
Cohere Labs
Exploring the unknown, together

Data as Leverage:

Improving Foundation Models Beyond Scaling

Cohere Labs

Daniel D'souza

Agenda

01 Motivation

02 Scaling Plateau

03 **Creating** Better Signal

04 **Revealing** Better Signal

05 Future Takeaways

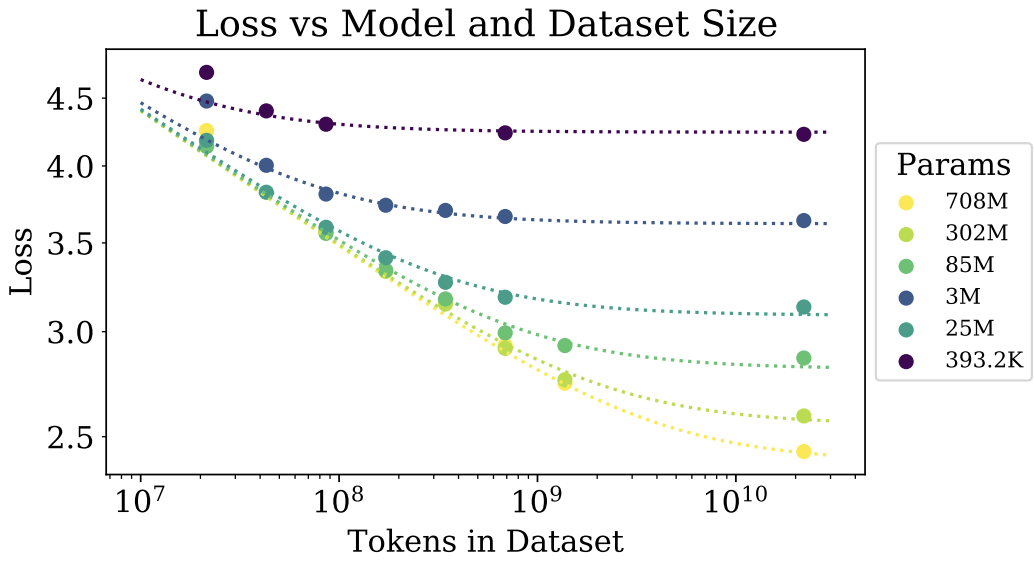
01

Motivation

Motivation

Trends that Scale

- High Frequency Use Cases
- High Resource Languages
- Broader Capabilities



Kaplan et al (https://arxiv.org/abs/2001.08361)

02

Scaling Plateau

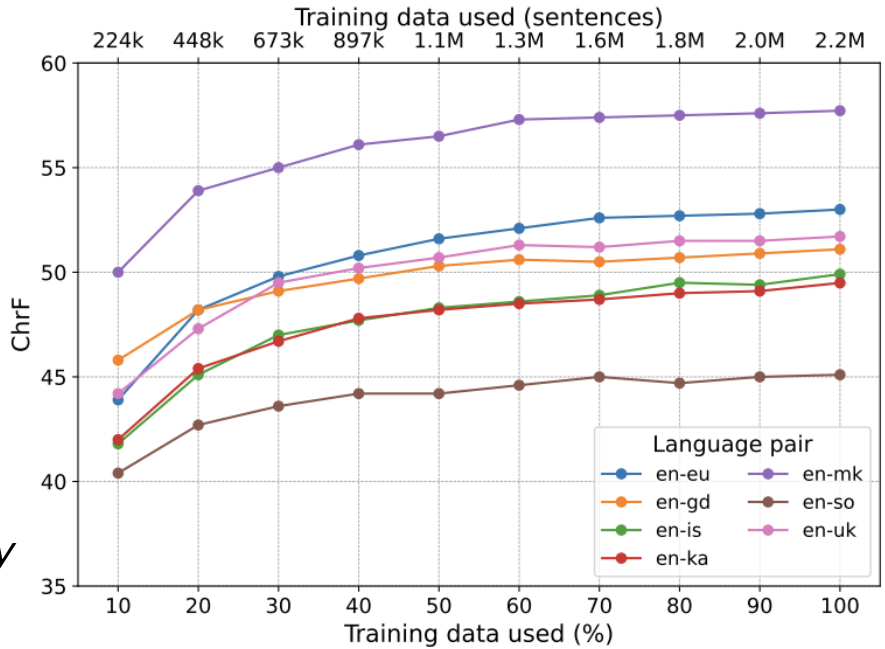
Scaling shifts the curve upward,
but it doesn't change which curve you're on

Where Scaling Plateaus

Is Scale the answer?

