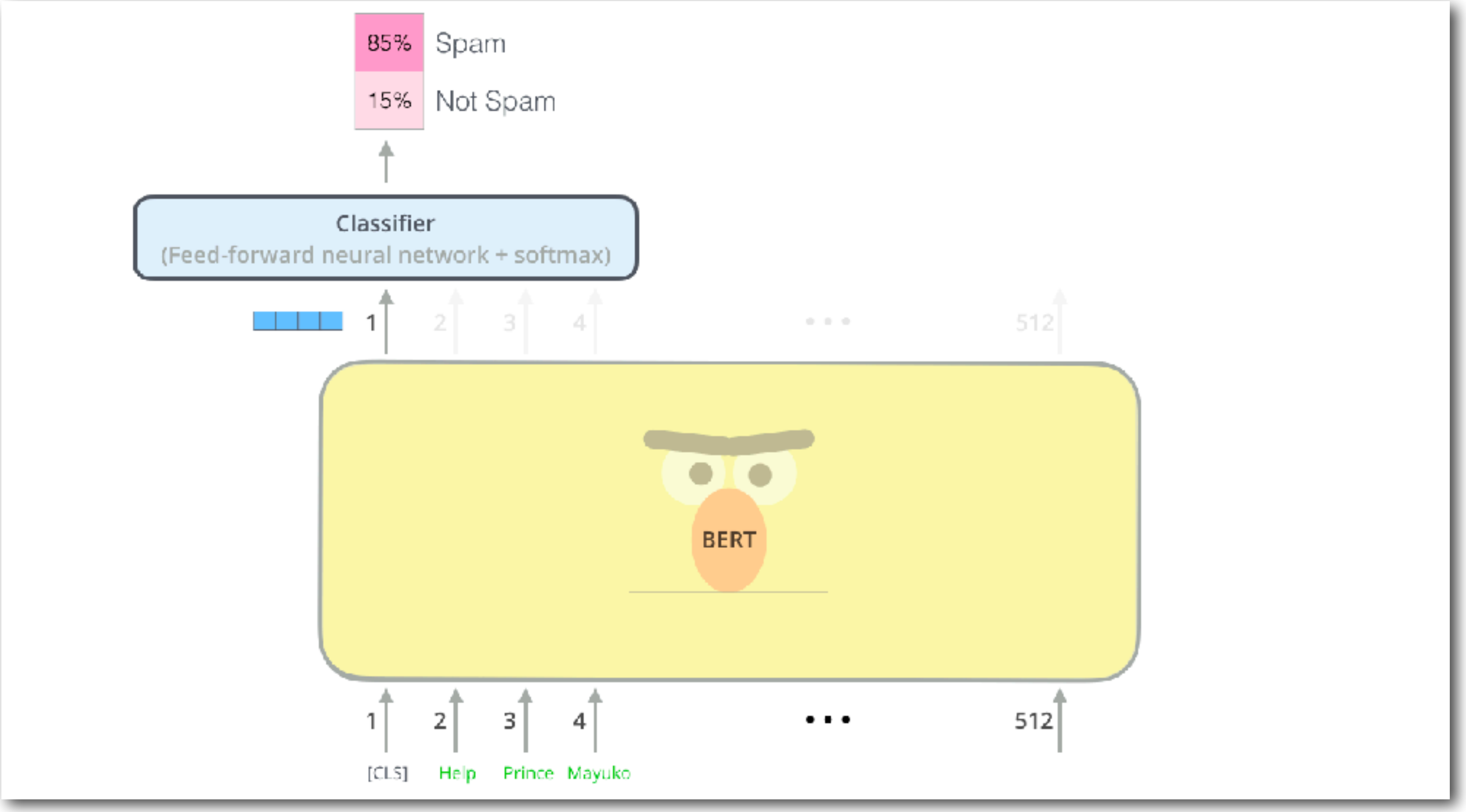
INRIA Paris - 18/11/2022

NLP beyond the top 100 languages

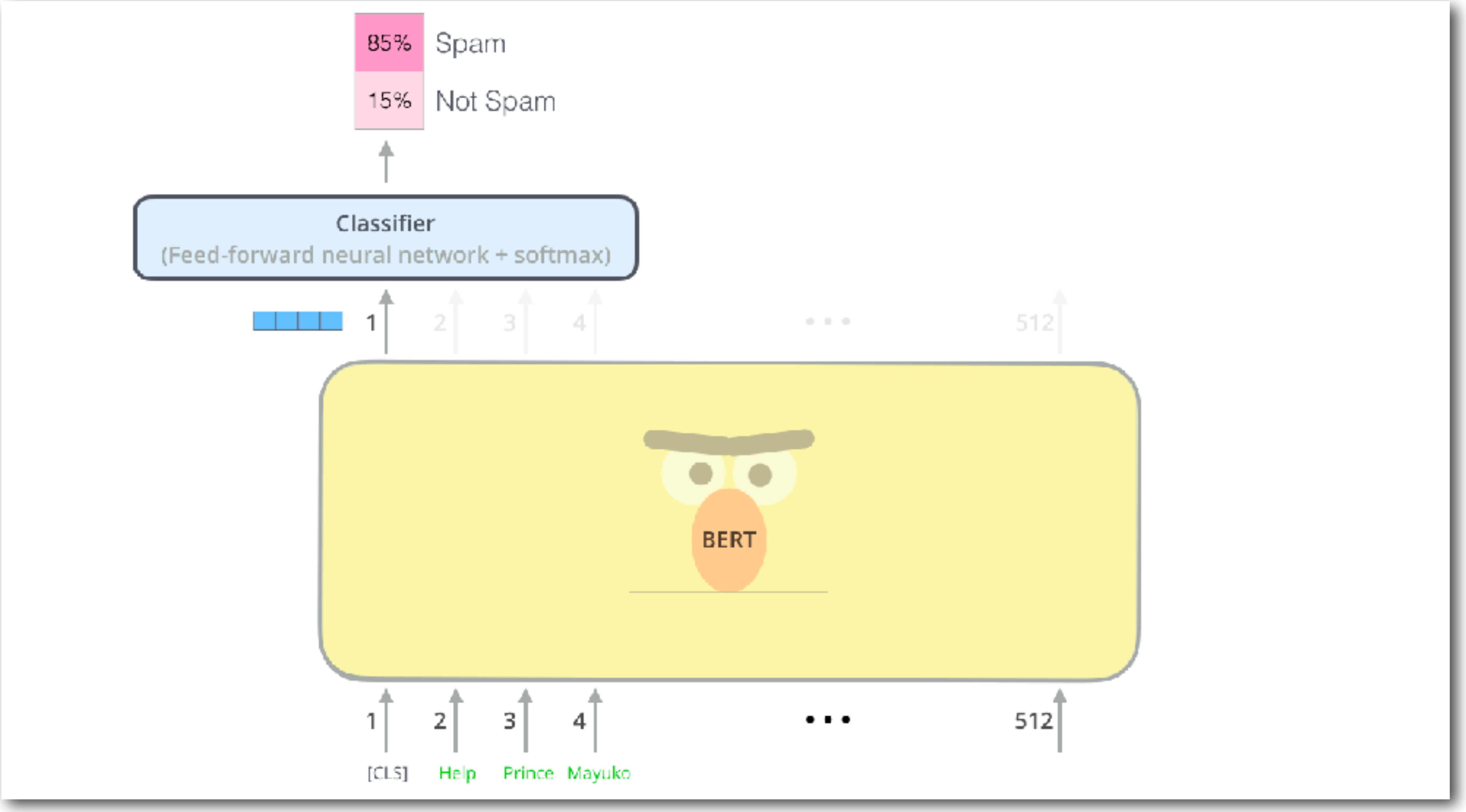
Antonis Anastasopoulos
antonis@gmu.edu

GEORGE MASON UNIVERSITY

Some recent trends



Some recent trends



Who cares

C~~hinchi~~la

~~PaLM~~

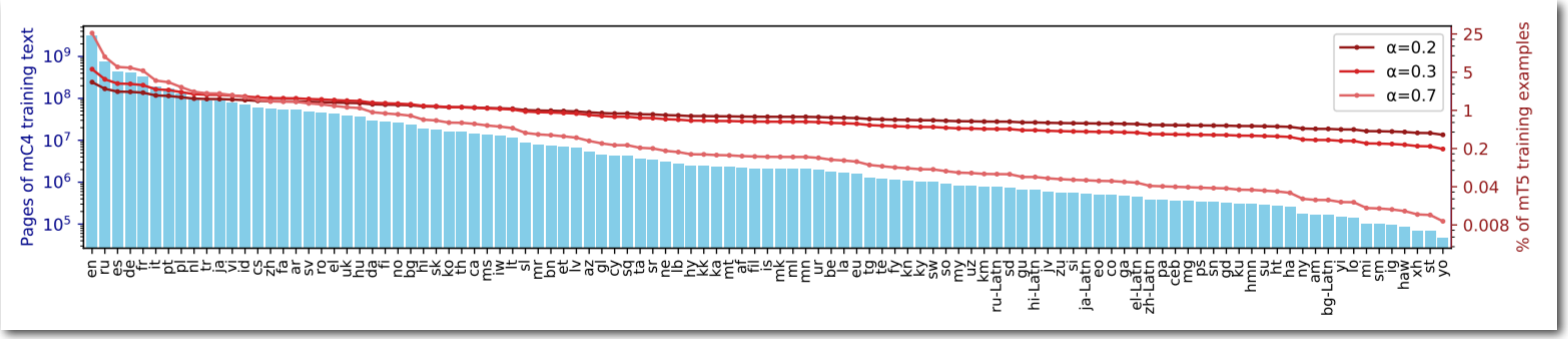
~~GPT-2~~

EL~~ECTR~~A

~~XLM-R~~

Ro~~BERT~~a

Make it multilingual!



Good recap of the current state of multilingual AI:
https://ruder.io/state-of-multilingual-ai/

Lang Tech utility is unequally distributed!

Lang Tech utility is unequally distributed!

Compare:

- American English speaker
- Arabic speaker
  - Tunisian vs Egyptian vs …
- Bemba speaker

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} \boxed{d_l^{(\tau)}} \cdot u_l$$

"normalized demand"

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} \boxed{d_l^{(\tau)}} \cdot \boxed{u_l}$$

"normalized demand" "utility"

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} \boxed{d_l^{(\tau)}} \cdot \boxed{u_l} \qquad d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

"normalized demand" "utility"

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} \underbrace{\boxed{d_l^{(\tau)}}}_{\text{"normalized demand"}} \cdot \underbrace{\boxed{u_l}}_{\text{"utility"}} \qquad d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

τ=1 : every person equal
("demographic-average utility")

Global Utility Metrics

Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi Harvard University dblasi@fas.harvard.edu
Antonios Anastasopoulos George Mason University antonis@gmu.edu
Graham Neubig Carnegie Mellon University gneubig@cs.cmu.edu

A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Two problems:

- **Problem 1:** how to measure utility of an NLP system?
  → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

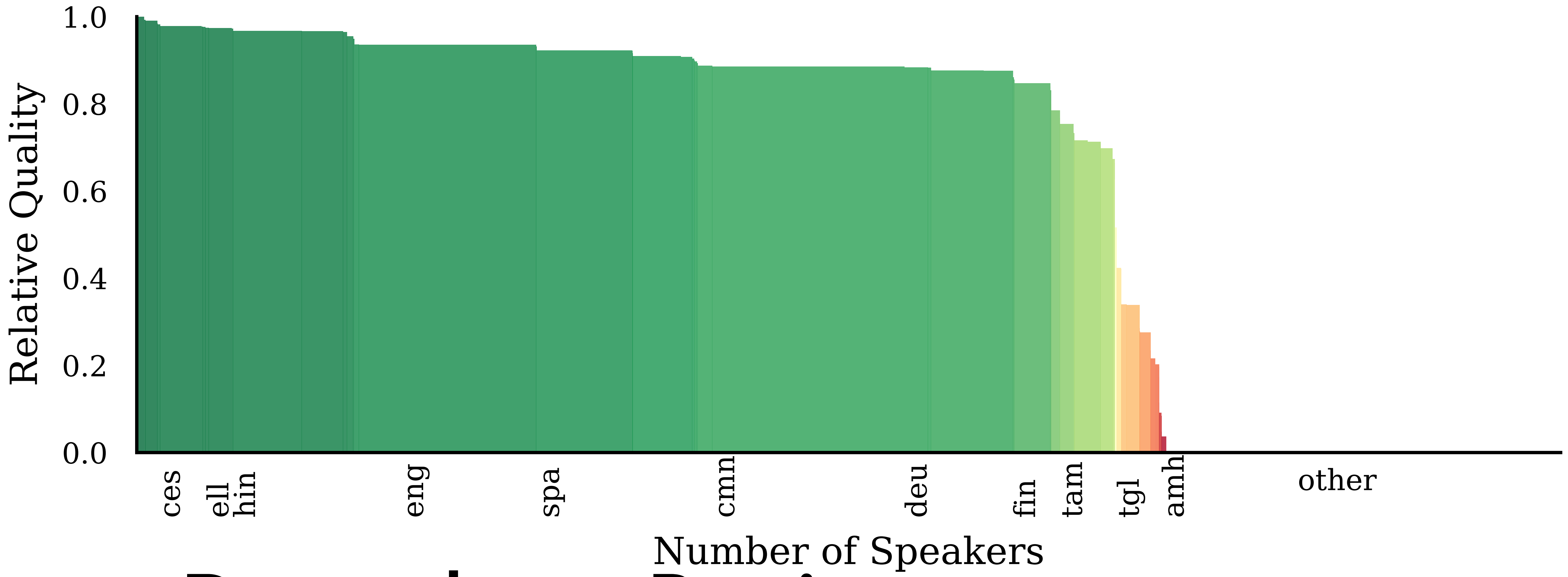
- **Problem 2:** how to consider different utility provided to every person in the world?
  → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} \boxed{d_l^{(\tau)}} \cdot \boxed{u_l} \qquad d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

"normalized demand" "utility"

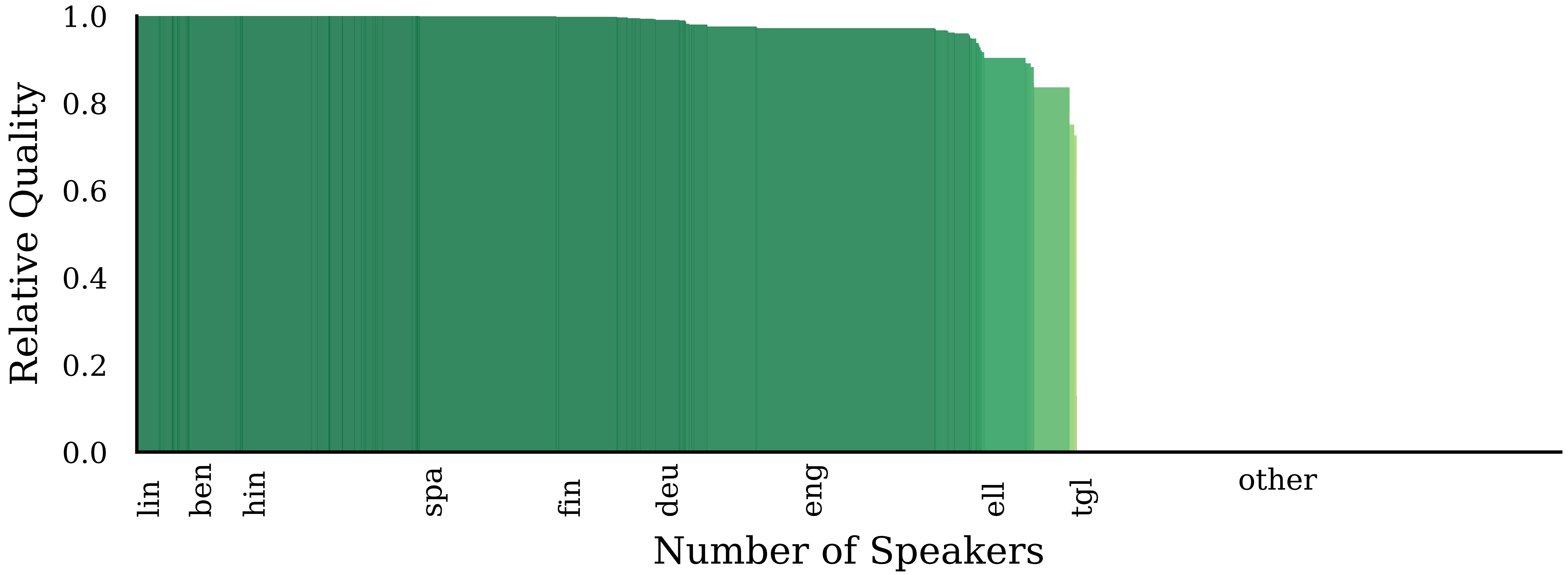
τ=1 : every person equal ("demographic-average utility")
τ=0 : every subgroup equal ("linguistic-average utility")

Zooming In (Analysis Tasks)

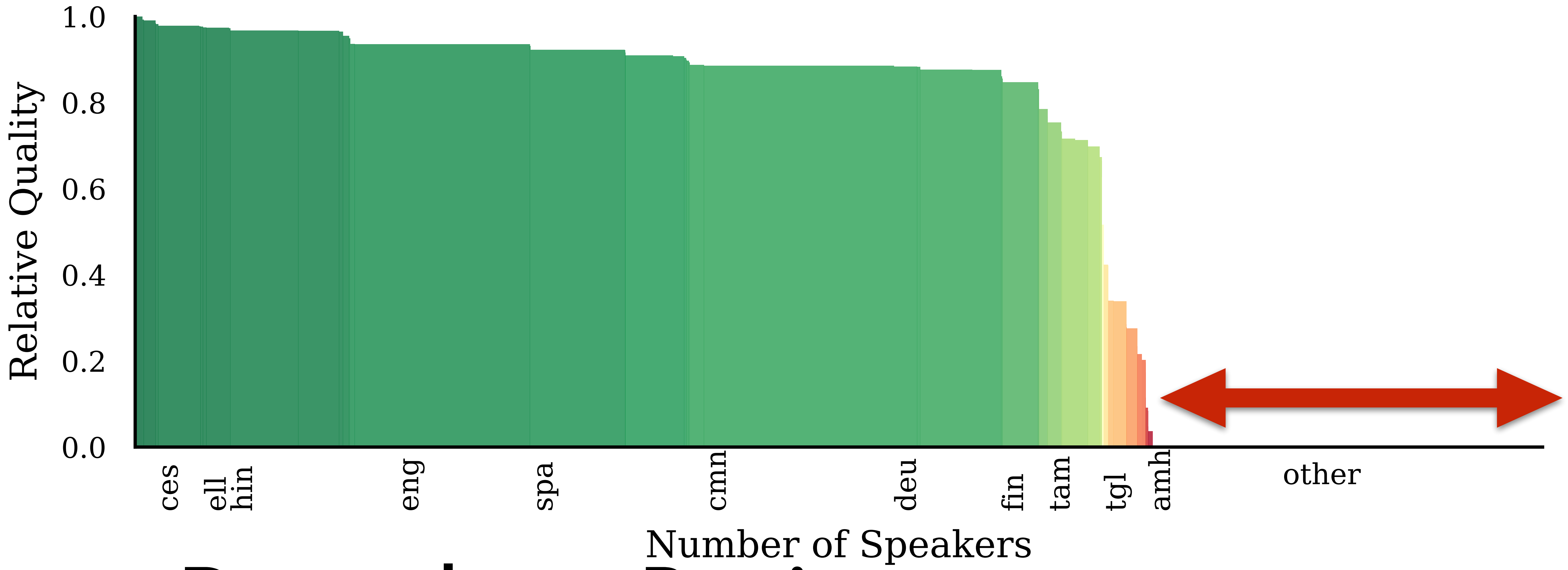


Dependency Parsing

Inflection

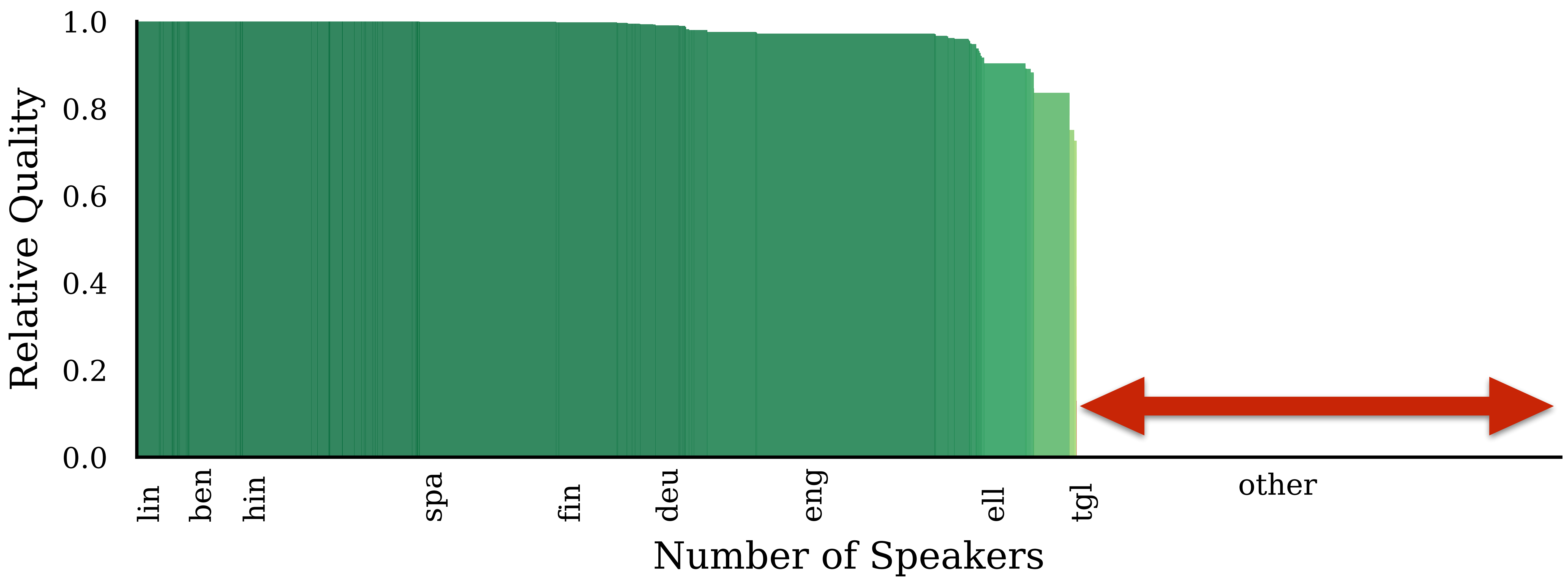


Zooming In (Analysis Tasks)

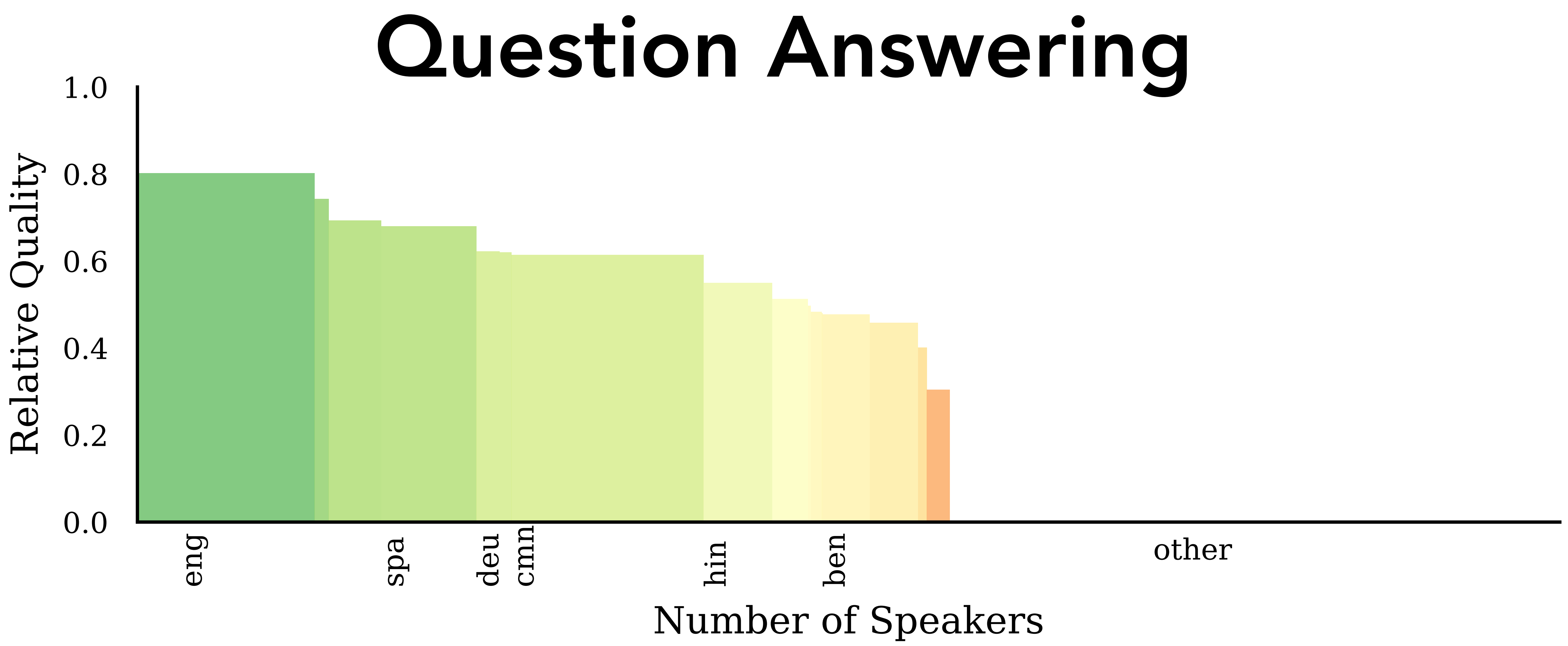


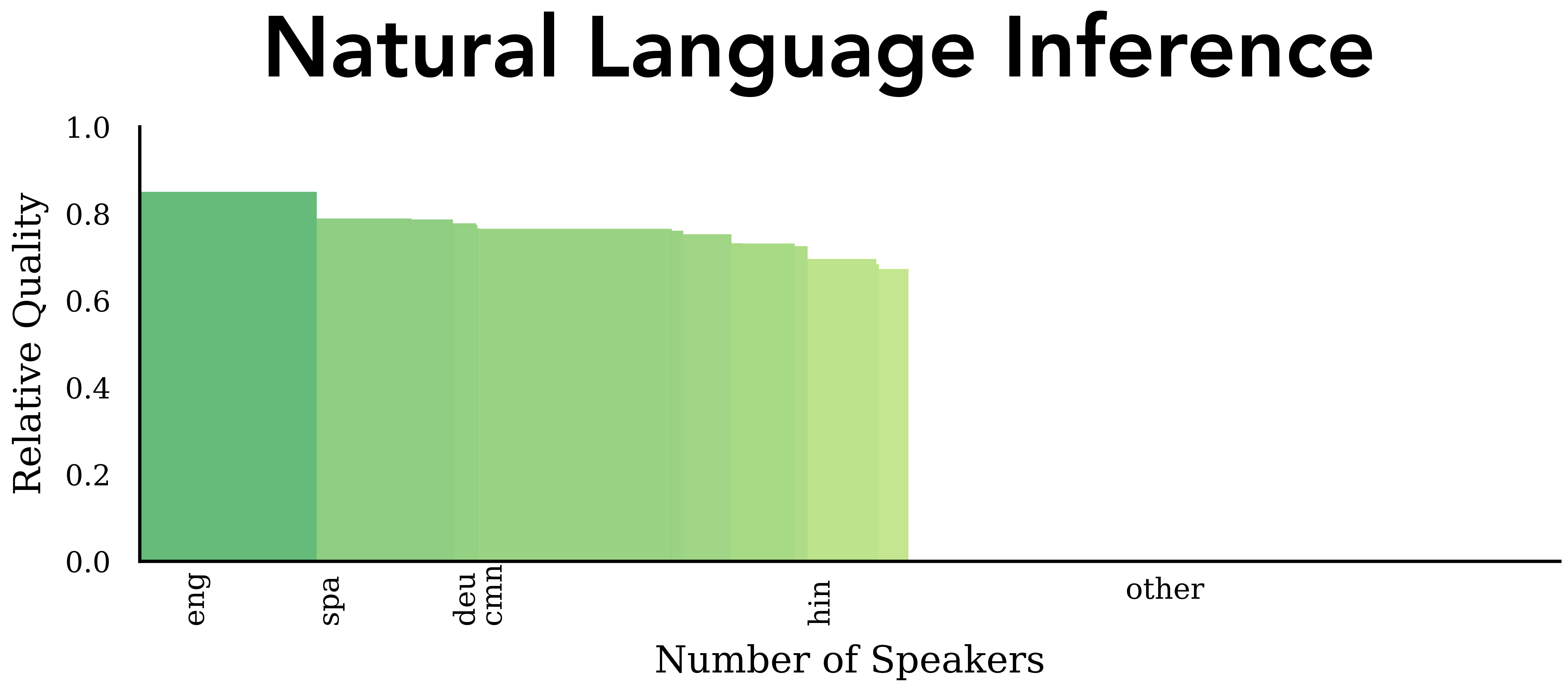
Dependency Parsing

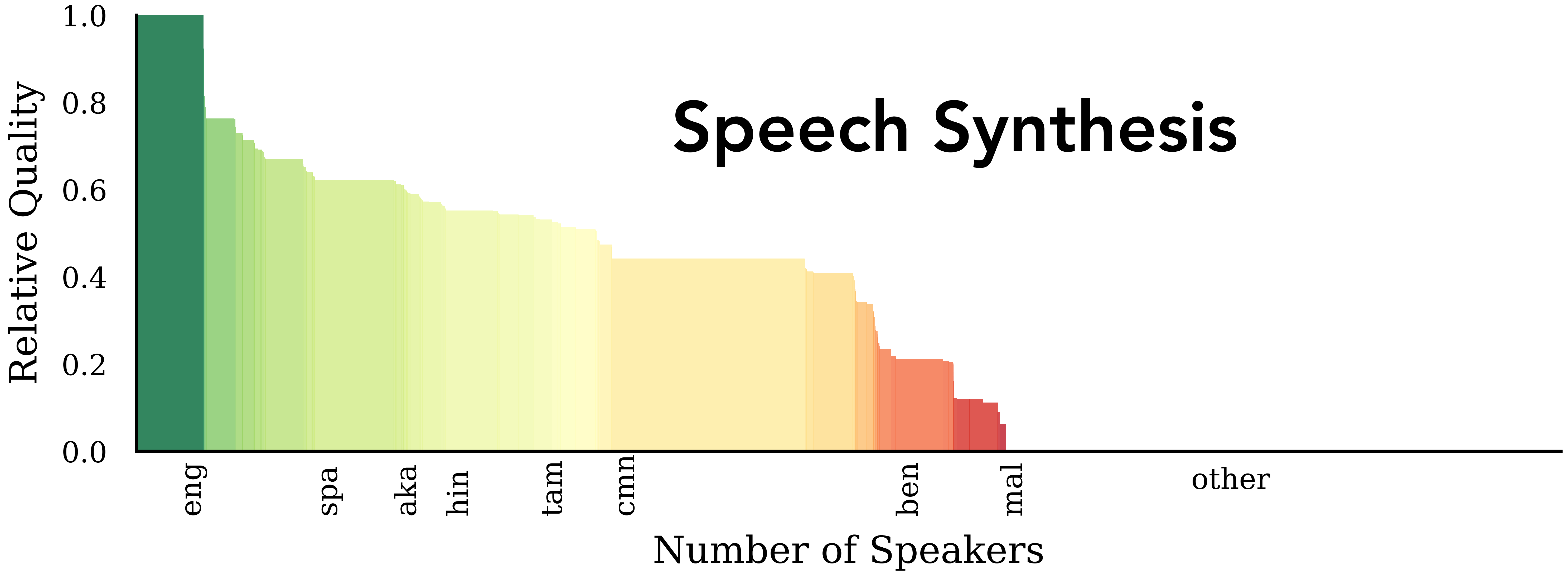
Inflection



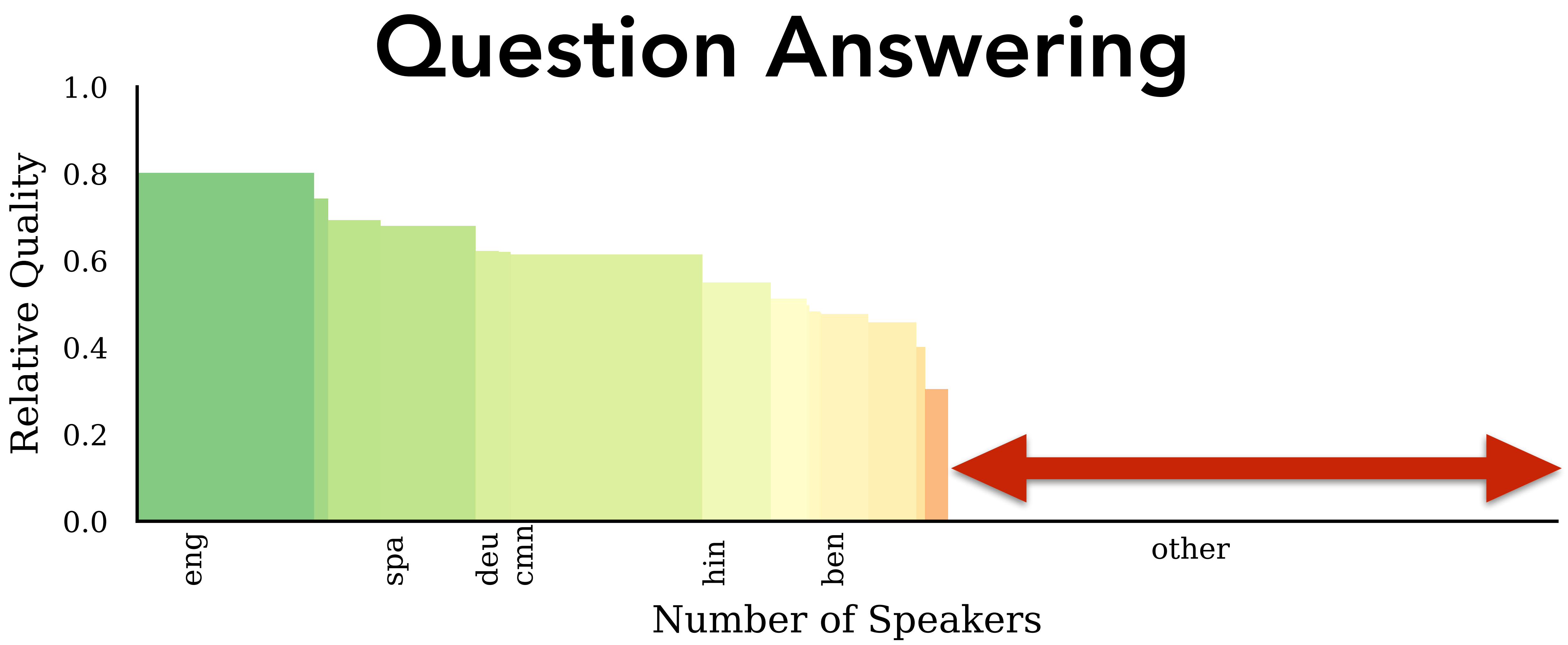
Zooming In (User-facing Tasks)

