Combining Observational Data and Language for Species Range Estimation

Max Hamilton[1]   **Christian Lange[2]**   **Elijah Cole[3]**   **Alexander Shepard[4]**

Samuel Heinrich[5]   **Oisin Mac Aodha[2]**   **Grant Van Horn[1]**   **Subhransu Maji[1]**

[1]UMass Amherst   [2]University of Edinburgh   [3]Altos Labs
[4]iNaturalist   [5]Cornell University

Abstract

Species range maps (SRMs) are essential tools for research and policy-making in ecology, conservation, and environmental management. However, traditional SRMs rely on the availability of environmental covariates and high-quality species location observation data, both of which can be challenging to obtain due to geographic inaccessibility and resource constraints. We propose a novel approach combining millions of citizen science species observations with textual descriptions from Wikipedia, covering habitat preferences and range descriptions for tens of thousands of species. Our framework maps locations, species, and text descriptions into a common space, facilitating the learning of rich spatial covariates at a global scale and enabling zero-shot range estimation from textual descriptions. Evaluated on held-out species, our zero-shot SRMs significantly outperform baselines and match the performance of SRMs obtained using tens of observations. Our approach also acts as a strong prior when combined with observational data, resulting in more accurate range estimation with less data. We present extensive quantitative and qualitative analyses of the learned representations in the context of range estimation and other spatial tasks, demonstrating the effectiveness of our approach.

1 Introduction

Collecting sufficient point-based observations of species in the wild allows us to infer species range maps (SRMs), which describe the spatial extent of where a species is likely to occur. These maps are invaluable, enhancing our understanding of natural history and informing land use and conservation decisions. Large-scale citizen science projects like iNaturalist [2] and eBird [30] have recently accelerated SRM generation by systematically consolidating millions of observations across tens of thousands of species. By combining these extensive databases with environmental covariates, we can produce accurate SRMs. However, there remains a 'long tail' of species with few observations, and current methods fall short of producing reliable range maps in such low-data settings.

We propose learning SRMs by combining citizen science observations with text descriptions of species from Wikipedia (see Figure 1). These text descriptions describe a wide array of properties including habitat preferences, range estimates, and visual attributes. Our framework learns to map location embeddings over the Earth's surface and text embeddings from Wikipedia [4] articles into a common space based on the species observations (see Figure 2). This provides a way to ground the information in text describing tens of thousands of species to spatial locations based on millions of observations. As shown in Figure 1, our resulting model allows us to estimate SRMs based on text descriptions that might be known to an ecologist, such as habitat preferences, even when no location observations are available.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

| Description of Species' Range | | Predicted Range |

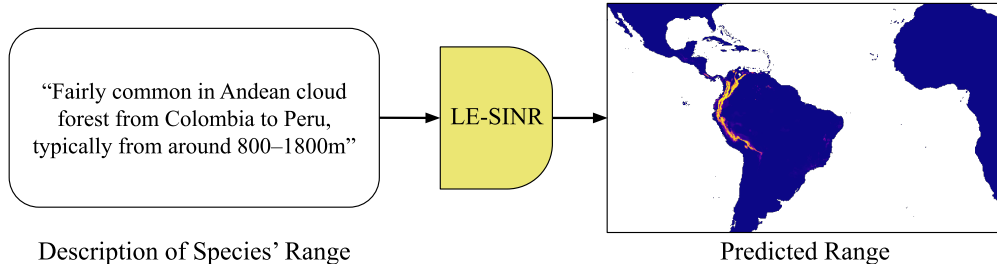



Figure 1: Our LE-SINR model takes as input free-form text describing aspects of a species' preferred habitat or range and geospatially grounds it to generate a plausible range map for that species.

We evaluate our text-driven approach for zero-shot estimation of SRMs for species from the IUCN [1] and eBird Status and Trends (S&T) [15] benchmarks from [10], which were excluded from the training data. Our model easily outperforms baselines (Table 1) and is competitive with SRMs estimated with ten observations for S&T species (Figure 3). Additionally, our model can be combined with observational data to achieve strong few-shot performance. Logistic regression models are regularized to be close to the species vector obtained from text descriptions, allowing us to match the performance of SRMs estimated using an order of magnitude more observations (Figure 3).

While training is based on Wikipedia articles, it is unrealistic to expect a biologist to provide such extensive information for a novel species. Therefore, we evaluate our model based on short summaries, often just a few sentences long, that describe various aspects of the species. We further organize the summaries based on descriptions of the range (e.g., where they appear) and habitat preferences. While a biologist might not know the full range of an unknown species, they may be familiar with some of the latter (e.g., a tree might be in a tropical forest, or a bird might prefer wetlands). However, training on both forms of data allows the location embeddings to capture a wide range of geographic concepts embedded in language, such as countries, continents, climate regions, topology, and biomes, as revealed in our visualizations (Figures 4 and 5). Even short text summaries provide a significant boost in estimation accuracy over baselines when there are zero or few observations available. Our code and data is publicly available at: `https://github.com/cvl-umass/le-sinr`

2 Related work

Species distribution models (SDMs). SDMs are a broad family of models, primarily from the ecological statistics literature, that are concerned with modeling and predicting different geospatial properties of a species of interest [14; 13]. These spatially varying properties encompass quantities such as occurrence (i.e., the presence or absence of a species) through to abundance (i.e., a count of the number of individuals from a given species present). By integrating occurrence predictions over the earth, we can generate the *range* of a species, which is defined as the geographic area in which a species can be found during its lifetime. Existing works can be broadly categorized based on the completeness of the data they are trained on (e.g., presence-only vs. presence-absence data), the number of species they simultaneously model (e.g., single vs. joint methods), or how interpretable they are (e.g., machine learning vs. mechanistic models). For an introduction to SDMs, we point interested readers to the following survey [6].

Most relevant to our work is the growing number of deep SDM approaches. These methods explore the core task from a representation learning perspective by training on raw observation data, typically from multiple species simultaneously, to learn an encoding of geographic space that is more predictive of species presence. [10] demonstrated the advantages of this approach by showing that model performance improves when trained on larger amounts of data, even if that data comes from disjoint species that do not appear in the evaluation set. Recent work has explored various challenges and design decisions related to training these models, addressing topics such as data imbalance [34], spatial biases [8], location encodings [27], the use of remote sensing data [11; 31], active learning [19], binarizing range maps [12], and modeling species co-occurrence [9]. However, the current literature has not extensively investigated the few-shot setting, where very limited or potentially no observation data is available. There are an estimated nine million species on Earth [23], and given that only a limited proportion of these have reliable range estimates, there is a need for methods that can reliably estimate geospatial properties of interest from few observations.

2

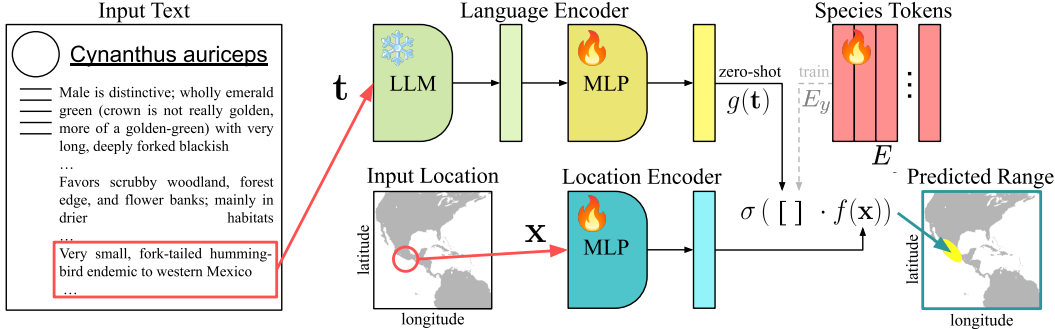



Figure 2: LE-SINR learns to align location and text representations at training time using presence-only observation data and habitat or range descriptions for a set of species. Optionally, we can also include a learnable species token E_y allowing range estimation for seen species y from the training set. The model is trained on millions of observations from iNaturalist and language data from Wikipedia articles across thousands of species. LE-SINR supports zero-shot range estimation based on text descriptions for novel species and can also be used as a prior for few-shot range estimation.