- Consistent improvements; trends upward
- Why does the “gap” remain consistent ?

If scale was enough, would these curves ideally converge ?

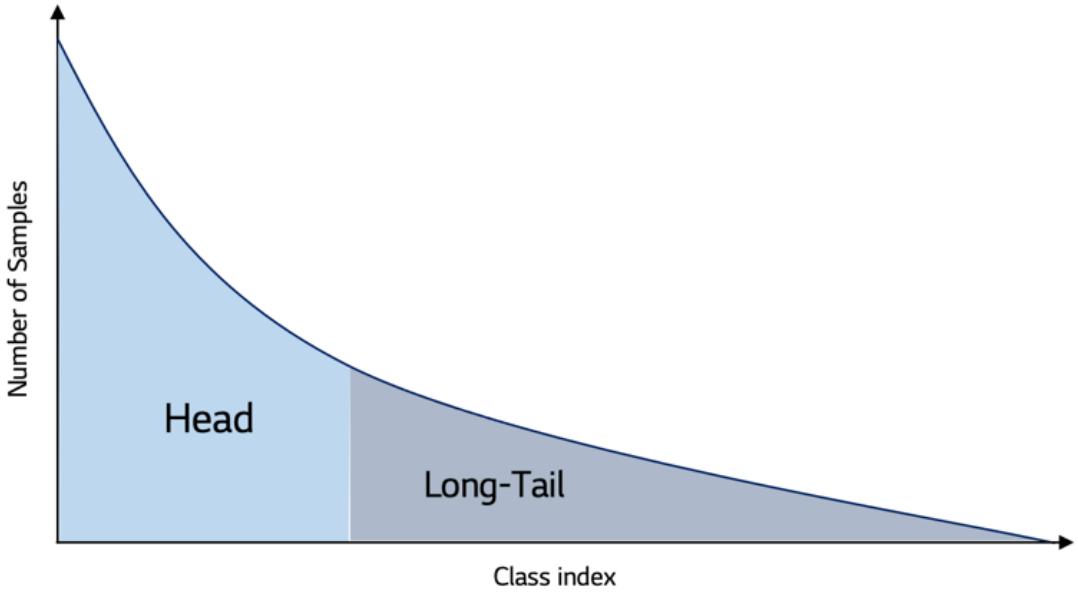


Gilbert et al
(https://arxiv.org/abs/2505.14423v2)

Where Scaling Plateaus

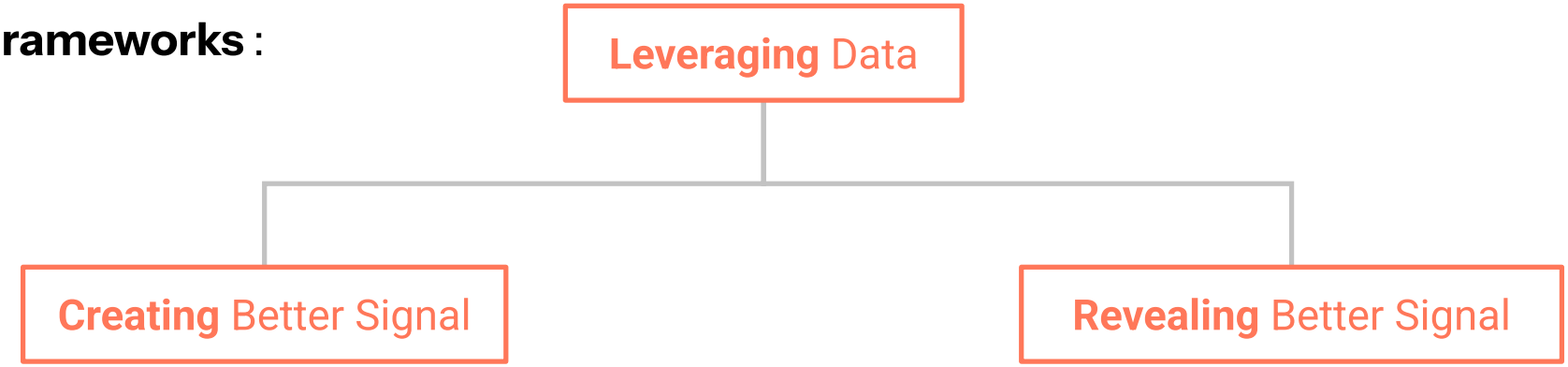
Trends that **don't** really scale well

- Low Frequency Use Cases
- Low Resource Languages
- Specific Targeted Capabilities



Can we do better?

