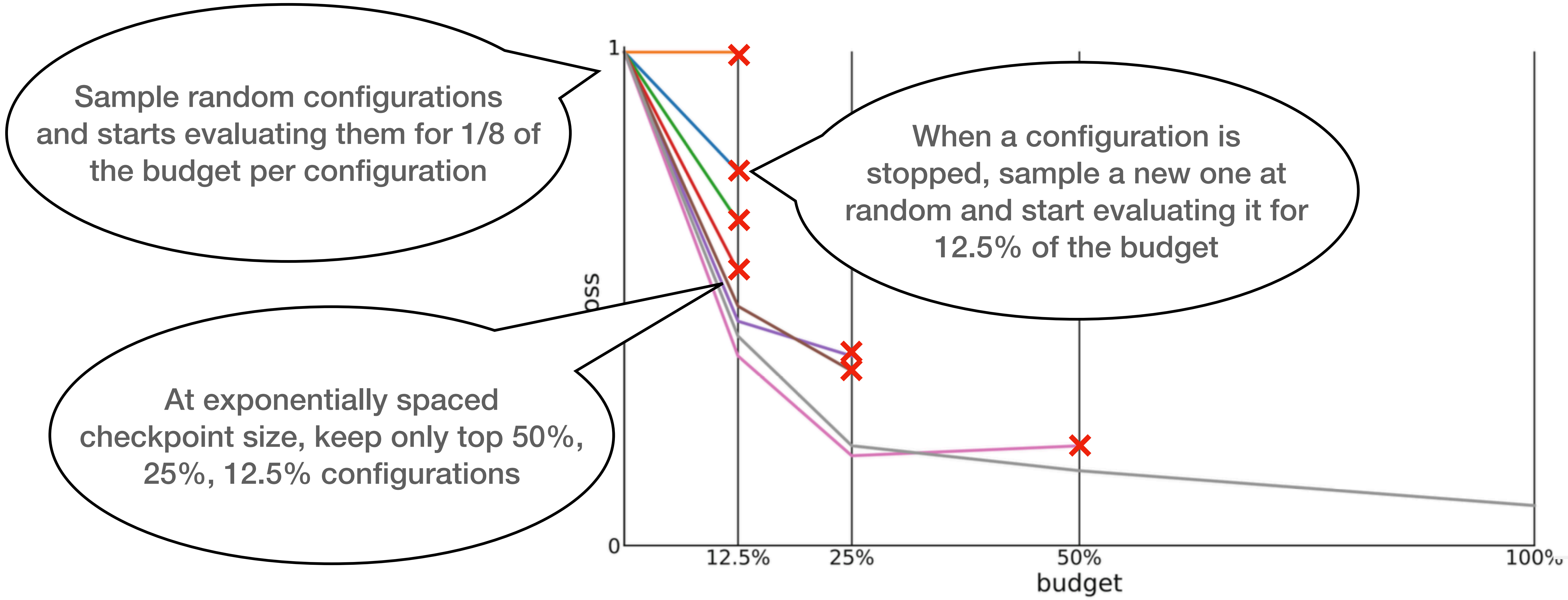
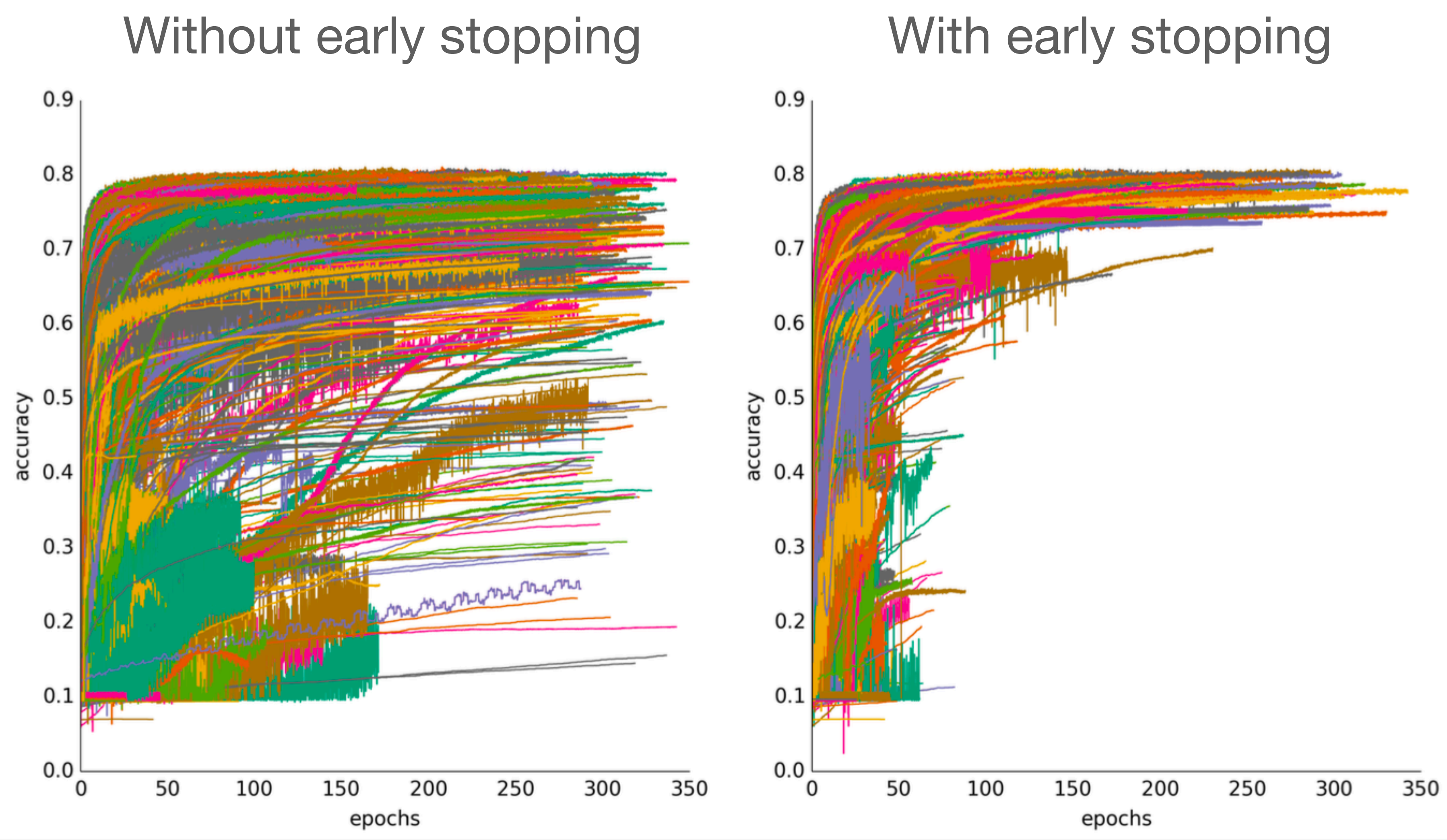


Image credit: Matthias Feurer.



Wistuba and Grabocka. Meta-Learning for Hyperparameter Optimization 2023

- 👍 **Great performance** in practice and one can use multiple workers
- 🤔 How can we extend the algorithm to handle multiple objectives?

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

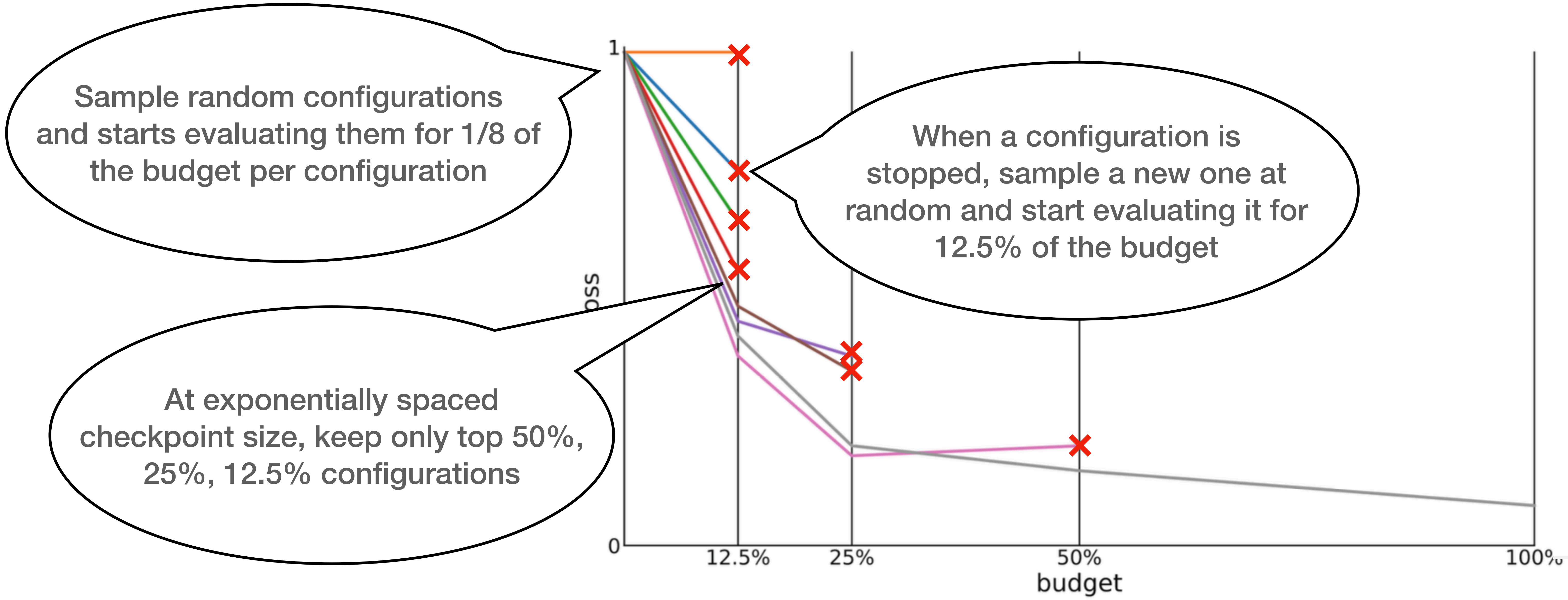
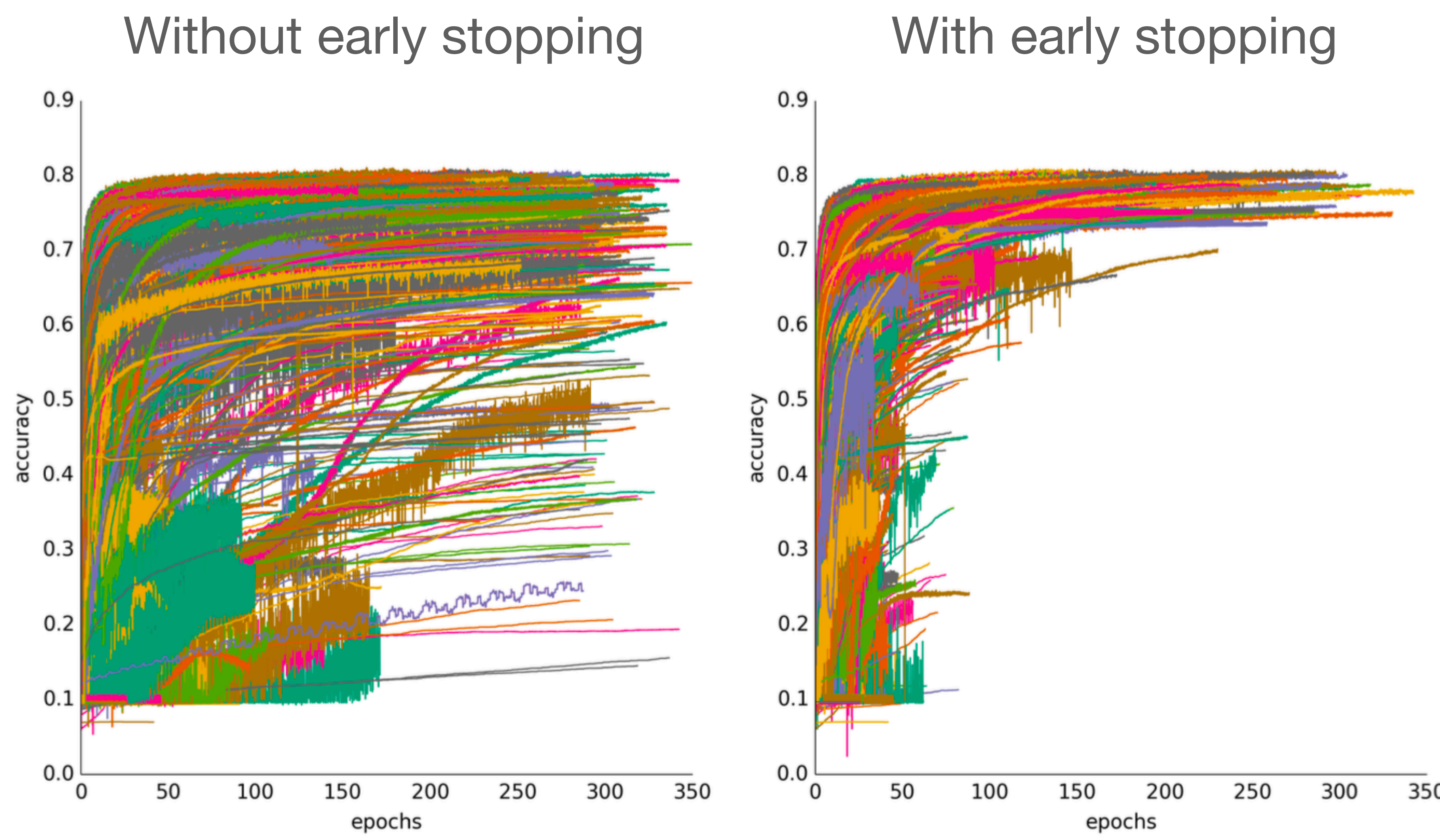


Image credit: Matthias Feurer.



Wistuba and Grabocka. Meta-Learning for Hyperparameter Optimization 2023

- 👍 **Great performance** in practice and one can use multiple workers
- 🤔 How can we extend the algorithm to handle multiple objectives?
- 🤔 We need to sort to discard the bottom half of configurations, how can we sort if we have multiple objectives?

Multifidelity

Asynchronous Succesful halving (ASHA)

- Typically, models are trained incrementally: can we stop bad configuration early?
- Early results (epochs) can be used to stop bad runs (early stopping)

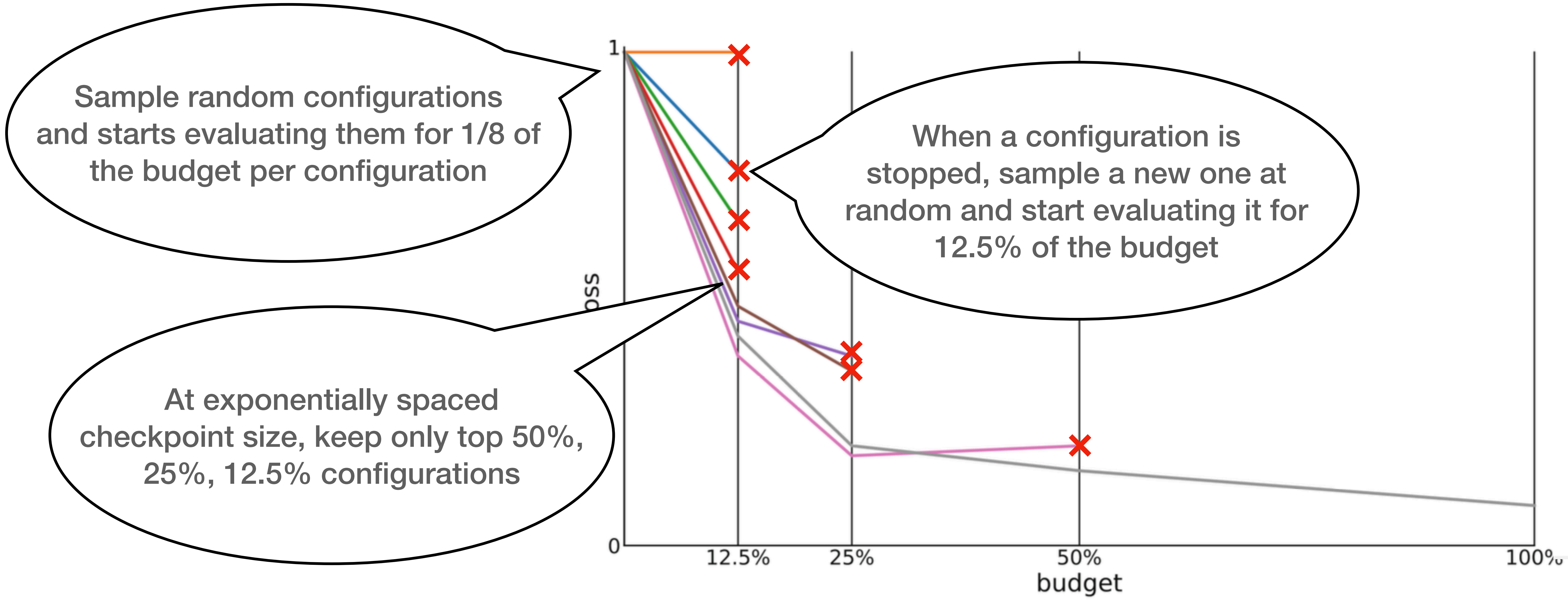
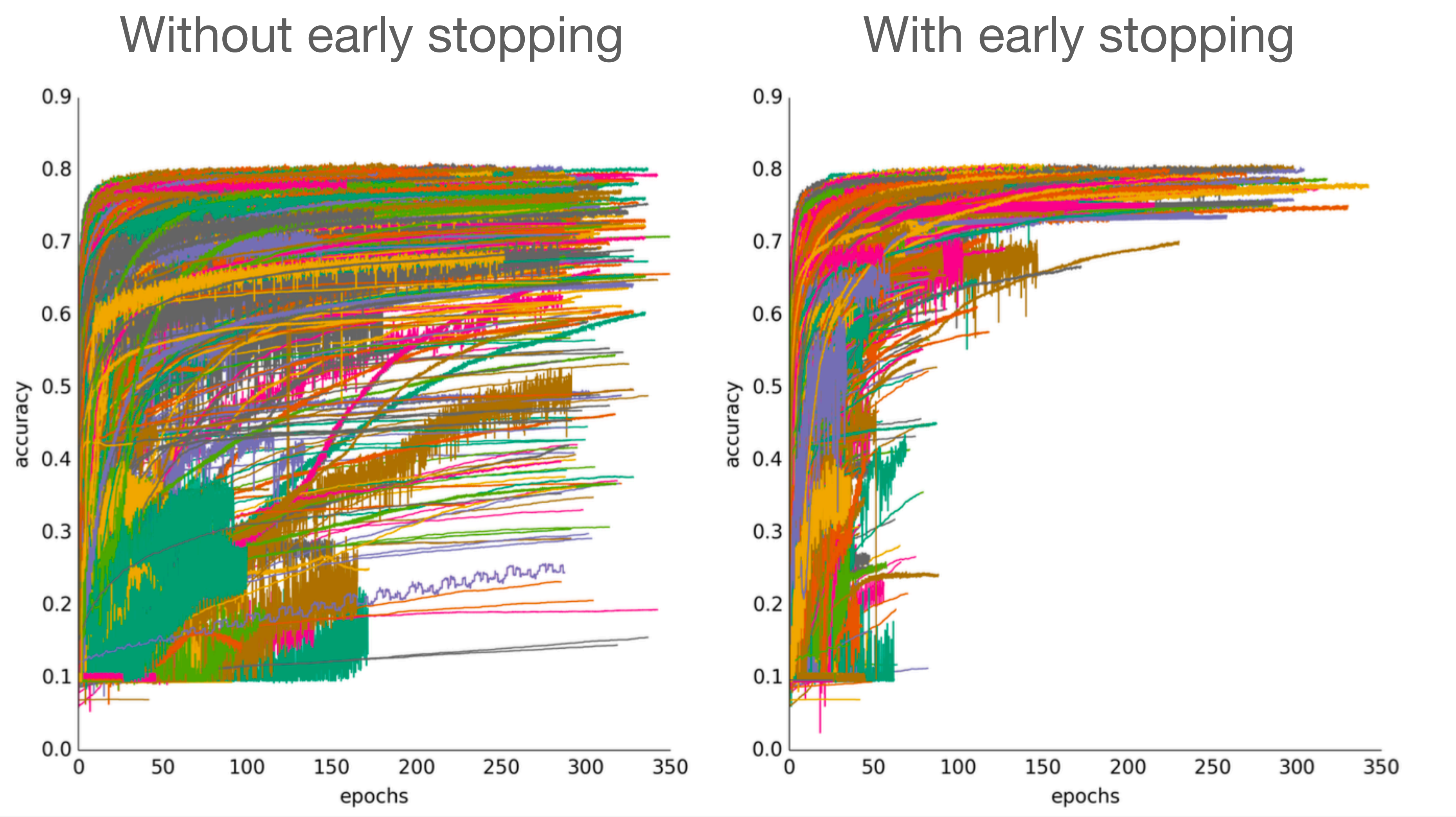


Image credit: Matthias Feurer.



Wistuba and Grabocka. Meta-Learning for Hyperparameter Optimization 2023

- 👍 **Great performance** in practice and one can use multiple workers
- 🤔 How can we extend the algorithm to handle multiple objectives?
- 🤔 We need to sort to discard the bottom half of configurations, how can we sort if we have multiple objectives?

💡 Non-dominated sort allows to sort even when we have multiple objectives

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
    - Compute the Pareto front of y, break ties with an heuristic

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathcal{P}(y)$, break ties with an heuristic

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathscr{P}(y)$, break ties with an heuristic

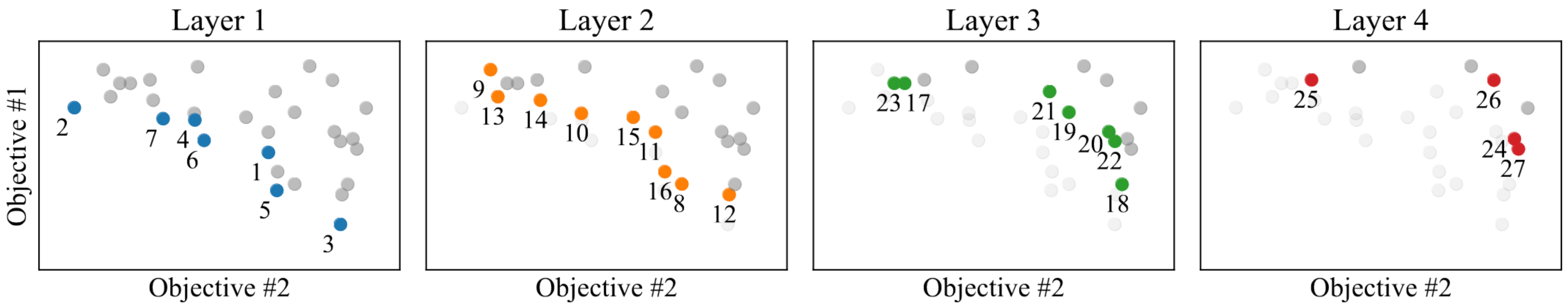


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathcal{P}(y)$, break ties with an heuristic

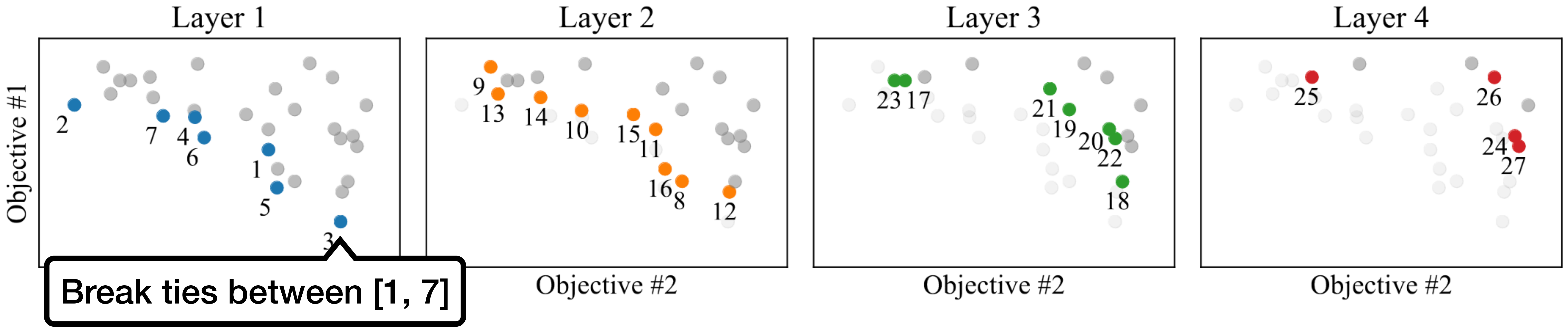


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathscr{P}(y)$, break ties with an heuristic
  - Heuristic choices aims at selecting a subset with a good coverage:

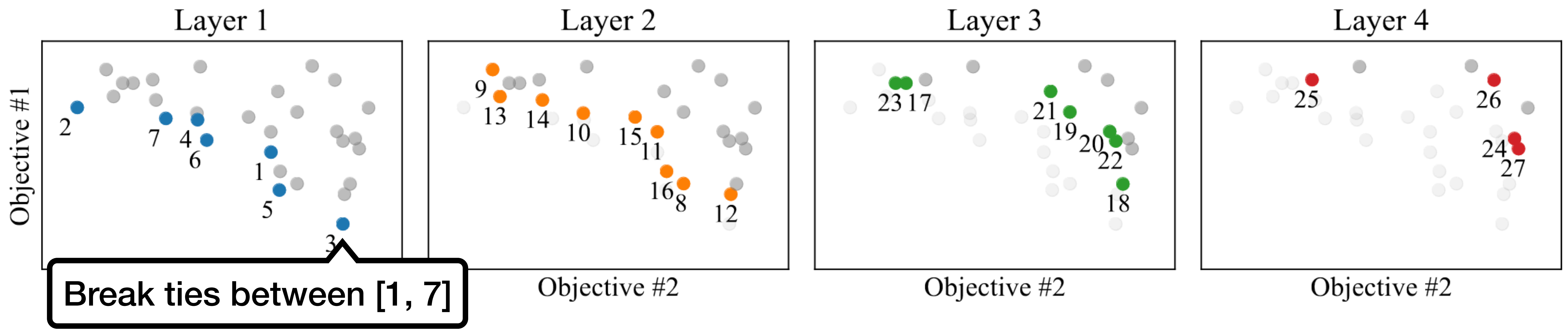


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n\times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathscr{P}(y)$, break ties with an heuristic
  - Heuristic choices aims at selecting a subset with a good coverage:

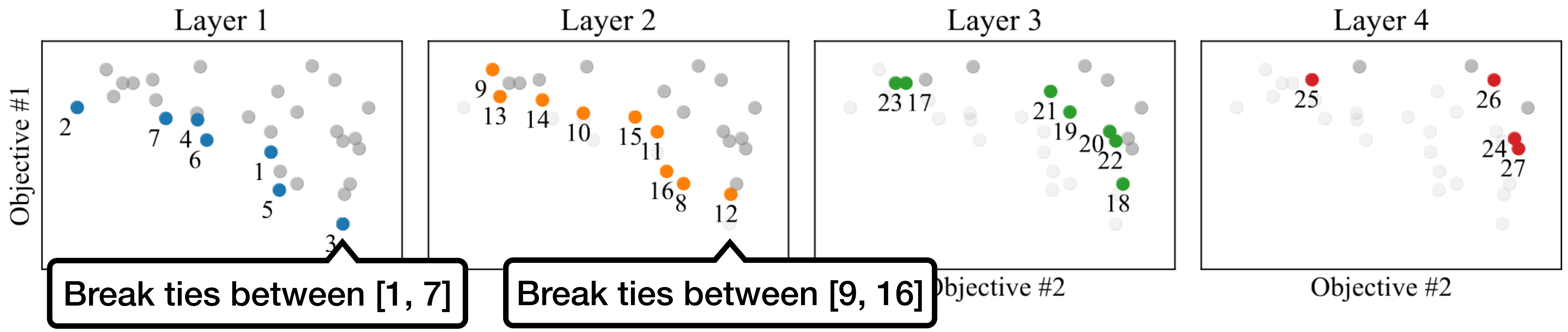


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathscr{P}(y)$, break ties with an heuristic
  - Heuristic choices aims at selecting a subset with a good coverage:
    - Crowding distance

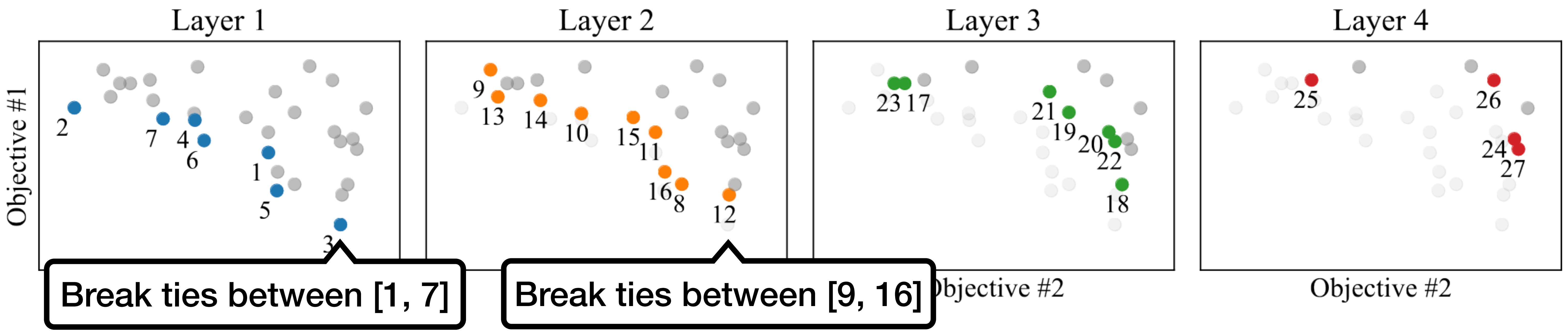


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathcal{P}(y)$, break ties with an heuristic
  - Heuristic choices aims at selecting a subset with a good coverage:
    - Crowding distance
    - Epsilon-net

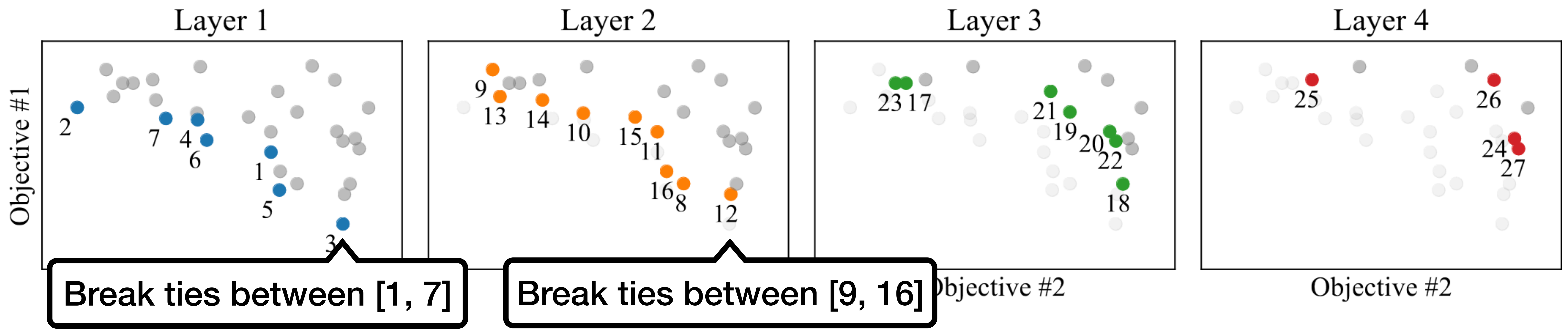


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Multiobjective optimization

Non dominated sort: extending sorting to multiple objectives

- How to *rank* n observations $y \in \mathbb{R}^{n \times d}$ when we have $d > 1$ objectives?
- Non dominated sort (aka onion sort):
  - Compute the Pareto front of y, break ties with an heuristic
  - Compute the Pareto front of $y \backslash \mathscr{P}(y)$, break ties with an heuristic
  - Heuristic choices aims at selecting a subset with a good coverage:
    - Crowding distance
    - Epsilon-net
    - …

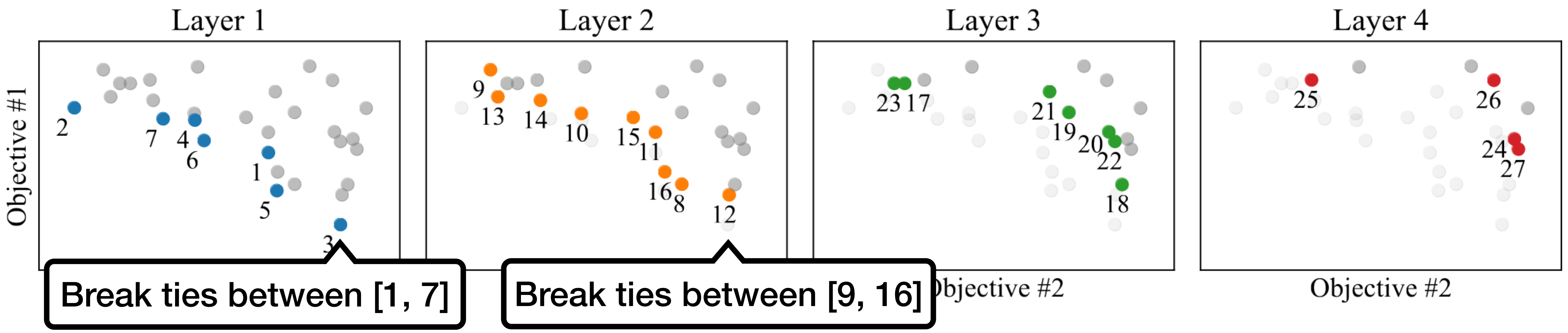
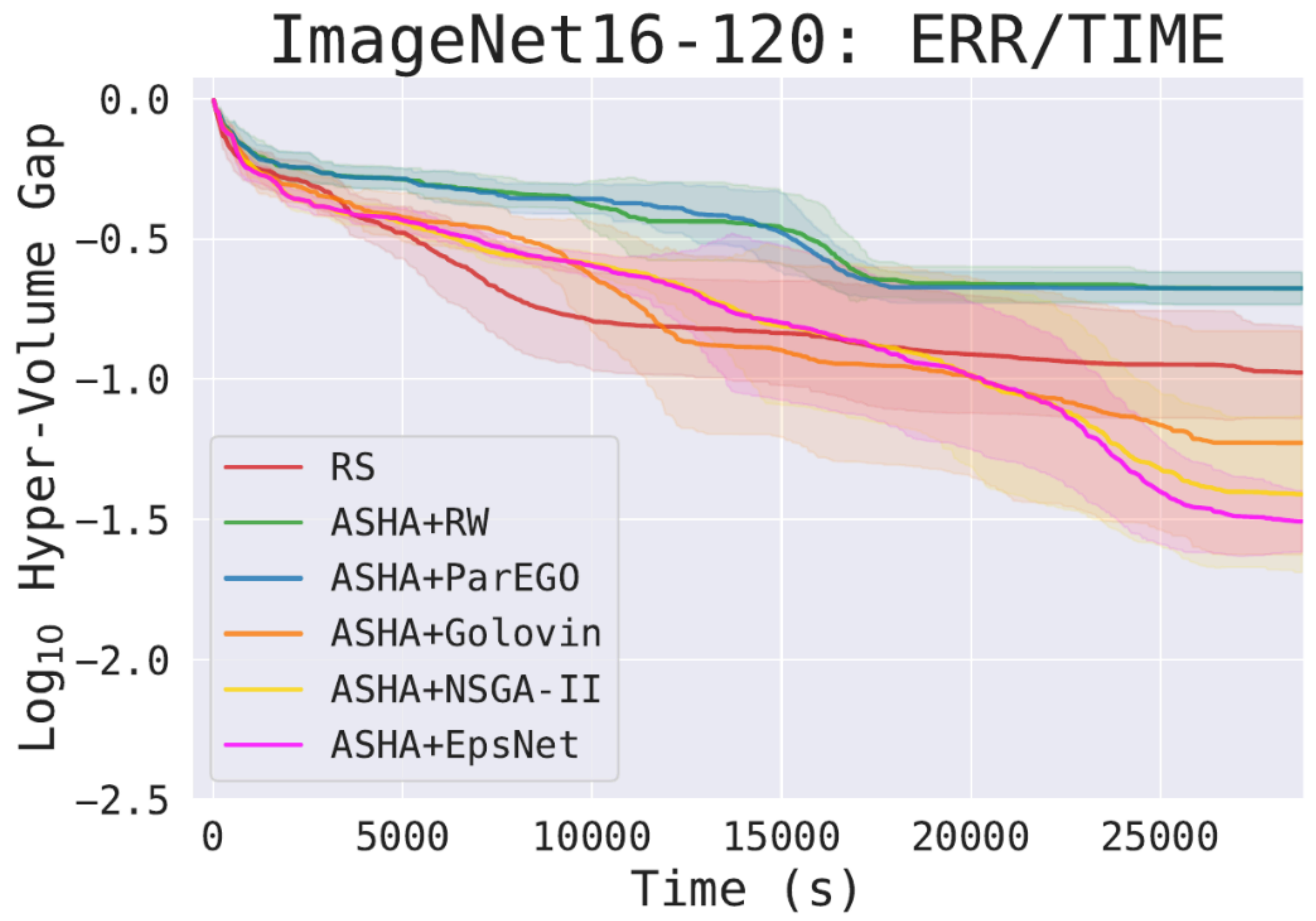
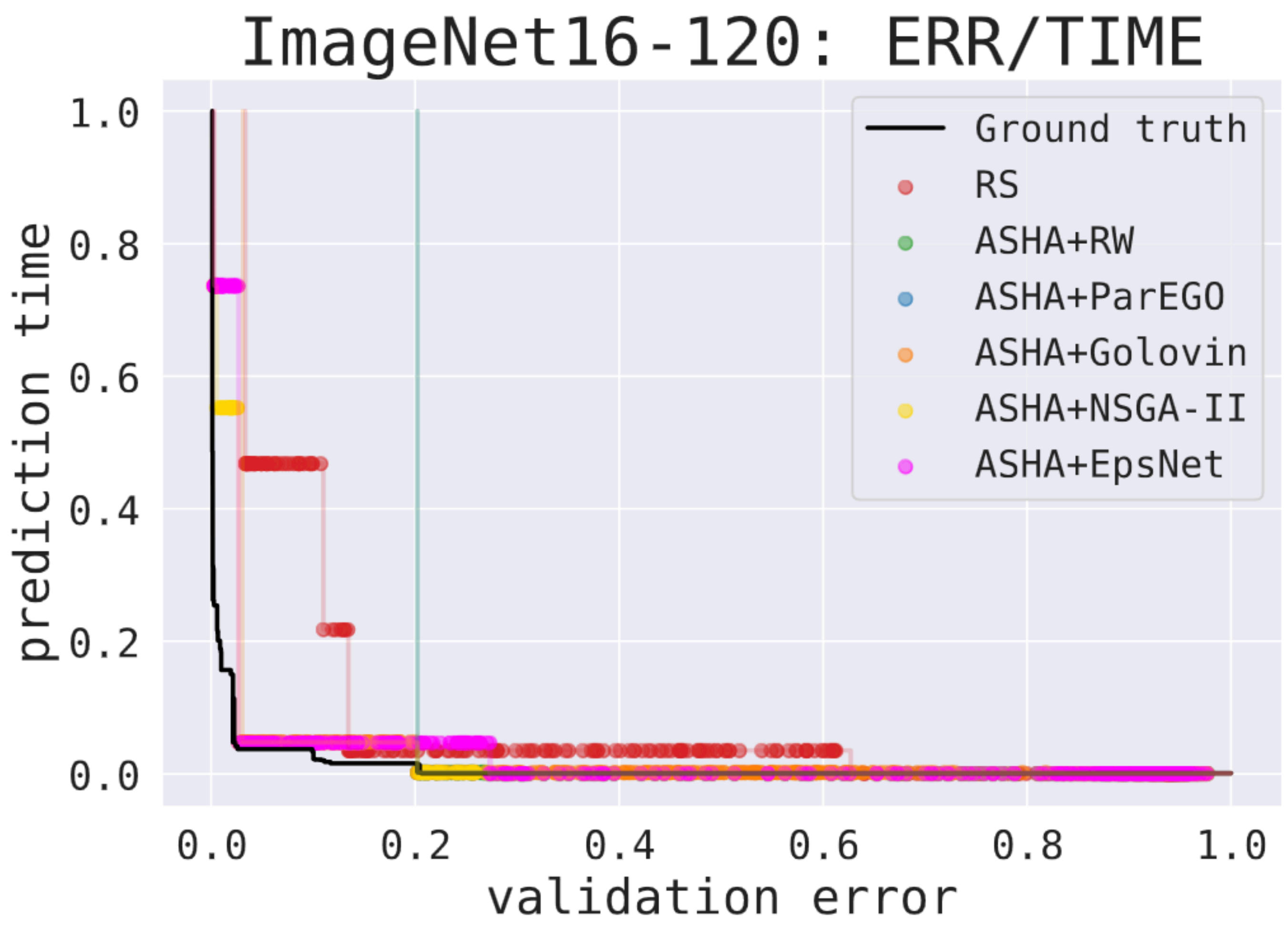


Figure 3: Illustration of non-dominated sorting. The layers show the partitioning of the data in Pareto fronts. The numbers depict the overall rank by computing the ϵ-net within each layer.