Geospatial data and large language models (LLMs). High-capacity transformer-based architectures [32], coupled with large web-sourced text training data, have largely contributed to recent advances in LLMs. LLMs have been demonstrated to be effective across a range of language-based reasoning tasks [35; 21; 7]. Inspired by this, recently there have been multiple attempts to explore what, if any, geospatial information is encoded inside of these models. For example, [25] evaluated a pre-trained closed-source LLM (i.e., GPT-4 [5]) on a range of geospatial tasks such as point based ones (e.g., location, distance, and elevation estimation) in addition to more complex path-based ones (e.g., geographic and outlines route planning) via carefully designed text prompts. They later extended this work to multimodal LLMs that can also take images as input [26]. In both cases, they observed impressive capabilities, but also some notable limitations.

In [22] the authors used pre-trained LLMs to map geographic coordinates to continuous geospatial properties (e.g., population, house value, etc.). They improved upon simple text prompts by engineering a prompt which provides spatial context in terms of relative distance to nearby named locations to the query coordinate of interest. By fine-tuning the LLM they demonstrated that their approach works better than the more naive encoding. However, their approach is expensive to evaluate at inference time as it requires a full forward pass of the LLM for every geographic coordinate of interest.

Most related to our work is LD-SDM [28] which also uses an LLM in conjunction with an SDM. Their goal is to predict the spatial range of a set of species of interest using location observation data at training time. However, instead of simply learning a per-species latent embedding vector as in [10], they employ a frozen LLM to map a text string that describes the explicit taxonomic hierarchy (i.e., species, genus, family, etc.) of a species of interest to a latent embedding. Unlike us, their species text description is not very expressive (i.e., it has a very specific hierarchical structure) and thus cannot generalize as effectively to distinct held-out species at evaluation time. Instead, in this work, we show that it is possible to predict the range of a previously unseen species from free-form, highly unstructured, internet sourced text that describes its habitat and/or range preferences. Furthermore, once trained, we show that our approach is also able to efficiently and densely geospatially ground non-species related text. Unlike [22], our approach is computationally efficient at inference time, requiring only one LLM forward pass for the text query of a species of interest, as opposed to one forward pass for each location of interest, which can number in the millions depending on the spatial resolution at which the evaluation is performed.

3   Methods

3.1   Problem Setup

We focus on the problem of estimating SRMs across multiple species, indicating their presence or absence at each location on Earth. For a given location $\mathbf{x}$, our goal is to predict the probability of each species being present there. Our observation dataset is composed of pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where

3

$\mathbf{x}_i = (lat, lon)$ is a geographic location and $y_i \in \{1, \dots, S\}$ is the observed species label. We use the dataset proposed in SINR [10], consisting of 35.5 million observations covering 47,375 species observed prior to 2022 on the iNaturalist platform. Included species have at least 50 observations. Removing those that are included in the S&T and IUCN evaluation benchmarks results in 44,181 species, which forms our training set.

We also consider estimating SRMs from text-based descriptions in addition to observational data. To this end we curate a text dataset $\{(\mathbf{t}_i, y_i)\}$, where $\mathbf{t}_i$ is a text description sourced from Wikipedia pertaining to species y_i. For a particular species y_i and corresponding text description $\mathbf{t}_i$, we aim to predict whether y_i is present at some location $\mathbf{x}_i$ given $\mathbf{t}_i$. It is important to emphasize that this prediction only depends on the text description $\mathbf{t}_i$ and query location $\mathbf{x}_i$. Thus, we are not constrained to only making range predictions for species that have been observed during training. This allows for zero-shot species range estimation, where we can generate a range map of a species not in our observation dataset by utilizing a text description for it.

3.2 Text Data

We source our text descriptions from Wikipedia [4]. For a particular species, we search for its Wikipedia article using its scientific name. We then extract the text and divide it into chunks by section. Every article starts with a lead section, which typically provides an overview, followed by a varying number of body sections. These body sections can describe a diverse range of attributes such as taxonomy, description, habitat, behavior, diet, etc. To improve data quality, we remove sections with "References," "Links," and "Bibliography" in their names. Since each species has multiple sections, at training time we randomly sample one section per iteration. Overall, our text dataset contains 127,484 sections from 37,889 species' articles. Note, not all species in our observation dataset have an associated text description.

3.3 Language Enhanced SINR (LE-SINR) Architecture

We utilize the Spatial Implicit Neural Representation (SINR) [10] framework for our approach, which models the probability of presence for a species y at given location $\mathbf{x}$ as $\sigma(f_\theta(\mathbf{x}) \cdot E_y)$, where $f_\theta(\mathbf{x})$ is a location encoding, E_y is an embedding of species y and σ is the sigmoid function. Our architecture, shown in Figure 2, is similarly composed of two branches: one for representing locations and one for representing species. Each branch outputs a 256-dimensional feature vector. The probability of occurrence is then estimated by the sigmoid of their dot product.

The location branch is a location encoder model, $f_\theta(\mathbf{x})$, with parameters θ, which takes a position embedding $\mathbf{x}$ (e.g., a location denoted by latitude and longitude) as input. The species branch is composed of two models, allowing us to generate species embeddings in two different ways. The first is a text-based species encoder, $g_\phi(\mathbf{t})$, with parameters ϕ, which takes text $\mathbf{t}$ from our Wikipedia text data as input. The second species representation is a batch of species tokens optimized directly, $E \in \mathbb{R}^{S \times 256}$. Given a known species $y \in \{1, \dots, S\}$ in the observation dataset, we can generate its representation with a simple lookup, E_y. Since we learn a unique species token for each species in the training set, these species tokens cannot be used in the zero-shot setting. However, we are able to maintain the ability to have true supervised evaluation on species seen at training time. Additionally, the species tokens are used when a species has no text description.

Our text-based species encoder has two parts: a frozen LLM used to extract text embeddings and a learned fully connected network. As mentioned previously, to perform zero-shot SDM, we cannot directly utilize the species tokens. Instead, we input segments of text sourced from Wikipedia articles. Since these text segments can be as long as multiple paragraphs, we utilize a pretrained Large Language Model, GritLM [24], to create a fixed-length text embedding. GritLM is a recent language model with strong performance on text embedding tasks like document classification and retrieval. Due to the length of text in our data, it is much easier to work with a fixed-length embedding than with per-token embeddings from generative models like Llama3 [3]. Given this text embedding, we then learn a three-layer fully connected language encoder network that outputs the final species embedding. As our approach incorporates language within the SINR framework for range estimation, we call it: **Language Enhanced SINR (LE-SINR)**.

4

3.4 Training LE-SINR

During training we only have access to presence observations, i.e., locations where species have been observed, and do not have any absence observations, i.e., locations where species have been confirmed to be absent. As a result, we train our model with a modified version of $\mathcal{L}_{\text{AN-FULL}}$ from SINR [10], which was one of the best-performing losses in their experiments. It minimizes binary cross-entropy with positives sampled from the observation dataset and negatives (i.e., 'pseudo absences') selected to be all other species at the same observation location, as well as all species at a uniformly sampled random location.