Frameworks :



Multilingual Arbitrage
[EMNLP'25]

Ayomide Odumakinde, Daniel D'souza*, Pat Verga, Beyza Ermis, Sara Hooker*

Treasure Hunt: Real-time Targeting of the Long Tail using Training-Time Markers
[NeurIPS'25]

Daniel D'souza, Julia Kreutzer, Adrien Morisot, Ahmet Üstün, Sara Hooker

03

Creating Better Signal

Why should we learn from just one ?

The Age of Synthetic Data

But is there a universal “best” multilingual model? 🤔

- 💭When was the last time your French Professor was your Chinese Professor?
- 💭When was the last time your Math Professor was your English Professor?

i.e *Why do we need to pick just one teacher ?* 🤔

The Age of Synthetic Data

But is there a universal “best” multilingual model? 🤔

- 💭When was the last time your French Professor was your Chinese Professor?
- 💭When was the last time your Math Professor was your English Professor?

i.e *Why do we need to pick just one teacher ?* 🤔

Multilingual Arbitrage

- There is no universal “best” multilingual model!
- Synthetic Data amplifies weakness i.e *mode collapse*
- Teacher Quality can vary by language, skill, task complexity, etc

i.e *Why not learn about specialized skills from specialized teachers?*

Cohere Labs

Multilingual Arbitrage

Setup

- Treat teachers as a portfolio.
- Route prompts to the best teacher available.
- Optimize data, not models

FROM

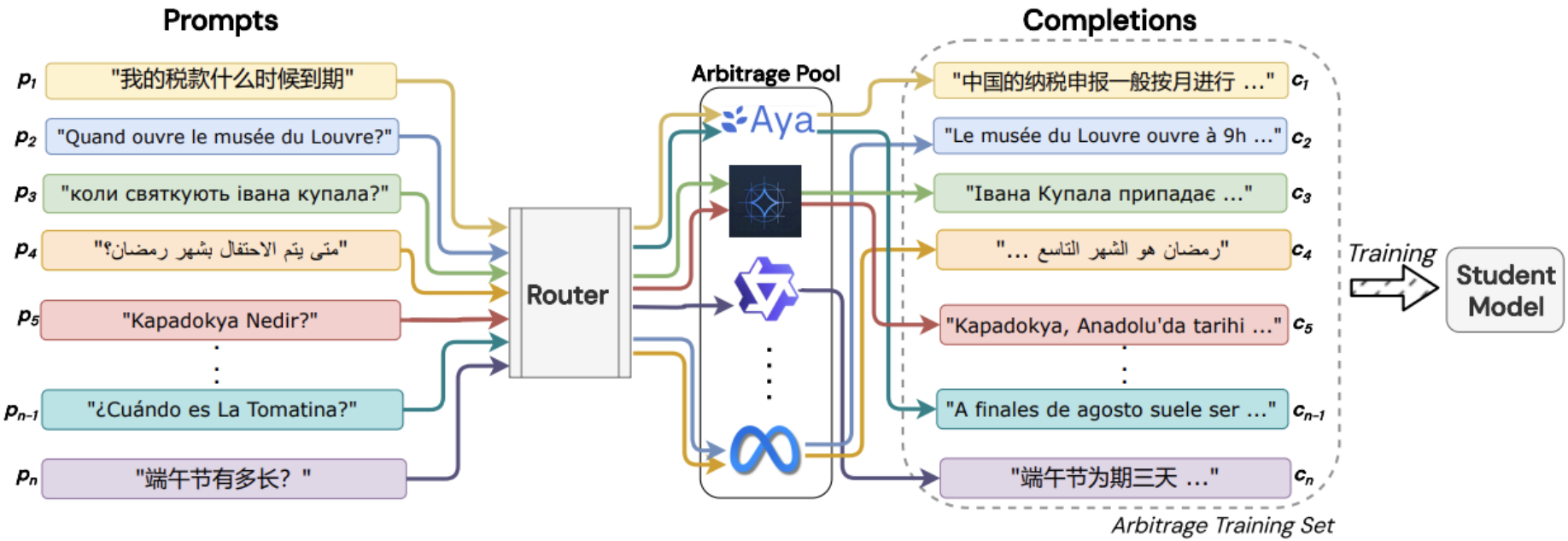
“*Which model is best overall ?*”

