



Exploring the unknown, together



# Data as Leverage:

*Improving Foundation Models Beyond Scaling*



# Agenda

01

Motivation

02

Scaling  
Plateau

03

**Creating**  
Better Signal

04

**Revealing**  
Better  
Signal

05

Future  
Takeaways

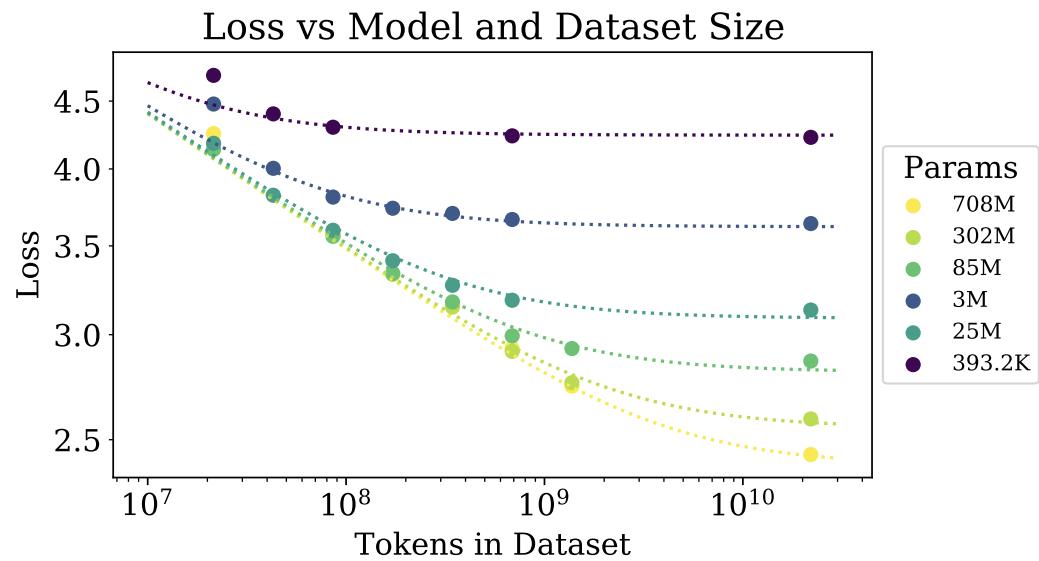
# 01

## 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

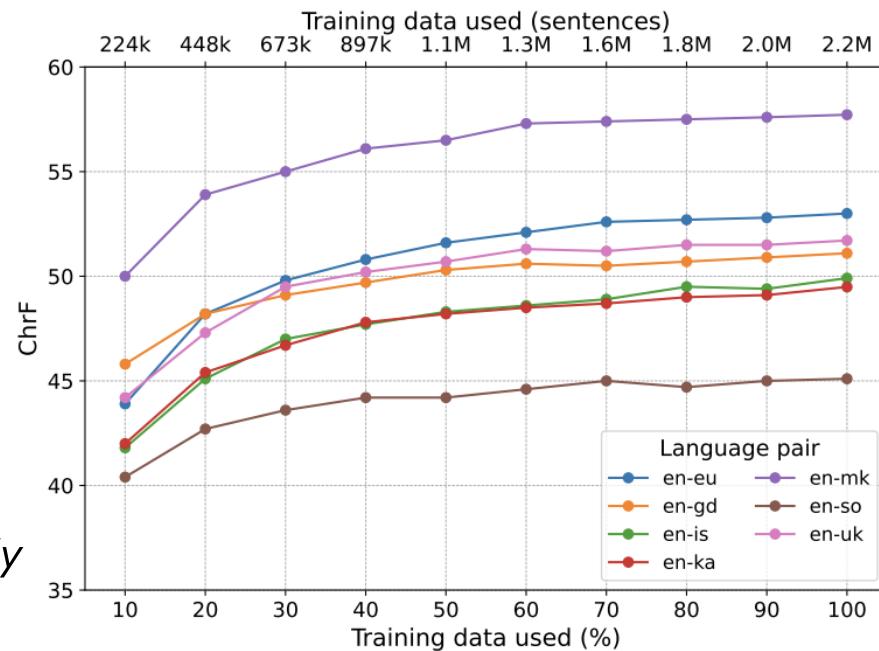
*Scaling shifts the curve upward,  
but it doesn't change which curve you're on*

