MultiBLiMP 1.0

A Multilingual Benchmark of Linguistic Minimal Pairs

Jaap Jumelet | Leonie Weissweiler | Joakim Nivre | Arianna Bisazza

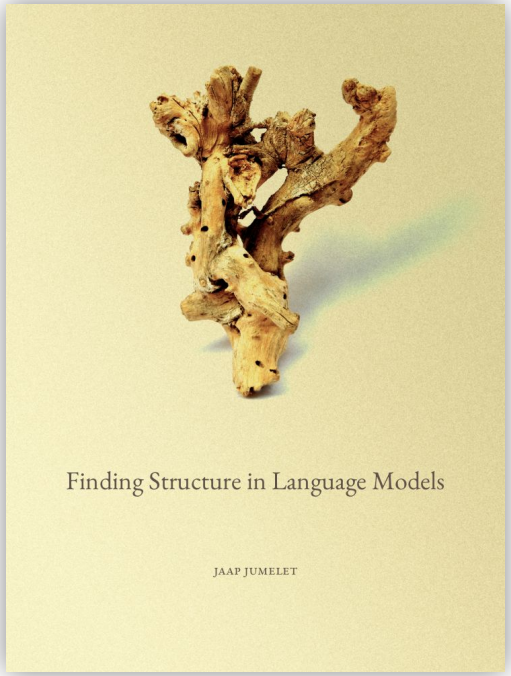
Personal Background

Bio

- BSc + MSc in **Artificial Intelligence** from the University of **Utrecht** and **Amsterdam**
- PhD at the *Institute for Logic Language, and Computation (ILLC)* at the University of Amsterdam with **Jelle Zuidema** and **Raquel Fernandez**

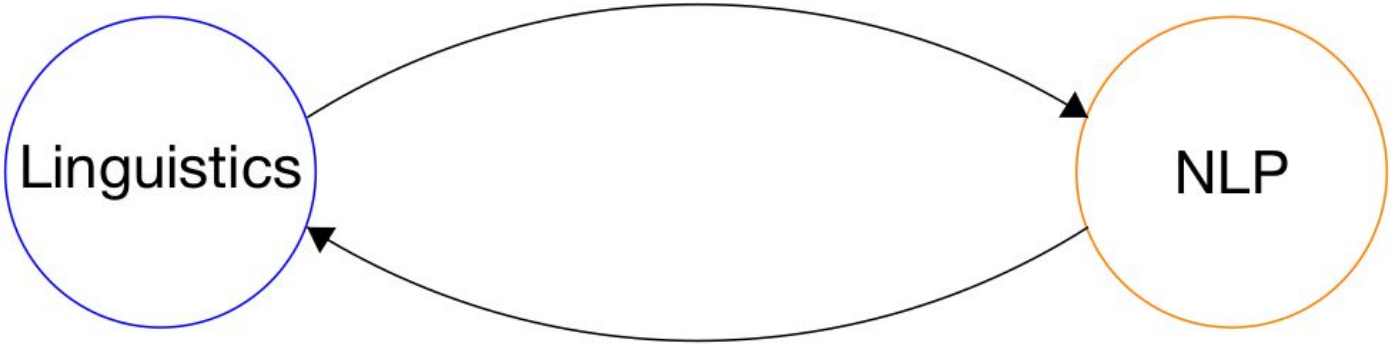
Bio

- BSc + MSc in **Artificial Intelligence** from the University of **Utrecht** and **Amsterdam**
- PhD at the *Institute for Logic Language, and Computation (ILLC)* at the University of Amsterdam with **Jelle Zuidema** and **Raquel Fernandez**
- PhD Thesis:
  Finding Structure in Language Models
- Now: postdoc at the University of Groningen with **Arianna Bisazza**



Main Research Interest

Improve understanding of model behaviour using linguistic concepts



Improve understanding of linguistic concepts using model behaviour(?)

Research Topics

Negative Polarity Items

Do Language Models Understand *Anything*?
On the Ability of LSTMs to Understand Negative Polarity Items

Jaap Jumelet
University of Amsterdam
jaap.jumelet@student.uva.nl

Dieuwke Hupkes
ILLC, University of Amsterdam
d.hupkes@uva.nl

Abstract

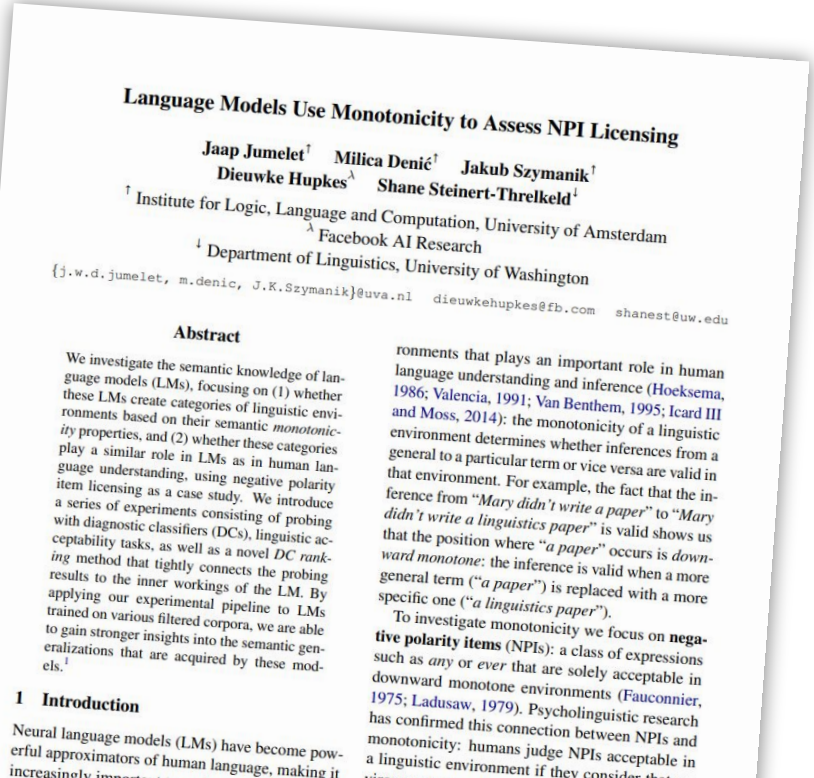
In this paper, we attempt to link the inner workings of a neural language model to linguistic theory, focusing on a complex phenomenon well discussed in formal linguistics: (negative) polarity items. We briefly discuss the leading hypotheses about the licensing contexts that allow negative polarity items and evaluate to what extent a neural language model has the ability to correctly process a subset of such constructions. We show that the model finds a relation between the licensing context and the negative polarity item and appears to be aware of the *scope* of this context, which we extract from a parse tree of the sentence. With this research, we hope to pave the way for other studies linking formal linguistics to deep learning.

1 Introduction

syntactic structure, yielding promising results. In this paper, we follow up on this research by studying a phenomenon that has received much attention by linguists and for which the model requires – besides knowledge of syntactic structure – also a *semantic* understanding of the sentence: negative polarity items (NPIs).

In short, NPIs are a class of words that bear the special feature that they need to be *licensed* by a specific licensing context (LC) (a more elaborate linguistic account of NPIs can be found in the next section). A common example of an NPI and LC in English are *any* and *not*, respectively: The sentence *He didn't buy any books* is correct, whereas *He did buy any books* is not. To properly process an NPI construction, a language model must be able to detect a relationship between a licensing context and an NPI.

Following Linzen et al. (2016) and Gulordava et al. (2018), we devise several tasks to assess



Language Models Use Monotonicity to Assess NPI Licensing

Jaap Jumelet[↑] **Milica Denić**[↑] **Jakub Szymanik**[↑]
Dieuwke Hupkes[λ] **Shane Steinert-Threlkeld**[↓]
[↑] Institute for Logic, Language and Computation, University of Amsterdam
[λ] Facebook AI Research
[↓] Department of Linguistics, University of Washington
{j.w.d.jumelet, m.denic, J.K.Szymanik}@uva.nl dieuwkehupkes@fb.com shanest@uw.edu

Abstract

We investigate the semantic knowledge of language models (LMs), focusing on (1) whether these LMs create categories of linguistic environments based on their semantic *monotonicity* properties, and (2) whether these categories play a similar role in LMs as in human language understanding, using negative polarity item licensing as a case study. We introduce a series of experiments consisting of probing with diagnostic classifiers (DCs), linguistic acceptability tasks, as well as a novel *DC ranking* method that tightly connects the probing results to the inner workings of the LM. By applying our experimental pipeline to LMs trained on various filtered corpora, we are able to gain stronger insights into the semantic generalizations that are acquired by these models.[1]

1 Introduction

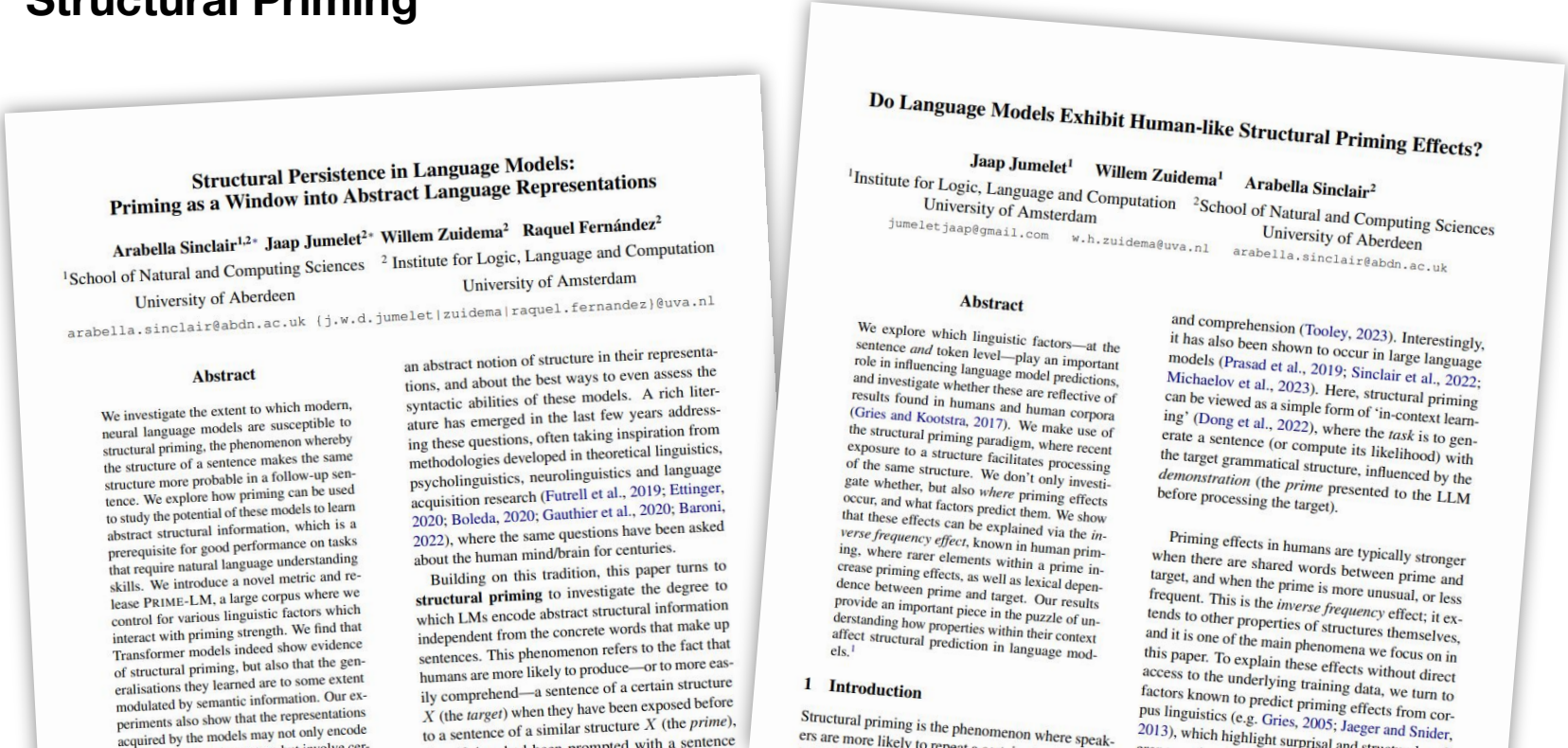
Neural language models (LMs) have become powerful approximators of human language, making it

ronments that plays an important role in human language understanding and inference (Hoeksema, 1986; Valencia, 1991; Van Benthem, 1995; Icard III and Moss, 2014): the monotonicity of a linguistic environment determines whether inferences from a general to a particular term or vice versa are valid in that environment. For example, the fact that the inference from "*Mary didn't write a paper*" to "*Mary didn't write a linguistics paper*" is valid shows us that the position where "*a paper*" occurs is *downward monotone*: the inference is valid when a more general term ("*a paper*") is replaced with a more specific one ("*a linguistics paper*").

