My research aims to understand what natural language processing (NLP) systems know about language. While generative language models (or LMs, à la ChatGPT) have become mainstream, little is known about *how* these models go from learning next-word prediction to performing complex tasks. Multilingual language models—LMs trained on many different languages simultaneously—are even more opaque as they also learn to transfer information between languages without explicit cross-lingual supervision. While multilingual LMs are much less studied than their English counterparts, these topics have become particularly relevant: almost all large LMs are trained in many languages and have multilingual capabilities, even if this is unintentional [1]. I develop and apply methods for **analyzing models of language** with a strong emphasis on **multilingual and low-resource settings**. In particular, I focus on how LMs capture linguistic phenomena, under the motivation that understanding what models infer about linguistics from pretraining is a good proxy for how they learn in general. I also apply insights from my analysis work to build methods for **more equitable modeling** of all languages, with the goal of extending generative LM's successes in English to other languages and speakers. Throughout my work, I address three research questions:

§1 What do LMs Learn From Next-Word Prediction? While language models are trained on a simple task—predicting the following word in a sequence of text—they acquire many unexpected capabilities. I analyze what models learn about language and linguistic structures from this training signal and how different factors, such as the choice of training data, affect this knowledge. [2, 3, 4]

§2 How do Multilingual Models Differ from Monolingual Ones? Multilingual LMs are trained on text from many different languages and act as the de facto models for non-English languages. This pretraining paradigm leads to different learning dynamics and model behaviors than observed in systems designed for English. My research quantifies these differences and evaluates the effect that multilingual pretraining has on individual languages. [5, 1, 6, 7].

§3 Can Analysis Inform Better Models for Low-Resource NLP? When we understand the limitations of our current systems and the reasons that they fail, we can use this knowledge to inform the development of better data, models, and learning algorithms. I leverage analysis to build better methods and new resources for low-resource languages and other limited data settings. [8, 9, 10]

1 What do LMs Learn From Next-Word Prediction?

While language models have drastically improved over the past decade, quantifying the information encoded by large LMs is a nontrivial task: the models consist of billions of parameters combined in complex functions, which makes directly interpreting independent portions of the model difficult. To address this, I develop and apply new methods for studying emerging NLP technologies. My analyses uncover novel information about LMs, such as how they learn to organize information in their parameters, and give insight into the causes of model shortcomings that can help solve them in the future (§3).

For example, my earlier work found that neural NLP models encode syntax hierarchically, with more general information found at later layers in the network [2]. I discovered this phenomenon with *structural probes*—or small, linear models that recover implicit attributes of text (e.g., part-of-speech) from a model's internal vector representations (Fig. 1)—applied to recurrent neural networks (RNNs) trained on different NLP tasks. Surprisingly, this finding holds for both supervised tasks and the self-supervised signal for language modeling, and it was later corroborated on pretrained LMs with a similar training objective [11]. Since then, probing internal model states has become a standard tool for interpreting transformer LMs.

However, structural probing comes with its own limitations: the method trains new parameters, which makes it hard to disentangle the task knowledge in the model from spurious information learned by the probe. To address this issue and extend probing to generative models, I developed a new *behavioral probing* method, which uses carefully designed inputs to test models for specific skills (Fig. 1, Behavioral Probing). Instead of training an auxiliary probe model, behavioral probing interprets LMs based on how they *behave*

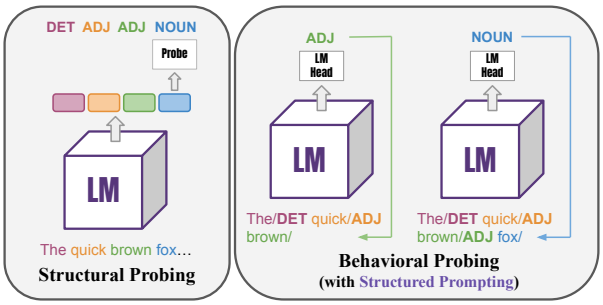

Figure 1: **Structural** and **Behavioral** probing tests language models for knowledge encoded in their parameters.

when given specific inputs in their native language modeling paradigm. My method, **structured prompting**, uses these specific inputs to probe large language models for linguistic knowledge [4]. Because these linguistic features are annotated at the word level, this means extending the prompting setting to be iterative so that the models label each word based on its previous predictions. This approach successfully recovers linguistic knowledge from model parameters, similar to the structural probes, without training new parameters.