# 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 ?*

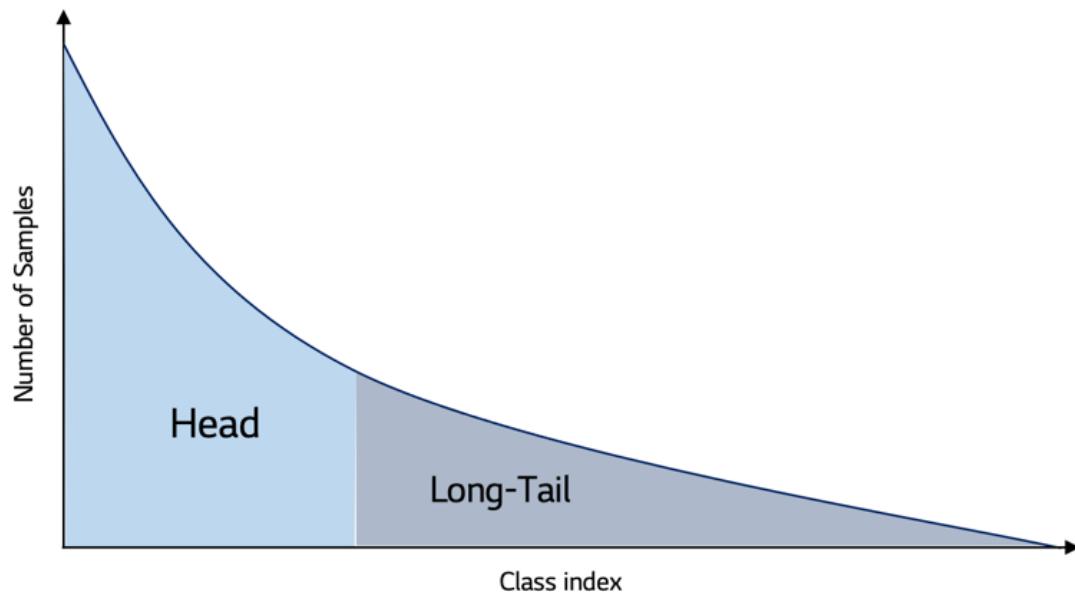


Gilbert et al  
(<https://arxiv.org/abs/2505.14423v2>)

# Where Scaling Plateaus

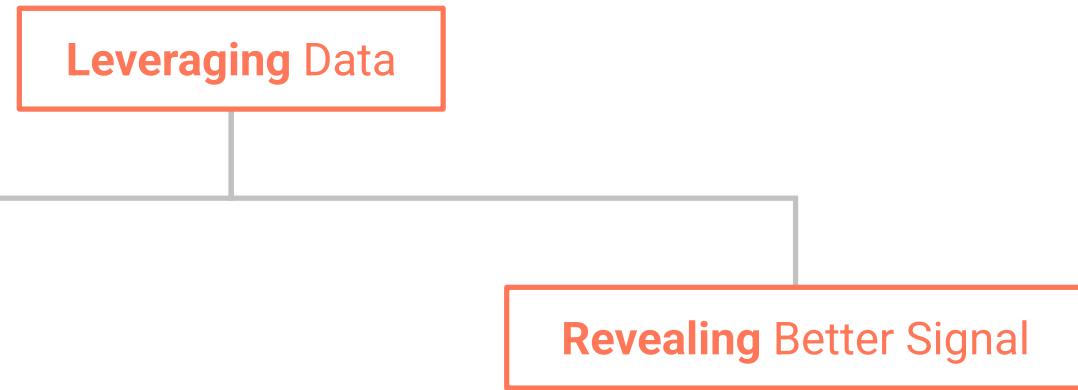
Trends that **don't** really scale well

- 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?*

# 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 ?

## Prompts

$p_1$  "我的税款什么时候到期"

$p_2$  "Quand ouvre le musée du Louvre?"

$p_3$  "коли святкують івана купала?"