To investigate monotonicity we focus on **negative polarity items** (NPIs): a class of expressions such as *any* or *ever* that are solely acceptable in downward monotone environments (Fauconnier, 1975; Ladusaw, 1979). Psycholinguistic research has confirmed this connection between NPIs and monotonicity: humans judge NPIs acceptable in a linguistic environment if they consider

Research Topics

Structural Priming



Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations

Arabella Sinclair[1,2*] Jaap Jumelet[2*] Willem Zuidema[2] Raquel Fernández[2]
[1]School of Natural and Computing Sciences, University of Aberdeen
[2] Institute for Logic, Language and Computation, University of Amsterdam
arabella.sinclair@abdn.ac.uk {j.w.d.jumelet|zuidema|raquel.fernandez}@uva.nl

Abstract

We investigate the extent to which modern, neural language models are susceptible to structural priming, the phenomenon whereby the structure of a sentence makes the same structure more probable in a follow-up sentence. We explore how priming can be used to study the potential of these models to learn abstract structural information, which is a prerequisite for good performance on tasks that require natural language understanding skills. We introduce a novel metric and release PRIME-LM, a large corpus where we control for various linguistic factors which interact with priming strength. We find that Transformer models indeed show evidence of structural priming, but also that the generalisations they learned are to some extent modulated by semantic information. Our experiments also show that the representations acquired by the models may not only encode

an abstract notion of structure in their representations, and about the best ways to even assess the syntactic abilities of these models. A rich literature has emerged in the last few years addressing these questions, often taking inspiration from methodologies developed in theoretical linguistics, psycholinguistics, neurolinguistics and language acquisition research (Futrell et al., 2019; Ettinger, 2020; Boleda, 2020; Gauthier et al., 2020; Baroni, 2022), where the same questions have been asked about the human mind/brain for centuries.

Building on this tradition, this paper turns to **structural priming** to investigate the degree to which LMs encode abstract structural information independent from the concrete words that make up sentences. This phenomenon refers to the fact that humans are more likely to produce—or to more easily comprehend—a sentence of a certain structure X (the *target*) when they have been exposed before to a sentence of a similar structure X (the *prime*),

Do Language Models Exhibit Human-like Structural Priming Effects?

Jaap Jumelet[1] Willem Zuidema[1] Arabella Sinclair[2]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]School of Natural and Computing Sciences, University of Aberdeen
jumeletjaap@gmail.com w.h.zuidema@uva.nl arabella.sinclair@abdn.ac.uk

Abstract

We explore which linguistic factors—at the sentence *and* token level—play an important role in influencing language model predictions, and investigate whether these are reflective of results found in humans and human corpora (Gries and Kootstra, 2017). We make use of the structural priming paradigm, where recent exposure to a structure facilitates processing of the same structure. We don't only investigate whether, but also *where* priming effects occur, and what factors predict them. We show that these effects can be explained via the *inverse frequency effect*, known in human priming, where rarer elements within a prime increase priming effects, as well as lexical dependence between prime and target. Our results provide an important piece in the puzzle of understanding how properties within their context affect structural prediction in language models.[1]

1 Introduction

Structural priming is the phenomenon where speakers are more likely to repeat a

and comprehension (Tooley, 2023). Interestingly, it has also been shown to occur in large language models (Prasad et al., 2019; Sinclair et al., 2022; Michaelov et al., 2023). Here, structural priming can be viewed as a simple form of 'in-context learning' (Dong et al., 2022), where the *task* is to generate a sentence (or compute its likelihood) with the target grammatical structure, influenced by the *demonstration* (the *prime* presented to the LLM before processing the target).

Priming effects in humans are typically stronger when there are shared words between prime and target, and when the prime is more unusual, or less frequent. This is the *inverse frequency* effect; it extends to other properties of structures themselves, and it is one of the main phenomena we focus on in this paper. To explain these effects without direct access to the underlying training data, we turn to factors known to predict priming effects from corpus linguistics (e.g. Gries, 2005; Jaeger and Snider, 2013), which highlight surprisal and

Research Topics

Interpretability

Feature Interactions Reveal Linguistic Structure in Language Models

Jaap Jumelet **Willem Zuidema**
Institute for Logic, Language and Computation
University of Amsterdam
{j.w.d.jumelet, w.zuidema}@uva.nl

Abstract

We study *feature interactions* in the context of *feature attribution* methods for post-hoc interpretability. In interpretability research, getting to grips with feature interactions is increasingly recognised as an important challenge, because interacting features are key to the success of neural networks. Feature interactions allow a model to build up hierarchical representations for its input, and might provide an ideal starting point for the investigation into linguistic structure in language models. However, uncovering the exact role that these interactions play is also difficult, and a diverse range of interaction attribution methods has been proposed. In this paper, we focus on the question which of these methods most *faithfully* reflects the inner workings of the target models. We work out a *grey box* methodology, in which we train models to perfection on a formal language classification task, using PCFGs. We show that under specific configurations, some methods are indeed

mostly ignore the existence of interactions between the effects of features on the prediction. This is problematic, because **Feature Interactions** are widely seen as a major factor in the success of neural networks (Goodfellow et al., 2016). This is all the more important in domains such as language and music processing, because feature interactions allow neural networks to model hierarchical representations of their input, which is considered a key design feature of language and music. To address these shortcomings, there is now an emerging literature on **feature interaction detection and attribution methods** (FIDAMs) that explain model predictions in terms of interacting features (Tsang et al., 2020; Janizek et al., 2021).

However, assessing the faithfulness of FIDAMs is even more challenging than assessing the faithfulness of feature attribution methods more generally (Jacovi and Goldberg, 2021). In this paper we present a systematic framework to characterise FIDAMs by

DecoderLens: Layerwise Interpretation of Encoder-Decoder Transformers

Anna Langedijk[1] **Hosein Mohebbi**[2] **Gabriele Sarti**[3] **Willem Zuidema**[1] **Jaap Jumelet**[1]
[1]ILLC, University of Amsterdam [2]CSAI, Tilburg University [3]CLCG, University of Groningen
annalangedijk@gmail.com h.mohebbi@tilburguniversity.edu
g.sarti@rug.nl {w.h.zuidema, j.w.d.jumelet}@uva.nl

Abstract

In recent years, several interpretability methods have been proposed to interpret the inner workings of Transformer models at different levels of precision and complexity. In this work, we propose a simple but effective technique to analyze encoder-decoder Transformers. Our method, which we name DecoderLens, allows the decoder to cross-attend representations of intermediate encoder activations instead of using the default final encoder output. The method thus maps uninterpretable intermediate vector representations to human-interpretable sequences of words or symbols, shedding new light on the information flow in this popular but understudied class of models. We apply DecoderLens to question answering, logical reasoning, speech recognition and machine translation models, finding that simpler subtasks are solved with high precision by low and intermediate encoder layers.

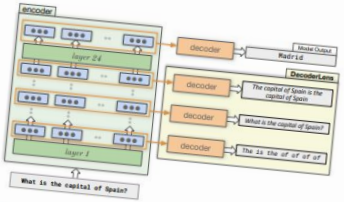


Figure 1: Schematic overview of the DecoderLens. By using the decoder to cross-attend intermediate encoder activation, we can gain qualitative insights into how representations evolve across encoder layers.

1 Introduction

Many methods for interpreting the

of encoder-decoder Transformers as a "lens" to explain the evolution of representations throughout model layers in these model architectures. Our method is directly inspired by the LogitLens (nostalgebraist, 2020), which leverages the *residual stream*[1] present in Transformer architectures. The LogitLens, however, is defined only for decoder

Research Topics

Relation of training data and model behaviour

Transparency at the Source: Evaluating and Interpreting Language Models With Access to the True Distribution

Jaap Jumelet **Willem Zuidema**
Institute for Logic, Language and Computation
University of Amsterdam
{j.w.d.jumelet, w.zuidema}@uva.nl

Abstract

We present a setup for training, evaluating and interpreting neural language models, that uses artificial, language-like data. The data is generated using a massive probabilistic grammar (based on state-split PCFGs), that is itself derived from a large natural language corpus, but also provides us complete control over the generative process. We describe and release both grammar and corpus, and test for the naturalness of our generated data. This approach allows us to define closed-form expressions to efficiently compute exact lower bounds on obtainable perplexity using both causal and masked language modelling. Our results show striking differences between neural language modelling architectures and training objectives in how closely they allow approximating the lower bound on perplexity. Our approach also allows us to directly compare learned repre-

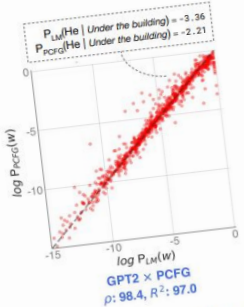


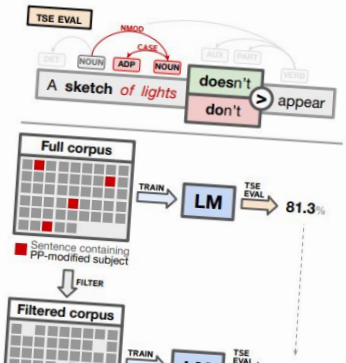
Figure 1: Spearman correlation between the log probabilities of a LM with GPT-2 architecture and the PCFG distribution. The LM has obtained a distribution that aligns significantly with the true PCFG distribution.

Filtered Corpus Training (FiCT) Shows that Language Models can Generalize from Indirect Evidence

Abhinav Patil[▲,*] and **Jaap Jumelet**[🚲,*]
Yu Ying Chiu[▲] and **Andy Lapastora**[☺,†] and **Peter Shen**[▲] and **Lexie Wang**[▲] and **Clevis Willrich**[π,†]
Shane Steinert-Threlkeld[▲]
▲: University of Washington 🚲: ILLC, University of Amsterdam
☺: AWS AI Labs, Amazon π: Microsoft

Abstract

This paper introduces **F**iltered **C**orpus **T**raining, a method that trains language models (LMs) on corpora with certain linguistic constructions filtered out from the training data, and uses it to measure the ability of LMs to perform linguistic generalization on the basis of indirect evidence. We apply the method to both LSTM and Transformer LMs (of roughly comparable size), developing filtered corpora that target a wide range of linguistic phenomena. Our results show that while transformers are better qua LMs (as measured by perplexity), both models perform equally and surprisingly well on linguistic generalization measures, suggesting that they are capable of generalizing from indirect evidence.



Research Topics Postdoc

PI: Arianna Bisazza

1. Improve language modelling efficiency for **low-resource languages** using insights from **language acquisition**
2. Investigate differences in **learnability** using **typological** features

Research Topics Postdoc

PI: Arianna Bisazza

1. Improve language modelling efficiency for **low-resource languages** using i
   language acquisition
2. Investigate differences in **learnabil**
   typological features
   - Organizer of **BabyLM**, released a **multilingual extension**

BabyBabelLM:
A Multilingual Benchmark of Developmentally Plausible Training Data

Jaap Jumelet[1], Abdellah Fourtassi[2], Akari Haga[3], Bastian Bunzeck[4], Bhargav Shandilya[5], Diana Galvan-Sosa[6, 12], Faiz Ghifari Haznitrama[7], Francesca Padovani[1], Francois Meyer[8], Hai Hu[9], Julen Etxaniz[10], Laurent Prévot[2], Linyang He[11], María Grandury[12], Mila Marcheva[6], Negar Foroutan[13], Nikitas Theodoropoulos[14], Pouya Sadeghi[15], Siyuan Song[16], Suchir Salhan[6], Susana Zhou[12], Yurii Paniv[17], Ziyin Zhang[18], Arianna Bisazza[1], Alex Warstadt[19], Leshem Choshen[20]

[1]University of Groningen, [2]Aix Marseille University, [3]Nara Institute of Science and Technology, [4]Bielefeld University, [5]University of Colorado Boulder, [6]University of Cambridge, [7]KAIST, [8]University of Cape Town, [9]City University of Hong Kong, [10]HiTZ, University of the Basque Country, [11]Columbia University, [12]SomosNLP, [13]EPFL, [14]Independent Researcher, [15]University of Tehran, [16]University of Texas at Austin, [17]Ukrainian Catholic University, [18]Shanghai Jiao Tong University, [19]University of California San Diego, [20]MIT, MIT-IBM Watson AI Lab

Correspondence to j.w.d.jumelet@rug.nl and leshem.choshen@mail.huji.ac.il*

Abstract

We present **BabyBabelLM**, a multilingual collection of datasets modeling the language a person observes from birth until they acquire a native language. We curate developmentally plausible pretraining data aiming to cover the equivalent of 100M English words of content in each of 45 languages. We compile evaluation suites and train baseline models in each language. BabyBabelLM aims to facilitate multilingual pretraining and cognitive modeling.[1]

created to redirect attention toward questions of data efficiency and developmental plausibility in language modeling. The shared task invites participants to propose data-efficient LMs pretrained on a fixed, developmentally plausible English corpus of child-directed speech (CDS), educational content, and other simplified texts. The top-performing submissions (Charpentier and Samuel, 2023, 2024) have significantly improved the state of the art for models trained on the same limited data budget, even surpassing LMs trained on much larger cor-

Today

MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs

Jaap Jumelet[1], **Leonie Weissweiler**[2], **Joakim Nivre**[2], and **Arianna Bisazza**[1]
[1]University of Groningen, The Netherlands [2]Uppsala University, Sweden
j.w.d.jumelet@rug.nl,
{leonie.weissweiler,joakim.nivre}@lingfil.uu.se, a.bisazza@rug.nl

Abstract

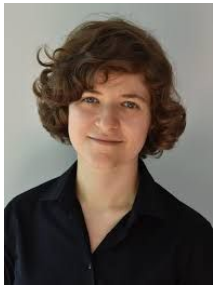
We introduce MultiBLiMP 1.0, a massively multilingual benchmark of linguistic minimal pairs, covering 101 languages and 2 types of subject-verb agreement, containing more than 128,000 minimal pairs. Our minimal pairs are created using a fully automated pipeline, leveraging the large-scale linguistic resources of Universal Dependencies and UniMorph. MultiBLiMP 1.0 evaluates abilities of LLMs at an unprecedented multilingual scale, and highlights the shortcomings of the current state-of-the-art in modelling low-resource languages.[1]

1 Introduction

Large language models (LLMs) are often trained on highly multilingual corpora, which enable users

syntactic aspect (Linzen et al., 2016; Warstadt et al., 2020), where a formally competent LM is expected to assign higher probability to the grammatical version. Such datasets, however, exist only for English and a few other, mostly high-resource languages (Gulordava et al., 2018; Mueller et al., 2020; Taktasheva et al., 2024).