This loss is computationally expensive with our model, as it requires computing a species embedding for every species and sample in the batch. To reduce this computation, we perform an approximation by selecting $M-1$ random negative species at the observation location and M random negatives from the random location, where $M \ll S$. We also weight the negatives by the inverse of the proportion selected. This ensures that the expectation over the randomly selected negative species equals the original loss. Our modified loss is given by:

$$\mathcal{L}'_{\text{AN-full}} = -\frac{1}{S} \sum_{j=1}^{M} [\mathbb{1}_{[z_j=1]} \lambda \log(\hat{y}_j) + \mathbb{1}_{[z_j \neq 1]} \frac{S-1}{M-1} \log(1-\hat{y}_j) + \frac{S}{M} \log(1-\hat{y}'_j)]. \quad (1)$$

Here, $\hat{\mathbf{y}} \in \mathbb{R}^M$ are predictions at one location for the ground truth species and $M-1$ random other species, $\mathbf{z} \in \mathbb{R}^M$ is the corresponding one-hot label, and $\hat{\mathbf{y}}' \in \mathbb{R}^M$ are predictions for a random species at a random location uniformly sampled over earth. For all models, we set $M = 192$ based on memory and compute considerations. We also tried values as large as 2,048 but saw significantly slower training times with no effect on zero-shot performance.

During training, we first use the location encoder to generate location features. We then use these to make two predictions: one using the species tokens and the other from the text-based species embeddings. Finally, we apply $\mathcal{L}'_{\text{AN-FULL}}$ to both of these predictions independently. We do not explicitly encourage the species tokens and text-based species embeddings to be close. In our preliminary experiments, this seemed to be too restrictive and hurt performance.

3.5 Evaluation

Similar to SINR, we evaluate our model using expert-derived range maps from the eBird Status & Trends (S&T) dataset [15], which covers 535 bird species with a focus on North America, as well as range maps for 2,418 species from the International Union for Conservation of Nature (IUCN) Red List [1]. During training, we exclude observations for these species to assess zero-shot and few-shot performance, measured using mean average precision (MAP), i.e., average precision (AP) averaged across all species in the set. SINR models trained with target species' observations provide an "upper bound" on performance.

Zero-shot evaluation. Our model naturally supports zero-shot evaluation by providing a text prompt from species not in the training data to the species model. The output species embedding can then be multiplied with position features to generate the probability of occurrence. We can then compute a precision-recall curve by varying the threshold over the score to generate the range map given an expert derived range map. We again report the mean average precision (MAP) across species in the S&T and IUCN datasets.

An important choice here is which text we provide for the evaluation. The Wikipedia section names are not consistent across articles and often the same information can appear under different headings. At the same time, it is unlikely that one can provide such detailed text for novel species. To standardize the information for a realistic evaluation, we use the open-source Llama-3 model [3] to generate two short summaries: a habitat description and a range description from the article text. The range description is most informative as it typically lists specific countries or regions where the species can be found. In practice, such a rich description might not be available, so we also generate a habitat description. During evaluation, we use these summaries instead of the Wikipedia text. Figure 4 shows some example summary texts along with zero-shot range predictions from LE-SINR. Further examples can be found in the Appendix.

5

Few-shot evaluation. While the original SINR model was trained with observations from the target species, we also consider a setting where position features are used to derive SRMs from sparse observational data. We achieve this by performing logistic regression with L_2 normalization on the position features to predict presence or absence. For each species, we sample n_p positives from the observation dataset and $n_n = 20,000$ negatives. To mimic the $\mathcal{L}_{\text{AN-FULL}}$ loss, we sample 10,000 negatives uniformly across the Earth and 10,000 negatives randomly from the training dataset species locations. Our logistic regression loss is given by,

$$L_{\text{reg}} = -\frac{1}{n_p}\sum_{i=1}^{n_p} -\log\left[\sigma(\mathbf{w}^\top f(\mathbf{x}_i))\right] - \frac{1}{n_n}\sum_{i=1}^{n_n}\log\left[1 - \sigma(\mathbf{w}^\top f(\mathbf{n}_i))\right] + \frac{\lambda}{n_p d}||\mathbf{w}||_2^2, \quad (2)$$

where $\mathbf{w} \in \mathbb{R}^{256}$ is the species parameter being optimized, λ is the regularization strength, $d = 256$, σ is the sigmoid function, $f(\mathbf{x}_i)$ is the output of the position branch for the i-th positive observation, and $f(\mathbf{n}_i)$ is the output of the position branch for the i-th negative observation. We use this loss to estimate a SRM when text is not provided at test time. The learned weights can then be applied to all positions to derive a range map.

To incorporate text at test time along with observational data, we modify the regularization to be the distance from the predicted text-based species embedding, resulting in the following loss,

$$L_{\text{reg-tx}} = -\frac{1}{n_p}\sum_{i=1}^{n_p} -\log\left[\sigma(\mathbf{w}^\top f(\mathbf{x}_i))\right] - \frac{1}{n_n}\sum_{i=1}^{n_n}\log\left[1 - \sigma(\mathbf{w}^\top f(\mathbf{n}_i))\right] + \frac{\lambda}{n_p d}||\mathbf{w} - \mathbf{w}_{\text{tx}}||_2^2, \quad (3)$$

where $\mathbf{w}_{\text{tx}}$ is the output of the text-based species encoder when provided a text summary. This encourages the learned species weight vector to be similar to the text derived one. This approach is simple to implement, and our experiments indicate that it is also effective.

3.6 Implementation Details

We closely follow the hyperparameters from SINR for a fair comparison. We train with the Adam optimizer for 10 epochs with a learning rate of 0.0005. The species network has three linear layers with ReLU activation. The input text embedding dimension is 4,096, the hidden dimension is 512, and the output species embedding dimension is 256. During training, at each iteration, we choose a random Wikipedia section to generate each species embedding. For logistic regression in the later few-shot evaluation experiments, we use a regularization strength $\lambda = 20$. Training a single LE-SINR model from scratch using all the text and observational data takes about 10 hours on a single NVIDIA RTX 2080ti GPU occupying about 10GB of VRAM. Wikipedia text embeddings and their summaries were generated once using a distributed GPU cluster.

4 Results

4.1 Zero-shot Range Estimation

In Table 1, we compare against several baseline methods to establish performance lower and upper bounds as zero-shot range prediction is a novel task. The constant prediction baseline assumes a species is present everywhere, while model mean predicts the mean species distribution map for all species. We also show the 'oracle' performance obtained by training a SINR which sees observations from evaluation species at training time. All models were trained with observations capped at $1,000$ with uniform negatives, using the AN_full loss and '+Env' indicates models that are trained using extra environmental features as input, as in [10]. We compare these to zero-shot range estimates obtained using various LE-SINR models and input text.

The results show that the zero-shot range estimates comfortably outperform the baselines, achieving non-trivial performance with no observations. Range text works the best, but even habitat text achieves strong performance on both IUCN and S&T species. Similar to SINR, we observe a boost when including environmental covariates. LE-SINR with explicit species tokens matches the Oracle SINR performance, i.e., when evaluation species are used during training, suggesting that LE-SINR does not lose performance on observed species. Therefore, LE-SINR can be used both to explicitly model observed species and for zero-shot prediction from text for novel species.

6

Table 1: **Zero-shot Range Estimation**. Our LE-SINR approach enables zero-shot range estimation given a range or habitat description for a particular species. All models were trained with a maximum of 1,000 observations per species and uniformly sampled negatives, using the AN_full loss. '+Env' indicates models trained with additional environmental features as input, as in SINR [10]. '+Eval Sp.' indicates models that had observed data from the evaluation species during training. LE-SINR matches Oracle SINR performance when explicit species tokens are available, i.e., the model is trained with evaluation species, and we also show the performance on the observed species using the text-based species encoders for completeness in the Appendix.

	Method	+Env	+Eval Sp.	IUCN	S&T
Oracle	SINR		✓	0.67	0.77
	LE-SINR (Species Token)		✓	0.69	0.78
	SINR	✓	✓	0.76	0.80
	LE-SINR (Species Token)	✓	✓	0.75	0.80
Baselines	Constant Prediction		✓	0.01	0.22
	Model Mean		✓	0.09	0.35
Zero-shot	LE-SINR (Habitat Text)			0.28	0.51
	LE-SINR (Range Text)			0.47	0.61
	LE-SINR (Habitat Text)	✓		0.32	0.52
	LE-SINR (Range Text)	✓		0.53	0.64

Table 2: **LE-SINR Ablations**. Impact of various design choices, such as species description (range and habitat vs. taxonomy in LD-SDM [28]), location encoders, and the loss function on the zero-shot performance of LE-SINR. All models were trained without 'Eval Sp' and using 'Env' when appropriate. LD-SDM† is our model and text encoder but trained with taxonomic text as in [28].