$p_4$  "متى يتم الاحتفال بشهر رمضان؟"

$p_5$  "Kapadokya Nedir?"

$\vdots$

$p_{n-1}$  "¿Cuándo es La Tomatina?"

$p_n$  "端午节有多长? "



## Arbitrage Pool



## Completions

"中国的纳税申报一般按月进行 ..."  $c_1$

"Le musée du Louvre ouvre à 9h ..."  $c_2$

"Івана Купала припадає ..."  $c_3$

"رمضان هو الشهر التاسع ..."  $c_4$

"Kapadokya, Anadolu'da tarihi ..."  $c_5$

$\vdots$

"A finales de agosto suele ser ..."  $c_{n-1}$

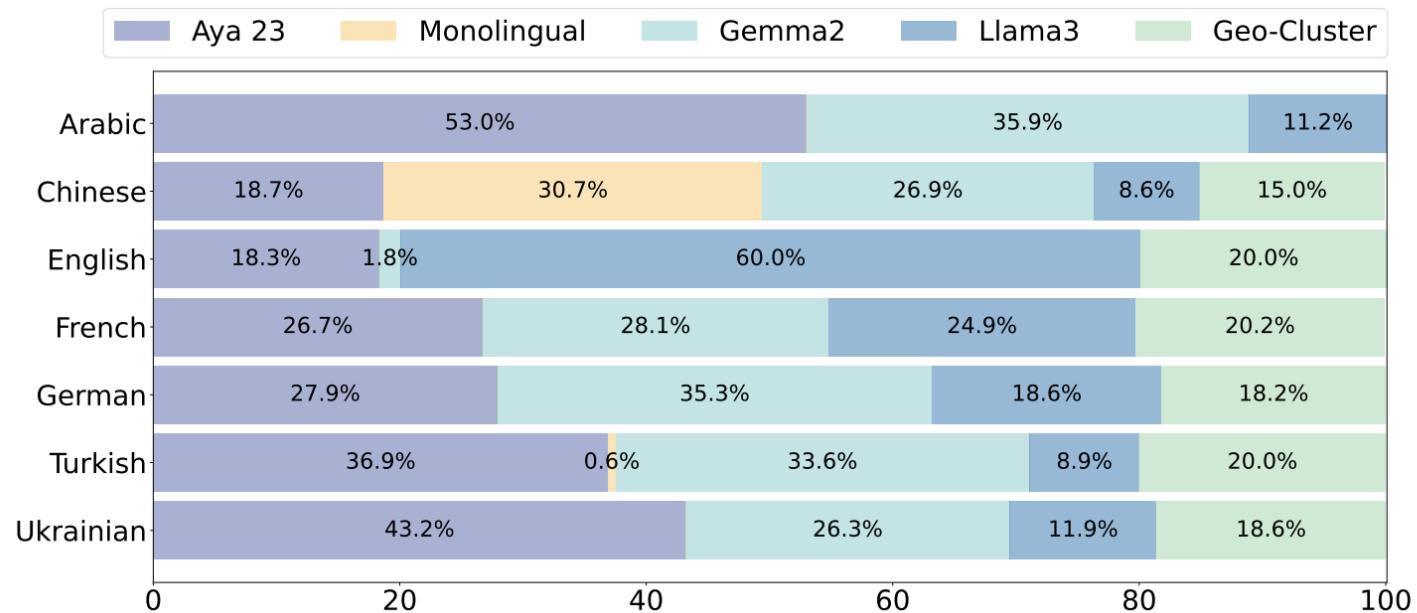