TO

“*Which model is best for this prompt?* “

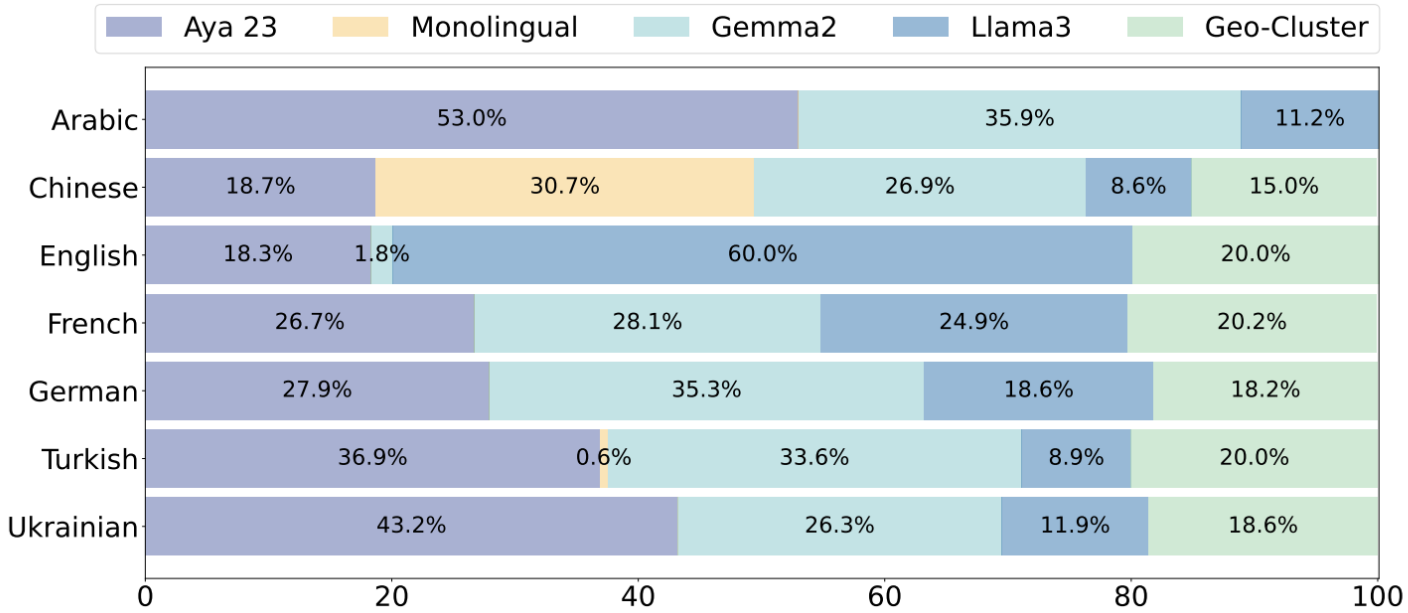
Multilingual Arbitrage

How does Routing work ?