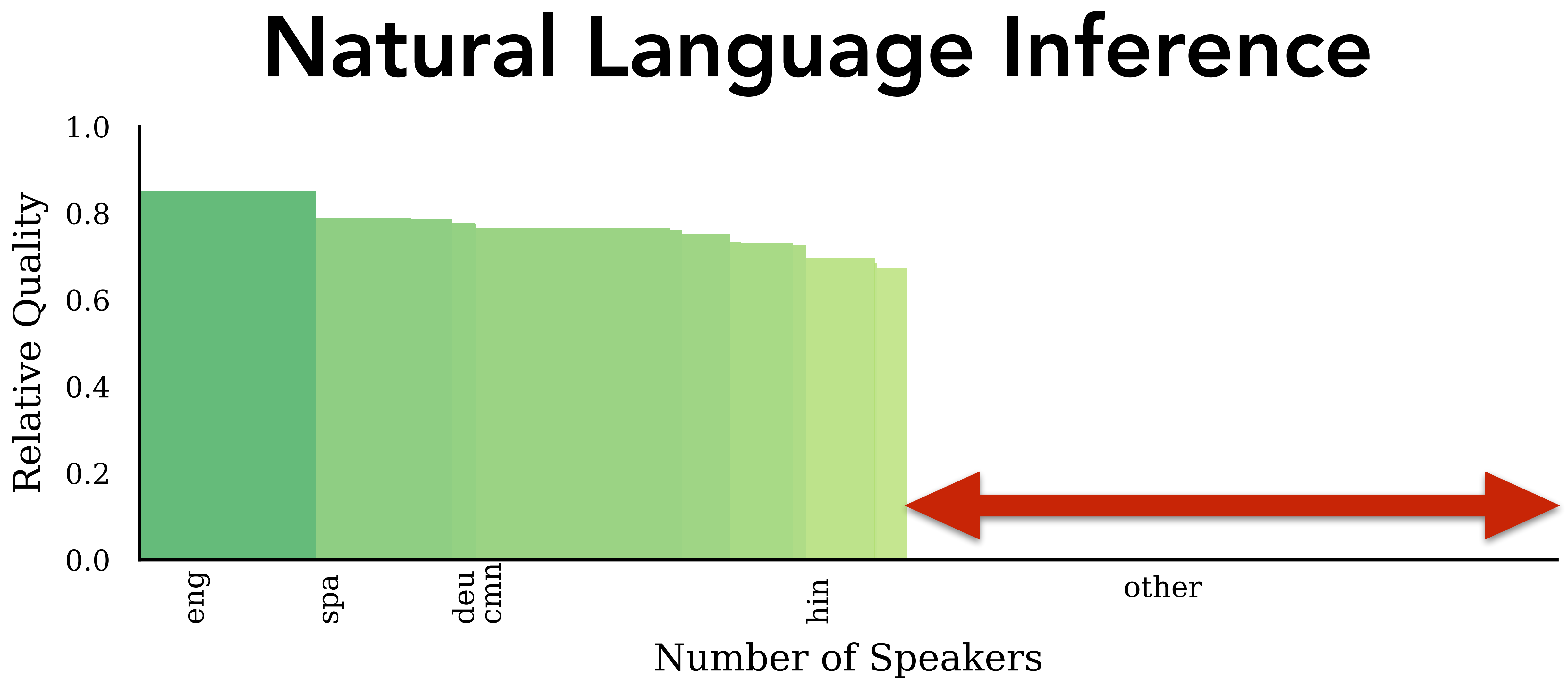


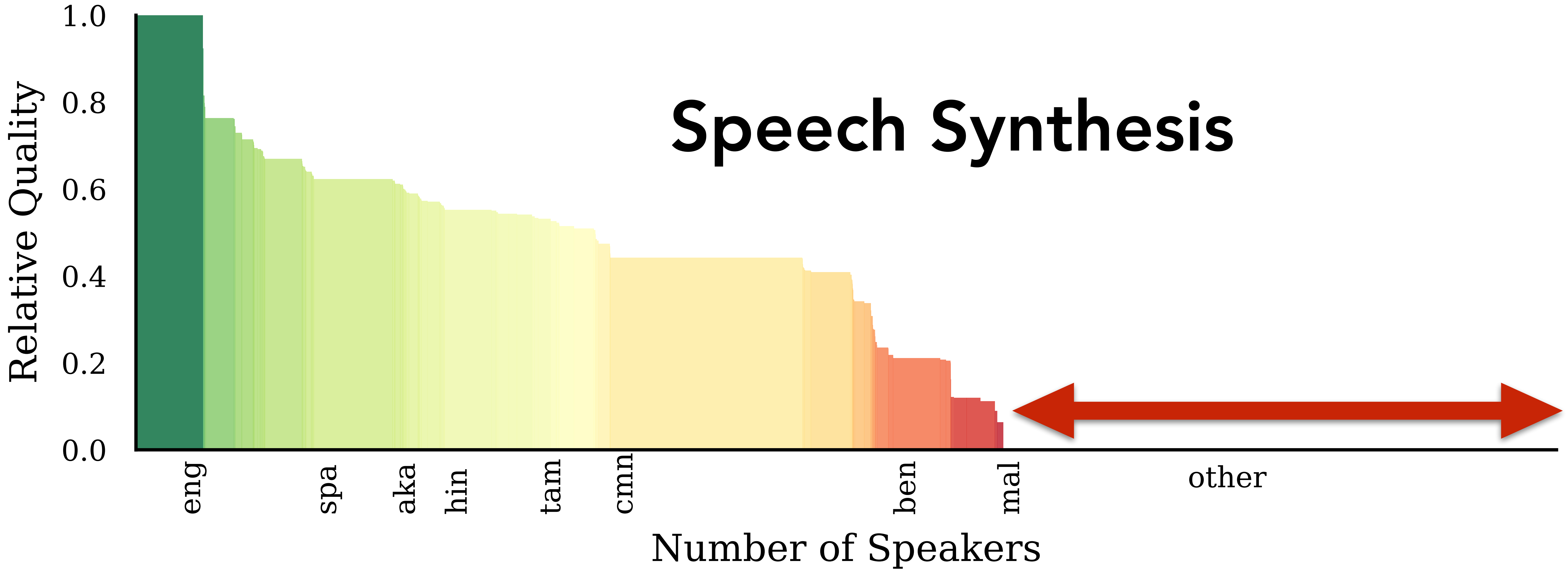




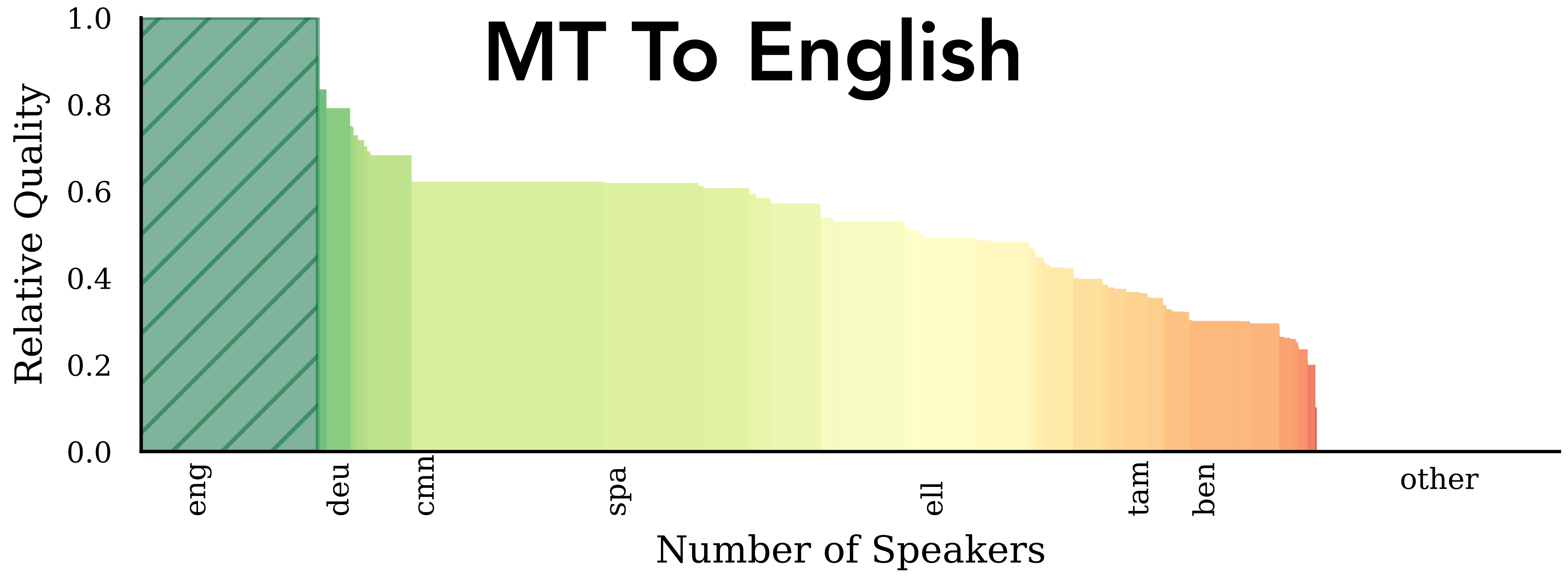
Zooming In (User-facing Tasks)

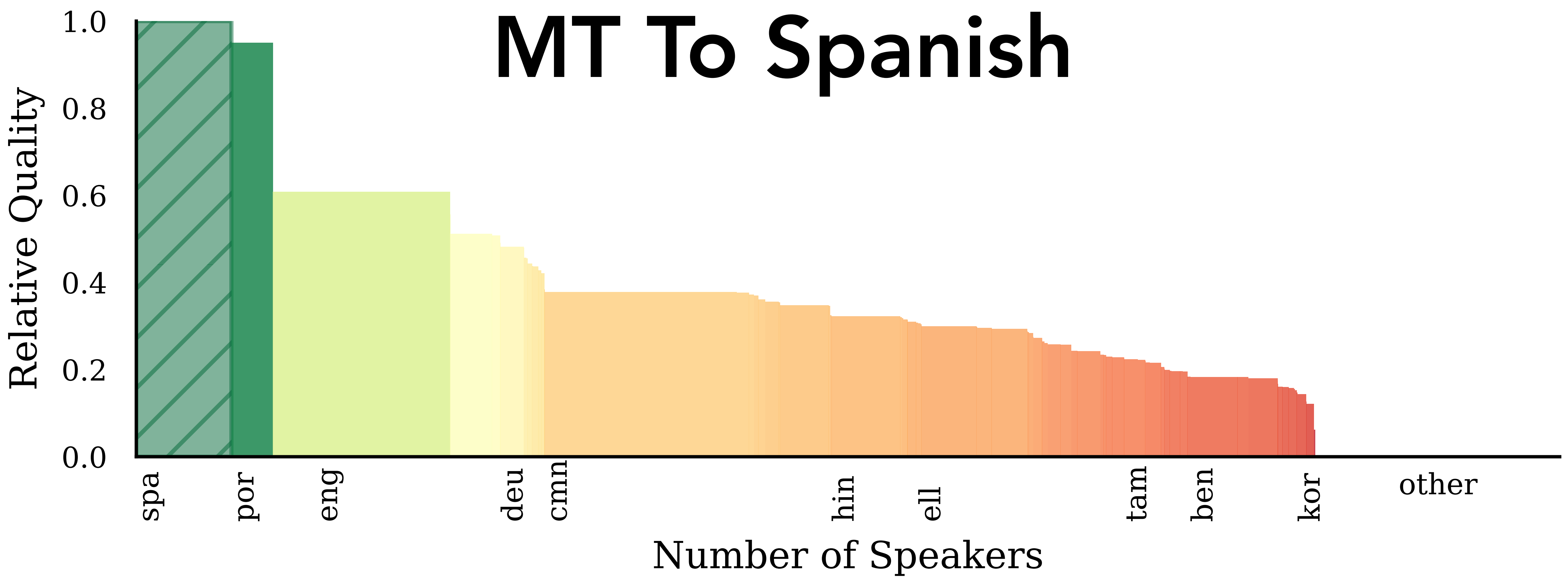


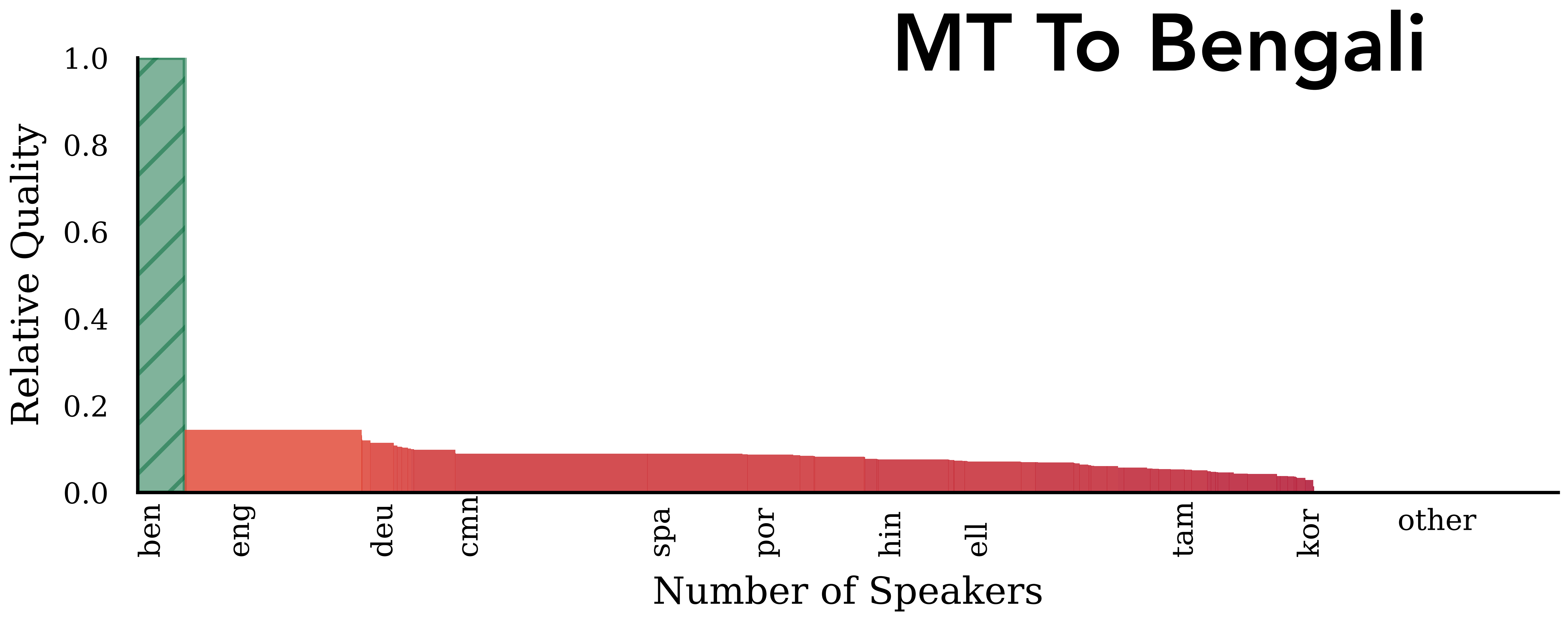




Zooming In (Machine Translation)







Going Deeper: Dialects

Going Deeper: Dialects

Very few languages are monoliths!

Going Deeper: Dialects

Very few languages are monoliths!

Need to model dialectal/regional/user variations.

Going Deeper: Dialects

Very few languages are monoliths!

Need to model dialectal/regional/user variations.

Problem: most are *spoken (like 45% of all languages)*

Going Deeper: Dialects

Very few languages are monoliths!

Need to model dialectal/regional/user variations.

Problem: most are *spoken (like 45% of all languages)*

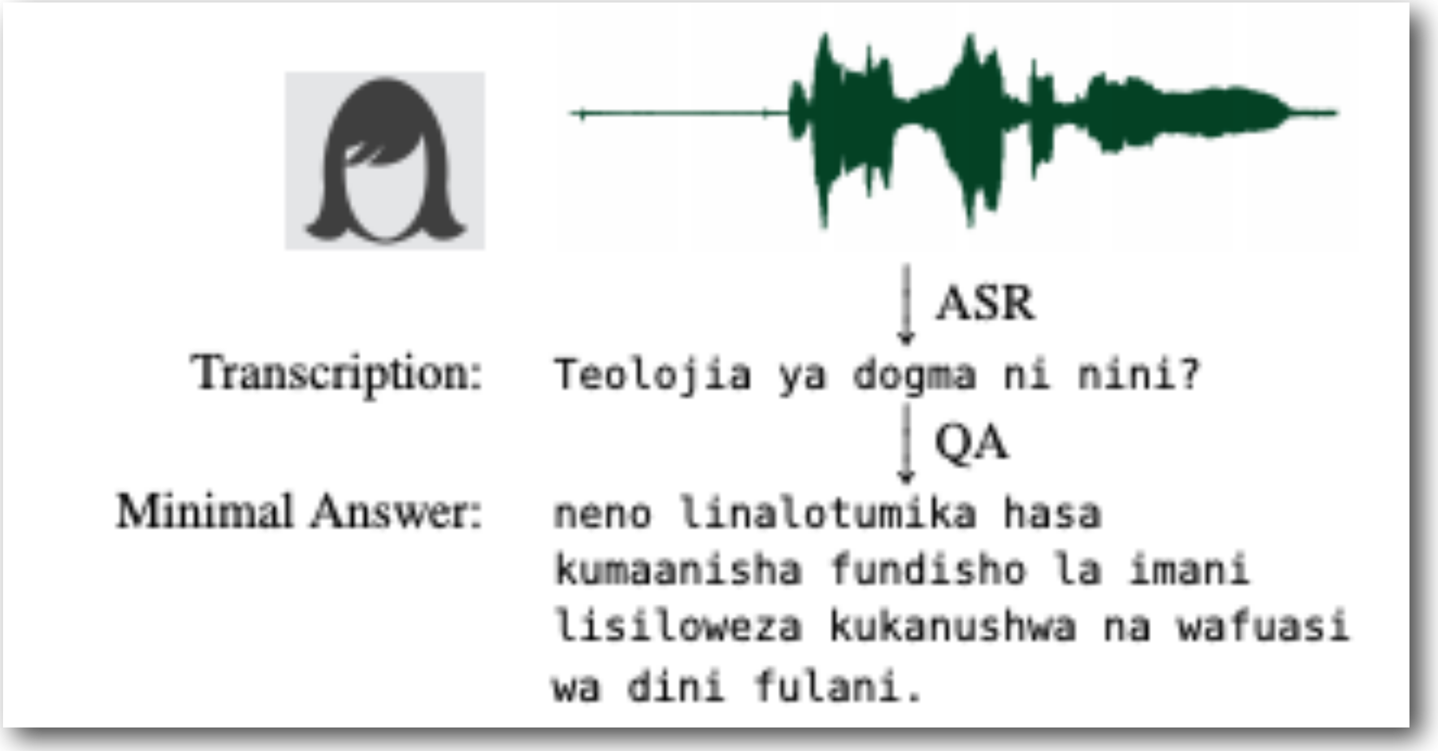
SD-QA: Spoken Dialectal Question Answering for the Real World

Fahim Faisal, Sharlina Keshava, Md Mahfuz ibn Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal,skeshav,malam21,antonis}@gmu.edu

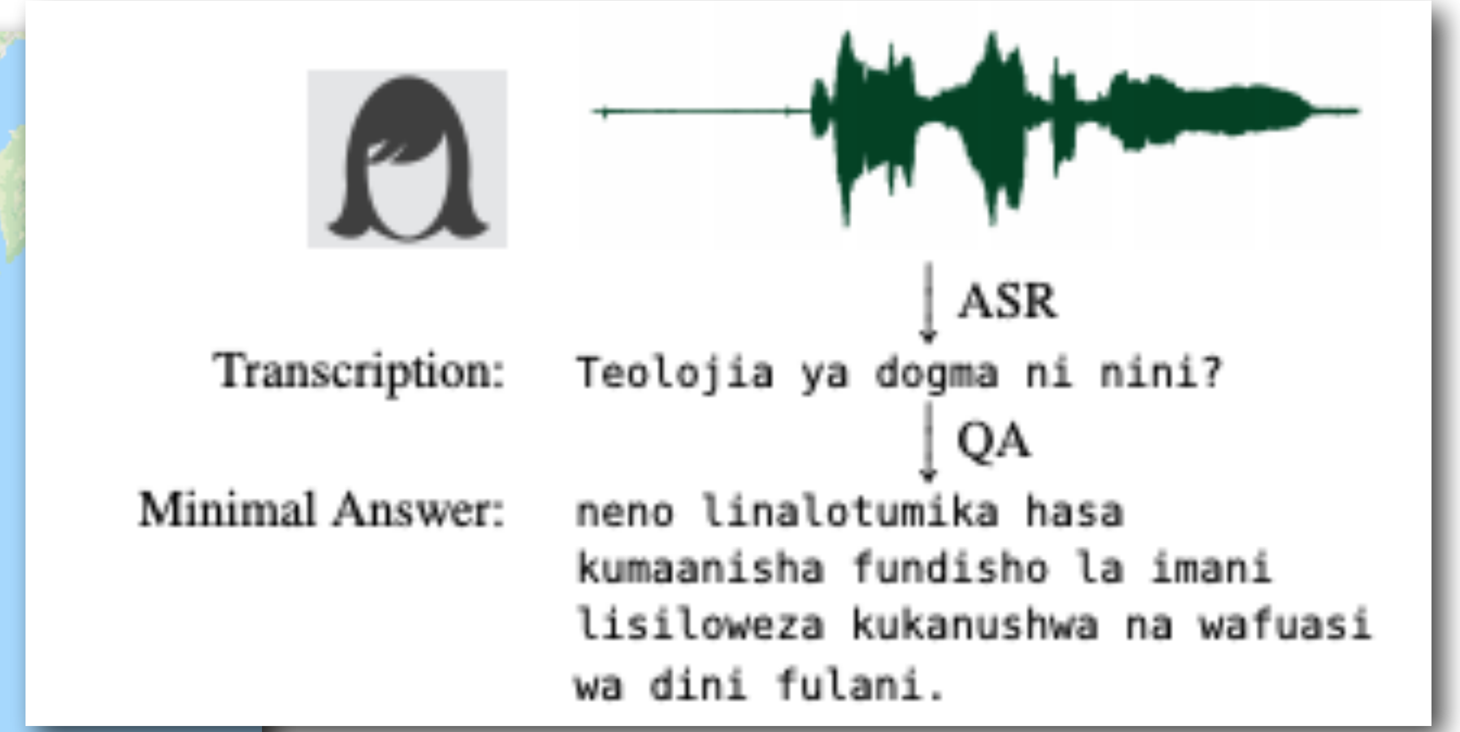
(EMNLP Findings 2021)

SD-QA: Spoken, Dialectal, Multilingual QA

SD-QA: Spoken, Dialectal, Multilingual QA

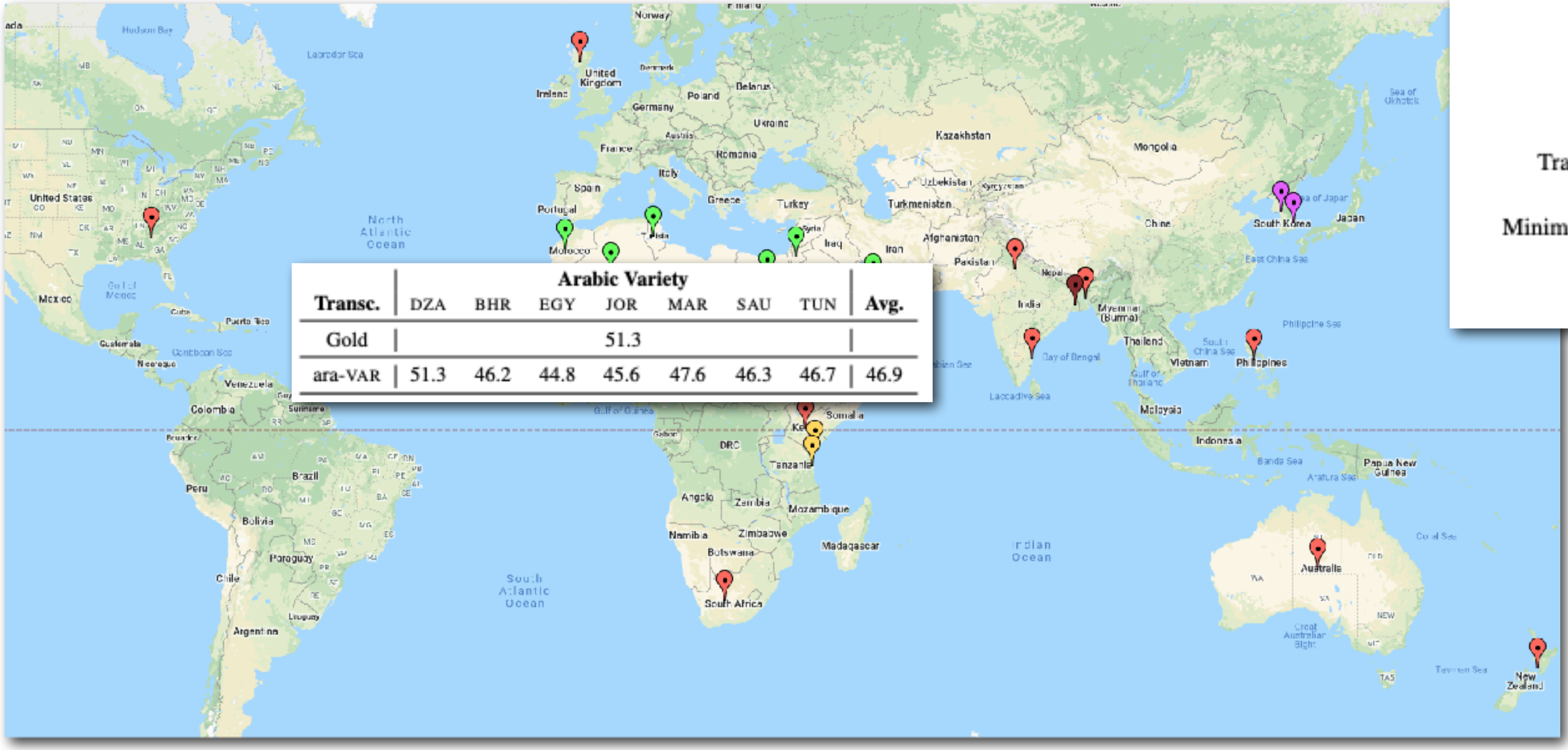


SD-QA: Spoken, Dialectal, Multilingual QA



NLP GEORGE MASON

SD-QA: Spoken, Dialectal, Multilingual QA



Transc.	Arabic Variety							Avg.
	DZA	BHR	EGY	JOR	MAR	SAU	TUN	
Gold	51.3							
ara-VAR	51.3	46.2	44.8	45.6	47.6	46.3	46.7	46.9

ASR

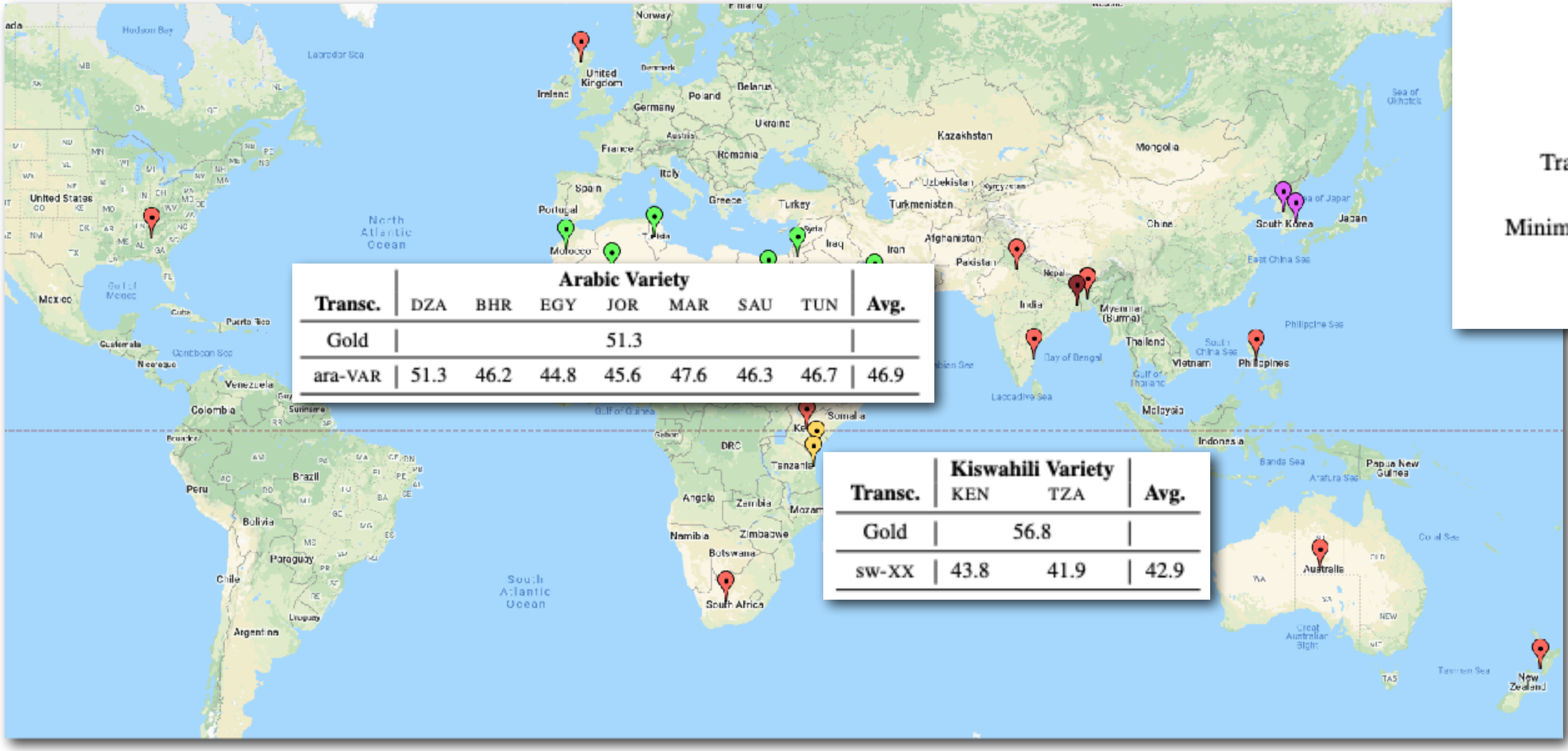
Transcription: Teolojia ya dogma ni nini?

QA

Minimal Answer: neno linalotumika hasa kumaanisha fundisho la imani lisiloweza kukanushwa na wafuasi wa dini fulani.

NLP GEORGE MASON

SD-QA: Spoken, Dialectal, Multilingual QA



Transc.	Arabic Variety DZA	BHR	EGY	JOR	MAR	SAU	TUN	Avg.
Gold	51.3							
ara-VAR	51.3	46.2	44.8	45.6	47.6	46.3	46.7	46.9

Transc.	Kiswahili Variety KEN	TZA	Avg.
Gold	56.8		
sw-XX	43.8	41.9	42.9

ASR

Transcription: `Teolojia ya dogma ni nini?`

QA

Minimal Answer: `neno linalotumika hasa kumaanisha fundisho la imani lisiloweza kukanushwa na wafuasi wa dini fulani.`

NLP beyond the top-100 languages

GEORGE MASON UNIVERSITY

Going beyond the top-100 languages

Going beyond the top-100 languages

Dominant
Written (Latin)
Standardized
high(ish)-resource

Going beyond the top-100 languages

Dominant
Written (Latin)
Standardized
high(ish)-resource

Local
Oral
non-Standardized
Very low-resource

NLP GEORGE MASON

Are all unseen languages equally hard?