Ablation Type	Method	+Env	IUCN	S&T
Text	LD-SDM† [28] (Genus Text)	✓	0.19	0.30
	LD-SDM† (Species Text)	✓	0.21	0.33
Position Encoder	GeoCLIP Backbone [33] (Habitat text)		0.23	0.49
	GeoCLIP Backbone (Range Text)		0.41	0.58
	Spherical Harmonics Encoding [27] (Habitat Text)	✓	0.31	0.52
	Spherical Harmonics Encoding (Range Text)	✓	**0.53**	0.63
Loss Function	SatCLIP Loss [18] (Habitat Text)	✓	0.26	0.52
	SatCLIP Loss (Range Text)	✓	0.46	0.63
	Ours (Habitat Text)	✓	0.32	0.52
	Ours (Range Text)	✓	**0.53**	**0.64**

Figure 3 compares our zero-shot performance with few-shot range estimation using logistic regression models trained with varying numbers of observations. On S&T the zero-shot performance with habitat text is worth two observations, while range text is worth ten. Figures 4 and A2 show range maps generated from text inputs for various species.

In Table 2 we compare against other training text, position encodings, and loss functions. For the LD-SDM [28] baseline we use our LE-SINR architecture but replace the Wikipedia-derived text with LD-SDM taxonomic classification strings. Unlike LD-SDM's autoregressive LLM, which generates an embedding for each token, this baseline, like our main method, encodes text using a model trained specifically for fixed-size text embeddings. For each species, we construct a string representing different levels of the taxonomic hierarchy, including class, order, family, genus, and species. During training, we randomly select one of these strings in each iteration. We then perform zero-shot evaluations at each taxonomic level. Performance decreases as we move up the hierarchy with species-level text performing the best. For the position encoder baselines we first replace our position encoder with the pre-trained GeoCLIP [33] model. The weights are then fine-tuned during training. We also try the Spherical Harmonics [27] coordinate encoding to encode the latitude and longitude. We experimented with multiple values for L as well as the SIREN network architecture and found $L = 10$ with our original MLP architecture worked best for this experiment. For the SatCLIP [18] baseline, we use our model but replace the AN_{full} loss with their contrastive loss applied to the position and text embeddings. Looking at Table 2, none of these experiments perform better than our method.

4.2 Text-driven Priors for Few-shot Range Estimation

In addition to zero-shot range map generation, we can also use LE-SINR in the few-shot setting, i.e., when there are a small number of location observations available for a previously unseen test species.

7

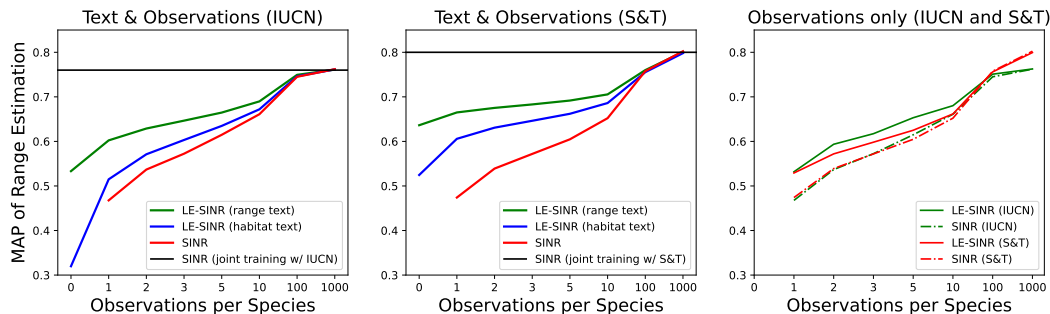



Figure 3: **Range Estimation from Text and Observations**. (Left) IUCN and (Middle) S&T results for zero-shot range estimation based on text, and few-shot estimation based on the text-driven prior. Both range and habitat texts improve few-shot performance over baseline SINR. (Right) Comparison of the position branch of SINR and LE-SINR for range estimation using a few examples. Language-driven covariates learned by LE-SINR lead to better generalization when observations are limited. We report the MAP for range estimation for species in the S&T and IUCN test sets.

Figure 3 shows few-shot results obtained using the text-driven prior from LE-SINR, as described in Section 3.5, on IUCN (left) and S&T (middle) species. Logistic regression models regularized toward species weights obtained from range and habitat text provide a significant boost over vanilla logistic models regularized toward zero. The gap is significant when the training data is limited, i.e., fewer than 100 observations. Range text priors reduce the need for observations on IUCN and S&T by a factor of 3× and 10×, respectively. Figure A2 in the Appendix shows some qualitative examples of how range estimates change with different numbers of observations.

4.3 Evaluation of Learned Positional Embeddings

Prior work on generating SRMs often rely on carefully-selected spatial covariates such as temperature, elevation, precipitation, land cover, etc., to predict the distribution of species based on a few observations. Our approach provides a way to learn a rich set of spatial covariates based on language data. Figure 3 (right) compares the position embeddings of SINR and LE-SINR as spatial covariates on the task of few-shot range estimation. In both cases, we use a model whose location branch is trained with position only (i.e., only latitude and longitude) as input to avoid conflation of the learned covariates with input ones. We find that the intermediate position embeddings learned when trained using language lead to better generalization, especially when training data is limited. Note that here we do not use any language input at test time, as both models are trained simply using observation data, same as the logistic regression baselines.

Figure A1 in the Appendix shows that the position embeddings of LE-SINR have a richer spatial structure than SINR. This figure was obtained by projecting the learned position embeddings to three dimensions using Independent Component Analysis and visualizing them as color in RGB space. Figure 5 visualizes a variety of maps generated from natural language. LE-SINR has learned about geographical regions, climate zones, and even abstract non-species concepts by aligning text representations with geographic locations through species observations. For example, the presence of a species such as the 'Fennec Fox' allows the model to learn that the species is associated with concepts such as the deserts of North Africa, particularly the Sahara Desert, sandy environments, and extreme temperatures based on information in Wikipedia text.

4.4 Limitations and Broader Impacts

While we show that language and location observations can be combined to estimate range maps from a few examples, we do not compare to other few-shot methods such as those based on meta-learning [29; 20; 16]. However, recent results for non-species range estimation tasks indicate that a good underlying representation is a key component to superior few-shot performance, even outperforming more advanced methods [17].

Another limitation is that our models are most applicable when range and habitat descriptions are available with very few observations, which is not common. While we evaluate our approach based on Wikipedia, a more realistic scenario involves obtaining descriptions of rarely observed species from domain experts. However, we note that many species have substantial Wikipedia articles despite having fewer than 50 observations on iNaturalist.

8

The hyacinth macaw prefers semi-open, somewhat wooded habitats. It usually avoids dense, humid forest, and in regions dominated by such habitats, it is generally restricted to the edge or relatively open sections (e.g., along major rivers). In different areas of their range, these parrots are found in savannah grasslands, in dry thorn forests known as caatinga, and in palm stands or swamps, particularly the moriche palm (Mauritia flexuosa).

The hyacinth macaw occurs today in three main areas in South America: In the Pantanal region of Brazil, and adjacent eastern Bolivia and northeastern Paraguay, in the cerrado regions of the eastern interior of Brazil (Maranhão, Piauí, Bahia, Tocantins, Goiás, Mato Grosso, Mato Grosso do Sul, and Minas Gerais), and in the relatively open areas associated with the Tocantins River, Xingu River, Tapajós River, and the Marajó island in the eastern Amazon Basin of Brazil.