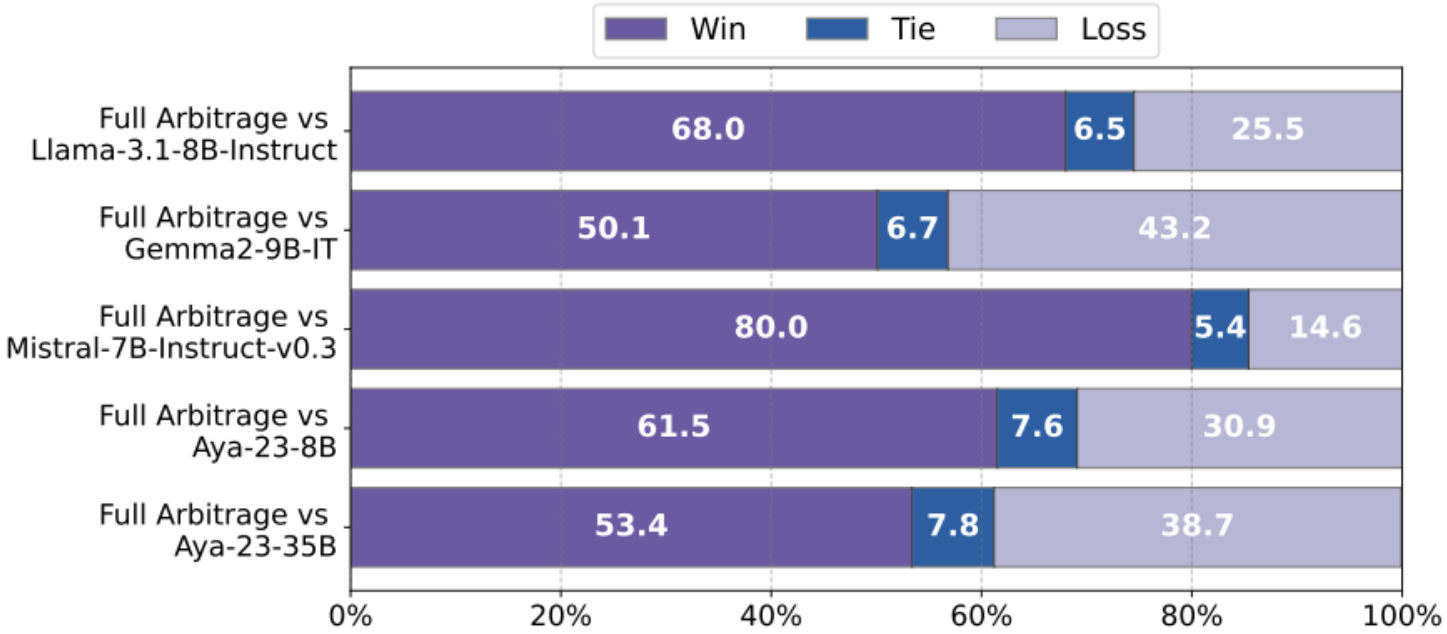
Multilingual Arbitrage

Compositions



Multilingual Arbitrage

Results



04

Revealing Better Signal

What if we we could learn a richer representation from existing data?

Why make learning implicit?

What else do we know ?

Prompt

```
Give me a one-liner law joke in Spanish.
```

Completion

```
¿Por qué los jueces no van al gimnasio?

Porque ya tienen suficiente poder en sus manos.
```

We actually know :

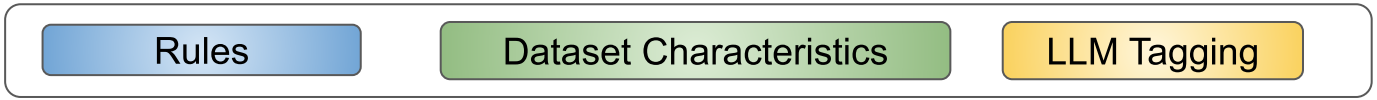
- The completion is in Spanish
- It's in the legal domain
- It's in the humour category
- It's 2 lines long

Defining the Taxonomy

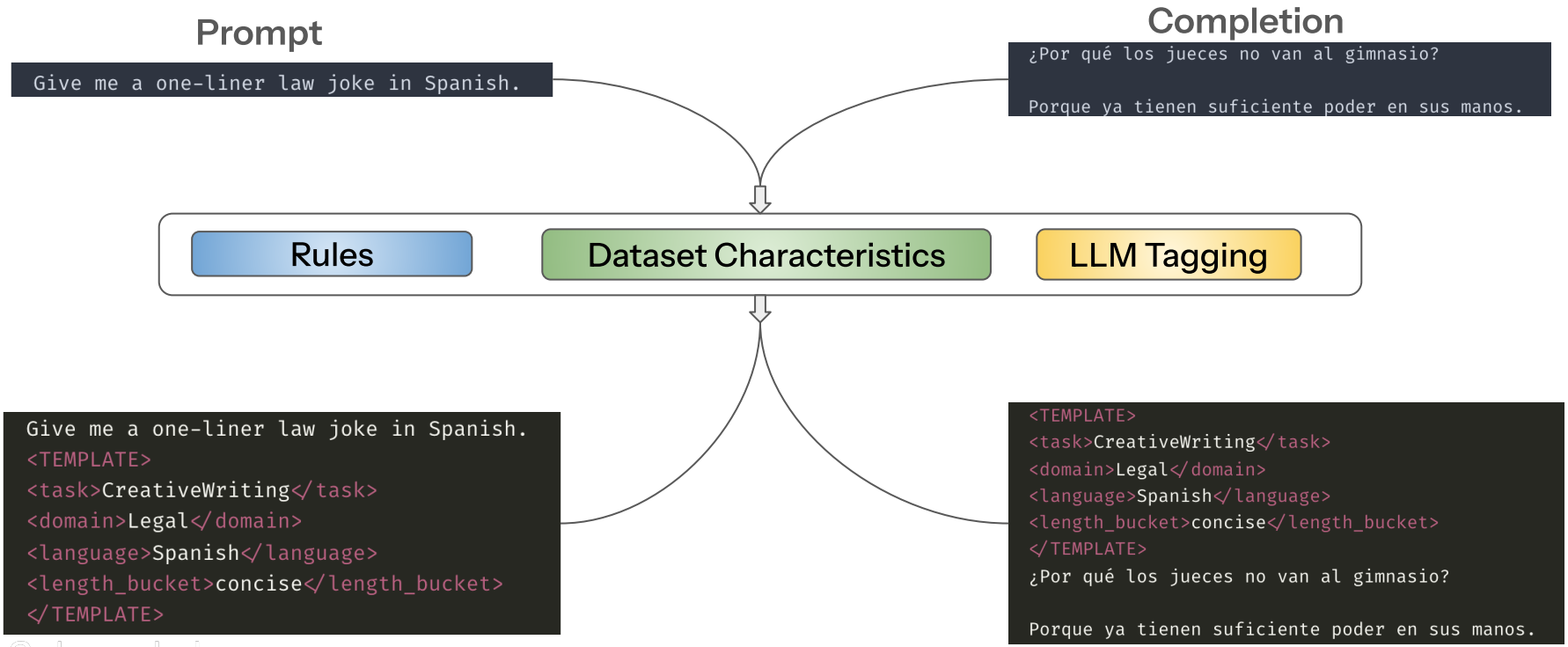
Taxonomy [13 markers → 90 possible values]

```
    <task>       ⟹ CodeGeneration, CodeTranslation, Summarization, ...
  <domain>       ⟹ Code, Math, Science, Legal, ...
<language>       ⟹ English, French, German, Spanish, Hindi, ...
<length_bucket>  ⟹ concise, medium, long
  < ... >        ⟹  ...
```