"端午节为期三天 ..."  $c_n$



Arbitrage Training Set

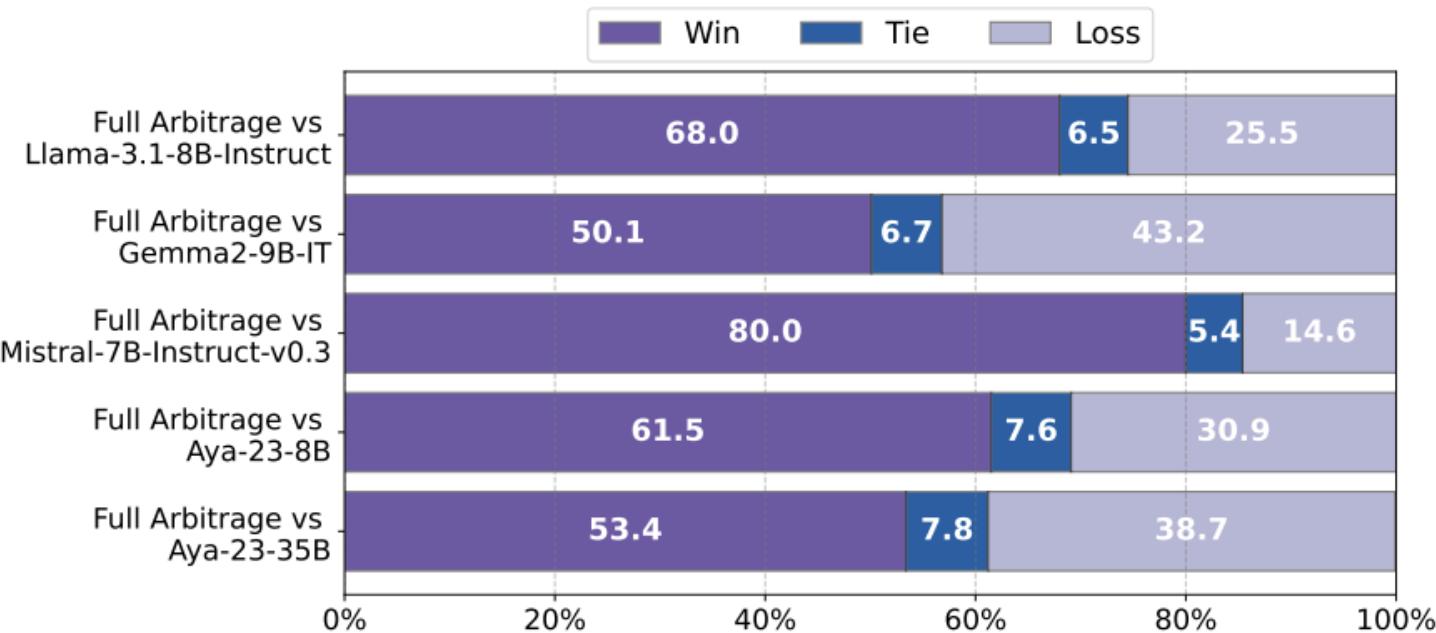
# Multilingual Arbitrage

## Compositions



# Multilingual Arbitrage

## Results



# 04

## Revealing Better Signal

*What if 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
< ... >       ==> ...
```

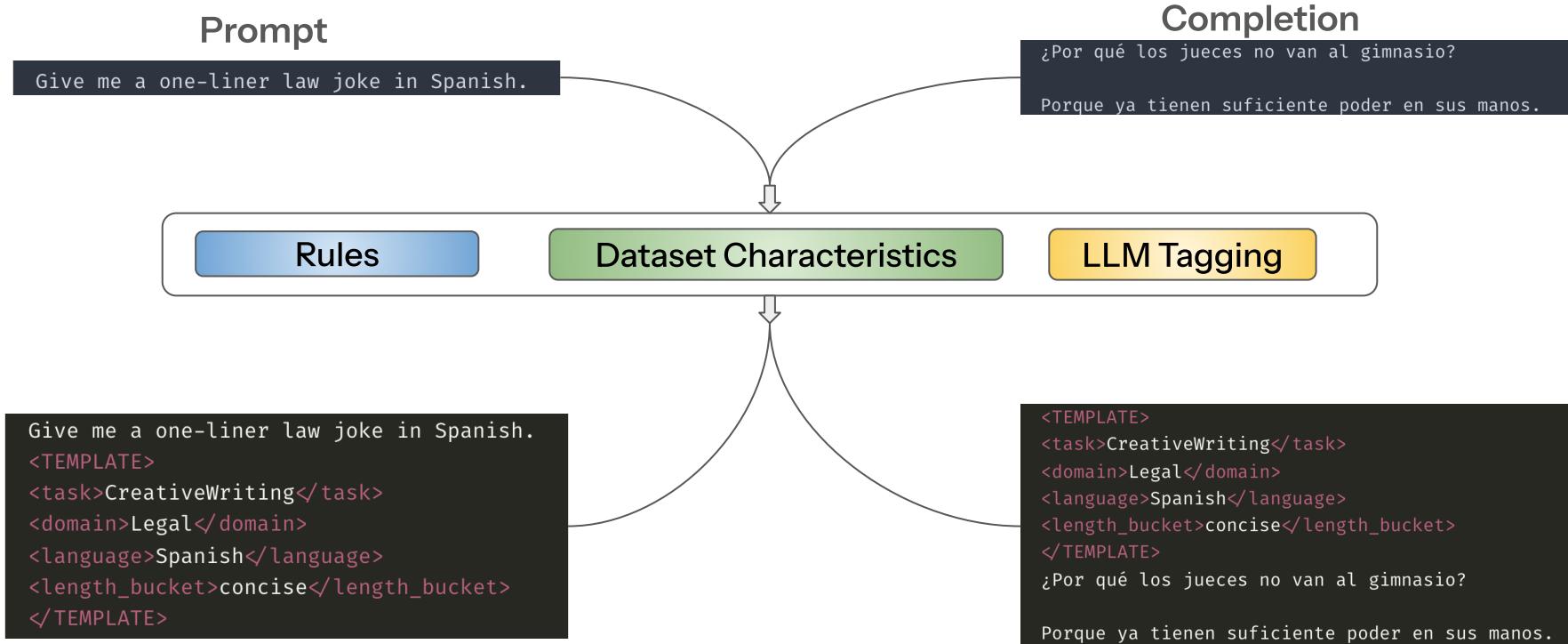
## Annotation

Rules

Dataset Characteristics

LLM Tagging

# 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.

# Treasure-marked Model



How do I check for a palindrome in python?



```
<TEMPLATE>
<code_type>python</code_type>
<domain>Code</domain>
<source>AA</source>
<task>CodeGeneration</task>
</TEMPLATE>
```

A palindrome is a word, phrase, number, or other that reads the same forward and backward (ignoring spaces and capitalization). To check if a given string is a palindrome, you can use the following code:

```
def is_palindrome(s):
    """
    Check if a string is a palindrome.
    