They are diurnal, terrestrial, and live in complex, mixed-gender social groups of 8 to 200 individuals per troop. They prefer savannas and light forests with a climate that is suitable for their omnivorous diet.

Yellow baboons inhabit savannas and light forests in eastern Africa, from Kenya and Tanzania to Zimbabwe and Botswana.

Habitat text description   *Range text description*

Figure 4: **Zero-Shot Range Estimation.** Here we show the 'Habitat' and 'Range' text descriptions and corresponding zero-shot range maps for the `Hyacinth Macaw` (top) and the `Yellow Baboon` (bottom), with expert derived range maps inset.

Our models rely on text embeddings and summaries generated from LLMs, and thus may inherit the biases contained within them. For example, our model could further amplify biases in the data by incorrectly spatially localizing specific text terms inappropriately. Both Wikipedia text and observational data from iNaturalist are biased toward the United States and Western Europe. As a result, our models may not generalize as well to other geographical regions. There could also be potential negative consequences associated with using the species range predictions from our model to inform conservation or policy decisions. While our results are promising, they may still fall short of the quality needed for such high-stakes use cases. Therefore, caution is encouraged when making decisions based on the model's predictions. Another risk is that the model could be used to locate threatened or endangered species. To mitigate this, we only train on publicly available observation data that has been deemed safe to redistribute by the iNaturalist platform.

5   Conclusion

The generation of detailed species range maps is often constrained by the need for extensive location observations, which can be expensive and time-consuming to collect. Our LE-SINR approach mitigates this issue by mapping species observations and text descriptions into the same space, enabling zero-shot range map generation from text alone. Unlike other methods that make use of text, we can generate a global range map for a species with a single forward pass of our language model, significantly increasing usability for potential downstream users.

Our extensive evaluation shows that zero-shot range maps produced by LE-SINR, derived solely from text descriptions, can outperform those produced by state-of-the-art models trained on tens of observations. Additionally, using LE-SINR as a prior also significantly enhances few-shot performance. By learning to relate text descriptions of species with locations where that species has been observed, we show that LE-SINR develops an understanding of a wide range of geographical and environmental features, as well as unrelated concepts not seen in the training data such as historical events and aspects of culture.

9

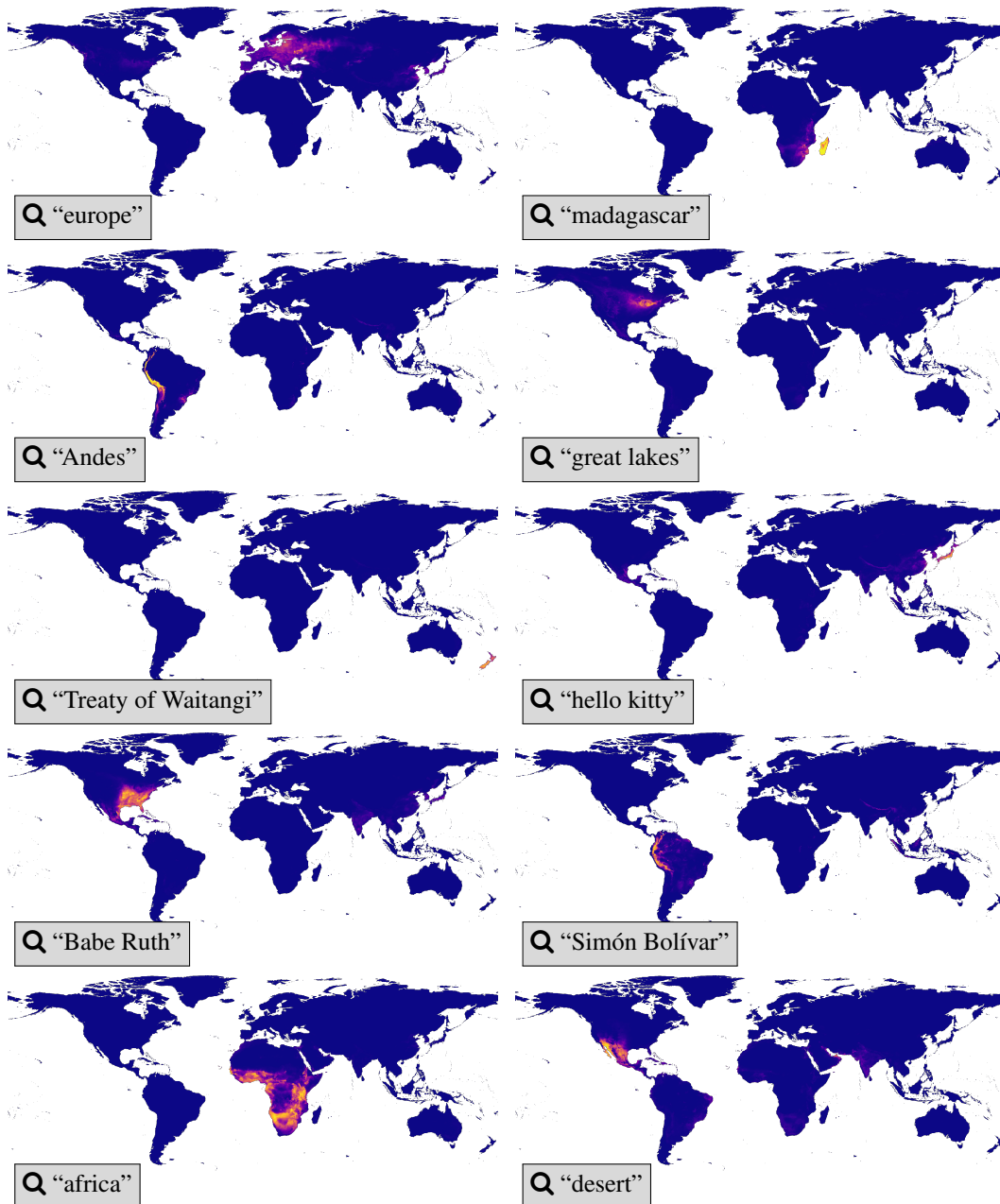



Figure 5: **Geospatial Grounding of Non-Species Concepts.** LE-SINR is able to geographically ground text prompts to locations on the earth. Here we display the inner product between the location encoder's features and the language model's encoding of the text displayed in the bottom left of each panel. This includes coarse concepts such as continents and countries (top row), geographic features such as specific lakes and mountain ranges (second row), in addition to concepts that do not appear in our species text training data but are likely already represented in the language model (third and fourth row). We do, however, observe some limitations resulting from the biases in our training data which favors North America, Europe, and Australasia (final row). Please zoom in to see more detail.

10

Acknowledgments and Disclosure of Funding

We thank the iNaturalist community for providing the data used for training our models. OMA was in part supported by a Royal Society Research Grant. SM and MH are supported by grant #2329927 from the National Science Foundation.

References

[1] The IUCN Red List of Threatened Species. Version 2022-2. `https://www.iucnredlist.org`. Accessed on 2024-05-22.

[2] iNaturalist. `https://www.inaturalist.org`. Accessed on 2024-05-22.

[3] Meta Llama 3. `https://llama.meta.com/llama3/`. Accessed on 2024-05-22.

[4] Wikipedia. `https://www.wikipedia.org/`. Accessed on 2024-05-22.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

[6] S. Beery, E. Cole, J. Parker, P. Perona, and K. Winner. Species distribution modeling for machine learning practitioners: a review. In *SIGCAS Conference on Computing and Sustainable Societies*, 2021.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *Transactions on Intelligent Systems and Technology*, 2024.

[8] D. Chen and C. P. Gomes. Bias reduction via end-to-end shift learning: Application to citizen science. In *AAAI*, 2019.

[9] D. Chen, Y. Xue, S. Chen, D. Fink, and C. Gomes. Deep multi-species embedding. In *IJCAI*, 2017.