To accelerate progress in this direction, we introduce **MultiBLiMP 1.0**, a massively multilingual benchmark of linguistic minimal pairs covering two types of subject-verb agreement (subject-finite-verb and subject-participle) for number, person, and gender; created automatically using two large-scale linguistic resources: Universal Dependencies (UD, Nivre et al., 2016, 2020; de Marneffe et al., 2021) and UniMorph (Batsuren et al., 2022). Multi-BLiMP is not only

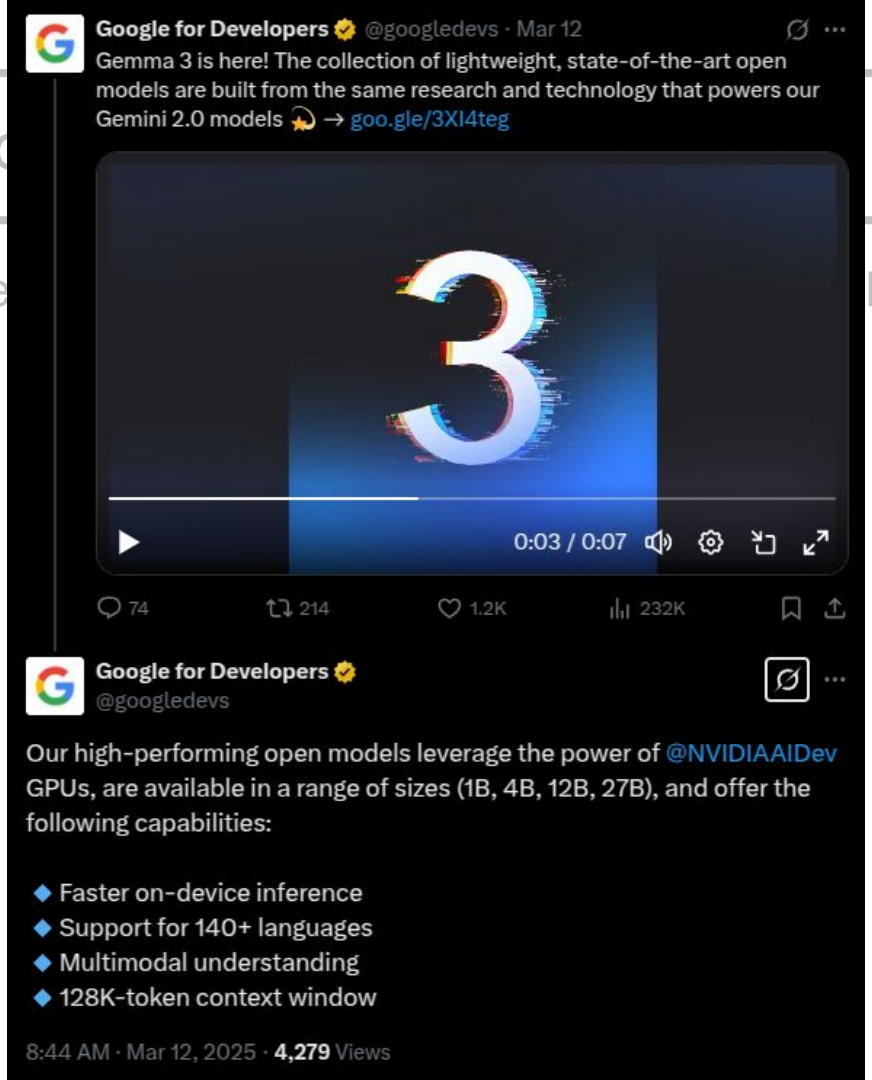


The State of Multilingual Evaluation

- How do we evaluate the *multilingual* capacities of LLMs?

The State

- How do we LMs?



The State

- How do we LMs?

Google for Developers @googledevs · Mar 12

Gemma 3 is here! The collection of lightweight, state-of-the-art open models are built from the same research and technology that powers our Gemini 2.0 models → goo.gle/3XI4teg

74 214 1.2K 232K

Google for Developers
@googledevs

Our high-performing open models leverage the power of @NVIDIAAIDev GPUs, are available in a range of sizes (1B, 4B, 12B, 27B), and offer the following capabilities:

- Faster on-device inference
- Support for 140+ languages
- Multimodal understanding
- 128K-token context window

?

8:44 AM · Mar 12, 2025 · **4,279** Views

The State of Multilingual Evaluation

- of LLMs?

	Gemma 2			Gemma 3			
	2B	9B	27B	1B	4B	12B	27B
MGSM	18.7	57.3	68.0	2.04	34.7	64.3	**74.3**
GMMLU	43.3	64.0	69.4	24.9	57.0	69.4	**75.7**
WMT24++	38.8	50.3	53.0	36.7	48.4	53.9	**55.7**
Flores	30.2	41.3	44.3	29.5	39.2	46.0	**48.8**
XQuAD	53.7	72.2	73.9	43.9	68.0	74.5	**76.8**
ECLeKTic	8.29	14.0	17.1	4.69	11.0	17.2	**24.4**
IndicGB	47.4	59.3	62.1	41.4	57.2	61.7	**63.4**

Table 13 | Multilingual performance after the pre-training phase. IndicGenBench is an average over benchmarks reported in Table 14.

Reasoning
MT
Knowledge retrieval

The State of Multilingual Evaluation

Aya-expanse (Cohere For AI, 2024):

Task	Dataset		Metric	Languages
Discriminative Tasks				
Coreference Resolution	XWinograd [Muennighoff et al., 2023]	0-shot	Acc.	6
Sentence Completion	XCOPA [Ponti et al., 2020]	0-shot	Acc.	11
	XStoryCloze [Lin et al., 2021]	0-shot	Acc.	10
Language Understanding	Global-MMLU [Singh et al., 2024a]	5-shot	Acc.	23
	INCLUDE [Romanou et al., 2024]	0-shot	Acc.	44
Generative Tasks				
Translation	FLORES-200 [Goyal et al., 2021; NLLB-Team et al., 2022]	0-shot	chrF++, xCOMET	23
Mathematical Reasoning	MGSM [Shi et al., 2023]	5-shot	Acc.	7
Open-Ended Generation	Multilingual ArenaHard	0-shot	win-rate	23
	Dolly Human-edited & Machine-translated [Singh et al., 2024b]	0-shot	win-rate	23

The State of Multilingual Evaluation

- How do we evaluate the ***multilingual*** capacities of LLMs?
- Evaluation focuses on **downstream** tasks
- Evaluation is often limited to **high-resource** languages

The State of Multilingual Evaluation

- How do we evaluate the ***multilingual*** capacities of LLMs?
- Evaluation focuses on **downstream** tasks
- Evaluation is often limited to **high-resource** languages
- **What about *linguistic evaluation*?**

Towards Multilingual Linguistic Evaluation

Why do we need *multilingual linguistic evaluation*?

Towards Multilingual Linguistic Evaluation

Why do we need *multilingual linguistic evaluation*?

- Competence on downstream tasks *depends* on linguistic ability

Tow

Why

ity

Trends in
Cognitive Sciences

CellPress

Feature Review

Dissociating language and thought in large language models

Kyle Mahowald[1,5,*], Anna A. Ivanova[2,5,*], Idan A. Blank[3,*], Nancy Kanwisher[4,*], Joshua B. Tenenbaum[4,*], and Evelina Fedorenko[4,*]

Large language models (LLMs) have come closest among all models to date to mastering human language, yet opinions about their linguistic and cognitive capabilities remain split. Here, we evaluate LLMs using a distinction between formal linguistic competence (knowledge of linguistic rules and patterns) and functional linguistic competence (understanding and using language in the world). We ground this distinction in human neuroscience, which has shown that formal and functional competence rely on different neural mechanisms. Although LLMs are surprisingly good at formal competence, their performance on functional competence tasks remains spotty and often requires specialized fine-tuning and/or coupling with external modules. We posit that models that use language in human-like ways would need to master both of these competence types, which, in turn, could require the emergence of separate mechanisms specialized for formal versus functional linguistic competence.

Highlights

Formal linguistic competence (getting the form of language right) and functional linguistic competence (using language to accomplish goals in the world) are distinct cognitive skills.

The human brain contains a network of areas that selectively support language processing (formal linguistic competence), but not other domains like logical or social reasoning (functional linguistic competence).

In the late 2010s, large language models trained on word prediction tasks began achieving unprecedented

Towards Multilingual Linguistic Evaluation

Why c

- C ty

- "We posit that models that use language in human-like ways would need to master both
  - **Formal linguistic competence** *(knowledge of linguistic rules and patterns)* and
  - **Functional linguistic competence** *(understanding and using language in the world)*"
- "*Both* formal and functional linguistic competence are essential components of human language use: an effective communicator needs to both generate grammatical, meaningful utterances and strategically use those utterances to achieve diverse, context-dependent goals"

— Mahowald et al. (2024)

Towards Multilingual Linguistic Evaluation

Why do we need *multilingual linguistic evaluation*?

- Competence on downstream tasks *depends* on linguistic ability
- Typological transfer

Towarc

Why do w

- Comp
- Typol

c ability

Do Llamas Work in English? On the Latent Language of Multilingual Transformers

Chris Wendler*, Veniamin Veselovsky*, Giovanni Monea*, Robert West*
EPFL
{chris.wendler, veniamin.veselovsky, giovanni.monea, robert.west}@epfl.ch

ACL 2024

Abstract

We ask whether multilingual language models trained on unbalanced, English-dominated corpora use English as an internal pivot language—a question of key importance for understanding how language models function and the origins of linguistic bias. Focusing on the Llama-2 family of transformer models, our study uses carefully constructed non-English prompts with a unique correct single-token continuation. From layer to layer, transformers gradually map an input embedding of the final prompt token to an output embedding from which next-token probabilities are computed. Tracking intermediate embeddings through their high-dimensional space reveals three distinct phases, whereby intermediate embeddings (1) start far away from output token embeddings; (2) already allow for decoding a semantically correct next token in middle layers, but give higher probability to its version in English than in the input language; (3) finally move into an input-language-specific region of the embedding space. We cast these results into a conceptual model where the three phases operate in "input space", "concept space", and "output space", respectively. Crucially, our evidence suggests that the abstract "concept space" lies closer to English than to

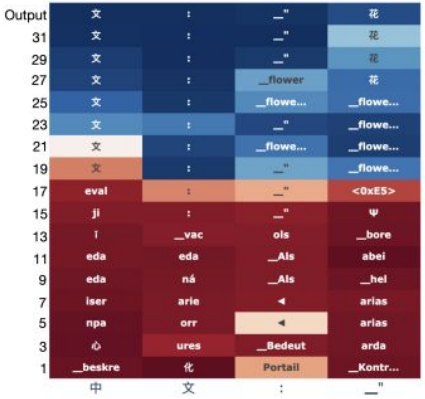


Figure 1: **Illustration of logit lens,** which applies language modeling head (here, Llama-2-7B) prematurely to latent embeddings in intermediate layers, yielding one next-token distribution per position (x-axis) and layer (y-axis). We show final tokens of translation prompt (cf. Sec. 3.3) ending with "Français: "fleur" - 中文: ""(where "中文" means "Chinese"). Final layer correctly ranks "花" (translation of "fleur") on top, whereas intermediate layers decode English "flower". Color indicates entropy of next-token distributions from low (blue) to high (red). (Plotting tool: Belrose et al. (2023).)