Extending Multifidelity to multi-objective

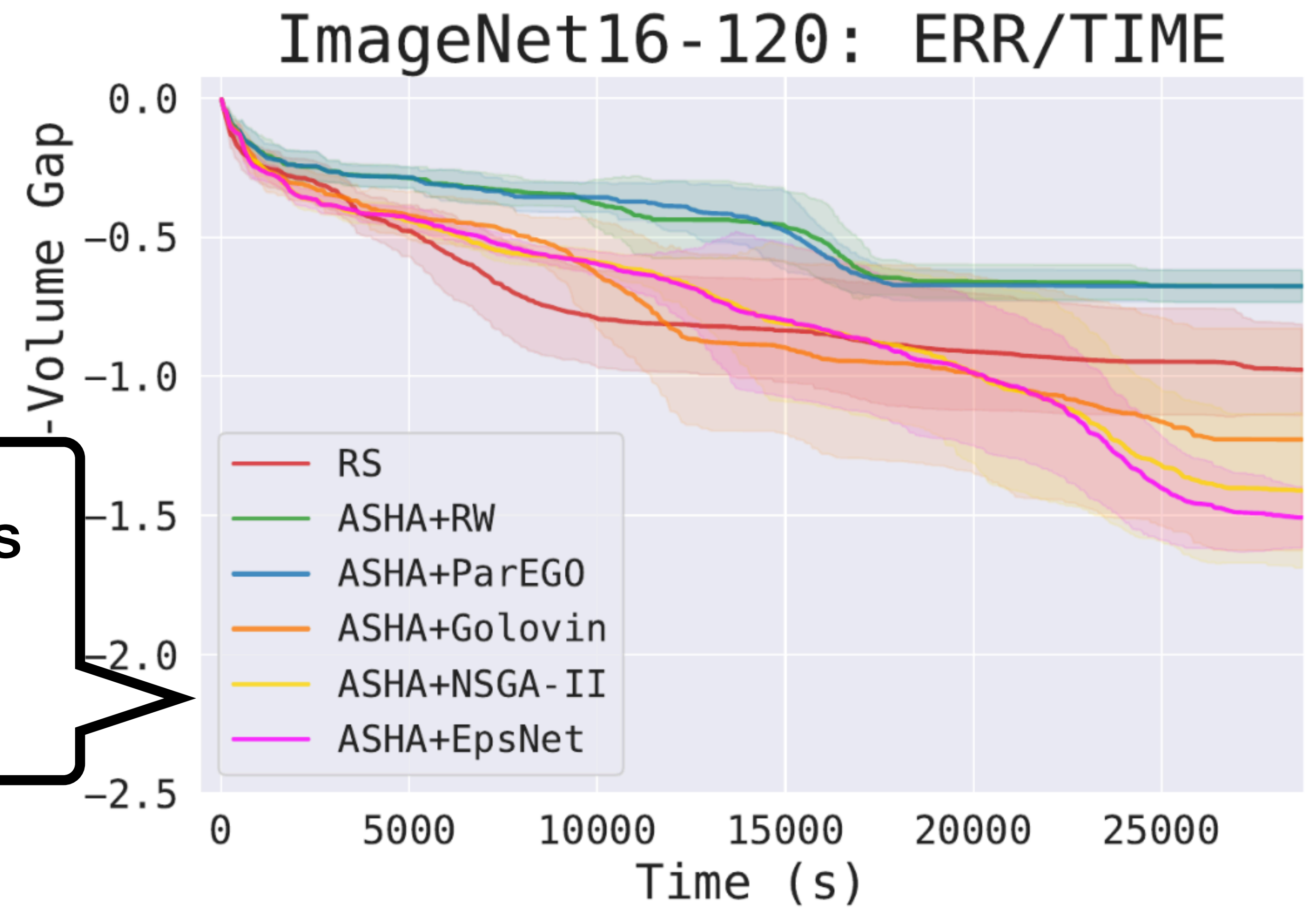


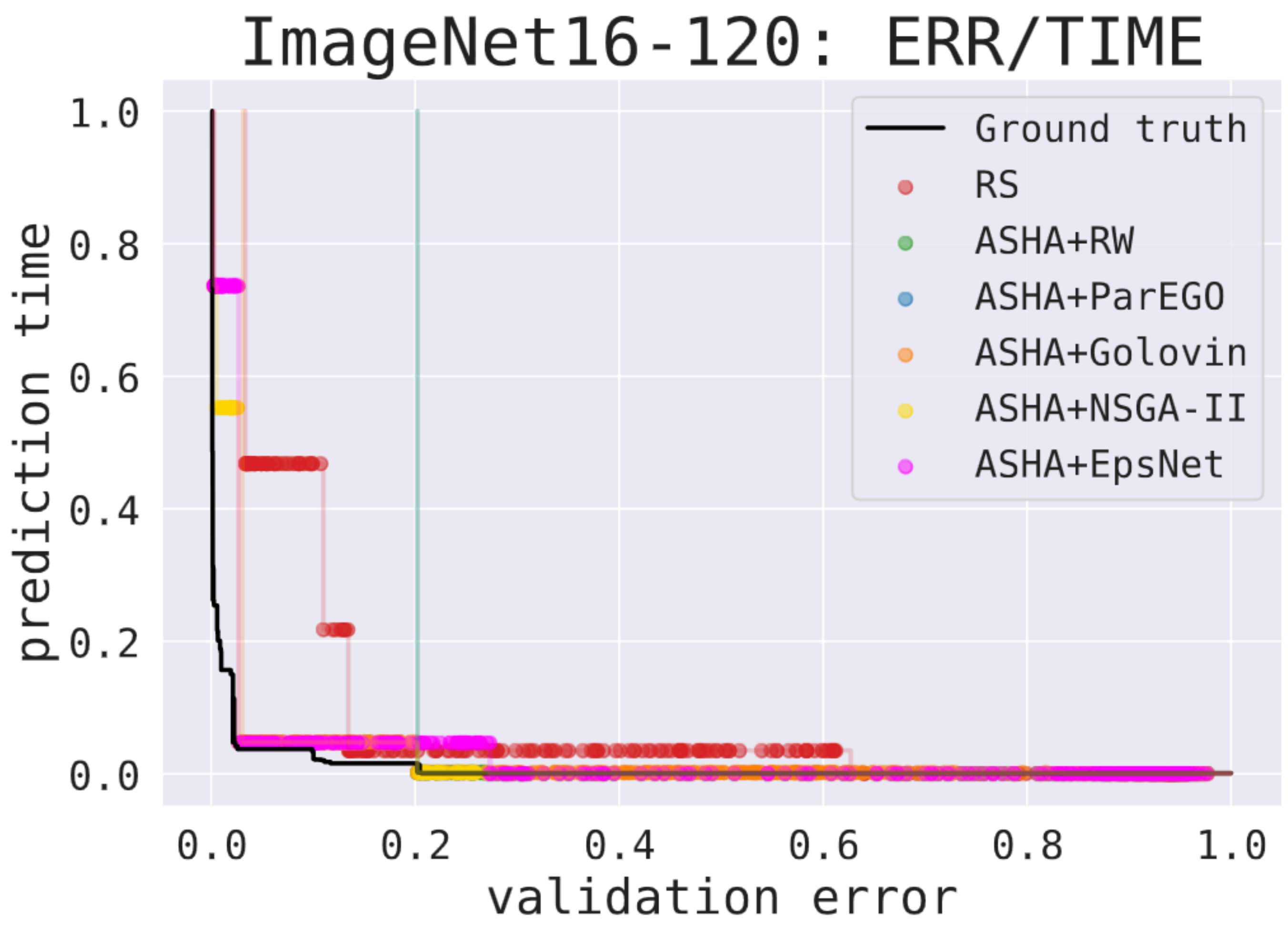


Multi-objective Asynchronous Successive Halving [Schmucker 2021]

Extending Multifidelity to multi-objective

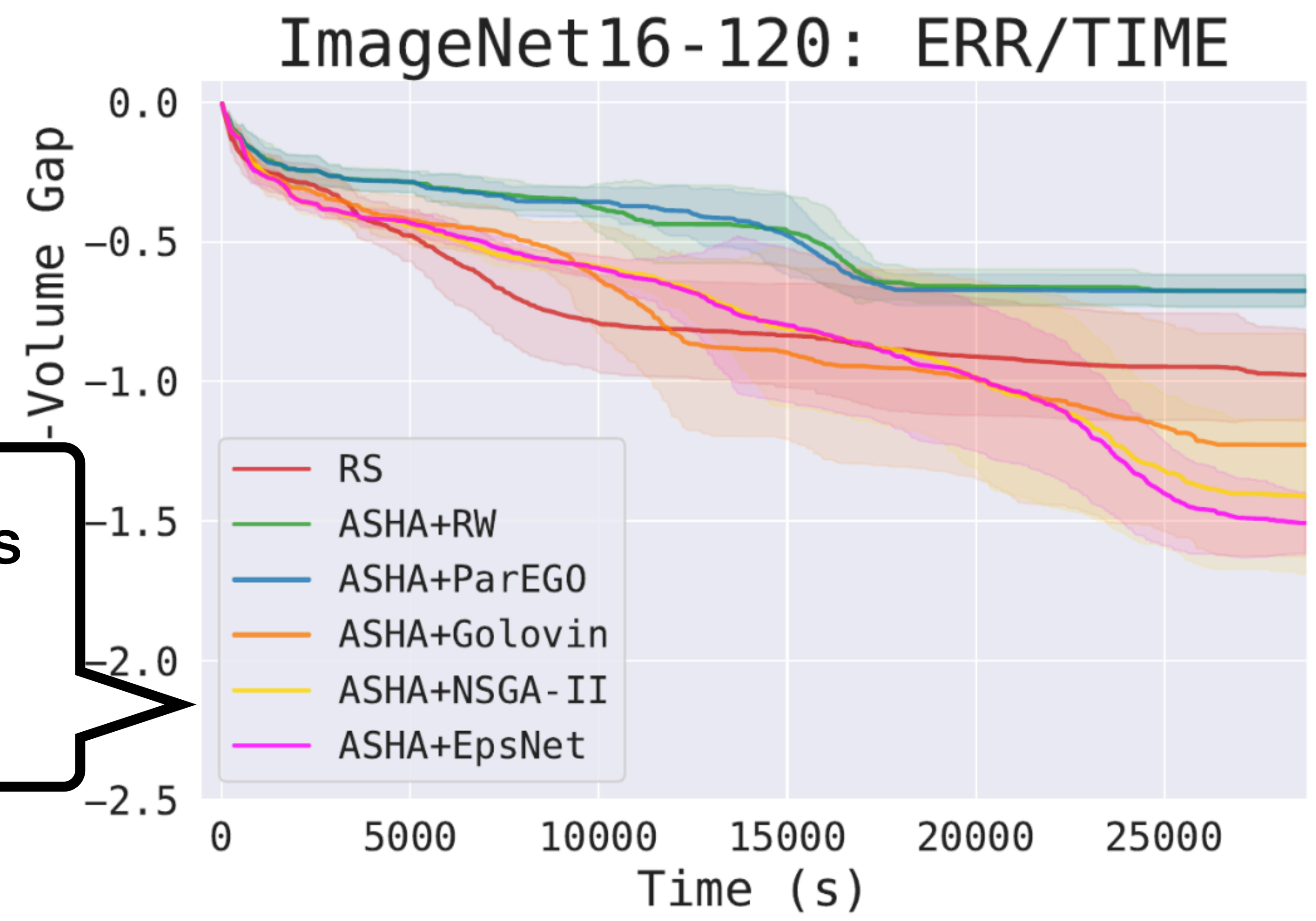
Non-dominated sort, works much better than scalarization!



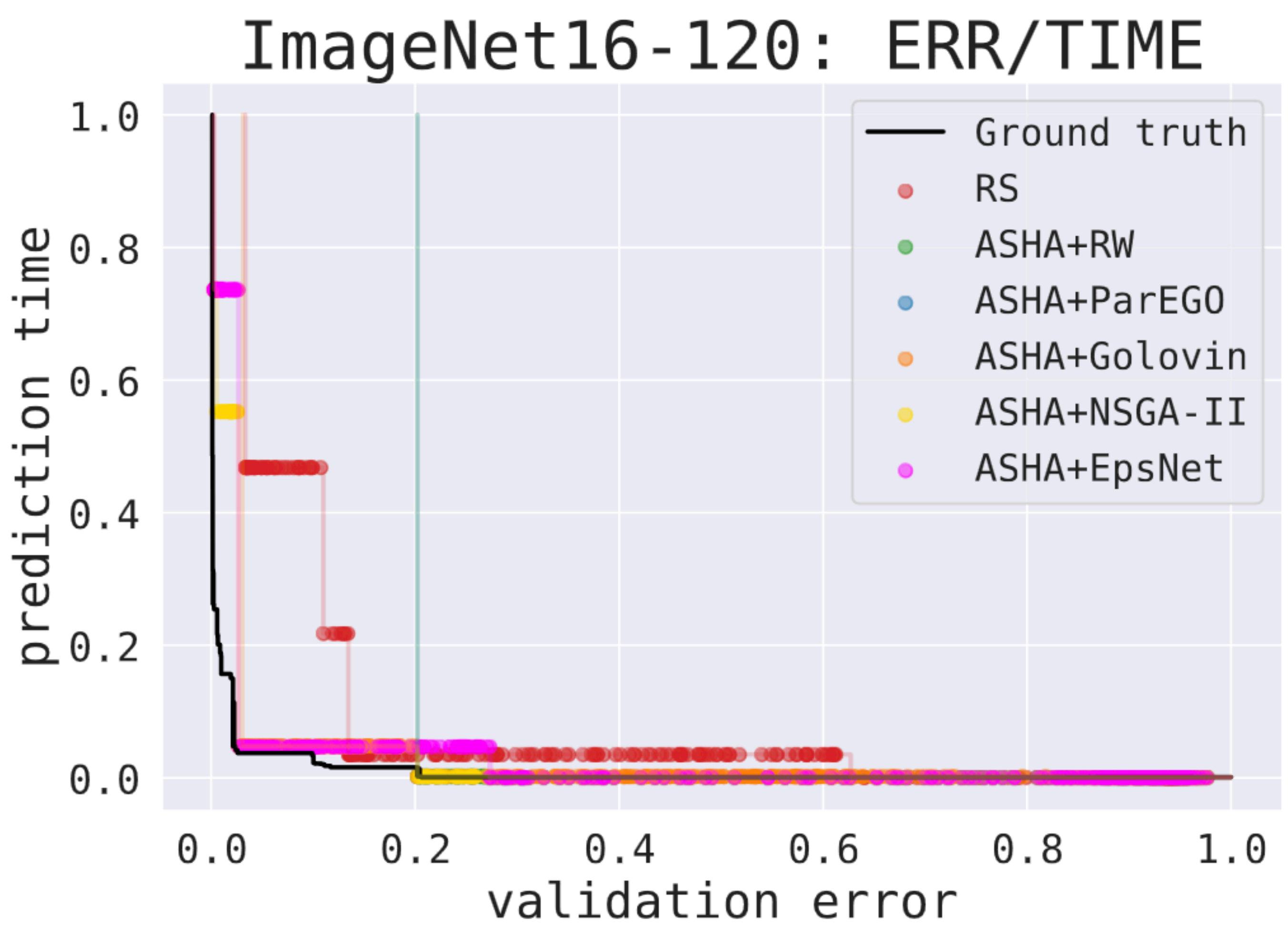


Multi-objective Asynchronous Successive Halving [Schmucker 2021]

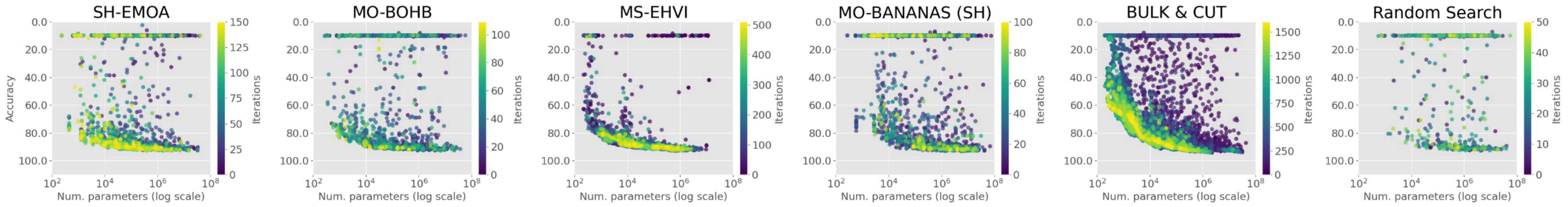
Extending Multifidelity to multi-objective



Non-dominated sort, works much better than scalarization!



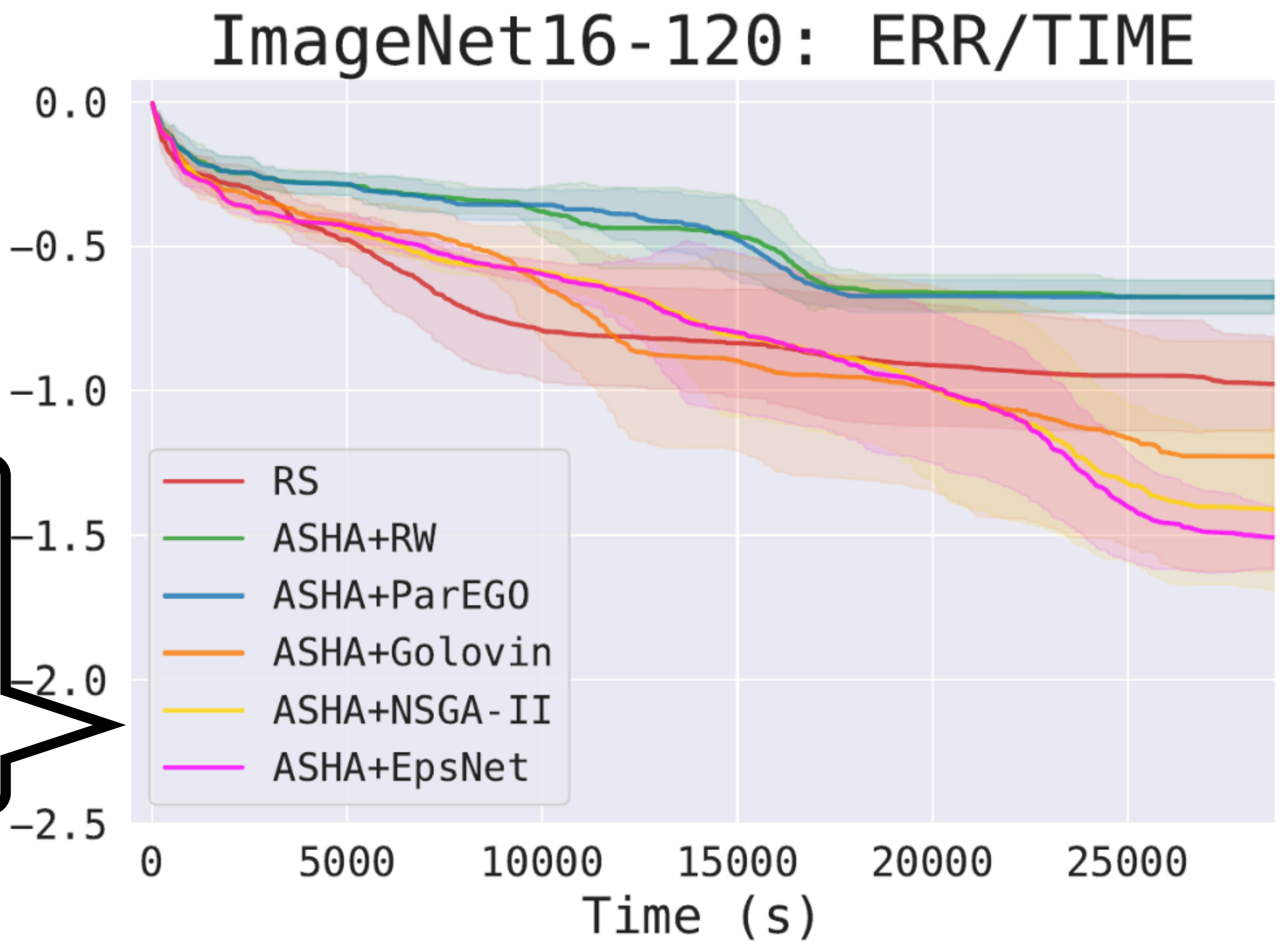
Multi-objective Asynchronous Successive Halving [Schmucker 2021]



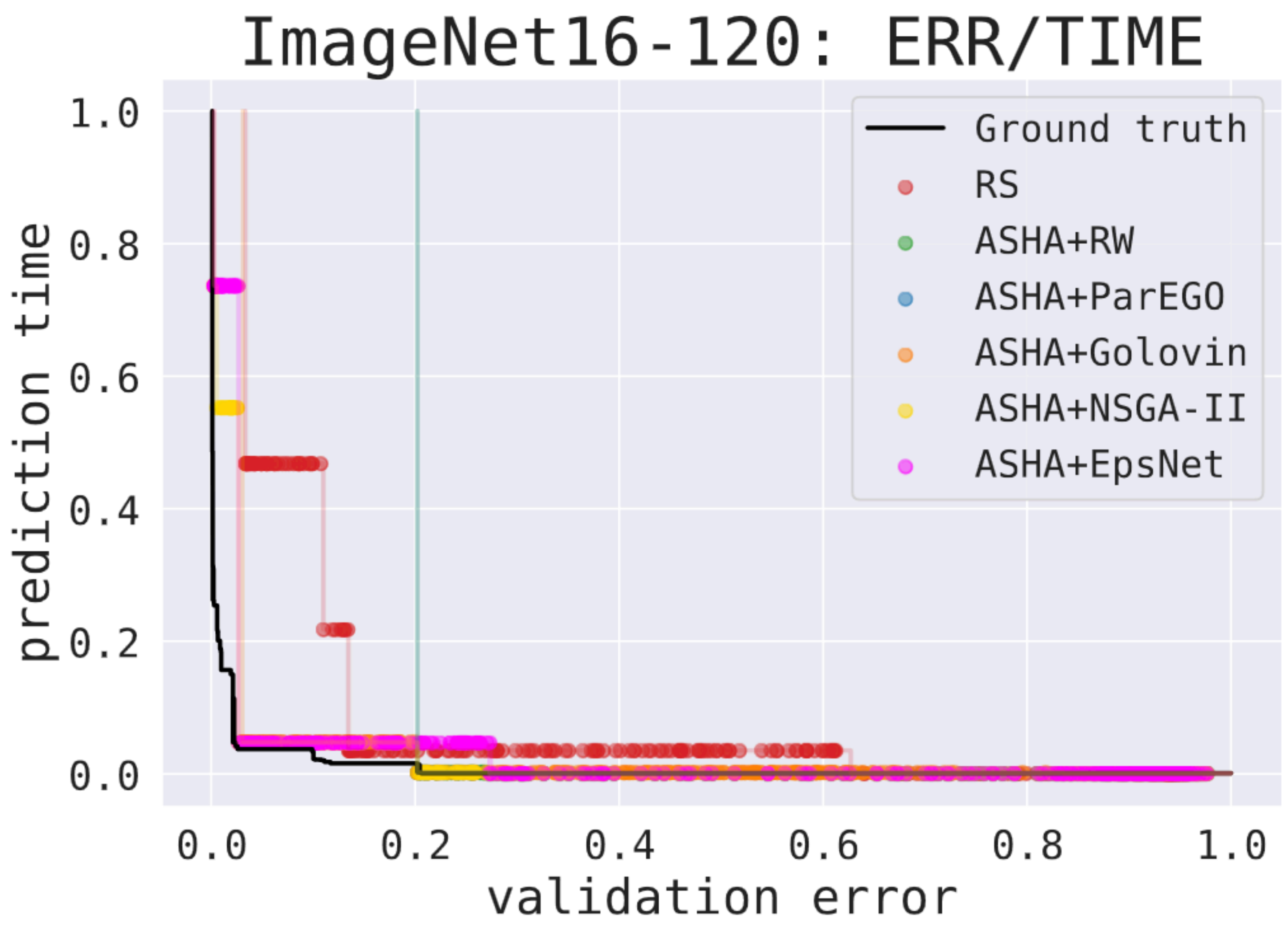
(b) Sampled configurations on Fashion-MNIST dataset.

Bag of Baselines for Multi-objective Joint Neural Architecture Search and Hyperparameter Optimization [Guererro Viu 2021]

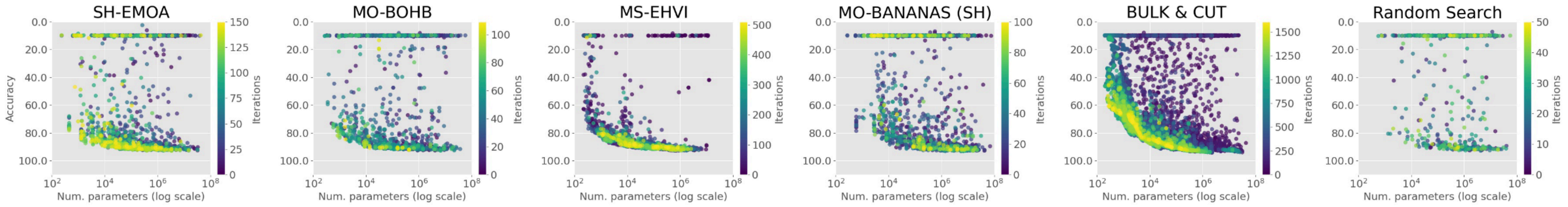
Extending Multifidelity to multi-objective



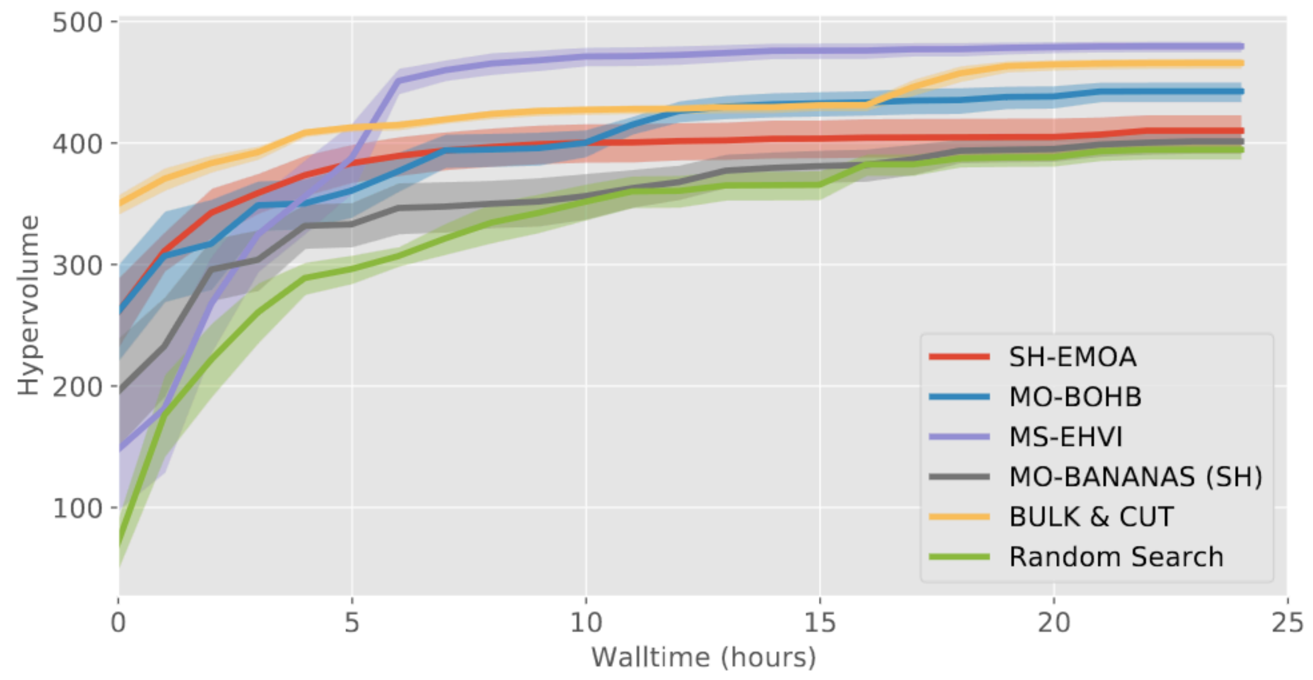
Non-dominated sort, works much better than scalarization!



Multi-objective Asynchronous Successive Halving [Schmucker 2021]



(b) Sampled configurations on Fashion-MNIST dataset.

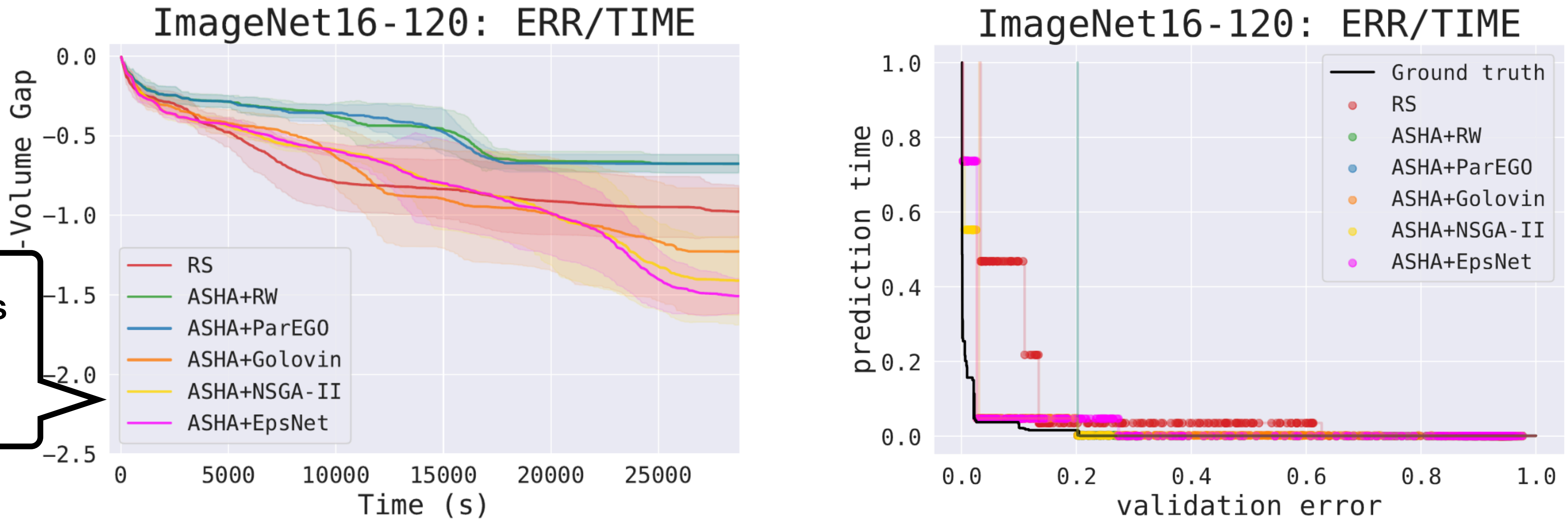


(d) Fashion-MNIST dataset on test split.

Bag of Baselines for Multi-objective Joint Neural Architecture Search and Hyperparameter Optimization [Guererro Viu 2021]

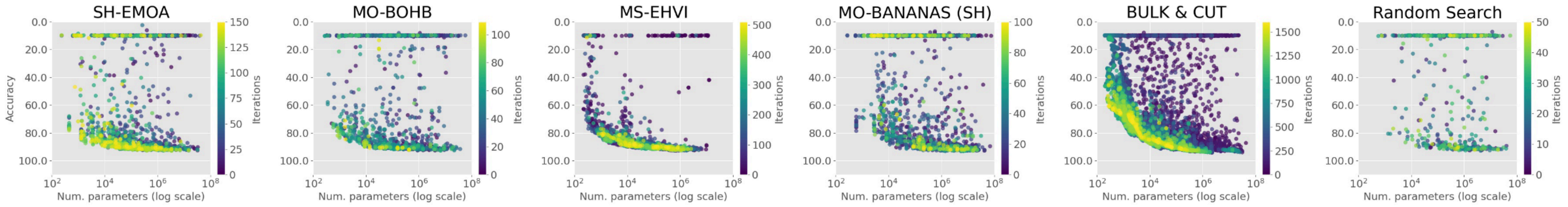
Extending Multifidelity to multi-objective

Non-dominated sort, works much better than scalarization!

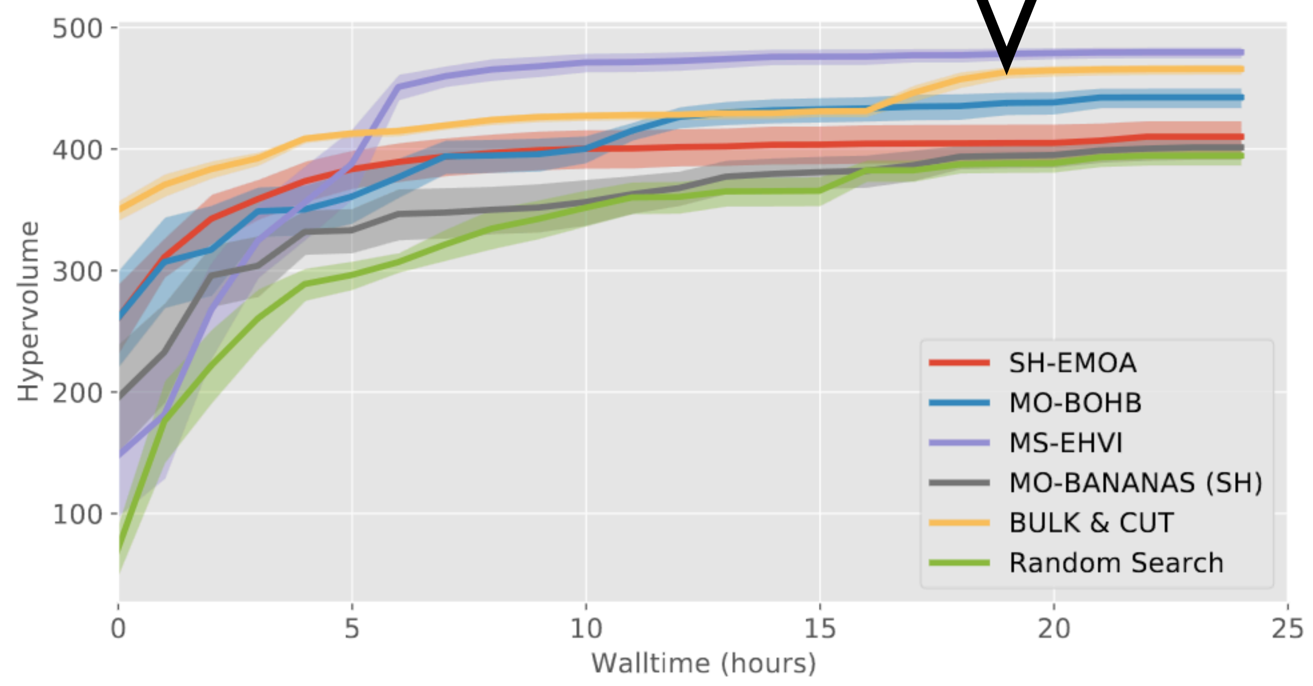


Multi-objective Asynchronous Successive Halving [Schmucker 2021]

Non-dominated sort, works well in general…



(b) Sampled configurations on Fashion-MNIST dataset.



(d) Fashion-MNIST dataset on test split.

Bag of Baselines for Multi-objective Joint Neural Architecture Search and Hyperparameter Optimization [Guererro Viu 2021]

A case study: tuning LLM judges

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

LLM →

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

LLM → "Model 1"

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

> **Please say which model is better when answering the question: "What is some cool music from the 1920s?"**
>
> **Model 1**: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …
>
> **Model 2**: The 1920s, often referred to as the "Roaring Twenties," was a period…

$\xrightarrow{\text{LLM}}$ "Model 1"

- Pros: cheaper than human annotations & can evaluate open-ended text

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

$\xrightarrow{\text{LLM}}$ "Model 1"

- Pros: cheaper than human annotations & can evaluate open-ended text
- Cons: can be biased, very often rely on GPT-4 or closed models

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

> **Please say which model is better when answering the question: "What is some cool music from the 1920s?"**
>
> **Model 1**: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …
>
> **Model 2**: The 1920s, often referred to as the "Roaring Twenties," was a period…

$\xrightarrow{\text{LLM}}$ "Model 1"

- Pros: cheaper than human annotations & can evaluate open-ended text
- Cons: can be biased, very often rely on GPT-4 or closed models
- Can be used for model selection, leaderboard…

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: "What is some cool music from the 1920s?"

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

LLM

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13)	57.5%	51.3%
2	GPT-4 Turbo (04/09)	55.0%	46.1%
3	Yi-Large Preview	51.9%	57.5%
4	GPT-4 Preview (11/06)	50.0%	50.0%
5	Claude 3 Opus (02/29)	40.5%	29.1%
6	GPT-4	38.1%	23.6%
7	Qwen1.5 72B Chat	36.6%	26.5%
8	GPT-4 (03/14)	35.3%	22.1%
9	Claude 3 Sonnet (02/29)	34.9%	25.6%
10	Llama 3 70B Instruct	34.4%	33.2%

Alpaca Eval Leaderboard: winrate against GPT4-turbo

- Pros: cheaper than human annotations & can evaluate open-ended text
- Cons: can be biased, very often rely on GPT-4 or closed models
- Can be used for model selection, leaderboard…

LLM as a judge

Quick recap

- Idea 💡: ask LLM to tell which LLM output is better

Please say which model is better when answering the question: “What is some cool music from the 1920s?”

Model 1: The 1920s was a fantastic decade for music, marked by the rise of jazz, blues …

Model 2: The 1920s, often referred to as the "Roaring Twenties," was a period…

LLM

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13)	57.5%	51.3%
2	GPT-4 Turbo (04/09)	55.0%	46.1%
3	Yi-Large Preview	51.9%	57.5%
4	GPT-4 Preview (11/06)	50.0%	50.0%
5	Claude 3 Opus (02/29)	40.5%	29.1%
6	GPT-4	38.1%	23.6%
7	Qwen1.5 72B Chat	36.6%	26.5%
8	GPT-4 (03/14)	35.3%	22.1%
9	Claude 3 Sonnet (02/29)	34.9%	25.6%
10	Llama 3 70B Instruct	34.4%	33.2%

Alpaca Eval Leaderboard: winrate against GPT4-turbo

- Pros: cheaper than human annotations & can evaluate open-ended text
- Cons: can be biased, very often rely on GPT-4 or closed models
- Can be used for model selection, leaderboard…
- **Our Goal**: Provide high-accuracy, low-cost LLM judges with **open** models

Tuning LLM judges

Search space

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user
prompt>,
 "explanation": <your explanation on
why you think A or B is better>,
 "score_A": <between 0 and 10 to
rate the quality of A>,
 "score_B": <between 0 and 10 to
rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user
prompt>,
 "explanation": <your explanation on
why you think A or B is better>,
 "score_A": <between 0 and 10 to
rate the quality of A>,
 "score_B": <between 0 and 10 to
rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user
prompt>,
 "explanation": <your explanation on
why you think A or B is better>,
 "score_A": <between 0 and 10 to
rate the quality of A>,
 "score_B": <between 0 and 10 to
rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
  "answer": <your answer to the user prompt>,
  "explanation": <your explanation on why you think A or B is better>,
  "score_A": <between 0 and 10 to rate the quality of A>,
  "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting
- Inference hyperparameters:

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting
- Inference hyperparameters:
  - 7 open-weight options (Llama3, Qwen2.5, Gemma 2 at different size)

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user prompt>,
 "explanation": <your explanation on why you think A or B is better>,
 "score_A": <between 0 and 10 to rate the quality of A>,
 "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting
- Inference hyperparameters:
  - 7 open-weight options (Llama3, Qwen2.5, Gemma 2 at different size)
  - 4 temperatures

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
  "answer": <your answer to the user prompt>,
  "explanation": <your explanation on why you think A or B is better>,
  "score_A": <between 0 and 10 to rate the quality of A>,
  "score_B": <between 0 and 10 to rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting
- Inference hyperparameters:
  - 7 open-weight options (Llama3, Qwen2.5, Gemma 2 at different size)
  - 4 temperatures
- Boolean to average predictions with both orders

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer":  <your answer to the user
prompt>,
 "explanation":  <your explanation on
why you think A or B is better>,
 "score_A":  <between 0 and 10 to
rate the quality of A>,
 "score_B":  <between 0 and 10 to
rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

Tuning LLM judges

Search space

- Prompt:
  - Output format: 5 options
  - 3 booleans to ask LLM to provide
    - answer: its own answer to the instruction as proposed
    - example: an example of a judgement
    - explanation: its explanation on the given preference
  - Boolean to use json formatting
- Inference hyperparameters:
  - 7 open-weight options (Llama3, Qwen2.5, Gemma 2 at different size)
  - 4 temperatures
- Boolean to average predictions with both orders

Prompt Template

You are a highly efficient assistant, please evaluate and select the best large language model based on the quality of their responses to a given instruction.

User Prompt: Who is Geoffrey Hinton?
Assistant A: Geoffrey Hinton is a research scientist.
Assistant B: I do not know who Geoffrey Hinton is.

 # Your Output
Format Description
Your output should follow this format:

```
{
 "answer": <your answer to the user
prompt>,
 "explanation": <your explanation on
why you think A or B is better>,
 "score_A": <between 0 and 10 to
rate the quality of A>,
 "score_B": <between 0 and 10 to
rate the quality of B>
}
```

Your output, do not repeat the input above.

Figure 3. Illustration of the prompt templating approach. We parametrize the prompt with the following hyperparameters: Provide answer, Provide explanation, Provide example, use JSON, output preference format. Given each of the $2^4 \times 5 = 80$ prompt hyperparameters, we generate a prompt like this one.

In total we have $5 \times 2^4 = 80$ prompts and $7 \times 4 \times 80 \times 2 = 4480$ possible judges

😱 Evaluating with Spearman correlation in brute force would cost \$2M (!)

LLM as a judge

Hyperparameter optimization

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…

Performance of some judge hyperparameters for Alpaca Eval

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation but increases cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation** but increases **cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but improves **cost**

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…
  - We could tune the judge with multi-objective optimization!

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation** but increases **cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but improves **cost**

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…
  - We could tune the judge with multi-objective optimization!
  - … but evaluating a single judge configuration costs ~370$ 🥶 (need to evaluate many models to compute Spearman correlation)

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation but increases cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but improves **cost**

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…
    - We could tune the judge with multi-objective optimization!
    - … but evaluating a single judge configuration costs ~370$ 🥶 (need to evaluate many models to compute Spearman correlation)
- Our approach:

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation but increases cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but **improves cost**

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…
  - We could tune the judge with multi-objective optimization!
  - … but evaluating a single judge configuration costs ~370$ 🥶 (need to evaluate many models to compute Spearman correlation)
- Our approach:
  - 1. Identify a cheaper and better metric than Spearman corr.

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation but increases cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but **improves cost**

LLM as a judge

Hyperparameter optimization

- Multiobjective: accuracy & cost per evaluation
- Currently configurations are manually selected and evaluated…
  - We could tune the judge with multi-objective optimization!
  - … but evaluating a single judge configuration costs ~370$ 🥶 (need to evaluate many models to compute Spearman correlation)
- Our approach:
  - 1. Identify a cheaper and better metric than Spearman corr.
  - 2. Use a multifidelity-multiobjective method to tune configurations

Performance of some judge hyperparameters for Alpaca Eval

Chain of thought improves **correlation but increases cost**

	Human agreement	Price [$/1000 examples]	Time [seconds/1000 examples]	Spearman corr.	Pearson corr.
alpaca_eval_gpt4	69.2	13.6	1455	0.97	0.93
alpaca_eval_cot_gpt4_turbo_fn	68.6	6.3	1989	0.97	0.90
alpaca_eval_gpt4_turbo_fn	68.1	5.5	864	0.93	0.82
alpaca_eval_llama3_70b_fn	67.5	0.4	209	0.90	0.86

Llama3-70B instead of GPT4-turbo worsen **correlation** but improves **cost**

How to compare judges?