```



Can you give me a 3 sentence paragraph on the French Revolution ?



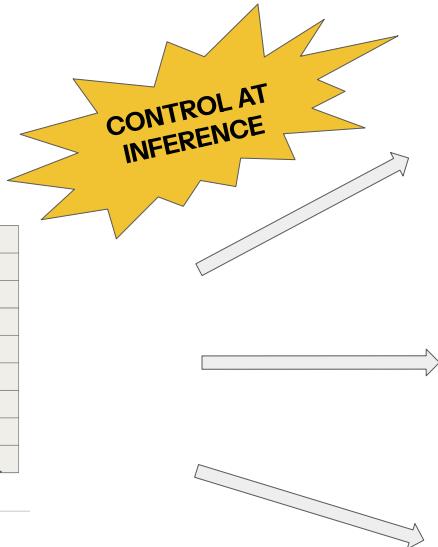
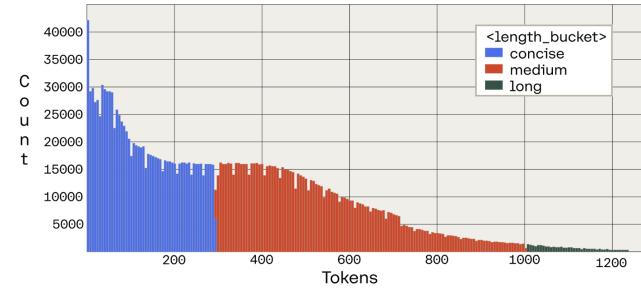
```
<TEMPLATE>
<domain>SocialSciences</domain>
<lang>English</lang>
<length_bucket>concise</length_bucket>
<length_paragraphs>1</length_paragraphs>
<length_sentences>3</length_sentences>
<length_tokens>81</length_tokens>
<quality>4.4</quality>
<quality_bucket>2</quality_bucket>
<source>CC</source>
<task>QuestionAnswering</task>
</TEMPLATE>
```

The French Revolution, which took place between 1789 and 1799, was a period of profound social and political upheaval in France. It marked the end of the French absolute monarchy and the creation of a republic, driven by ideals of liberty, equality, and fraternity. The revolution had a significant impact on European history, spreading the ideas of Enlightenment and leading to the Napoleonic Era and subsequent modernization of European political systems.

# Treasure-marked Model

*“Controlling” generations at inference*

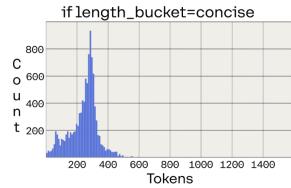
Training Distribution



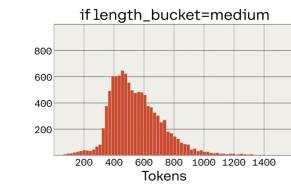
At inference time, **changing a single marker** can lead to **vastly different test-time distributions**