Annotation



Enriching the Dataset



Marker Dropout

```
Give me a one-liner law joke in Spanish.
<TEMPLATE>
<domain>Legal</domain>
</TEMPLATE>
```

```
<TEMPLATE>
<task>CreativeWriting</task>
<domain>Legal</domain>
<language>Spanish</language>
<length_bucket>concise</length_bucket>
</TEMPLATE>
¿Por qué los jueces no van al gimnasio?

Porque ya tienen suficiente poder en sus manos.
```

```
Give me a one-liner law joke in Spanish.
<TEMPLATE>
<task>CreativeWriting</task>
</TEMPLATE>
```

```
<TEMPLATE>
<task>CreativeWriting</task>
<domain>Legal</domain>
<language>Spanish</language>
<length_bucket>concise</length_bucket>
</TEMPLATE>
¿Por qué los jueces no van al gimnasio?

Porque ya tienen suficiente poder en sus manos.
```

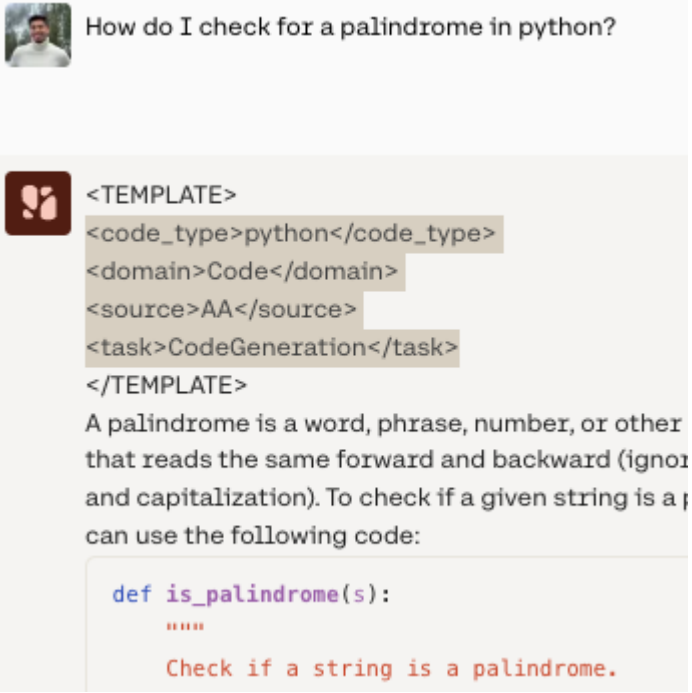
```
Give me a one-liner law joke in Spanish.
<TEMPLATE>
<language>Spanish</language>
<length_bucket>concise</length_bucket>
</TEMPLATE>
```

```
<TEMPLATE>
<task>CreativeWriting</task>
<domain>Legal</domain>
<language>Spanish</language>
<length_bucket>concise</length_bucket>
</TEMPLATE>
¿Por qué los jueces no van al gimnasio?

Porque ya tienen suficiente poder en sus manos.
```