Identifying a better metric

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
    - 1. Evaluate a list of LLMs on a list of prompts
    - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
    - 3. Compute winrate
    - 4. Measure Spearman correlation with Chatbot Arena ranking

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
    - 1. Evaluate a list of LLMs on a list of prompts
    - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
    - 3. Compute winrate
    - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human

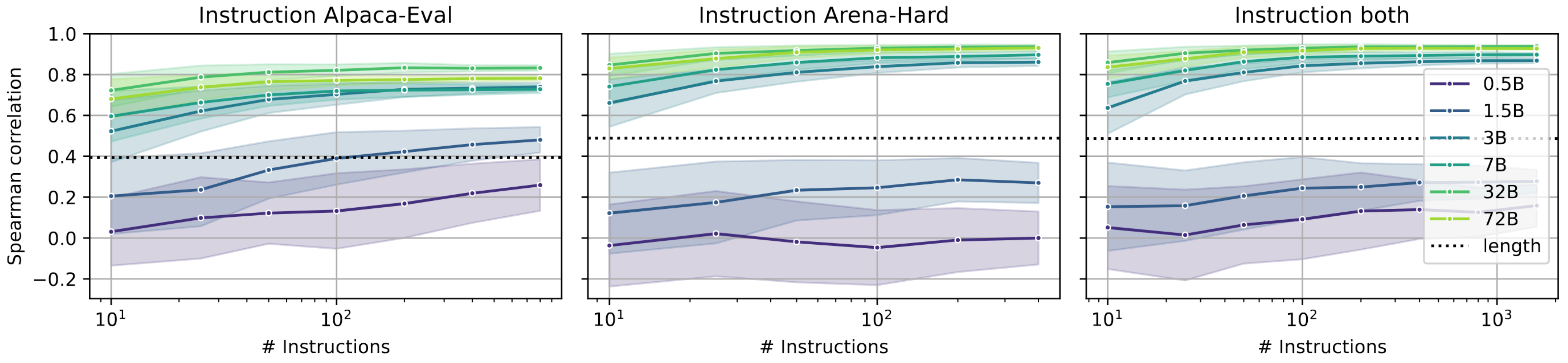
How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

How to compare judges?

Identifying a better metric

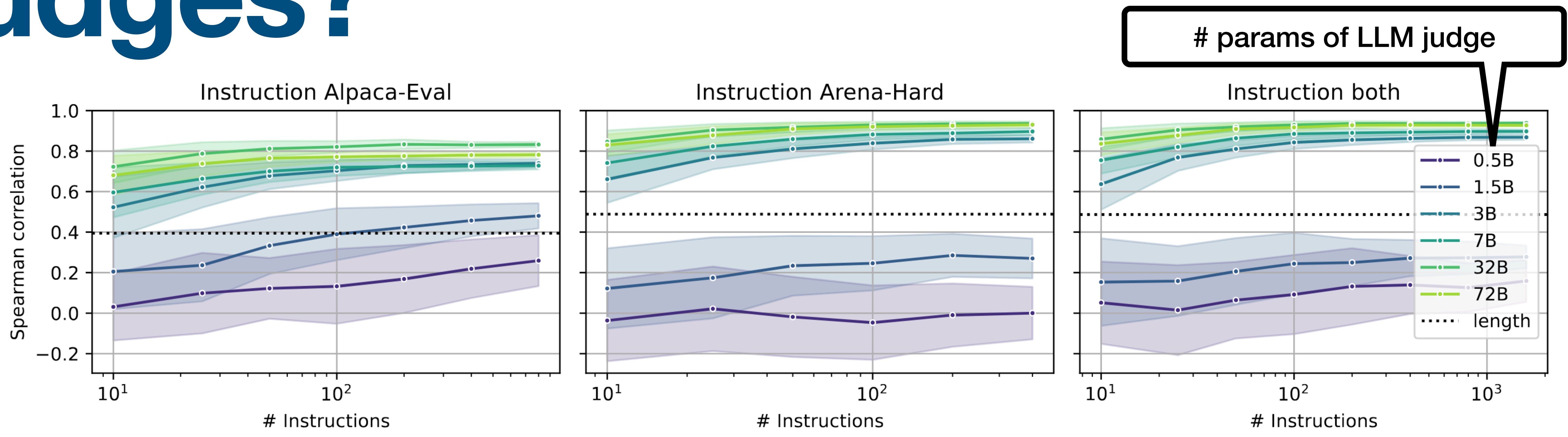


- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

How to compare judges?

Identifying a better metric

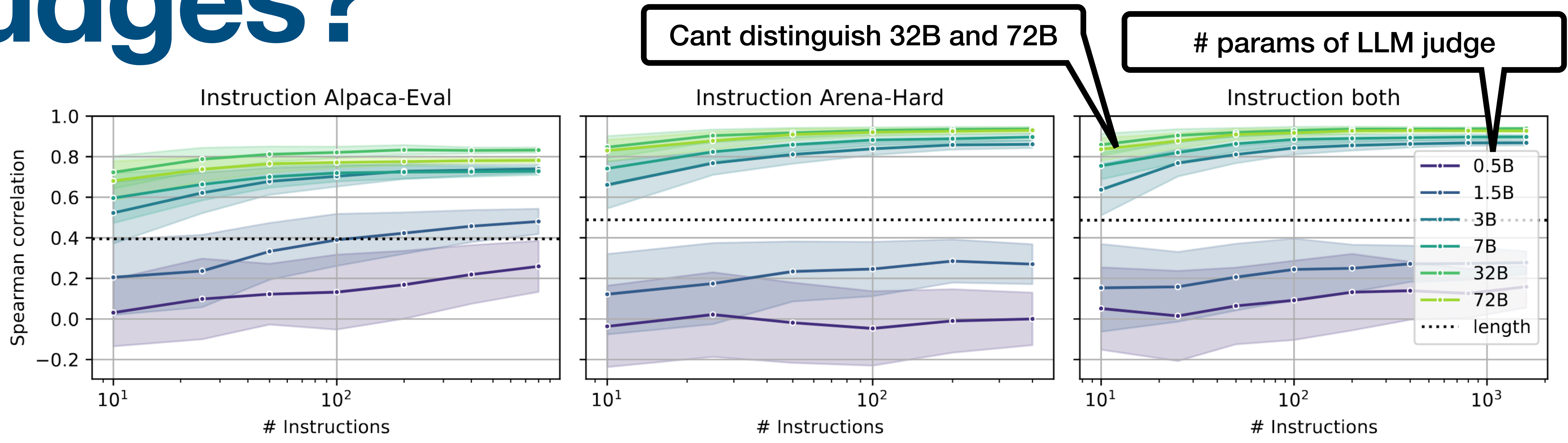
- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?



How to compare judges?

Identifying a better metric

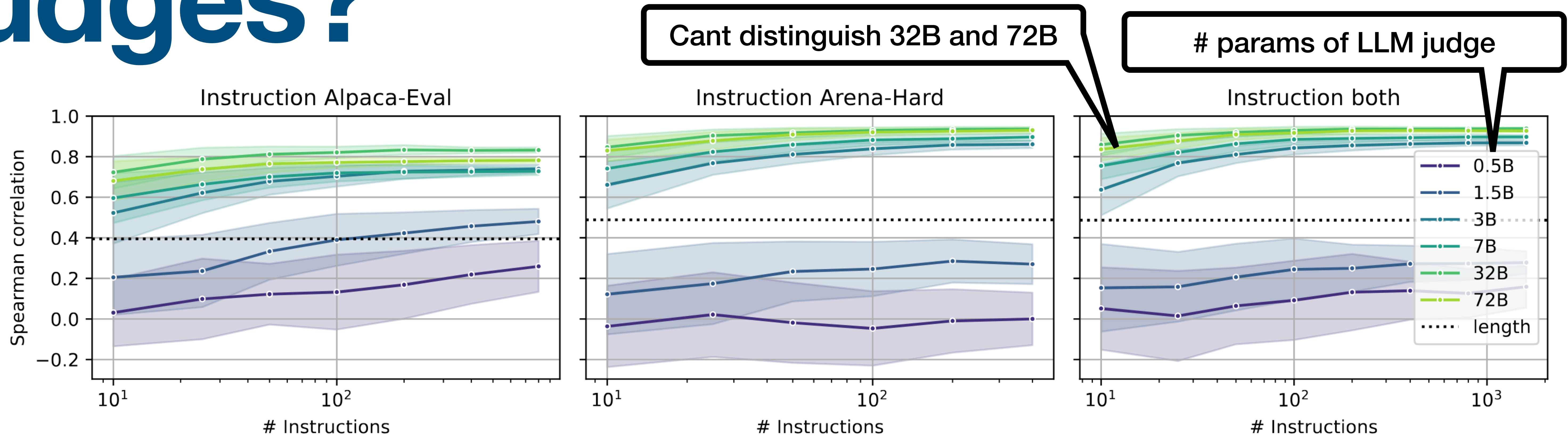
- Option 1: Spearman-correlation
    - 1. Evaluate a list of LLMs on a list of prompts
    - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
    - 3. Compute winrate
    - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
    - Collect list of annotated preferences by human
    - Call LLM judge to annotate preference
    - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

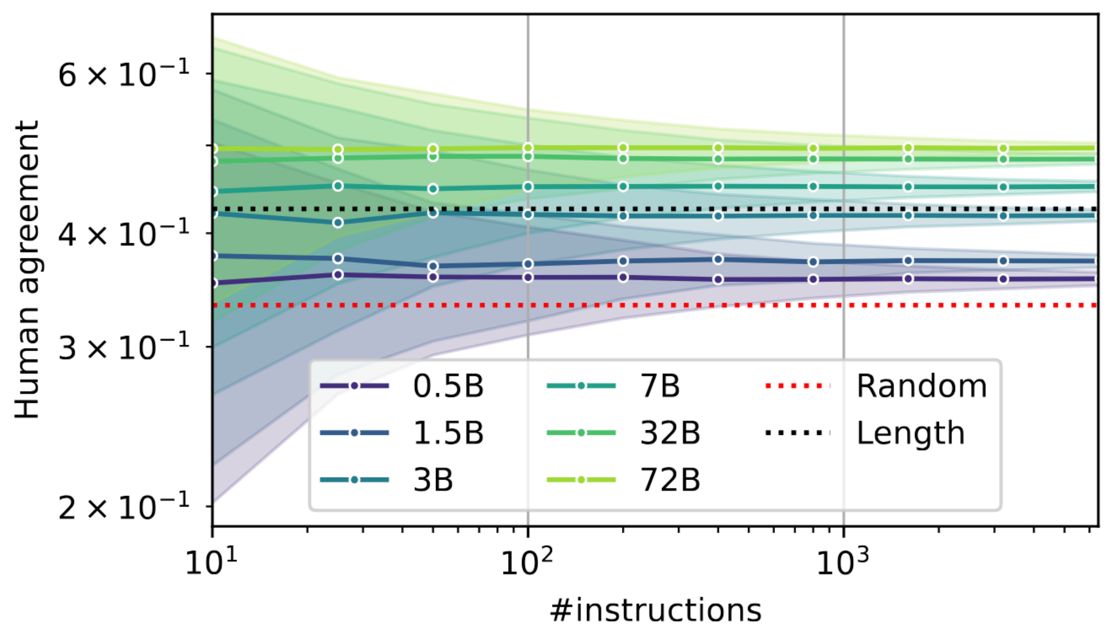


How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
    - 1. Evaluate a list of LLMs on a list of prompts
    - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
    - 3. Compute winrate
    - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
    - Collect list of annotated preferences by human
    - Call LLM judge to annotate preference
    - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

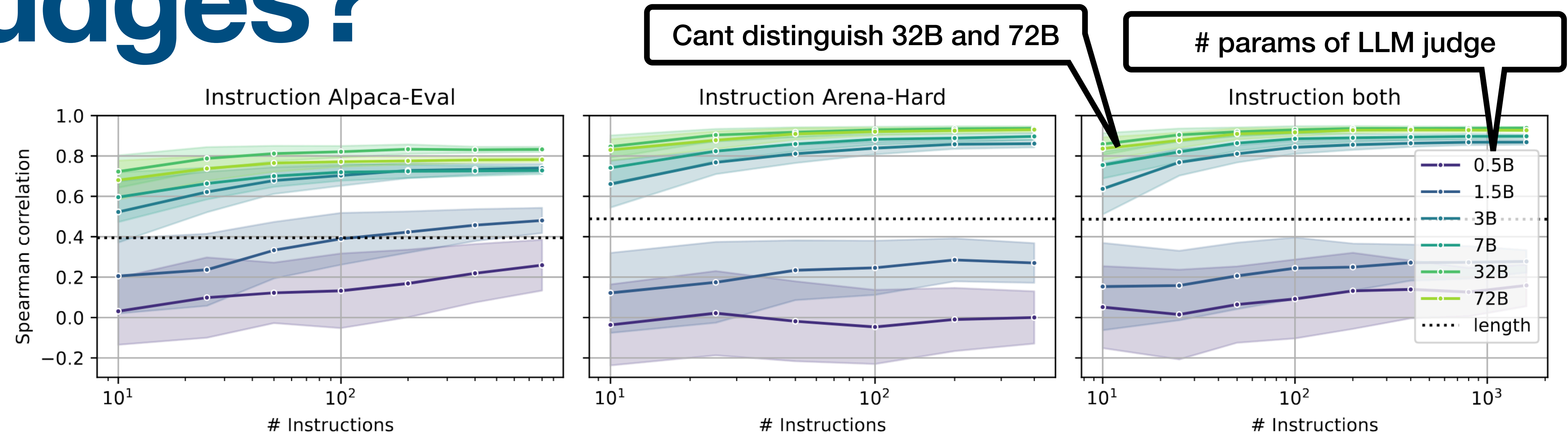


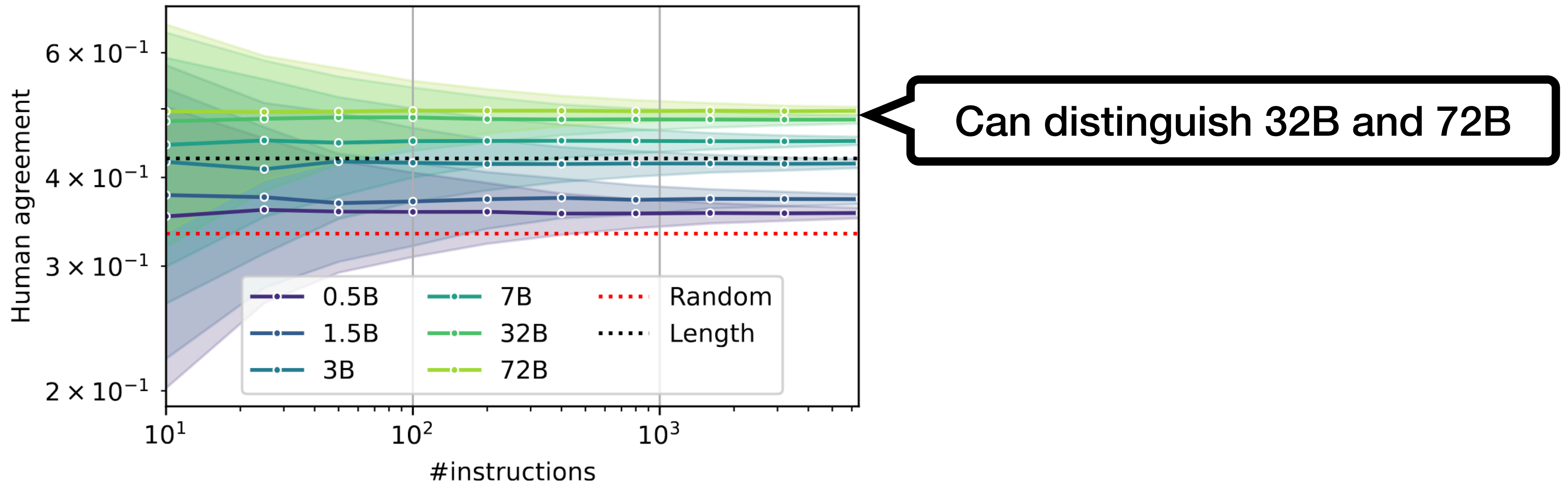


How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

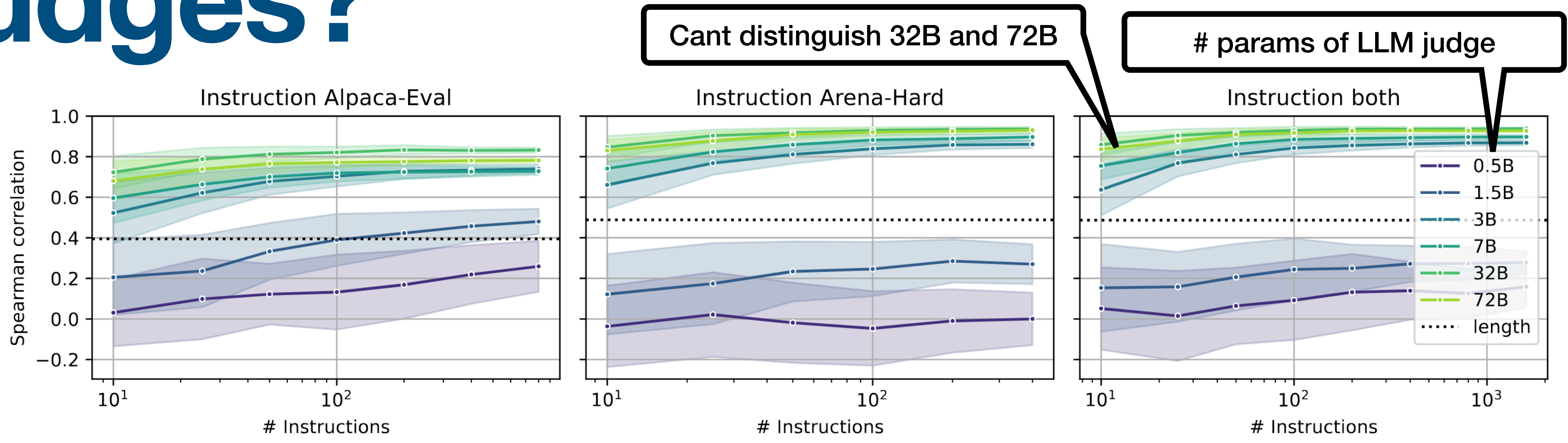


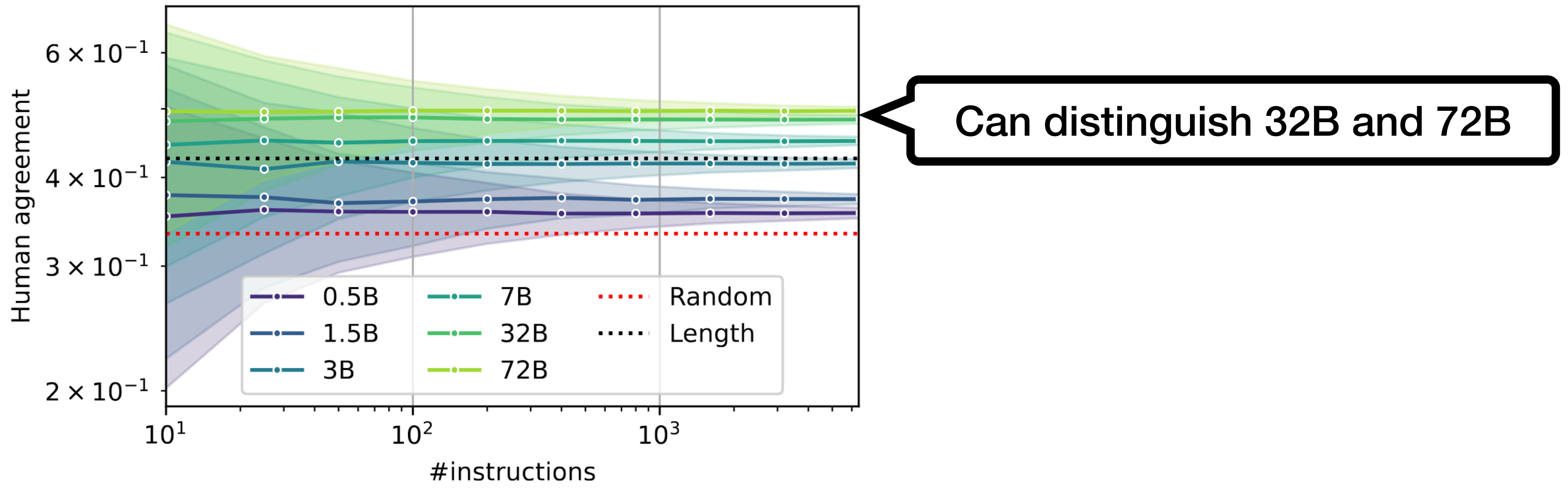


How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?





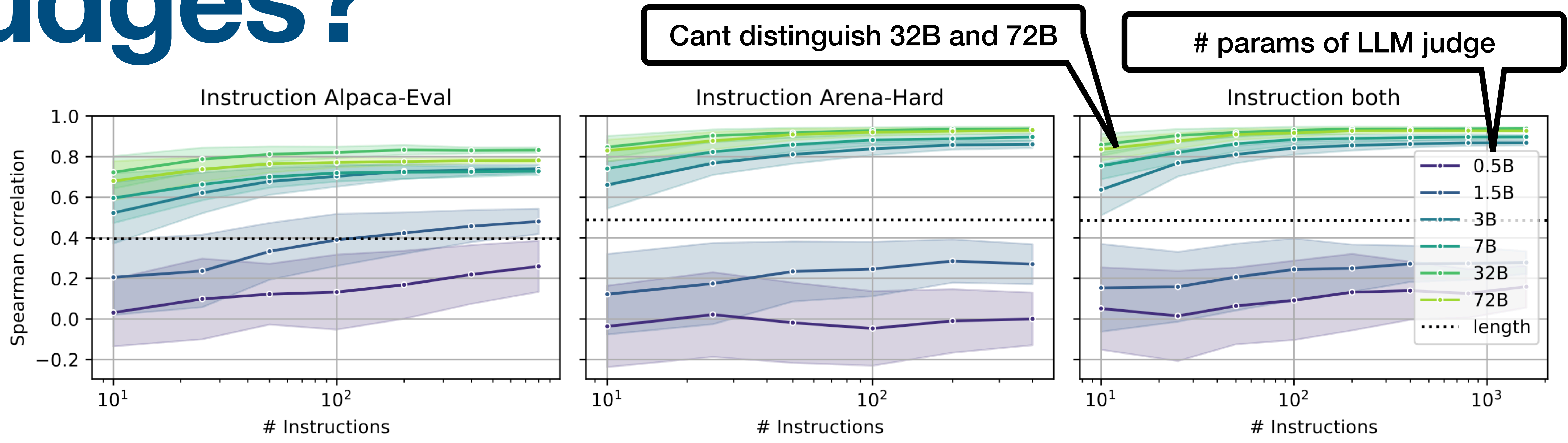
#params	Sp. corr. (↑)	CV(↓)	Hum. agr. (↑)	CV(↓)
0.5B	0.09 ± 0.189	207.60	0.36 ± 0.006	1.70
1.5B	0.33 ± 0.137	41.26	0.37 ± 0.006	1.61
3B	0.82 ± 0.066	8.11	0.42 ± 0.006	1.45
7B	0.83 ± 0.082	9.84	0.45 ± 0.006	1.33
32B	0.90 ± 0.052	5.75	0.48 ± 0.006	1.29
72B	0.88 ± 0.074	8.43	0.50 ± 0.006	1.27

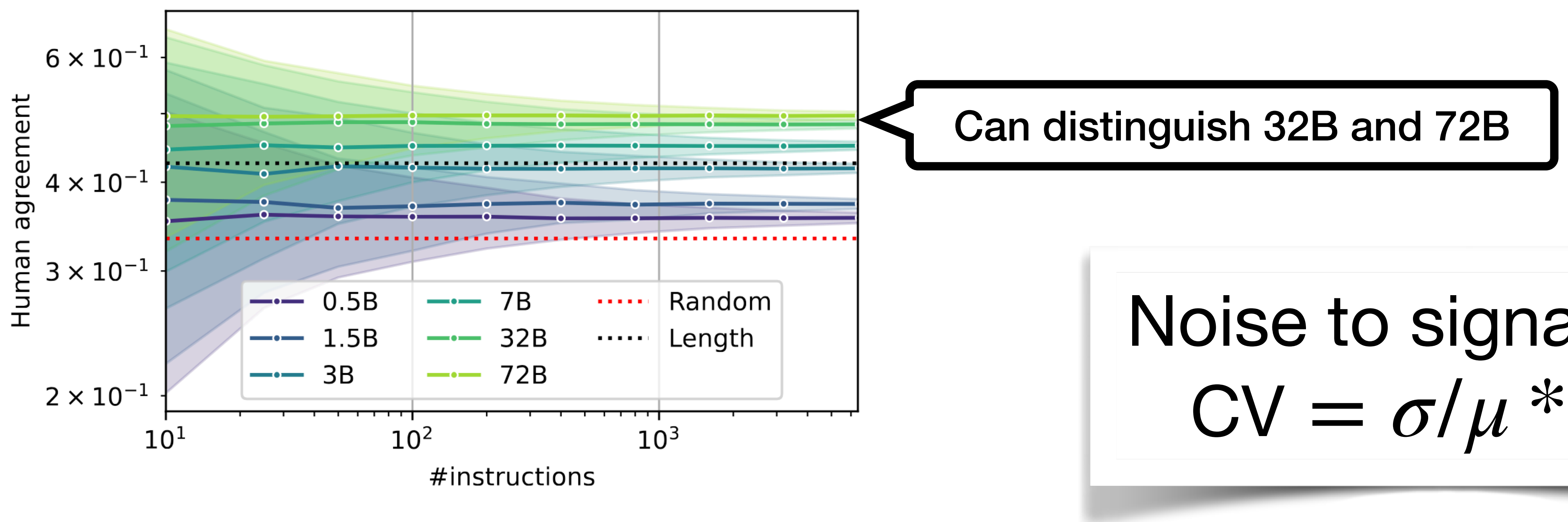
Table 1. Comparison of the variability of Spearman-correlation and Human-agreement metrics when using 6500 random annotations and the same default prompt for all model sizes.

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?





Noise to signal ratio:
$CV = \sigma / \mu * 100$

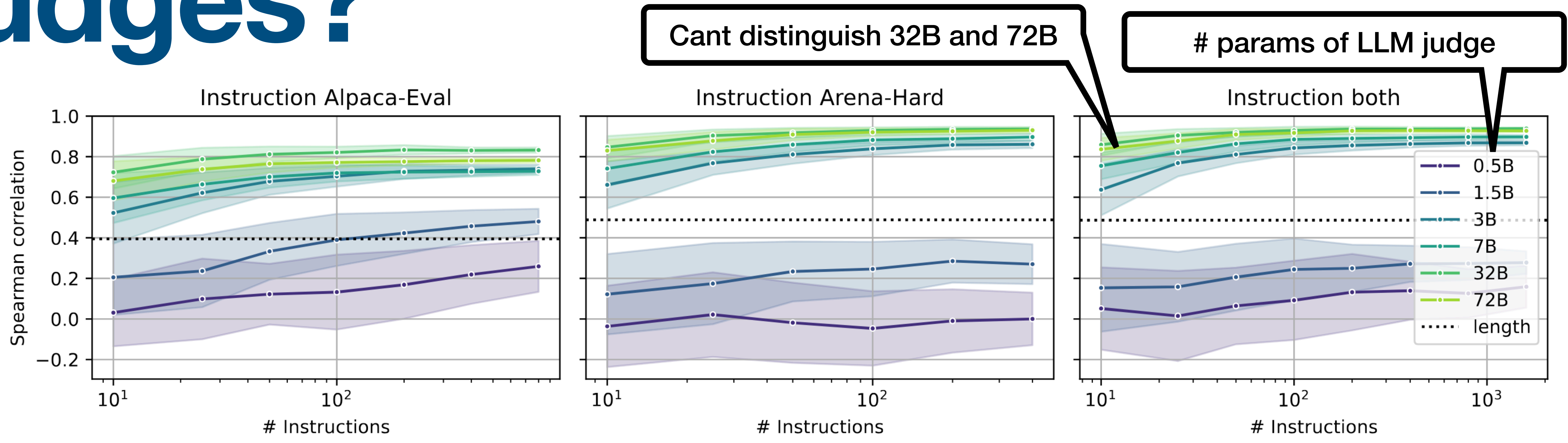
#params	Sp. corr. (↑)	CV(↓)	Hum. agr. (↑)	CV(↓)
0.5B	0.09 ± 0.189	207.60	0.36 ± 0.006	1.70
1.5B	0.33 ± 0.137	41.26	0.37 ± 0.006	1.61
3B	0.82 ± 0.066	8.11	0.42 ± 0.006	1.45
7B	0.83 ± 0.082	9.84	0.45 ± 0.006	1.33
32B	0.90 ± 0.052	5.75	0.48 ± 0.006	1.29
72B	0.88 ± 0.074	8.43	0.50 ± 0.006	1.27

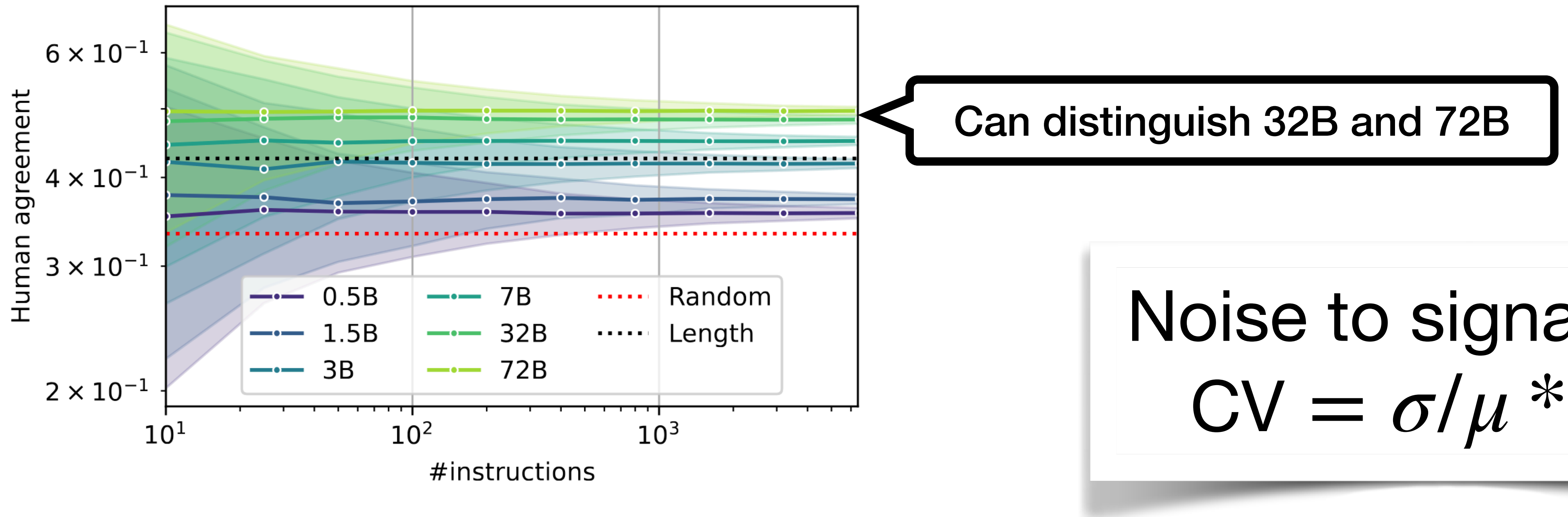
Table 1. Comparison of the variability of Spearman-correlation and Human-agreement metrics when using 6500 random annotations and the same default prompt for all model sizes.

How to compare judges?

Identifying a better metric

- Option 1: Spearman-correlation
    - 1. Evaluate a list of LLMs on a list of prompts
    - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
    - 3. Compute winrate
    - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
    - Collect list of annotated preferences by human
    - Call LLM judge to annotate preference
    - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?





Noise to signal ratio:
$CV = \sigma/\mu * 100$

#params	Sp. corr. (↑)	CV(↓)	Hum. agr. (↑)	CV(↓)
0.5B	0.09 ± 0.189	207.60	0.36 ± 0.006	1.70
1.5B	0.33 ± 0.137	41.26	0.37 ± 0.006	1.61
3B	0.82 ± 0.066	8.11	0.42 ± 0.006	1.45
7B	0.83 ± 0.082	9.84	0.45 ± 0.006	1.33
32B	0.90 ± 0.052	5.75	0.48 ± 0.006	1.29
72B	0.88 ± 0.074	8.43	0.50 ± 0.006	1.27

Table 1. Comparison of the variability of Spearman-correlation and Human-agreement metrics when using 6500 random annotations and the same default prompt for all model sizes.

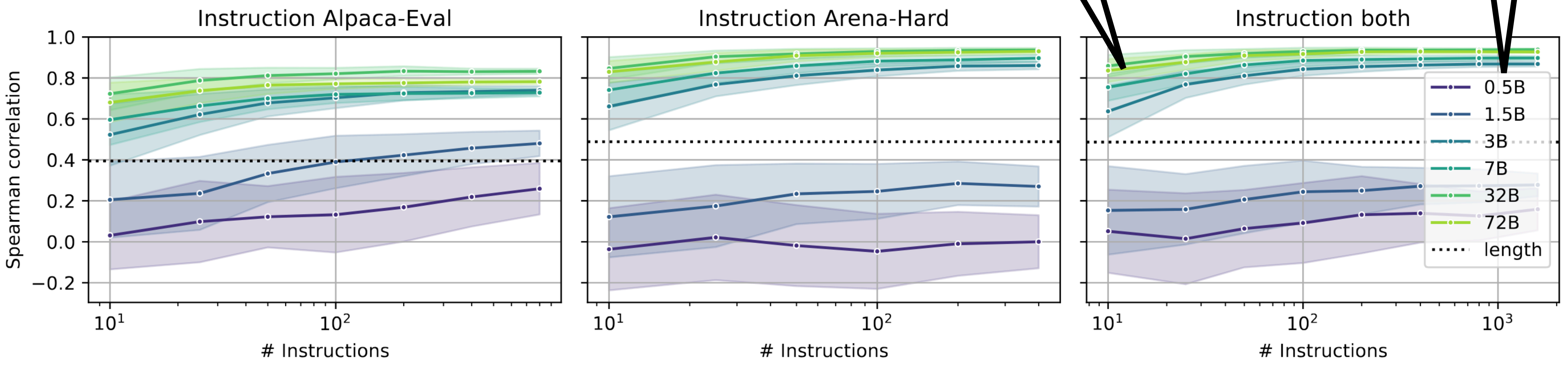
How to compare judges?

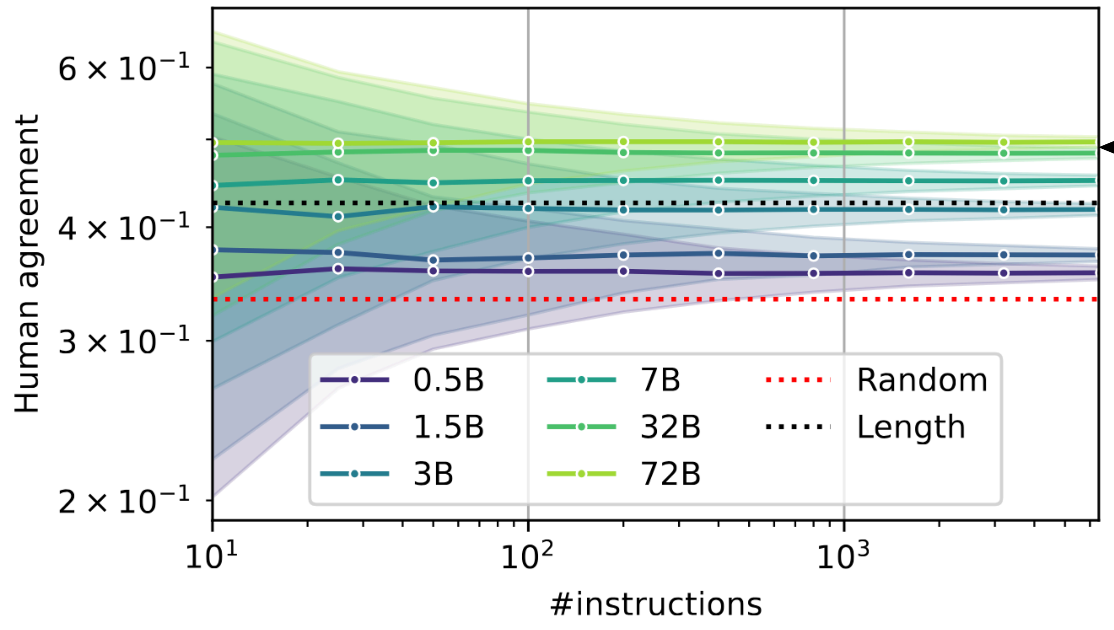
Identifying a better metric

- Option 1: Spearman-correlation
  - 1. Evaluate a list of LLMs on a list of prompts
  - 2. Call LLM judge to compare with baseline (GPT-4) on all annotations
  - 3. Compute winrate
  - 4. Measure Spearman correlation with Chatbot Arena ranking
- Option 2: Human-agreement
  - Collect list of annotated preferences by human
  - Call LLM judge to annotate preference
  - Compare LLM judge preference with human
- How both options signal to noise ratio scale with the number of annotations?

Cant distinguish 32B and 72B

params of LLM judge





Can distinguish 32B and 72B

Noise to signal ratio: CV = $\sigma/\mu * 100$

#params	Sp. corr. (↑)	CV(↓)	Hum. agr. (↑)	CV(↓)
0.5B	0.09 ± 0.189	207.60	0.36 ± 0.006	1.70
1.5B	0.33 ± 0.137	41.26	0.37 ± 0.006	1.61
3B	0.82 ± 0.066	8.11	0.42 ± 0.006	1.45
7B	0.83 ± 0.082	9.84	0.45 ± 0.006	1.33
32B	0.90 ± 0.052	5.75	0.48 ± 0.006	1.29
72B	0.88 ± 0.074	8.43	0.50 ± 0.006	1.27

Table 1. Comparison of the variability of Spearman-correlation and Human-agreement metrics when using 6500 random annotations and the same default prompt for all model sizes.

Human agreement has **much** better signal to noise ratio!

Tuning LLM judges

Multifidelity & multiobjective optimization

Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation

Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!

Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
  - Evaluate all 4480 configurations on 400 instructions

Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
  - Evaluate all 4480 configurations on 400 instructions
  - Pick top 1200 and evaluate on 1200 instructions

Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
  - Evaluate all 4480 configurations on 400 instructions
  - Pick top 1200 and evaluate on 1200 instructions
  - Pick again top 400 and evaluate on 3548 Instructions

Tuning LLM judges

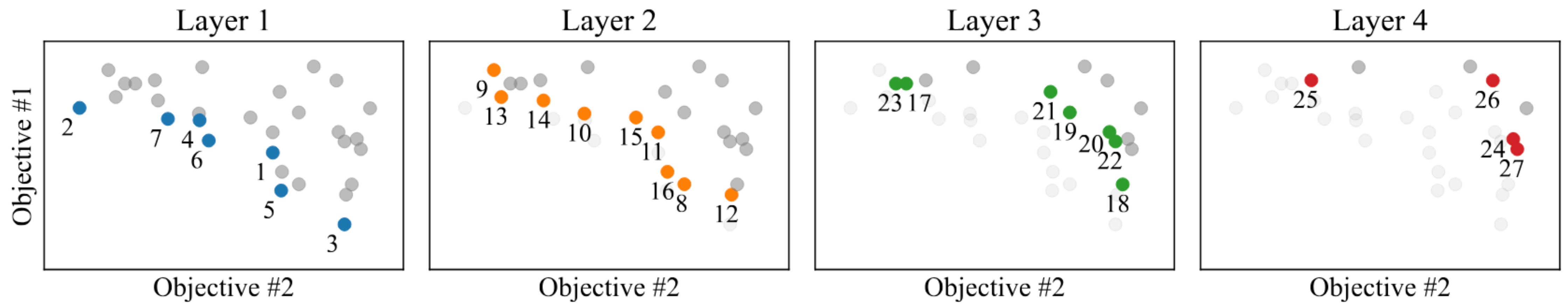
Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
    - Evaluate all 4480 configurations on 400 instructions
    - Pick top 1200 and evaluate on 1200 instructions
    - Pick again top 400 and evaluate on 3548 Instructions
- Use Non dominated sort to determine top configurations

Tuning LLM judges

Multifidelity & multiobjective optimization

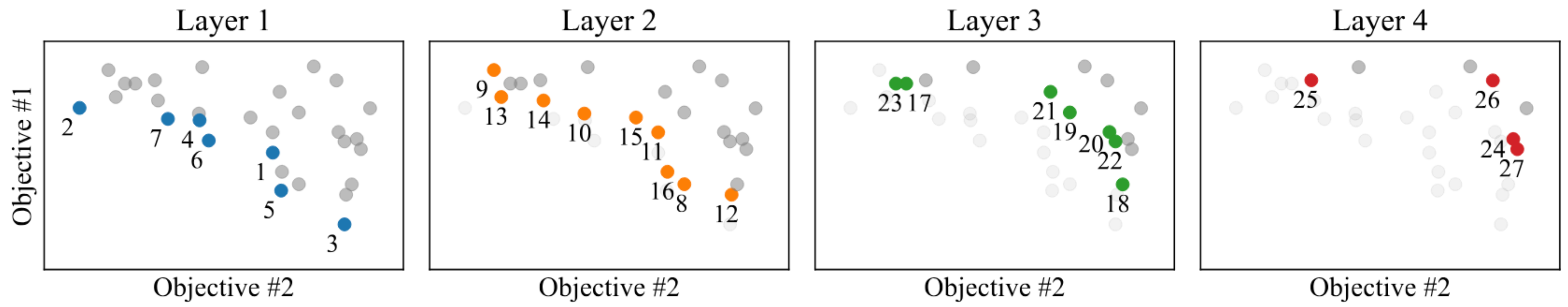
- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
    - Evaluate all 4480 configurations on 400 instructions
    - Pick top 1200 and evaluate on 1200 instructions
    - Pick again top 400 and evaluate on 3548 Instructions
- Use Non dominated sort to determine top configurations



Tuning LLM judges

Multifidelity & multiobjective optimization

- Two objectives: human-agreement & cost per annotation
- Multi-fidelity multi-objective to the rescue!
  - Evaluate all 4480 configurations on 400 instructions
  - Pick top 1200 and evaluate on 1200 instructions
  - Pick again top 400 and evaluate on 3548 Instructions
- Use Non dominated sort to determine top configurations



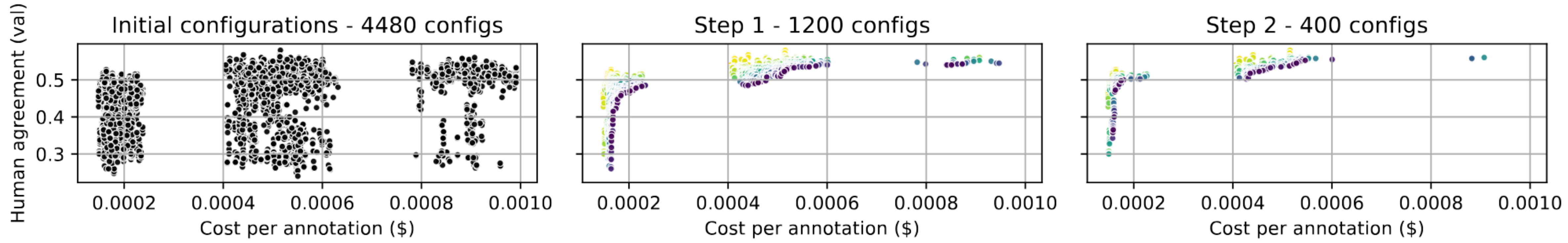


Figure 4. Illustration of the selection process. All 4 480 configurations are first evaluated on 400 instructions (left), the top 1 200 configurations are then evaluated on 1 200 instructions (center) and finally the top 400 configurations are evaluated on 3 548 instructions (right). The color denotes the ranking assigned by the non-dominated sort procedure.