Towards Multilingual Linguistic Evaluation

Why do we need *multilingual linguistic evaluation*?

- Competence on downstream tasks *depends* on linguistic ability
- Typological transfer
- Investigating *learnability* of languages and *inductive bias*

Towa

Why do

- Co
- Typ
- Inv

bility

Mission: Impossible Language Models

Julie Kallini[1], **Isabel Papadimitriou**[1], **Richard Futrell**[2], **Kyle Mahowald**[3], **Christopher Potts**[1]

[1]Stanford University; [2]University of California, Irvine; [3]University of Texas, Austin

kallini@stanford.edu

Abstract

Chomsky and others have very directly claimed that large language models (LLMs) are equally capable of learning languages that are possible and impossible for humans to learn. However, there is very little published experimental evidence to support such a claim. Here, we develop a set of synthetic *impossible languages* of differing complexity, each designed by systematically altering English data with unnatural word orders and grammar rules. These languages lie on an impossibility continuum: at one end are languages that are inherently impossible, such as random and irreversible shuffles of English words, and on the other, languages that may not be intuitively impossible but are often considered so in linguistics, particularly those with rules based on counting word positions. We report on a wide range of evaluations to assess the capacity of GPT-2 small models to learn these uncontroversially impossible languages, and crucially, we perform these assessments at various stages throughout training to compare the learning process for each language.

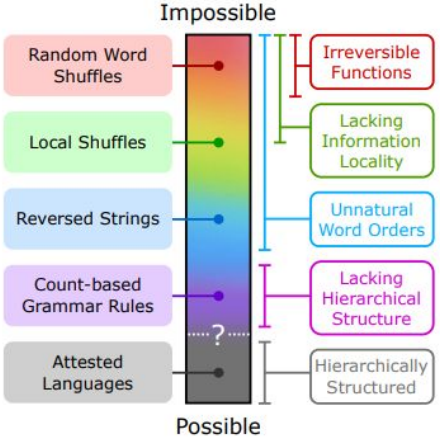


Figure 1: Partial impossibility continuum of languages based on complexity. We assess the learnability of languages at different points in the continuum and push the (currently unclear) boundary between possible and impossible.

Multilingual Linguistic Evaluation

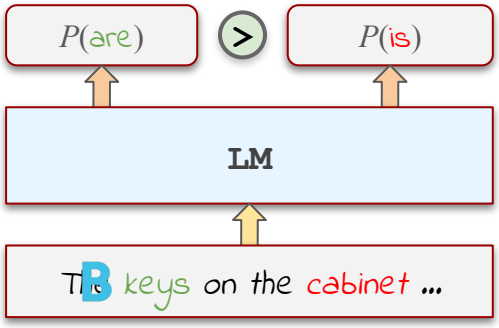
How can we understand a model's *linguistic abilities*?

- Using carefully crafted **minimal pairs** we can investigate a model's performance on a specific phenomenon.

Multilingual Linguistic Evaluation

How can we understand a model's *linguistic abilities*?

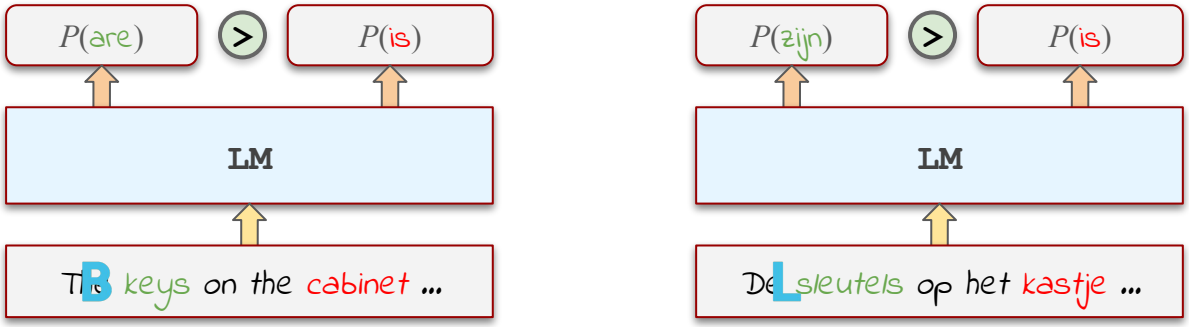
- Using carefully crafted **minimal pairs** we can investigate a model's performance on a specific phenomenon.
- This type of experiment only requires access to the **output probabilities** of the model.



Multilingual Linguistic Evaluation

How can we understand a model's *linguistic abilities*?

- Using carefully crafted **minimal pairs** we can investigate a model's performance on a specific phenomenon.
- This type of experiment only requires access to the **output probabilities** of the model.



BLiMP

- Warstadt et al. (2020): **Benchmark of Linguistic Minimal Pairs for English**
- Tests the capacity of language models for a wide range of *linguistic phenomena*
- Allows us to test and compare language model performance regardless of size
- Comparison done based on *sentence probability:*

$$\mathrm{P}(\text{grammatical sentence}) > \mathrm{P}(\text{ungrammatical sentence})$$

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

BLiMP

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Table 1: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category.

Warstadt et al. (2020) - BLiMP: The Benchmark of Linguistic Minimal Pairs for English

Extending BLiMP to other languages

- Recently many extensions of BLiMP have been proposed for other languages:
  - Dutch: BLiMP-NL
  - Mandarin Chinese: ZhoBLiMP
  - Russian: RuBLiMP
  - Japanese: JBLiMP

Extending BLiMP to other languages

- Recently many extensions of BLiMP have been proposed for other languages:
  - Dutch: BLiMP-NL
  - Mandarin Chinese: ZhoBLiMP
  - Russian: RuBLiMP
  - Japanese: JBLiMP
- All these efforts rely on **linguistic experts** of that language
- How can we scale linguistic evaluation beyond this?

Introducing **MultiBLiMP 1.0**

- Pipeline for automatic minimal pair creation
- Coverage of...
  - **... 101 languages**
  - **... over 128.000 minimal pairs**

MultiBLiMP Logic

- We focus on **agreement**, which can be expressed as a violation of a morphological marker:
  - The boy^{SG} walks^{SG} vs. *The boy^{SG} walk^{PL}

MultiBLiMP Logic

- We focus on **agreement**, which can be expressed as a violation of a morphological marker:
  - The boy^{SG} walks^{SG} vs. *The boy^{SG} walk^{PL}
- We consider **number**, **gender**, and **person** agreement between subject-verb and subject-participle

MultiBLiMP Logic

- We focus on **agreement**, which can be expressed as a violation of a morphological marker:
  - The boy^{SG} walks^{SG} vs. *The boy^{SG} walk^{PL}
- We consider **number**, **gender**, and **person** agreement between subject-verb and subject-participle
- We create multilingual minimal pairs using two resources:
  - **Universal Dependencies**:
    Provides syntactic structure to identify agreement relations
  - **UniMorph**:
    Provides morphological feature to form ungrammatical **inflections**

MultiBLiMP Pipeline

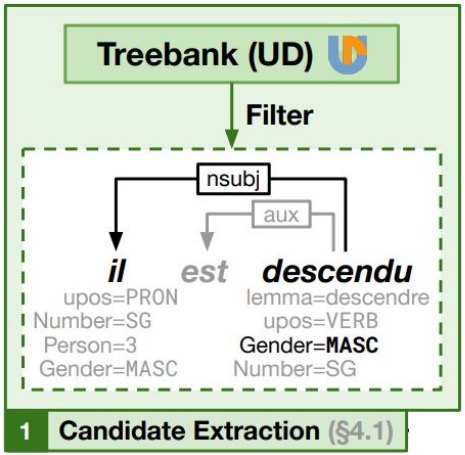
Language L **French**	$L \in$ {Abkhaz, Albanian ··· Wolof, Yakut}
Dependency D **Subject-participle**	$D \in$ {subject-verb, subject-participle}
Feature ϕ **Gender**	$\phi \in$ {number, gender, person}

0 **Configuration**

MultiBLiMP Pipeline

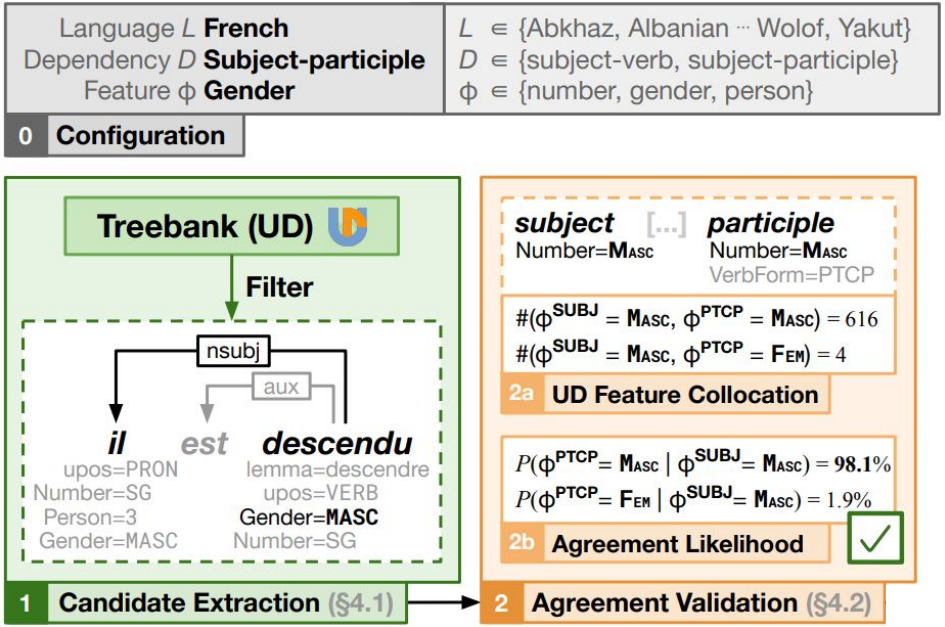
Language L **French**	$L \in$ {Abkhaz, Albanian ··· Wolof, Yakut}
Dependency D **Subject-participle**	$D \in$ {subject-verb, subject-participle}
Feature ϕ **Gender**	$\phi \in$ {number, gender, person}

0 Configuration



1 Candidate Extraction (§4.1)

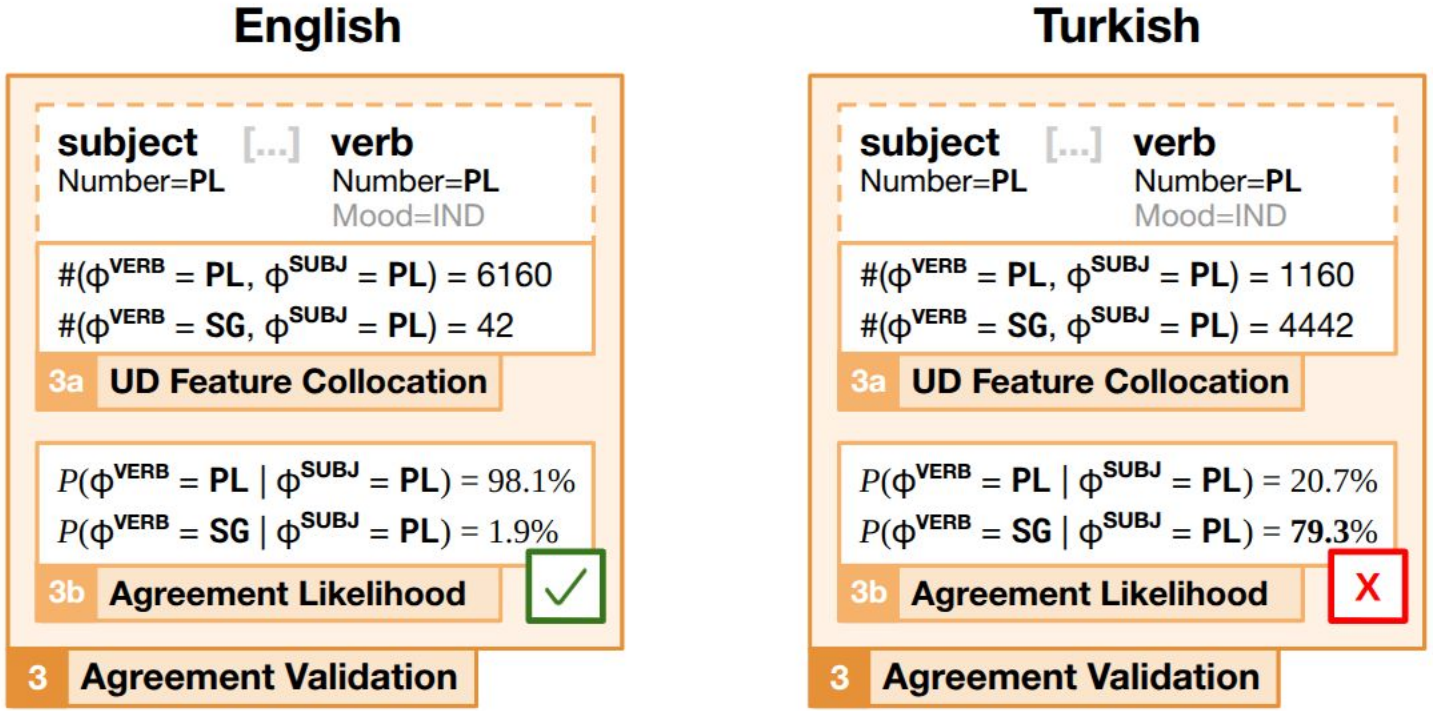
MultiBLiMP Pipeline



Agreement Validation

- How do we ensure our minimal pairs capture *grammaticality*?
- The inflected form **must** raise an agreement violation
  - ... otherwise we'd "punish" the LM for preferring the inflected form even if it would be correct!
- English: *The boys walk* ⇒ **The boys walks*
  N;PL V;**PL** N;PL V;**SG**
- Turkish: *Oğlanlar yürüyorlar* ⇒ *Oğlanlar yürüyor*
  N;PL V;**PL** N;PL V;**SG**

Agreement Validation



Agreement Validation

- We condition agreement on **word order** (SV / VS)
  - E.g. in Dutch 2nd person: ***jij maakt*** / ***maak jij***
  - Future work: condition on *animacy*, *lemma*, *definiteness*, etc.