Cohere Labs

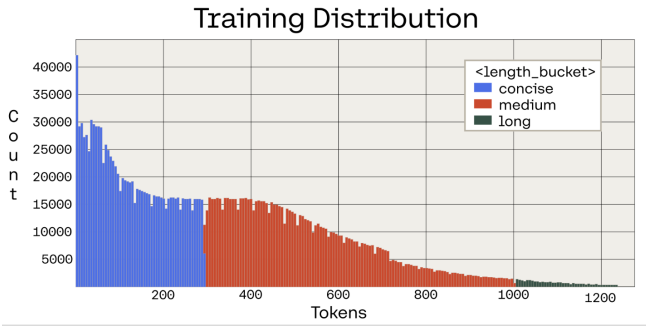
Treasure-marked Model



Cohere Labs

Treasure-marked Model

*"**Controlling**" generations at inference*



CONTROL AT INFERENCE

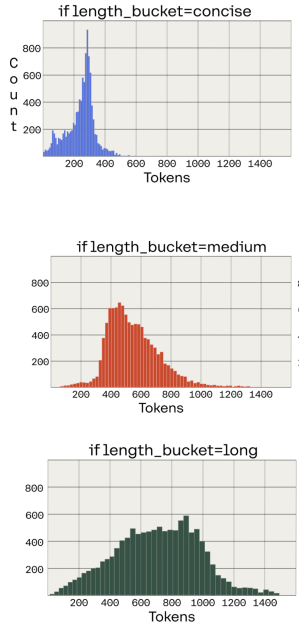
At inference time, **changing a single marker** can lead to **vastly different test-time distributions**

```
Tell me about the French Revolution
<TEMPLATE>
<length_bucket>concise</length_bucket>
</TEMPLATE>
```

```
Tell me about the French Revolution
<TEMPLATE>
<length_bucket>medium</length_bucket>
</TEMPLATE>
```

```
Tell me about the French Revolution
<TEMPLATE>
<length_bucket>long</length_bucket>
</TEMPLATE>
```

Distribution at Inference



Cohere Labs

Treasure Hunt

*"**Controlling**" generations at inference*

Reward Model Annotation Markers

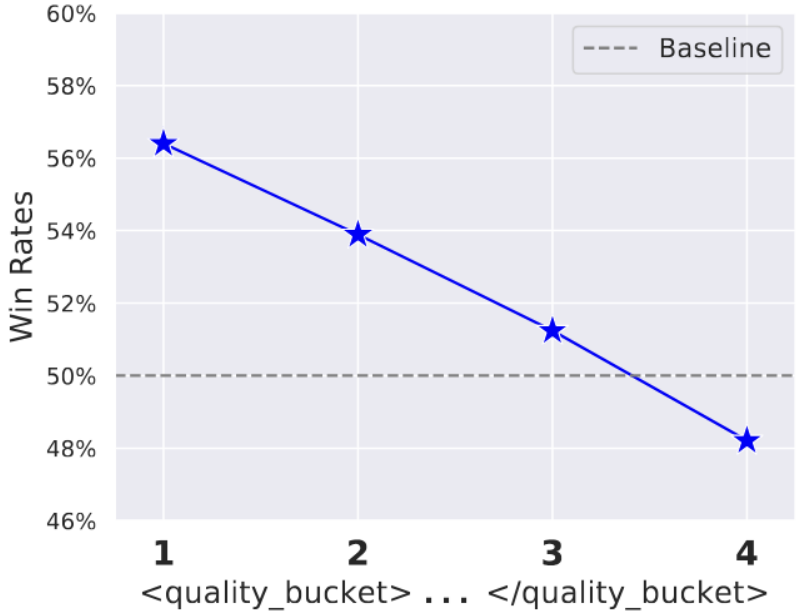
<quality_bucket> 1 ⇒ **Highest** Quality

<quality_bucket> 2 ⇒ ...

<quality_bucket> 3 ⇒ ...