Results of top judges

How are tuned judges from open-weights compared to previous approaches?

- How are tuned judges from open-weights compared to previous approaches?
- With tuning, we can outperform close-weight and fine-tune judges while having a much lower cost

Results of top judges

How are tuned judges from open-weights compared to previous approaches?

- How are tuned judges from open-weights compared to previous approaches?
- With tuning, we can outperform close-weight and fine-tune judges while having a much lower cost

Judge	Human agr. (↑)	Cost per 1K ann. (↓)
Random	0.33 +/- 0.01	-
Length	0.42 +/- 0.01	-
PandaLM-7B	0.38 +/- 0.01	6.0
JudgeLM-7B	0.42 +/- 0.01	8.6
Arena-Hard	0.50 +/- 0.01	1.2
Ours-small	0.45 +/- 0.01	0.21
Ours-medium	0.47 +/- 0.01	0.48
Ours-large	0.49 +/- 0.01	0.48

Table 2. Comparison of judges on LMSys test instructions. For each judge, we report the bootstrap mean and std for human-agreement on 3K test instructions with 100 seeds.

LMsys test set

Results of top judges

How are tuned judges from open-weights compared to previous approaches?

- How are tuned judges from open-weights compared to previous approaches?
- With tuning, we can outperform close-weight and fine-tune judges while having a much lower cost

Judge	Human agr. (↑)	Cost per 1K ann. (↓)
Random	0.33 +/- 0.01	-
Length	0.42 +/- 0.01	-
PandaLM-7B	0.38 +/- 0.01	6.0
JudgeLM-7B	0.42 +/- 0.01	8.6
Arena-Hard	0.50 +/- 0.01	1.2
Ours-small	0.45 +/- 0.01	0.21
Ours-medium	0.47 +/- 0.01	0.48
Ours-large	0.49 +/- 0.01	0.48

Table 2. Comparison of judges on LMSys test instructions. For each judge, we report the bootstrap mean and std for human-agreement on 3K test instructions with 100 seeds.

Judge	Human agr. (↑)
Random	0.33
Length	0.60
GPT-3.5	0.63
GPT-4	0.67
PandaLM-7B	0.60
PandaLM-70B	0.67
Ours-small	0.67
Ours-medium	0.78
Ours-large	0.76

Table 3. Comparison with PandaLM on PandaLM test set. Note that our method is not fine-tuned as opposed to PandaLM.

LMsys test set

PandaLM test set

Results of top judges

How are tuned judges from open-weights compared to previous approaches?

- How are tuned judges from open-weights compared to previous approaches?
- With tuning, we can outperform close-weight and fine-tune judges while having a much lower cost

Judge	Human agr. (↑)	Cost per 1K ann. (↓)
Random	0.33 +/- 0.01	-
Length	0.42 +/- 0.01	-
PandaLM-7B	0.38 +/- 0.01	6.0
JudgeLM-7B	0.42 +/- 0.01	8.6
Arena-Hard	0.50 +/- 0.01	1.2
Ours-small	0.45 +/- 0.01	0.21
Ours-medium	0.47 +/- 0.01	0.48
Ours-large	0.49 +/- 0.01	0.48

Table 2. Comparison of judges on LMSys test instructions. For each judge, we report the bootstrap mean and std for human-agreement on 3K test instructions with 100 seeds.

LMsys test set

Judge	Human agr. (↑)
Random	0.33
Length	0.60
GPT-3.5	0.63
GPT-4	0.67
PandaLM-7B	0.60
PandaLM-70B	0.67
Ours-small	0.67
Ours-medium	0.78
Ours-large	0.76

Table 3. Comparison with PandaLM on PandaLM test set. Note that our method is not fine-tuned as opposed to PandaLM.

PandaLM test set

Judge	Sp. corr. (↑)	Cost per 1K ann. (↓)
Length	0.50 +/- 0.21	-
Arena-hard + Claude	0.82 +/- 0.12	75.0
Arena-hard + GPT4	0.90 +/- 0.06	50.0
Ours-small	0.81 +/- 0.10	0.21
Ours-medium	0.93 +/- 0.05	0.48
Ours-large	0.86 +/- 0.09	0.48

Table 4. For each judge, we compute the Spearman correlation between win-rates using the protocol of Arena-Hard and ELO-ratings computed from human annotations from Chatbot Arena. We report the mean and std over 100 boostraps of the set of models.

ArenaHard

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

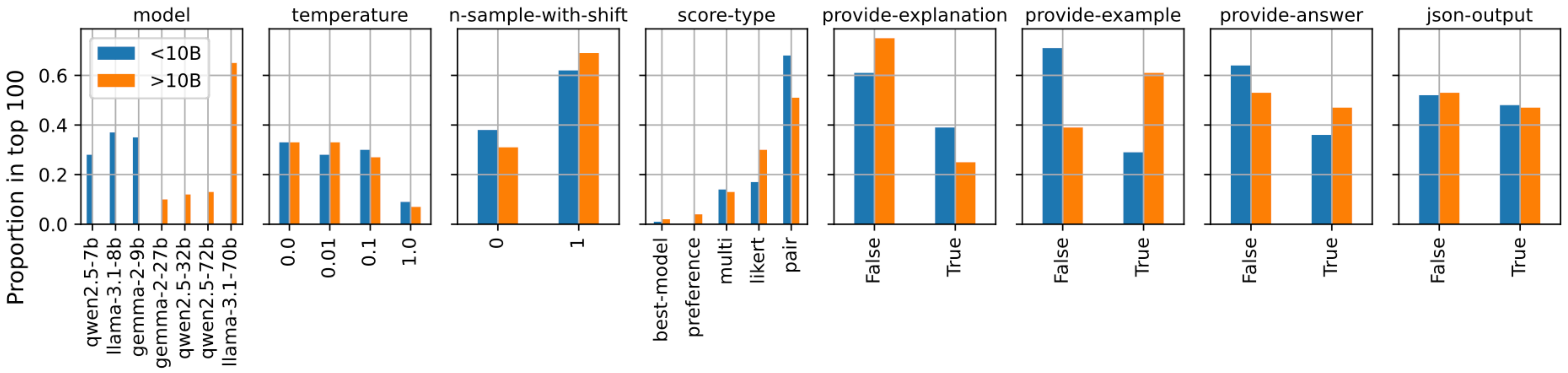


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

How many hyperparameters in the top 100 from the 4480 judges?

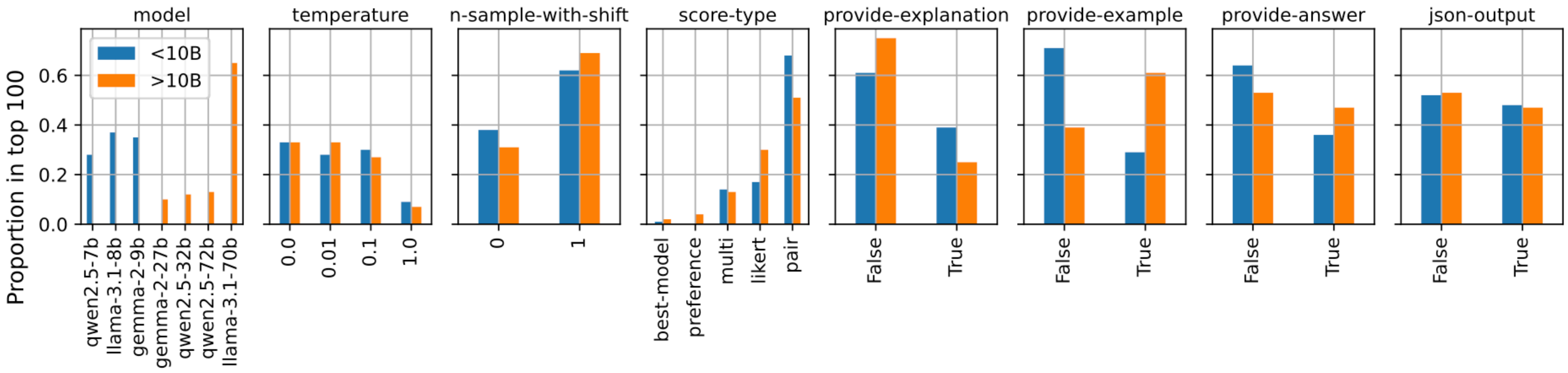


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

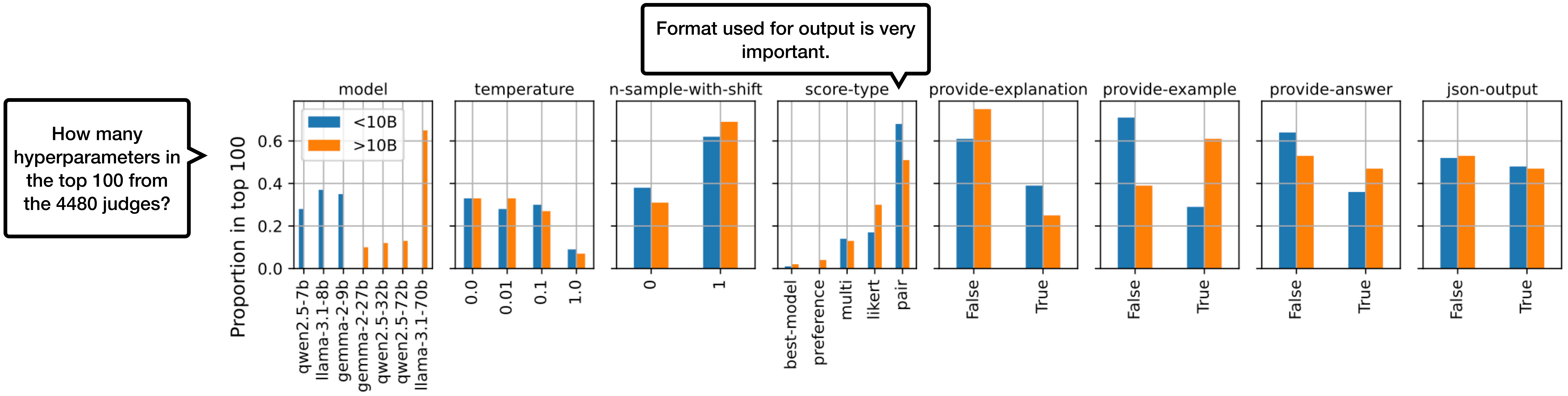


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

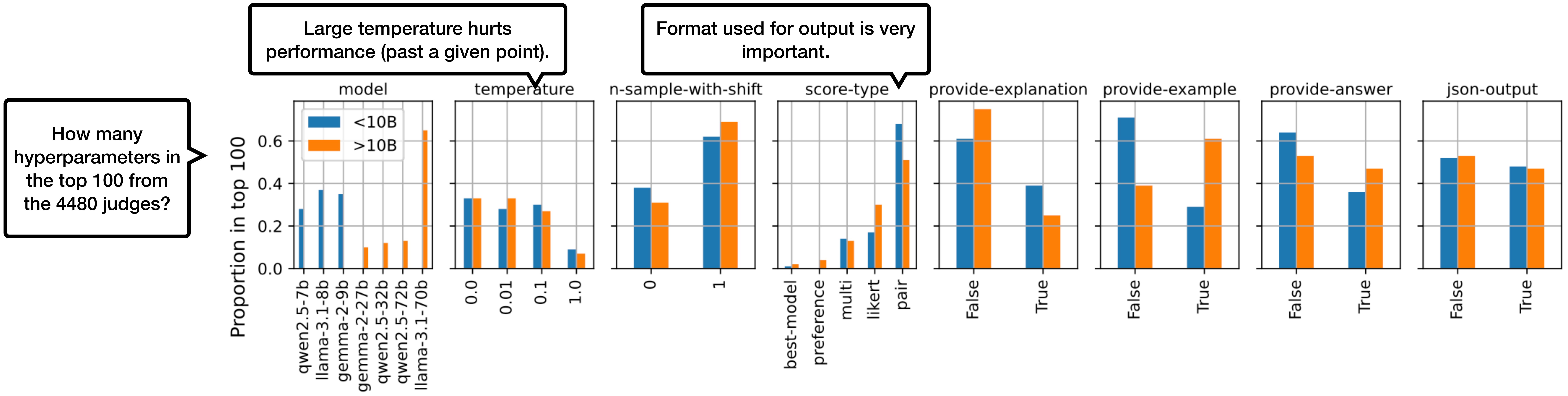


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

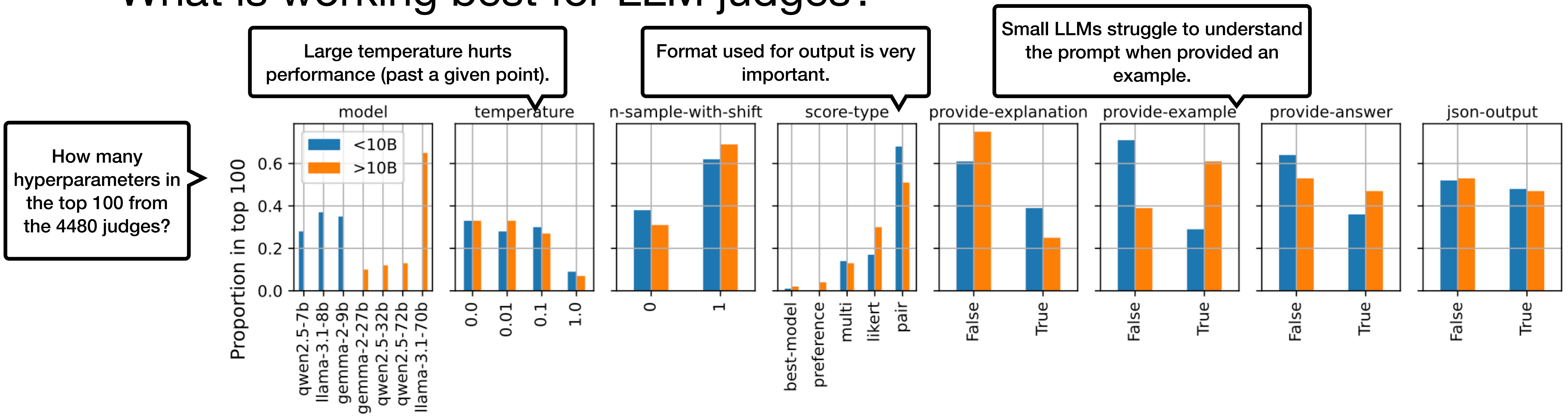


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Analysing judge performance

What hyperparameter/prompt works best?

- What is working best for LLM judges?

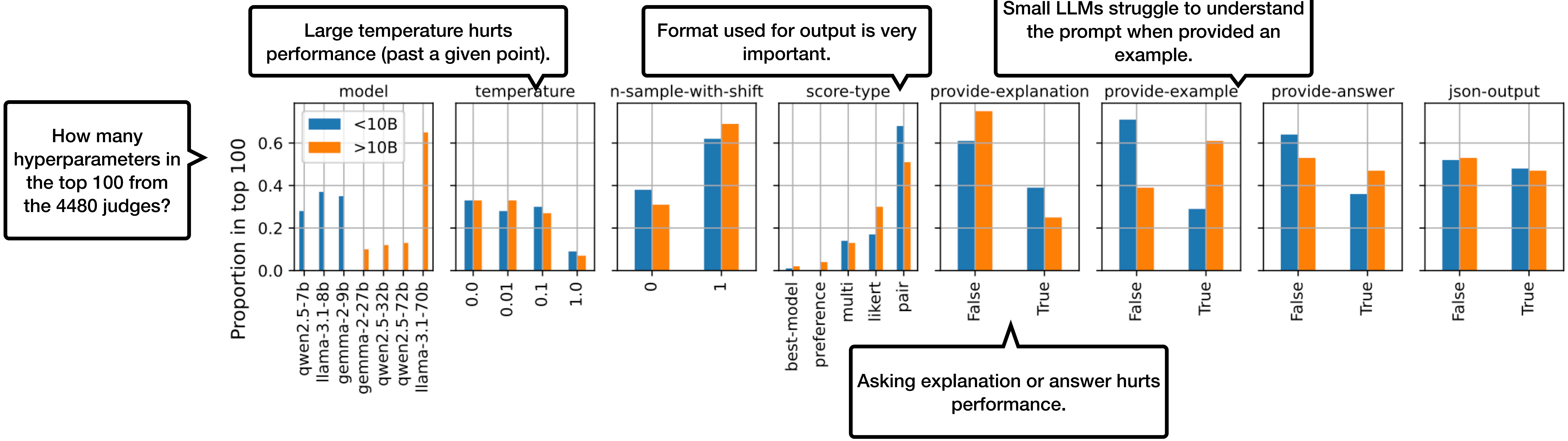


Figure 6. Fraction of time each hyperparameter appears in the top 100 configurations for small (<10B) and large models (>10B).

Prompt generalization

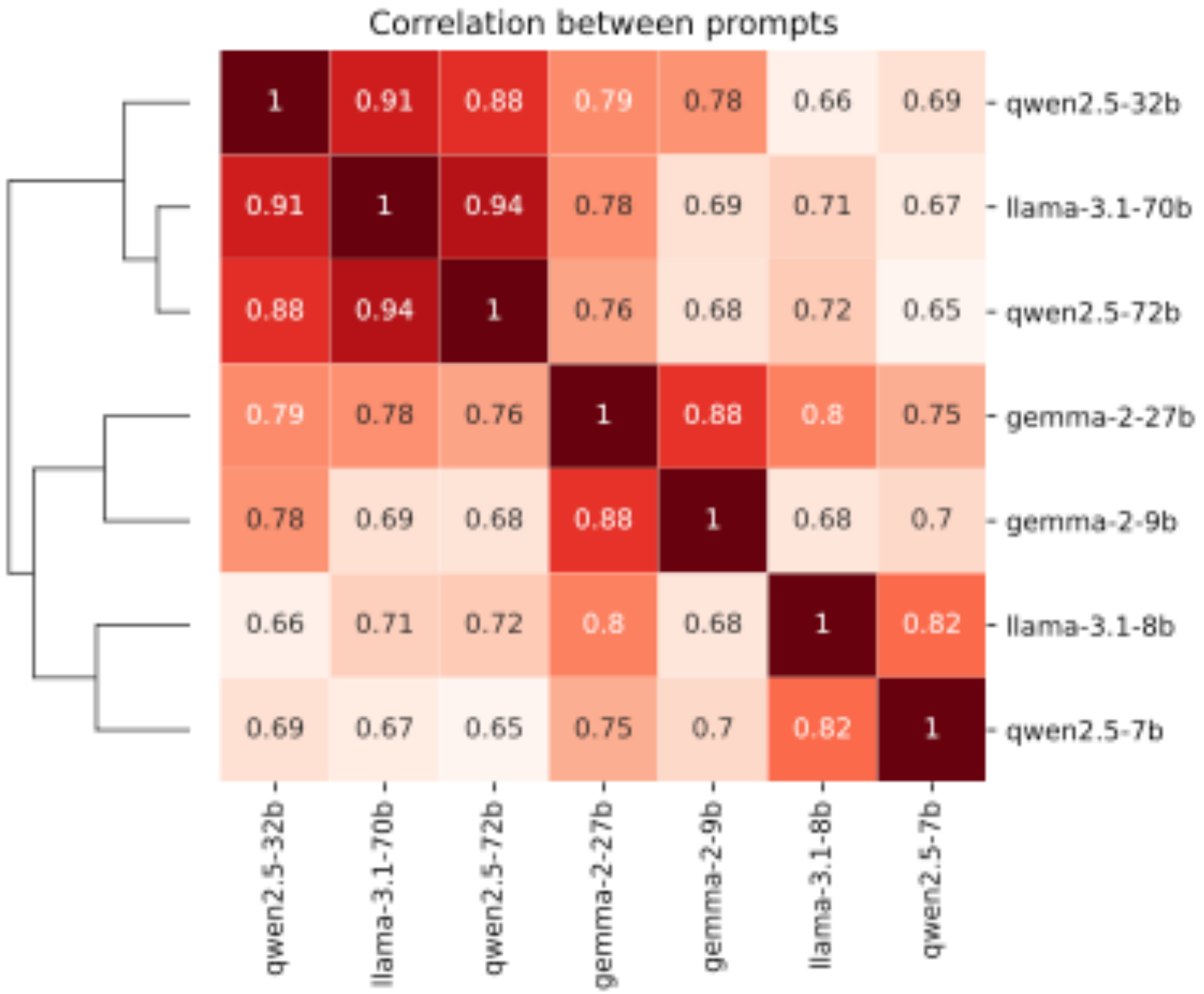
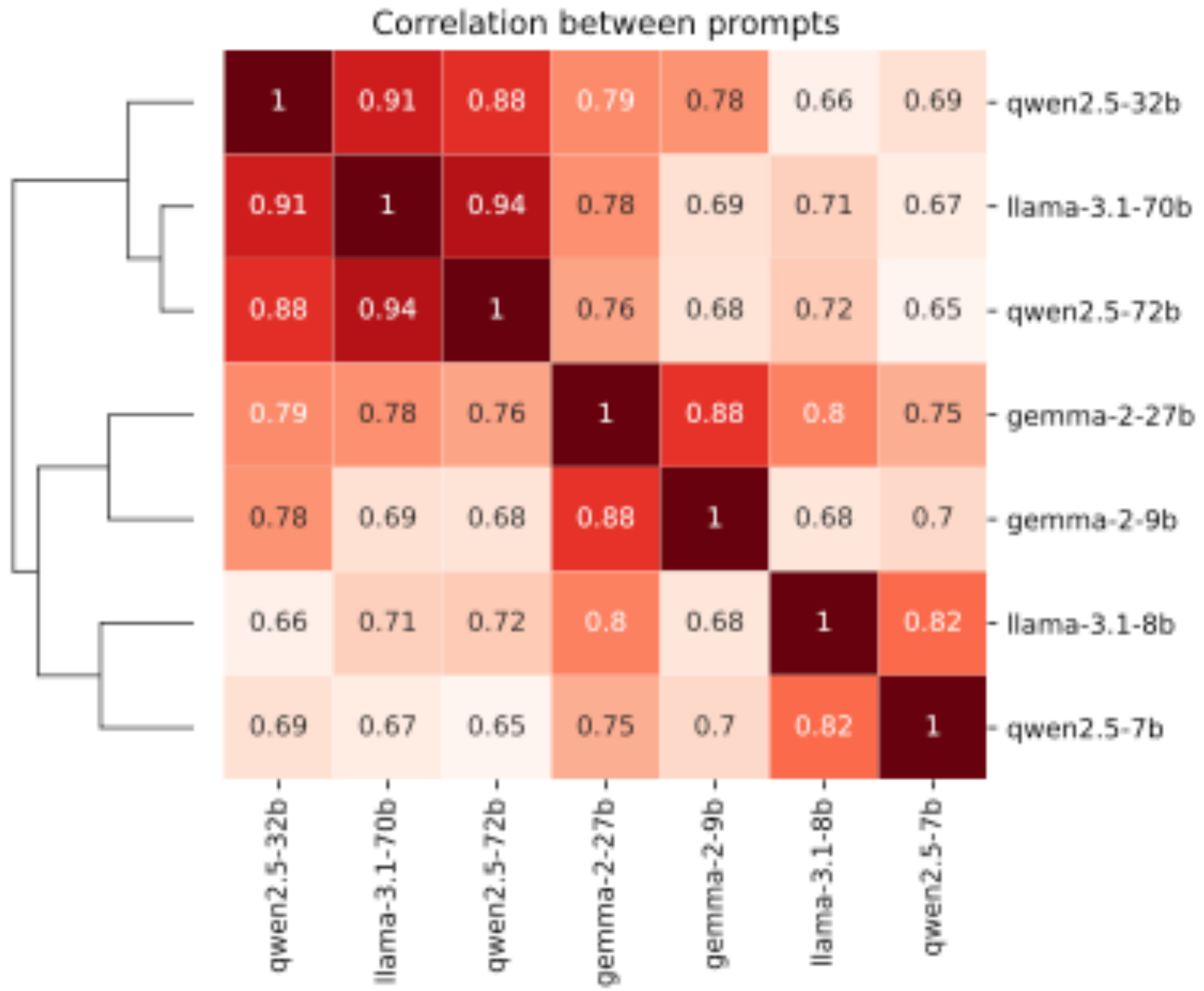


Figure 7. Prompt performance stability across different models. We show the correlation matrix between models when looking at their performance on all of the 80 different prompts.

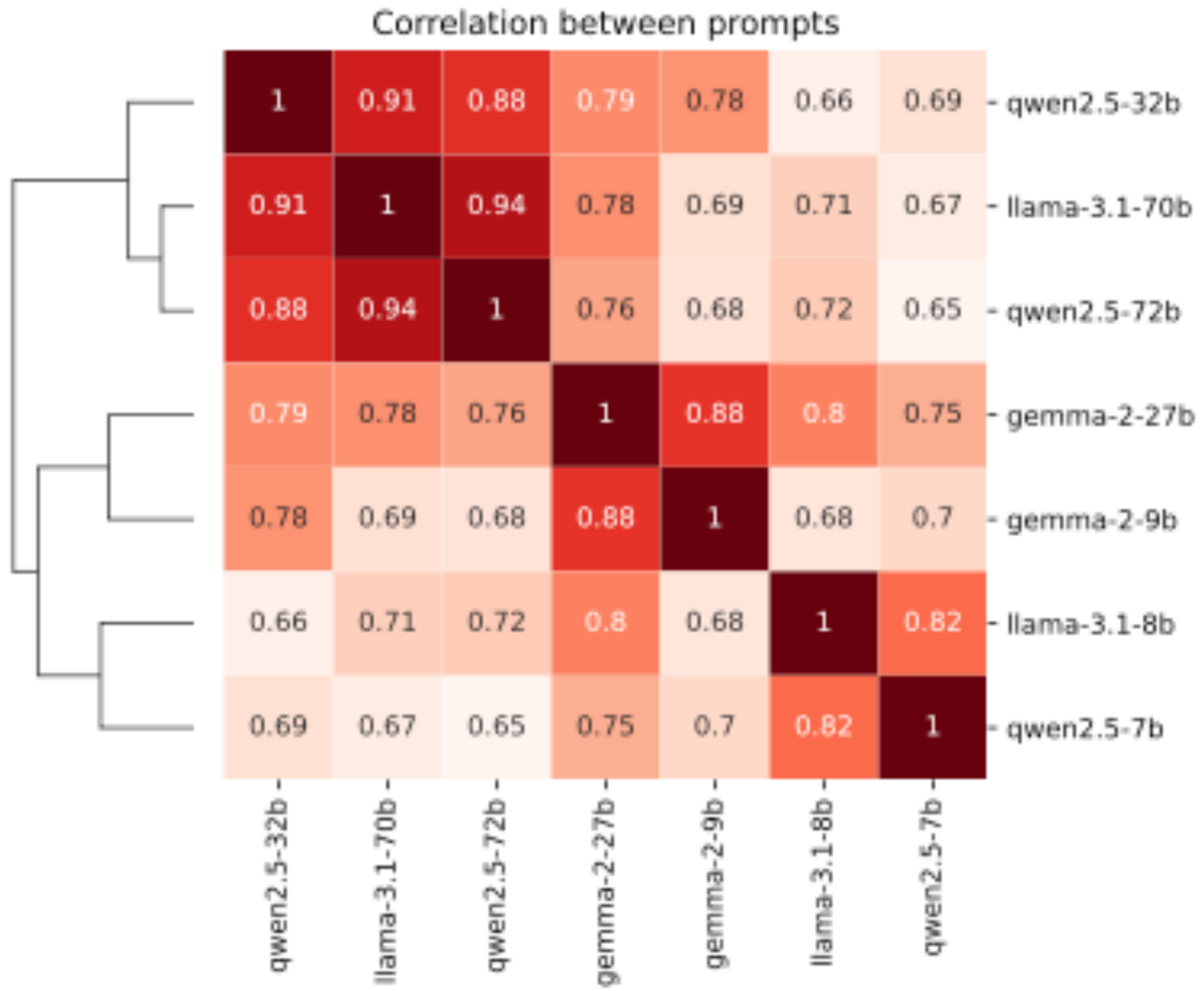
Prompt generalization



Correlation between rank of prompt configurations across models

Figure 7. Prompt performance stability across different models. We show the correlation matrix between models when looking at their performance on all of the 80 different prompts.

Prompt generalization

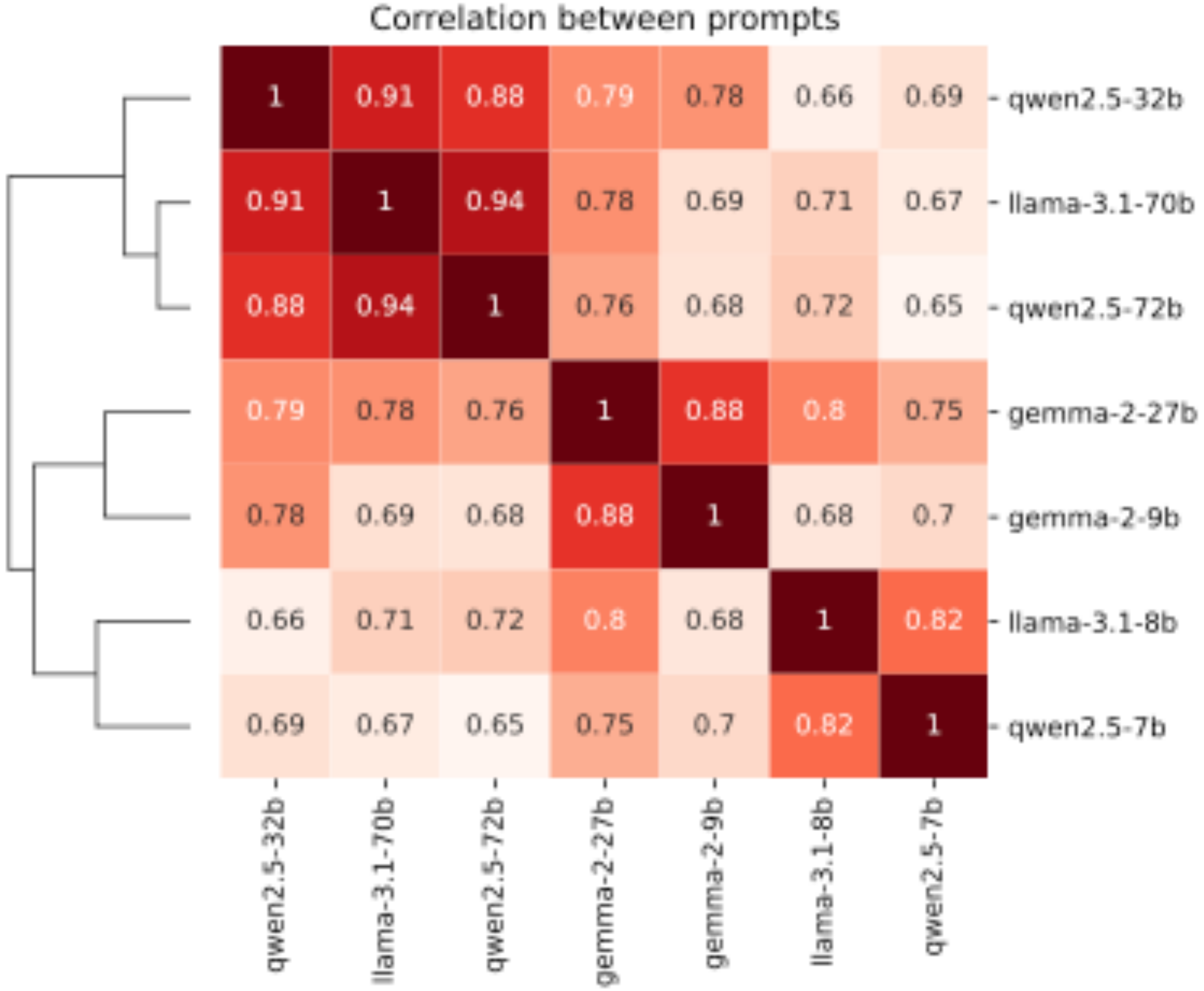


Correlation between rank of prompt configurations across models

High correlation => best prompt relatively stable across models

Figure 7. Prompt performance stability across different models. We show the correlation matrix between models when looking at their performance on all of the 80 different prompts.

Prompt generalization



	qwen2.5-32b	llama-3.1-70b	qwen2.5-72b	gemma-2-27b	gemma-2-9b	llama-3.1-8b	qwen2.5-7b
qwen2.5-32b	1	0.91	0.88	0.79	0.78	0.66	0.69
llama-3.1-70b	0.91	1	0.94	0.78	0.69	0.71	0.67
qwen2.5-72b	0.88	0.94	1	0.76	0.68	0.72	0.65
gemma-2-27b	0.79	0.78	0.76	1	0.88	0.8	0.75
gemma-2-9b	0.78	0.69	0.68	0.88	1	0.68	0.7
llama-3.1-8b	0.66	0.71	0.72	0.8	0.68	1	0.82
qwen2.5-7b	0.69	0.67	0.65	0.75	0.7	0.82	1

Figure 7. Prompt performance stability across different models. We show the correlation matrix between models when looking at their performance on all of the 80 different prompts.

Correlation between rank of prompt configurations across models

High correlation => best prompt relatively stable across models

Correlation higher when model sizes and family are close

````
You are a highly efficient assistant, who evaluates and selects the best large language
↪ model based on the quality of their responses to a given instruction.
You will be shown one instruction and the output of Assistant A and Assistant B and will
↪ have to decide which one was best.
Make sure to not over-confidently prefer one assistant or the other and also make sure to
↪ not bias your preference based on the ordering or on the length of the answers.

<|User Prompt|>
Who is Barack Obama?

<|The Start of Assistant A's Answer|>
Barack Obama is a former US president.
<|The End of Assistant A's Answer|>

<|The Start of Assistant B's Answer|>
I do not know who Barack Obama is.
<|The End of Assistant B's Answer|>

# Your output

## Format description
Your output should follow this format:
```
answer: <your answer to the user prompt>
score: <one of A>>B, A>B, A=B, A<B, A<<B, see instruction bellow>
```
The "score" value should indicate your preference for the assistant. You must output only
↪ one of the following choices as your final verdict with a label:

A>>B: Assistant A is significantly better
A>B: Assistant A is slightly better
A=B: Tie, relatively the same
B>A: Assistant B is significantly better
B>>A: Assistant B is significantly better

## Your output, do not repeat the input above
```
````

Figure 10. Example of a prompt for the user prompt "Who is Barack Obama?". In this case, the judge is asked to provide its answer. It is asked to use the Likert format and provide its answer in raw text.
````

Conclusion

Conclusion

- LLM judges can be tuned at a reasonable cost

Conclusion

- LLM judges can be tuned at a reasonable cost
- Tuning hyperparameters allows to match or outperform previous approaches

Conclusion

- LLM judges can be tuned at a reasonable cost
- Tuning hyperparameters allows to match or outperform previous approaches
  - while diminishing the cost significantly

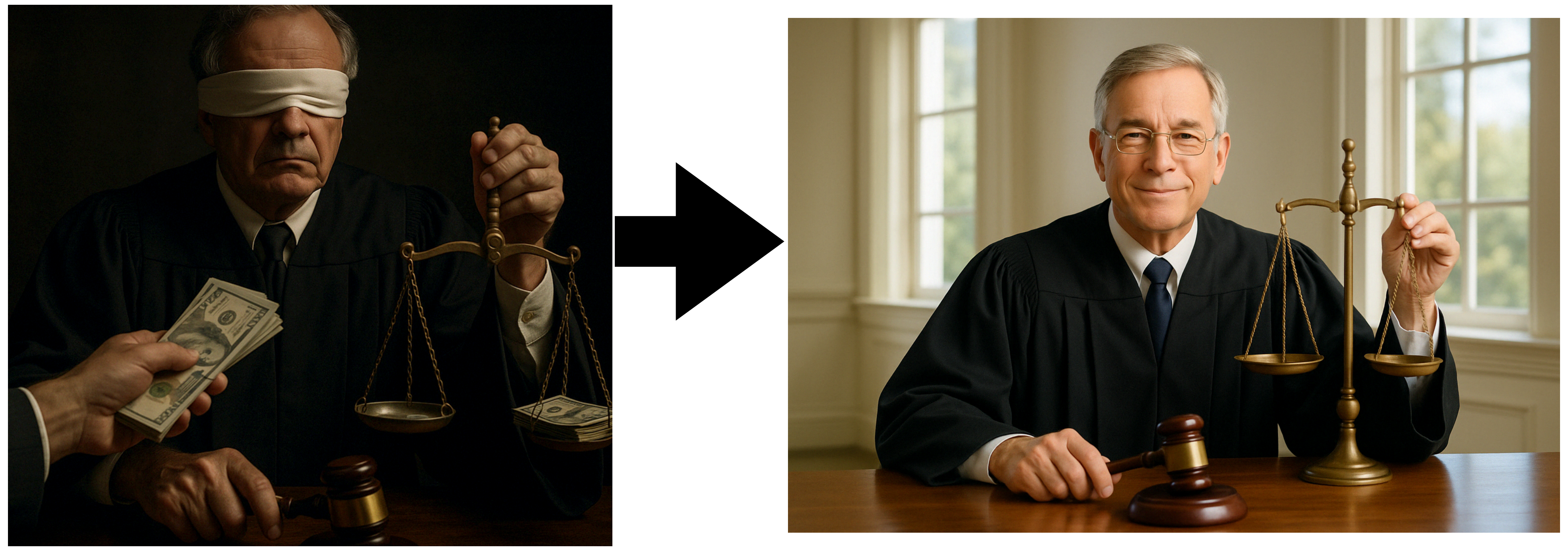
Conclusion

- LLM judges can be tuned at a reasonable cost
- Tuning hyperparameters allows to match or outperform previous approaches
    - while diminishing the cost significantly
    - … and using only open-weight models

Next steps / future work

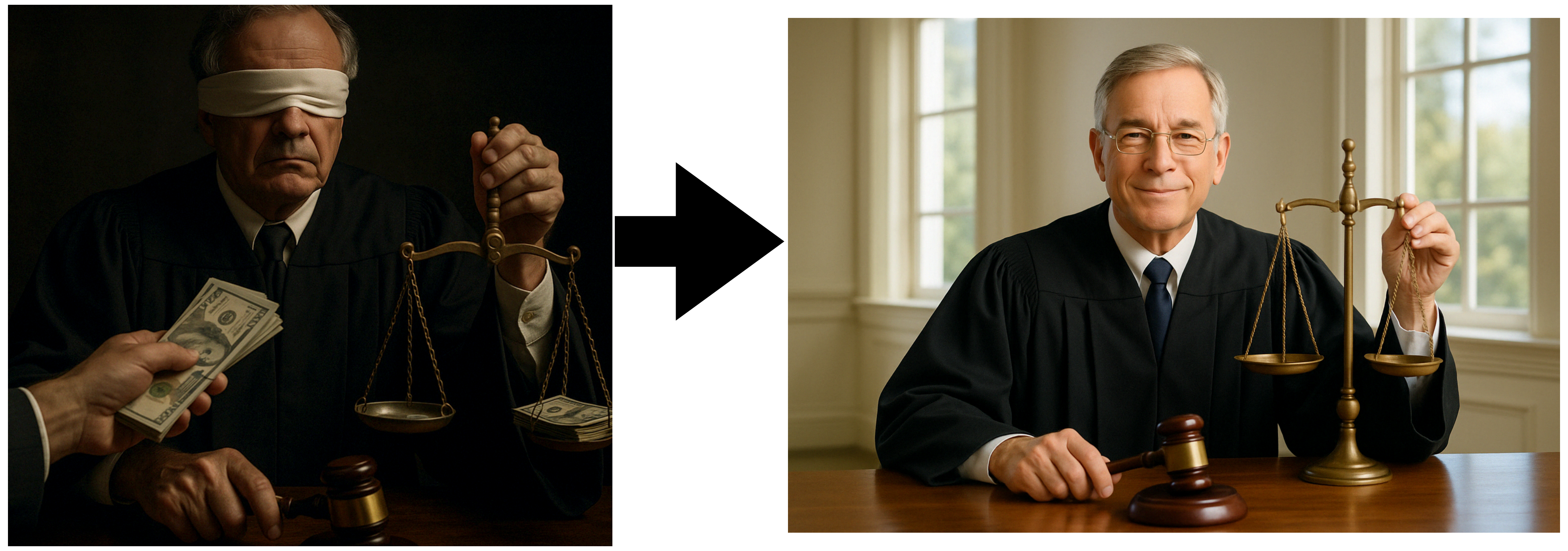
Next steps / future work

Use open-weight models to evaluate OpenEuroLLM instruction tuned models



Next steps / future work

Use open-weight models to evaluate OpenEuroLLM instruction tuned models



Optimize instruction tuning hyperparameters (DPO, learning-rate) that maximizes judge ratings

Pretrained model → Instruction-tuned model

Next steps / future work

Use open-weight models to evaluate OpenEuroLLM instruction tuned models

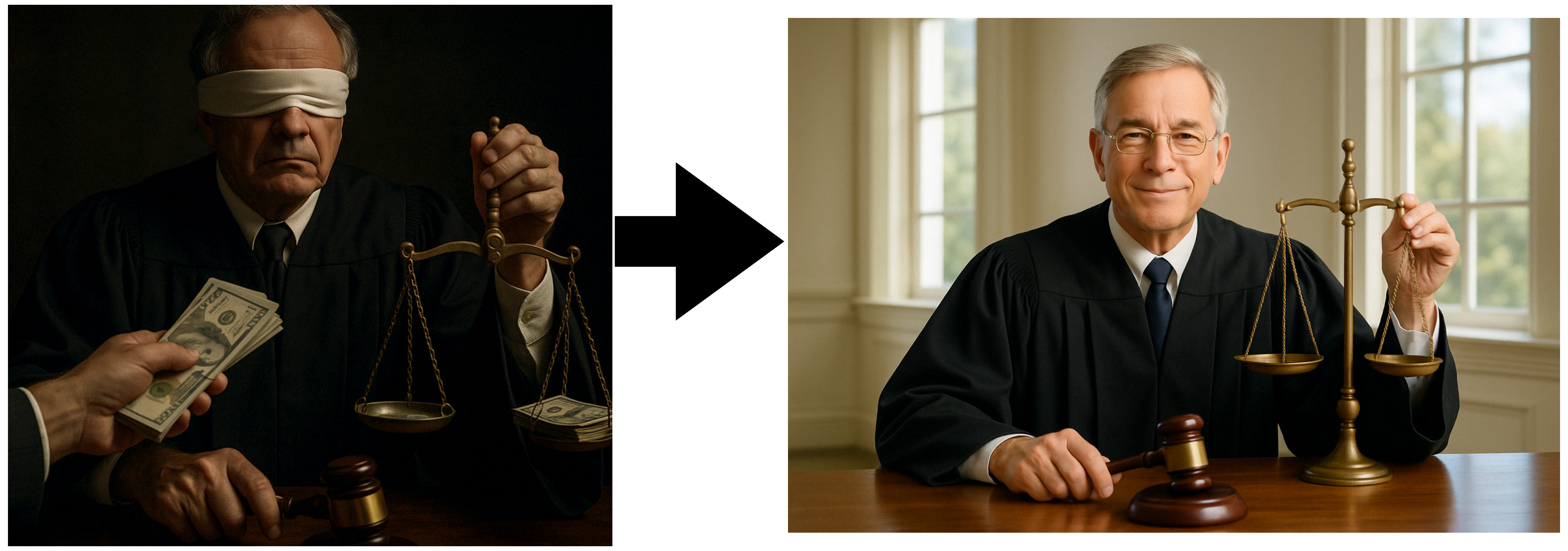
Optimize instruction tuning hyperparameters (DPO, learning-rate) that maximizes judge ratings

Pretrained model → Instruction-tuned model

Better AutoML for judges

Next steps / future work

Use open-weight models to evaluate OpenEuroLLM instruction tuned models



Optimize instruction tuning hyperparameters (DPO, learning-rate) that maximizes judge ratings

Pretrained model → Instruction-tuned model

Better AutoML for judges

- AutoML: tabular benchmark released https://github.com/geoalgo/judgetuning (together with code to reproduce results)
- Ensemble? Portfolio? Model-based optimizers?