Tell me about the French Revolution  
<TEMPLATE>  
<length\_bucket>concise</length\_bucket>  
</TEMPLATE>

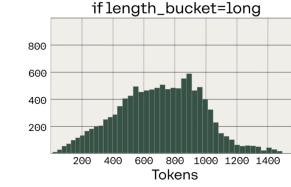
Distribution at Inference



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>



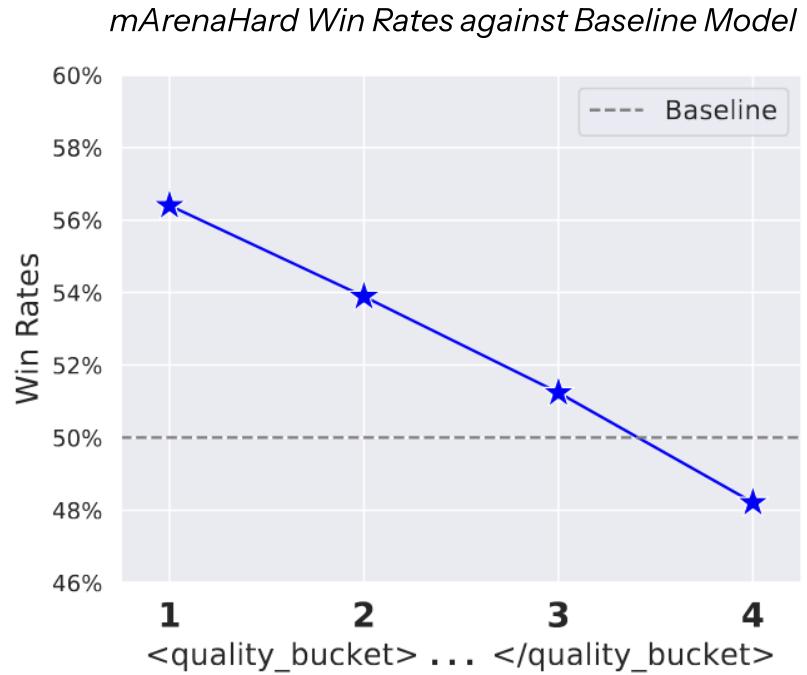
# Treasure Hunt

**“Controlling” generations at inference**



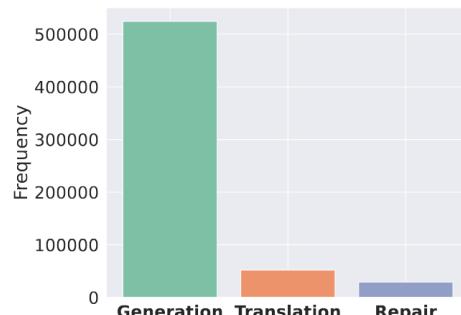
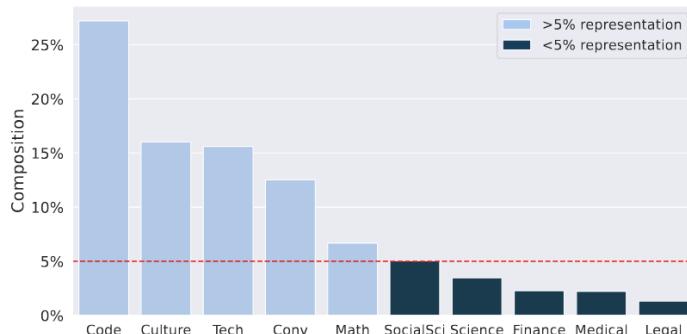
## Reward Model Annotation Markers

<quality\_bucket> 1 ⇒ **Highest** Quality  
<quality\_bucket> 2 ⇒ ...  
<quality\_bucket> 3 ⇒ ...  
<quality\_bucket> 4 ⇒ **Lowest** Quality

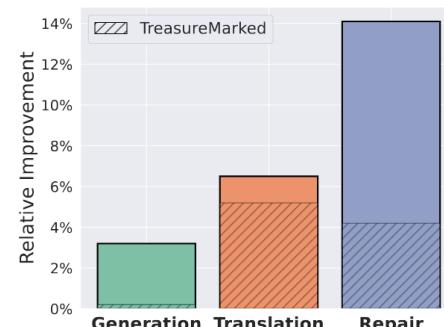
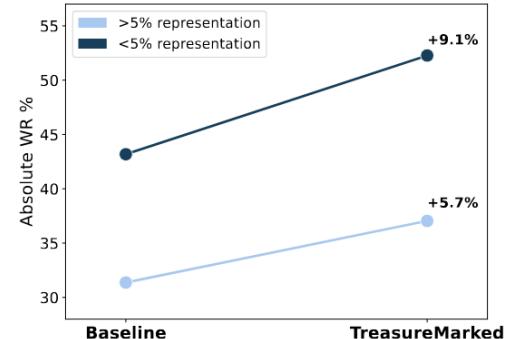


# Results

## Training Distribution



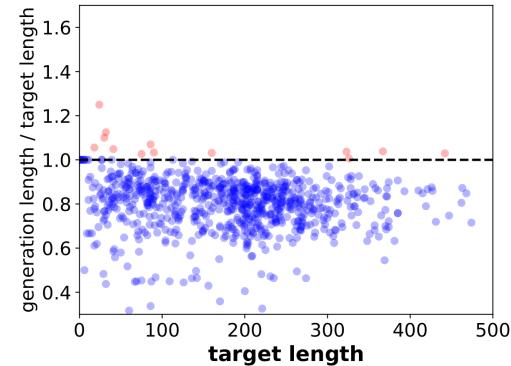
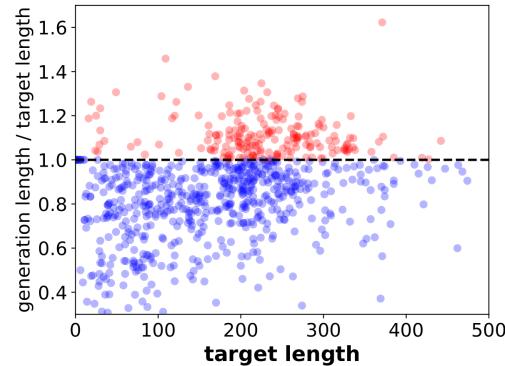
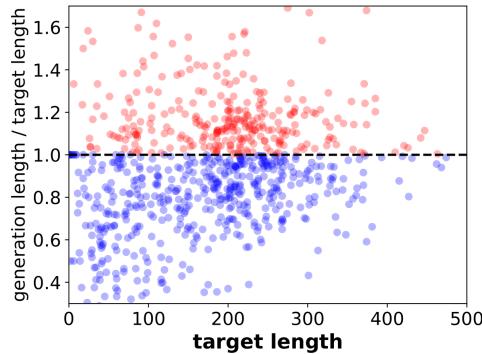
## Improvement at Inference



*The long tail benefits the most!*

# Results

## Alpaca Eval Length Instruct



	<b>Baseline</b>	<b>TreasureMarked</b>	<b>+fixed</b>
<i>Violation Rate(%)</i>	36.6 %	24.7 %	<b>1.3 % </b>
<i>Win Rate(%)</i>	14.4 %	19.5 %	<b>21.2 % </b>

# 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](mailto:ddsouza@umich.edu)