When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models

Benjamin Muller[†] **Antonis Anastasopoulos**[‡] **Benoît Sagot**[†] **Djamé Seddah**[†]
[†]Inria, Paris, France
[‡]Department of Computer Science, George Mason University, USA
firstname.lastname@inria.fr antonis@gmu.edu

(NAACL 2021)

https://github.com/benjamin-mlr/mbert-unseen-languages.git

NLP GEORGE MASON

Are all unseen languages hard?

Are all unseen languages hard?

Some are “easy”

Are all unseen languages hard?

Some are "easy"

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Are all unseen languages hard?

Some are “easy”

- Similar languages in pre-training + same script
- e.g. Faroese, Swiss German

Some are “intermediate”

Are all unseen languages hard?

Some are "easy"

- Similar languages in pre-training + same script
- e.g. Faroese, Swiss German

Some are "intermediate"

- Simple approach (continued fine-tuning) leads to good results
- e.g. Maltese, Bambara, Wolof

Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Some are “intermediate”

Simple approach (continued fine-tuning) leads to good results

e.g. Maltese, Bambara, Wolof

Some seem “hard”

Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Some are “intermediate”

Simple approach (continued fine-tuning) leads to good results

e.g. Maltese, Bambara, Wolof

Some seem “hard”

Hard == Static monolingual embeddings >> mBERT adaptation

Are all unseen languages hard?

Some are "easy"

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Some are "intermediate"

Simple approach (continued fine-tuning) leads to good results

e.g. Maltese, Bambara, Wolof

Some seem "hard"

Hard == Static monolingual embeddings >> mBERT adaptation

Similar languages in pre-training BUT different script

Are all unseen languages hard?

Some are “easy”

Similar languages in pre-training + same script

e.g. Faroese, Swiss German

Some are “intermediate”

Simple approach (continued fine-tuning) leads to good results

e.g. Maltese, Bambara, Wolof

Some seem “hard”

Hard == Static monolingual embeddings >> mBERT adaptation

Similar languages in pre-training BUT different script

e.g. Uyghur, Sorani, Sindhi

Are all unseen languages hard?

Some are "easy"

- Similar languages in pre-training + same script
- e.g. Faroese, Swiss German

Some are "intermediate"

- Simple approach (continued fine-tuning) leads to good results
- e.g. Maltese, Bambara, Wolof

Some seem "hard"

- Hard == Static monolingual embeddings >> mBERT adaptation
- Similar languages in pre-training BUT different script
- e.g. Uyghur, Sorani, Sindhi
- Transliteration helps

Doing better by hard-coding linguistic information

Phylogeny-Inspired Adaptation of Multilingual Models to New Languages

Fahim Faisal, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal,antonis}@gmu.edu

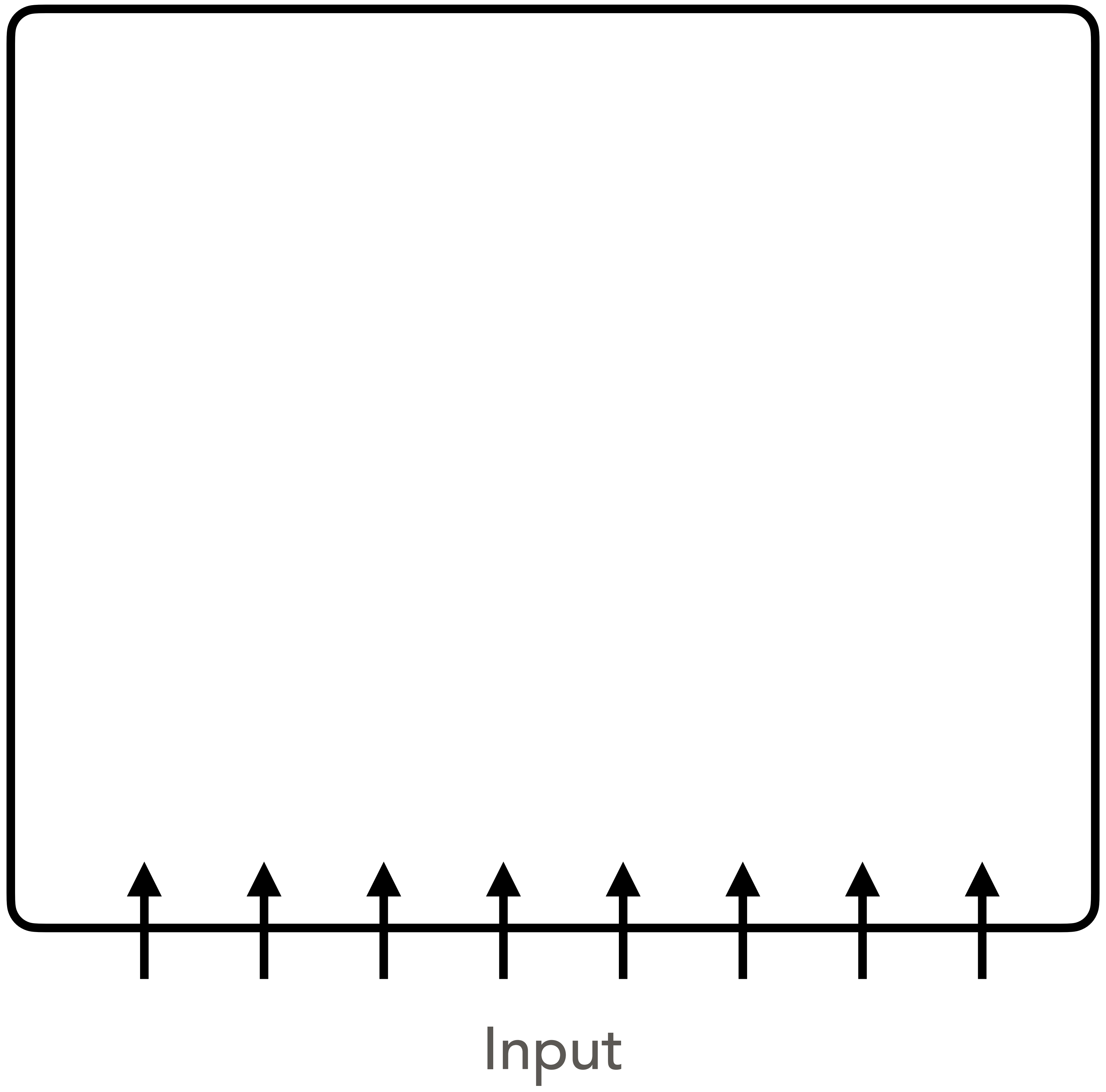
(AACL 2022)