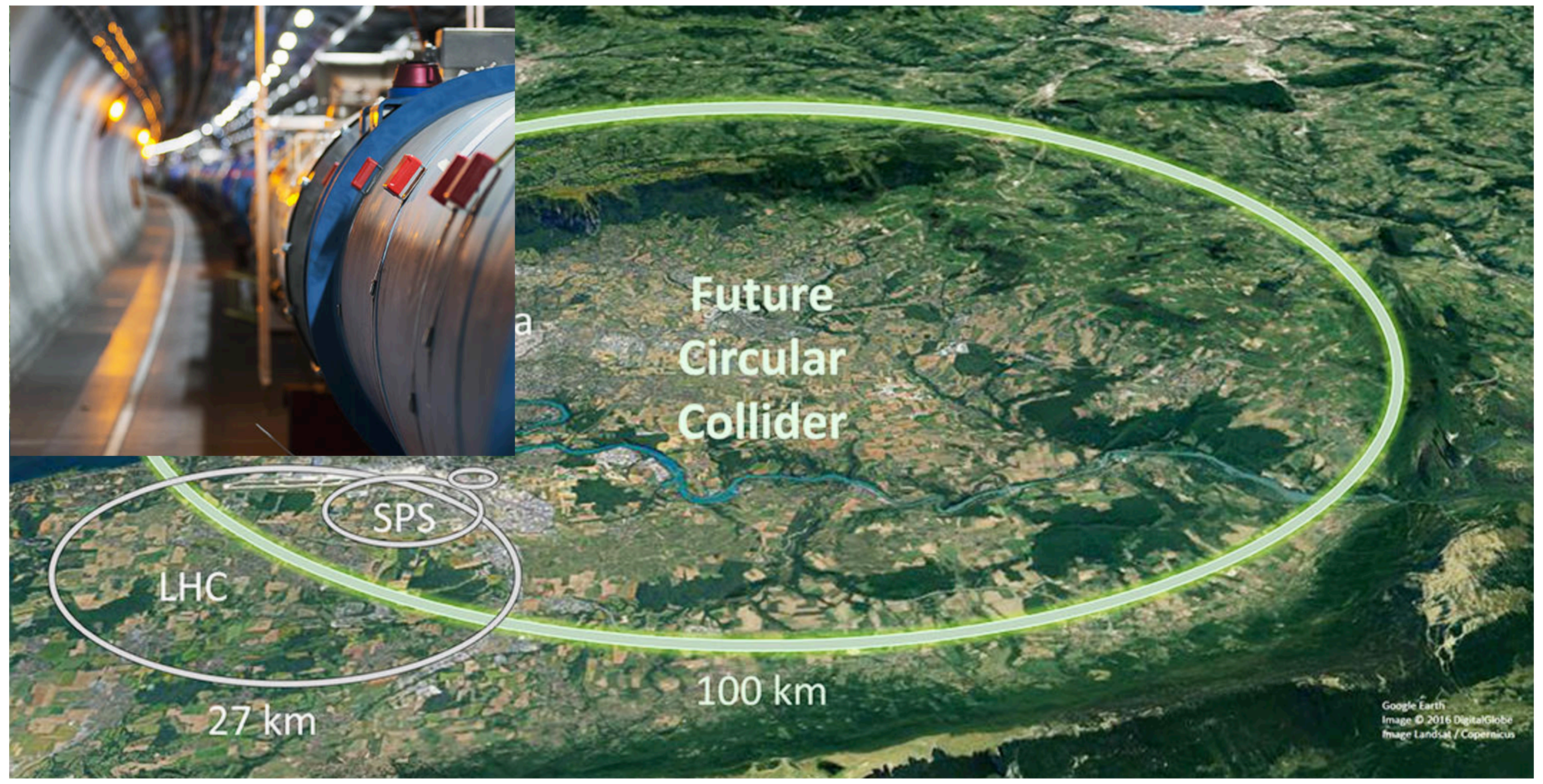
A case of openness

A case of openness

Some of humanity largest projects

A case of openness

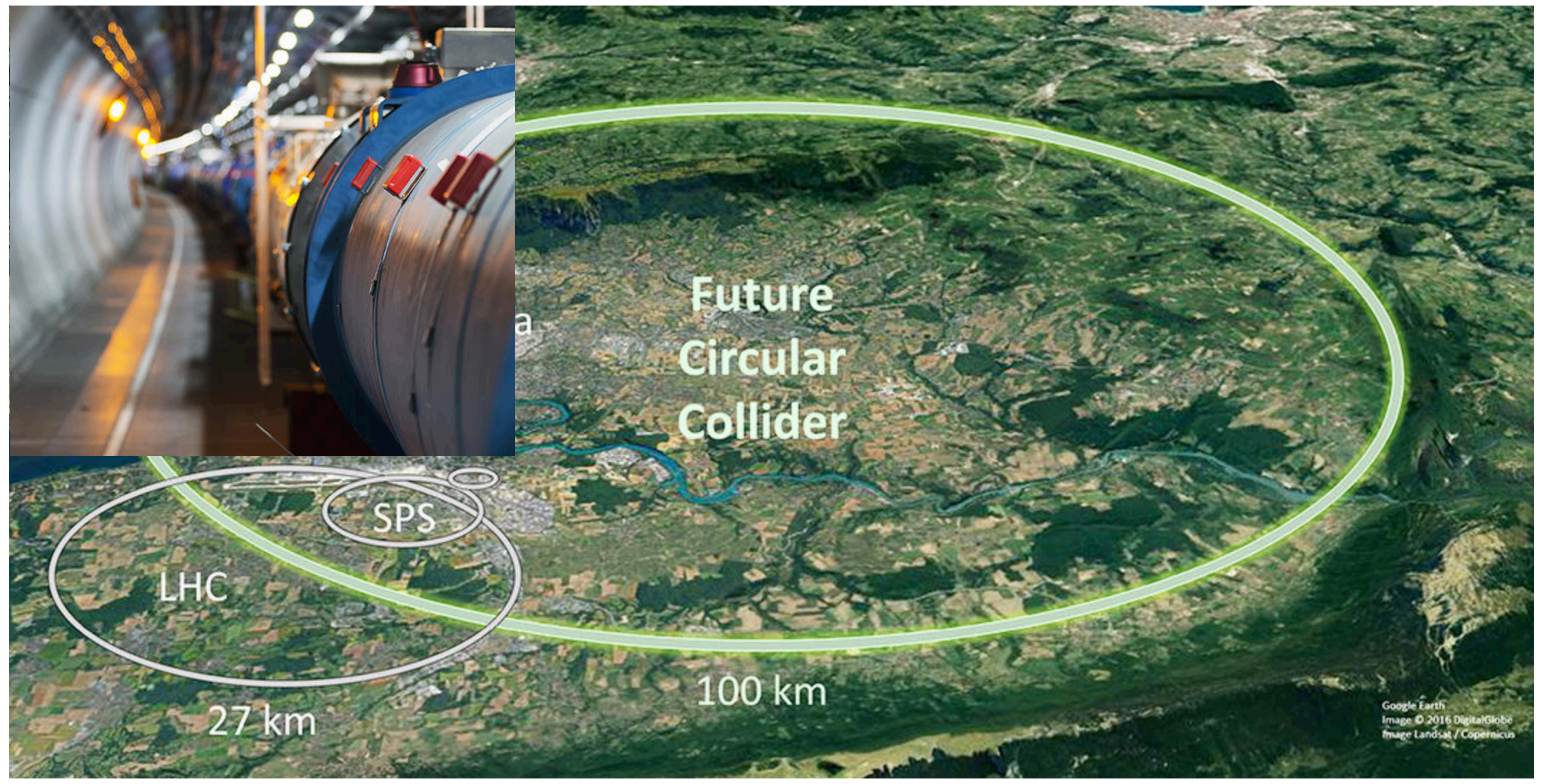
Some of humanity largest projects

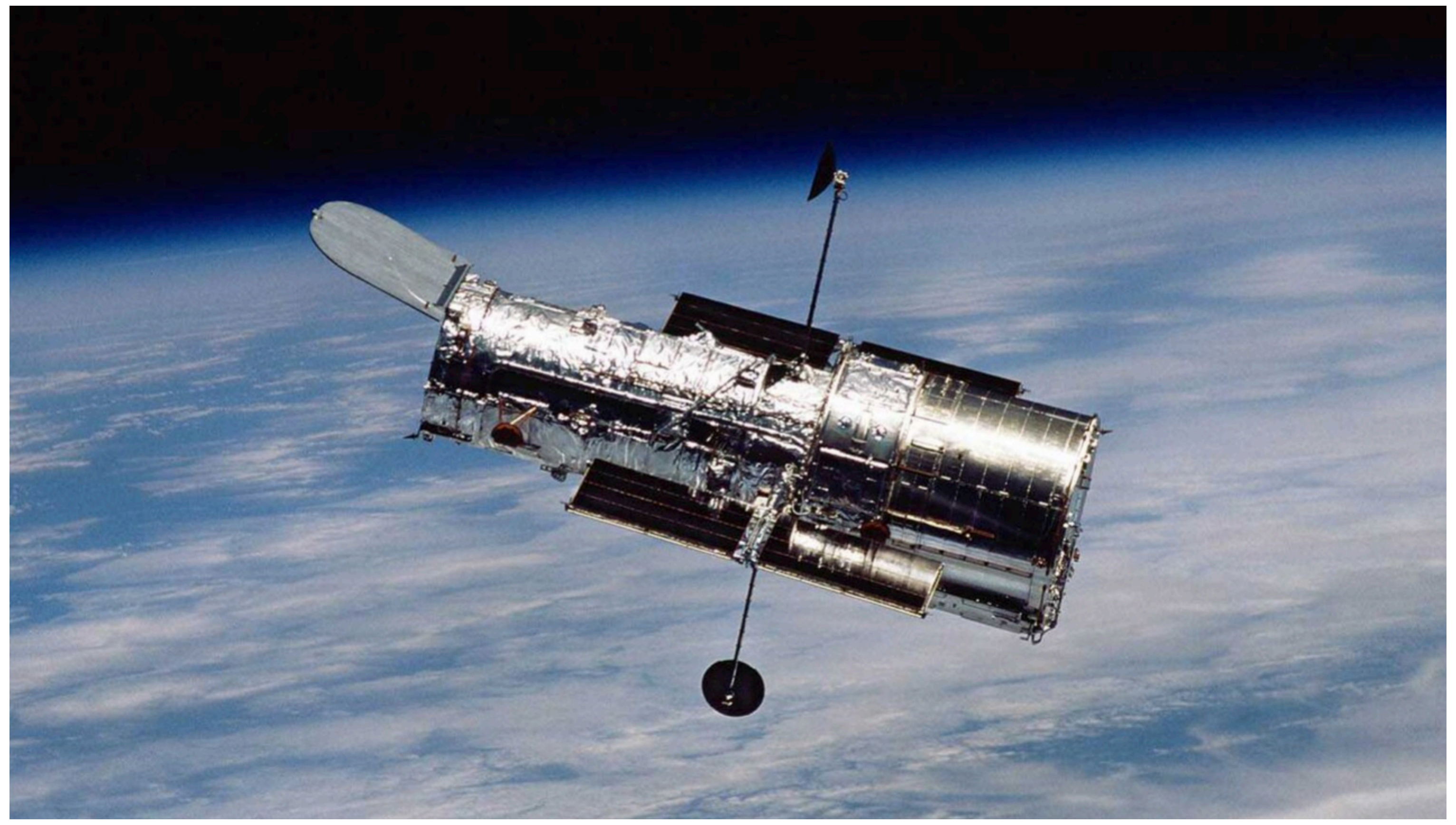


LHC: $5 Billion, 23 countries

A case of openness

Some of humanity largest projects

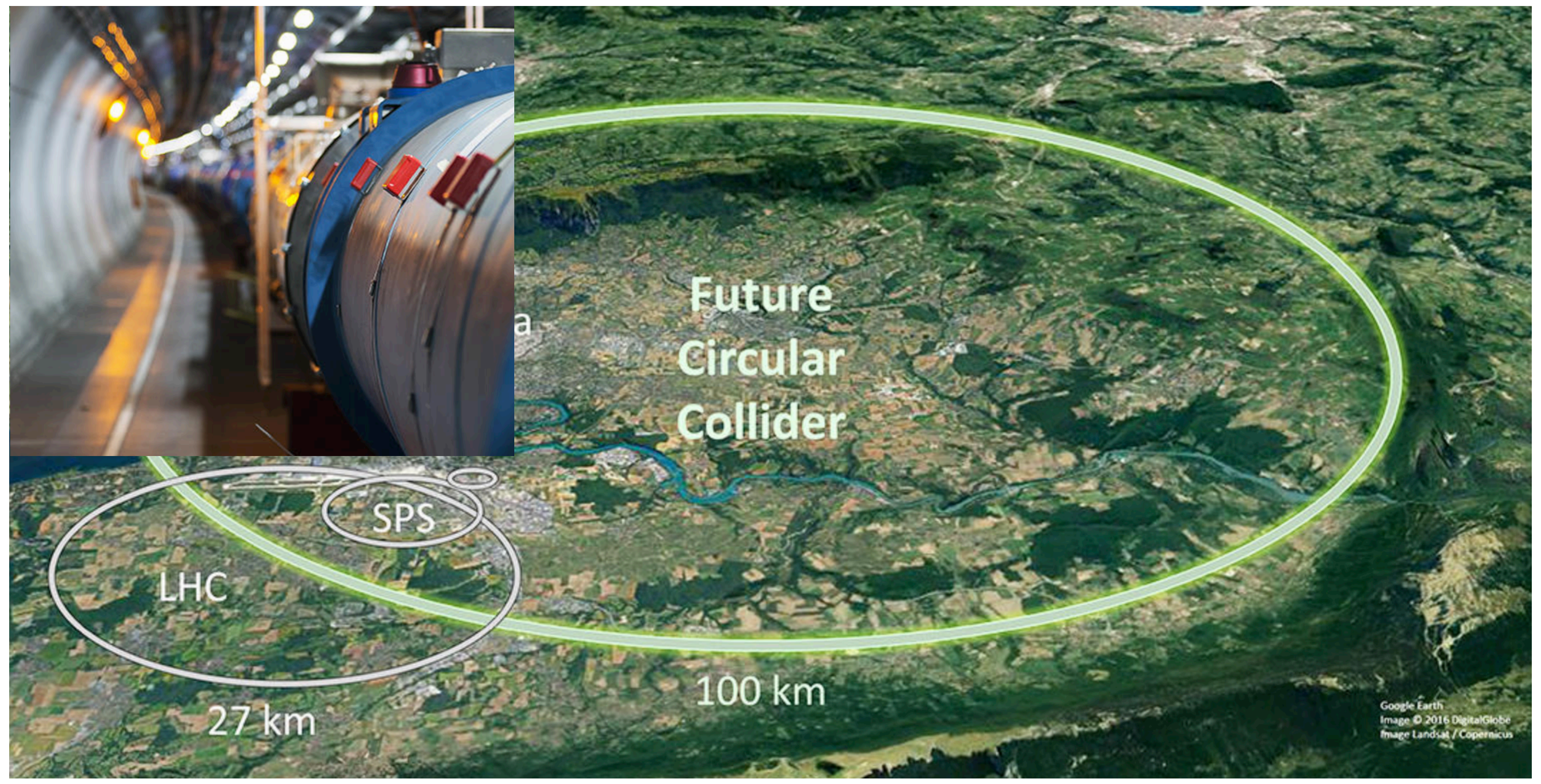




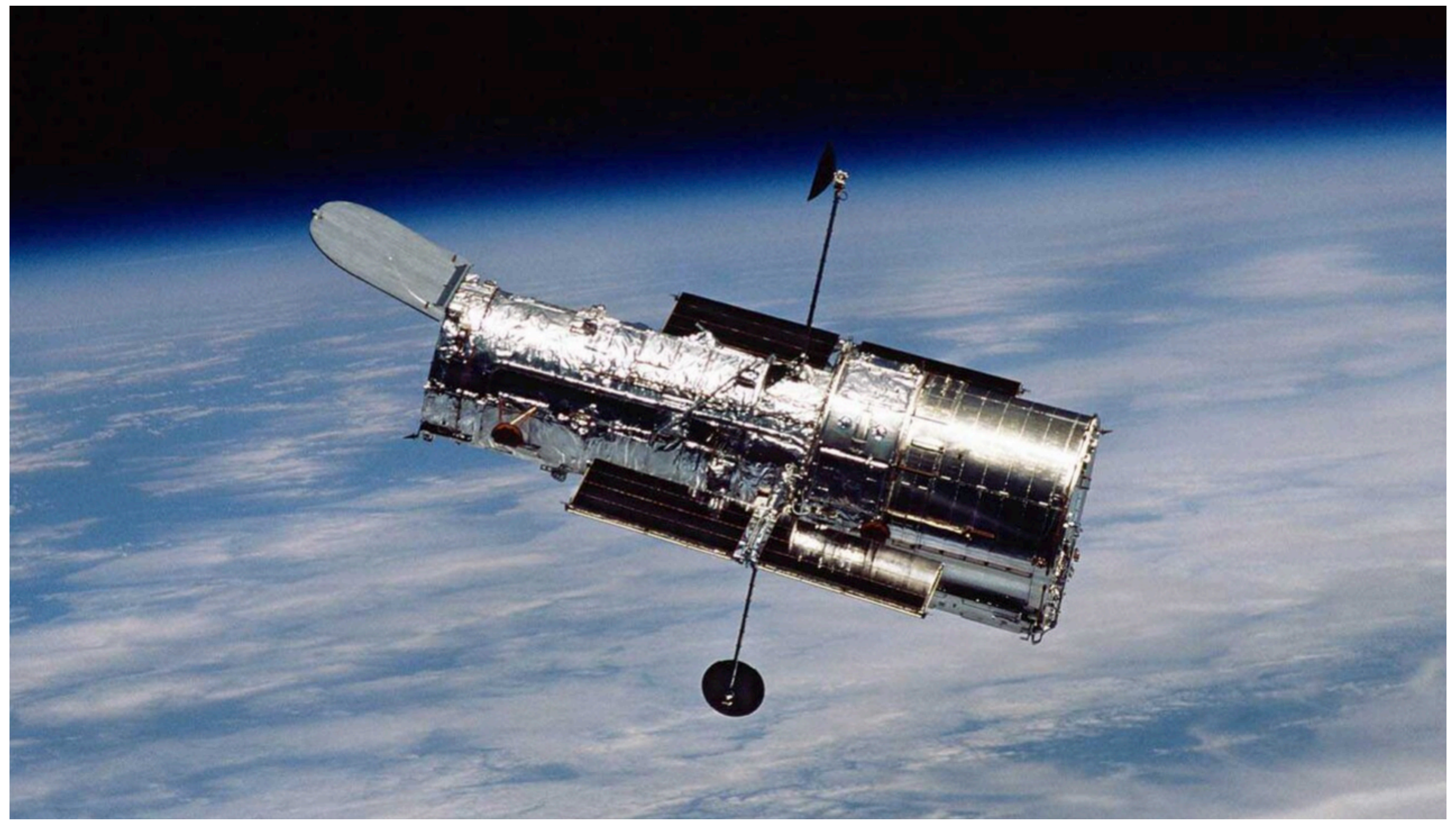
LHC: $5 Billion, 23 countries

A case of openness

Some of humanity largest projects



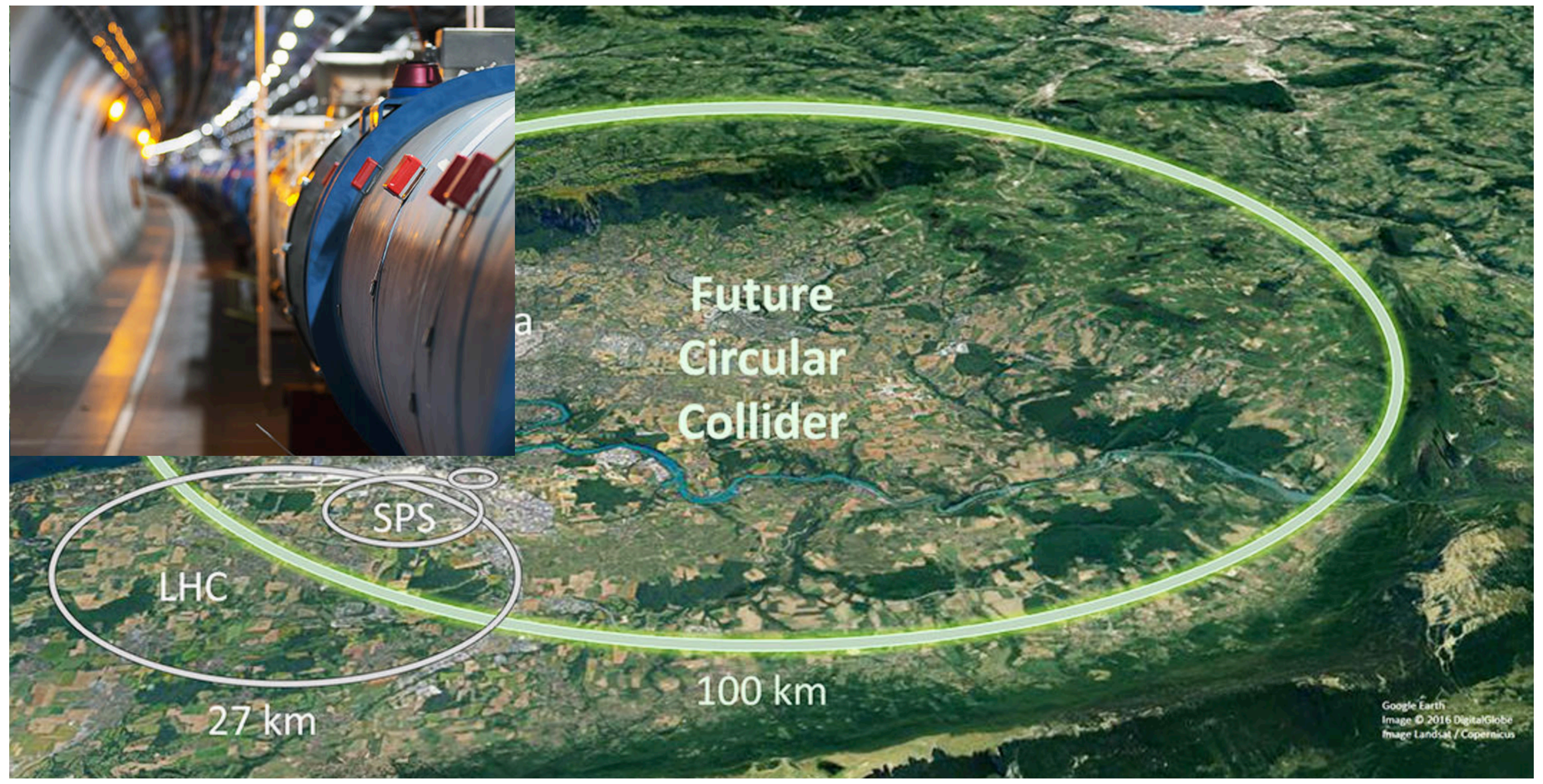
LHC: $5 Billion, 23 countries



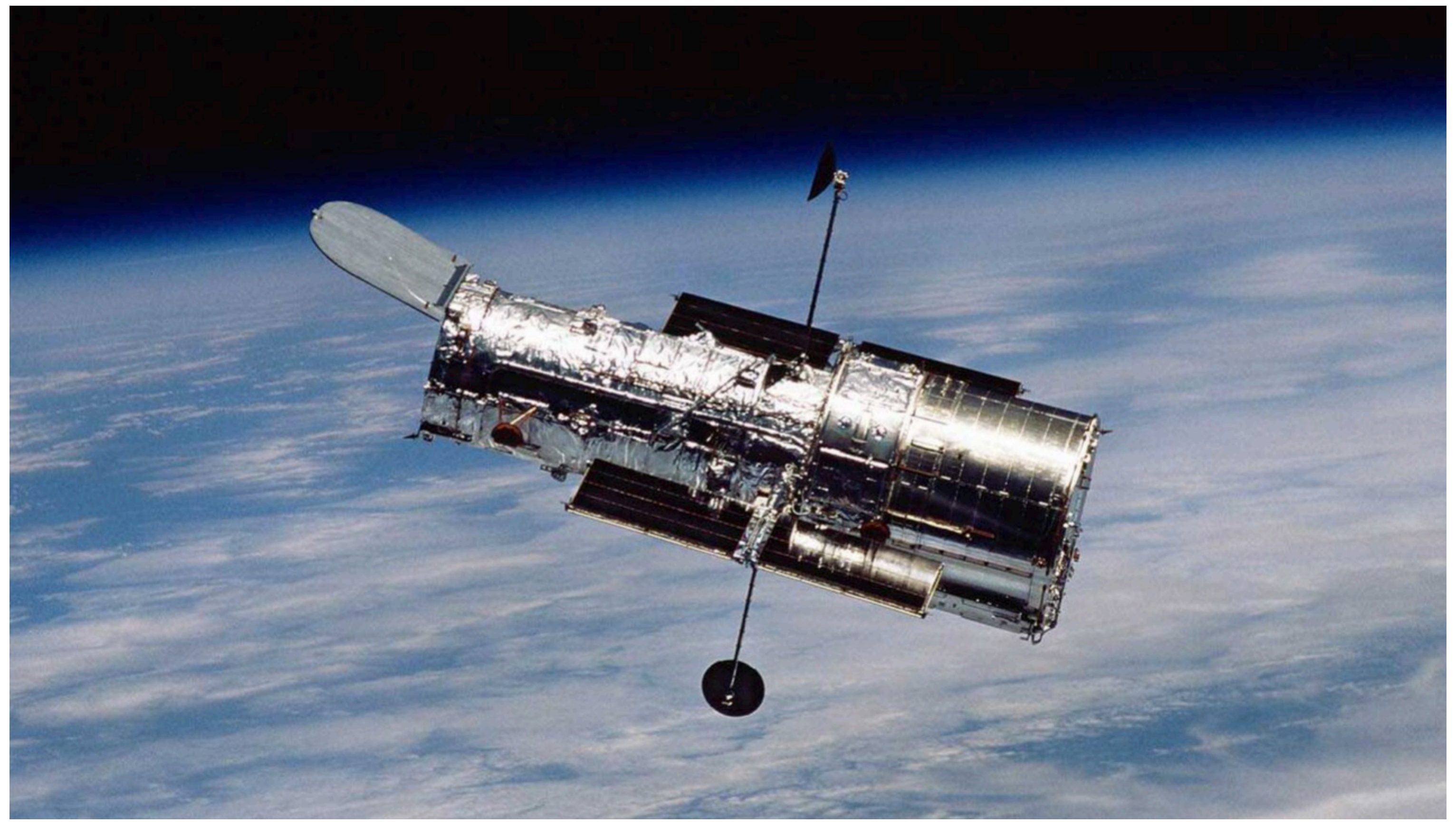
Hubble $16 billion, 11 countries

A case of openness

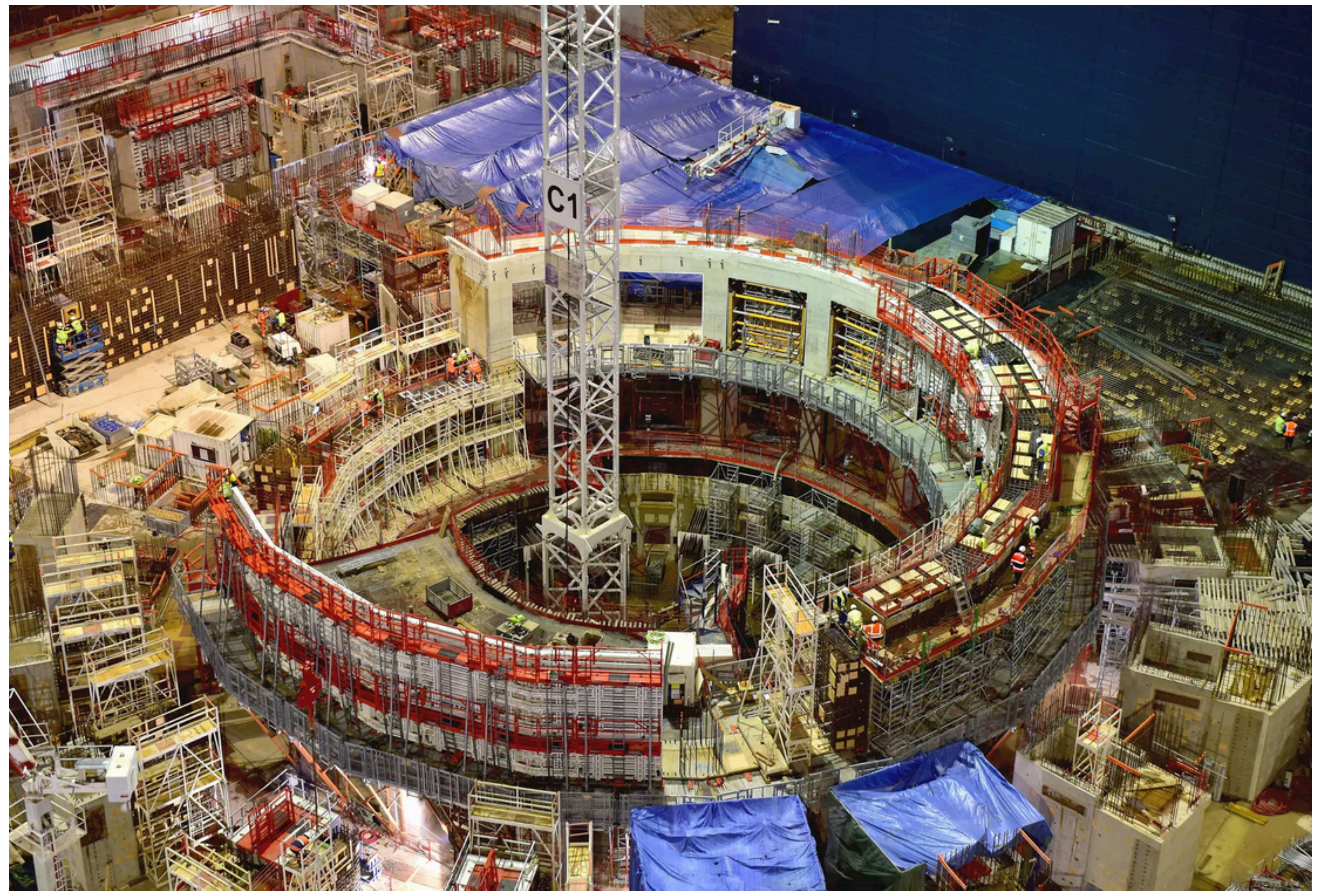
Some of humanity largest projects



LHC: $5 Billion, 23 countries

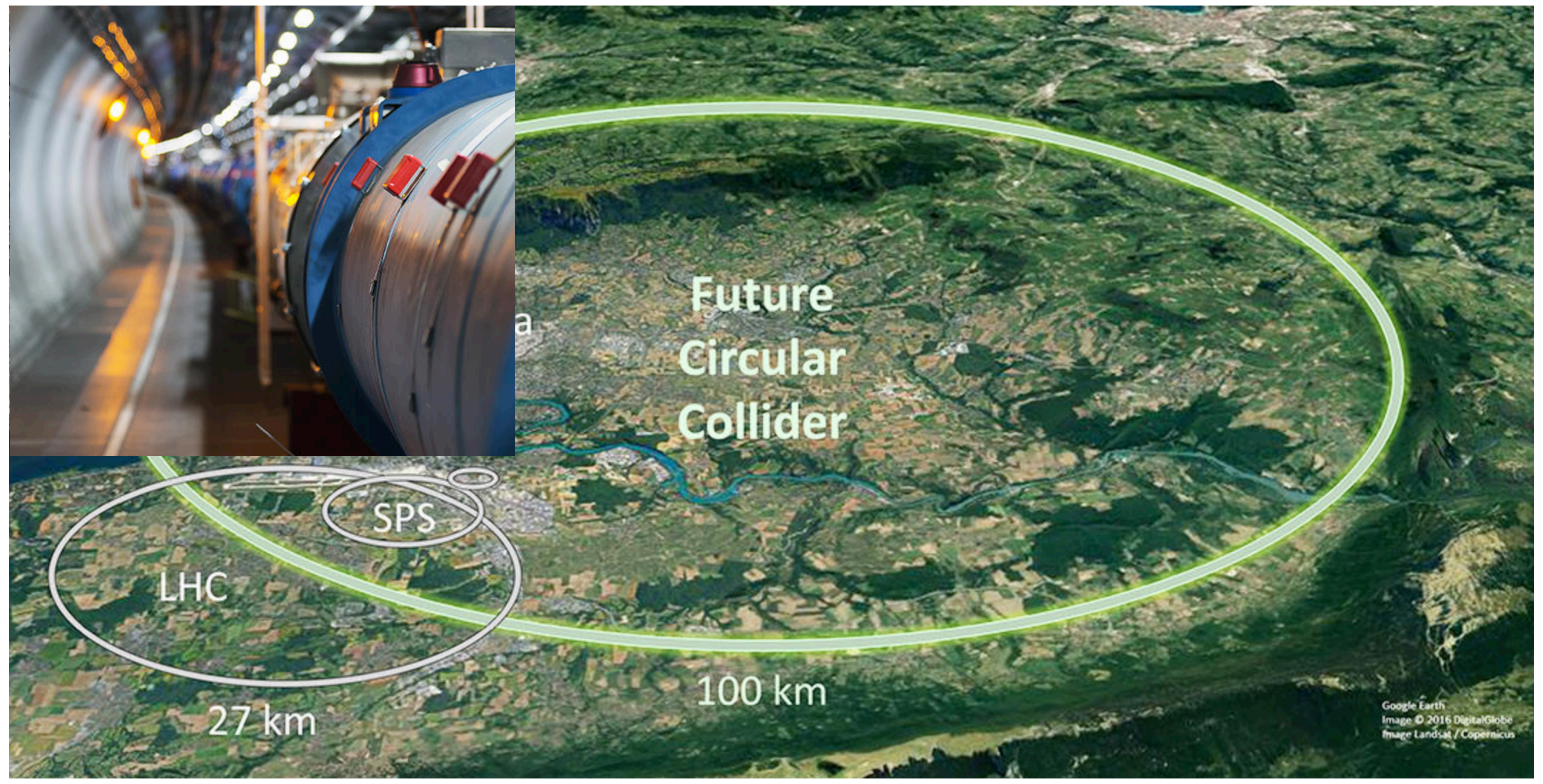


Hubble $16 billion, 11 countries

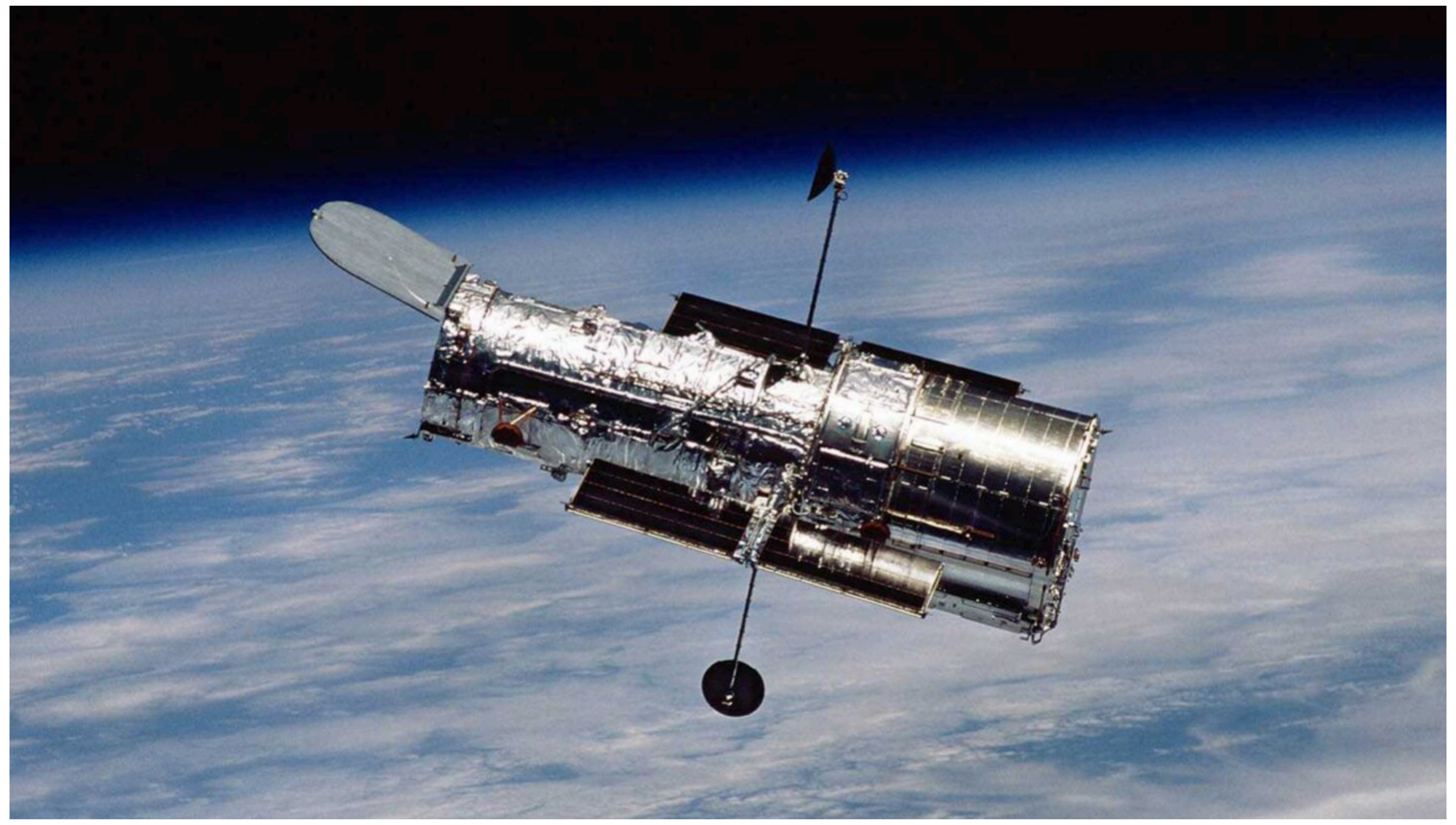


A case of openness

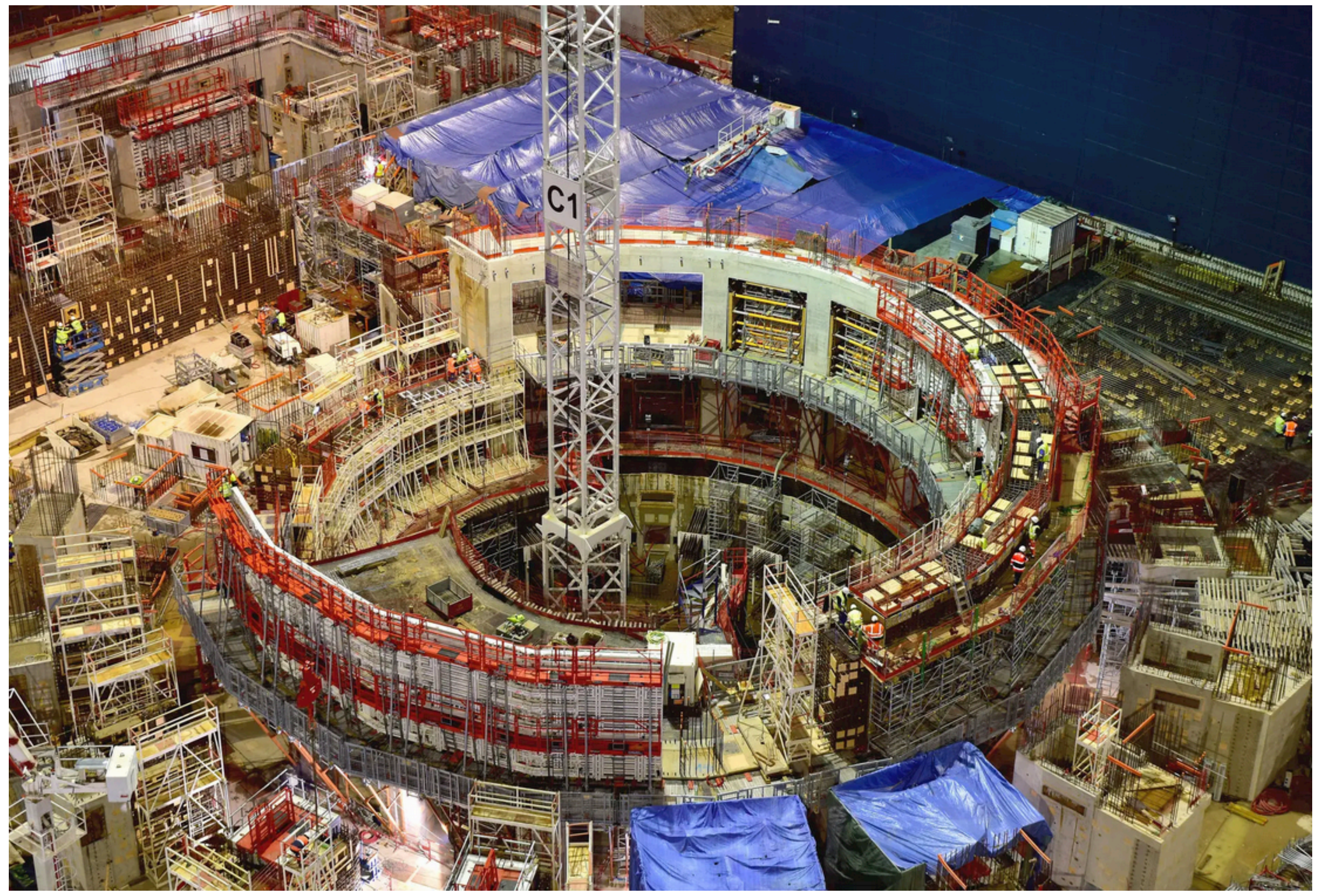
Some of humanity largest projects



LHC: $5 Billion, 23 countries



