Quantification of Biodiversity from Historical Survey Text with LLM-based Best-Worst Scaling

Thomas Haider, Tobias Perschl, Malte Rehbein
Chair of Computational Humanities
University of Passau
firstname.lastname@uni-passau.de

Abstract

In this study, we evaluate methods to determine the frequency of species via quantity estimation from historical survey text. To that end, we formulate classification tasks and finally show that this problem can be adequately framed as a regression task using Best-Worst Scaling (BWS) with Large Language Models (LLMs). We test Ministral-8B, DeepSeek-V3, and GPT-4, finding that the latter two have reasonable agreement with humans and each other. We conclude that this approach is more cost-effective and similarly robust compared to a fine-grained multi-class approach, allowing automated quantity estimation across species.

1 Introduction

Long-term observation data plays a vital role in shaping policies for preventing biodiversity loss caused by habitat destruction, climate change, pollution, or resource overexploitation (Dornelas et al., 2013; Hoque and Sultana, 2024). However, these efforts depend on the availability of reliable and relevant historical data and robust analytical methods, a significant challenge due to the heterogeneity of records representing such data.

The available biodiversity data varies widely in resolution, ranging from detailed records (e.g., point occurrences, trait measurements) to aggregated compilations (e.g., Floras, taxonomic monographs) (König et al., 2019). Many projects, such as the *Global Biodiversity Information Facility* (GBIF), focus largely on the disaggregated end of the spectrum, particularly with presence/absence data (Dorazio et al., 2011; Iknayan et al., 2014). Furthermore, despite their utility, longitudinal data is largely confined to records from after 1970 (van Goethem and van Zanden, 2021), leaving significant historical gaps.

Natural history collections and records from the archives of societies present valuable opportunities to extend data further back in time (Johnson et al., 2011; Brönnimann et al., 2018). Such sources are rich, but typically unstructured and require sophisticated extraction tools to produce meaningful quantitative information. Recent advances in NLP have shown promising potential for retrieval-based biodiversity detection from (mostly scientific) literature (Kommineni et al., 2024; Langer et al., 2024; Lücking et al., 2022).

This paper focuses on evaluating methods for biodiversity quantification from semi-structured historical survey texts. To achieve this, we test tasks to distill meaningful metrics from textual information found in survey records. A particular focus lies on the feasibility of Best-Worst Scaling (BWS) with a Large Language Model (LLM) as an annotator, which promises greater efficiency and cost-effectiveness compared to manual annotation (Bagdon et al., 2024). In the following, we describe the data, outline the tasks and machine learning methods, and finally present a case study.

2 Data

In 1845, the Bavarian Ministry of Finance issued a survey to evaluate biodiversity in the Bavarian Kingdom, a region that encompasses a variety of different ecosystems and landscapes. To that end, 119 forestry offices were contacted to complete a standardized questionnaire. Namely, trained local foresters recorded in free text how frequently 44 selected vertebrate species occurred in the respective administrative territory, and in which habitats and locations they could be found.

Figure 1 shows the facsimile of a digitized survey page. It features a header containing instructions and a number of records describing animal species with their respective responses. These historical survey documents are preserved by the Bavarian State Archives (cf. Rehbein et al., 2024).

Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), pages 61–67
March 2, 2025 ©2025 University of Tartu Library