Agreement Validation

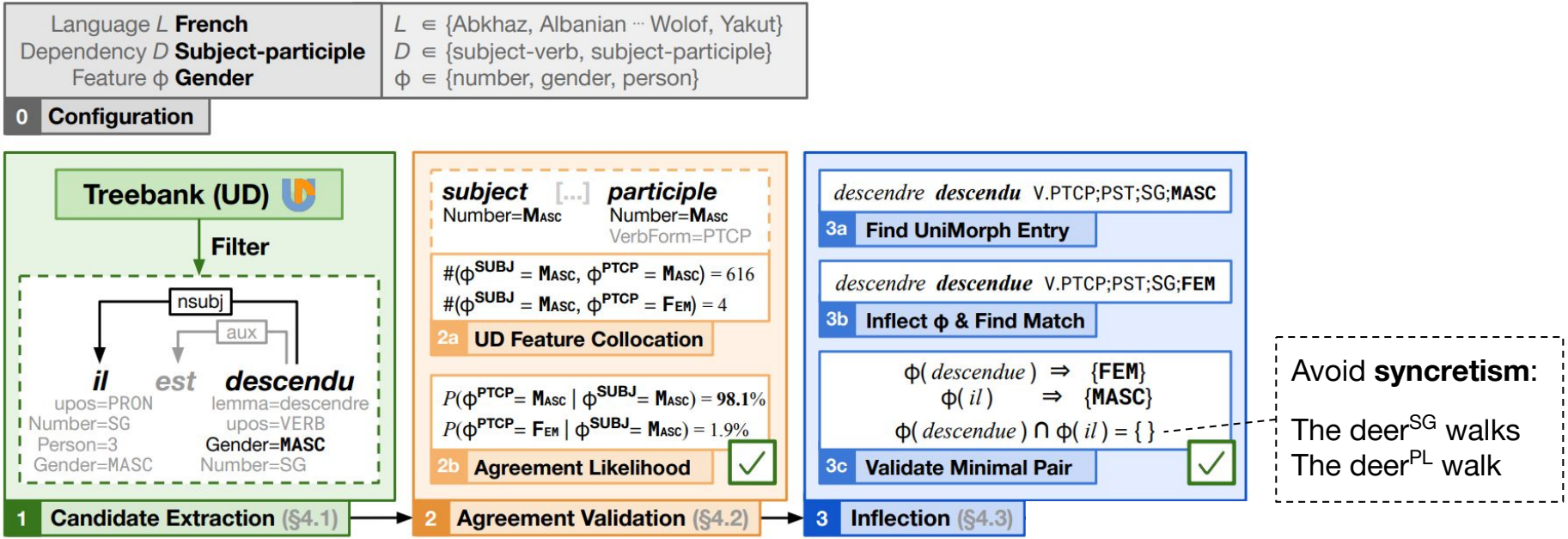
- We condition agreement on **word order** (SV / VS)
  - E.g. in Dutch 2nd person: ***jij maakt*** / ***maak jij***
  - Future work: condition on *animacy*, *lemma*, *definiteness*, etc.
- Measure significance with binomial test: $p(P(X \mid X) > p_0) > \alpha$
- Three conditions ($p_0 = 0.9$, $\alpha = 0.1$):
  - **Certain agreement**:
    Binomial test significant
  - **No agreement**:
    Binomial test significant for $p(P(X \mid X)) < p_0$
  - **Uncertain agreement**:
    Binomial test insignificant in both directions

Agreement Validation

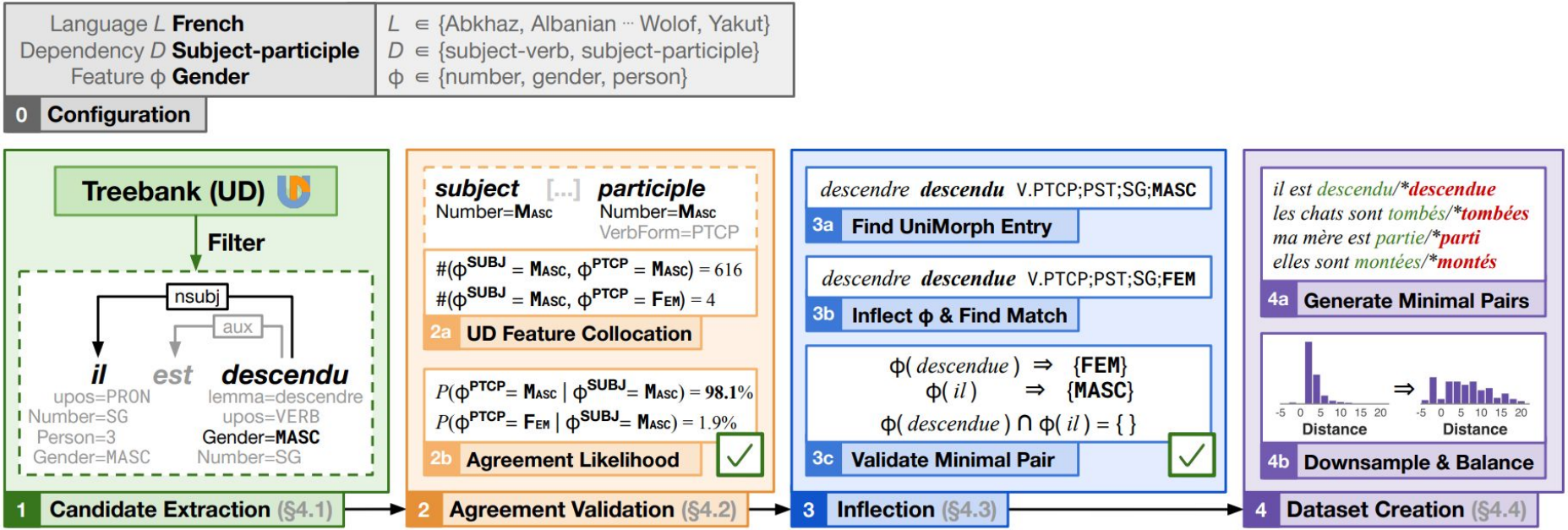
			$P_{agr}(\phi^v \mid \phi^n, \text{WO})$	
ϕ^n	WO	ϕ^v	Dutch	Turkish
PL	SV	PL	**0.989**	0.161
		SG	0.011	0.839
PL	VS	PL	**0.990**	0.000
		SG	0.010	1.000
SG	SV	PL	0.012	0.025
		SG	**0.988**	**0.975**
SG	VS	PL	0.008	0.018
		SG	**0.992**	**0.982**

Table 1: Dutch and Turkish agreement probabilities for subject-verb number agreement. Significant agreement is denoted in **boldface**. WO stands for word order, either subject verb (SV) or verb subject (VS).