https://github.com/ffaisal93/adapt_lang_phylogeny

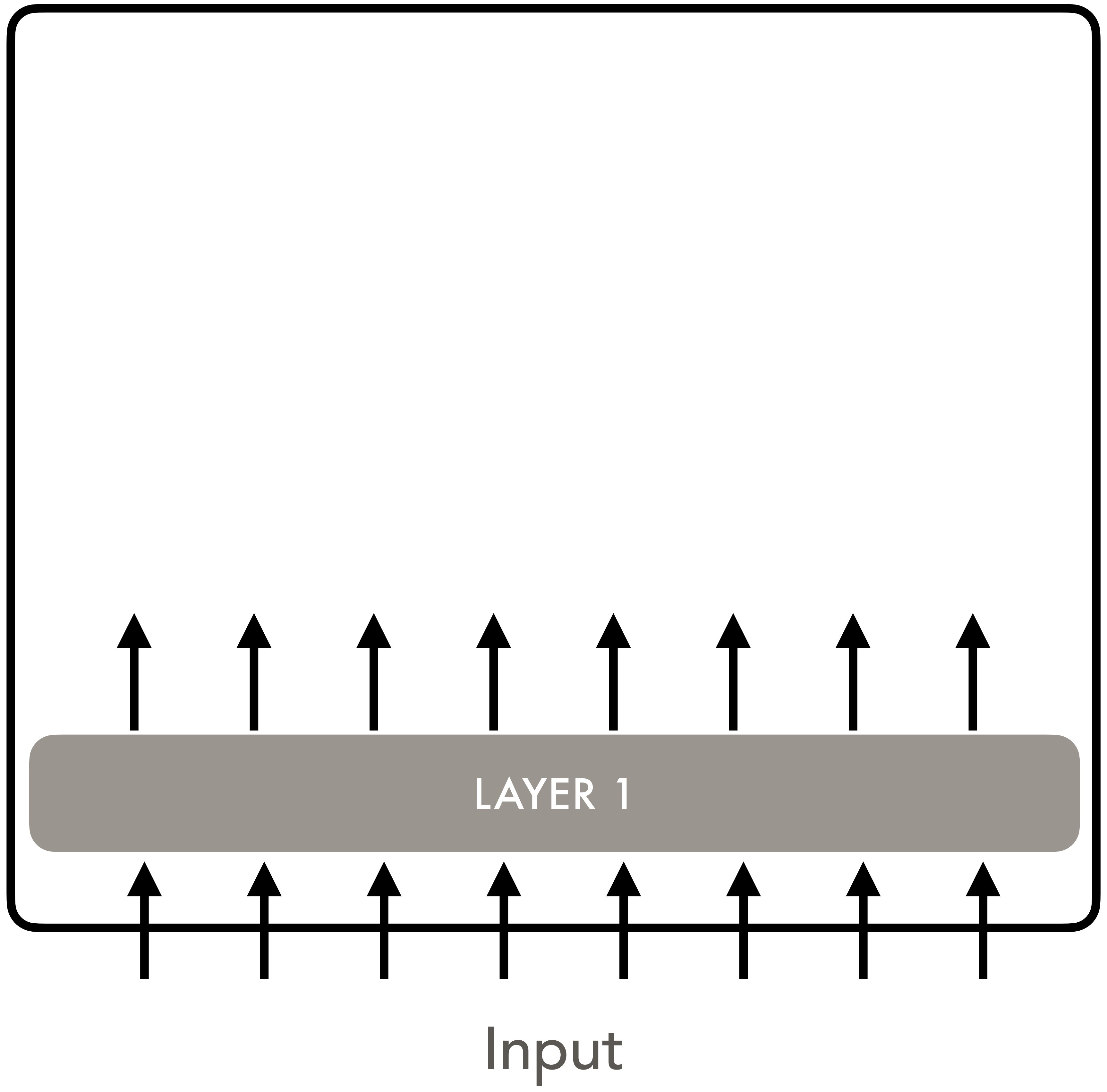
Revisiting Adapters

Revisiting Adapters

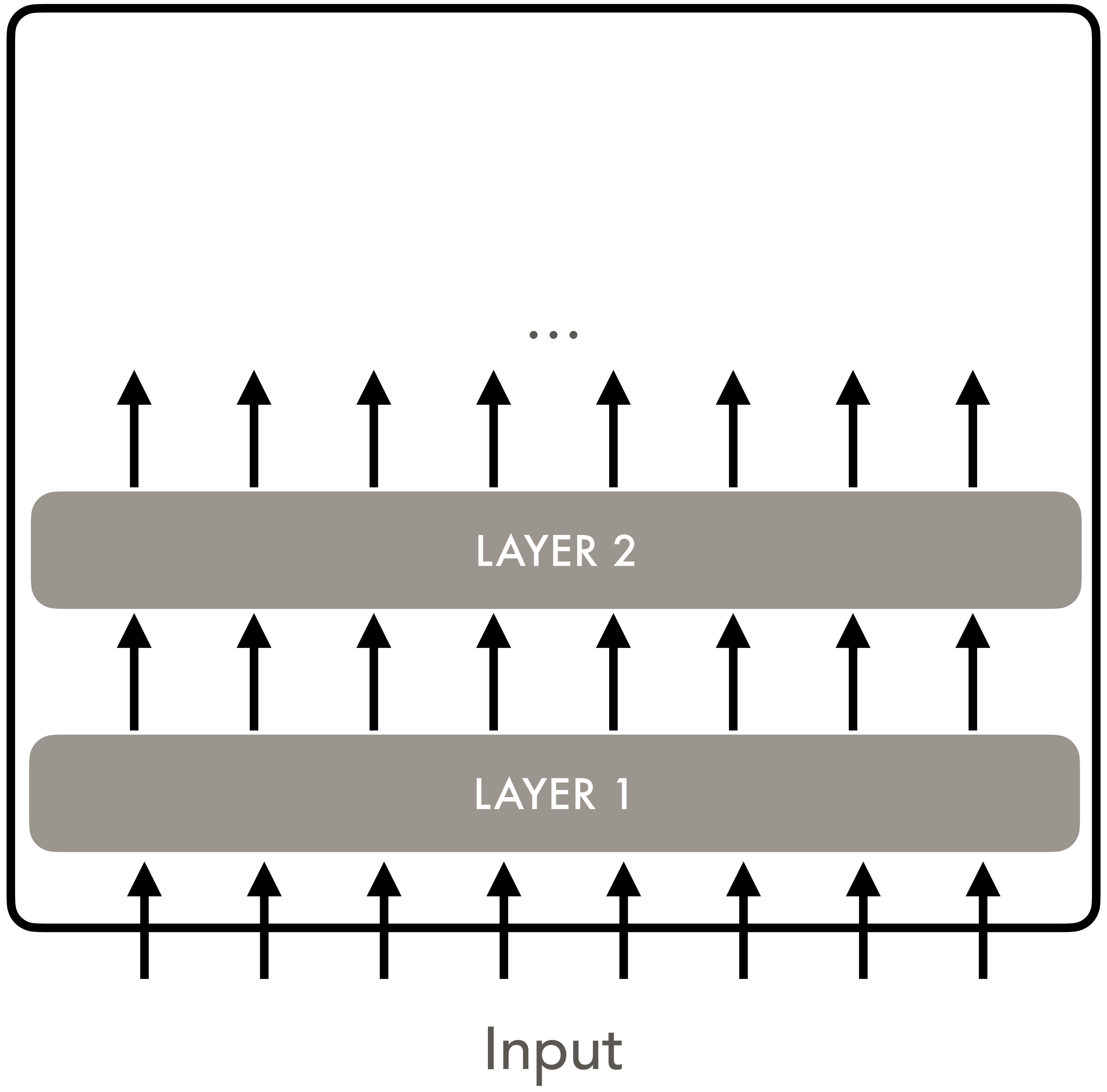
Revisiting Adapters



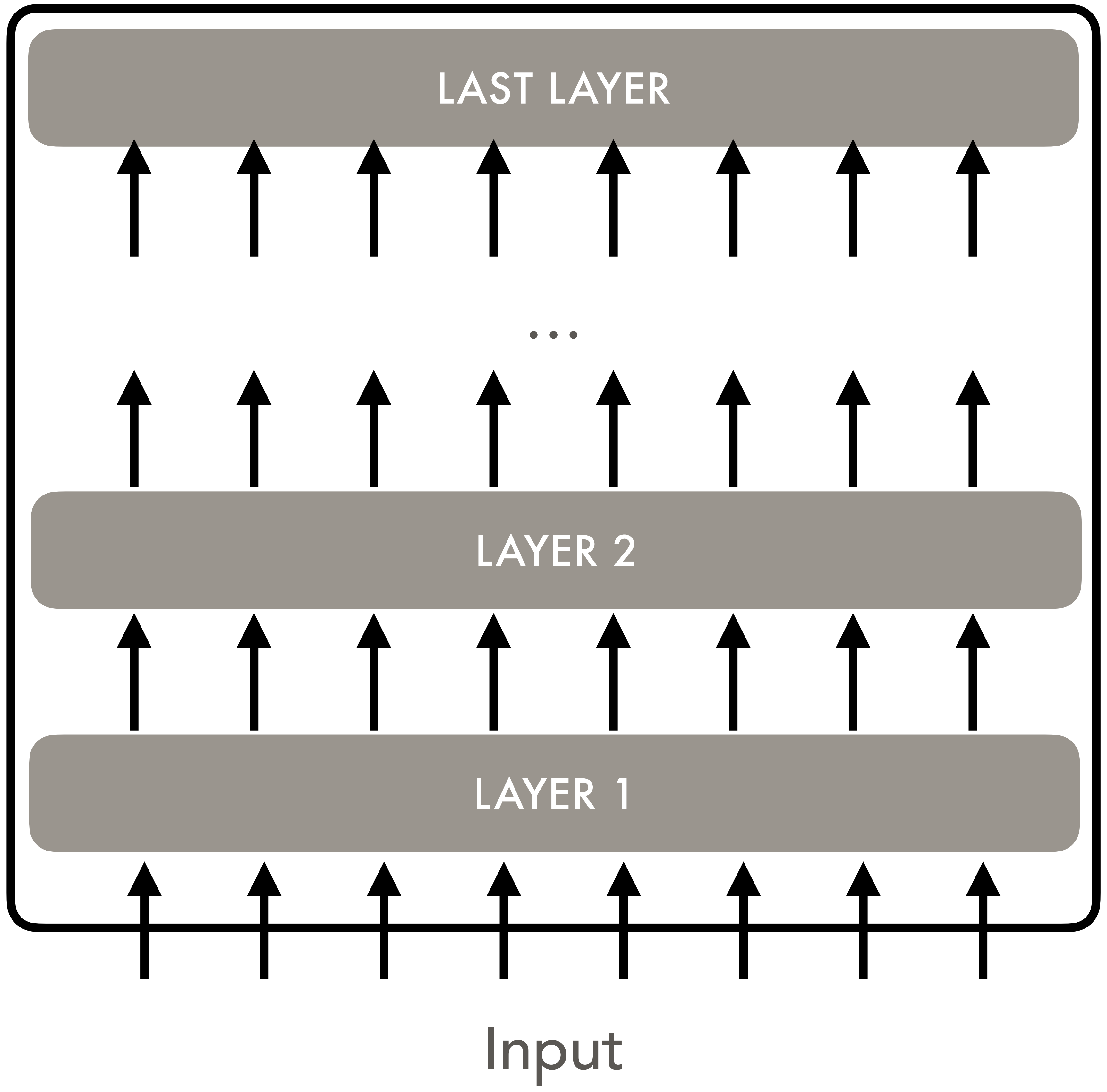
Revisiting Adapters



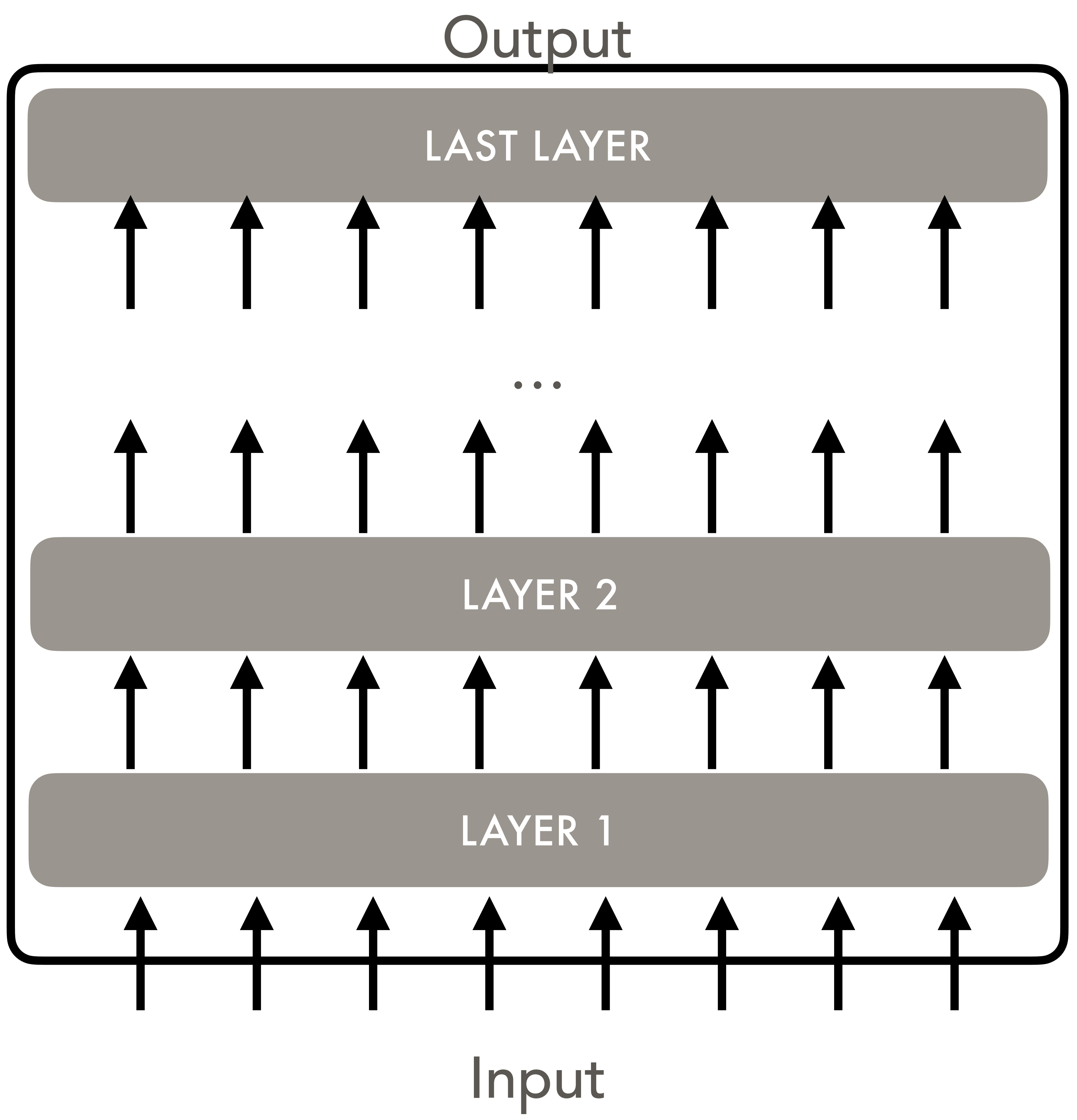
Revisiting Adapters



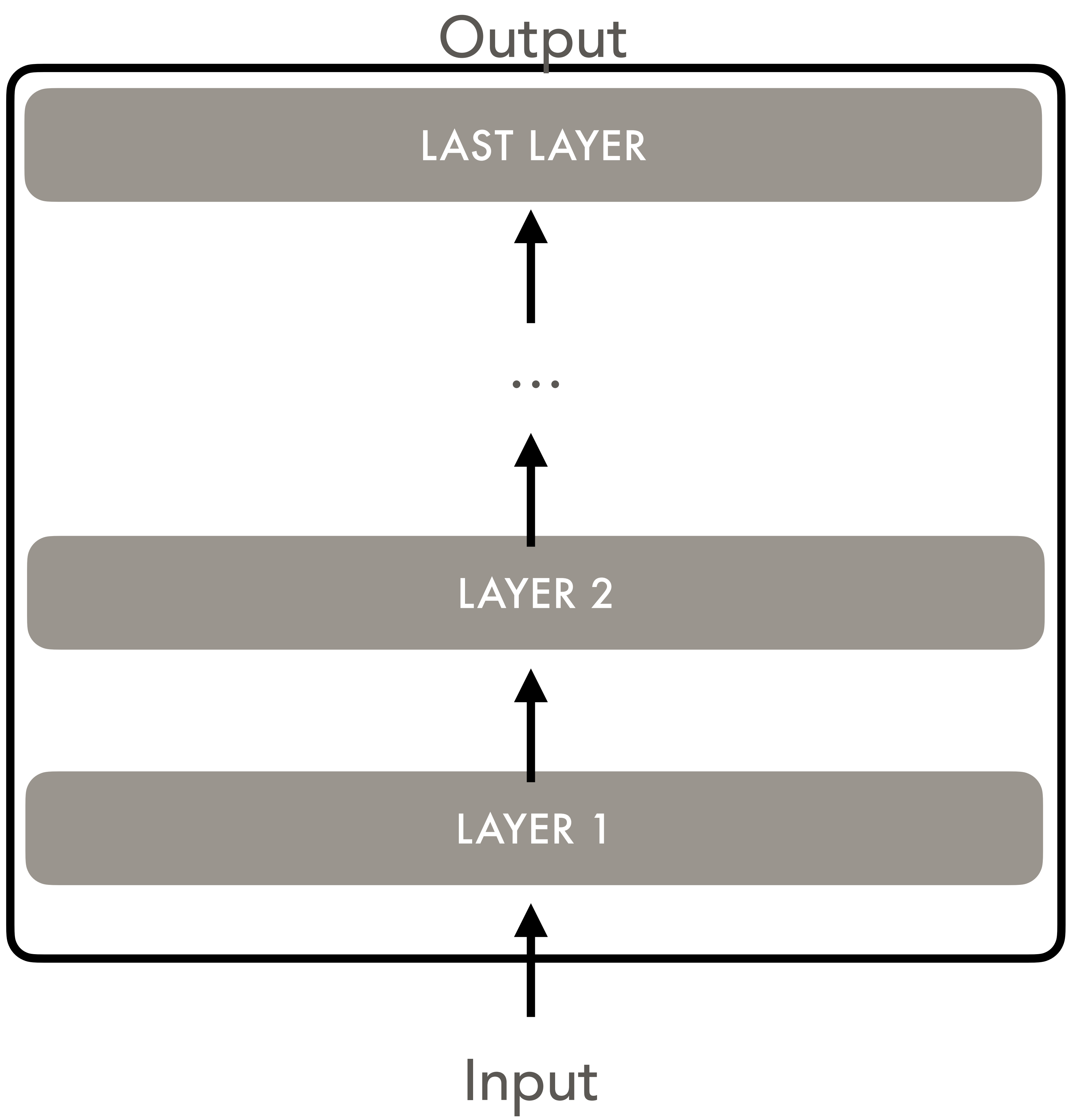
Revisiting Adapters



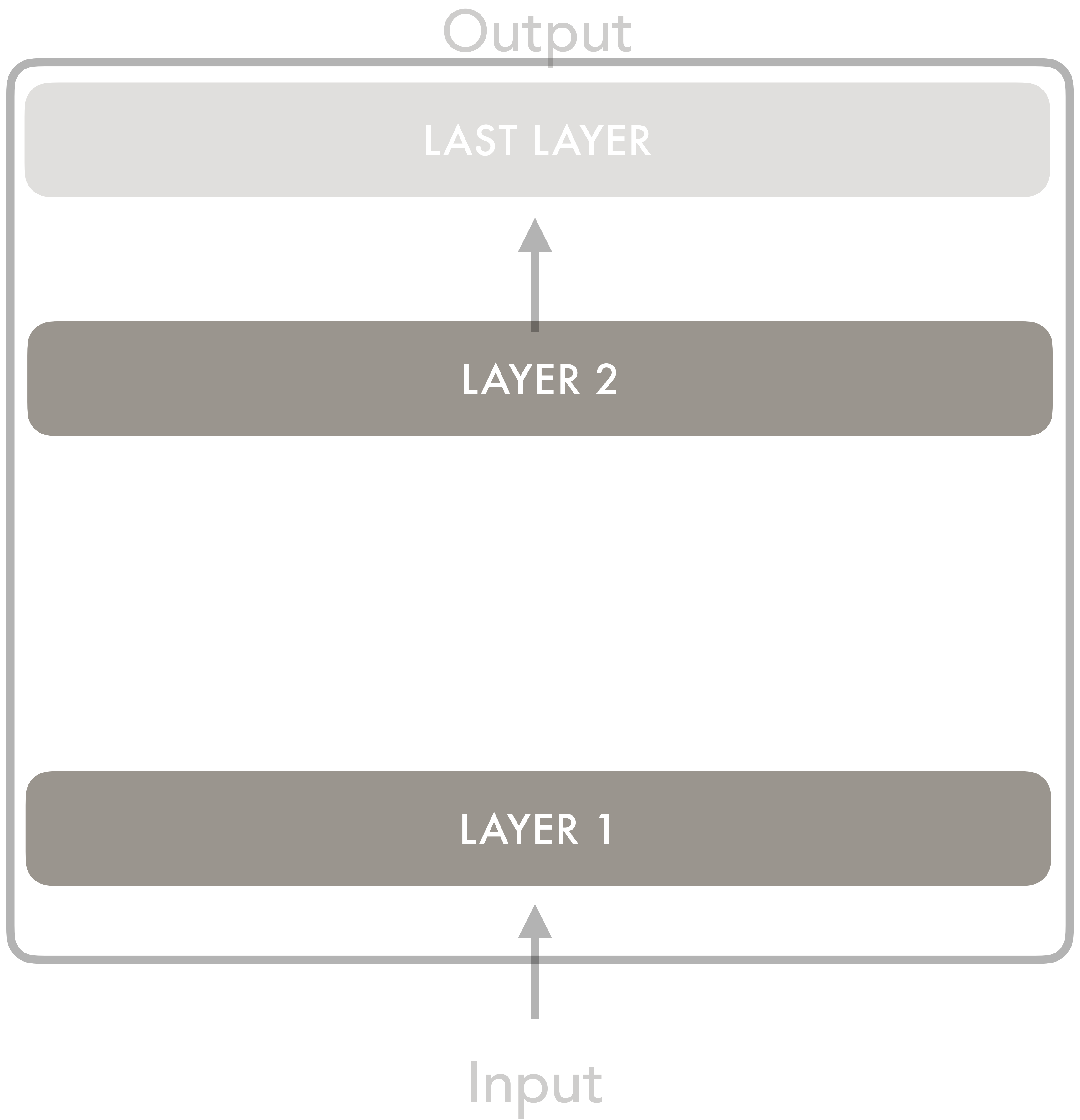
Revisiting Adapters



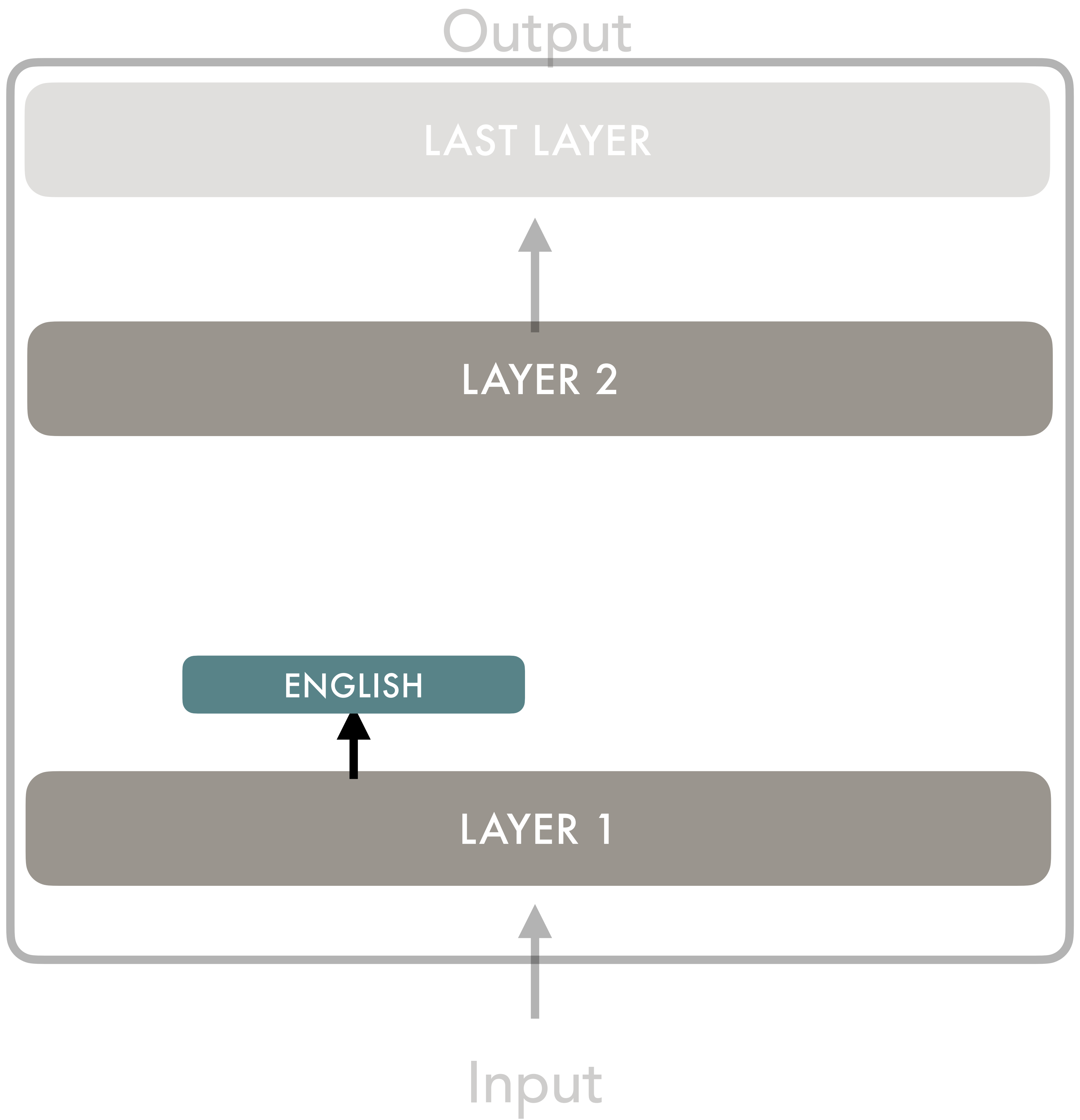
Revisiting Adapters



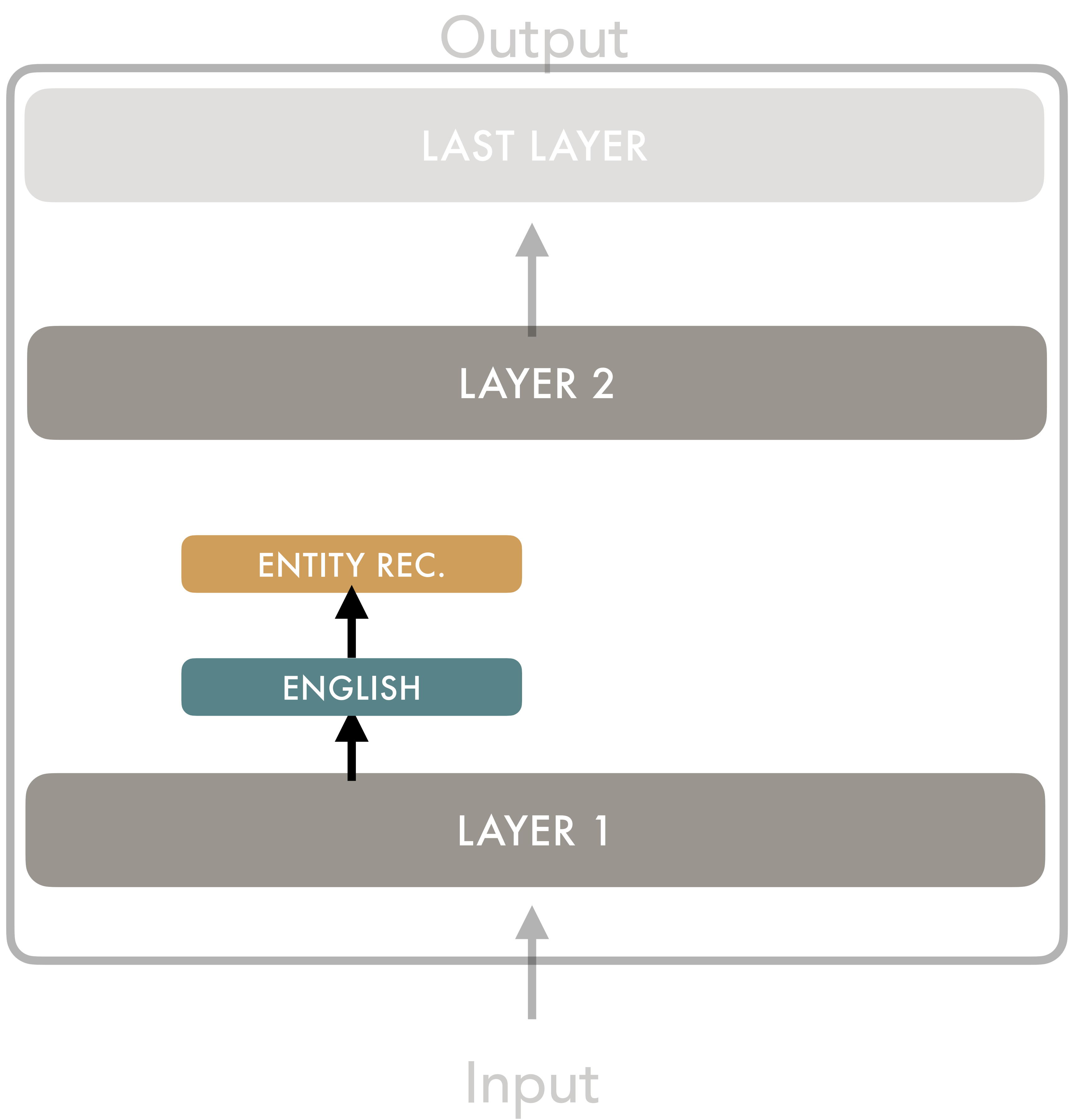
Revisiting Adapters



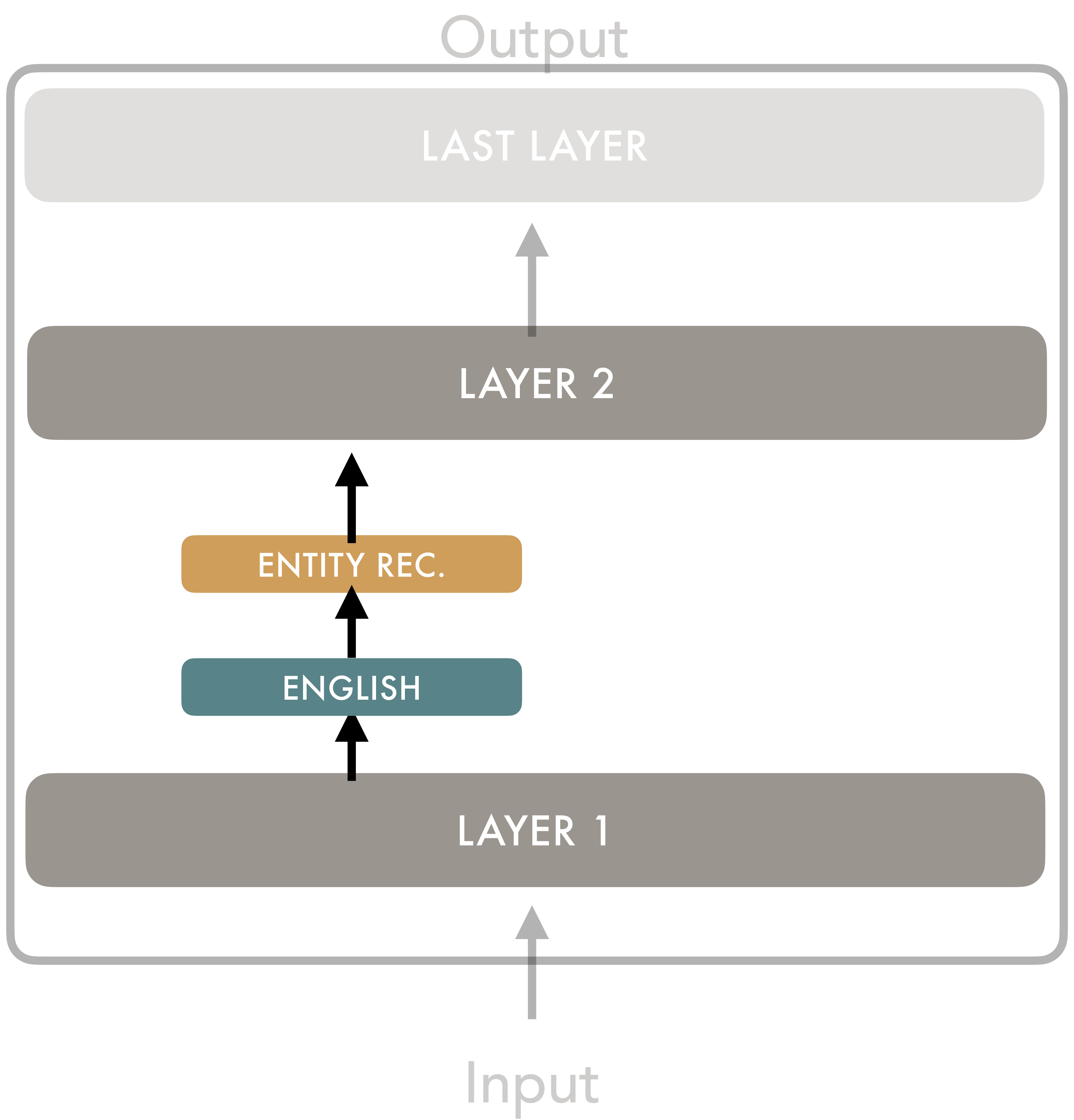
Revisiting Adapters



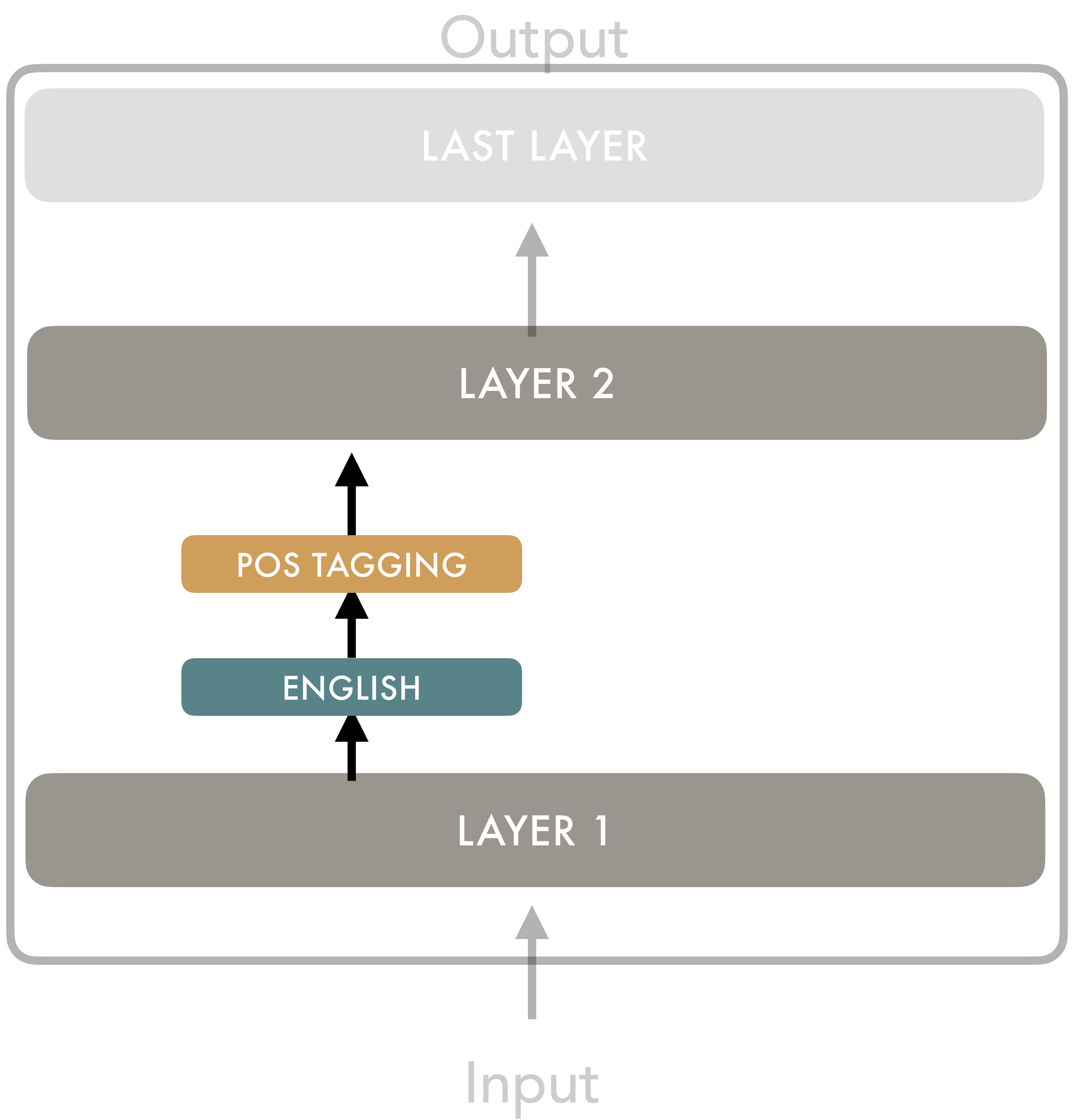
Revisiting Adapters



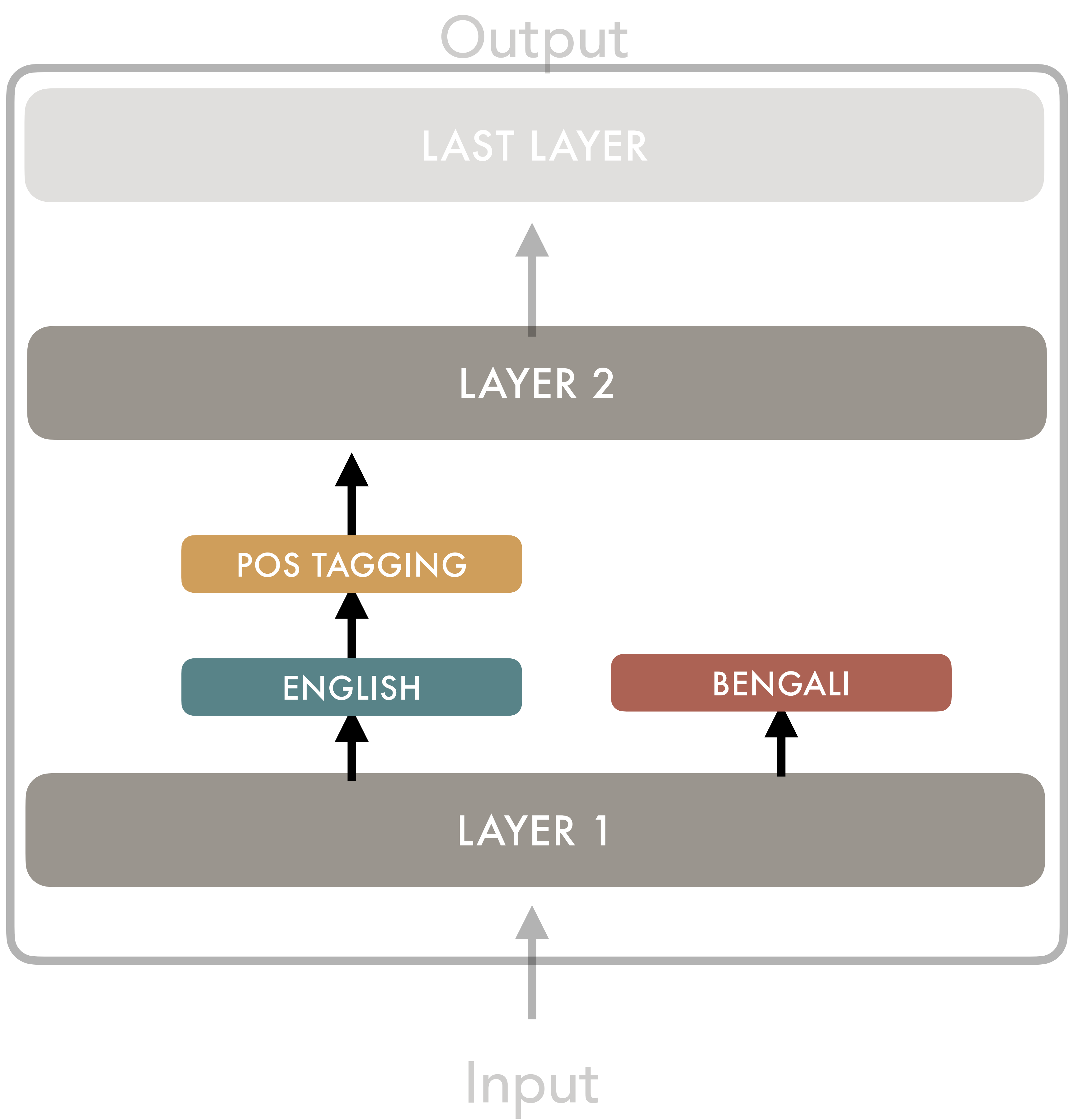
Revisiting Adapters



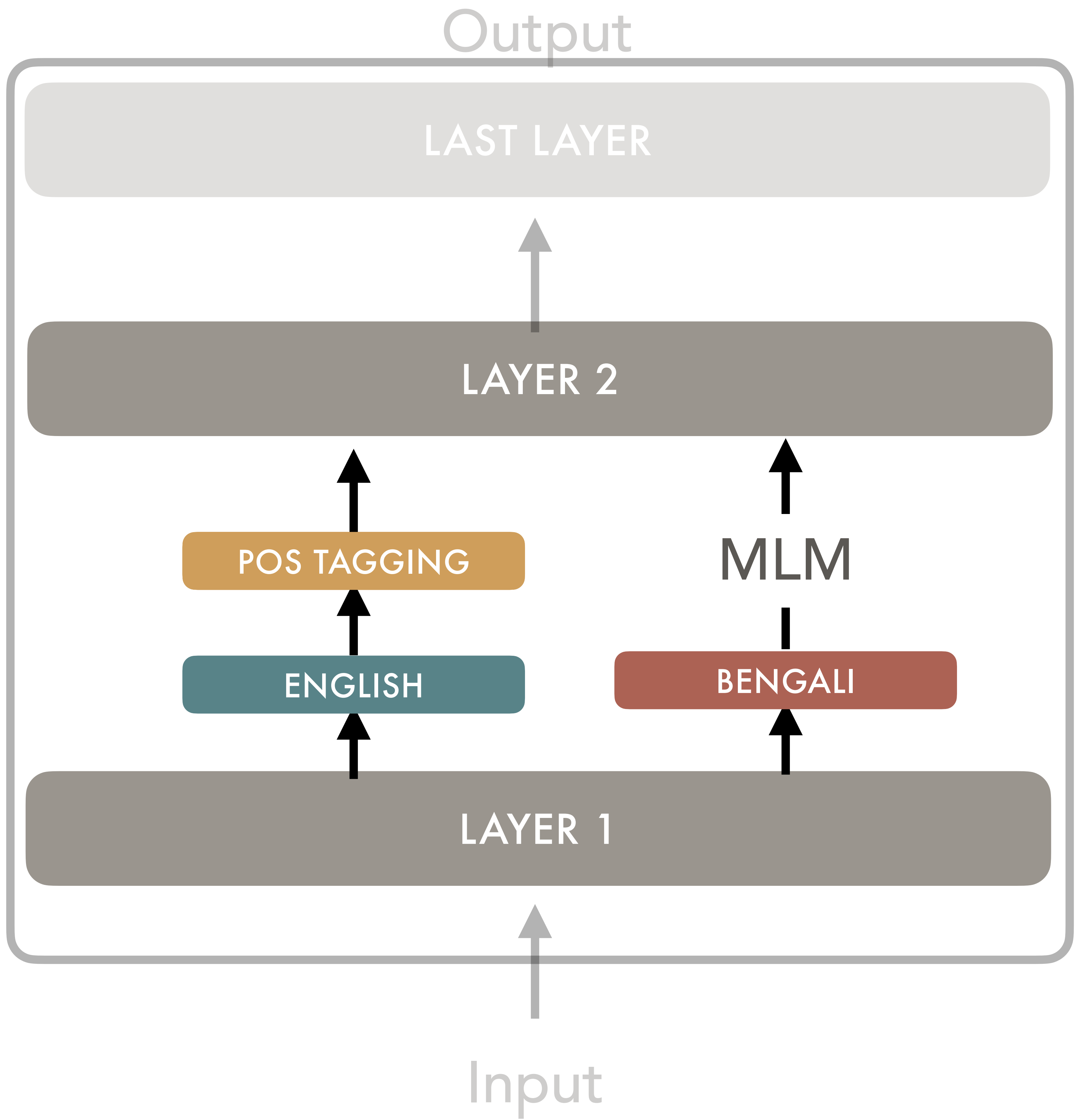
Revisiting Adapters



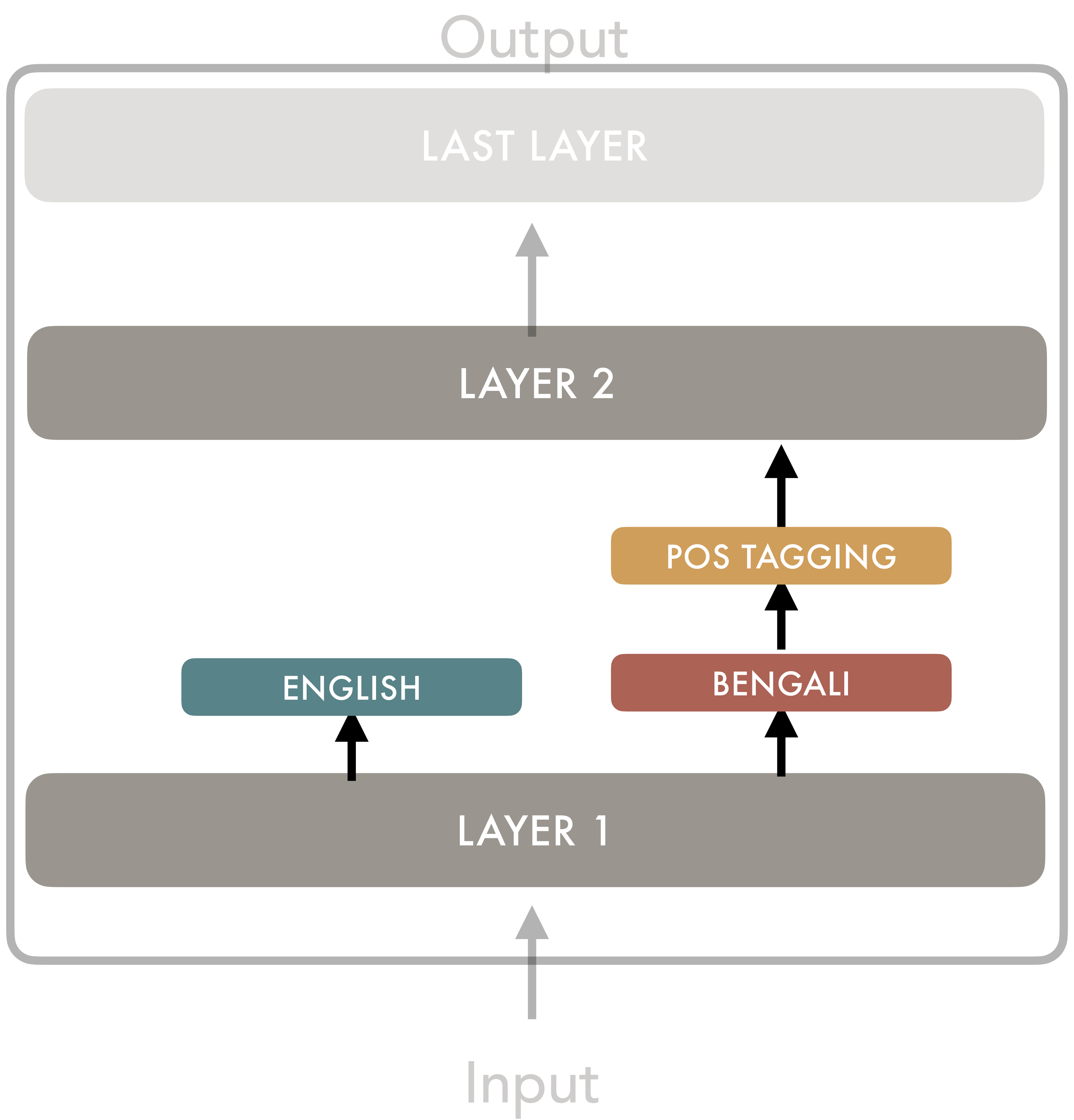
Revisiting Adapters



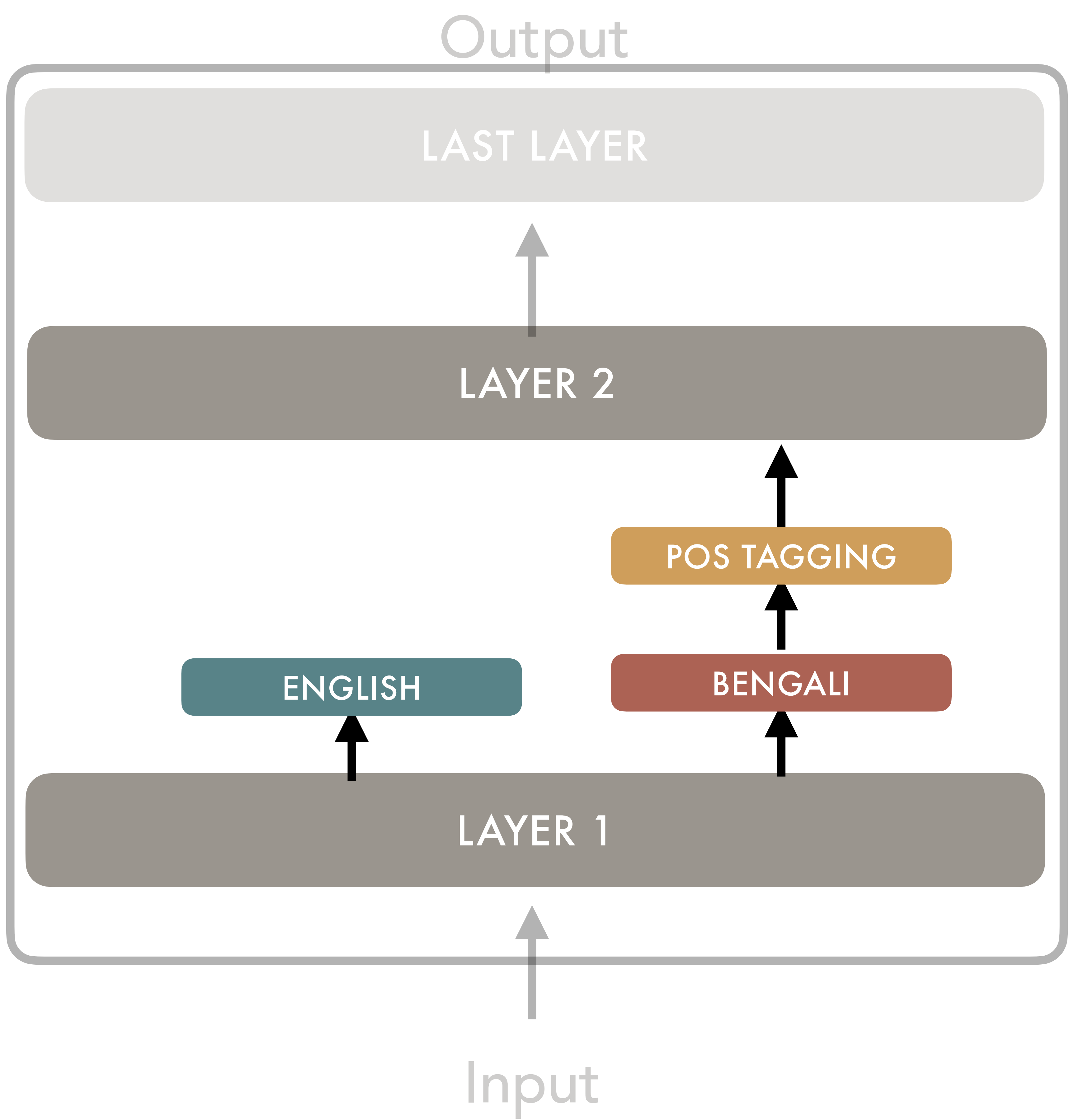
Revisiting Adapters



Revisiting Adapters

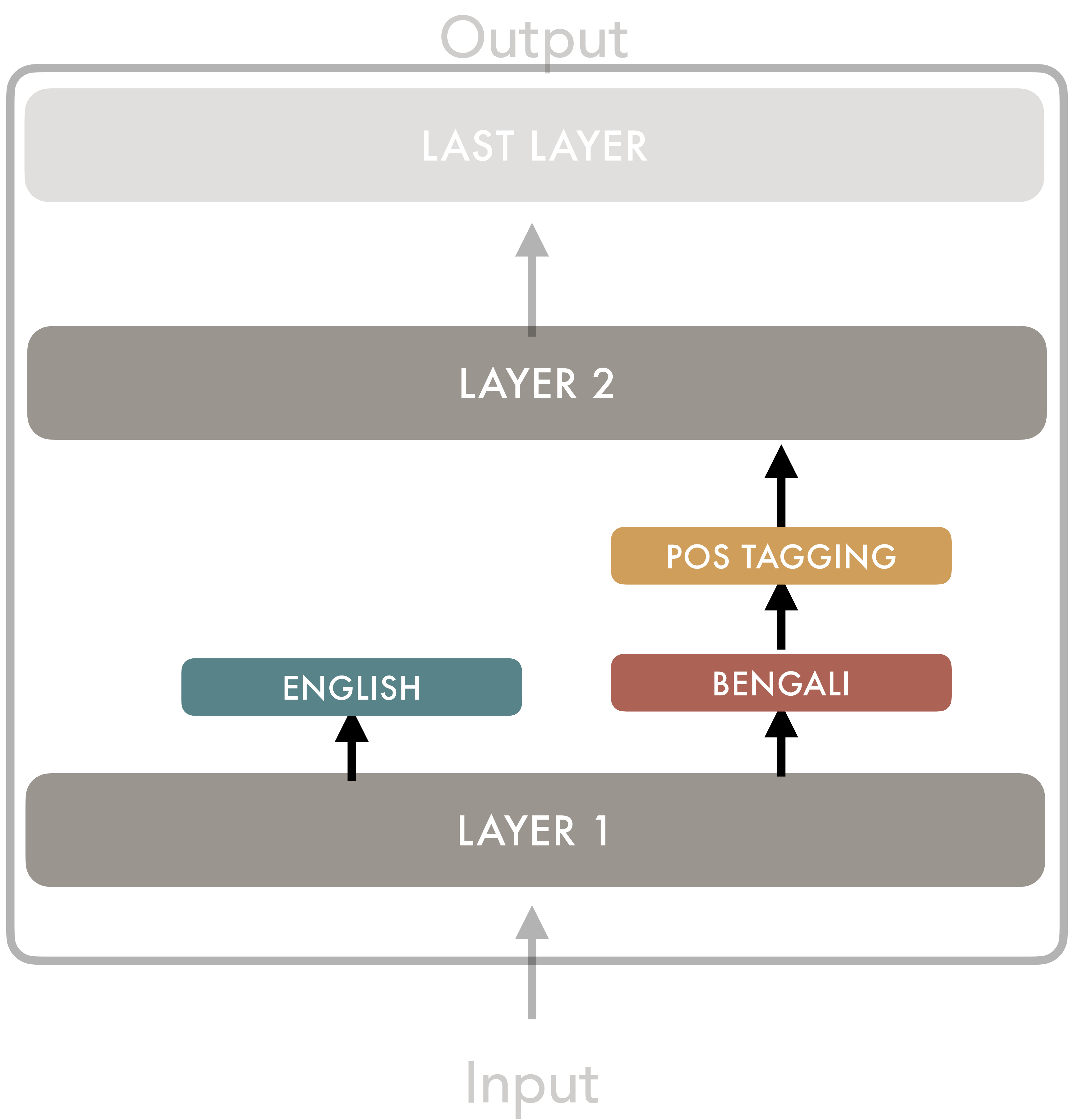


Revisiting Adapters



Easy zero-shot adaptation to new languages at a low cost (additional parameters)