[10] E. Cole, G. Van Horn, C. Lange, A. Shepard, P. Leary, P. Perona, S. Loarie, and O. Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *ICML*, 2023.

[11] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 2021.

[12] F. Dorm, C. Lange, S. Loarie, and O. Mac Aodha. Generating Binary Species Range Maps. In *Computer Vision for Ecology Workshop at ECCV*, 2024.

[13] J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 2009.

[14] J. Elith, C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 2006.

[15] D. Fink, T. Auer, A. Johnston, M. Strimas-Mackey, O. Robinson, S. Ligocki, W. Hochachka, L. Jaromczyk, C. Wood, I. Davies, M. Iliff, and L. Seitz. ebird status and trends, data version: 2020; released: 2021. *Cornell Lab of Ornithology, Ithaca, New York*, 2020.

[16] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[17] S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, 2022.

[18] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv:2311.17179*, 2023.

11

[19] C. Lange, E. Cole, G. Horn, and O. Mac Aodha. Active learning-based species range estimation. In *NeurIPS*, 2023.

[20] K. Lee, S. Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In *CVPR*, 2019.

[21] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. In *TMLR*, 2023.

[22] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. Geollm: Extracting geospatial knowledge from large language models. In *ICLR*, 2024.

[23] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm. How many species are there on earth and in the ocean? *PLoS Biology*, 2011.

[24] N. Muennighoff, H. Su, L. Wang, N. Yang, F. Wei, T. Yu, A. Singh, and D. Kiela. Generative representational instruction tuning. *arXiv:2402.09906*, 2024.

[25] J. Roberts, T. Lüddecke, S. Das, K. Han, and S. Albanie. Gpt4geo: How a language model sees the world's geography. *arXiv:2306.00020*, 2023.

[26] J. Roberts, T. Lüddecke, R. Sheikh, K. Han, and S. Albanie. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *CVPR Workshops*, 2024.

[27] M. Rußwurm, K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *ICLR*, 2024.

[28] S. Sastry, X. Xing, A. Dhakal, S. Khanal, A. Ahmad, and N. Jacobs. Ld-sdm: Language-driven hierarchical species distribution modeling. *arXiv:2312.08334*, 2023.

[29] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *NeurIPS*, 2017.

[30] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 2009.

[31] M. Teng, A. Elmustafa, B. Akera, Y. Bengio, H. Radi, H. Larochelle, and D. Rolnick. Satbird: a dataset for bird species distribution modeling using remote sensing and citizen science data. *NeurIPS*, 2023.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[33] V. Vivanco Cepeda, G. K. Nayak, and M. Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *NeurIPS*, 2024.

[34] R. Zbinden, N. van Tiel, M. Rußwurm, and D. Tuia. Imbalance-aware presence-only loss function for species distribution modeling. In *ICLR Workshop on Tackling Climate Change with Machine Learning*, 2024.

[35] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023.

12

Appendix

A   Additional Results

In Table A1 we evaluate our oracle model, that was trained on evaluation species observations, on the habitat and range text summaries. These results show that some of the performance gap between the zero-shot and supervised methods can be explained by inherent ambiguities of our evaluation text descriptions to describe the range.

In Figure A1 we display the intermediate representations learned by our location encoder and compare it to the standard SINR model that does not make use of any language information. There appears to be more higher frequency spatial information encoded in our representations which is consistent with the improved few-shot performance we observe when using these features in conjunction with a logistic regression classifier in Figure 3 from the main paper.

In Figure A2 we show the range estimates obtained from text (zero-shot setting) as well as a few examples (few-shot setting corresponding to Table 1 from the main paper) for three different species, namely the `Northern Yellow-shouldered Bat`, `Lark Bunting`, and `European Serin` along with the associated range maps curated by experts.

In Figure A3 we show the range estimates obtained from both habitat and range text embeddings (zero-shot setting) alongside the associated range maps curated by experts for six additional species, namely the `Striated Babbler`, `Striped Sticky Frog`, `Common Hawk-Cuckoo`, `Cape Griffon`, `Madagascar Hoopoe` and `Raucous Toad`.

In Figure A4 we show how zero-shot range estimates for the `Collared Bush Robin` change as we use different parts of a piece of text.

Table A1: **Evaluation Text Oracles**. We can get a tighter upper bound on performance by evaluating our oracle model with the habitat and range text summaries rather than species tokens. These summaries are not seen in the training data.

Method	+Eval Sp.	IUCN	S&T
LE-SINR (Habitat Text)		0.32	0.52
LE-SINR (Range Text)		0.53	0.64
LE-SINR (Habitat Text)	✓	0.39	0.61
LE-SINR (Range Text)	✓	0.60	0.69

B   Dataset Examples

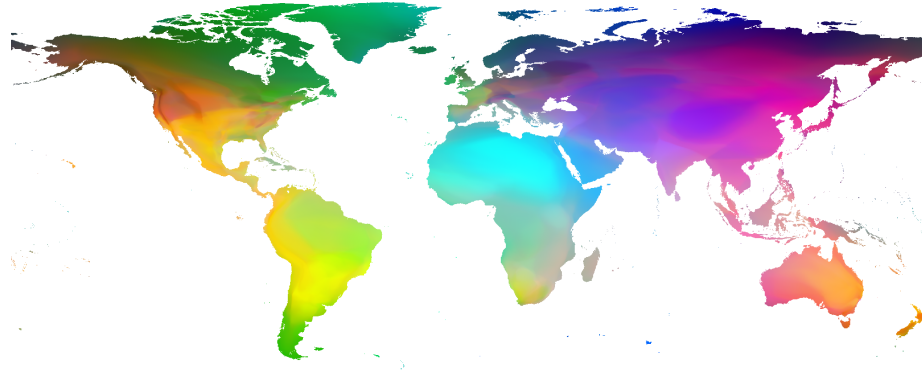
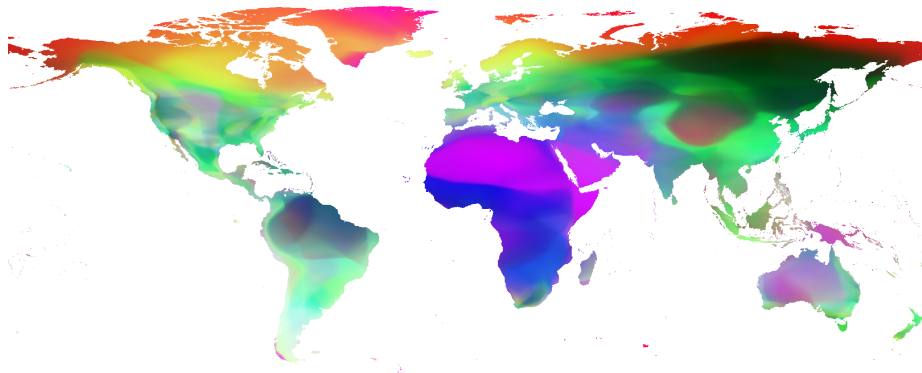
Below we show 'range', 'habitat' text for several species used for our zero-shot range estimation experiments. These were generated by using Llama3 language model to summarize the Wikipedia text into content that describes the range, i.e., the geographical extent where the species is found, including names of countries, continents, and geographic regions, as well as its habitat, which indicates specific environmental conditions and types of habitats that a species thrives in such as descriptions of climate, vegetation, topography, soil types, food resources, etc.

1. `Gray Kingbird (Tyrannus dominicensis)`

    - *Range*: The gray kingbird is found in the southeast USA, Colombia, and Venezuela, with two recognized subspecies: T. d. dominicensis and T. d. vorax. It breeds from the extreme southeast of the United States, mainly in Florida, as well as Central America, and through the West Indies south to Venezuela, Trinidad and Tobago, the Guianas, and Colombia. Northern populations are migratory, wintering on the Caribbean coast of Central America and northern South America.