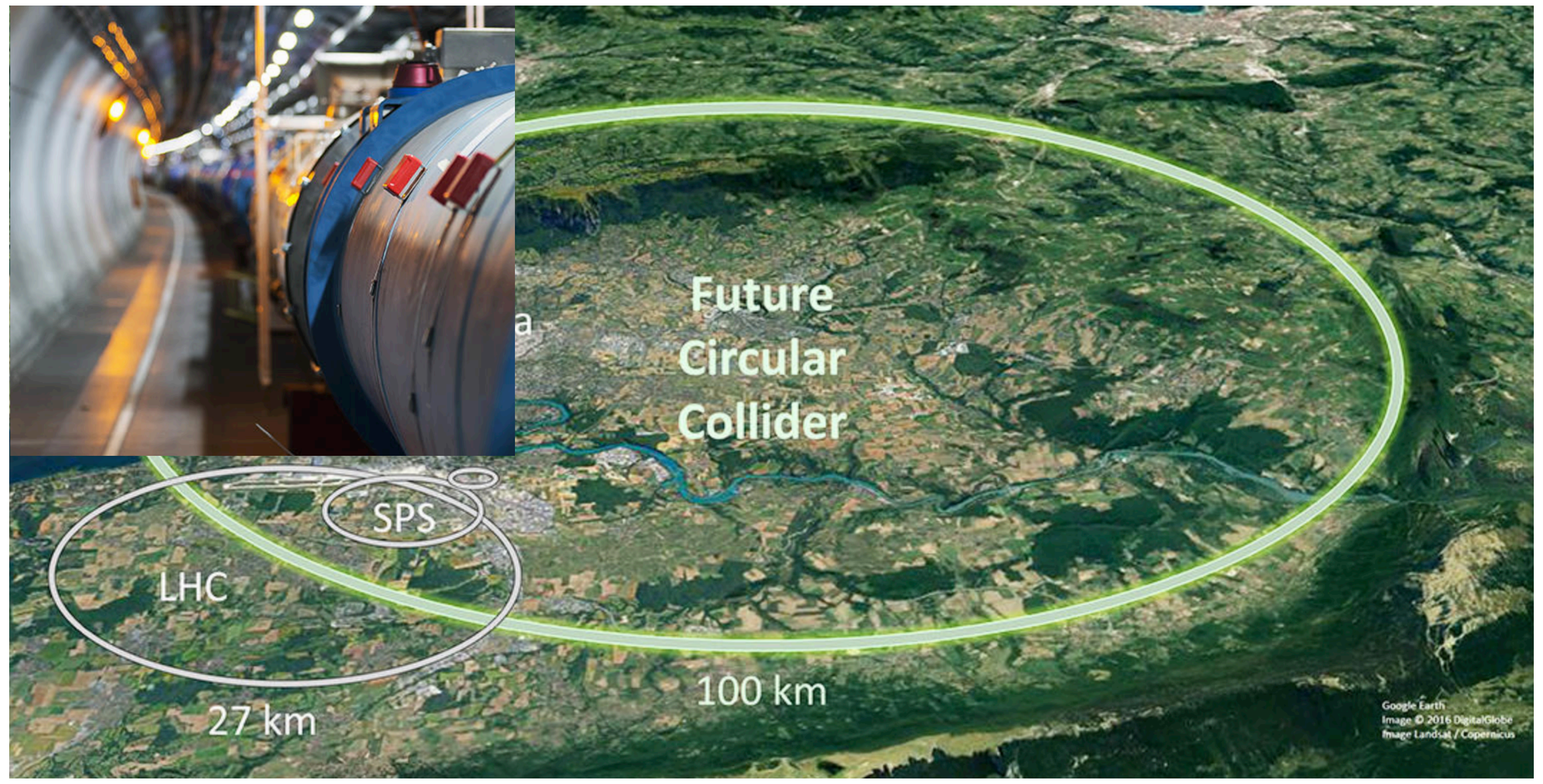
Hubble $16 billion, 11 countries



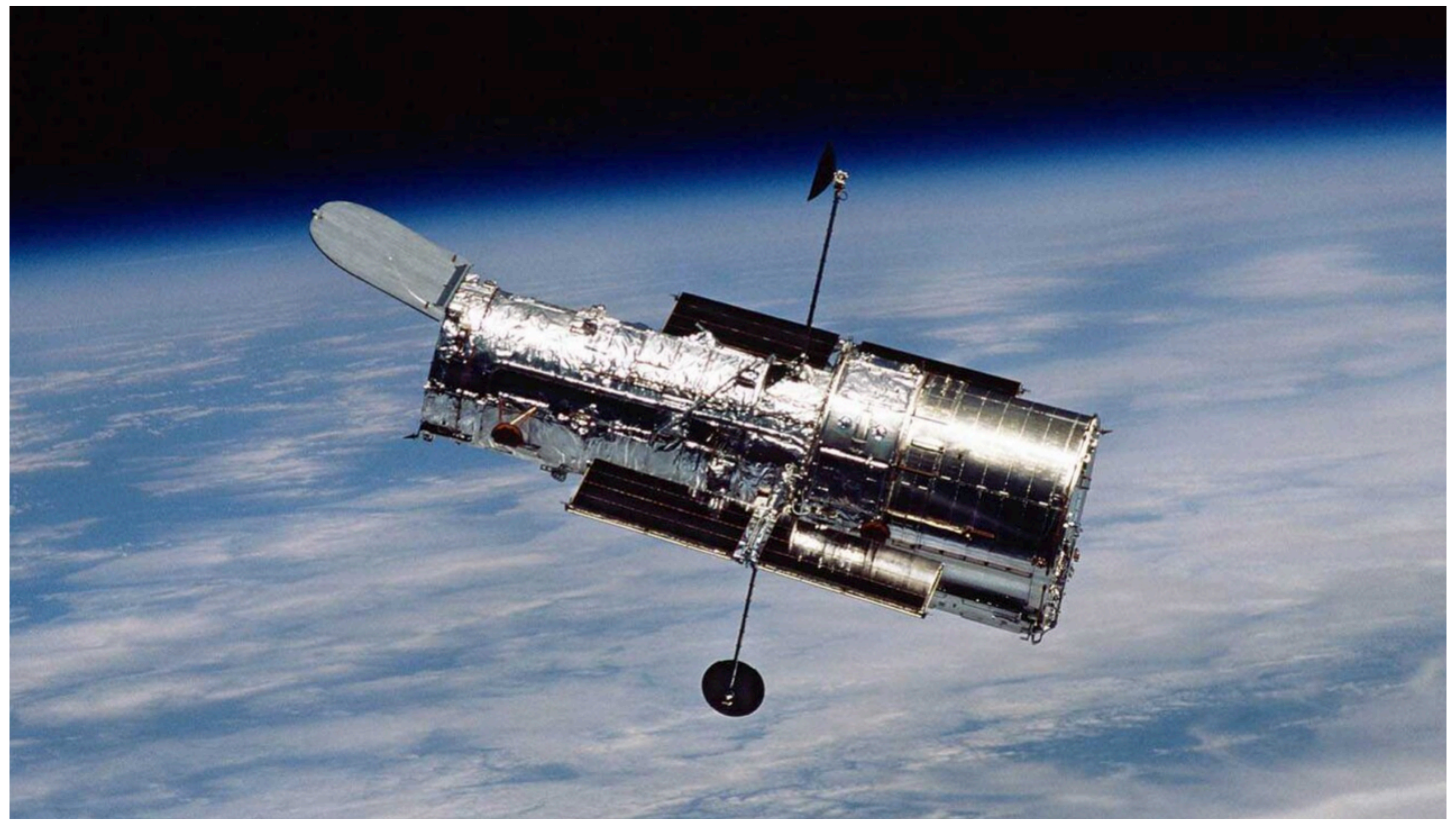
ITER: $45 Billion, 35 countries

A case of openness

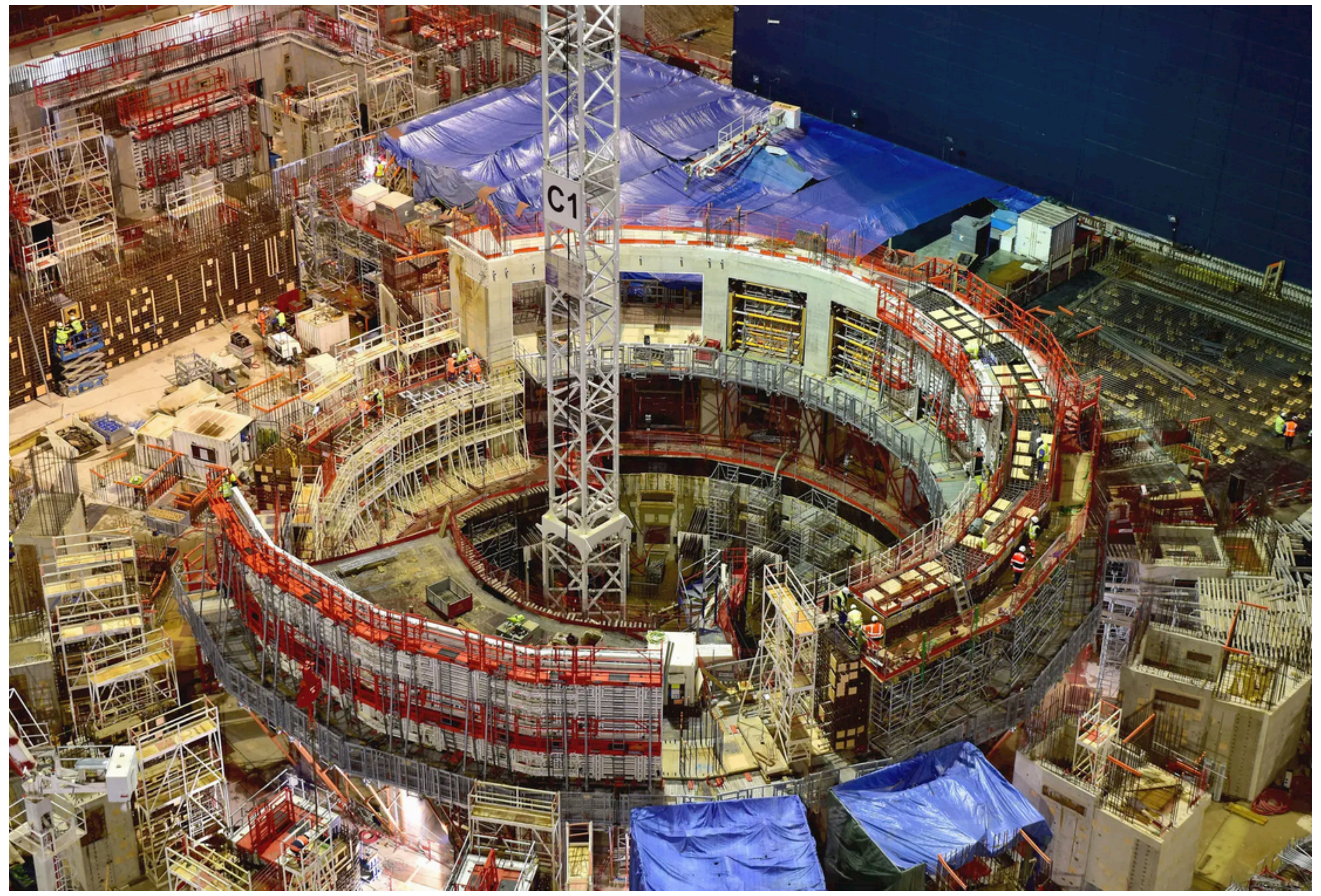
Some of humanity largest projects



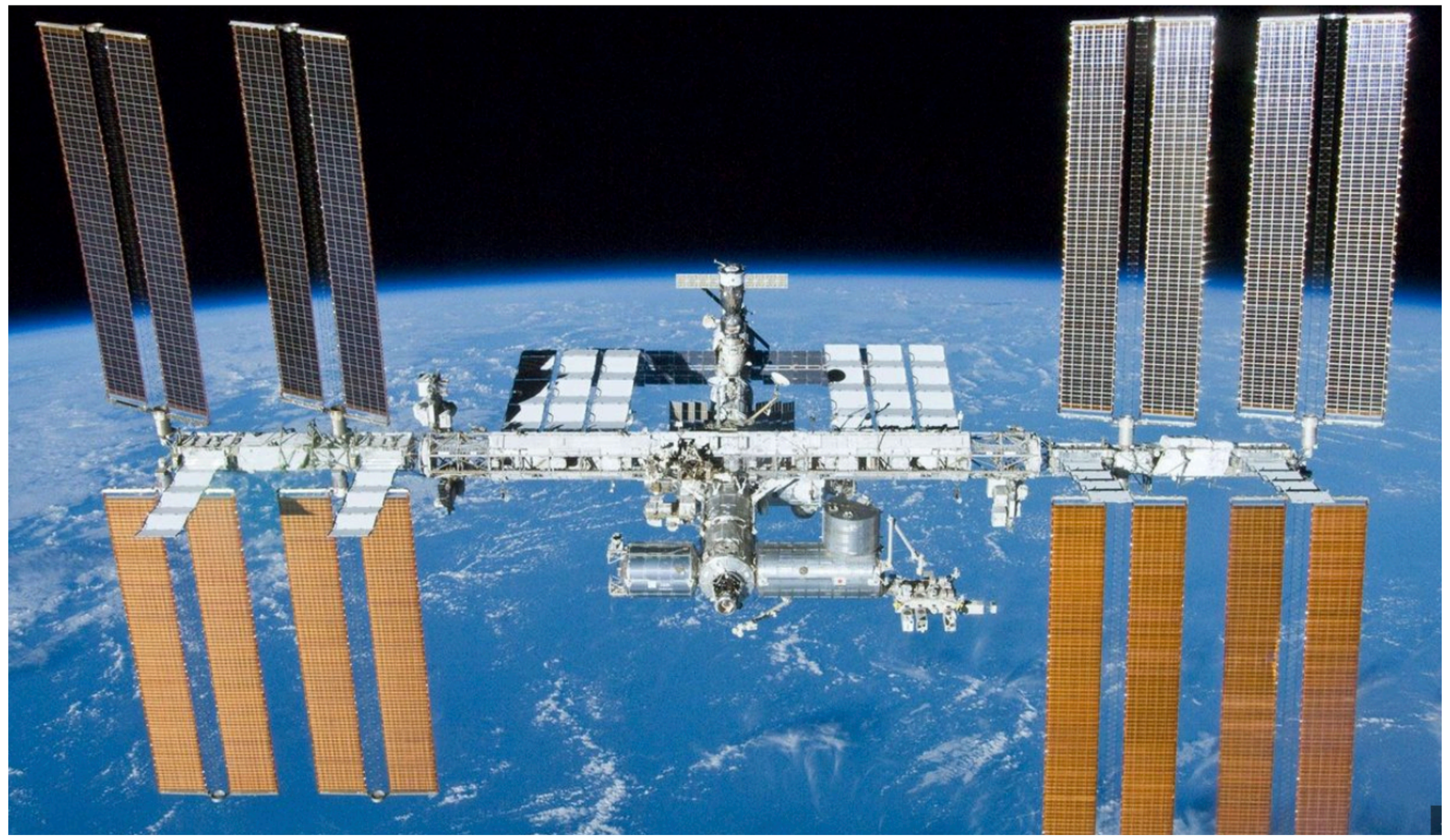
LHC: $5 Billion, 23 countries



Hubble $16 billion, 11 countries

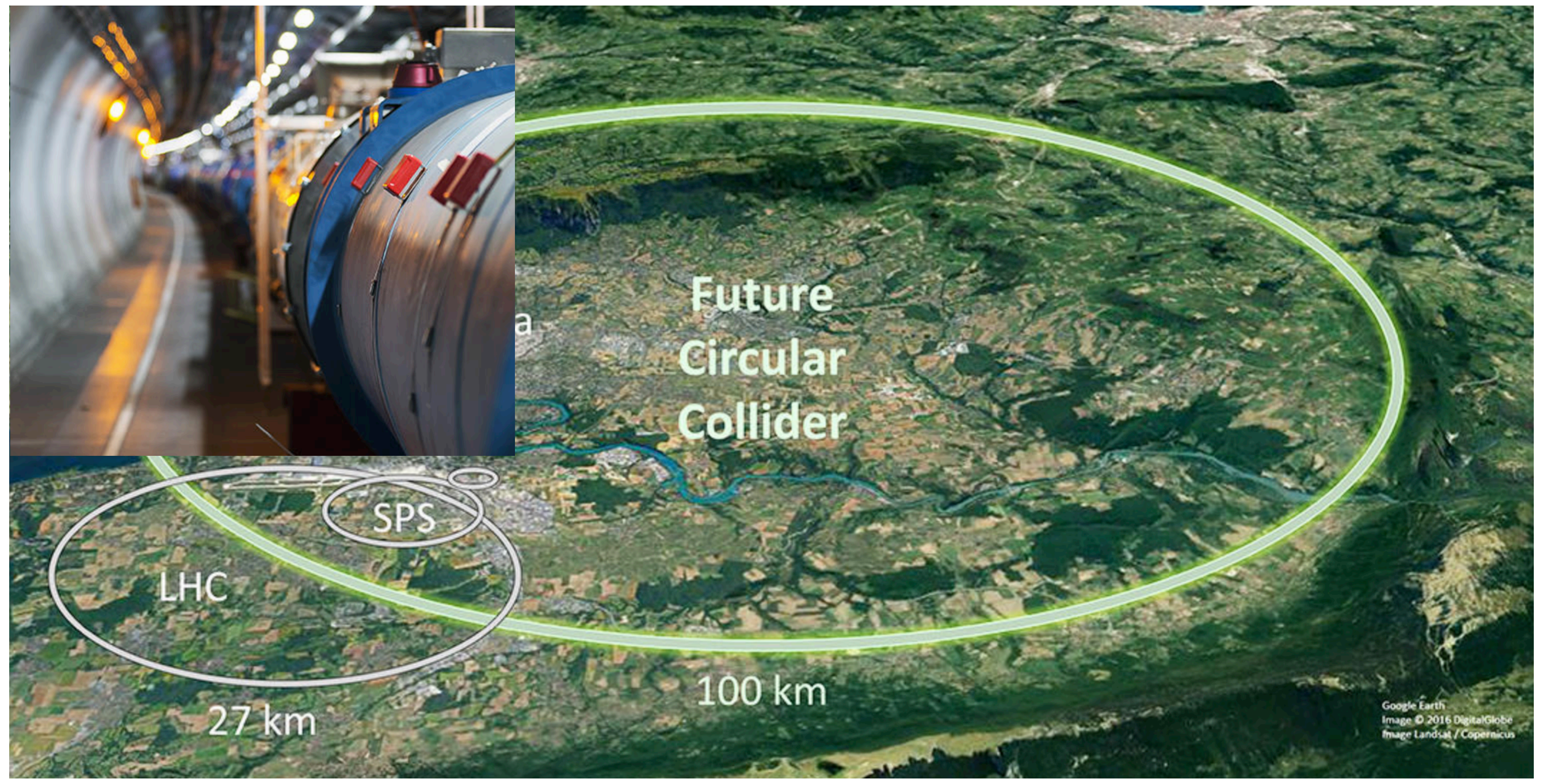


ITER: $45 Billion, 35 countries

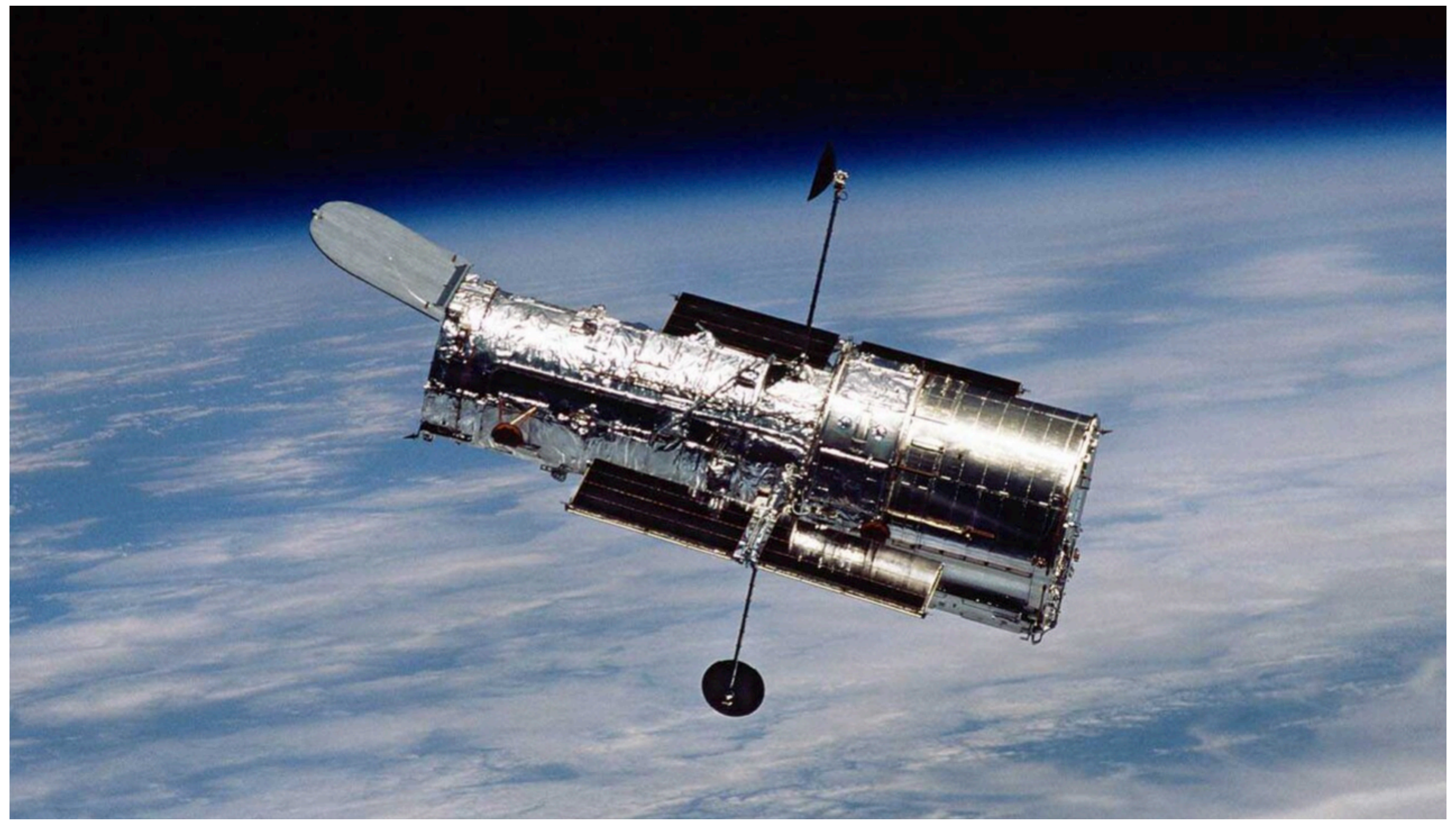


A case of openness

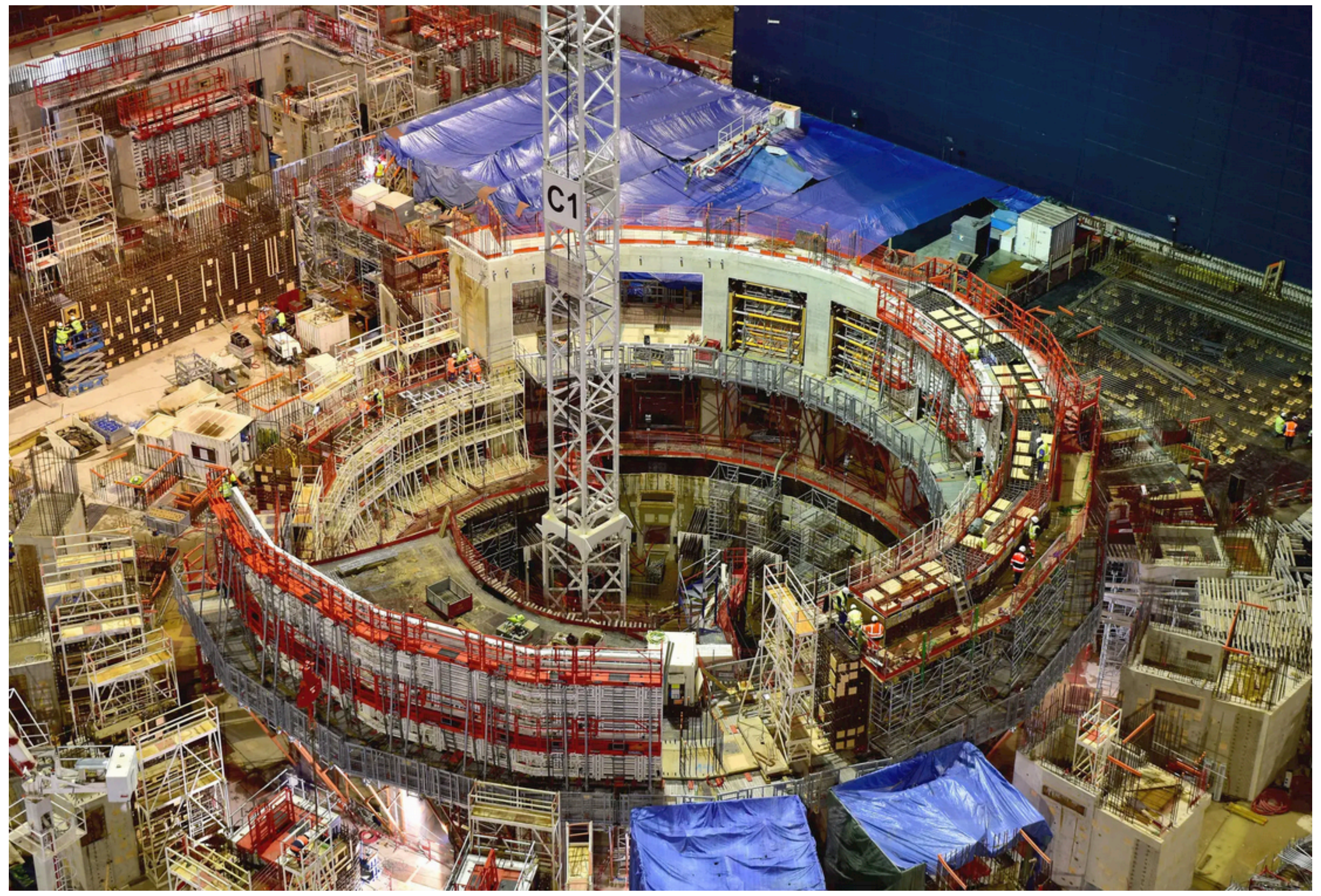
Some of humanity largest projects



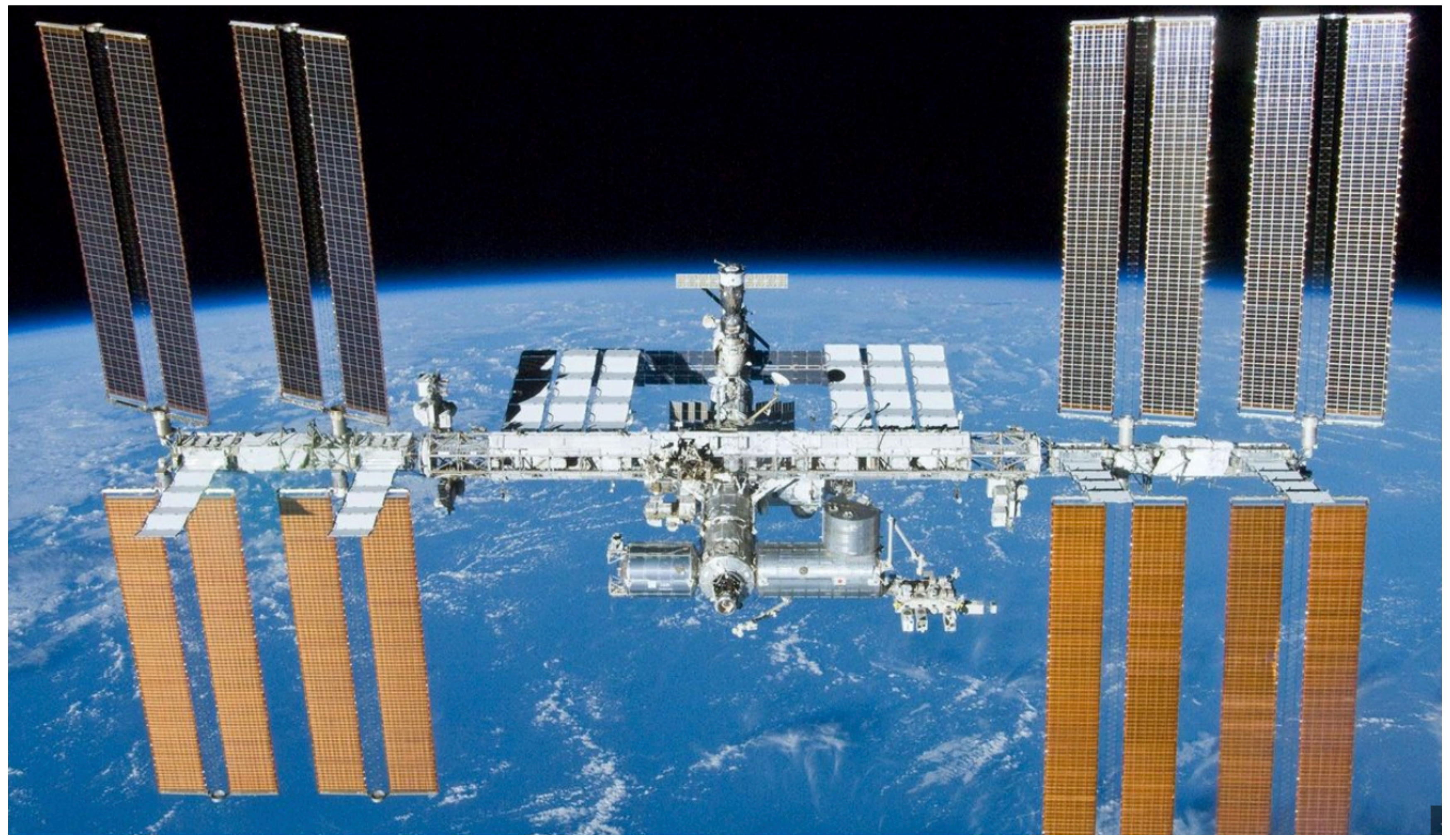
LHC: $5 Billion, 23 countries



Hubble $16 billion, 11 countries



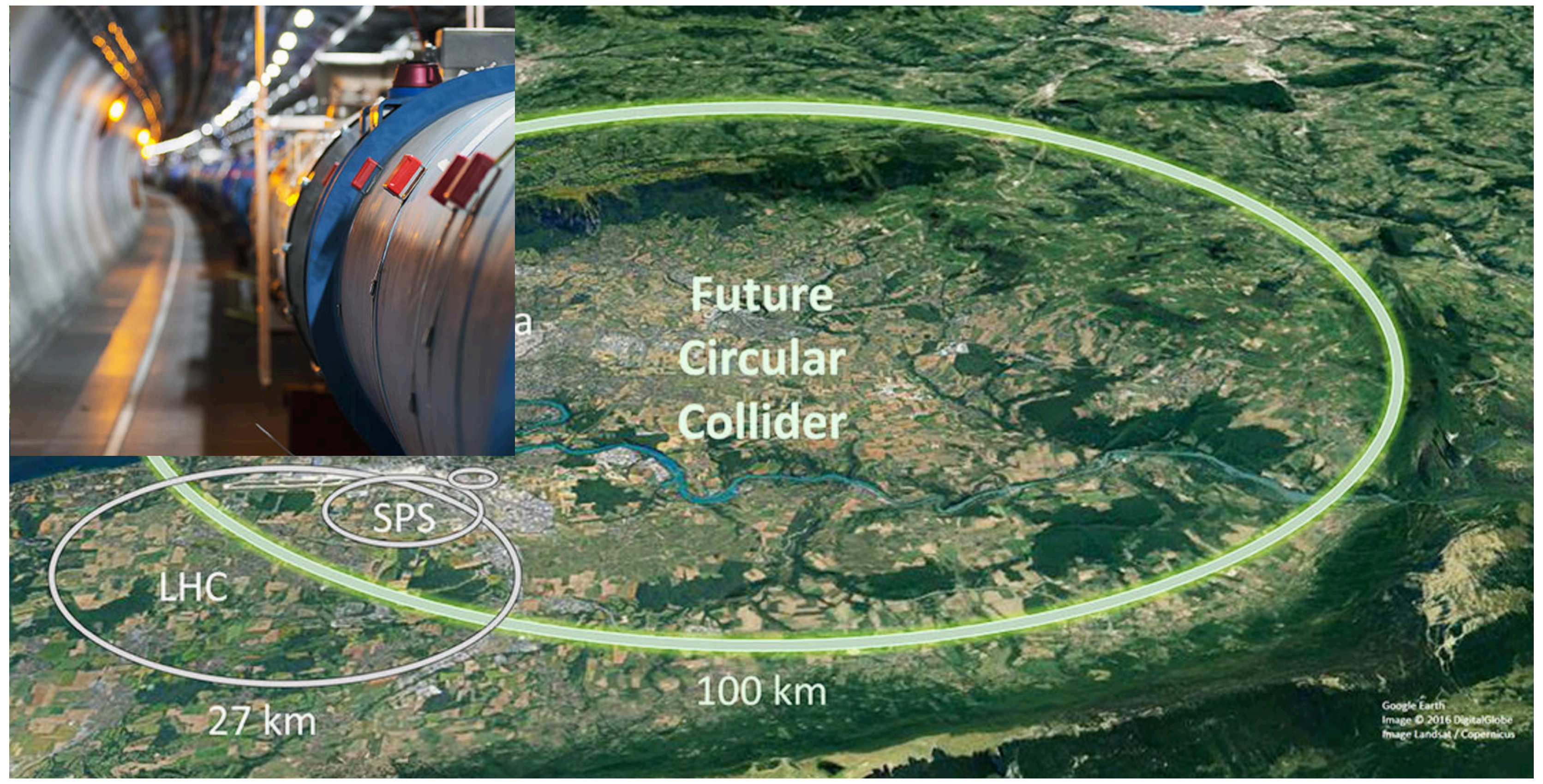
ITER: $45 Billion, 35 countries



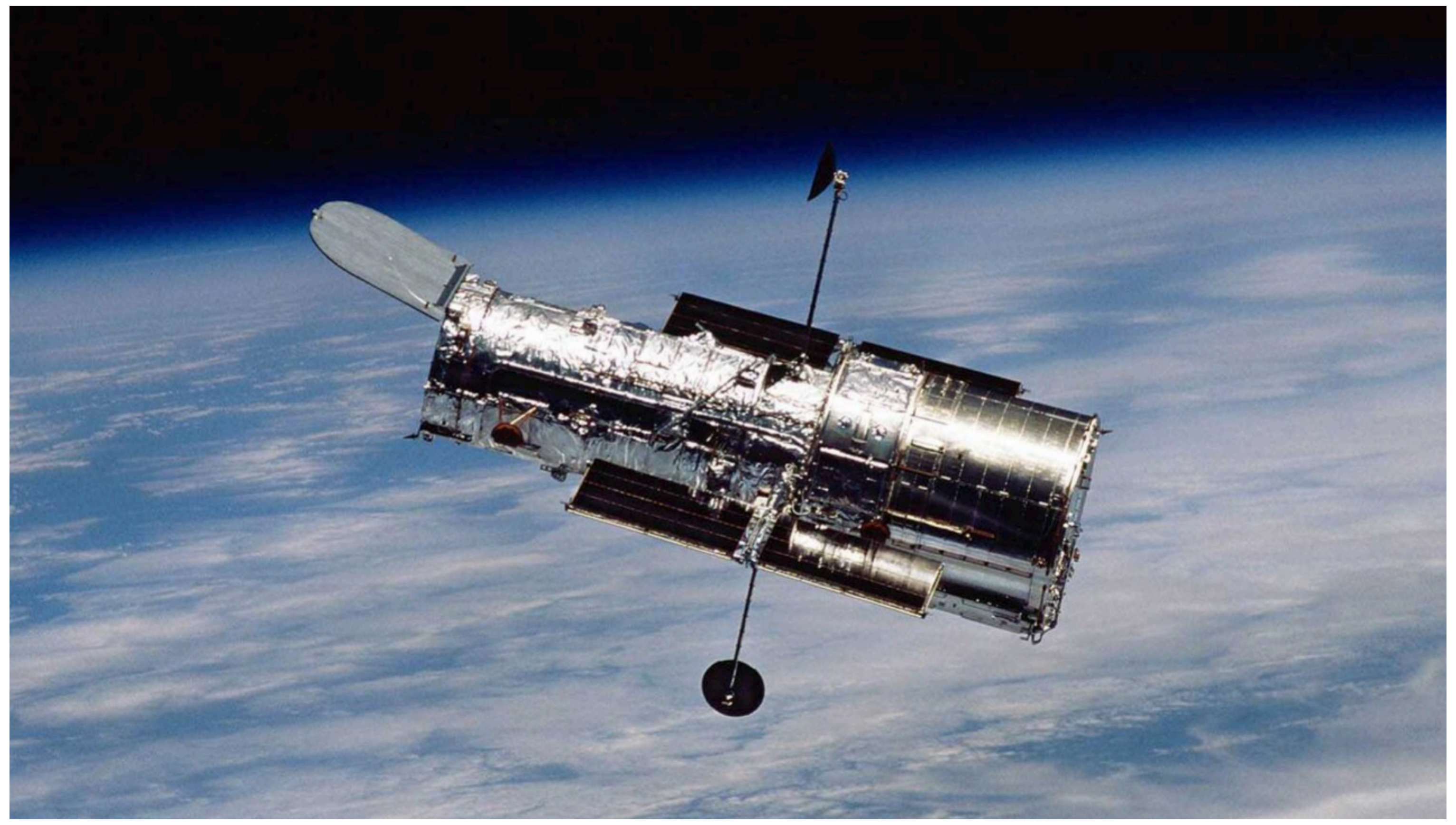
ISS: $100 Billion, 16 countries

A case of openness

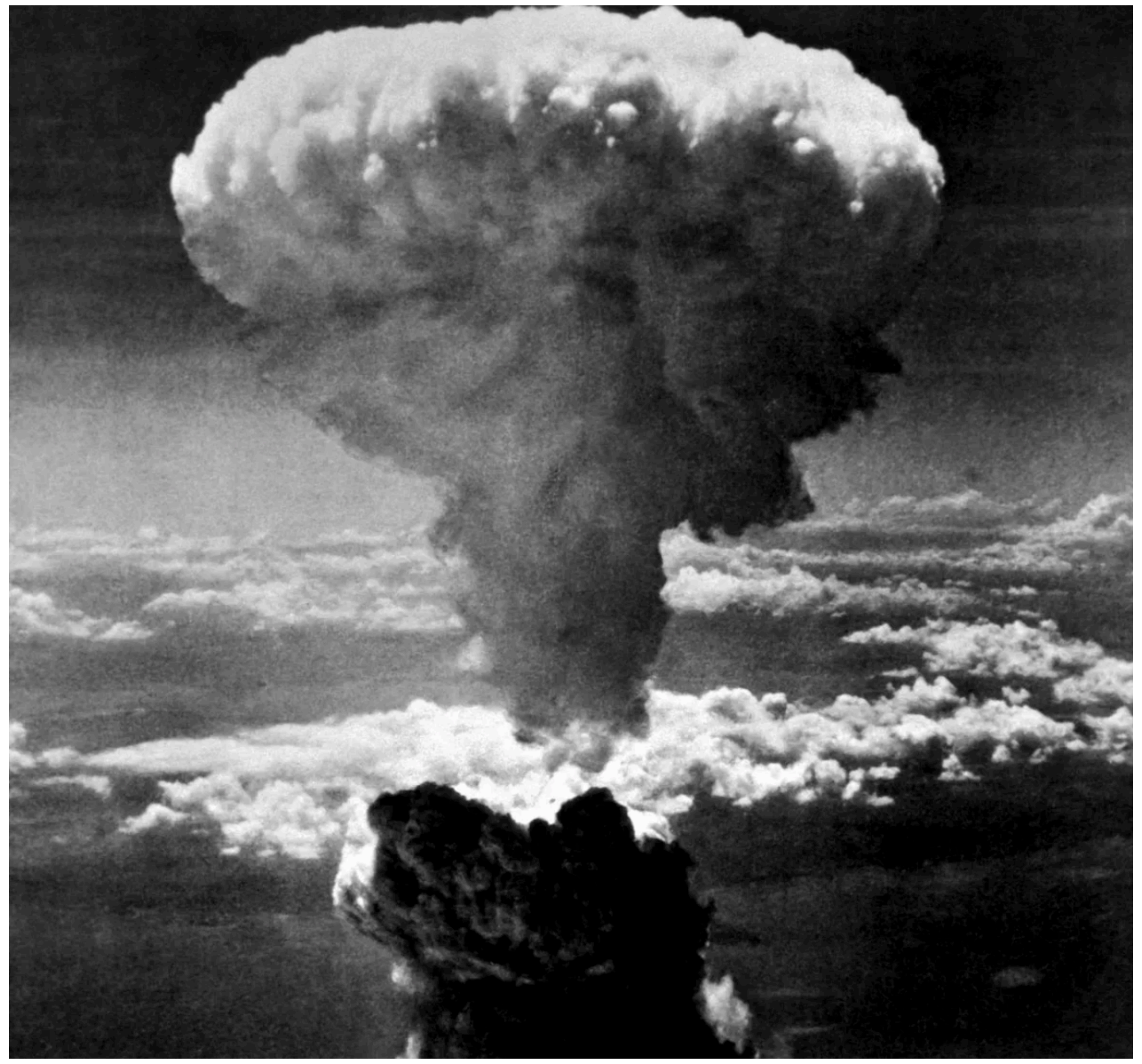
Some of humanity largest projects

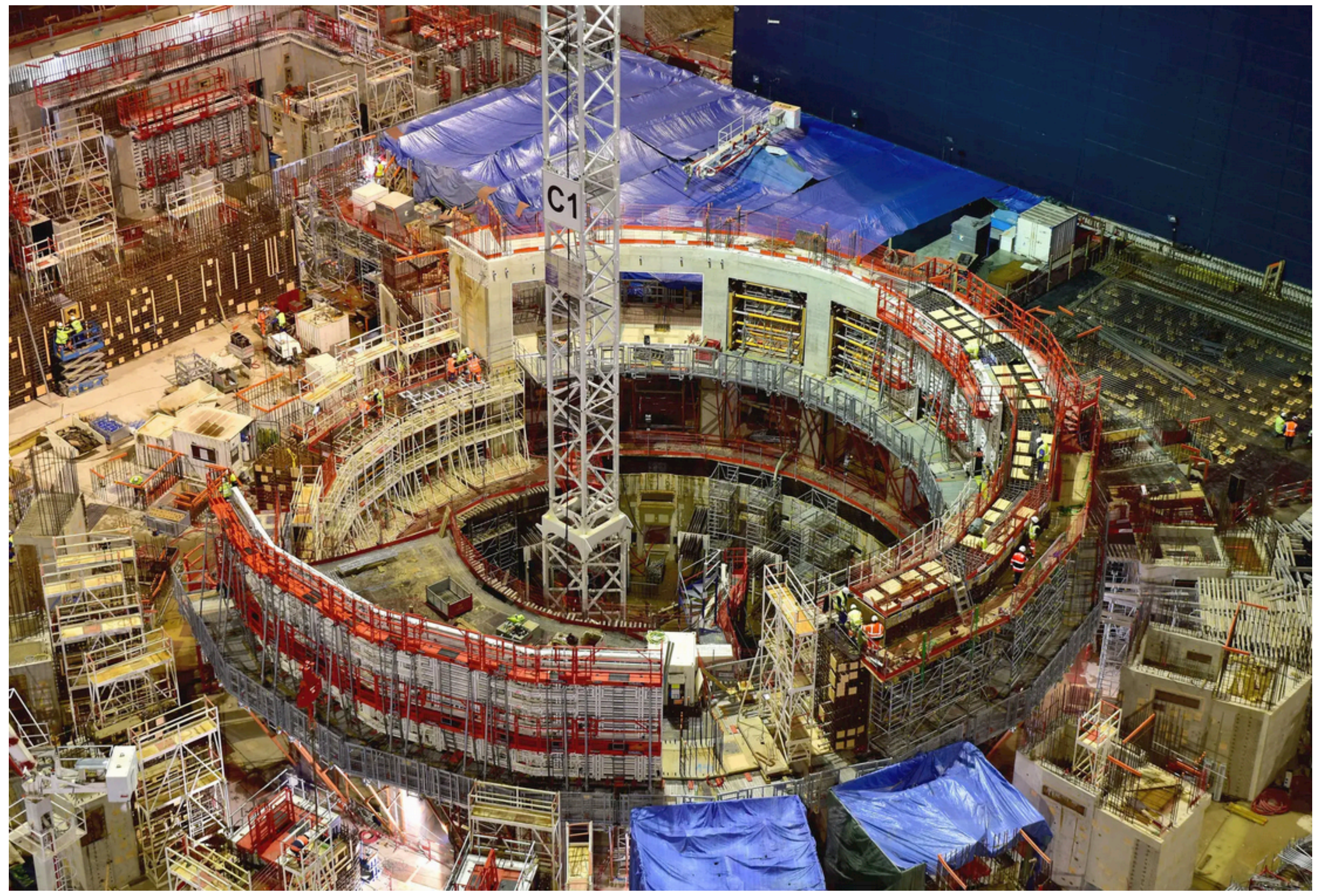


LHC: $5 Billion, 23 countries