MultiBLiMP Pipeline



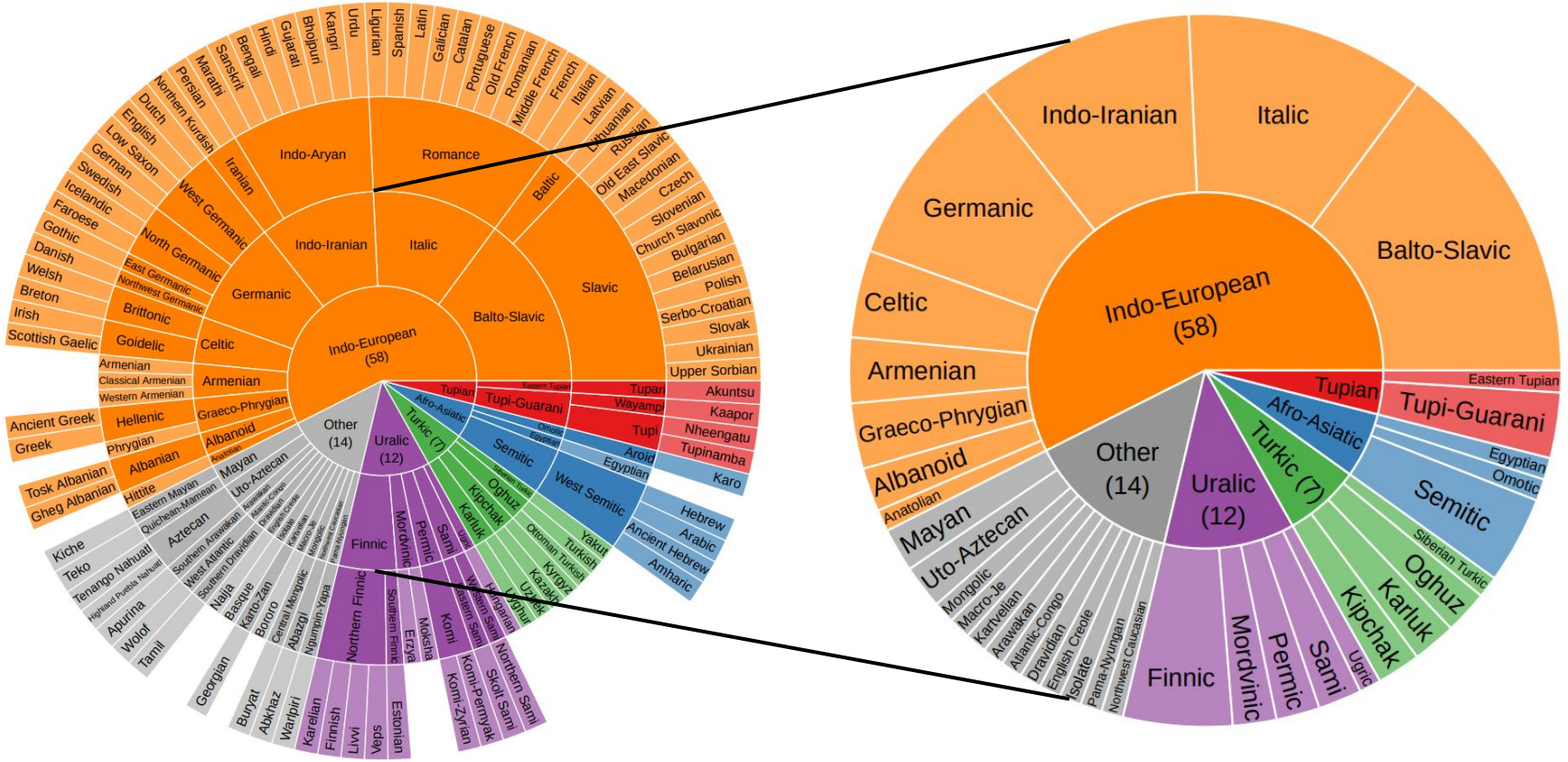
MultiBLiMP Pipeline



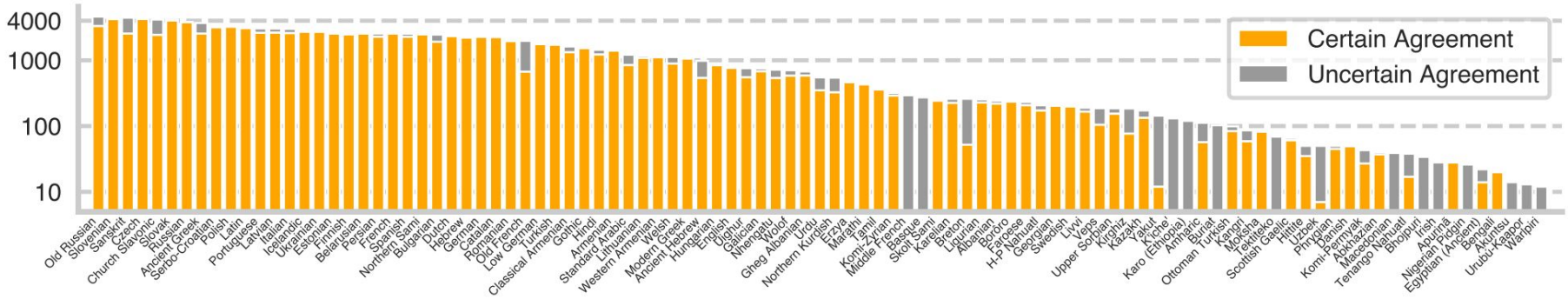
Minimal Pairs

Nhengatu	SV-P	Awá kurí ti uruyari, [**tauyuká**/***peyuká**] kurí arupí aé.
Gheg Albanian	SV-P	dhe ata e [**pan**/***pam**] se ky isht ërrxue .
Wolof	SV-#	Njiitam Séydu Nuura Njaay woyof [**na**/***nañu**]!
Low German	SV-#	De jungens [**lachen**/***lach**] luudhals: »Ney, dat büst du!«
Faroese	SV-#	Í 2008 [**var**/***vóru**] ASFALT tó avlýst.
Latin	SV-G	sese propediem cum magno exercitu ad urbem [**accessurum**/***accessuram**].
Breton	SV-#	Ne [**lennan**/***lennomp**]-me ket al lizher.
Kirghiz	SV-P	Балдар ыйлакташып, кечирим сурашып, экинчи кайталабайбыз деп убада [**беришет**/***бердик**].
Hebrew	SP-#	"מימרה טורקית [**אומרת**/***אומרות**] ,"מי ש יוצא הרבה מביא חזרה לכלוך רב.
Spanish	SP-#	Ninguno de los dos escritores ha [**colaborado**/***colaborados**] en los guiones.
Moksha	SV-#	Весть очижить карта попсь алашаса кудрядс [**ёгась**/***ёгазь**].
Skolt Sami	SV-P	Son [**vuâlgg**/***vuâlgam**], tõid neävveez kuâđđ.
French	SP-G	L'argent qui devait financer le film n'est jamais [**arrivé**/***arrivée**].

Language Distribution



Language Distribution



Minimal Pairs

NUMBER	SG		PL		DU		Total		
	SV	VS	SV	VS	SV	VS	SV	VS	BOTH
S-Verb	88 (74)	82 (63)	73 (58)	58 (43)	8 (3)	3 (2)	89 (79)	84 (65)	90 (80)
S-Participle	35 (33)	28 (21)	23 (18)	19 (13)	3 (0)	1 (0)	35 (33)	30 (22)	35 (33)

63 languages with significant SG|SG agreement in **V**erb-**S**ubject word order

LLM Experiments

Language Models:

- Llama 3
- Gemma 3
- Aya
- OLMo 2
- Qwen3
- EuroLLM

- GoldFish monolingual models (125M parameters, 1GB data / language)
  - Serves as a *monolingual baseline*

Metric:

Accuracy(LM) = %[P_{LM}(gram. sen.) > P_{LM}(ungram. sen.)]

LLM Results

			Subject-Verb			Subject-Participle			Resources			Language Subset					
Model	Size	Version	N	P	G	N	P	G	Low	Mid	High	GF	Aya	EU	Eng	All	#best
Llama3	8B	base	84.2	87.8	88.0	89.4	94.0	88.0	77.2	90.3	96.2	89.4	95.2	92.7	99.4	86.9	0
	70B	base															
	70B	chat															
Aya	32B	chat															
Gemma3	27B	base															
	27B	chat															
OLMo2	32B	base															
	32B	chat															
Qwen3	14B	chat															
EuroLLM	9B	base															
Goldfish	125M																

Low-, mid-, high-resource language subsets

Number of languages significantly better than *all* other models

Language subsets for GoldFish, Aya and EuroLLM models

LLM Results

			Subject-Verb			Subject-Participle			Resources			Language Subset					
Model	Size	Version	N	P	G	N	P	G	Low	Mid	High	GF	Aya	EU	Eng	All	#best
Llama3	8B	base	84.2	87.8	88.0	89.4	94.0	88.0	77.2	90.3	96.2	89.4	95.2	92.7	99.4	86.9	0
	70B	base	**87.4**	**91.2**	91.1	92.1	**97.5**	91.2	**81.1**	93.9	97.8	92.6	97.1	95.5	99.0	**90.2**	2
	70B	chat	86.5	90.3	90.3	91.7	97.2	90.6	80.3	93.1	96.9	91.9	96.2	94.6	98.3	89.3	0
Aya	32B	chat	82.9	87.8	88.1	87.4	95.1	87.0	75.7	89.4	97.7	89.0	97.3	92.8	98.4	86.4	1
Gemma3	27B	base	87.2	91.1	**91.3**	**92.8**	96.6	**91.7**	78.3	**96.3**	**98.0**	93.2	**97.4**	**97.1**	98.6	**90.2**	3
	27B	chat	82.7	87.0	86.3	88.7	95.8	86.7	73.1	92.3	94.4	88.9	93.7	93.3	96.0	85.8	0
OLMo2	32B	base	79.8	85.0	80.2	85.7	88.2	80.9	74.4	84.6	92.5	85.1	90.8	87.8	**99.5**	82.7	0
	32B	chat	78.2	83.9	79.1	84.2	87.1	80.0	72.5	83.6	91.9	84.0	90.0	86.9	99.1	81.5	0
Qwen3	14B	chat	82.2	86.4	86.4	87.8	91.5	86.6	74.1	89.9	94.8	88.2	93.8	92.2	98.3	85.3	0
EuroLLM	9B	base	82.7	86.5	87.6	89.1	71.5	89.4	72.6	92.0	95.7	88.9	94.9	96.7	99.4	85.8	0
Goldfish	125M		92.4	95.3	92.2	95.2	98.2	90.9	88.0	95.6	95.9	**93.8**	95.2	95.8	96.4	93.8	14

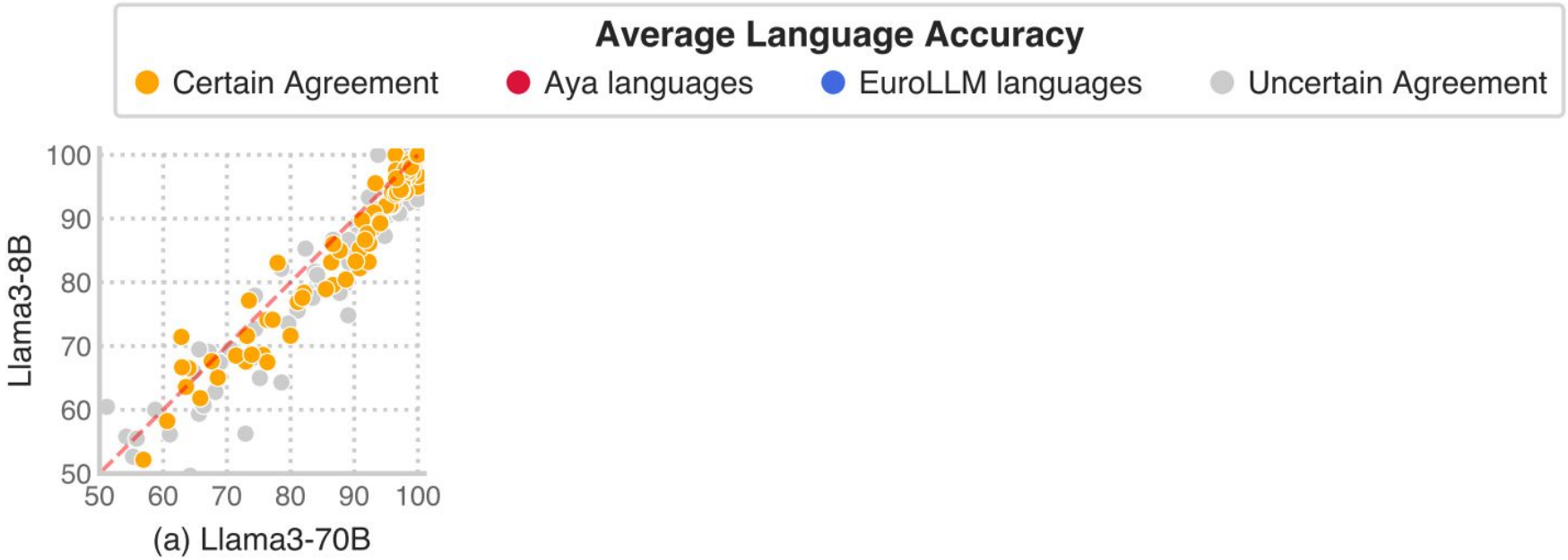
Best LM per Language

Model	Language	Accuracy
Goldfish	Albanian	99.2 ($p = 0.041$)
	Buriat	91.3 ($p = 0.002$)
	Catalan	97.7 ($p = 0.004$)
	Erzya	73.7 ($p = 0.000$)
	Faroese	99.6 ($p = 0.006$)
	Galician	98.3 ($p = 0.001$)
	Icelandic	98.4 ($p = 0.000$)
	Ligurian	94.5 ($p = 0.000$)
	Marathi	95.2 ($p = 0.001$)
	Northern Kurdish	94.7 ($p = 0.033$)
	Northern Sami	96.9 ($p = 0.000$)
	Welsh	99.4 ($p = 0.000$)
	Wolof	97.0 ($p = 0.000$)
	Yakut	95.8 ($p = 0.029$)
Llama3-70B	Church Slavonic	68.1 ($p = 0.009$)
	Gothic	75.4 ($p = 0.000$)
	Komi-Zyrian	77.2 ($p = 0.070$)
	Latin	92.1 ($p = 0.026$)
	Old French	81.8 ($p = 0.001$)
	Old Russian	80.0 ($p = 0.024$)
	Sanskrit	81.4 ($p = 0.000$)
aya-32b	French	99.5 ($p = 0.063$)
gemma3-27b	Bulgarian	99.2 ($p = 0.000$)
	Gheg Albanian	80.4 ($p = 0.050$)
	Polish	98.8 ($p = 0.002$)