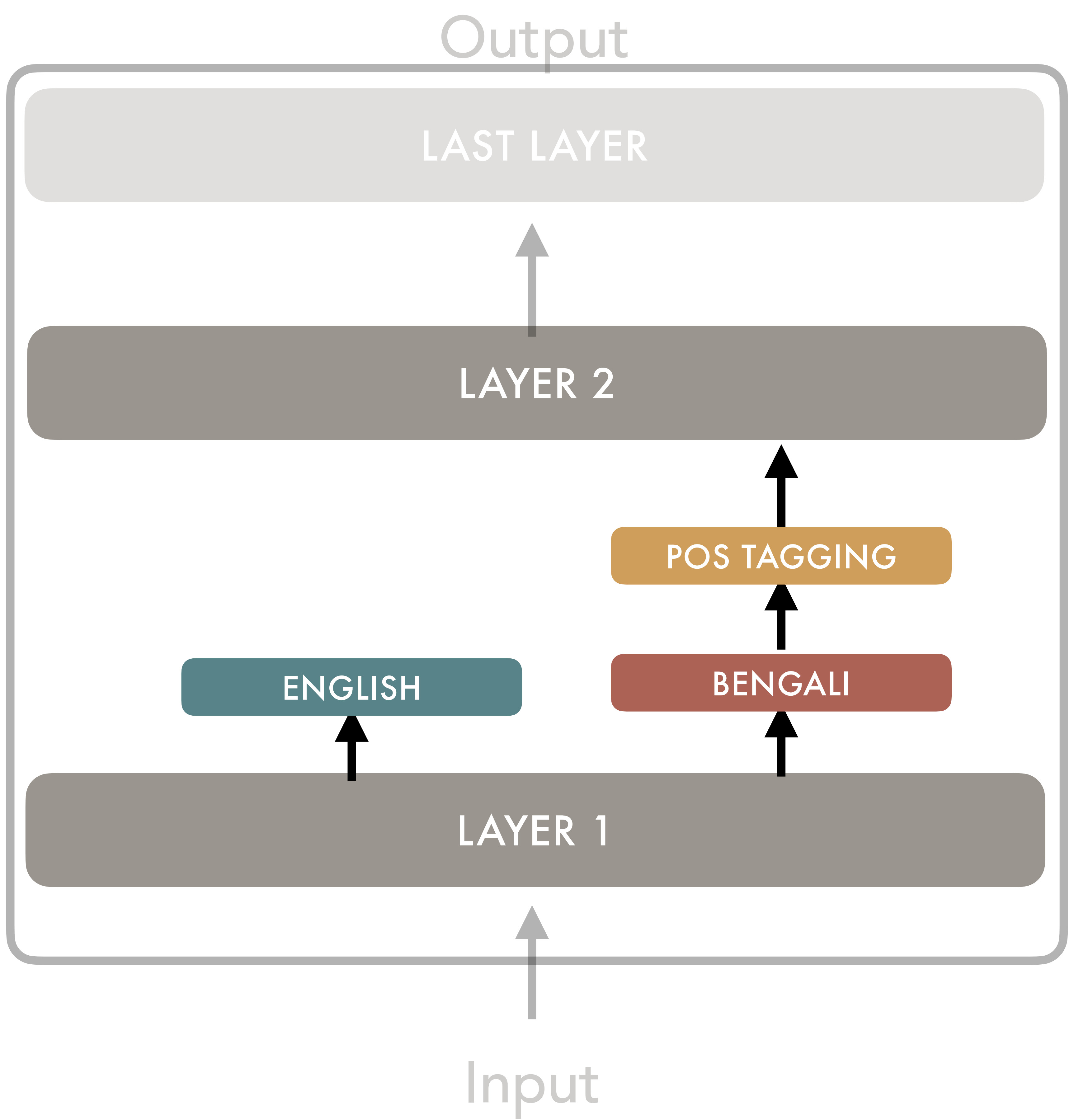
Revisiting Adapters



Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting

Revisiting Adapters

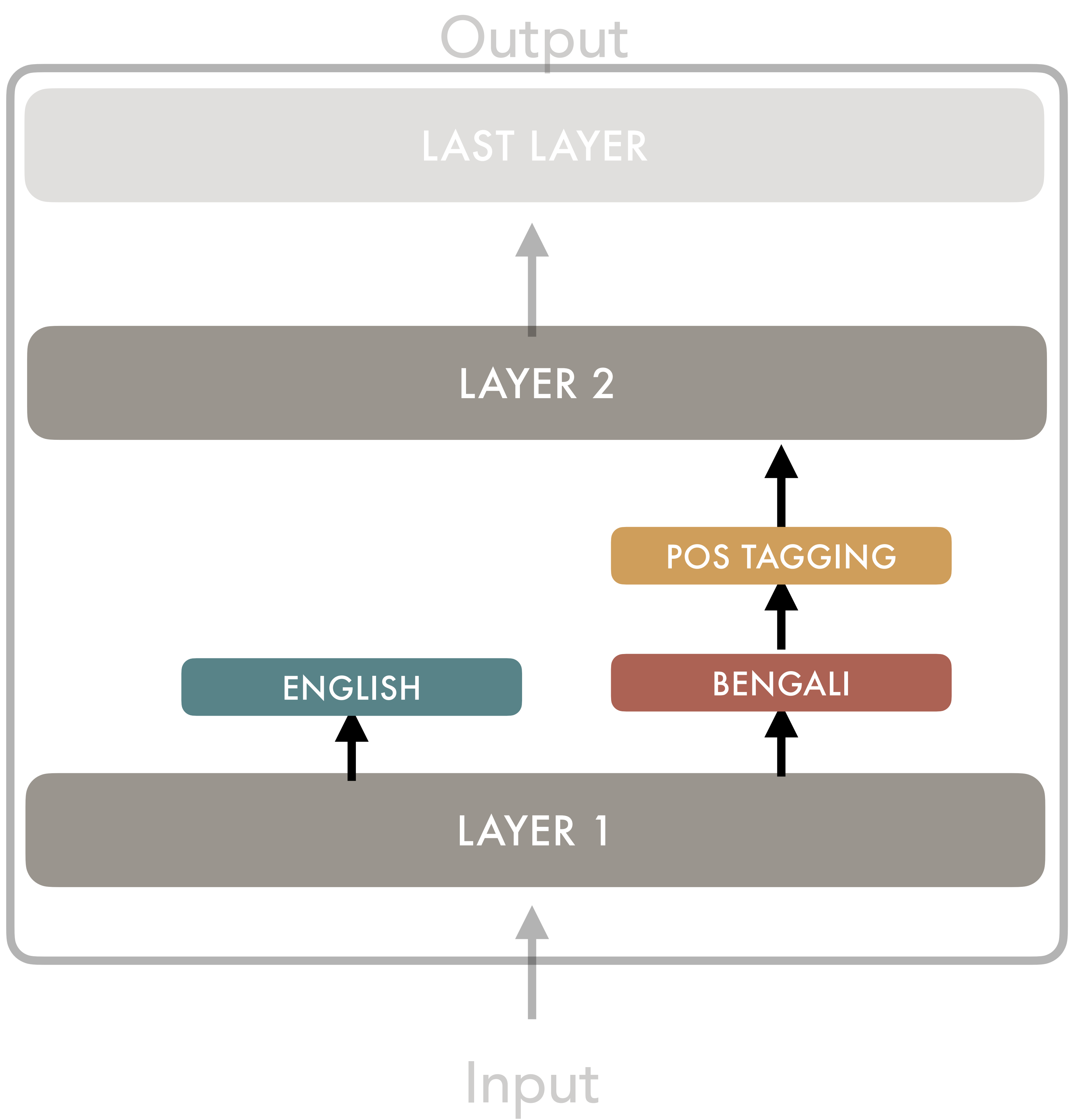


Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting

Performance comparable to full-model fine-tuning

Revisiting Adapters



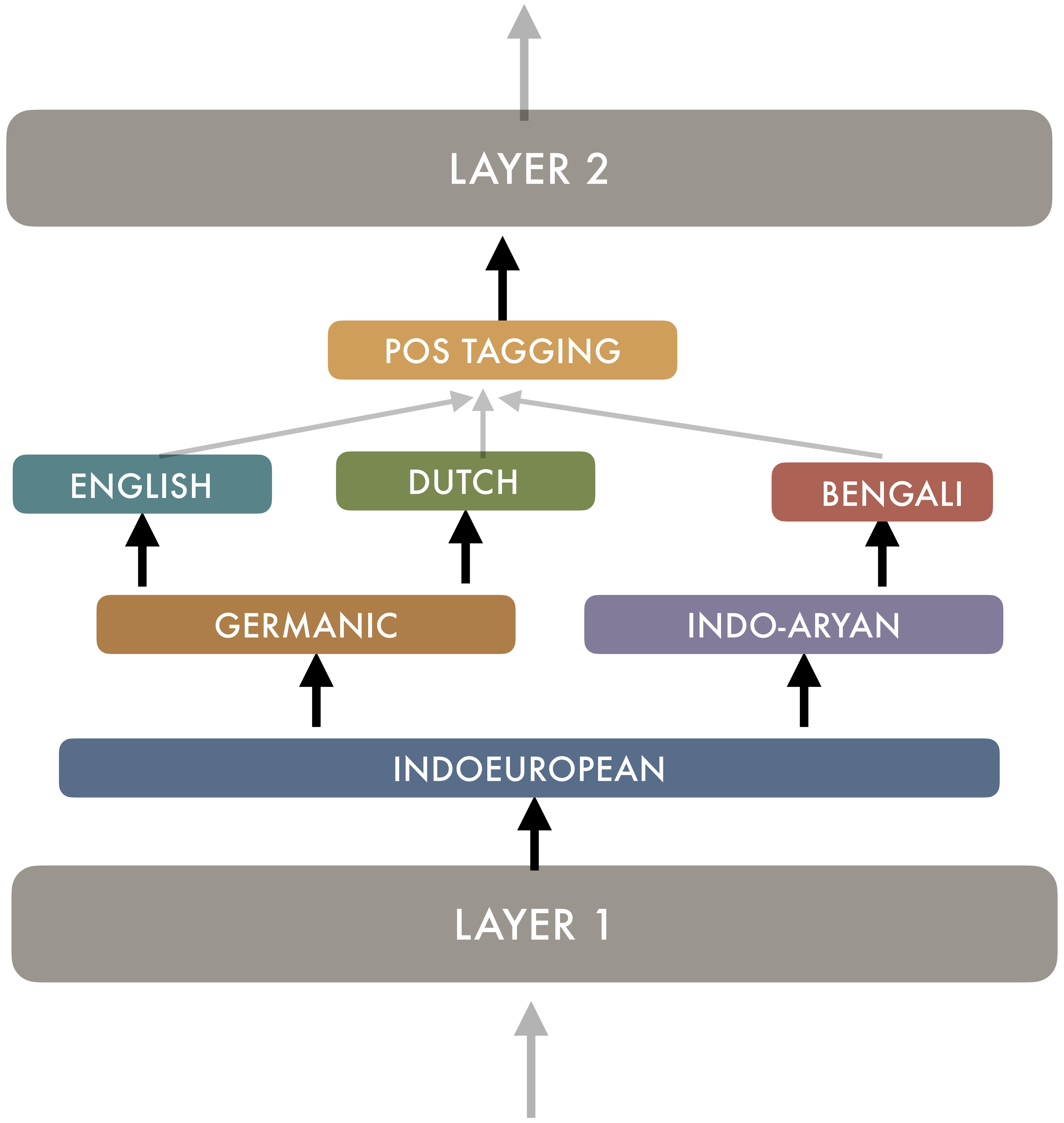
Easy zero-shot adaptation to new languages at a low cost (additional parameters)

Avoids catastrophic forgetting

Performance comparable to full-model fine-tuning

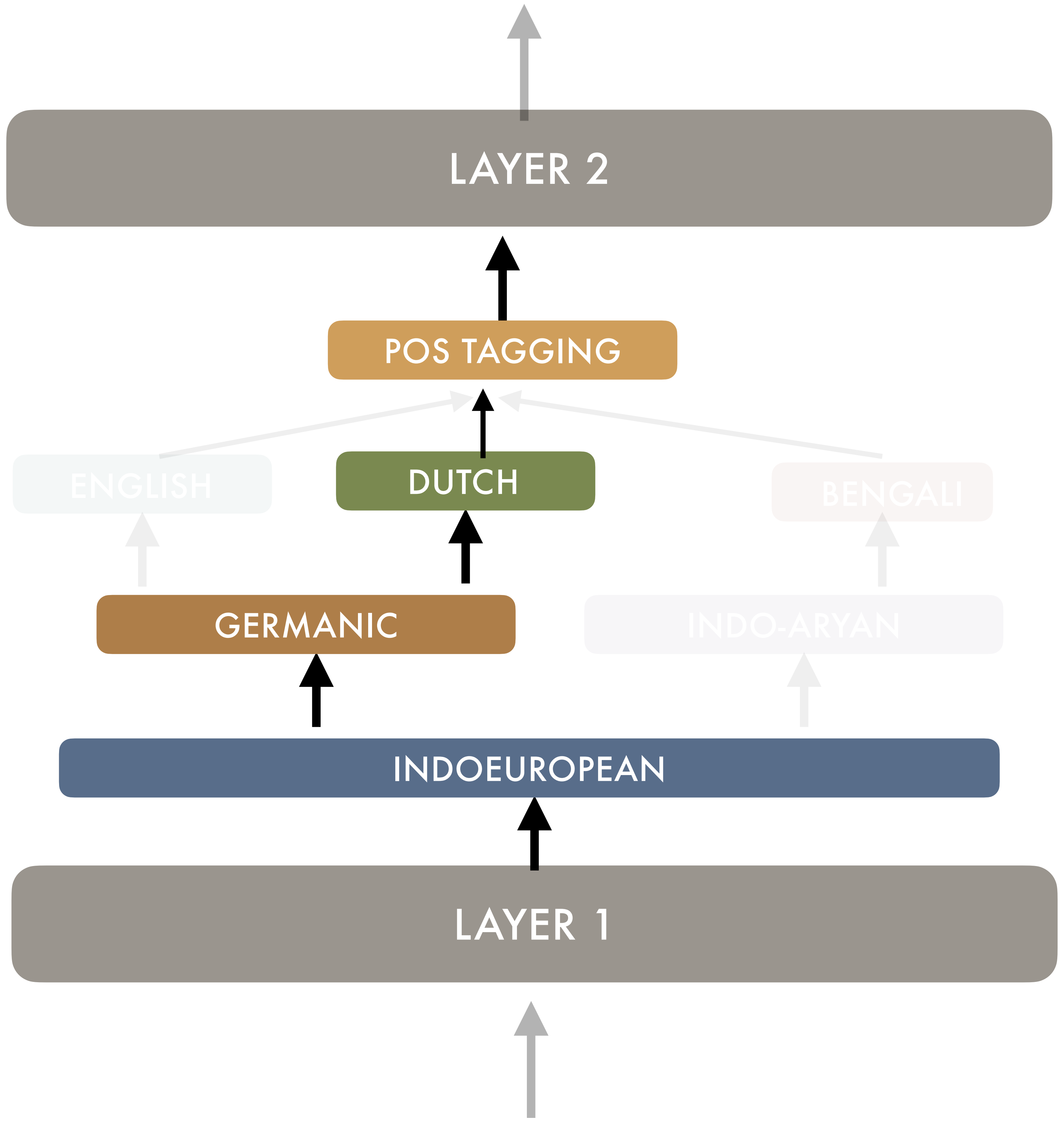
Can we do better?

Follow Phylogeny for Parameter Sharing



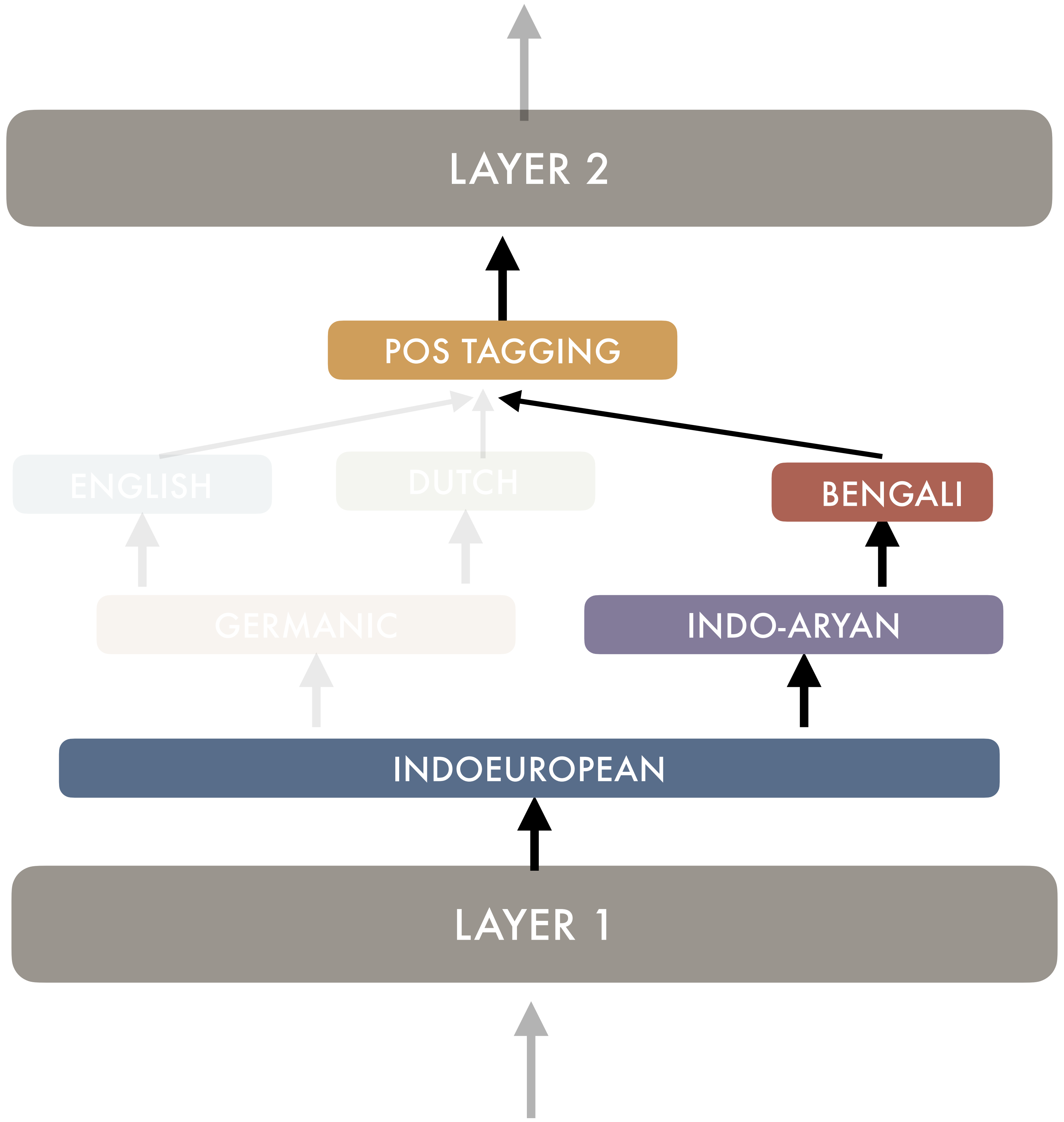
Follow Phylogeny for Parameter Sharing

For Dutch input



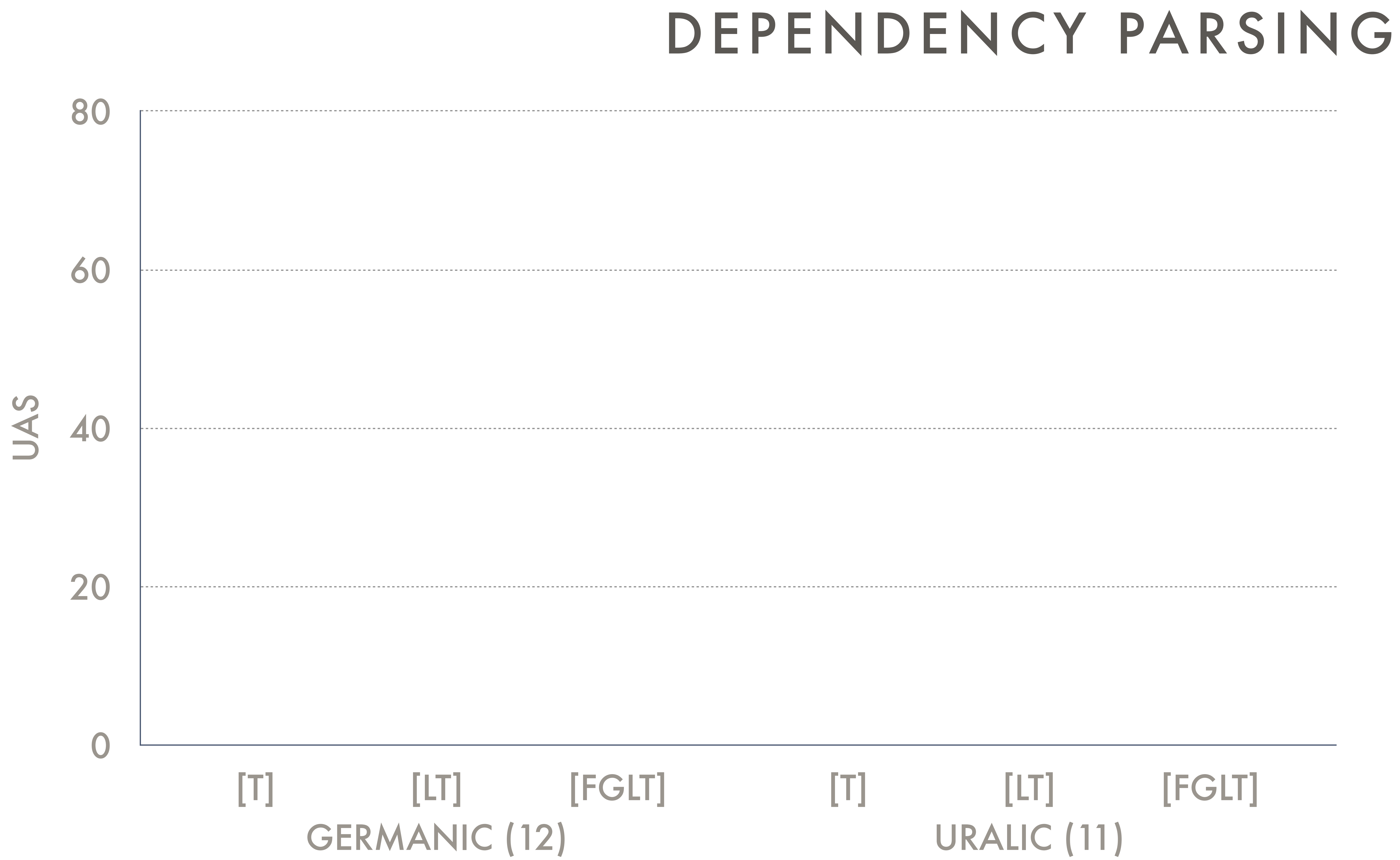
Follow Phylogeny for Parameter Sharing

For Bengali input

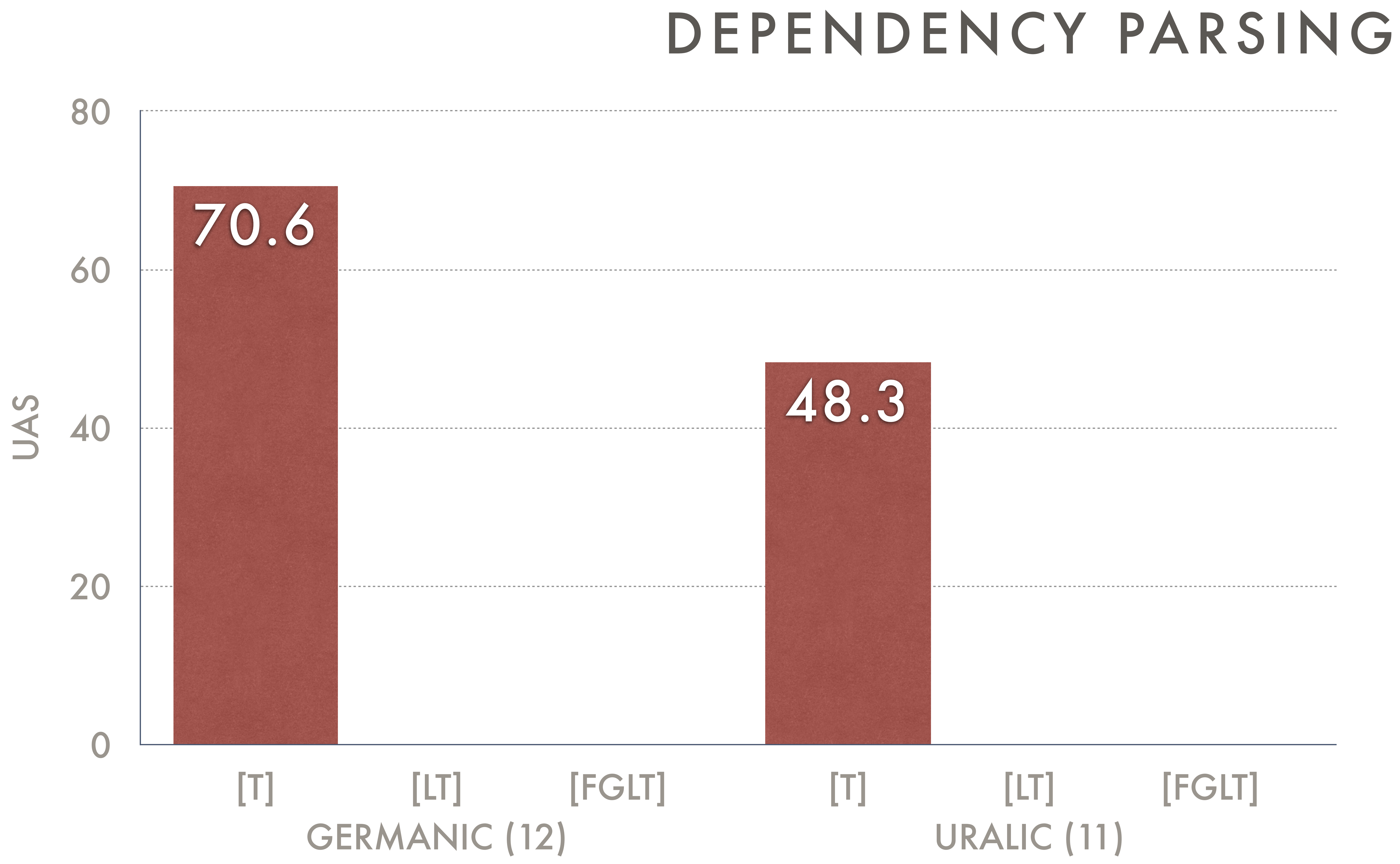


Results

Results

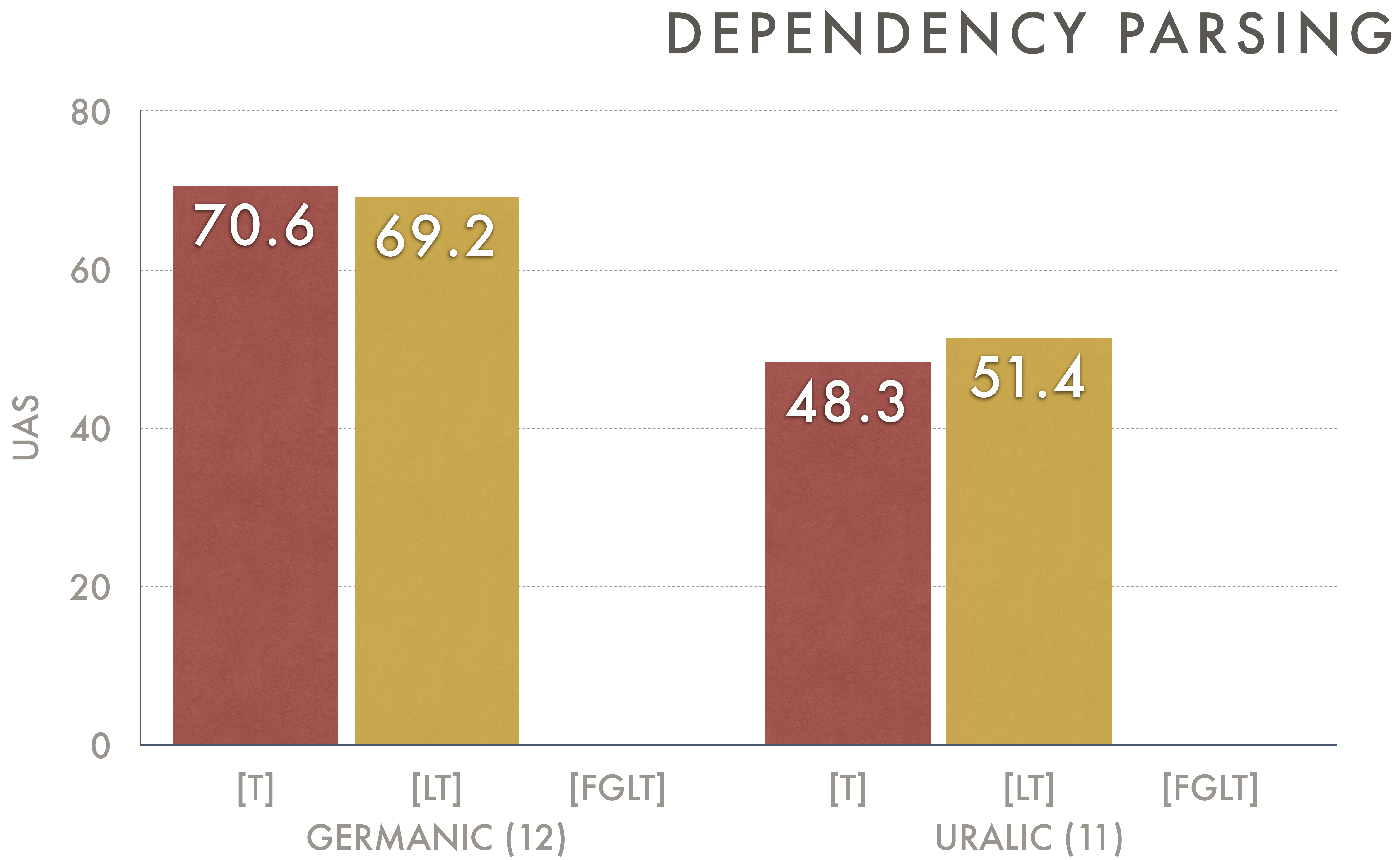

DEPENDENCY PARSING
UAS
80
60
40
20
0
[T]
[LT]
[FGLT]
GERMANIC (12)
[T]
[LT]
[FGLT]
URALIC (11)

Results



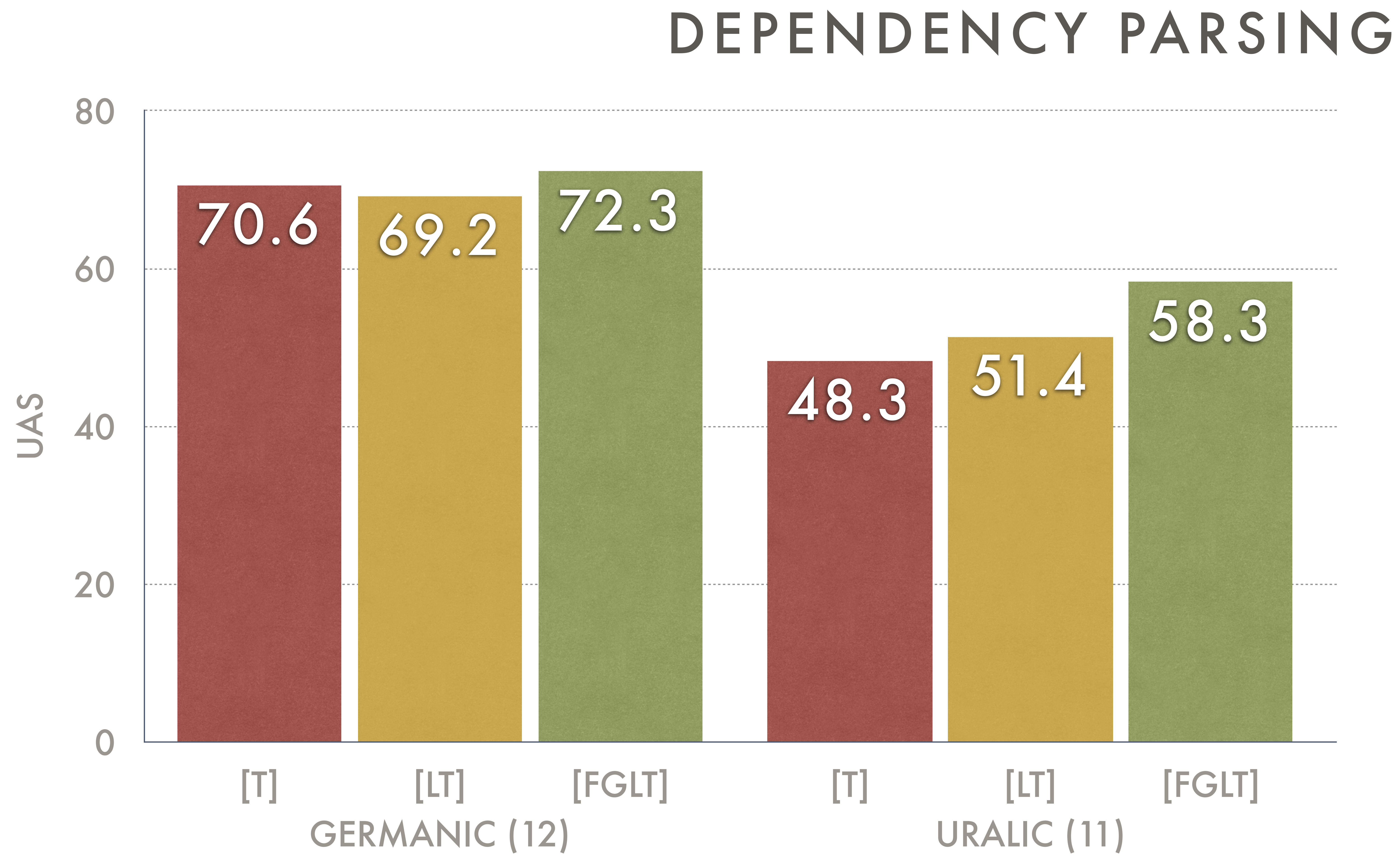
DEPENDENCY PARSING
80
60
40
20
0
UAS
70.6
48.3
[T]
[LT]
[FGLT]
GERMANIC (12)
[T]
[LT]
[FGLT]
URALIC (11)

Results



DEPENDENCY PARSING
80
60
40
20
0
UAS
70.6
69.2
48.3
51.4
[T]
[LT]
[FGLT]
GERMANIC (12)
[T]
[LT]
[FGLT]
URALIC (11)