In addition to model analysis, we can look to the training data of language models to explain their behavior. Current LMs are trained on terabytes of text data primarily scraped from the web. The exact composition of these data is unknown, and how particular data attributes affect the resulting LM's behavior is often unclear. For example, in [4], I ablated the structured prompting set-up to show that the models primarily learn to perform structured prediction tasks through leaked information about the task, such as labeled examples, that appears in the web-crawled corpus. In [1], I similarly showed that errors in data filtering expose ostensibly English-only models to many other languages during pretraining, effectively making them multilingual. My data analysis work demonstrates that it is crucial to consider the effect of pretraining data when working with large LMs, as it establishes the effect of language and task contamination on model behavior; it has also motivated follow-up works that audit other aspects of the training data [12].

2 How Do Multilingual Models Differ from Monolingual Ones?

As models grow and become more resource intensive, the ability to represent multiple languages within the same pretrained model is increasingly important; it would be infeasible to train individual large LMs for the approximately 7,000 languages worldwide. However, trying to encode many languages in the same model comes at a cost, as the absolute performance on individual languages decreases [13]. This limitation is referred to as the *curse of multilinguality*, which occurs when individual languages compete for limited model capacity. This curse affects languages unequally, with lower-resource languages being much more susceptible [14]. My research characterizes how this competition causes multilingual LMs to differ from English ones, and how these differences affect low-resource languages in particular.

Multilingual LMs also learn to perform *cross-lingual transfer*, or translate information between languages. This skill is essential, as we often have data in English that we would like to leverage when interacting with the model in other languages. However, it is unclear why multilingual models learn to do this. To gain insight into this, I studied the training dynamics of multilingual LMs to understand *when* cross-lingual transfer arises [5]. Specifically, I trained a multilingual LM from scratch on 100 languages and stored intermediate models as snapshots of different pretraining steps. I then probed the intermediate models to form a timeline of when various multilingual skills arise. While language-specific features are acquired early on, cross-lingual transfer is learned and refined throughout pretraining. Surprisingly, model performance can degrade between the intermediate checkpoints and the final model, and some languages perform better over time while others suffer. These fluctuations in performance demonstrate inter-language competition for parameters, highlighting how the curse of multilinguality develops during training.

My research has also found that the issues of multilinguality extend to ostensibly monolingual models: English-only LMs are unintentionally trained on trace amounts of non-English text that they manage to generalize from. This *language contamination* is due to automatic filtering errors during data preprocessing, which leads to out-of-English generalization by these models [1]. This result also sheds light on large LMs' recently observed multilingual abilities. While these models are trained on webscraped data that do not consider language distribution, I showed LMs can capture new languages from even tiny proportions of

text. This finding indicates that **all large LMs are multilingual**, albeit unoptimized for most languages.

An intuitive solution to the curse of multilinguality is to simply increase the model's capacity. However, though recent LMs contain hundreds of billions of parameters, non-English performance continues to lag behind. Based on current trends, model size alone will be unable to solve the problems of multilingual modeling. Instead, we need more creative solutions that leverage domain knowledge and a solid understanding of current model limitations, as I describe in the next section.

3 Can Analysis Inform Better Models for Low-Resource NLP?

As large language models increase in size and become less interpretable, understanding *why* they have certain limitations through their behavior and their data remains the best way to build systems that are fair, safe, and robust. I use insights about the shortcomings of current models to develop better data and methods, particularly for settings with (very) limited data. In this vein, I have built multilingual LMs that better model low-resource languages and improved how NLP systems handle rare language phenomena more generally.