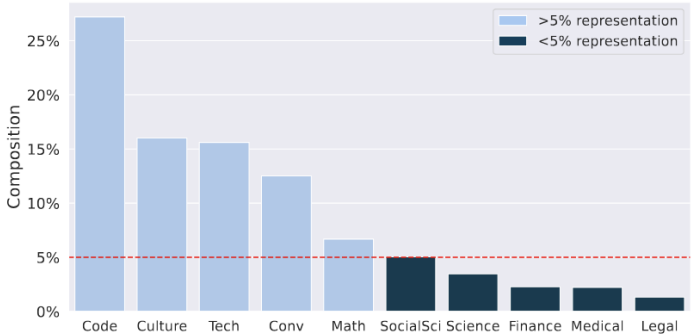
<quality_bucket> 4 ⇒ **Lowest** Quality

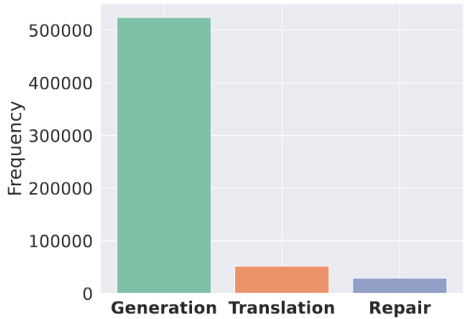
mArenaHard Win Rates against Baseline Model



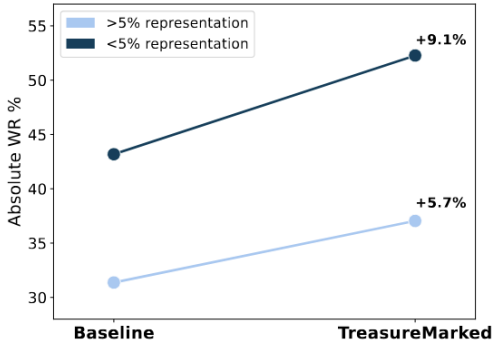
Results

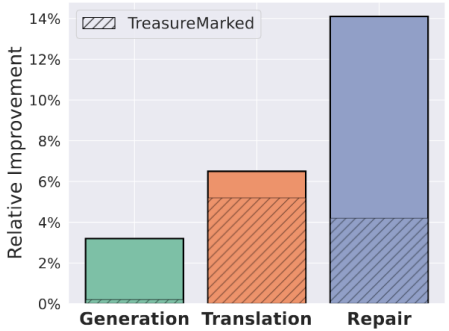
Training Distribution





Improvement at Inference

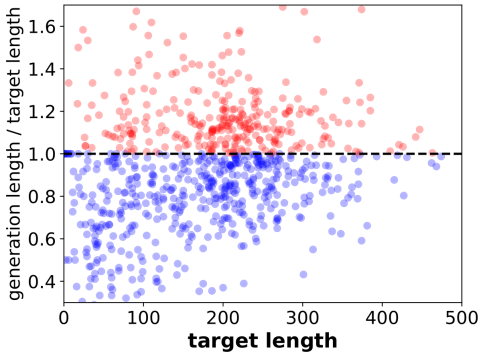


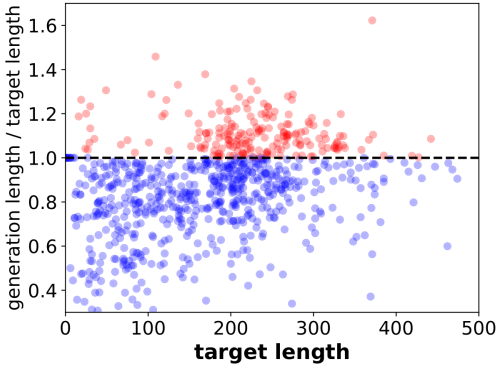


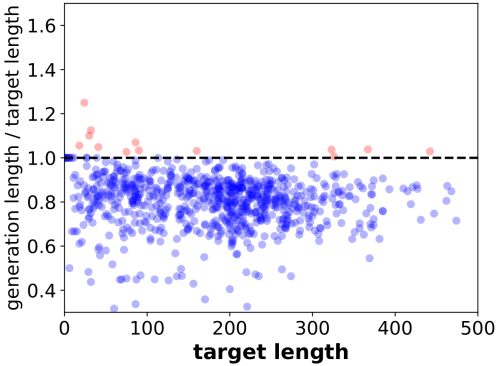
The long tail benefits the most!

Results

Alpaca Eval Length Instruct







	Baseline	**TreasureMarked**	***+fixed***
Violation Rate(%)	36.6 %	24.7 % ⬇️	**1.3** % ⬇️
Win Rate(%)	14.4 %	19.5 % ⬆️	**21.2** % ⬆️

05

Future Takeaways

Use Data as Leverage

- There are significant gains when focus shifts to **data**

- Simple optimizations in the data space are far more achievable than in the model space.

- Open Problems:
    - What are other avenues to generate "***better***" signal ?
    - How do we learn more "***explicit***" signals ?
    - What other "***assumptions***" do we hold about our data ?

Open for Questions!

Correspondence: ddsouza@umich.edu