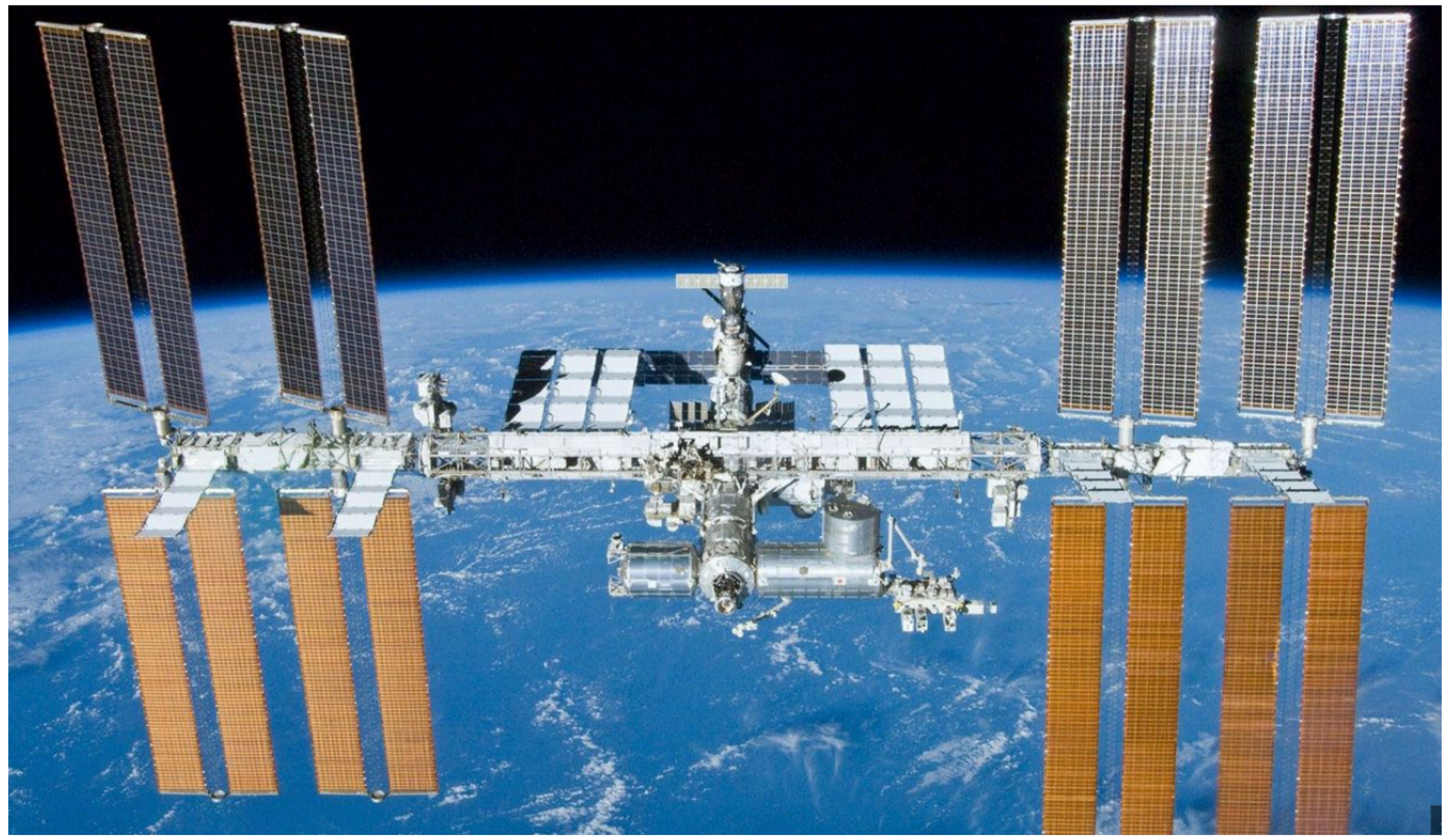


Hubble $16 billion, 11 countries





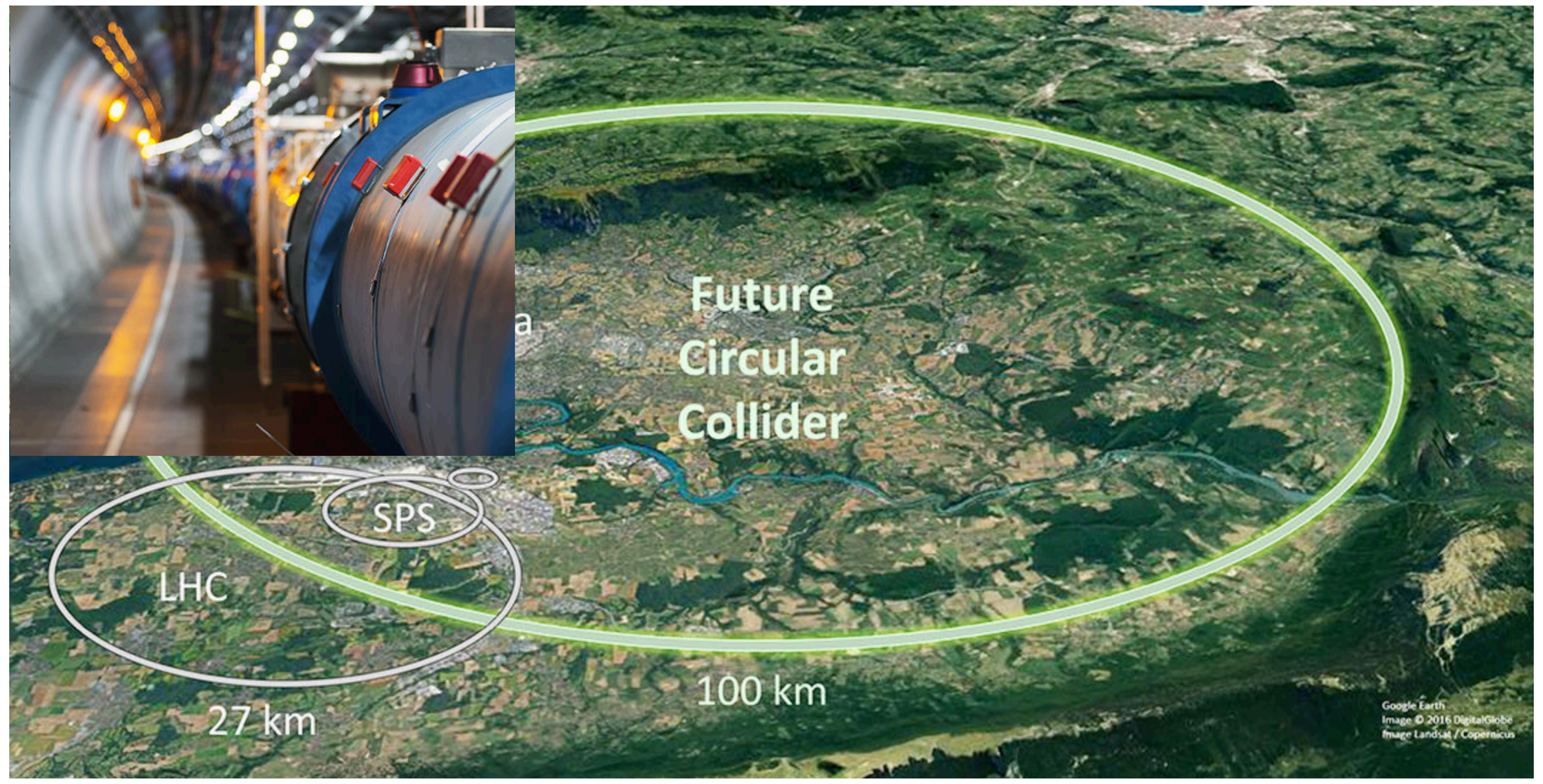
ITER: $45 Billion, 35 countries



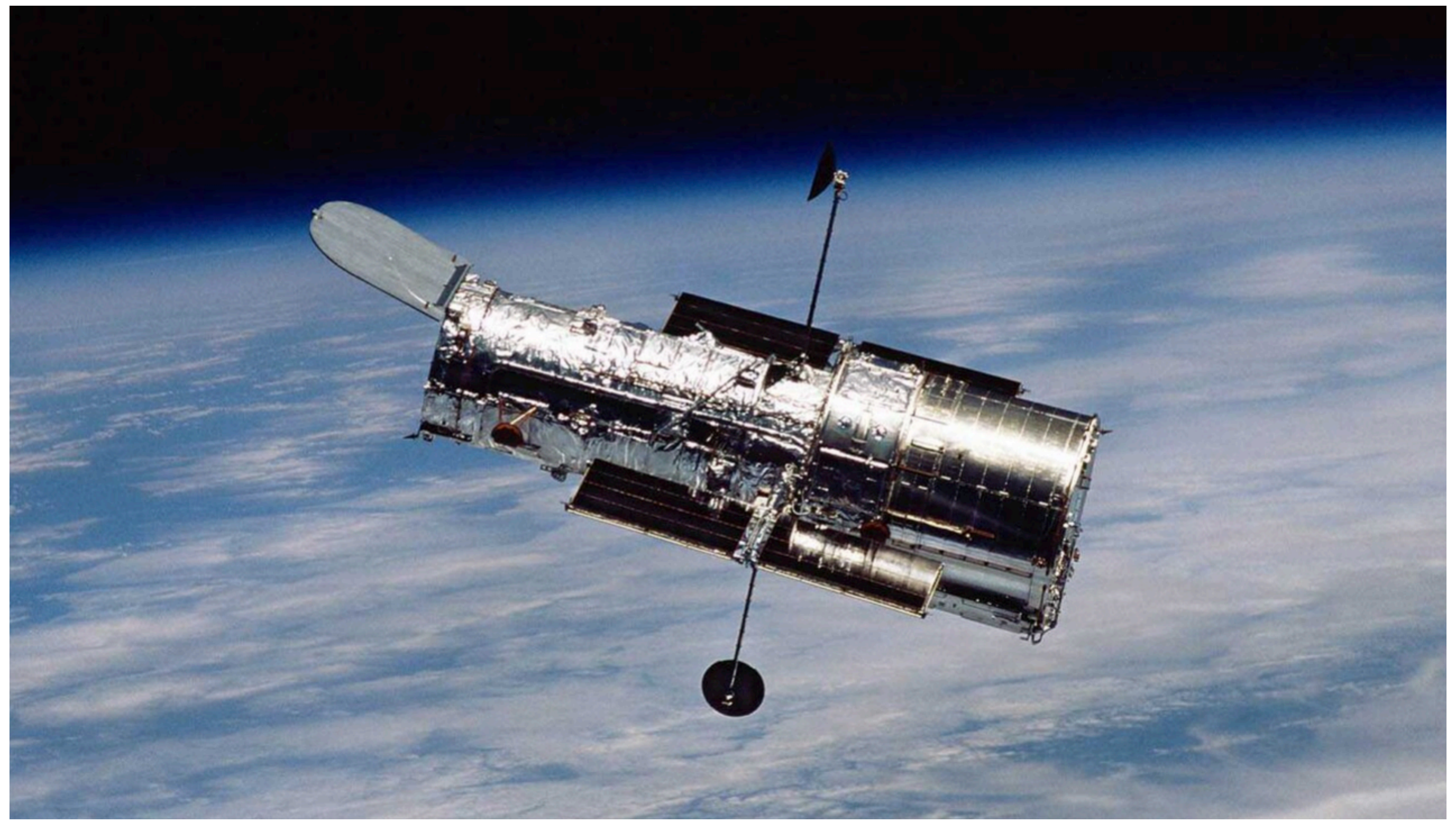
ISS: $100 Billion, 16 countries

A case of openness

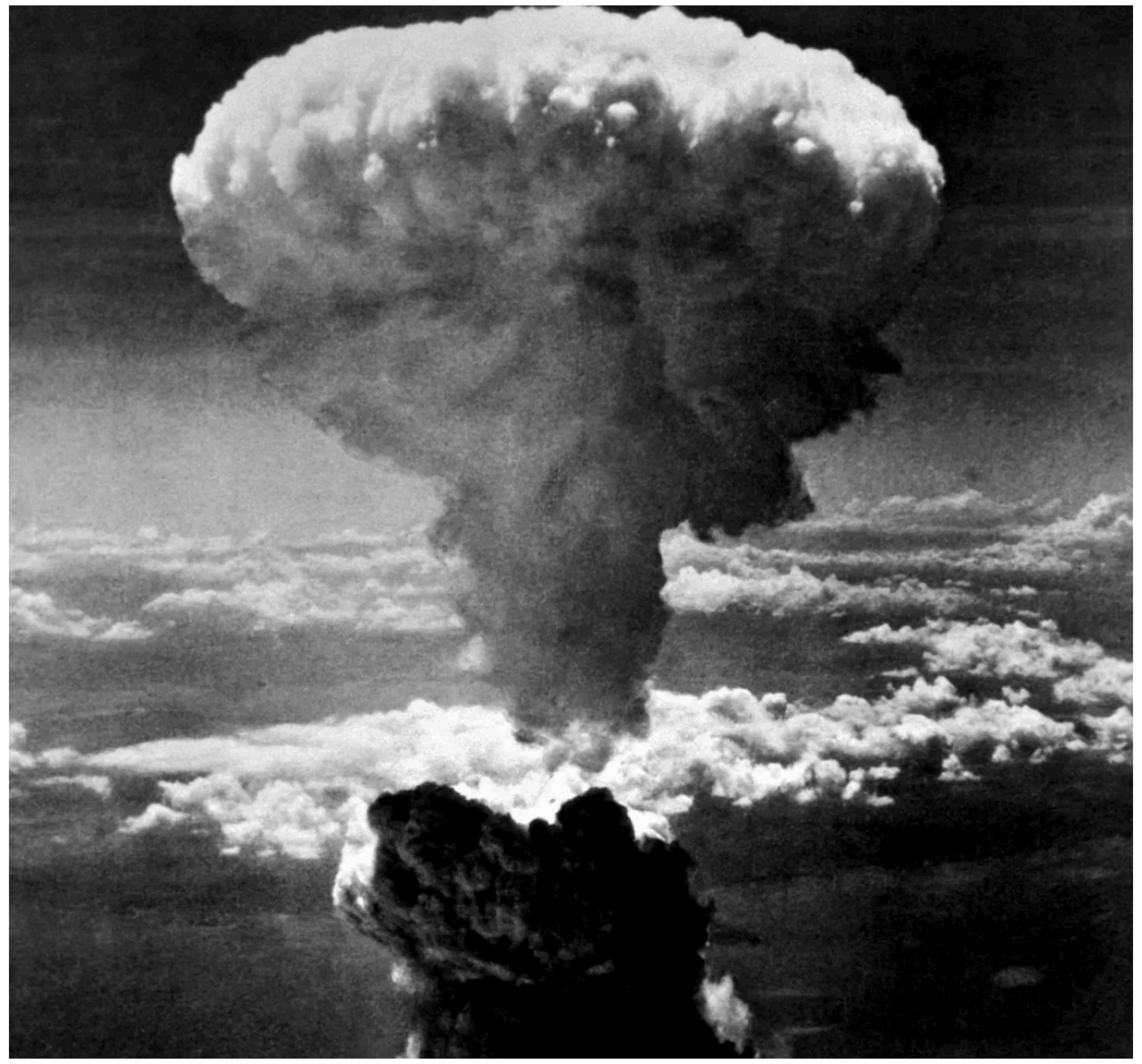
Some of humanity largest projects



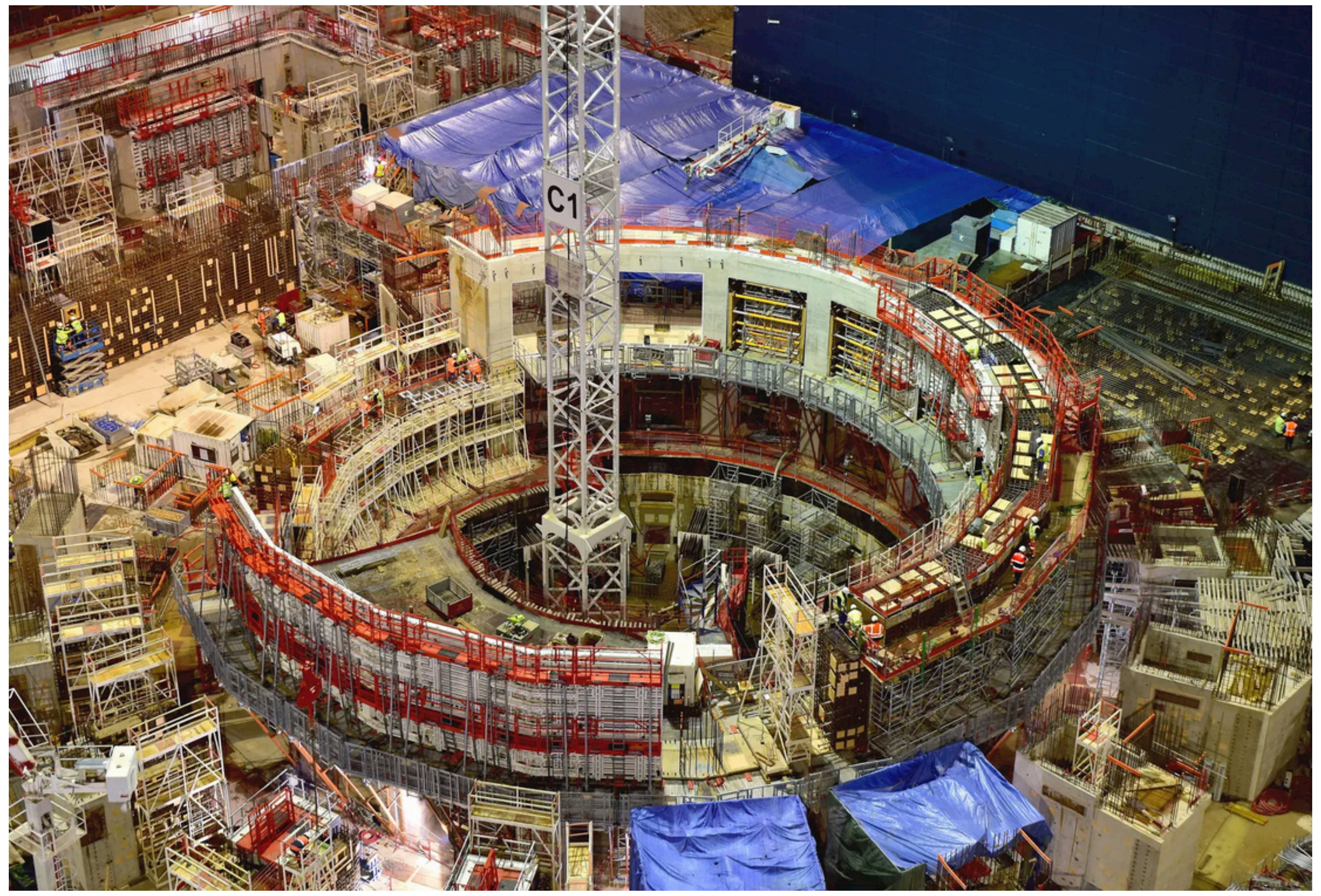
LHC: $5 Billion, 23 countries



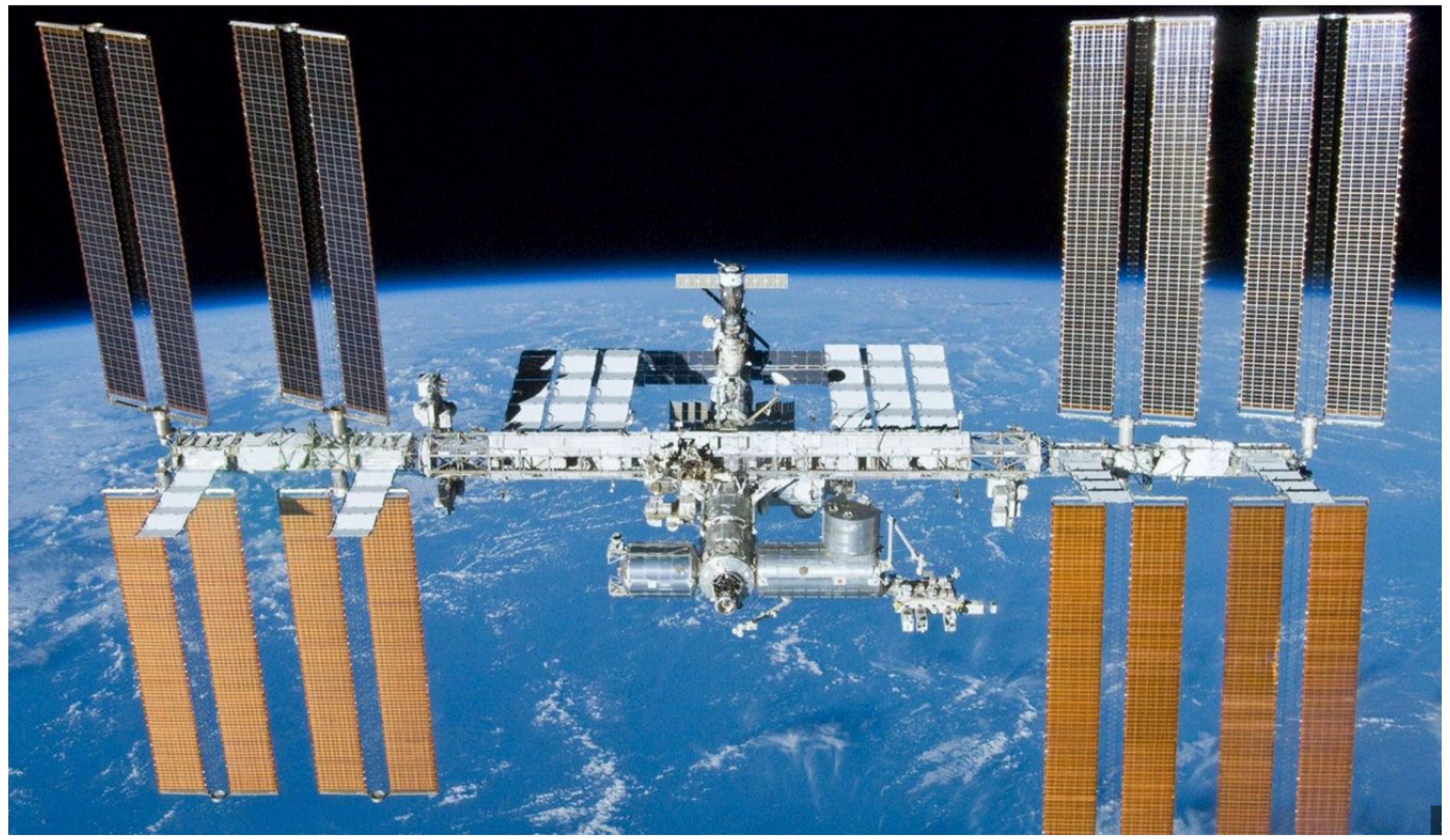
Hubble $16 billion, 11 countries



Manhattan project $30 billion, 3 countries



ITER: $45 Billion, 35 countries



ISS: $100 Billion, 16 countries

The case of LLMs

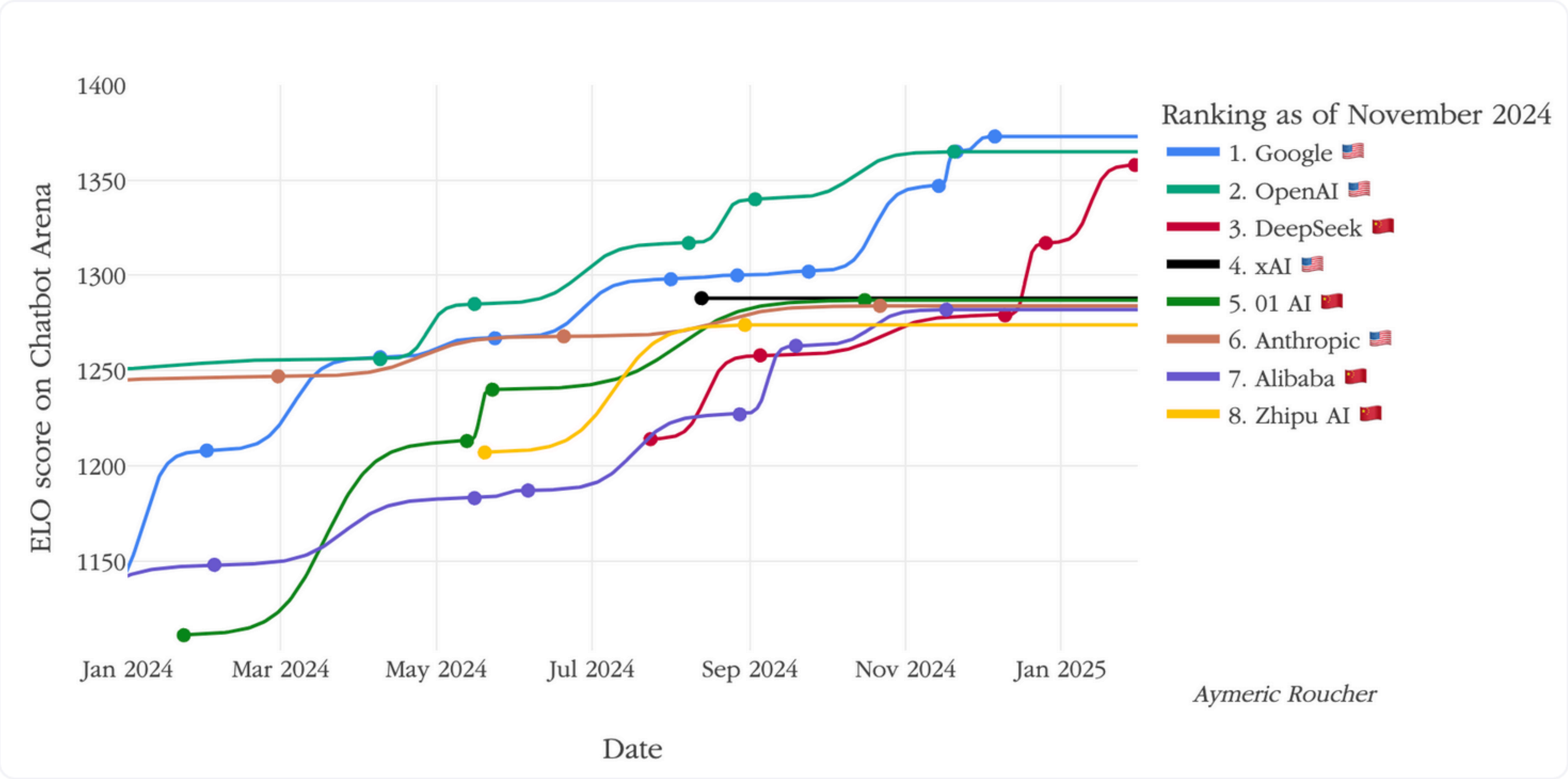
The case of LLMs

- Currently an arm race

The case of LLMs

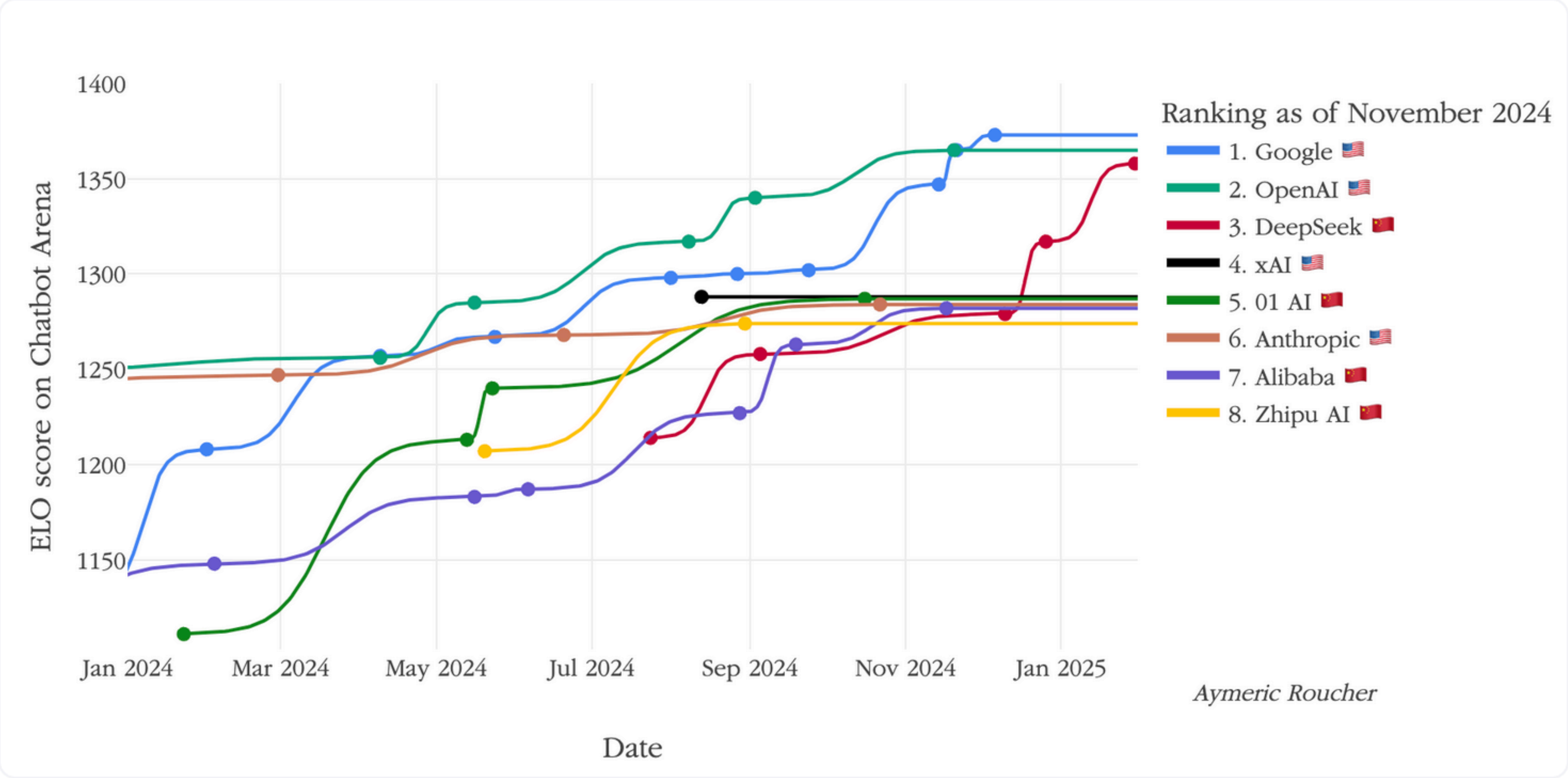
- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time

The case of LLMs



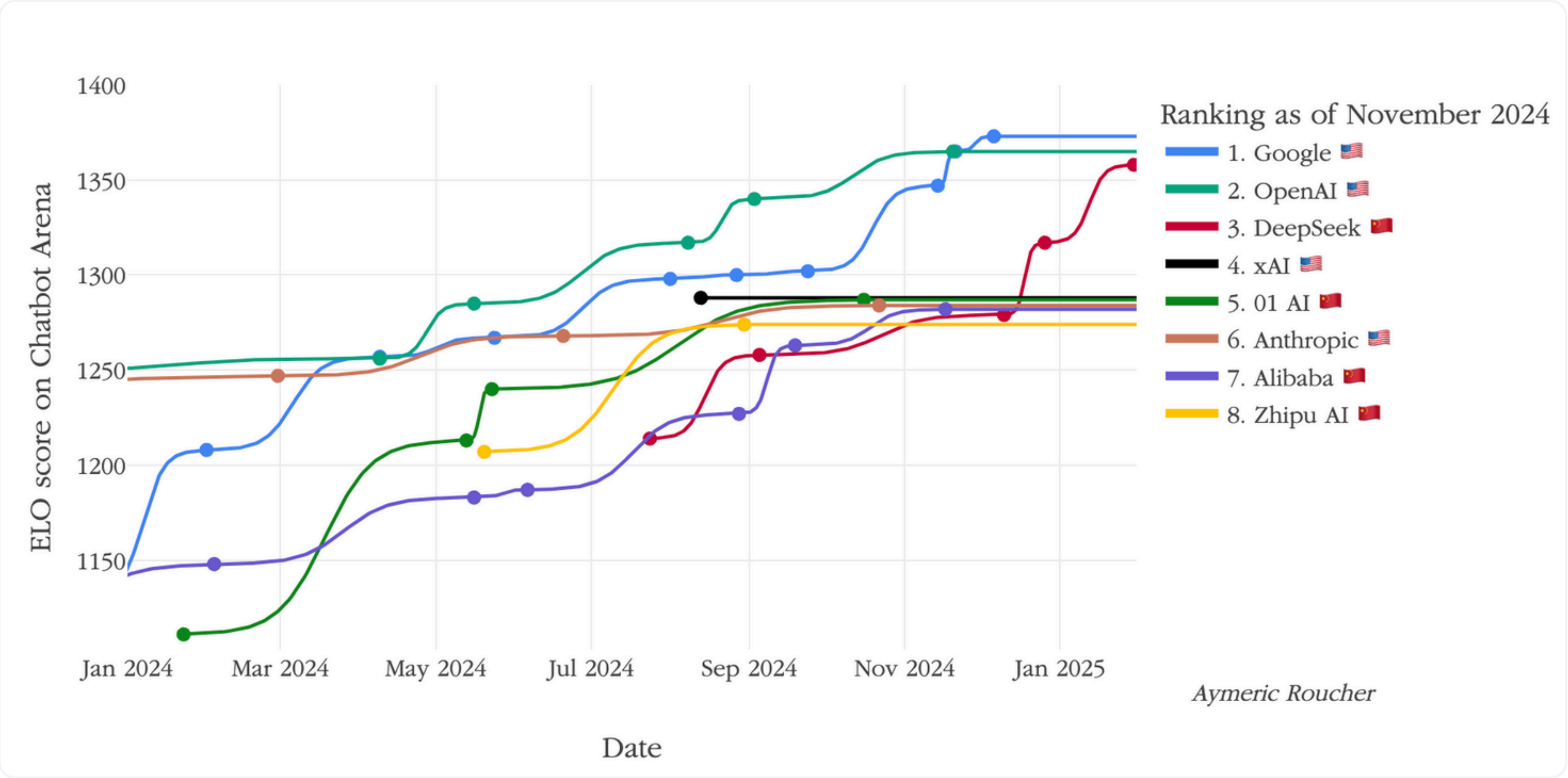
- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time

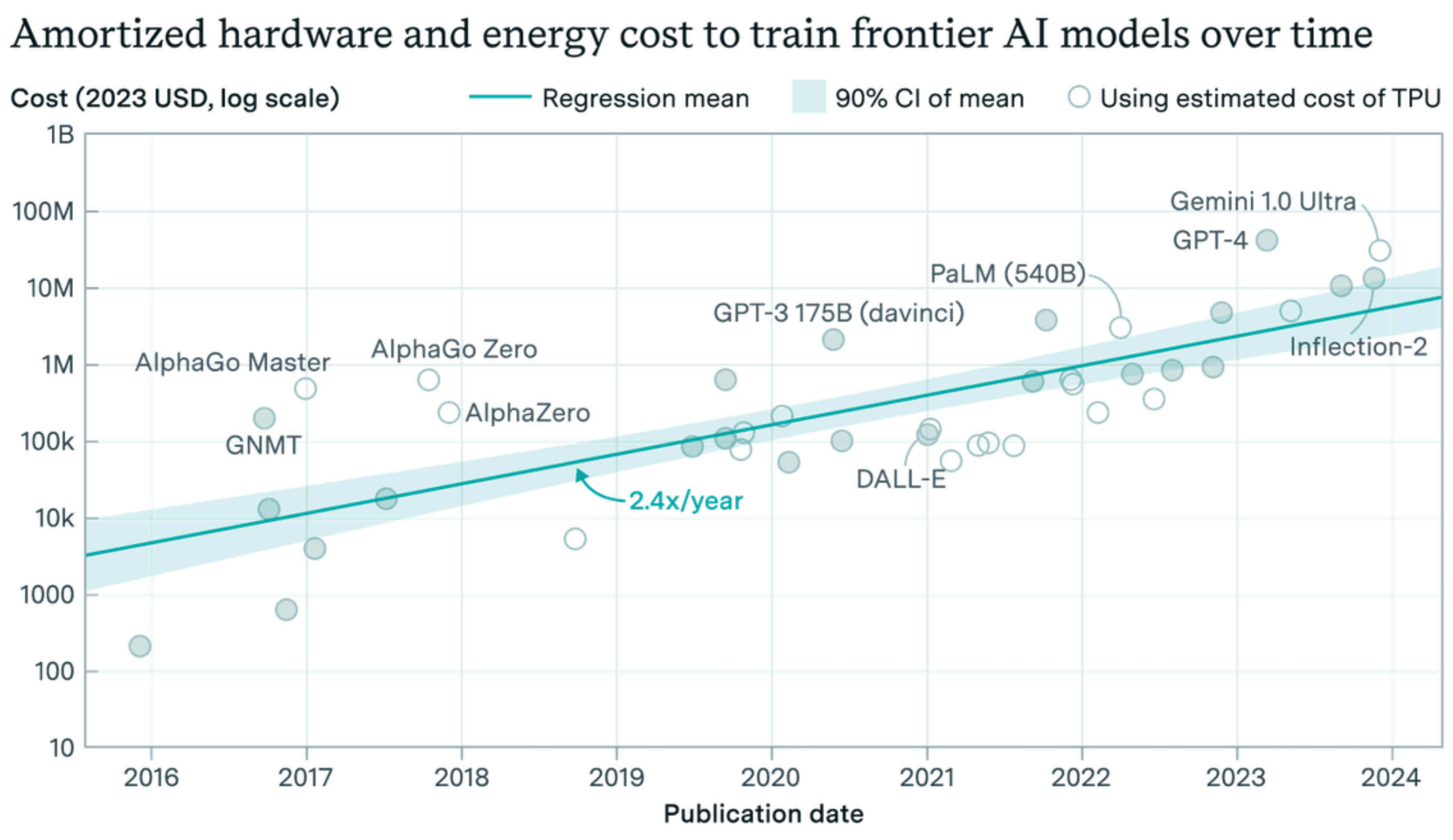
The case of LLMs



- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)