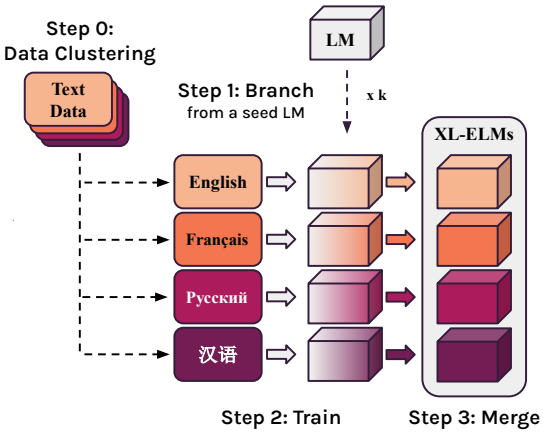

Figure 2: Specializing cross-lingual expert LMs (XL-ELMs) to different sets of multilingual data.

I have illustrated how the curse of multilinguality leads to sub-par representation and performance on most languages in multilingual LMs [5, 6]. Inspired by this, and in particular by the insight that the models *can* learn many features for low-resource languages that are forgotten in the final model state, I proposed a new method for training multilingual LMs that explicitly distributes model parameters to different sets of languages with sparse language modeling, called **cross-lingual expert language models** (XL-ELMs, Fig. 2). First, I automatically cluster a multilingual pretraining corpus into different subsets where similar languages (e.g., Russian and Bulgarian) share a cluster. I then initialize a set of language models with a pretrained seed LM and train separate, independent models on each data subset. This setup allows each expert model to specialize on a specific data cluster, resulting in better language modeling performance than training one model on *every* language [10]. Furthermore, the proposed approach is much more computationally efficient than traditional LM training on the same data.

I have also used findings from model analysis to build better systems for word sense disambiguation (WSD), which is the task of identifying the meaning, or sense, of a word given a specific context (such as distinguishing when the word "bank" refers to a financial institution rather than a river "bank"). A common weakness in WSD systems is poor performance on infrequent senses of words compared to more common senses. I demonstrated that this issue still occurs when we apply LMs to word sense disambiguation, despite the increased information the models see during pretraining [15]. We find that this imbalance stems from limited data for uncommon senses—so in [8], I built a new dataset focusing on rare senses using Wiktionary, which contains many specialized and new senses not covered by existing WSD resources. I then showed that augmenting the models from [15] on this new data improves performance on rare senses, even on existing benchmarks. My work in this area has thus exposed issues in capturing all senses by pretrained models and existing WSD benchmarks, and has since motivated new methods that better handle rare senses.

More recently, larger language models have performed poorly when prompted for word sense information despite the paradigm's success on many other tasks. I built upon insights from these models' performance on contextual word-level translation to design a more natural prompting setup for zero-shot, cross-lingual WSD [9]. As a result, I show that the more artificial prior prompting approaches misestimated sense knowledge in LMs. While large LMs *do* contain significant word sense knowledge in many languages, we need to probe them appropriately to retrieve this knowledge.

Future Work

The future of NLP is multilingual. As language models and their training corpora grow in scale and diversity, the divide between monolingual and multilingual systems has waned, and the need to understand their data and behavior across languages becomes increasingly important. I will expand my work on how LMs learn new abilities from pretraining and develop new lines of research in multilingual modeling, documenting the composition of pretraining data, and analyzing the linguistics of generative models.

Breaking the Curse of Multilinguality NLP for non-English languages continues to lag far behind the impressive results we see touted in English. This gap is significant, both in terms of providing usable NLP systems outside of English and because issues in one language have compounding effects in a multilingual system. For example, recent work found gaps in GPT-4's safety constraints when prompted in low-resource languages, which anyone can use to generate harmful content via a translation API [16]. Therefore, building robust multilingual systems is critical for safer and more equitable NLP technology in general.