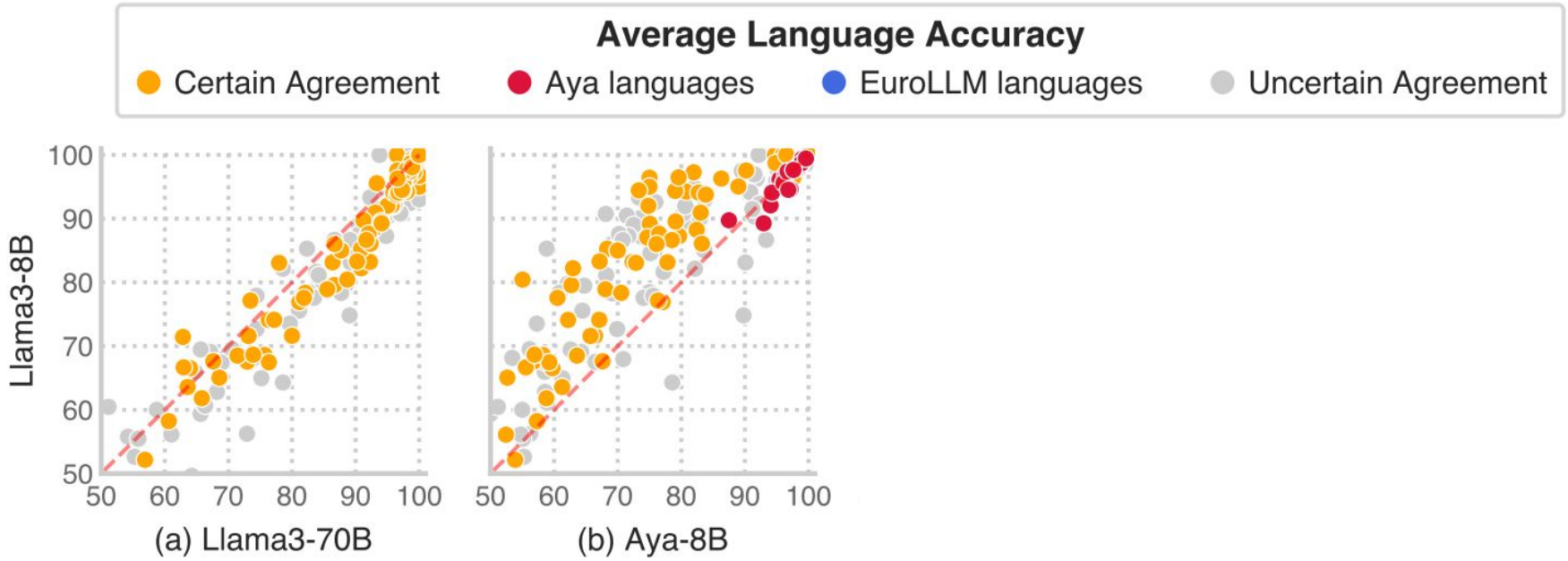
LLM Results

			Llama3				Aya		Gemma3						OLMo2		EuLLM	GPT2	GF
ISO	Language	n	8B	8B-it	tülu3	70B	8B-it	32B-it	4B	4B-it	12B	12B-it	27B	27B-it	32B	32B-it	9B	xl	500M
abk	Abkhazian	40	65.0	62.5	62.5	70.0	55.0	67.5	55.0	27.5	42.5	42.5	**77.5**	70.0	70.0	67.5	47.5	52.5	**77.5**
aqz	Akuntsu	14	35.7	42.9	28.6	42.9	50.0	28.6	35.7	50.0	21.4	21.4	35.7	14.3	35.7	28.6	**57.1**	35.7	—
sqi	Albanian	243	86.0	86.8	84.0	88.5	69.5	83.1	91.4	79.8	95.9	90.9	96.7	92.2	80.7	81.9	61.3	58.0	**99.2**
amh	Amharic	112	**100.0**	97.3	76.8	98.2	97.3	99.1	95.5	82.1	96.4	89.3	96.4	91.1	96.4	93.8	94.6	91.1	96.4
grc	Ancient Greek	3695	86.8	86.3	85.4	90.8	78.9	87.7	77.8	64.8	85.7	73.2	87.7	79.9	**91.2**	90.3	87.7	64.5	88.1
hbo	Ancient Hebrew	983	86.9	85.1	82.8	91.8	75.4	90.9	81.8	63.5	91.6	74.4	**92.9**	83.1	90.0	85.7	80.7	66.0	—
apu	Apurinã	28	**96.4**	**96.4**	**96.4**	**96.4**	75.0	92.9	92.9	75.0	**96.4**	89.3	85.7	92.9	**96.4**	89.3	**96.4**	67.9	—
hye	Armenian	1415	96.5	95.3	94.3	**98.4**	79.6	93.7	95.1	84.2	97.3	90.0	97.9	93.9	86.9	84.9	72.7	67.8	**98.4**
eus	Basque	273	94.1	95.2	93.0	95.2	90.8	91.2	93.4	89.0	96.0	94.1	97.1	89.4	90.1	90.5	91.6	89.4	**98.9**
bel	Belarusian	2570	89.2	86.6	86.5	93.1	75.1	84.6	93.4	81.0	95.2	87.6	96.9	92.3	74.7	74.4	80.5	50.4	**97.3**
ben	Bengali	21	90.5	95.2	**100.0**	95.2	71.4	95.2	95.2	90.5	95.2	90.5	95.2	85.7	81.0	85.7	85.7	47.6	**100.0**
bho	Bhojpuri	34	85.3	76.5	70.6	82.4	58.8	82.4	70.6	61.8	76.5	67.6	82.4	79.4	55.9	67.6	67.6	58.8	**88.2**
bor	Borôro	241	66.0	66.0	65.1	64.7	58.5	67.6	61.0	62.2	58.9	58.9	61.0	61.8	63.1	60.2	65.6	**68.5**	—
bre	Breton	260	94.6	93.1	92.7	95.0	81.5	94.2	94.6	78.5	96.9	92.7	97.3	93.5	75.0	64.6	86.2	50.4	**99.2**
bul	Bulgarian	2458	93.6	91.3	90.4	96.0	76.4	85.5	96.8	87.9	97.8	93.9	**99.2**	95.8	89.7	87.8	97.6	60.7	97.2
bua	Buriat	103	68.0	67.0	66.0	73.8	70.9	69.9	71.8	68.9	70.9	67.0	77.7	70.9	67.0	66.0	68.0	68.0	**91.3**
cat	Catalan	2284	94.8	93.3	94.4	96.1	88.8	95.1	94.7	85.3	96.2	90.5	96.4	92.6	90.9	90.2	95.5	65.0	**97.7**
chu	Church Slavonic	4166	63.7	61.2	61.6	**68.1**	59.9	64.2	59.8	56.1	63.8	62.1	66.3	62.3	64.0	61.1	63.0	61.3	—
xcl	Classical Armenian	1623	70.0	67.7	65.6	75.4	60.7	69.6	67.3	58.7	73.2	61.6	**76.9**	70.5	66.4	65.1	64.1	57.7	—
ces	Czech	4256	92.1	91.0	89.5	95.7	94.0	97.0	92.0	81.8	95.2	88.5	96.2	91.5	83.7	82.8	**97.2**	59.2	92.2
dan	Danish	50	**100.0**	98.0	98.0	**100.0**	96.0	**100.0**	**100.0**	86.0	**100.0**	98.0	**100.0**	**100.0**	90.0	88.0	**100.0**	88.0	**100.0**
nld	Dutch	2331	96.1	95.2	95.2	97.7	95.5	98.0	96.3	82.9	97.5	88.8	97.6	89.8	90.8	90.5	**98.2**	66.5	97.3
egy	Egyptian (Ancient)	22	**50.0**	45.5	**50.0**	45.5	45.5	**50.0**	45.5	**50.0**	40.9	45.5	**50.0**	45.5	45.5	40.9	40.9	40.9	—
eng	English	770	99.4	99.4	98.6	99.0	99.0	98.4	98.2	93.4	99.0	95.6	98.6	96.0	**99.5**	99.1	99.4	98.3	96.4
myv	Erzya	464	52.2	55.2	55.6	56.9	53.9	54.1	51.9	50.6	46.3	44.4	52.6	46.3	57.1	55.8	55.2	54.7	**73.7**
est	Estonian	2575	79.6	77.6	76.3	86.7	62.7	74.9	87.8	73.3	92.7	85.1	95.4	89.3	71.3	70.6	94.4	56.0	**96.2**
fao	Faroese	232	79.7	82.8	80.2	88.4	62.5	75.9	81.9	69.8	87.9	83.6	94.8	89.7	83.2	78.4	74.6	56.0	**99.6**
fin	Finnish	2570	91.4	90.6	89.8	94.5	74.5	86.0	94.6	85.7	96.3	91.4	**96.4**	93.9	91.5	91.4	95.2	58.8	96.2