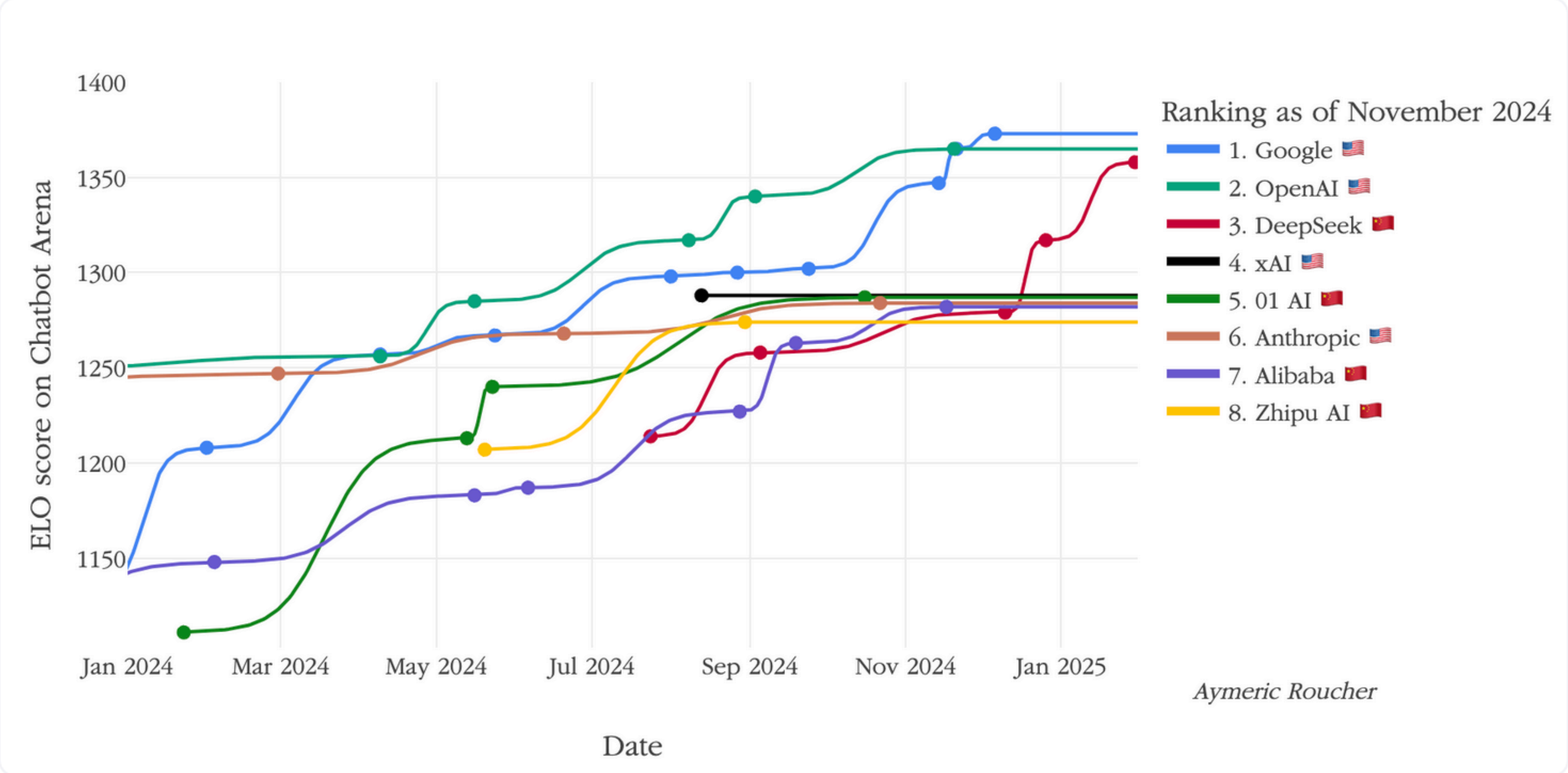
The case of LLMs

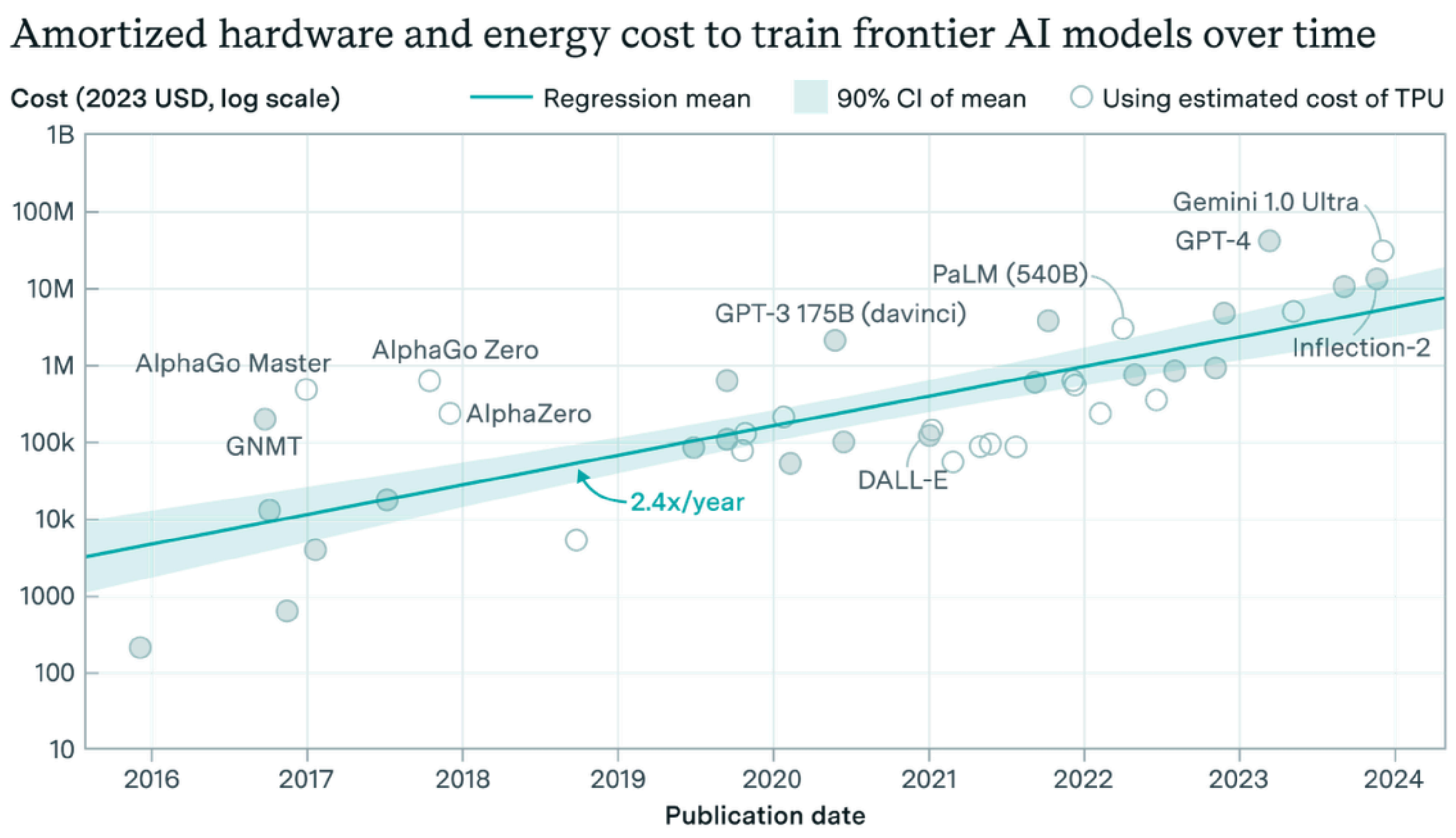




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)

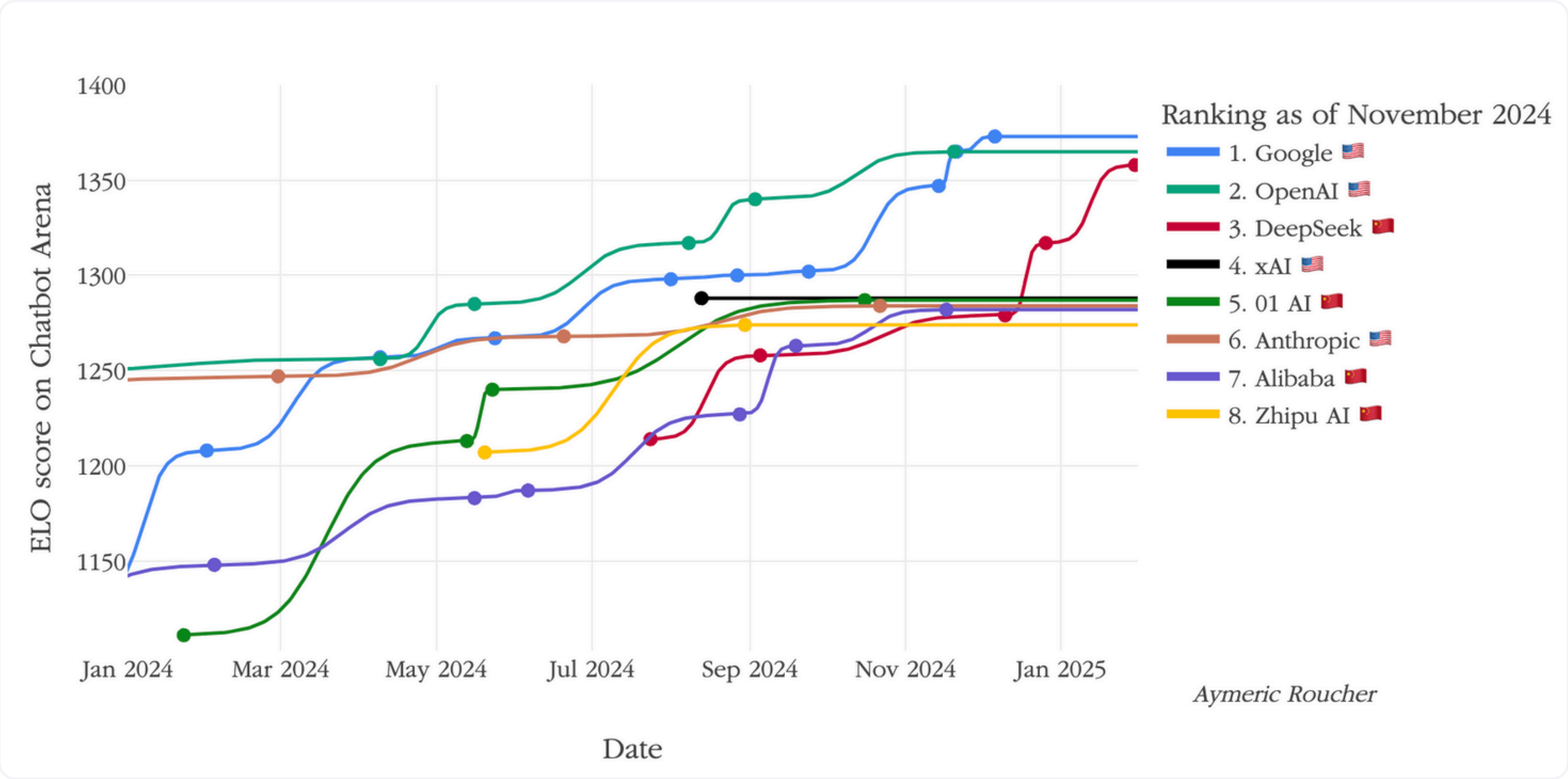
The case of LLMs

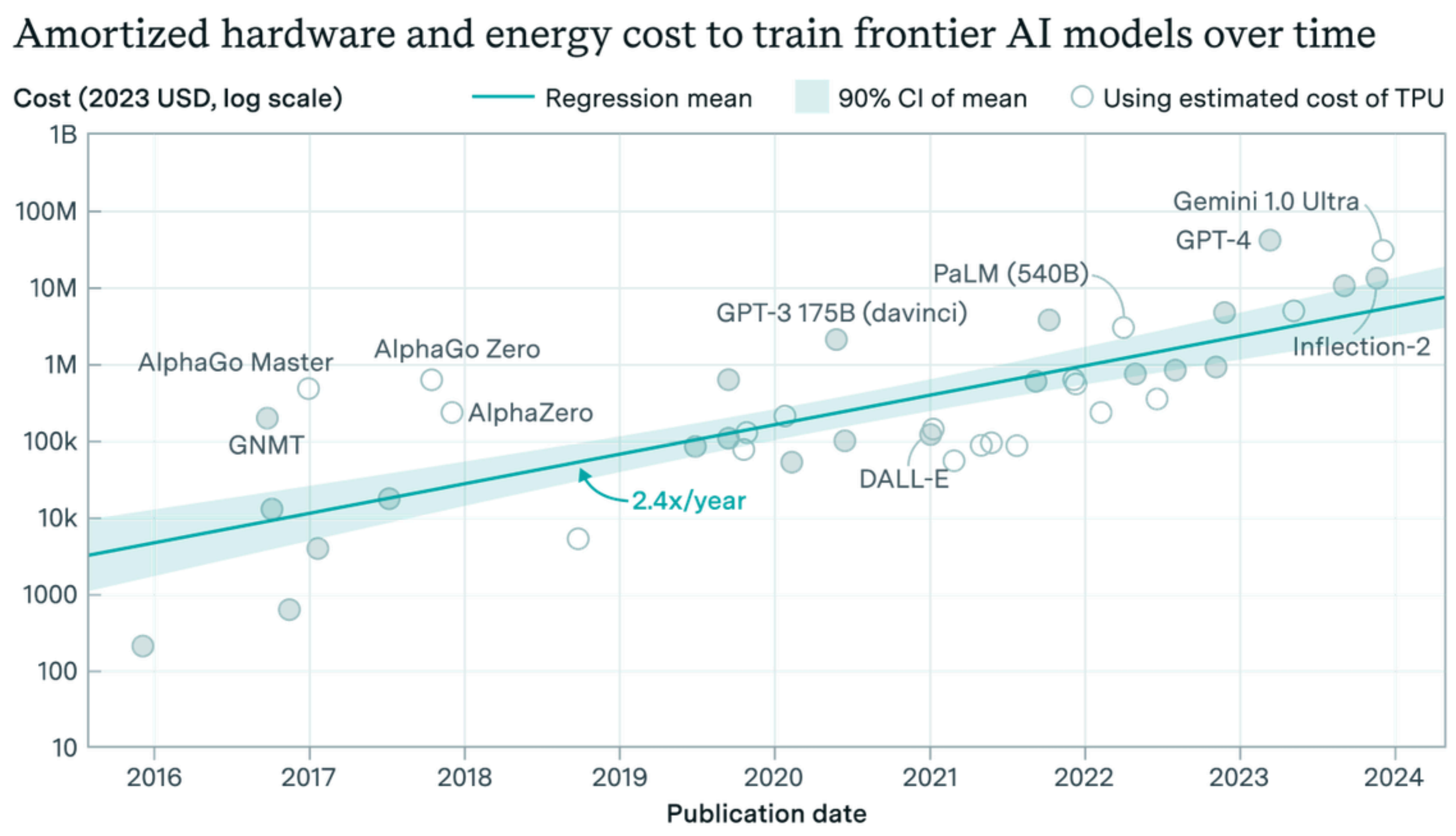




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls

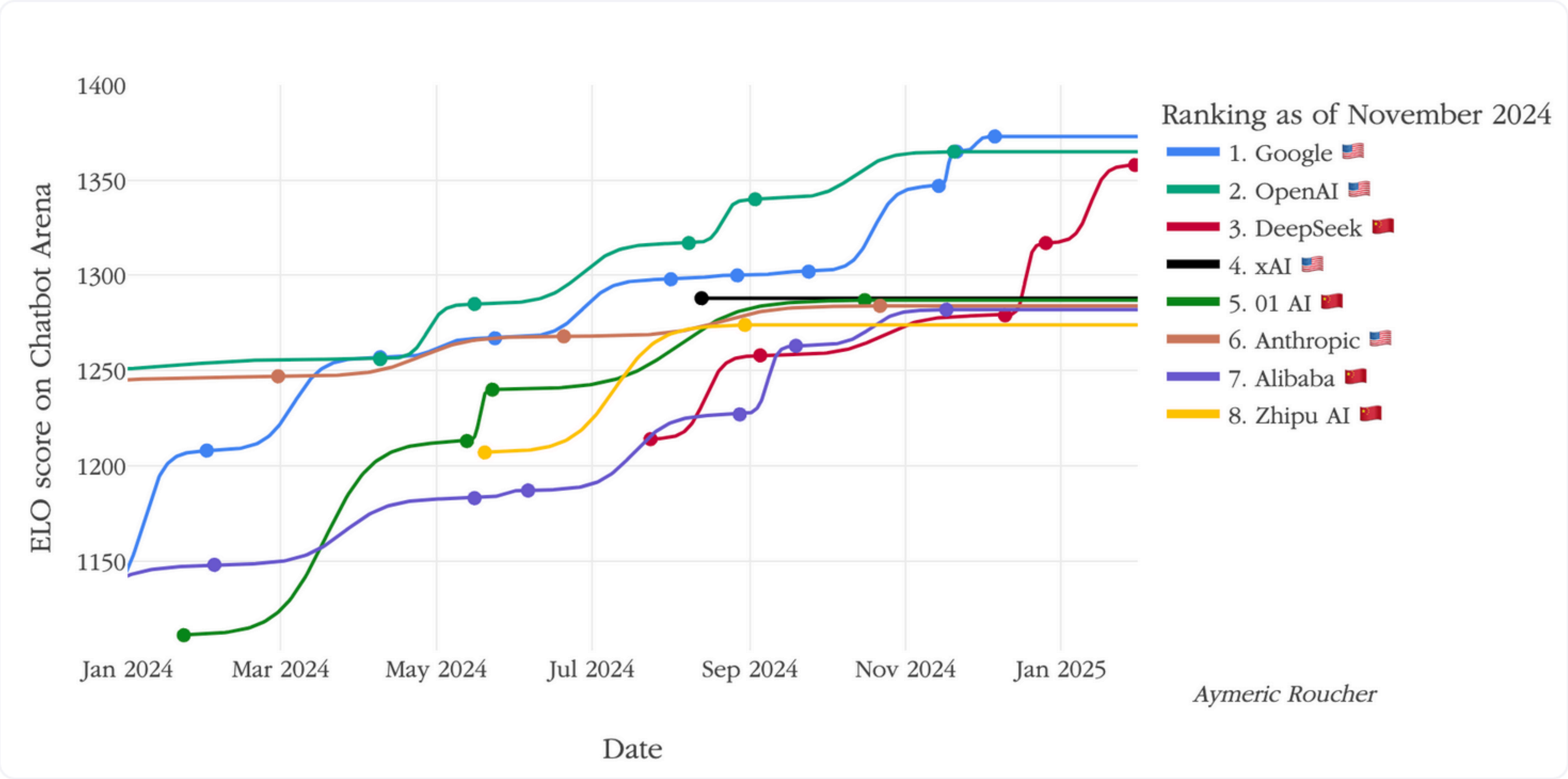
The case of LLMs

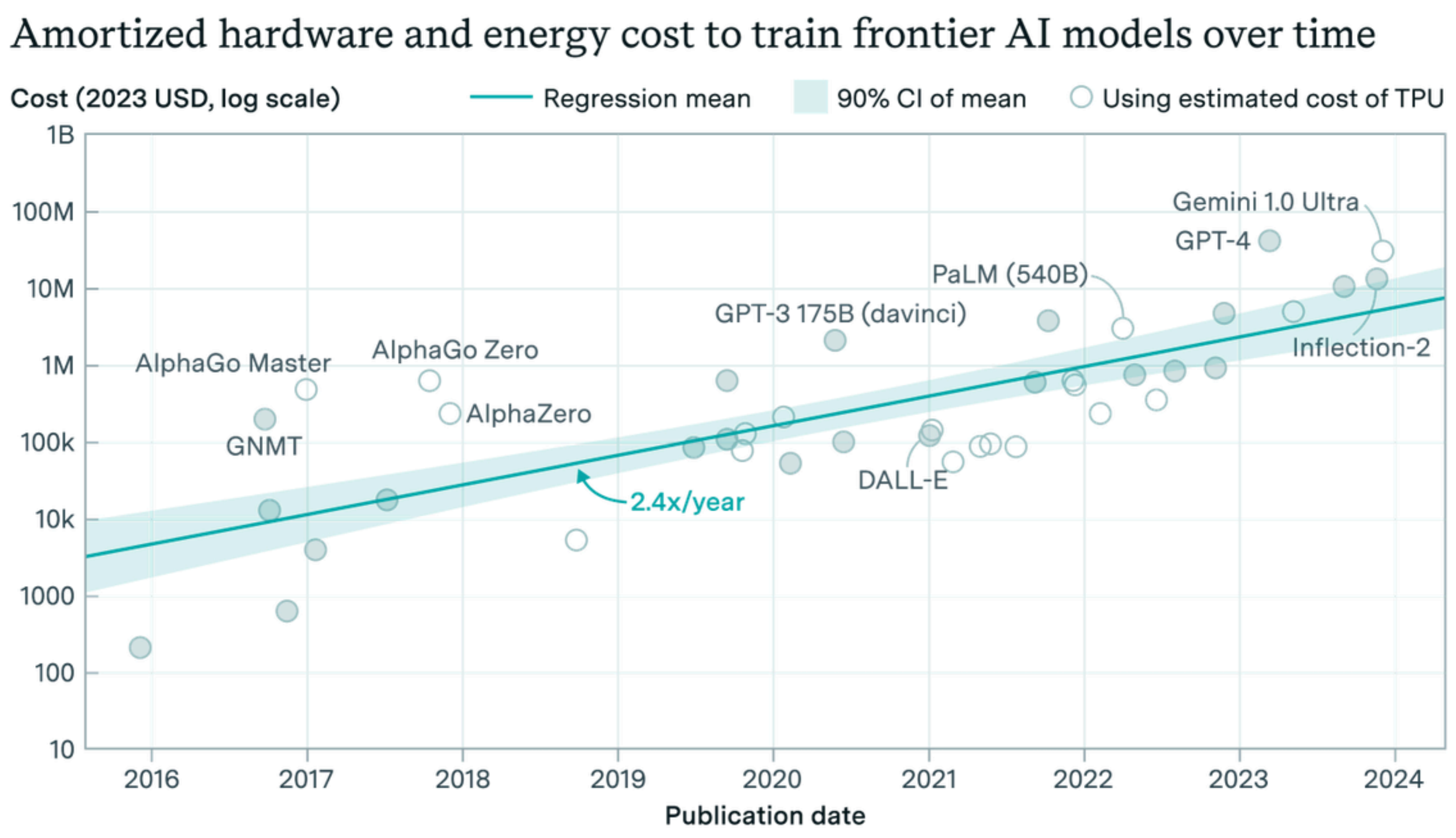




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models

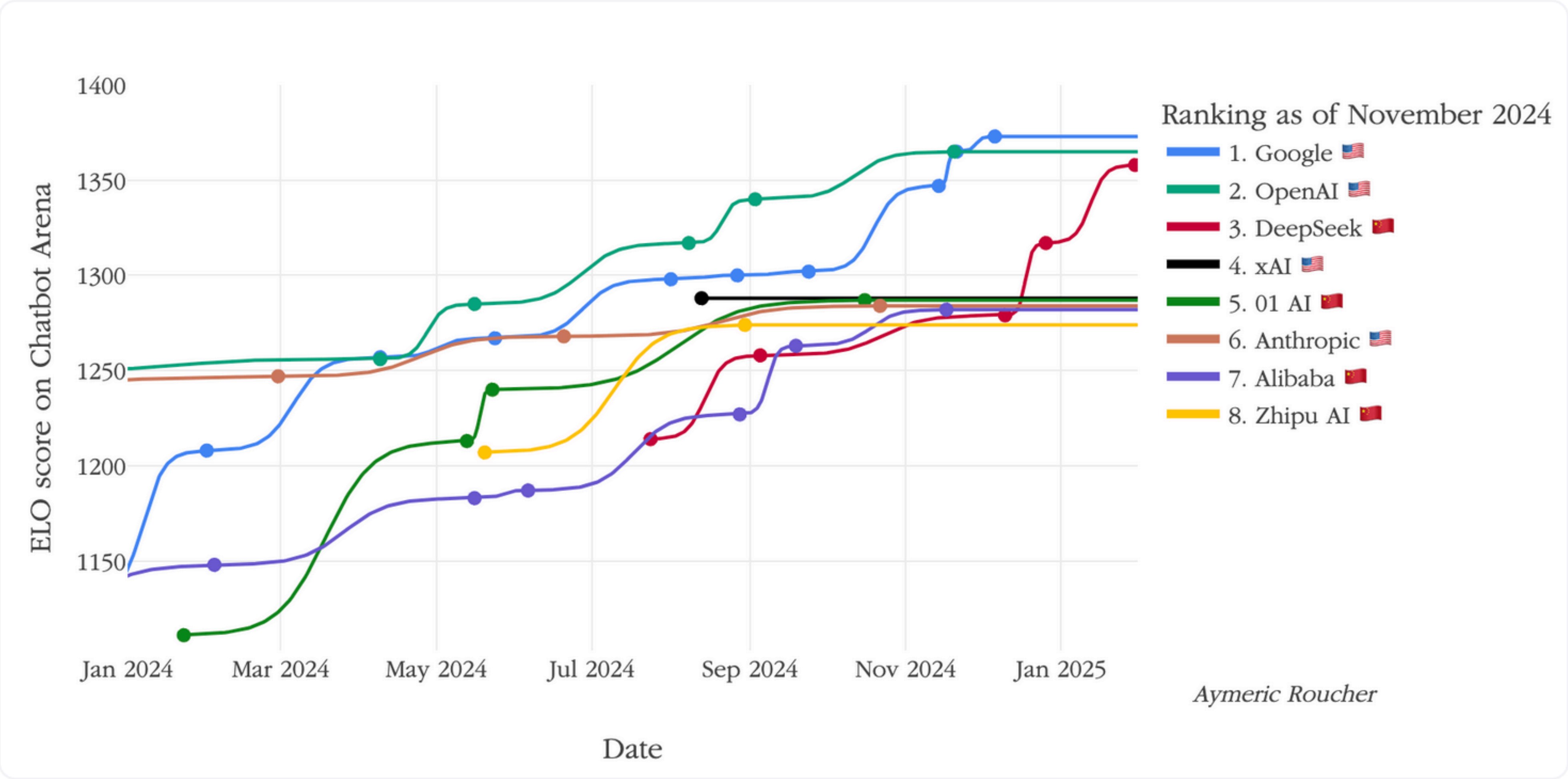
The case of LLMs

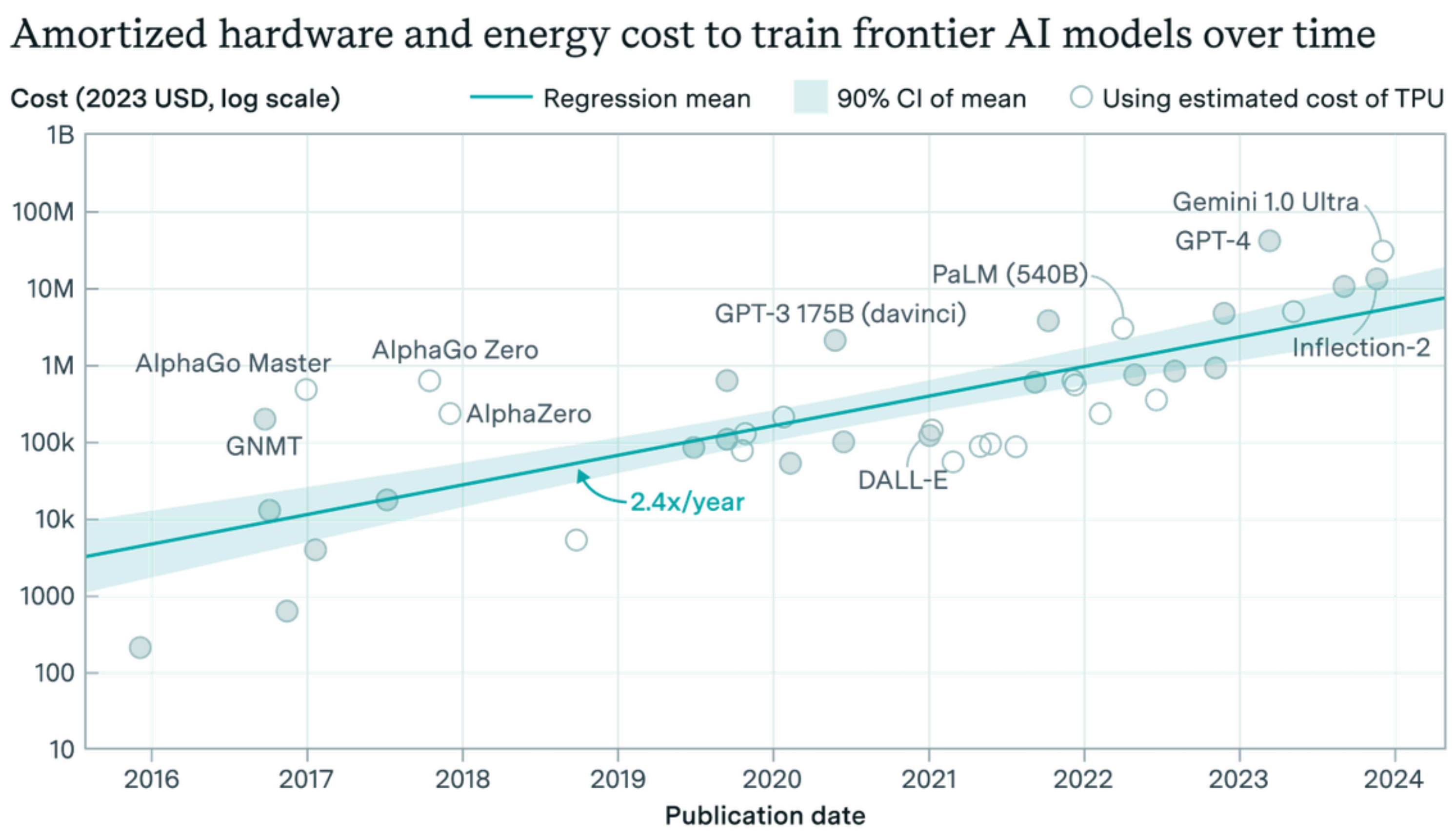




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models
  - Open-weights: Meta, Google, Mistral, …

The case of LLMs





- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models
  - Open-weights: Meta, Google, Mistral, …
  - Fully open models: Stanford, AllenAI institute, Apple …

Collaboration for profit is possible!

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!

- What is this technology? 🤔

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI

An EUV machine, $380 million

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI

An EUV machine, $380 million

It prints features of just a few nanometers; it is the key technology required to build GPU chips

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI
- EUV development costed ~$14-21 billion and involved **many** companies and countries over decades of research

An EUV machine, $380 million

It prints features of just a few nanometers; it is the key technology required to build GPU chips

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI
- EUV development costed ~$14-21 billion and involved **many** companies and countries over decades of research
- Currently built only by ASML, an EU company 🇪🇺🇳🇱🌷

An EUV machine, $380 million

It prints features of just a few nanometers; it is the key technology required to build GPU chips

Collaboration for profit is possible!

- AI rely on high-performance GPUs which relies on one of the biggest industrial collaboration in history!
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core** technology powering AI
- EUV development costed ~$14-21 billion and involved **many** companies and countries over decades of research
- Currently built only by ASML, an EU company 🇪🇺🇳🇱🌷
- Key geopolitical stake

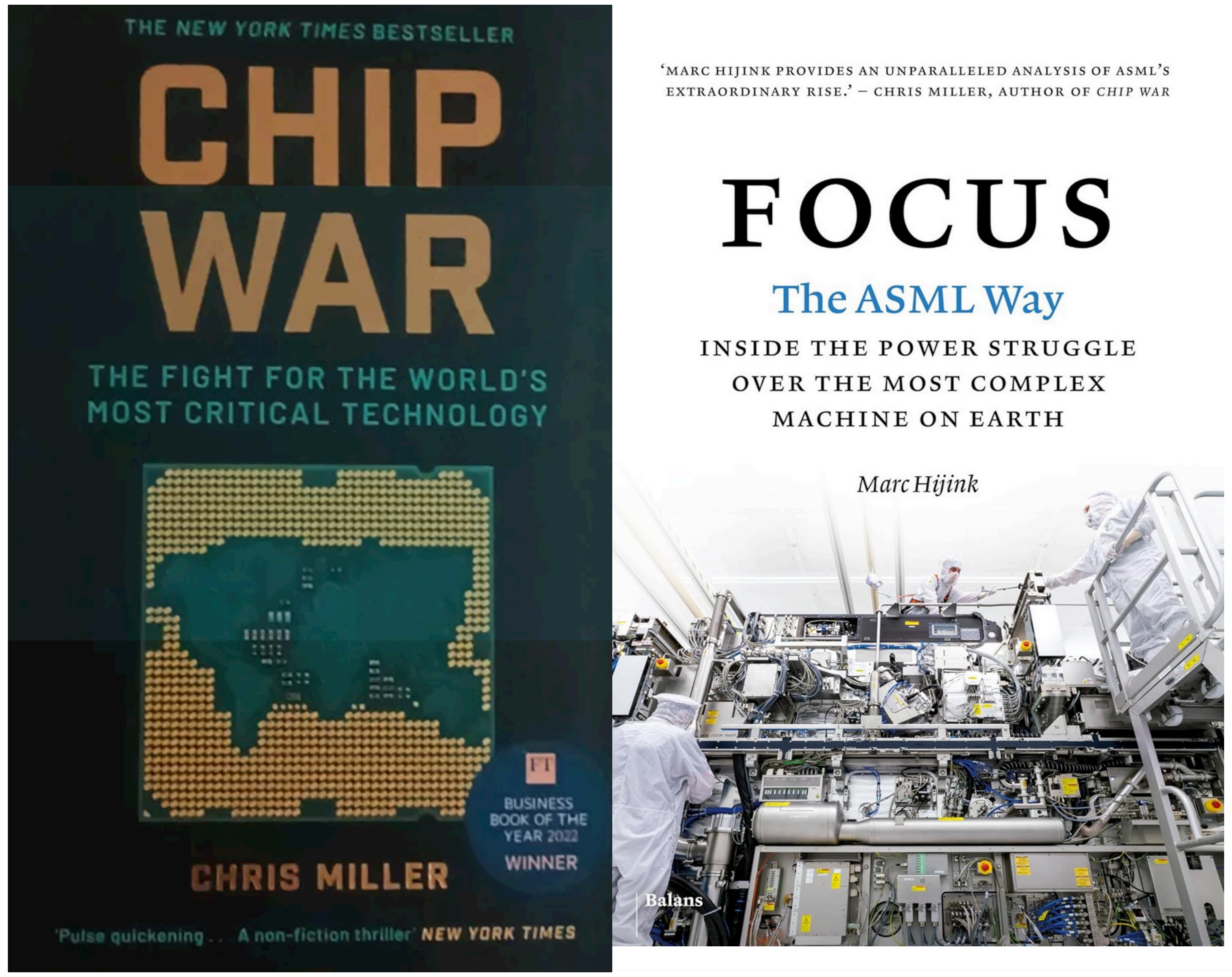
An EUV machine, $380 million

It prints features of just a few nanometers; it is the key technology required to build GPU chips

Collaboration for profit is possible!

- AI rely on high-performance GPUs wh
  biggest industrial collaboration in histo
- What is this technology? 🤔
- Extreme Ultraviolet (EUV) is **the core t**
- EUV development costed ~$14-21 bill
  companies and countries over decade
- Currently built only by ASML, an EU c
- Key geopolitical stake

Recommended reading 📚



An EUV machine, $380 million

It prints features of just a few nanometers; it is the key technology required to build GPU chips

OpenEuroLLM

Universities and Research Organizations

Companies

Co-funded by
the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours in EuroHPC

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models

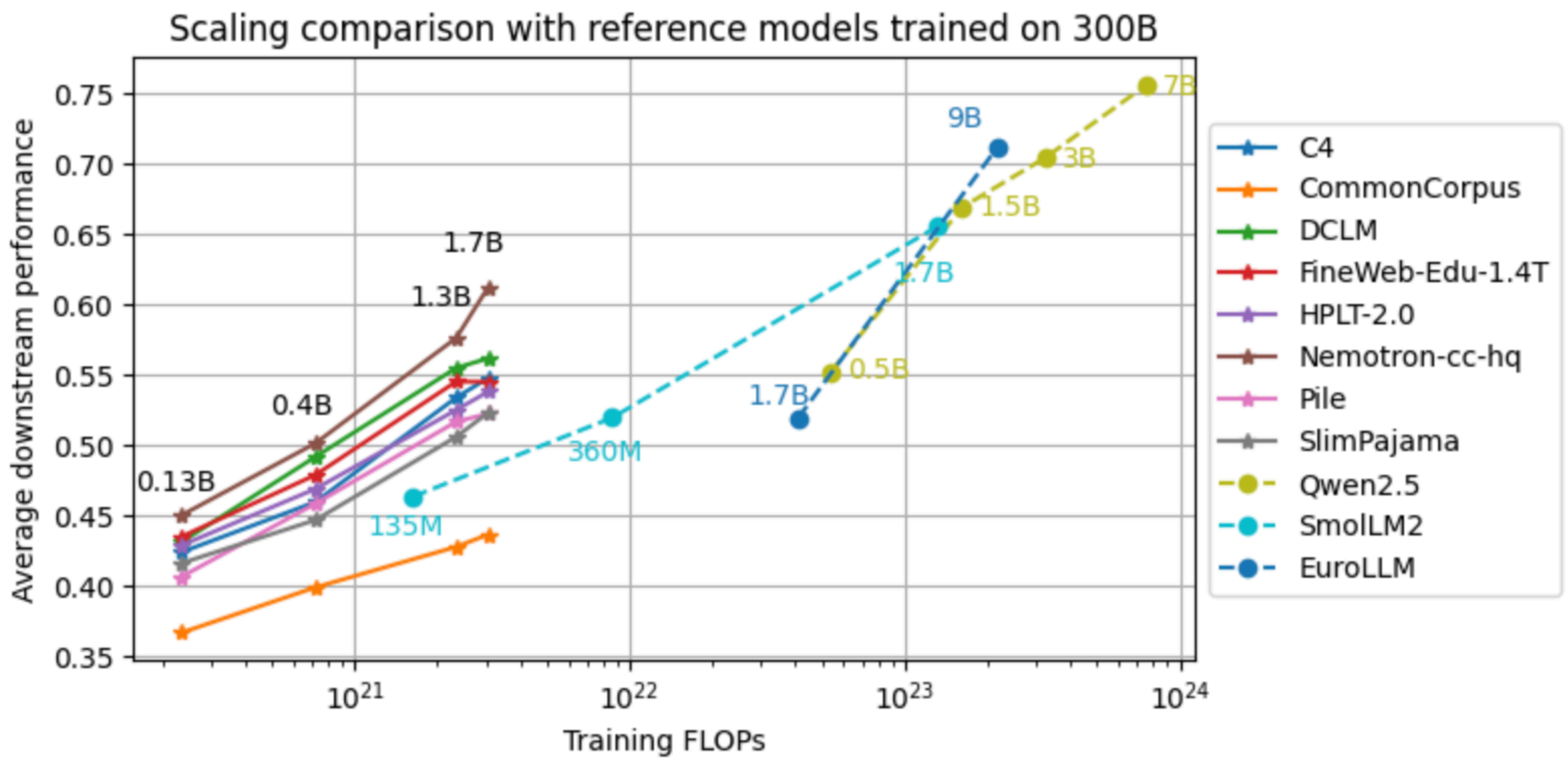
Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models

Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋

Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Charles University
AI SWEDEN
alt-edic
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
ellis INSTITUTE TÜBINGEN
Fraunhofer IAIS
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
JÜLICH Forschungszentrum
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF HELSINKI
UNIVERSITY OF TURKU

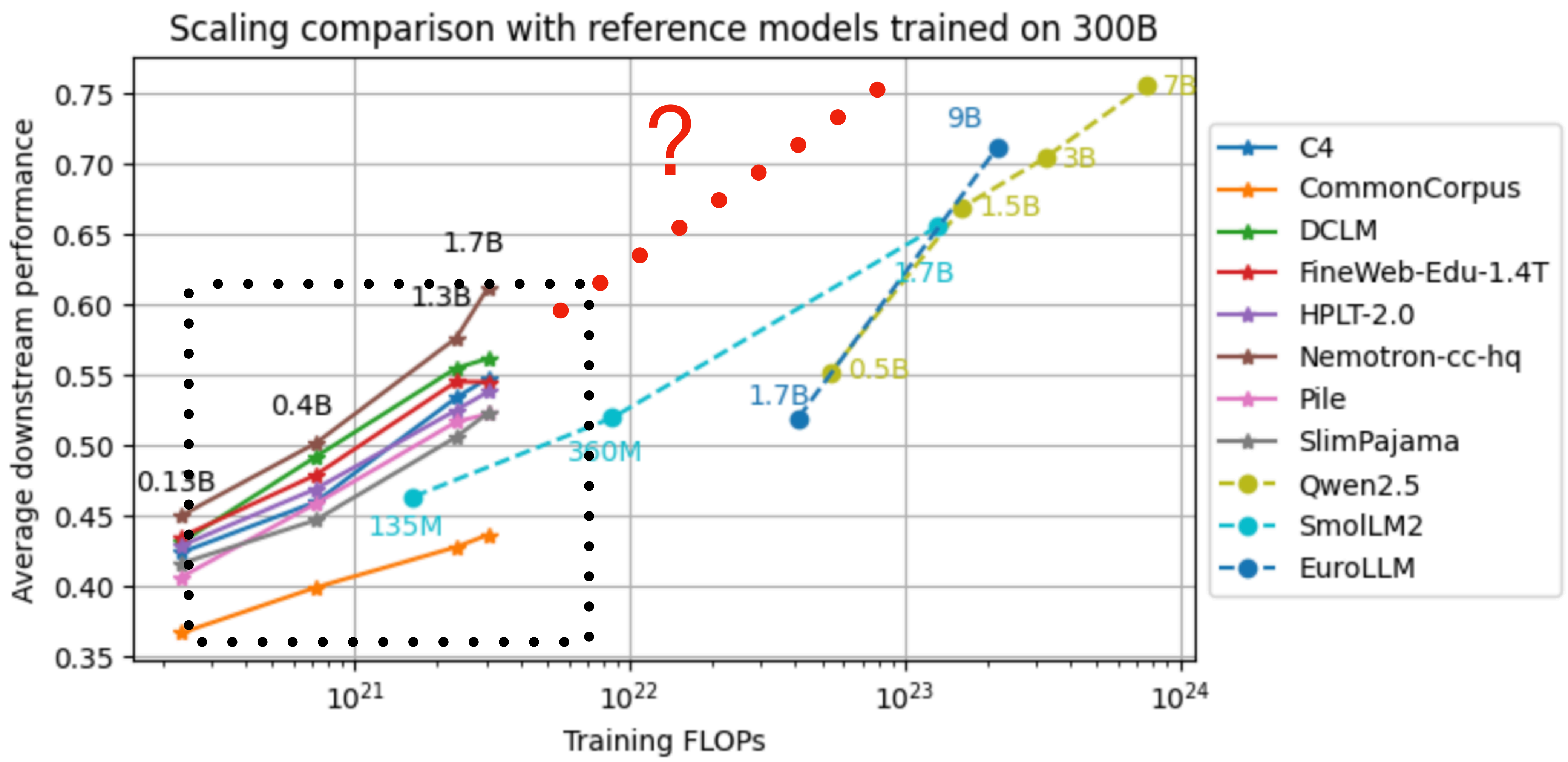
Companies

ALEPH ALPHA
AMD SILO AI
ellamind
LightOn
prompsit

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋
- Ping me and Aaron if interested 🤗

Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours in EuroHPC
- Will release soon reference 1.7B models with SOTA performance among fully open models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋
- Ping me and Aaron if interested 🤗
- Lots of areas for AutoML in pre-training, post-training, evaluation 🎉

Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Companies

Co-funded by the European Union

Conclusion

Conclusion

- LLM evaluation is an open problem
  - Fundamental difficulty to evaluate open-ended answers
  - Need to handle many dimensions (objectives, languages, cost)

Conclusion

- LLM evaluation is an open problem
  - Fundamental difficulty to evaluate open-ended answers
  - Need to handle many dimensions (objectives, languages, cost)
- AutoML has a lot to say
  - Many objectives => multiobjective optimization
  - Costly =>
    - Multifidelity optimization
    - Transfer/meta-learning, portfolio, …

Conclusion

- LLM evaluation is an open problem
    - Fundamental difficulty to evaluate open-ended answers
    - Need to handle many dimensions (objectives, languages, cost)
- AutoML has a lot to say
    - Many objectives => multiobjective optimization
    - Costly =>
        - Multifidelity optimization
        - Transfer/meta-learning, portfolio, …
- AutoML all the way for evaluations (LLM-judge), instruction tuning and maybe pretraining?

Questions