While current multilingual models trade individual language performance for cross-lingual information sharing, my future research aims to allow models to achieve both through **data and algorithmic improvements**. A promising direction is the development of methods for allocating parameters and other computational resources across languages to prevent low-resource languages from being overwhelmed by higher-resource ones. I will also address the open question of *how* to perform **rigorous cross-lingual evaluation of LMs** by following up my work on standardizing multilingual evaluation ([7]) and on crowdsourcing new data, as we are doing to obtain gold annotations in many languages through the Universal NER project [17]. Beyond better representation of languages, robustly multilingual models will also afford novel settings for testing linguistic questions about what LMs can learn outside of a single language.

Connecting Data to Model Behavior Training on vast amounts of text is critical to the success of large LMs, and understanding the text we put into our models is crucial for understanding their behavior. However, the composition of these data is unknown, leading to misinterpretations of model performance [1, 4]. I will study the effects of pretraining data on generative AI by **characterizing influential aspects of the text corpus** and identifying **how these aspects affect model behavior**. For instance, it is unclear whether the context or language in which a model learns specific facts affects how it expresses that information. Understanding the underlying data and its effect on the model will provide ways to better extract information from LMs—such as by choosing a specific query language per fact—and verify its accuracy. In cases where pretraining data is not publicly available, I am interested in developing **methods for inferring the training data composition** from models directly, building on our work in [3, 18].

The Linguistics of Generative Models While text generated from large LMs is usually high-quality, these models often adopt a particular writing style that is easily recognizable by humans. This raises the question of **how machine-generated text differs from that of humans** and presents a new field of study into a communication landscape that includes machines. This field will involve questions such as whether certain phenomena more commonly occur in machine-generated text due to the pretraining process—for example, whether specific syntax structures are more common in generated text due to the autoregressive nature of LMs. These new linguistic questions extend to **human-model interaction**. For instance, I plan to assess the extent to which humans *accommodate* (or adapt their speech style to match) LM agents when interacting with them. Research into this new area of linguistics will inform us of differences in how humans and models learn language as well as aid new methods for detecting AI-generated text.

Overall, my goal is to perform empirically driven research into what computational models learn from natural language and what these patterns of learning in return tell us about the languages we speak. In doing so, I hope to foster an ethos of open science in the models and data we choose to research and to facilitate equitable access to NLP technology, particularly in under-represented languages and communities.

References

[1] **Terra Blevins** and Luke Zettlemoyer. Language contamination helps explain the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022.

[2] **Terra Blevins**, Omer Levy, and Luke Zettlemoyer. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018.

[3] Hila Gonen, Srini Iyer, **Terra Blevins**, Noah A. Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023.

[4] **Terra Blevins**, Hila Gonen, and Luke Zettlemoyer. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023.

[5] **Terra Blevins**, Hila Gonen, and Luke Zettlemoyer. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022.

[6] CM Downey, **Terra Blevins**, Nora Goldfine, and Shane Steinert-Threlkeld. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Multilingual Representation Learning (MRL) Workshop*. Association for Computational Linguistics, 2023.

[7] Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, **Terra Blevins**, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*, 2023.

[8] **Terra Blevins**, Mandar Joshi, and Luke Zettlemoyer. FEWS: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[9] Haoqiang Kang, **Terra Blevins**, and Luke Zettlemoyer. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. *arXiv*, 2023.

[10] **Terra Blevins**, Margaret Li, Suchin Gururangan, Hila Gonen, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models. *In Preparation*, 2023.

[11] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[12] Eleftheria Briakou, Colin Cherry, and George Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.524. URL https://aclanthology.org/2023.acl-long.524.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[14] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, Online, July 2020. Association for Computational Linguistics.

[15] **Terra Blevins** and Luke Zettlemoyer. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[16] Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv*, 2023.

[17] Stephen Mayhew, **Terra Blevins**, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. Universal NER: A gold-standard multilingual named entity recognition benchmark. *arXiv*, 2023.

[18] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, **Terra Blevins**, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv*, 2023.