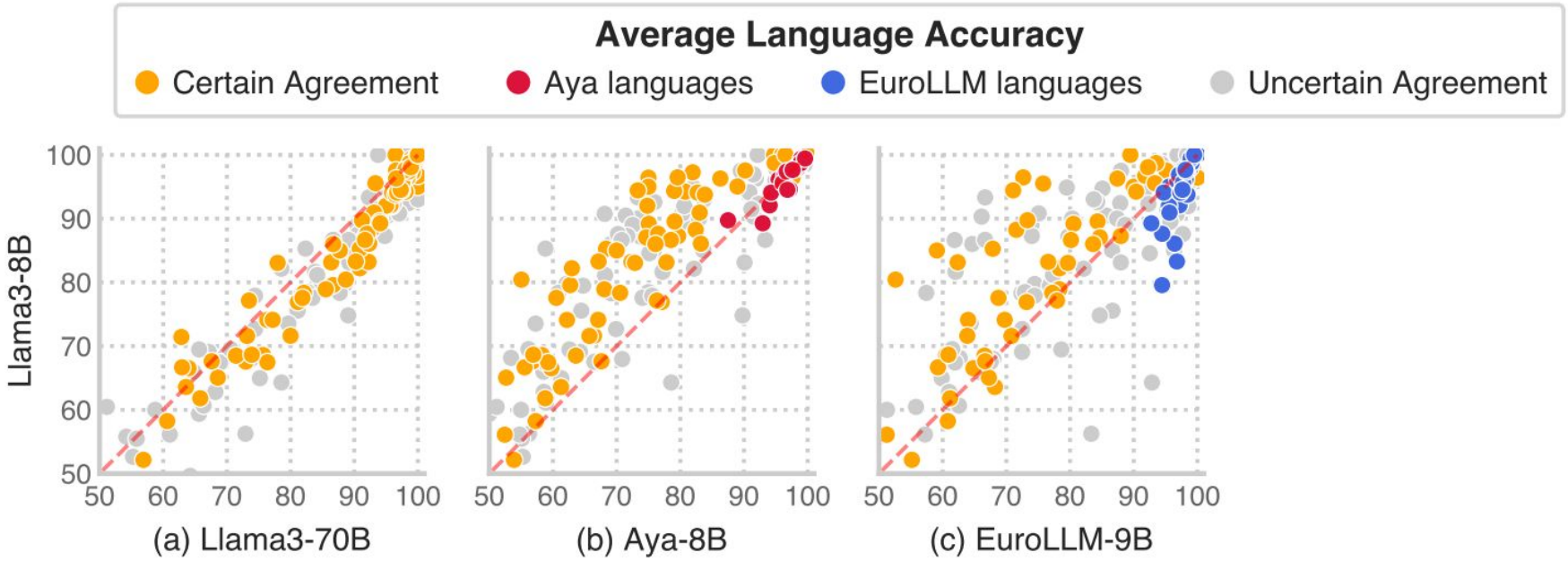
LM Comparisons



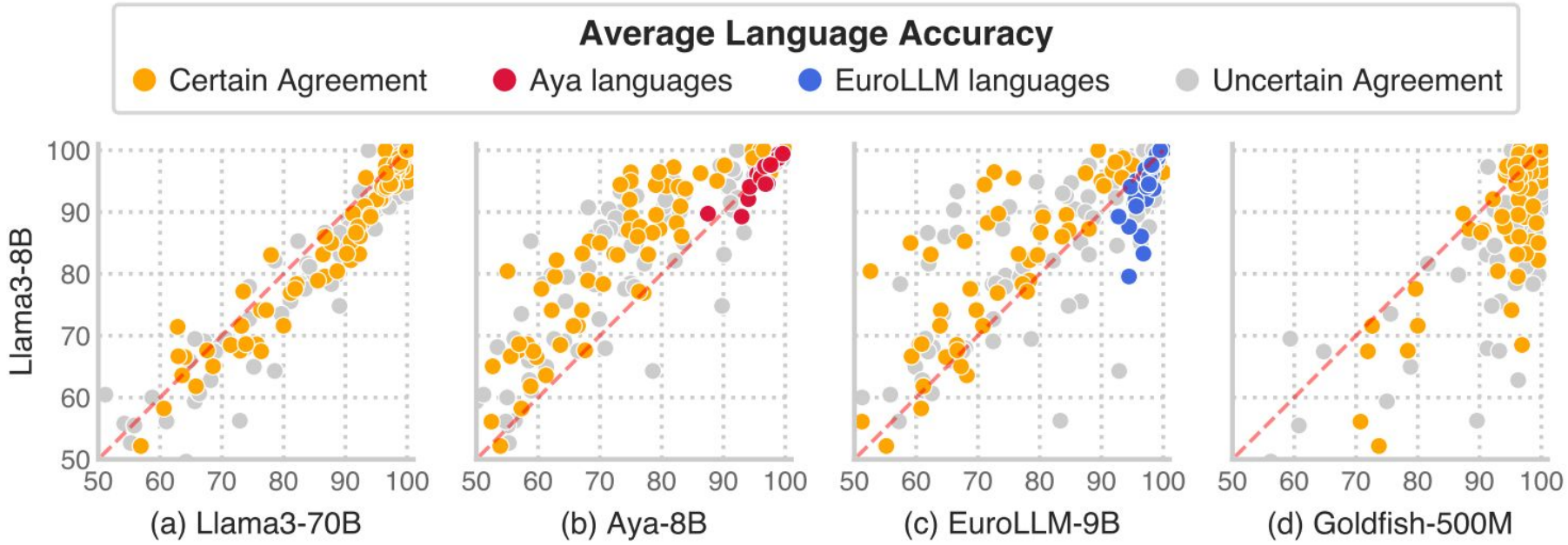
LM Comparisons



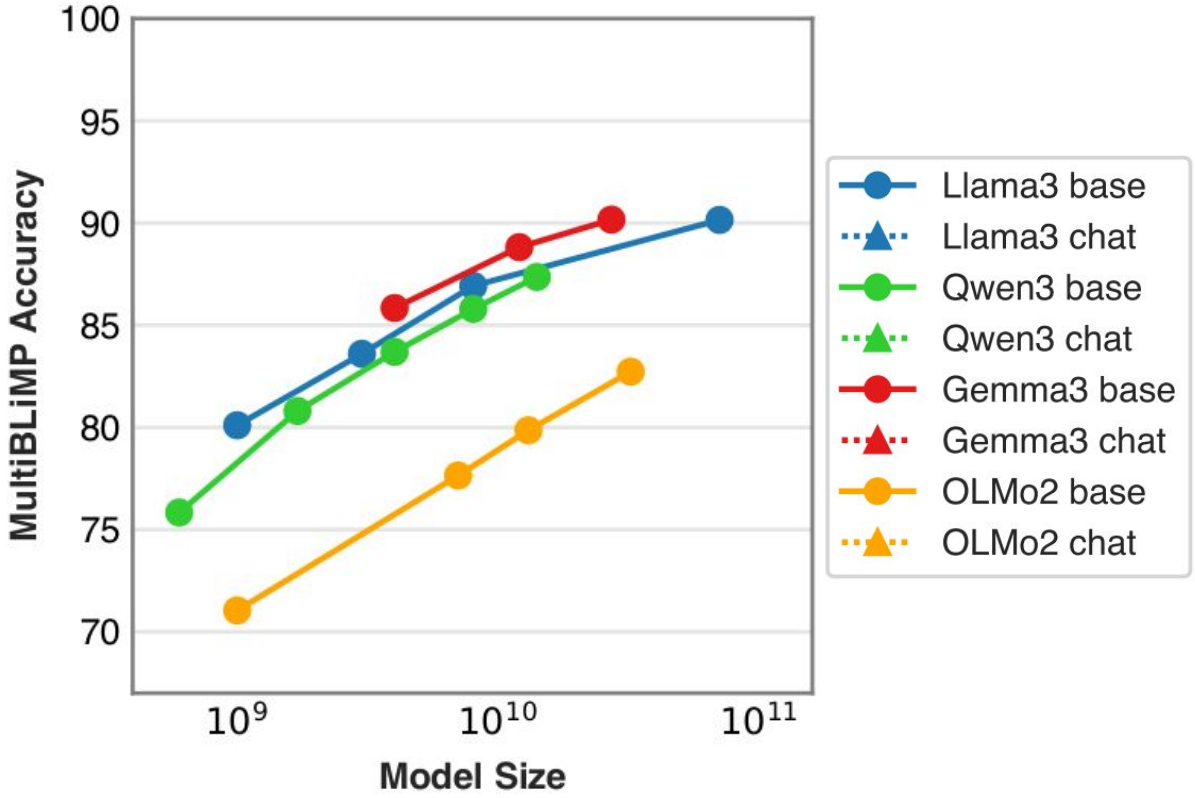
LM Comparisons



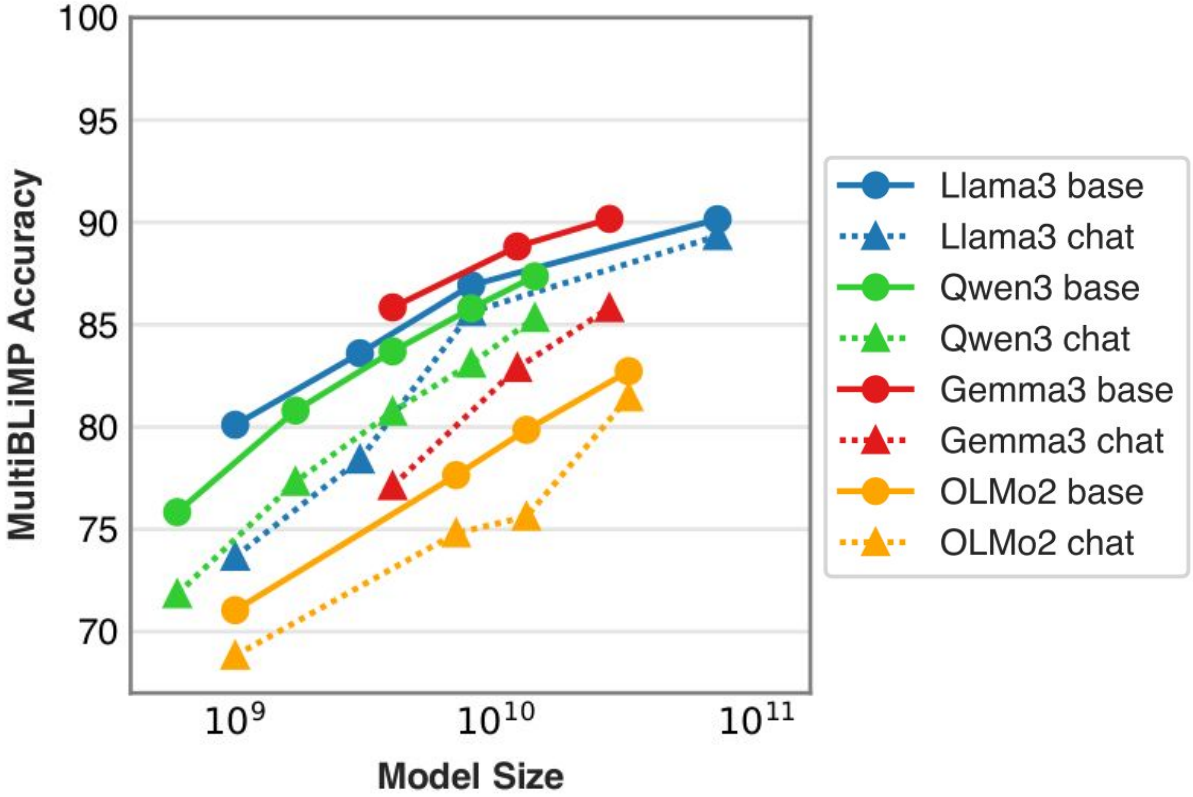
LM Comparisons



Impact of Size & Post-Training

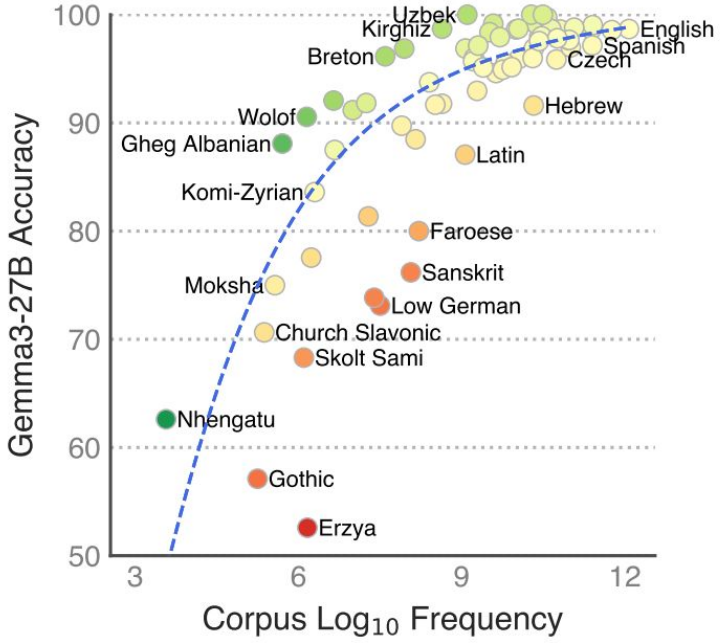


Impact of Size & Post-Training



Impact of language resources

- Language classification on Common Crawl (3.9T bytes)*
- Future work: investigate whether *underperformance* of language is driven by typological features



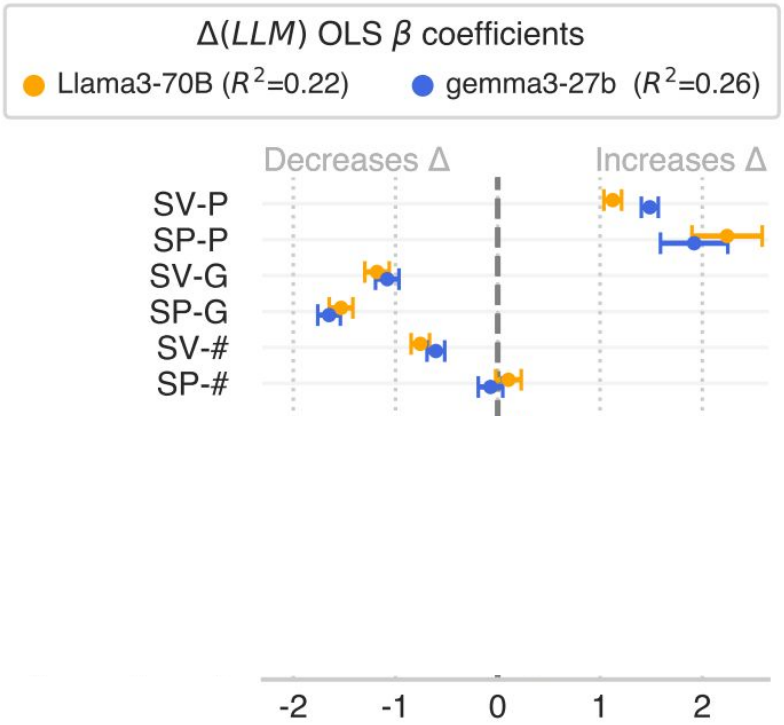
* language frequencies provided by Amir H. Kargaran

Linear Modelling

- What drives linguistic ability in LLMs?
- We fit a linear model to predict the TSE_{Δ}:
  $\ln P(sen^{gram}) - \ln P(sen^{ungram})$
- Coefficients tell us if a feature has a positive impact on grammaticality judgments

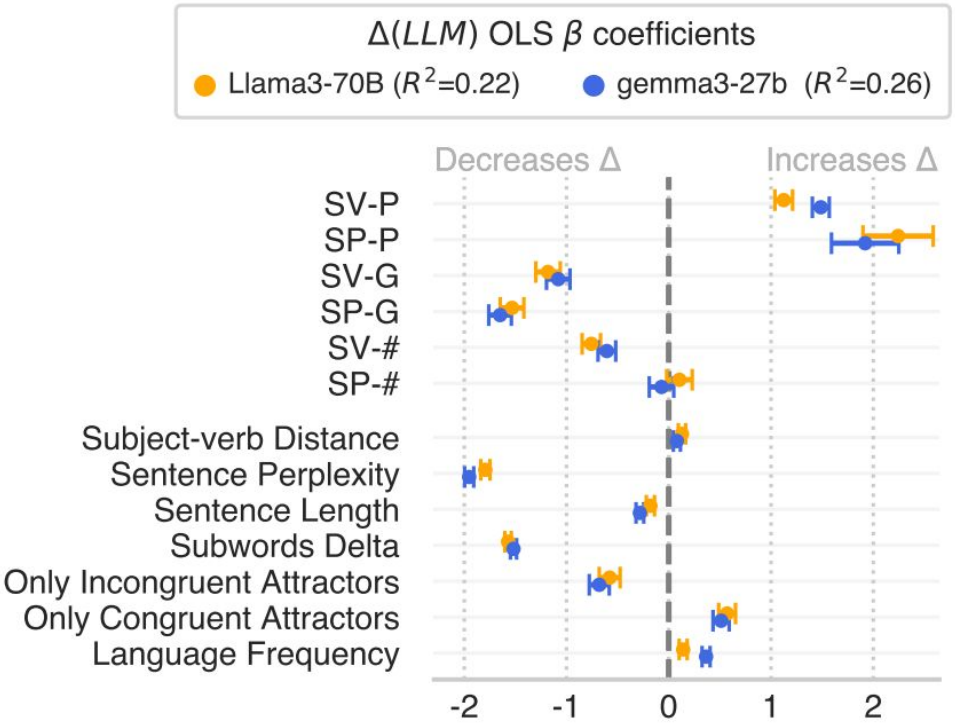
Linear Modelling

- What drives linguistic ability in LLMs?
- We fit a linear model to predict the TSE_{Δ}:
  $\ln P(sen^{gram}) - \ln P(sen^{ungram})$
- Coefficients tell us if a feature has a positive impact on grammaticality judgments



Linear Modelling

- What drives linguistic ability in LLMs?
- We fit a linear model to predict the TSE_{Δ}:
  $\ln P(sen^{gram}) - \ln P(sen^{ungram})$
- Coefficients tell us if a feature has a positive impact on grammaticality judgments



Discussion / Outlook

- Moving beyond agreement:
  - Agreement is very present in Indo-European languages
  - How can we target a wider range of linguistic phenomena?
  - MultiBLiMP v2.0: **word order**

Discussion / Outlook

- Moving beyond agreement:
  - Agreement is very present in Indo-European languages
  - How can we target a wider range of linguistic phenomena?
  - MultiBLiMP v2.0: **word order**
- Use MultiBLiMP for...
  - Multilingual interpretability:
    Do LLMs use similar circuits for the same phenomenon?
    Does English always act as a pivot language?

Discussion / Outlook

- Moving beyond agreement:
  - Agreement is very present in Indo-European languages
  - How can we target a wider range of linguistic phenomena?
  - MultiBLiMP v2.0: **word order**
- Use MultiBLiMP for...
  - Multilingual interpretability:
    Do LLMs use similar circuits for the same phenomenon?
    Does English always act as a pivot language?
  - Language Learnability:
    - Train LLMs on controlled (parallel) corpora
    - What typological features drive language acquisition?

Thanks for listening!

Contact info:
jumeletjaap@gmail.com
@JumeletJ / @jumelet.bsky.social
https://jumelet.ai

Paper: arxiv.org/abs/2504.02768