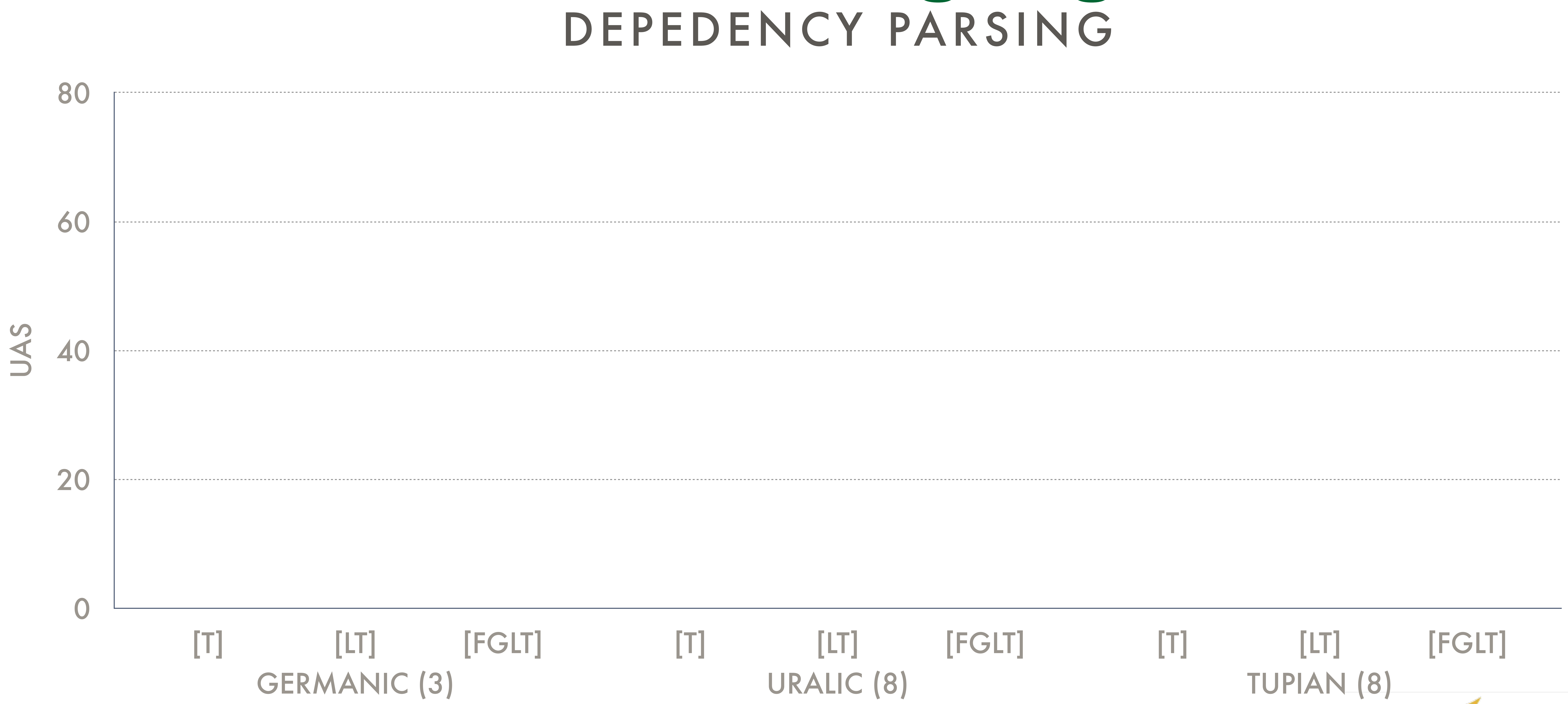
Results



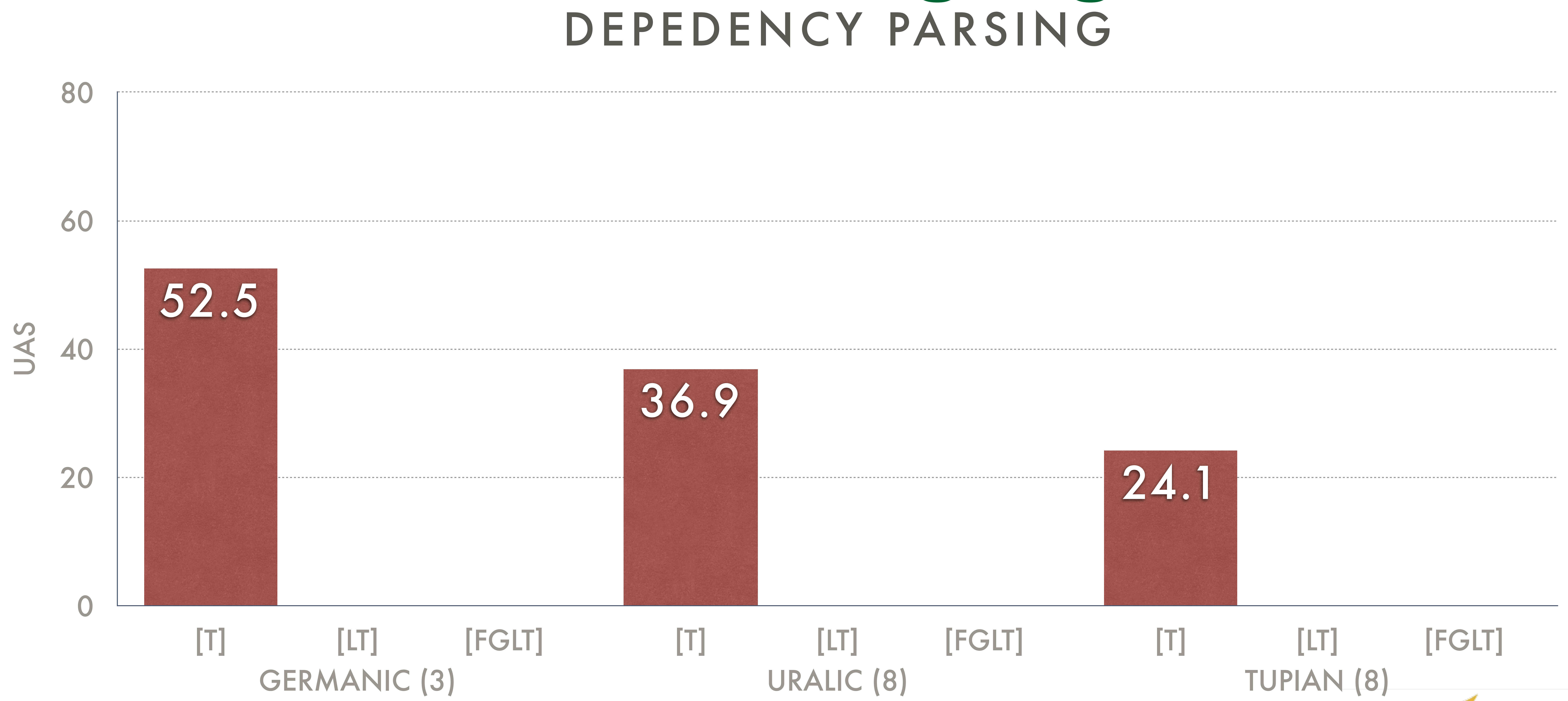
DEPENDENCY PARSING
80
60
40
20
0
UAS
70.6
69.2
72.3
48.3
51.4
58.3
[T]
[LT]
[FGLT]
[T]
[LT]
[FGLT]
GERMANIC (12)
URALIC (11)

Results on unseen languages

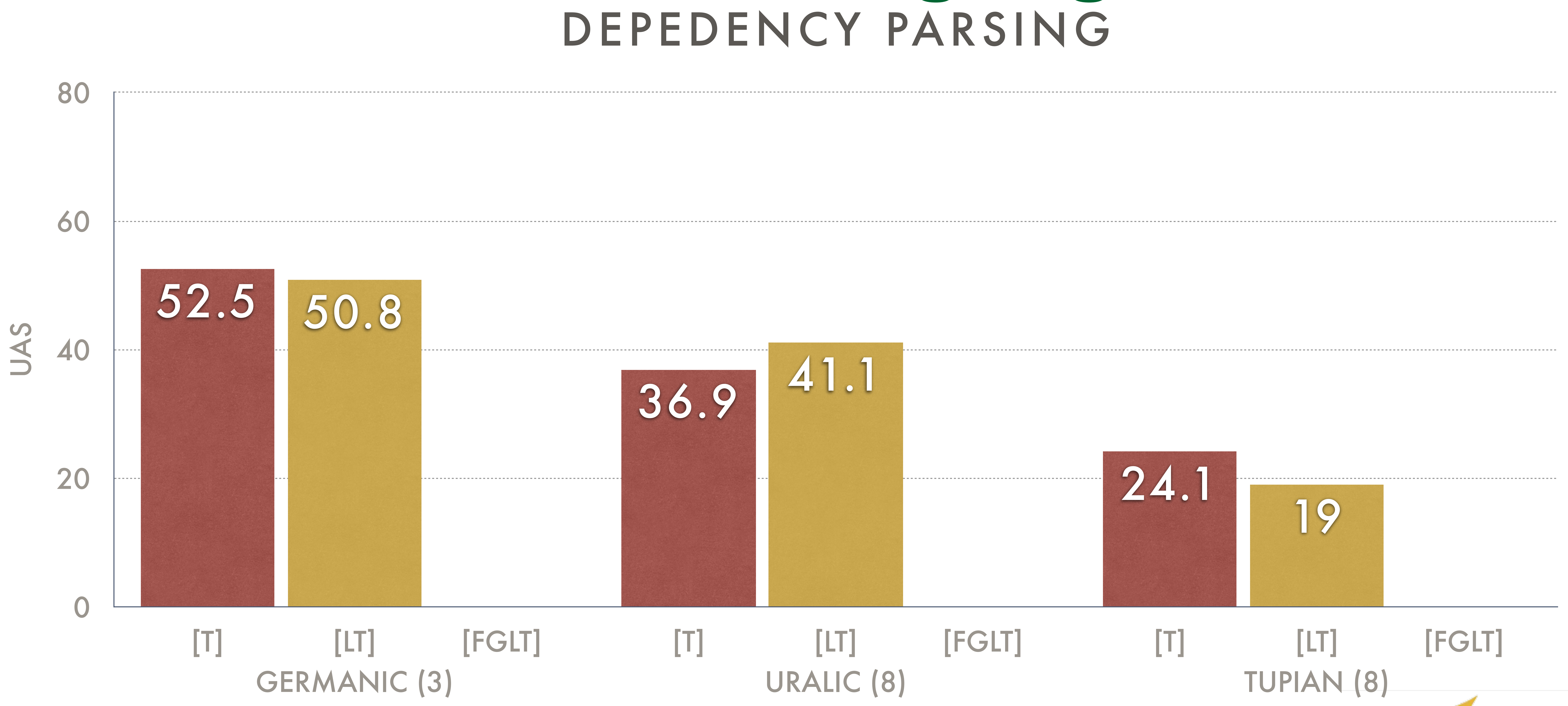
Results on unseen languages



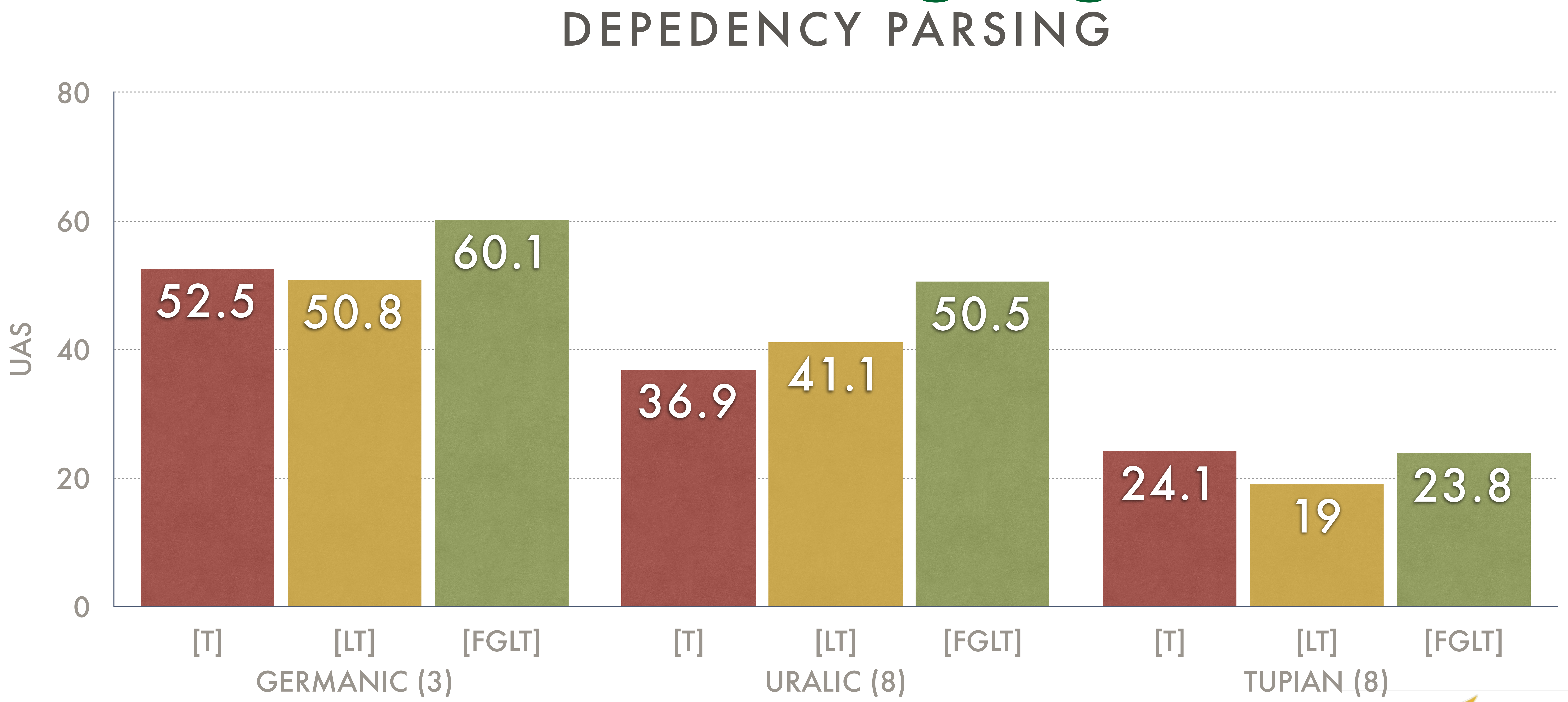
Results on unseen languages



Results on unseen languages

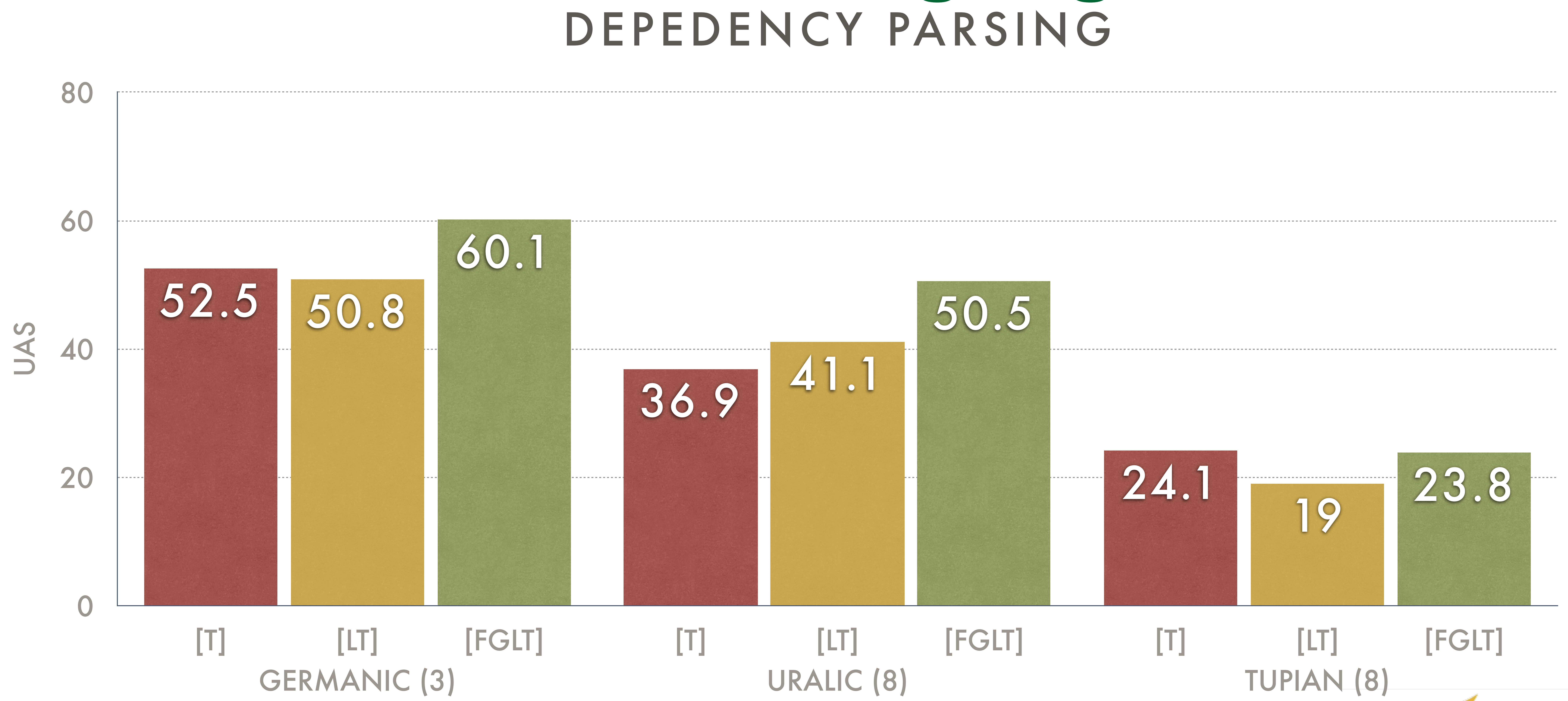


Results on unseen languages



NLP GEORGE MASON

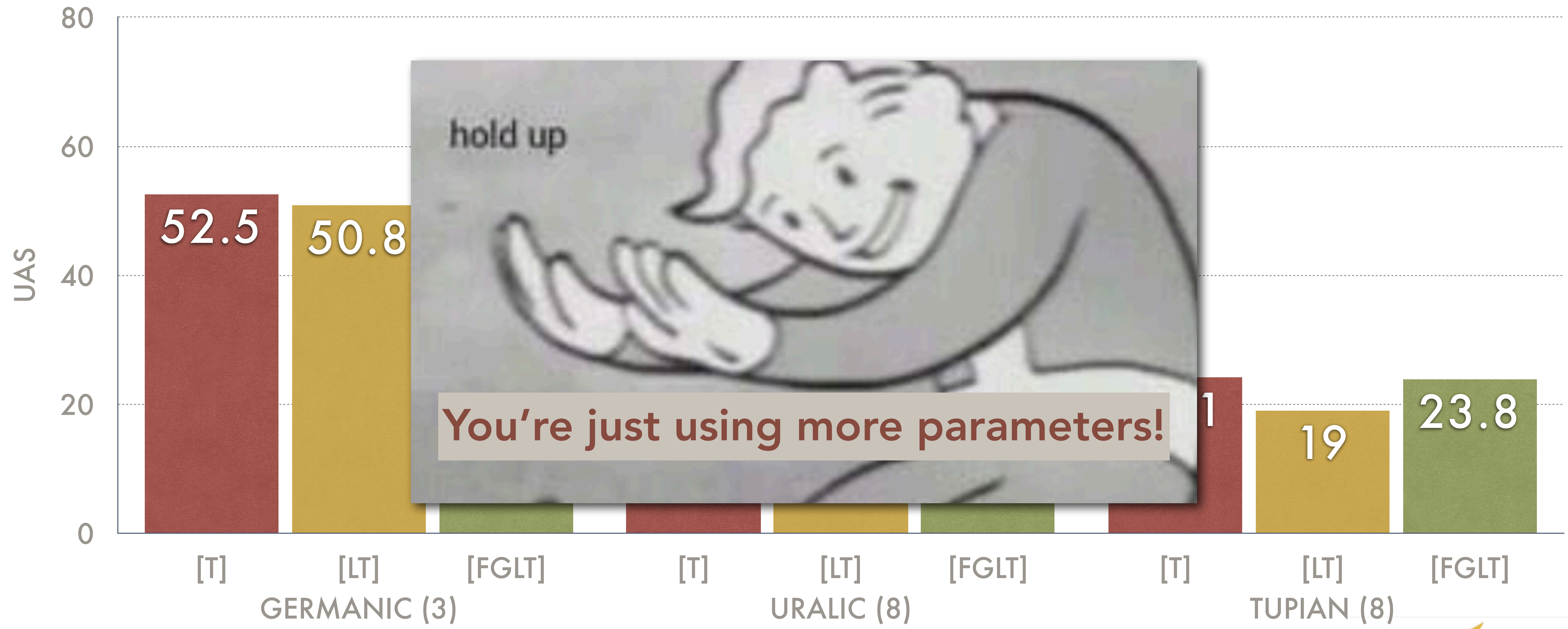
Results on unseen languages



Much larger improvements for *new, unseen* languages

Results on unseen languages

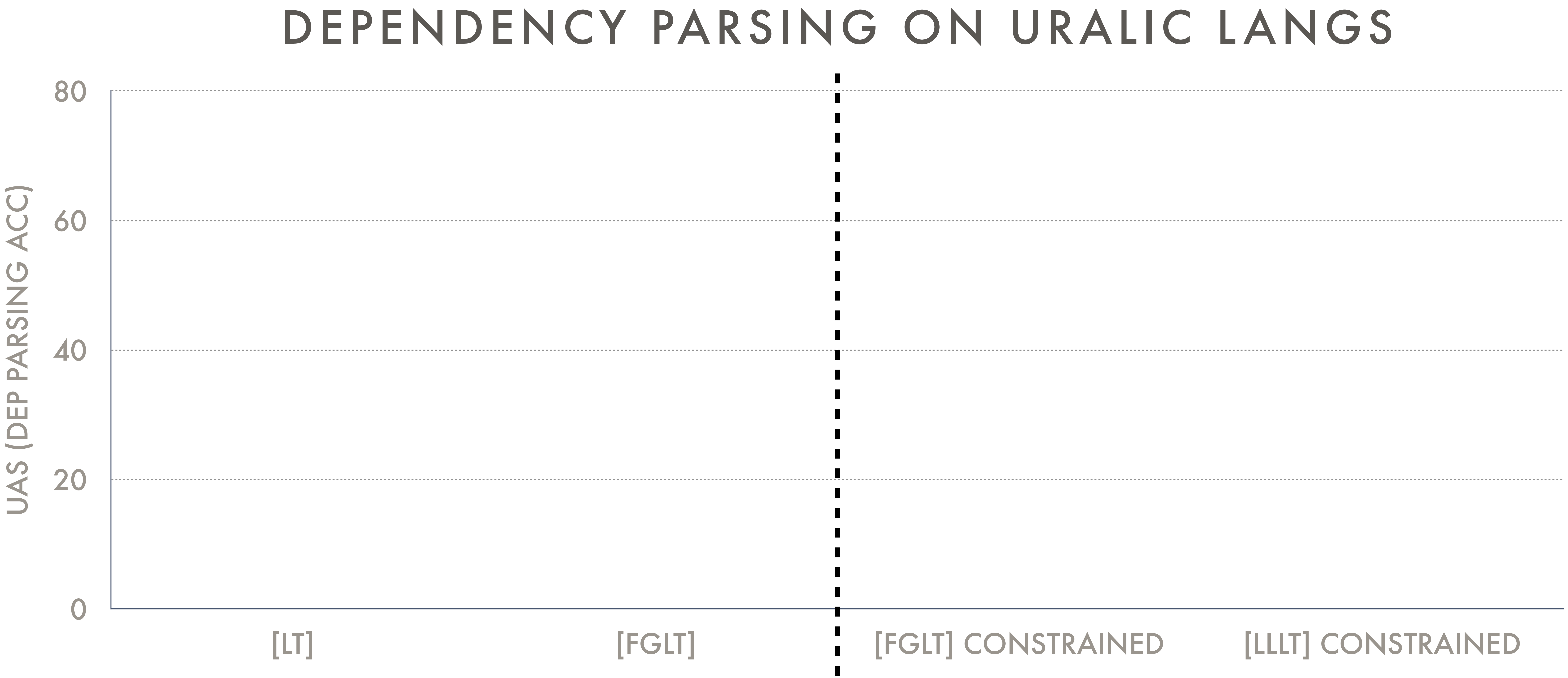
DEPEDENCY PARSING



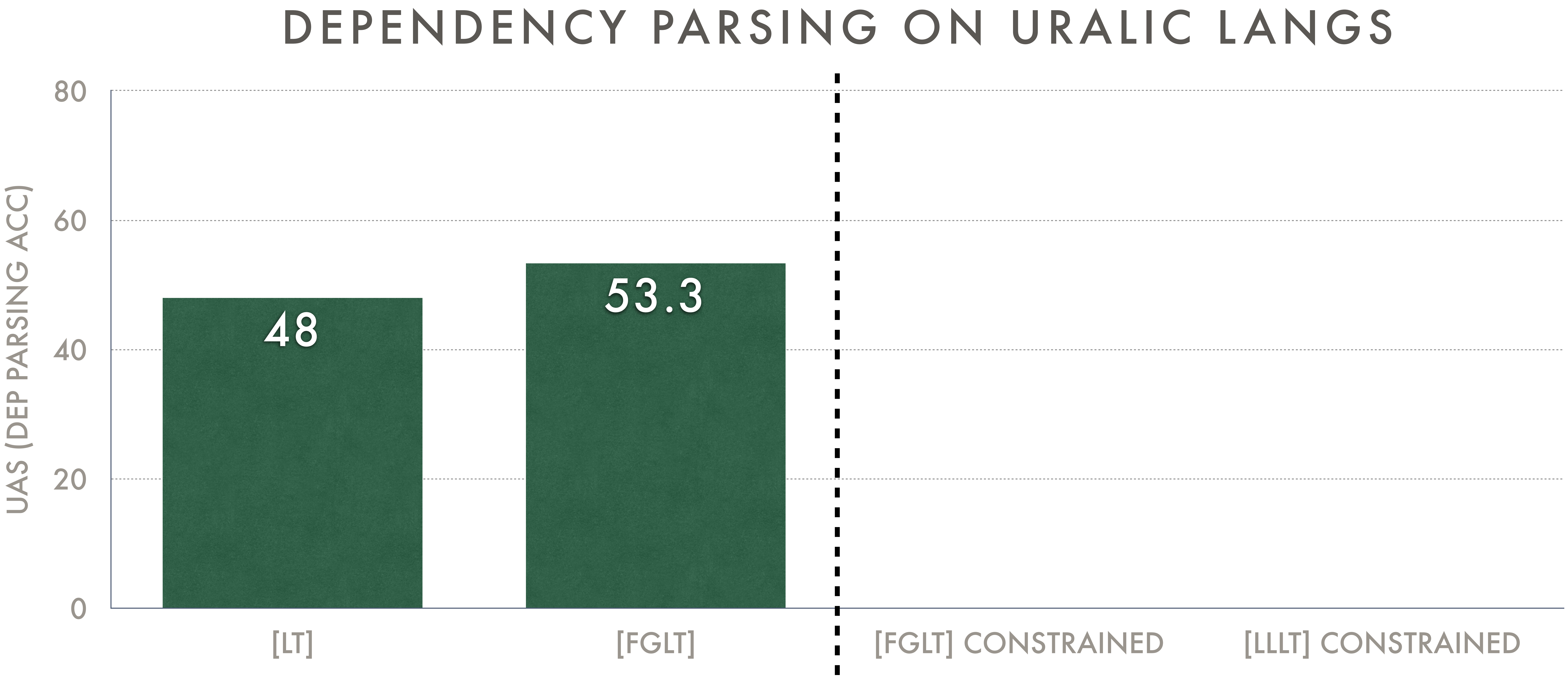
Much larger improvements for *new, unseen* languages

Ablations

Ablations

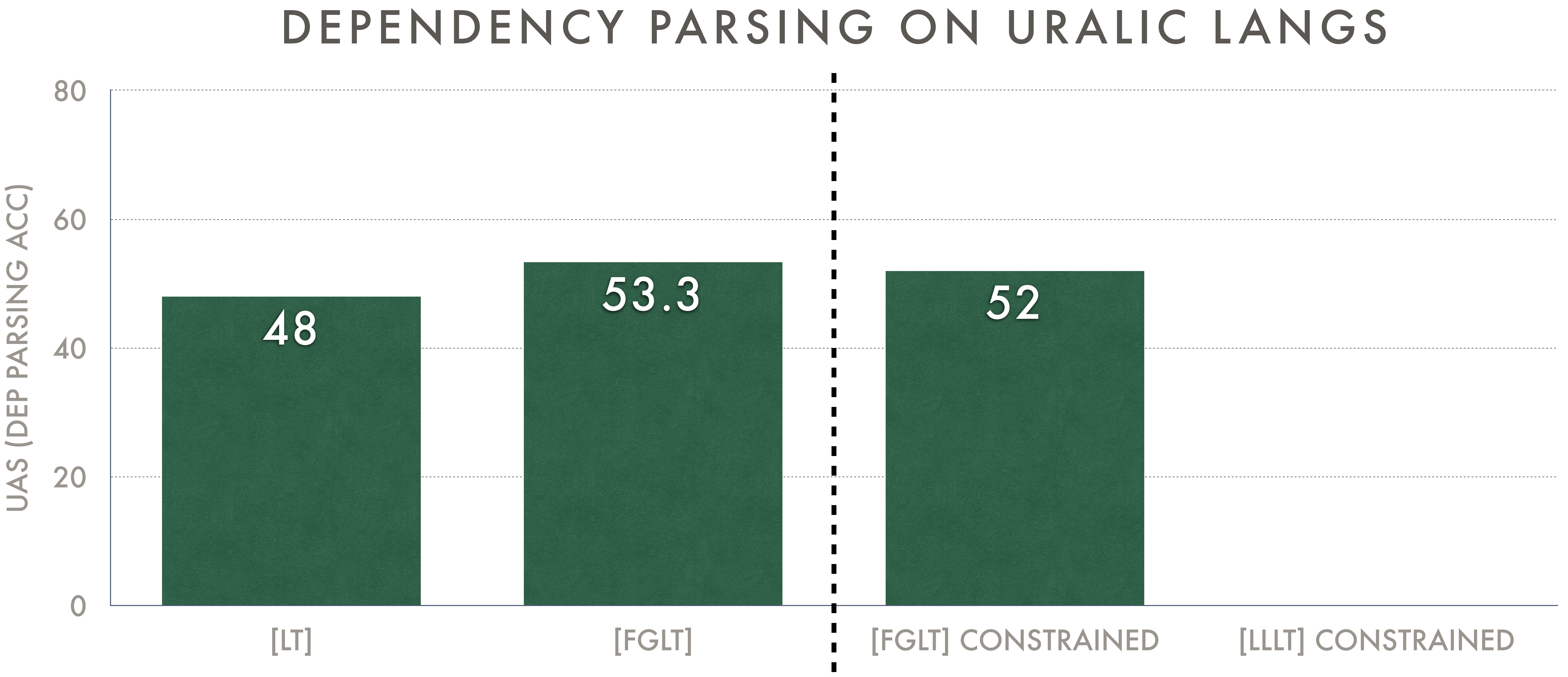

DEPENDENCY PARSING ON URALIC LANGS
UAS (DEP PARSING ACC)
80
60
40
20
0
[LT]
[FGLT]
[FGLT] CONSTRAINED
[LLLT] CONSTRAINED

Ablations



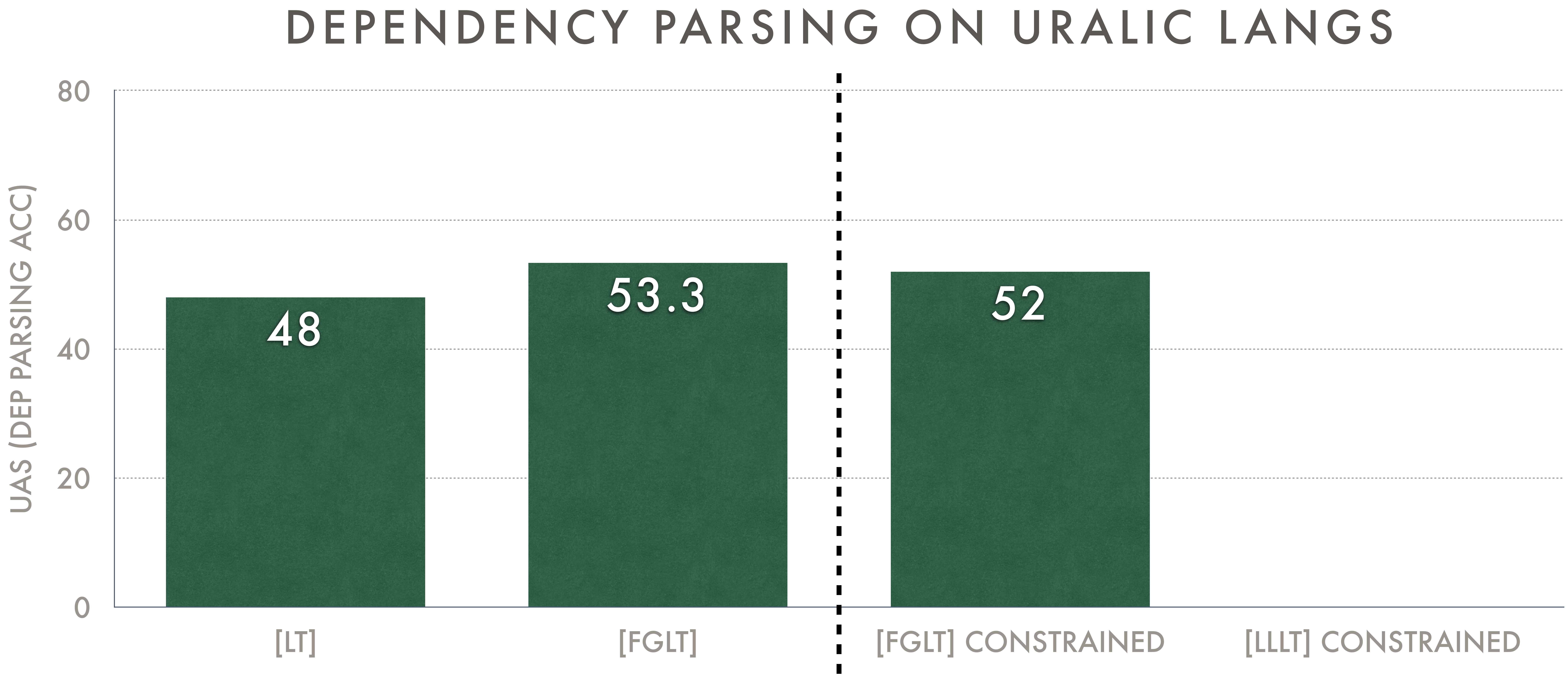
DEPENDENCY PARSING ON URALIC LANGS
UAS (DEP PARSING ACC)
80
60
40
20
0
48
53.3
[LT]
[FGLT]
[FGLT] CONSTRAINED
[LLLT] CONSTRAINED