    - *Habitat*: The gray kingbird favors tall trees and shrubs, including the edges of savanna and marshes. It is found in increasing numbers in the state of Florida, and is more often found inland though it had been previously restricted to the coast.

2. `Cape Weaver (Ploceus capensis)`

13

SINR

LE-SINR

Figure A1: Visualization of the intermediate position representation learned by the network projected to three dimensions using Independent Component Analysis. (Top) Standard SINR model. (Bottom) Our LE-SINR model. Both models were trained with position features as the only input.

- *Range*: The Cape weaver is endemic to South Africa, Lesotho, and Eswatini, occurring across much of the area excluding the Kalahari Desert from the Orange River in the Northern Cape south to the Cape of Good Hope, then east to northern KwaZulu Natal, and inland almost to Bloemfontein in the Free State.
- *Habitat*: The Cape weaver occurs in open grassland, lowland fynbos, coastal thicket, and farmland, so long as there is permanent water and trees. In the more arid, hotter regions, it is restricted to upland areas and never occurs in forest.

3. `Plate-billed Mountain-Toucan (Andigena laminirostris)`

- *Range*: The plate-billed mountain toucan is found in the western foothills of the Andes of western Ecuador and far southwestern Colombia, specifically from Pita Canyon (Narino) in southwestern Colombia and south to the northwestern border of Morona-Santiago Province, in Ecuador.
- *Habitat*: The species inhabits the humid forest and edges of the temperate forest of the lateral slope of the Andes Mountains, featuring abundant epiphytes, bromeliads, and mosses. The forests receive an average of 14 feet of rainfall per year, and the canopy ranges from 6 to 10 meters high. Their altitudinal range is between 1600 and 2600 meters above sea level.

4. `Harlequin Racerunner (Plica umbra)`

- *Range*: The blue-lipped tree lizard or harlequin racerunner (Plica umbra) is found in South America, specifically in Colombia, Venezuela, Guyana, Suriname, French Guiana, Brazil, Bolivia, Peru, and Ecuador.
- *Habitat*: The species inhabits tropical rainforests, savannas, and dry forests, typically at elevations below 500 meters. It prefers areas with dense vegetation, rocky outcrops, and sandy or clay soils.

14

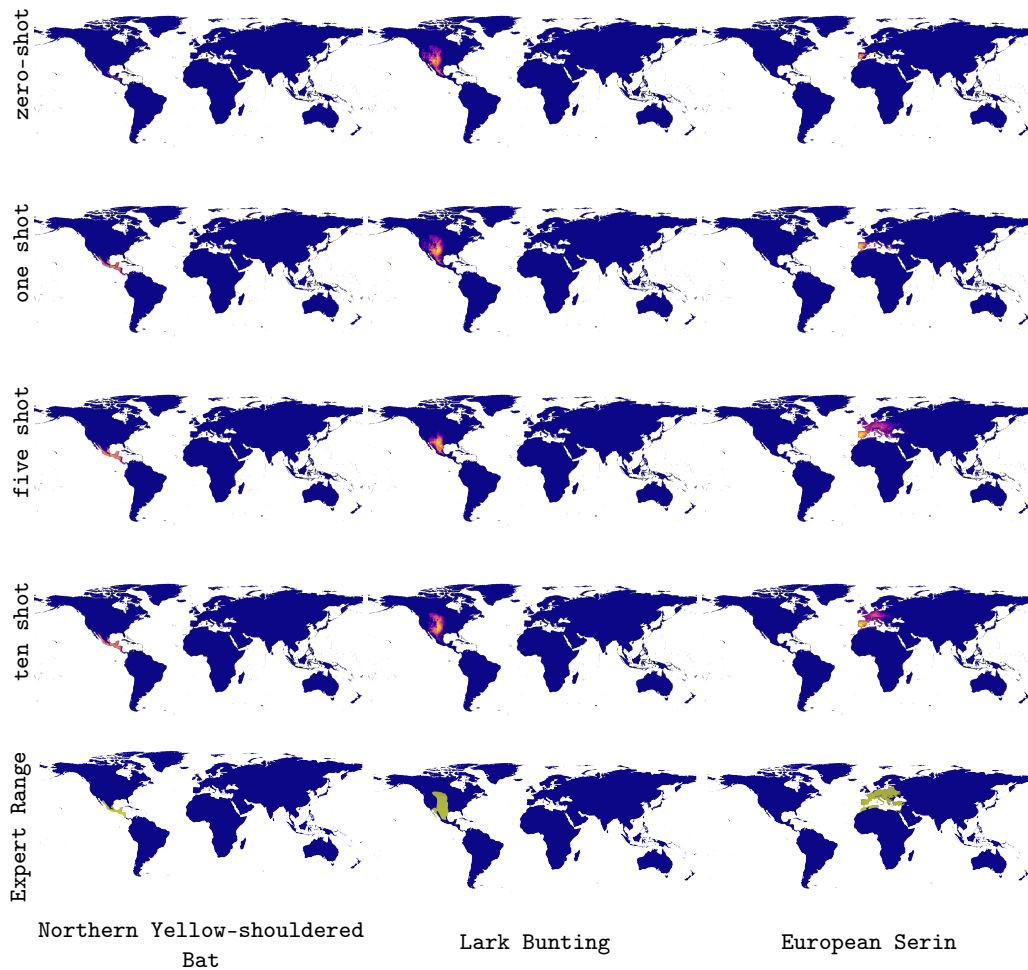



Figure A2: **Visualization of Zero-Shot and Few-Shot Range Estimates.** Range estimates obtained from text only (zero-shot) for three different species is shown on the top row. Range estimates with one, five, and ten observations are shown in rows two, three, and four respectively, while the ground truth range shown at the bottom row along with the name of the species. Zero-shot predictions were made from the 'range' text. Zoom in for details.

5. Dusky Rattlesnake (Crotalus triseriatus)

- *Range*: The Mexican dusky rattlesnake (Crotalus triseriatus) is found in Mexico, along the southern edge of the Mexican Plateau in the highlands of the Transverse Volcanic Cordillera, including the states of Jalisco, México, Michoacán, Morelos, Nayarit, Puebla, Tlaxcala, and Veracruz.
- *Habitat*: Crotalus triseriatus occurs in pine-oak forest, boreal forest, coniferous forest, and bunchgrass grasslands. On Volcán Orizaba, it is found at very high altitudes, with the species being found within the zone where the snow line comes down to about 4,572 m (15,000 ft), and green plants can be found up to 4,573 m (15,003 ft).

6. Western Ghats Flying Lizard (Draco dussumieri)

- *Range*: The species is found principally in the Western Ghats and some other hill forests of Southern India, including Karnataka, Kerala, Tamil Nadu, Goa, and Maharashtra. It is also reported from some parts of the Eastern Ghats in Andhra Pradesh.
- *Habitat*: The southern flying lizard is almost entirely arboreal, found on trees in forests and adjoining palm plantations. It climbs trees in search of insect prey on the trunks and leaps off when it reaches the top to land on adjoining trees. The species is active during the day after it has warmed up in the early morning sun.