Ablations



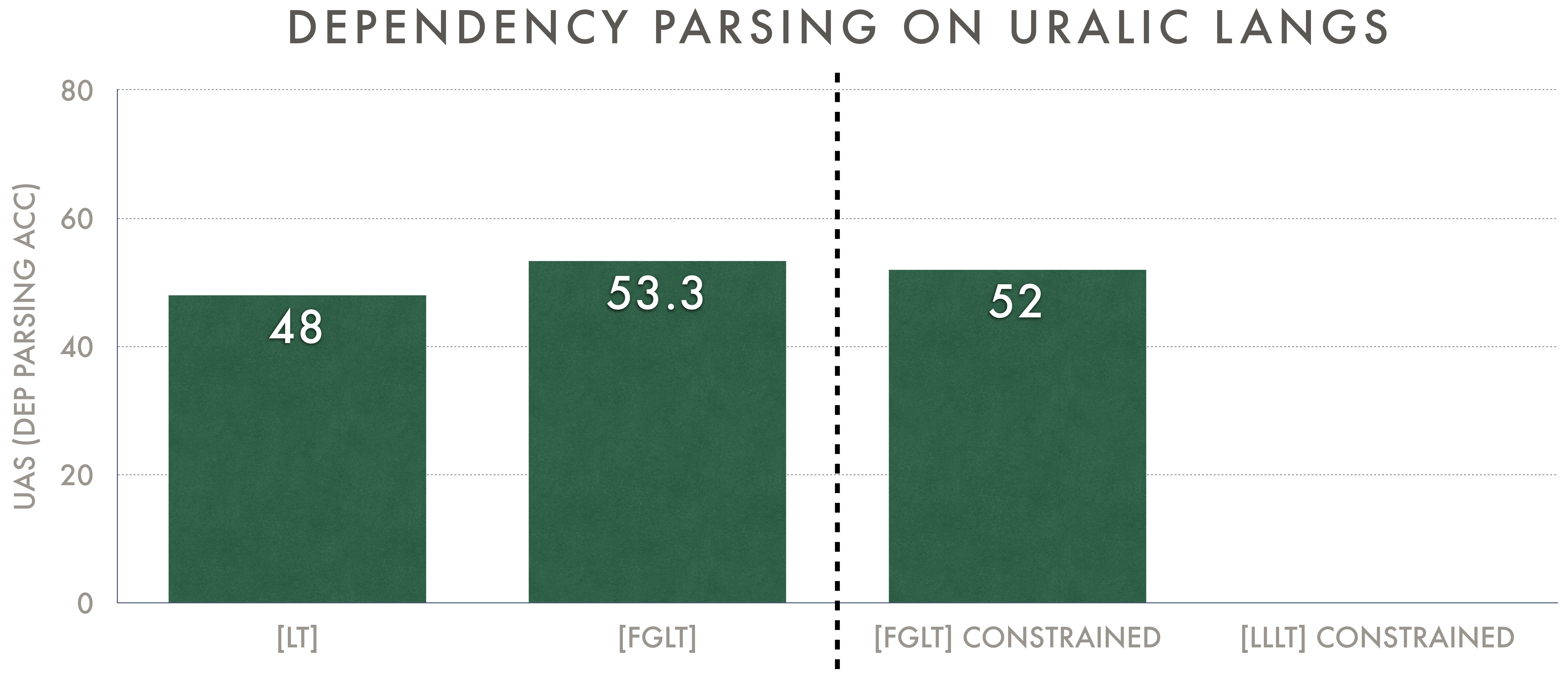
DEPENDENCY PARSING ON URALIC LANGS
UAS (DEP PARSING ACC)
80
60
40
20
0
48
53.3
52
[LT]
[FGLT]
[FGLT] CONSTRAINED
[LLLT] CONSTRAINED

Ablations



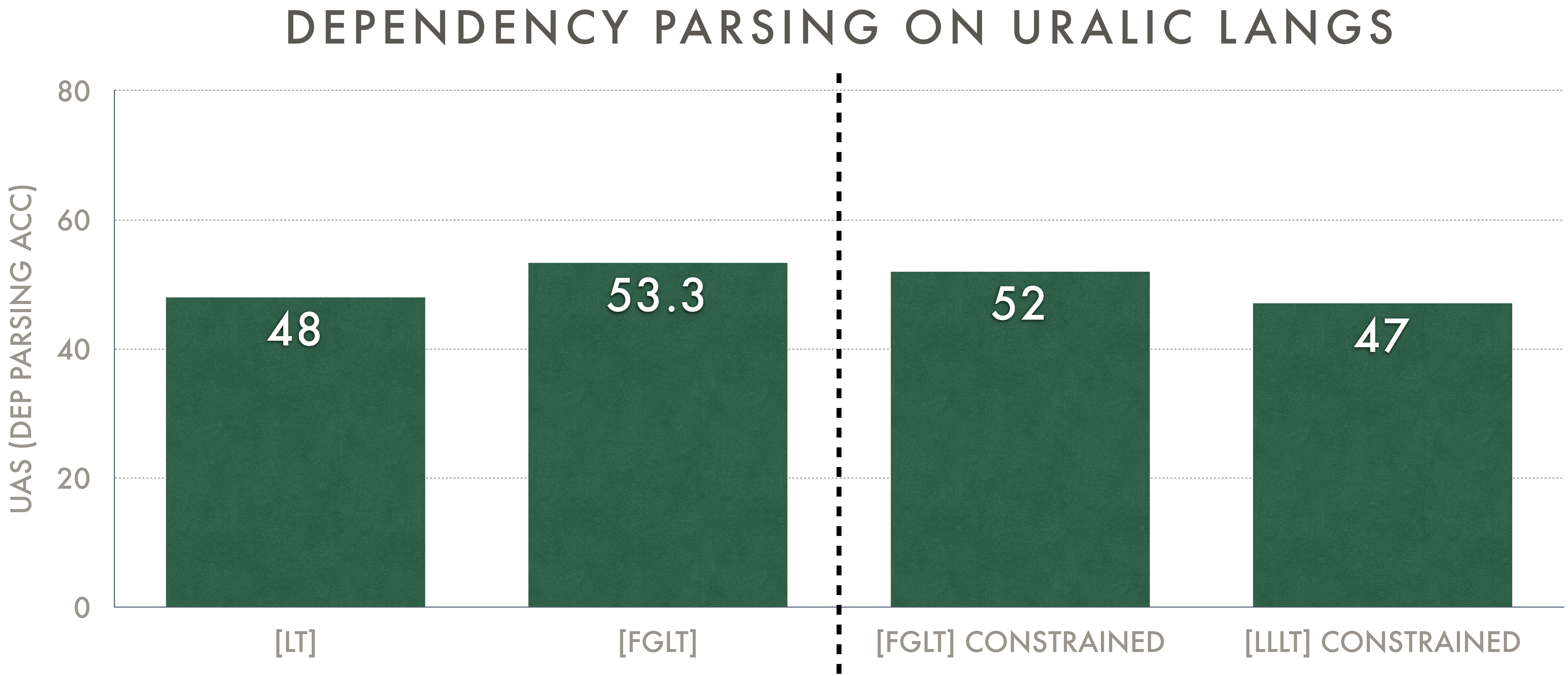
Even constraining to the same number of parameters, still improvements!

Ablations

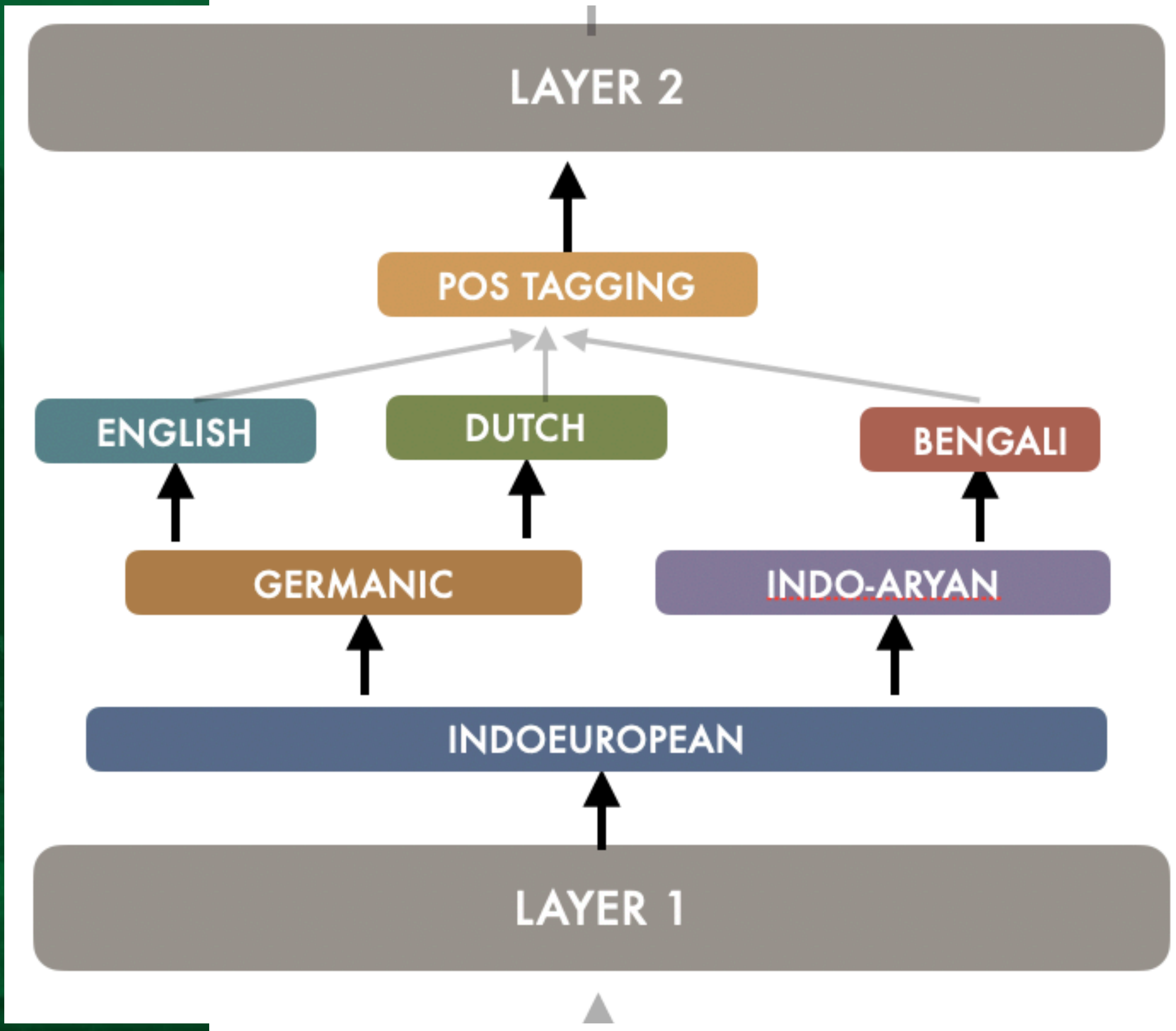


Even constraining to the same number of parameters, still improvements!
Is it language sharing or network depth?

Ablations



Even constraining to the same number of parameters, still improvements!
Is it language sharing or network depth?



Same idea applied to Translation:

- 2^{nd} best constrained system at WMT Shared Task on Large-Scale Multilingual Systems for African Languages!

Going forward and beyond

No matter what, we need data in these languages.

What data do we need, though?

Few-Shot is the way

Let's leave script issues aside for a minute

—since we can find solutions, e.g.

MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer

Jonas Pfeiffer[1], Ivan Vulić[2], Iryna Gurevych[1], Sebastian Ruder[3]
[1]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt
[2]Language Technology Lab, University of Cambridge
[3]DeepMind

CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation

Jonathan H. Clark, Dan Garrette, Iulia Turc, John Wieting
Google Research

When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models

Benjamin Muller[†] Antonis Anastasopoulos[‡] Benoît Sagot[†] Djamé Seddah[†]
[†]Inria, Paris, France
[‡]Department of Computer Science, George Mason University, USA
firstname.lastname@inria.fr antonis@gmu.edu

Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank

Ethan C. Chau[†◇] Lucy H. Lin[†] Noah A. Smith[†⋆]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Department of Linguistics, University of Washington
[⋆]Allen Institute for Artificial Intelligence

It seems that we can do ~~great~~ well with just a few in-domain in-language task data + data augmentation!

Going forward and beyond

Going forward and beyond

Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{adebnath,nrajabi,falam5,antonis}@gmu.edu

(ACL 2021)

How much data?

Going forward and beyond

Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{adebnath,nrajabi,falam5,antonis}@gmu.edu

(ACL 2021)

Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos
Department of Computer Science, George Mason University, USA
{ffaisal, ywang88, antonis}@gmu.edu

(ACL 2022)

How much data?

What data?

Example: (Extractive) Question Answering

We have a lot of English-only datasets for QA (e.g. SQuAD)

Can we leverage them, and investigate few-shot approaches in new languages?

Study on TyDi-QA dataset (7 languages)

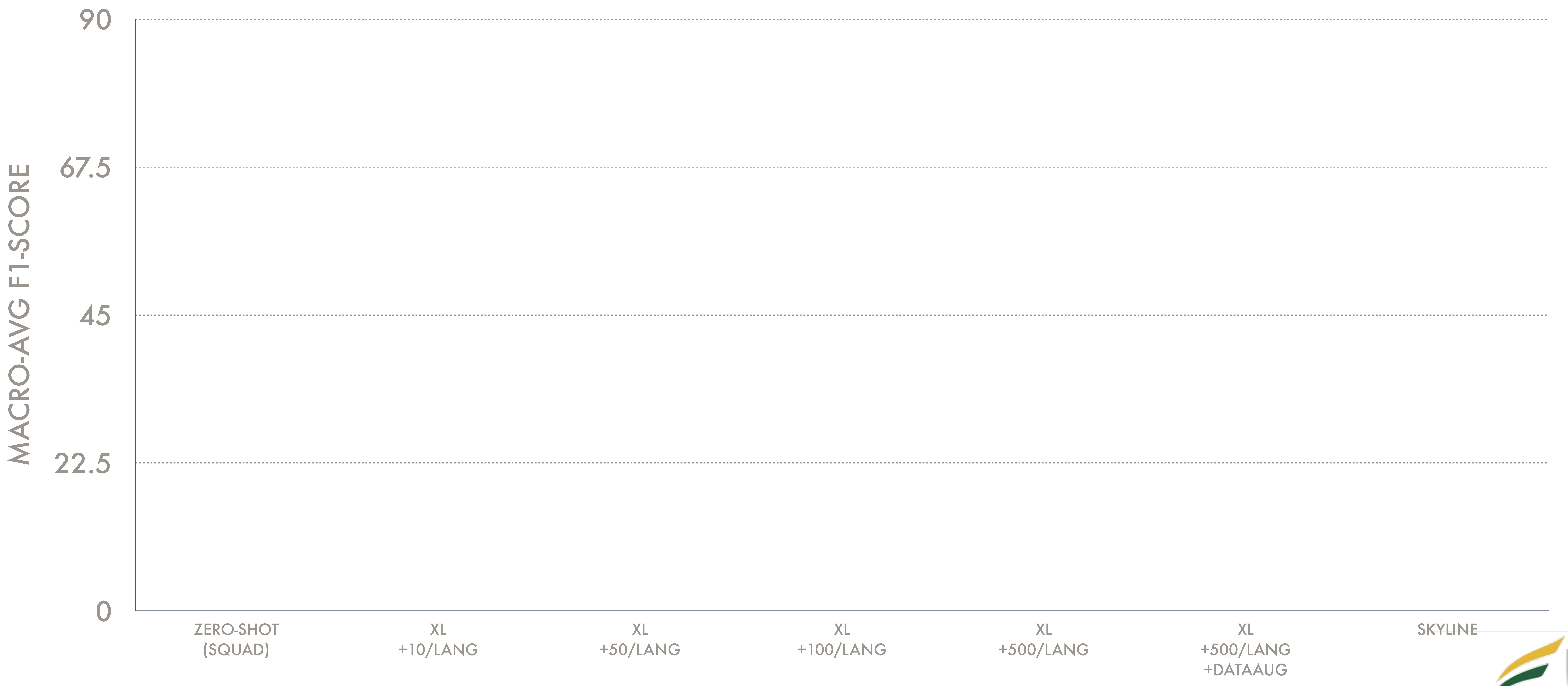
Towards More Equitable Question Answering Systems: How Much More Data Do you Need?

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{adebnath,nrajabi,falam5,antonis}@gmu.edu

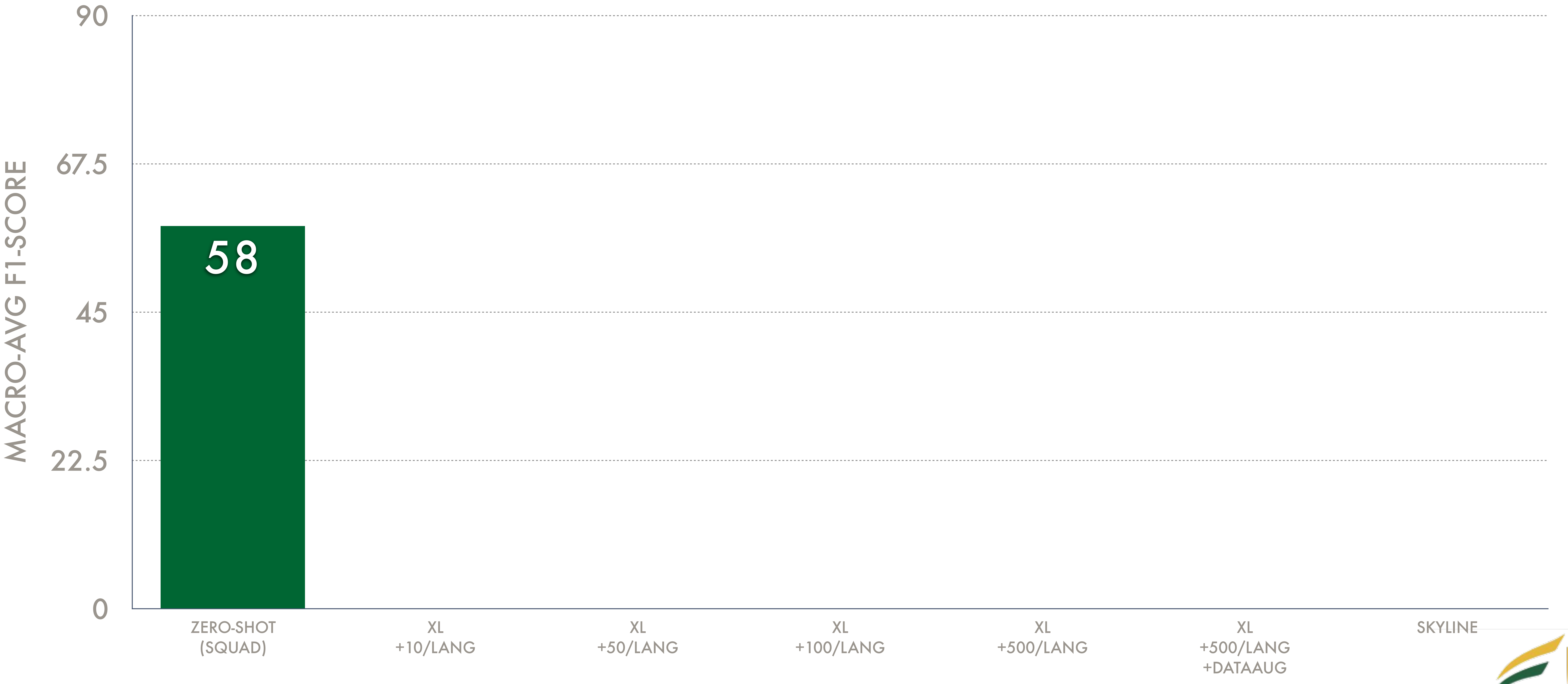
(ACL 2021)

Few-Shot adaptation

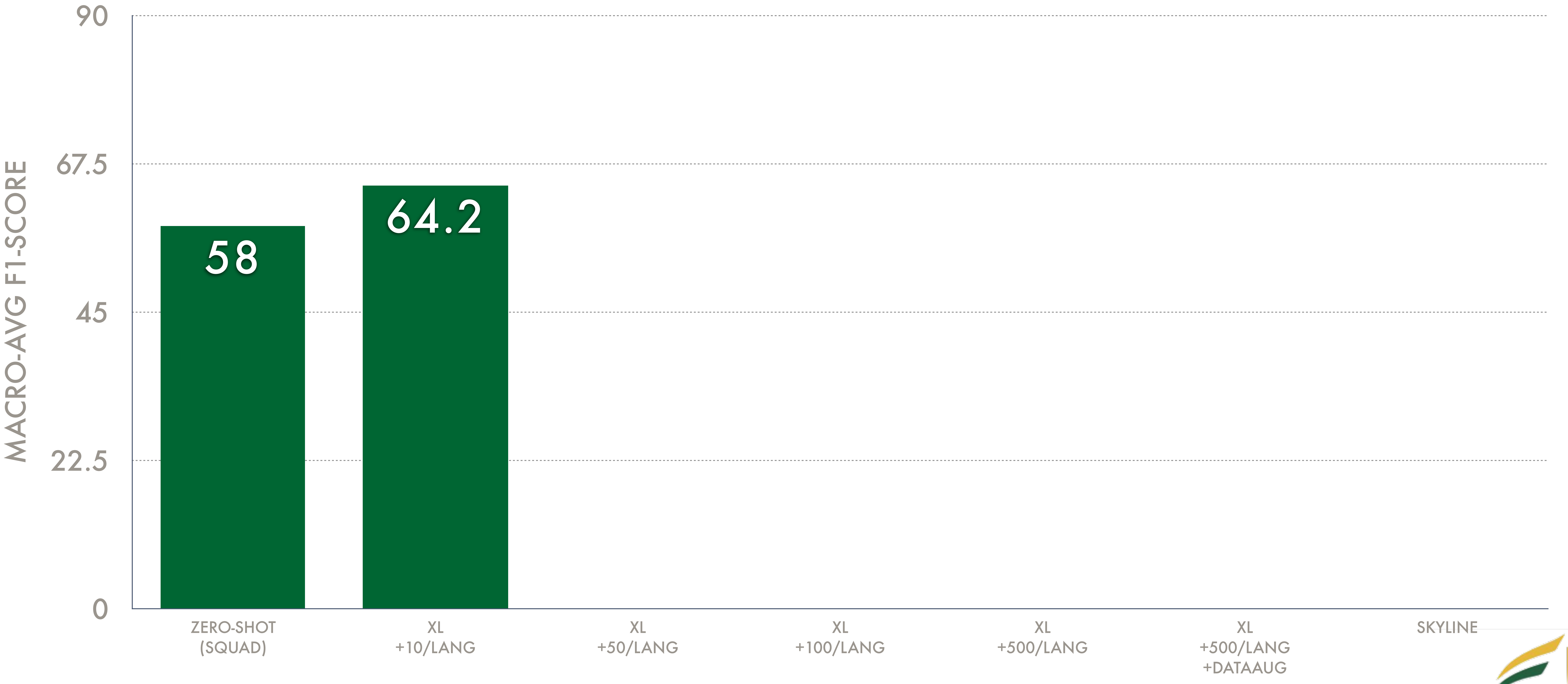
Few-Shot adaptation



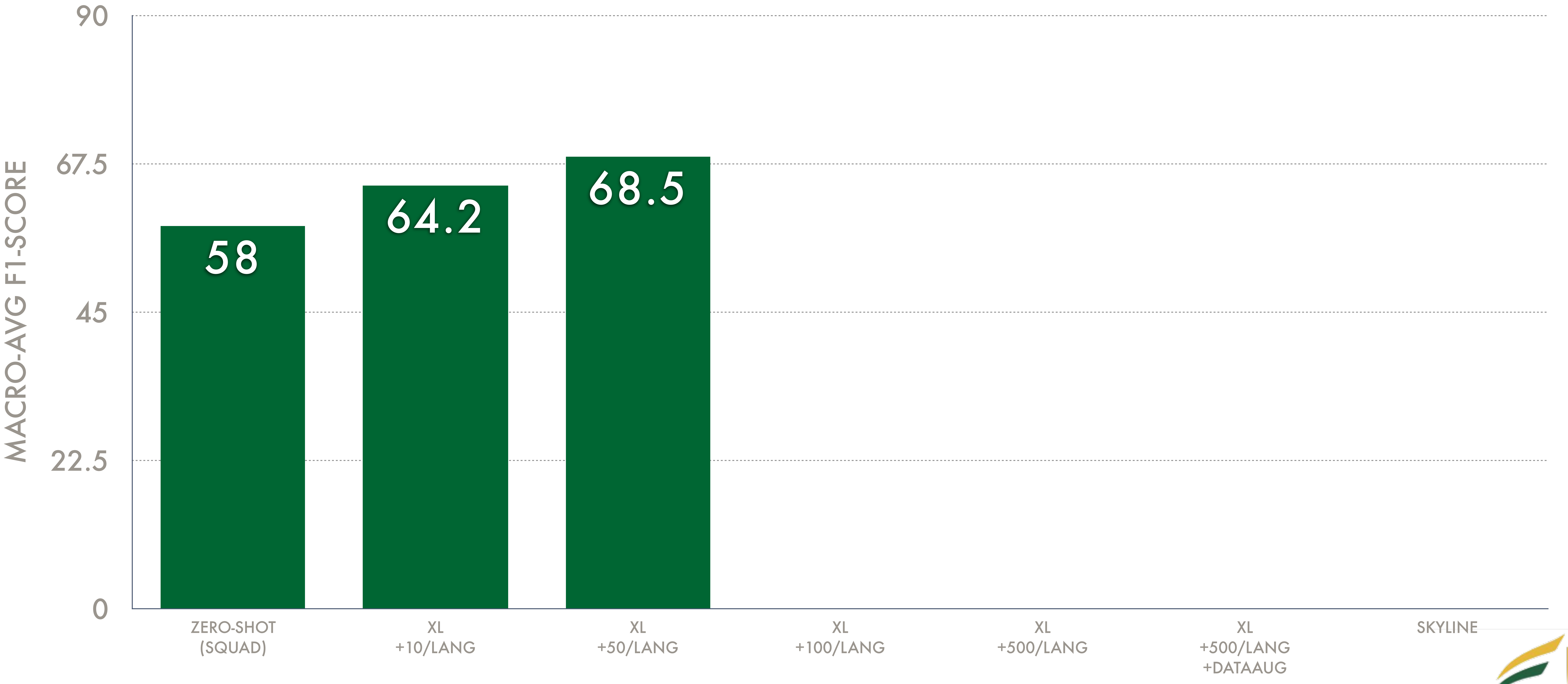
Few-Shot adaptation



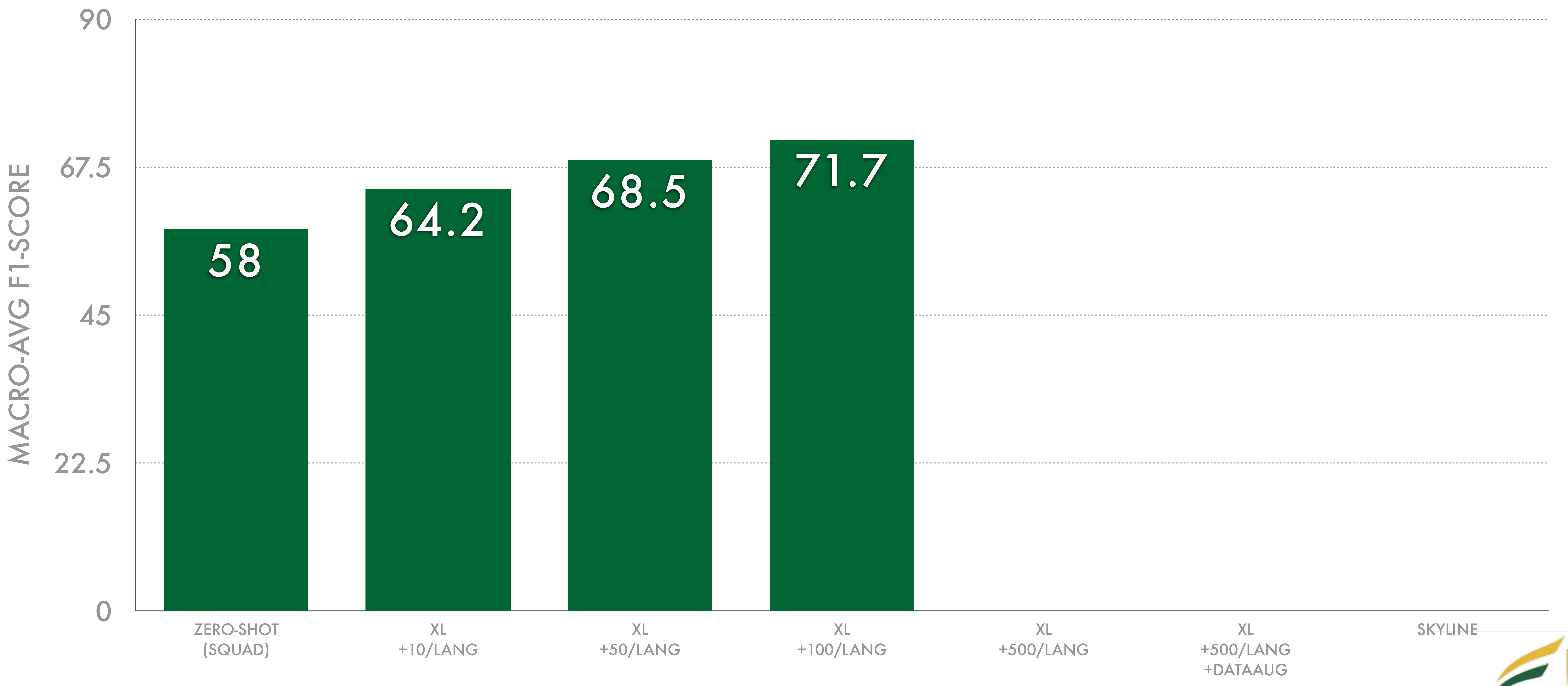
Few-Shot adaptation



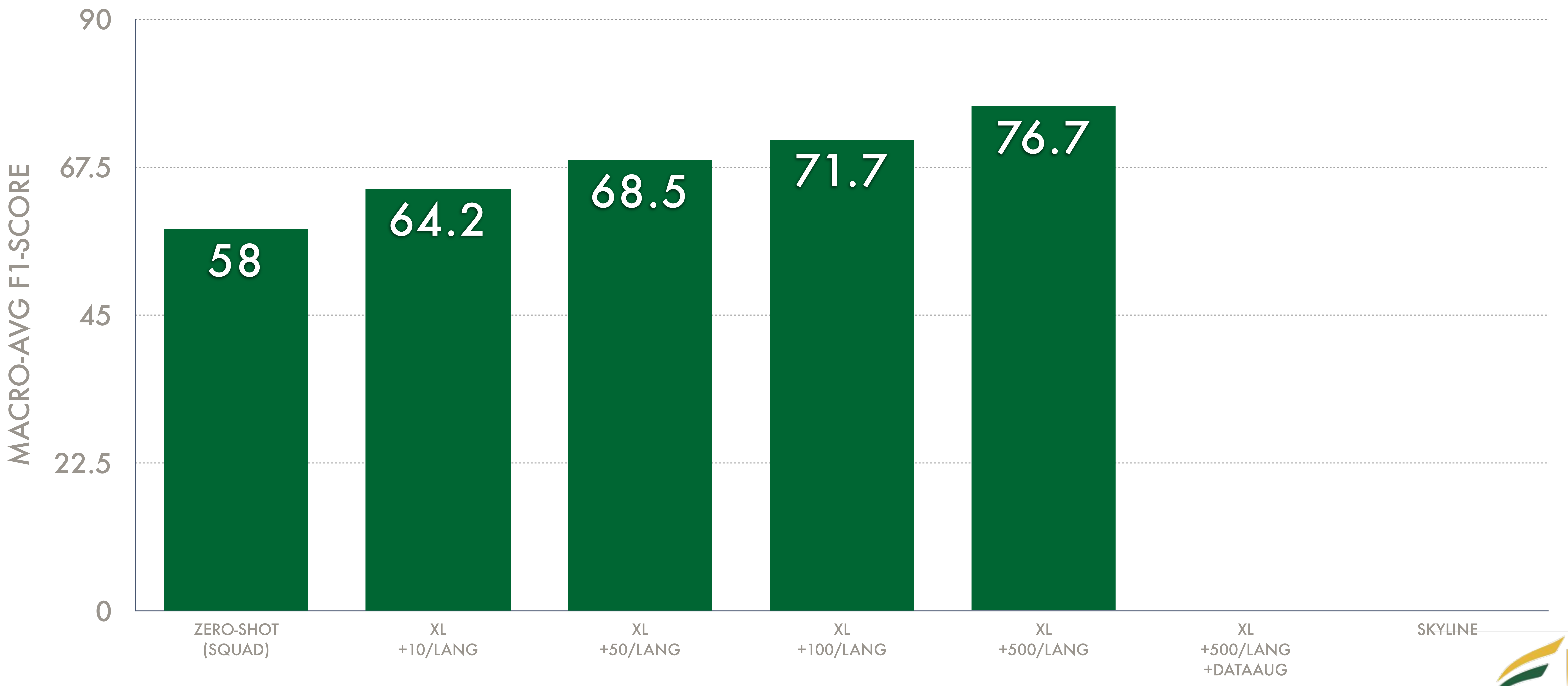
Few-Shot adaptation



Few-Shot adaptation

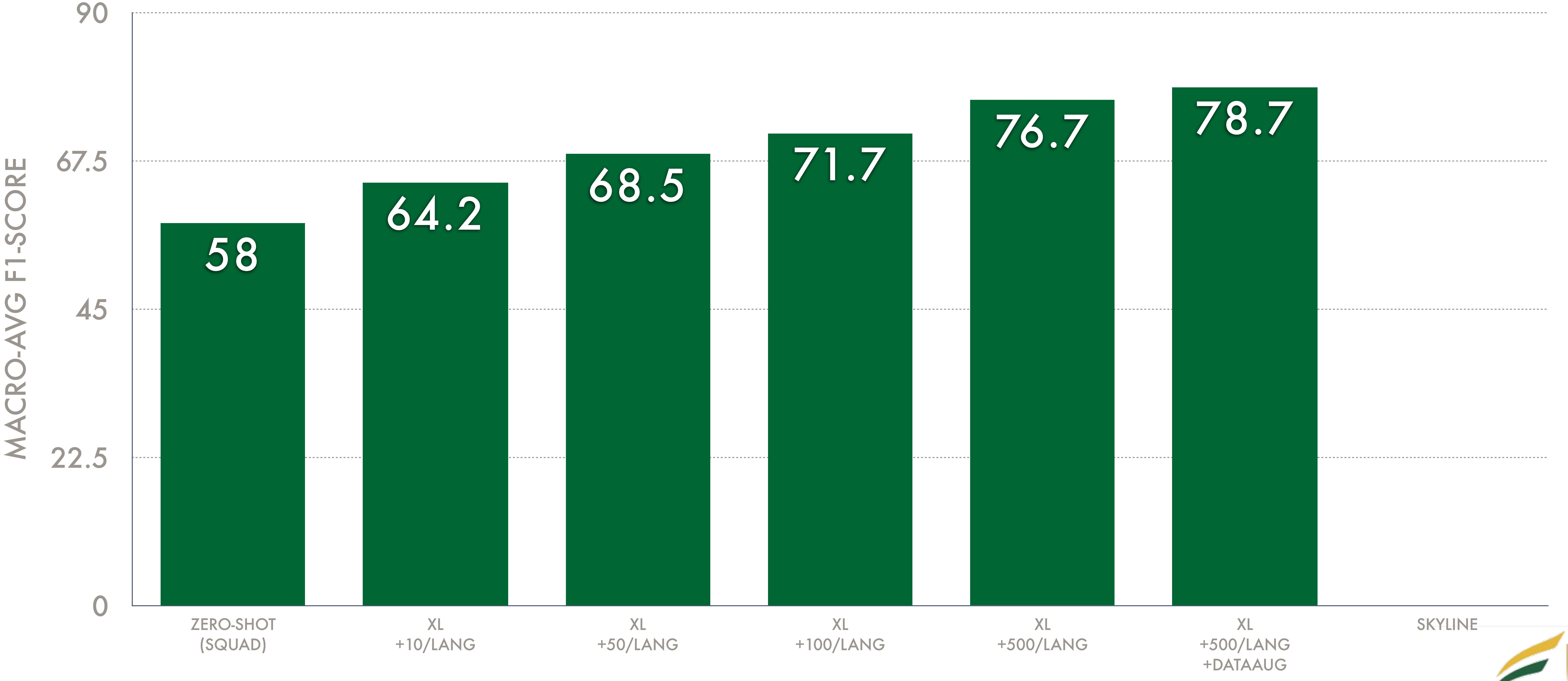


Few-Shot adaptation

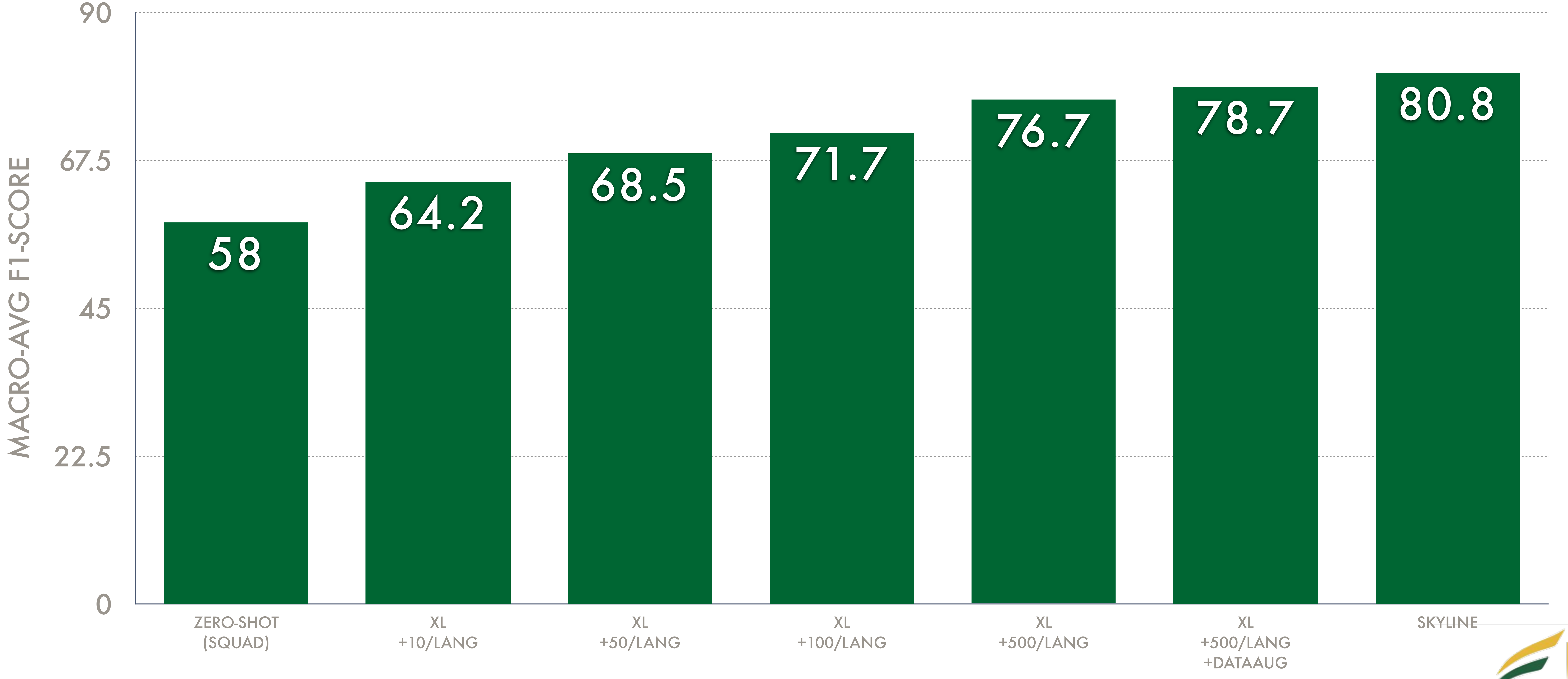


NLP GEORGE MASON

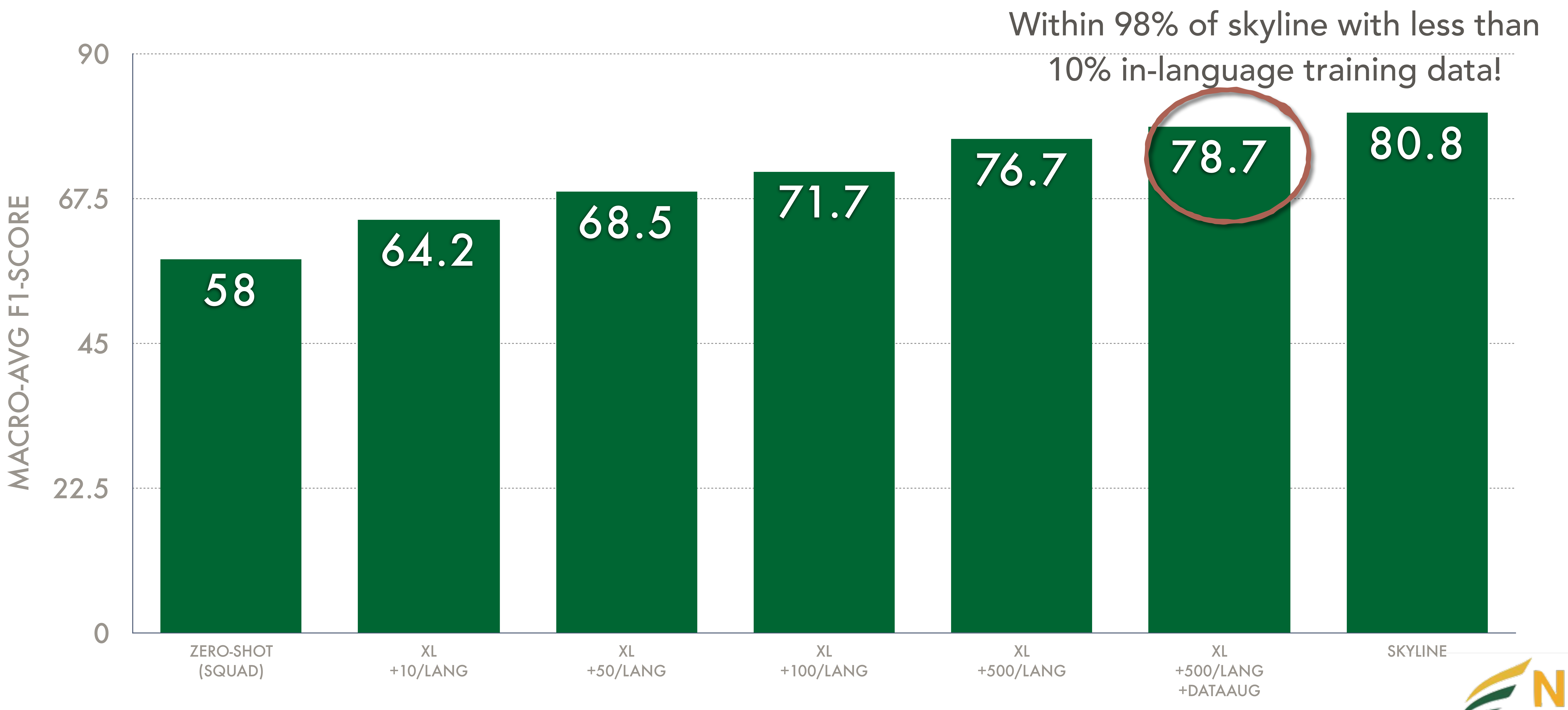
Few-Shot adaptation



Few-Shot adaptation



Few-Shot adaptation



So how should we spend our annotation budget?

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

4500 training examples in 1 language

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3
<	
1500 training examples in 3 languages	

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3
<	
1500 training examples in 3 languages	74.5

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3
<	
1500 training examples in 3 languages	74.5
<	
500 training examples in 6 languages	

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3
<	
1500 training examples in 3 languages	74.5
<	
500 training examples in 6 languages	78.7

So how should we spend our annotation budget?

My view:

Focus on building high-quality evaluation sets

Spend only a fraction of your budget on training data — combine with stronger baselines

***Terms and Conditions apply*

	Avg F-score on 6 other languages:
4500 training examples in 1 language	72.3
<	
1500 training examples in 3 languages	74.5
<	
500 training examples in 6 languages	78.7
<	
250 training examples in 12 languages	…

How Representative Are your Data?

How Representative Are your Data?

Where does your data come from?

How Representative Are your Data?

Where does your data come from?

Which speakers are modeled?

How Representative Are your Data?

Where does your data come from?

Which speakers are modeled?

Study for country-level representation

Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos
Department of Computer Science, George Mason University, USA
{ffaisal, ywang88, antonis}@gmu.edu

(ACL 2022)

Idea

Idea

Named Entities can reveal the information we need!

Idea

Named Entities can reveal the information we need!

Idea

Named Entities can reveal the information we need!

Idea

Named Entities can reveal the information we need!

For a given dataset

RUI COSTA FROM AMADORA PLAYED FOR FIORENTINA

Idea

Named Entities can reveal the information we need!

For a given dataset

- ☑ Identify named entities

RUI COSTA FROM AMADORA PLAYED FOR FIORENTINA

Idea

Named Entities can reveal the information we need!

For a given dataset

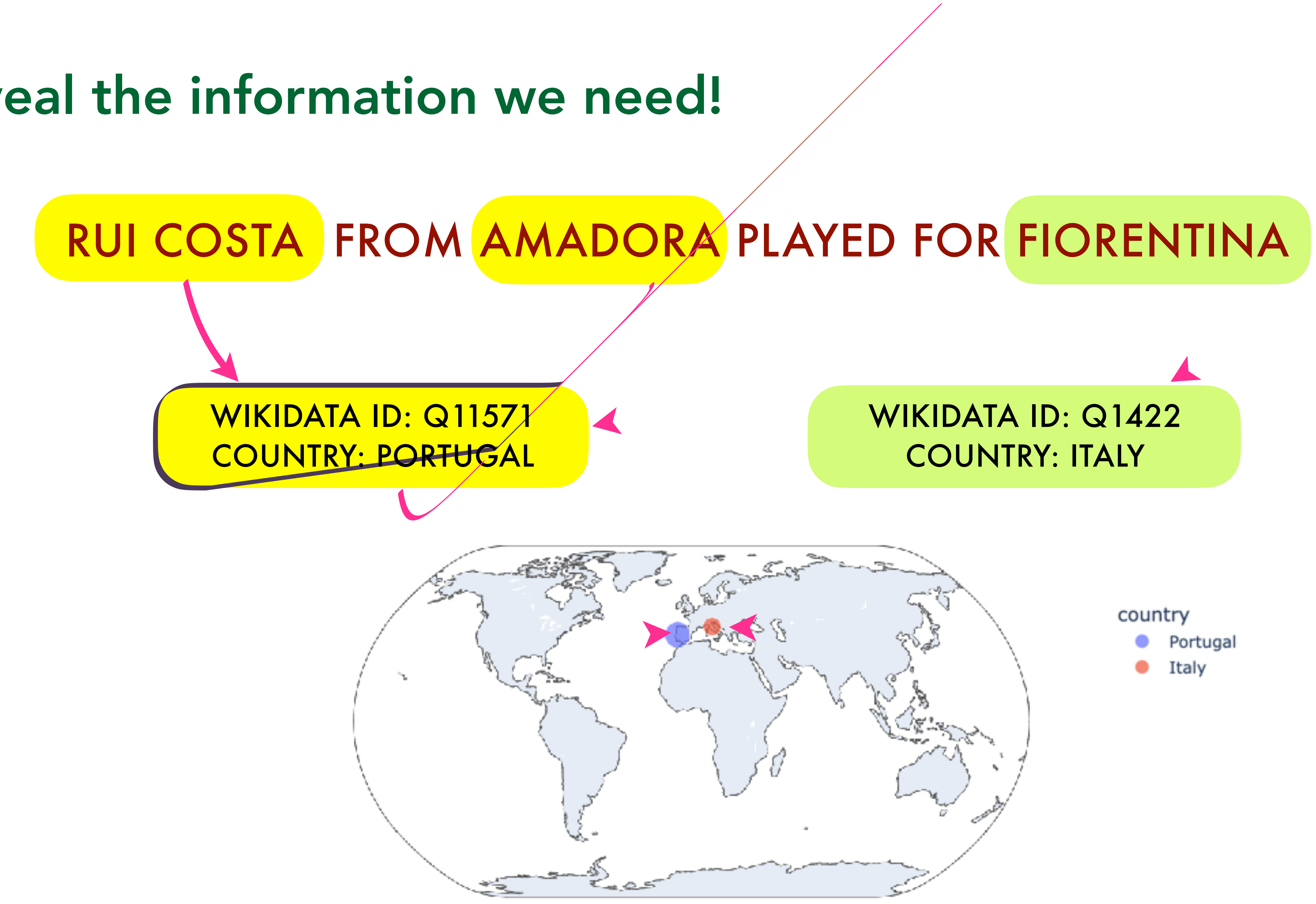
- ☑ Identify named entities
- ☑ Link entities to countries through wikidata

Idea

Named Entities can reveal the information we need!

For a given dataset

- ☑ Identify named entities
- ☑ Link entities to countries through wikidata
- ☑ Aggregate through dataset
  - ☑ Representativeness measures
  - ☑ Fairness measures
  - ☑ Visualizations



Dataset Geography

Code
&
Dataset

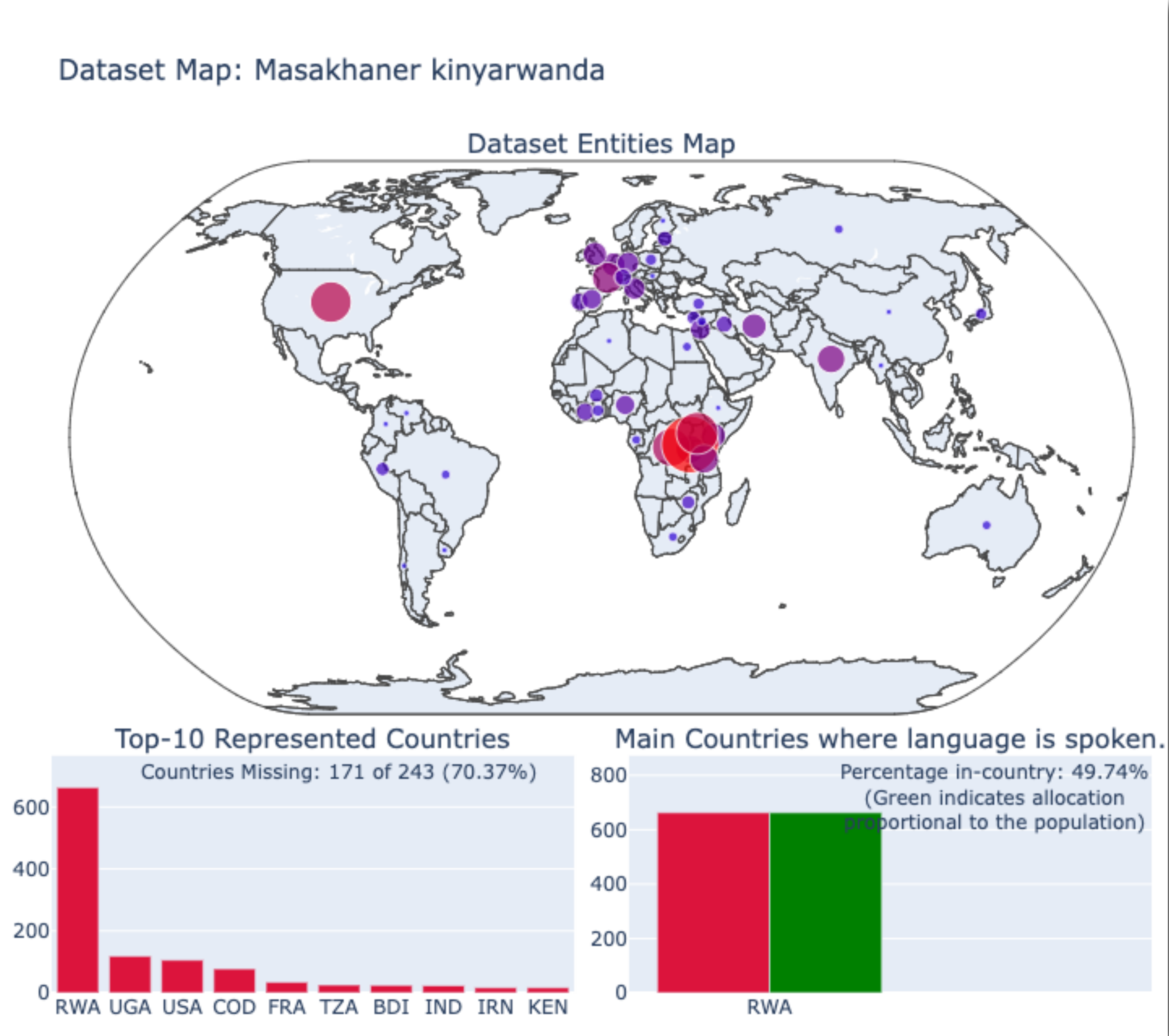
https://github.com/ffaisal93/dataset_geography

Project Webpage
&
Additional Visualizations

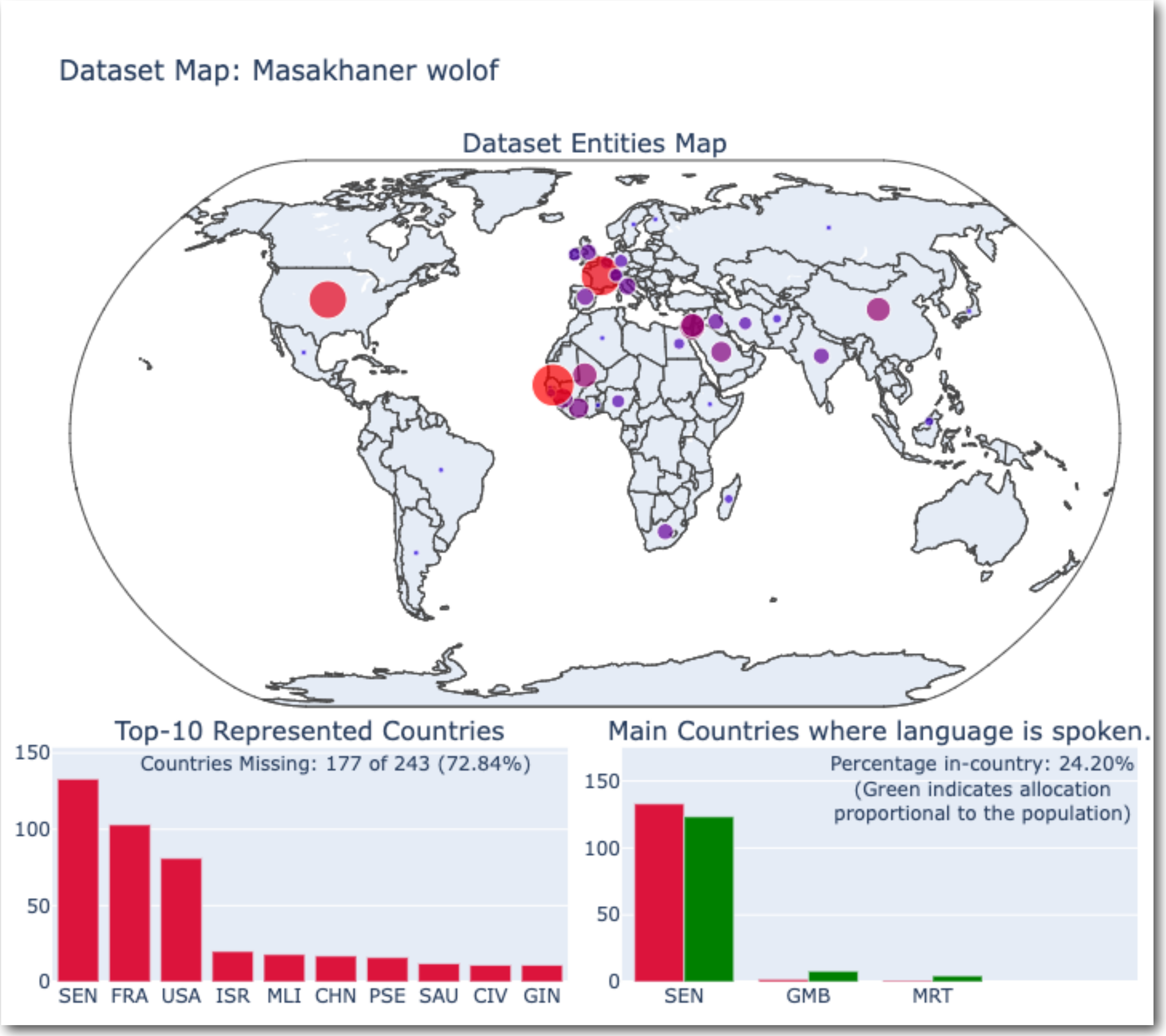
https://nlp.cs.gmu.edu/project/datasetmaps

Dataset Geography

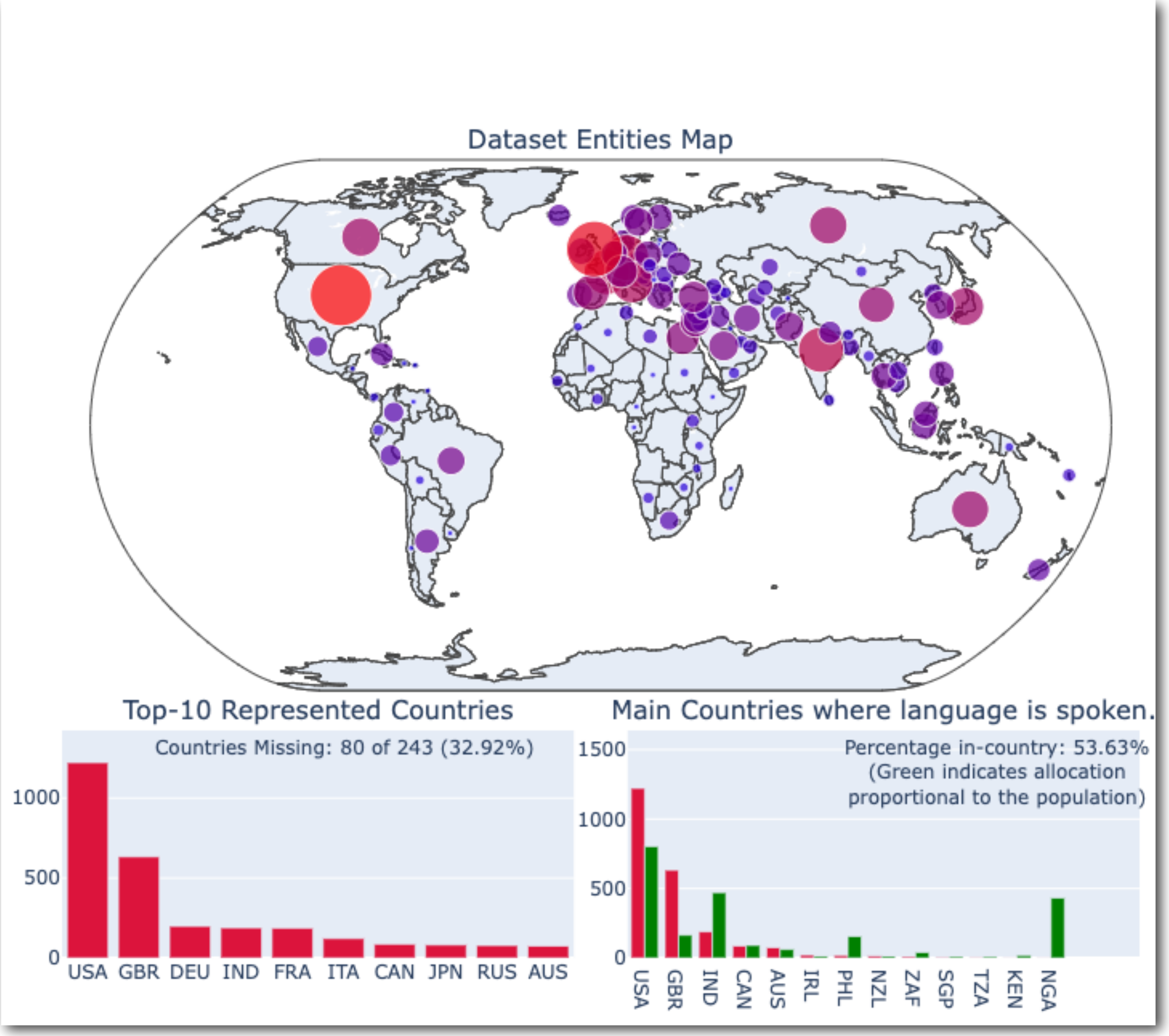
Dataset Geography



Dataset Geography



Dataset Geography



What do communities need/want?

What do communities need/want?

Work ***with*** the communities ***for*** the communities

What do communities need/want?

Work ***with*** the communities ***for*** the communities

What do communities need/want?

Work ***with*** the communities ***for*** the communities

What do communities need/want?

Work ***with*** the communities ***for*** the communities

Educational Tools for Mapuzugun

Cristian Ahumada[1] **Claudio Gutierrez**[1] **Antonios Anastasopoulos**[2]
[1]Department of Computer Science, Universidad de Chile
[2]Computer Science Department, George Mason University
ahumada.860@gmail.com cgutierr@dcc.uchile.cl antonis@gmu.edu

What do communities need/want?

Work ***with*** the communities ***for*** the communities

What do communities need/want?

Work ***with*** the communities ***for*** the communities

BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote
Department of Computer Science
University of Zambia
Zambia
claytone.sikasote@cs.unza.zm

Antonios Anastasopoulos
Department of Computer Science
George Mason University
USA
antonis@gmu.edu

BIG-C: Multimodal Dataset for the Bemba Language

Claytone Sikasote[1], Eunice Mukonde[1], and Antonios Anastasopoulos[2]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

NLP GEORGE MASON

What do communities need/want?

Work ***with*** the communities ***for*** the communities

BembaSpeech: A Speech Recognition Corpus for the Bemba Language

Claytone Sikasote
Department of Computer Science
University of Zambia
Zambia
claytone.sikasote@cs.unza.zm

Antonios Anastasopoulos
Department of Computer Science
George Mason University
USA
antonis@gmu.edu

BIG-C: Multimodal Dataset for the Bemba Language

Claytone Sikasote[1], Eunice Mukonde[1], and Antonios Anastasopoulos[2]
[1]Department of Computer Science, University of Zambia, Zambia
[2]Department of Computer Science, George Mason University, USA
claytone.sikasote@cs.unza.zm, antonis@gmu.edu

Thank you!

Shoutout to collaborators:

Graham Neubig, Damian Blasi, Benjamin Muller, Benoît Sagot, Djamé Seddah

And students:

Fahim Faisal, Sharlina Keshava, Mahfuz ibn Alam, Yinkai Wang

Other things I’m working on:

- NLP for endangered languages (e.g. OCR for scanned documents from Latin America, building basic tools for Griko, Mapudungun, Pomak)
- NLP for linguists (Machine-aided annotation)
- Machine Translation from/into dialects
- Cross-Lingual and Cross-Cultural Fairness
- Geospatial Language Understanding and Navigation
- SLT for Crisis Response
- …

GMU and GMNLP is hiring!

Faculty/postdocs/PhD students

NLP GEORGE MASON

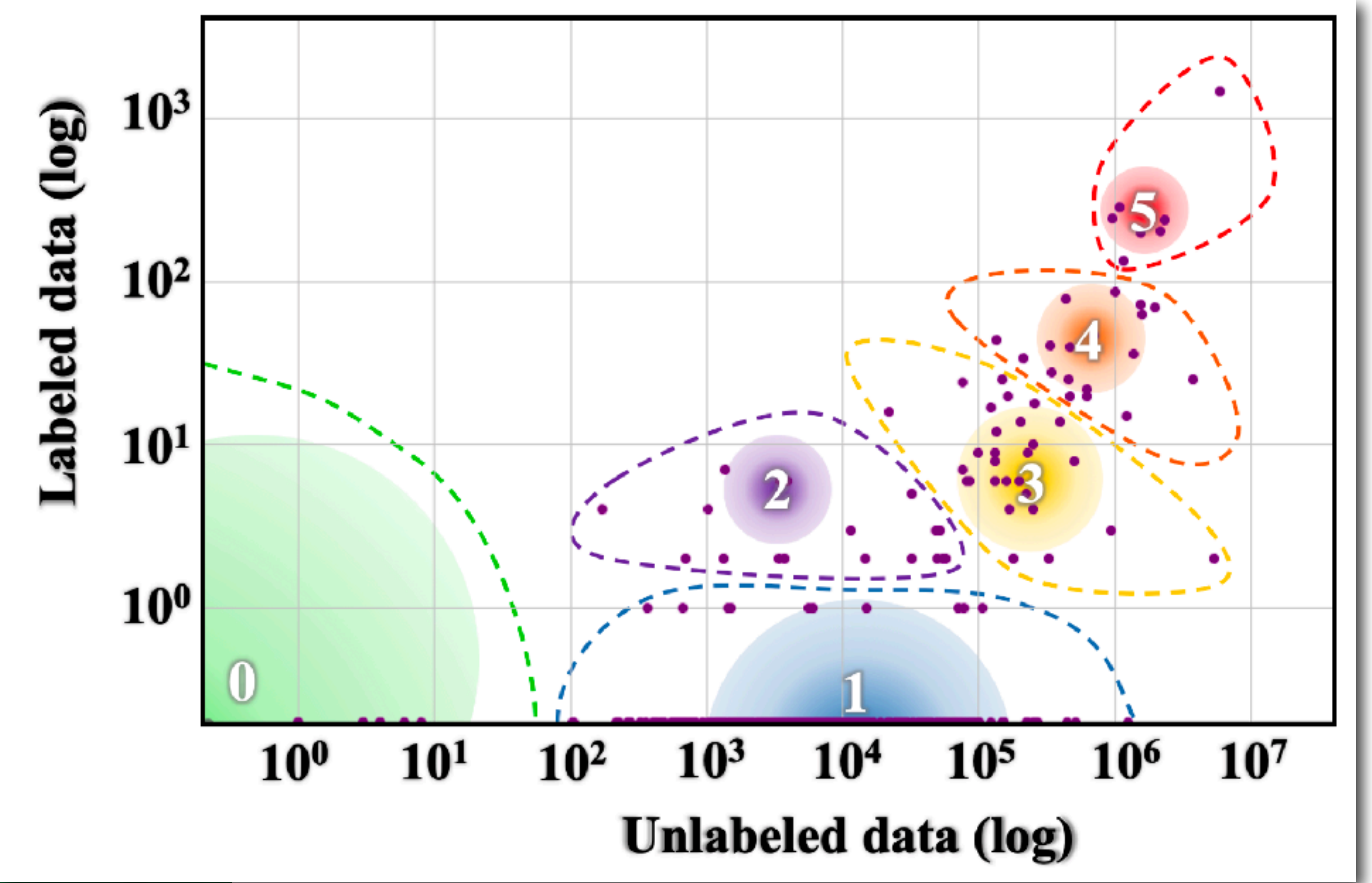


Image from Joshi et al 2020

The amount of data, labeled or unlabeled, varies wildly across languages!