15

Habitat text   Range text   Expert Derived Range

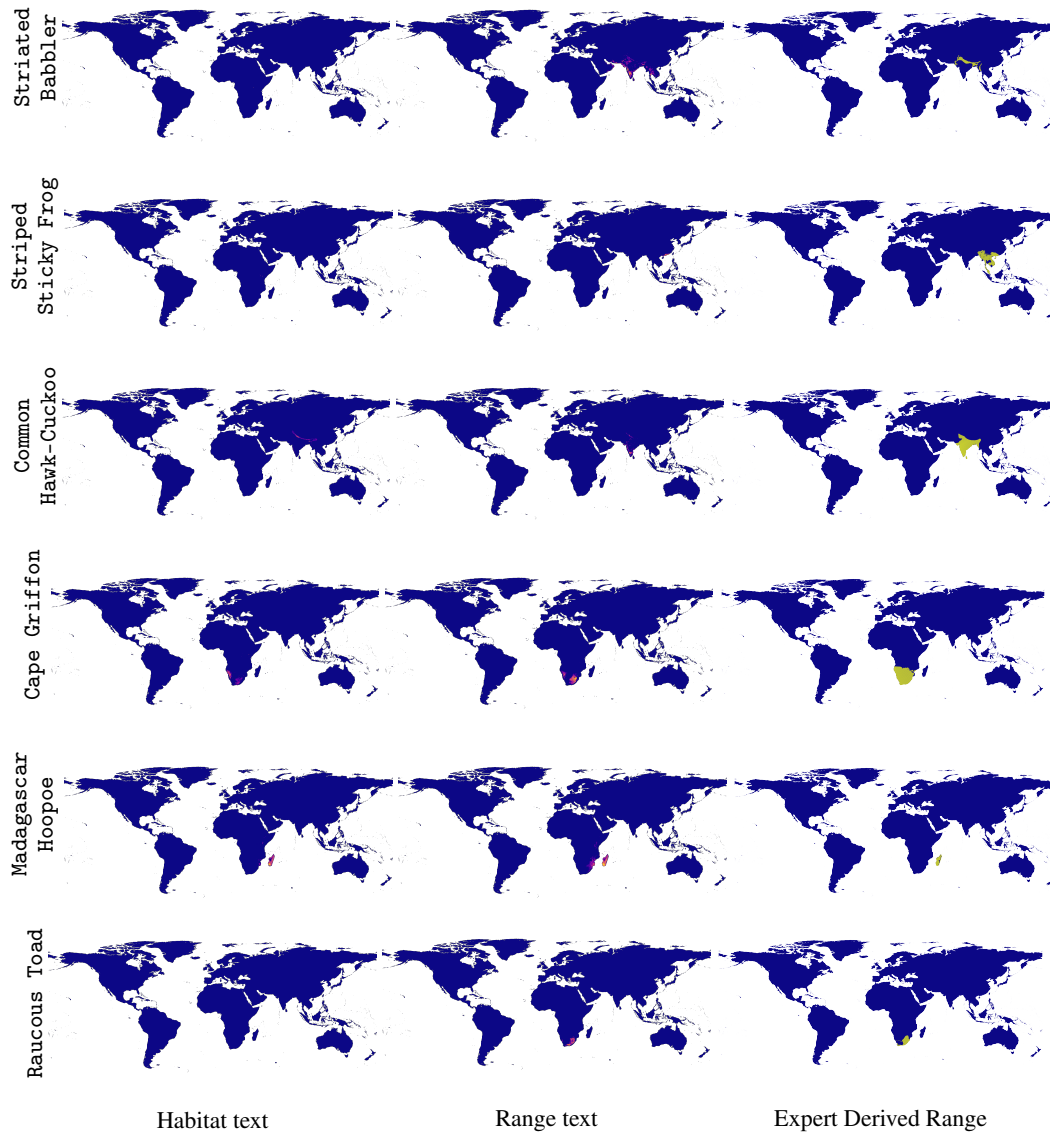



Figure A3: **Additional Zero-Shot Range Estimation.** Here we show the 'Habitat' (left) and 'Range' (center) text zero-shot range maps with associated expert derived range maps (right), for a variety of species. While the 'Range' text provides a strong prior in all cases with high probability assigned to regions within the expert derived range, the 'Habitat' text does not always do this, with the `Striated Babbler`, `Striped Sticky Frog`, and the `Raucous Toad` providing no strong prior. Zoom in to see details.

16

Habitat Text   **Range Text**

The species inhabits undergrowth of coniferous forests with shrubs and bamboo, and can also be seen in parks and along roads.

The collared bush robin is endemic to Taiwan,

It typically resides at elevations of 2,000 to 2,800m (6,600 to 9,200 ft) and sometimes above the tree line, descending to lower elevations in winter.

living in montane and subalpine forests.

The species inhabits undergrowth of coniferous forests with shrubs and bamboo, and can also be seen in parks and along roads. It typically resides at elevations of 2,000 to 2,800m (6,600 to 9,200 ft) and sometimes above the tree line, descending to lower elevations in winter.

The collared bush robin is endemic to Taiwan, living in montane and subalpine forests.

Figure A4: **Zero-Shot Range Estimation From Parts of Text.** Here we show the zero-shot predicted ranges when different parts of the habitat text (left column) and range text (right column) are given to our model for the Collared Bush Robin, which is found in Taiwan. From top to bottom, the rows show the predicted range for the first part of the text, the second part of the text, and the entire text, i.e., the first and second parts concatenated. For the habitat text we see that the first part of the text loosely identifies several forested areas but not Taiwan, while the second part loosely identifies several mountainous areas including Taiwan. Combining these parts reduces the false positives in the range map produced while still correctly including Taiwan. For the range text, we again see that the second part of the text identifies some mountainous areas, while the first part seems to locate Taiwan effectively. Together the range is correctly limited to the inland highland areas of Taiwan. Please zoom in to see more detail.

17

NeurIPS Paper Checklist

1. **Claims**

    Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

    Answer: [Yes]

    Justification: Section 4 provide detailed experimental evidence quantifying the accuracy of predicted range maps using expert-derived range maps for species in the S&T and IUCN lists. We show the value of language supervision by demonstrating zero-shot range estimation from descriptions of habitat and range preferences. We also demonstrate the value of learning from language with superior performance over the baseline SINR model trained with observations only in the few-shot setting

    Guidelines:

    - The answer NA means that the abstract and introduction do not include the claims made in the paper.
    - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
    - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
    - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

    Question: Does the paper discuss the limitations of the work performed by the authors?

    Answer: [Yes]

    Justification: Section 4.4 describes some of the limitations of our work, which include the need for a more thorough evaluation of the few-shot experiments, inheriting biases from language models and text data on the internet, and the lack of precision required for high-stakes use cases in conservation and planning.

    Guidelines:

    - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
    - The authors are encouraged to create a separate "Limitations" section in their paper.
    - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
    - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
    - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
    - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
    - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
    - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

18

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The work closely follows the experimental setup, source code, and training data of SINR using iNaturalist, both of which are publicly available. The novel part is the incorporation of Wikipedia text and their short summaries, which are provided in the Supplementary Material. We have described the model architecture and hyperparameters for training in our paper, and will publicly release the dataset and the evaluation framework for the zero-shot and few-shot experiments upon publication of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

19

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include a link to a publicly available github in the intro that contains instructions and code needed to reproduce the experimental results. We also provide pre-trained weights of the models used in this work as well as the training and evaluation data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We follow the same evaluation splits and metrics as previous work, which is publicly available. The training data is also identical, except for the zero-shot and few-shot settings where the evaluation species are held out.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

20

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not report error bars, but the MAP numbers are the result of averaging AP across hundreds or thousands of species for S&T and IUCN, respectively, and the confidence intervals for MAP are extremely narrow ($\approx$0.1 MAP). The improvements from language supervision in the zero-shot and few-shot settings in Figure 3 are 10 to 15 points MAP, so the improvements are statistically significant.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 3.6 describes the computational resources needed for training.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have conformed to ethics code to the best of our ability and knowledge.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

21

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 4.4 discusses the broader impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: While we do scrape text from the internet, the datasets do not contain images, and are less extensive focusing on a few thousand Wikipedia articles covering a wide range of species in natural world. Thus we believe it poses low safety risks such as those associated with unsafe images or text.

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

22

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Our benchmark builds on data from iNaturalist and Wikipedia. The iNaturalist data is already publicly available from the original SINR paper, as we noted in our paper. Meanwhile, Wikipedia grants permission to copy, distribute and/or modify Wikipedia's text under the terms of the Creative Commons Attribution-Share Alike 4.0 International License. We will include these details in the public release of our benchmark.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Documentation for the datasets and models is provided in the github repository.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

23

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

24