Animal	Text	Binary	BWS	Multi-Classification	
Ducks	Bedecken Isar-Strom, wie Amper und Moosach in ganzen Schwärmen. *Cover Isar-stream, likewise Amper and Moosach in whole swarms.*	1	1.00	5	ABUNDANT
Roe Deer	Ist hier zu Hause, und beinahe in allen Waldtheilen zu finden. *Is at home here and can be found in almost all parts of the forest.*	1	0.88	4	COMMON
European Adder	Kommt wohl aber eben nicht häufig vor. *Does indeed appear but just not that often.*	1	0.44	3	COMMON TO RARE
Lynx	Höchst selten wechseln derlei Thiere von Tyrol herüber. *Very rarely do such animals cross over from Tyrol.*	1	0.12	2	RARE
Wild Goose	Kommt nur äußerst selten zur Winterszeit vor. *Occurs only very rarely at winter time.*	1	0.06	1	VERY RARE
Owl	Horstet dahier nicht und verstreicht sich auch nicht in diese Gegend. *Does not nest here and does not stray into this area.*	0	0.00	0	ABSENT
Wolf	Kommt nicht mehr vor. *No longer occurs.*	0	0.00	-1	EXTINCT

Table 1: Data Examples with Annotation (our own translations)

The archival sources were digitized, transcribed from the handwritten original and enriched with metadata, including, among others, taxonomic norm data according to the GBIF-database[1] (Telenius, 2011) and geographical references to forestry offices. This data set is freely available on Zenodo (Rehbein et al., 2024).

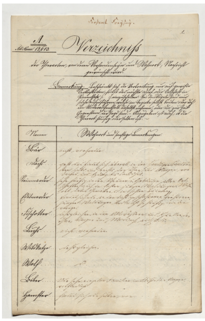

Figure 1: Facsimile of a survey page, Freysing forestry office in the Upper Bavaria district.

In total, the data set contains 5,467 entries[2] among which are also a number of empty (striked out) or 'see above'-type responses. The unique set we used for our experiments contains 2,555 texts. We find that the foresters' replies vary considerably in length where most texts contain 3 to 10 tokens and only a few texts more than 20 tokens. Table 1 provides examples with annotation according to the tasks detailed in the next section.

3 Tasks & Experiments

The main task in this paper is to assign a quantity label to a text, indicating the frequency with which an animal species occurs in a specific area. This can be operationalized in various ways, either through a classification task or through regression. In both, it can be as difficult to obtain consistent labels by asking humans to assign a value from a rating scale (Schuman and Presser, 1996; Likert, 1932). Likewise, it is also difficult for researchers to design rating scales, considering design decisions such as scale point descriptions or granularity may bias the annotators.

We evaluate three different task setups,[3] as detailed in Table 1: Binary 'Presence vs. Absence' Classification, a 7-ary Multi-Class setup (Abundant to Extinct), and continuous values scaled to $[0, 1]$. For the first two tasks, we use manual annotation, while continuous values are derived through BWS with LLMs (Bagdon et al., 2024).

3.1 Binary Classification

The simplest form of animal occurrence quantification is a binary distinction between the absence (0) or presence (1) of a given species, an annotation scheme as popular as it is problematic in biodiversity estimation.[4] In our annotation, the label PRESENT is given when a species is described in the historical dataset as having been observed in that particular locality at the time of the survey (thus excluding mentions of past occurrences, i.e., extinctions). The annotation workflow consists of iterative steps with discussions. Agreement is nearly perfect. Overall, from the set of 2,555 unique texts, 1,992 (78%) fall into class PRESENT, 563 (22%) into ABSENT.[5]

[1] gbif.org

[2] Including species that were not explicity prompted.

[3] Code: github.org/maelkolb/biodivquant

[4] Since ABSENCE may just stem from non-detection, rather than real absence (Dorazio et al., 2011; Iknayan et al., 2014; Kestemont et al., 2022).

[5] In the complete dataset, absence texts make up more than half of all text descriptions, but often amount to empty or 'strike-out' responses. Thus, the task would be easier on the full dataset, because many instances are trivial to predict.

To test the feasibility of the binary task, we create training curves with different models, namely BERT against Logistic Regression, SVM, and Random Forest on Unigrams. We use 20% of the data for testing, and take another 20% from the training set for hyperparameter search at each cumulative 100 text increment. Despite the 78% majority baseline, we find that the models perform well and training requires only a few hundred texts to reach an F1-macro score in the high 90s.

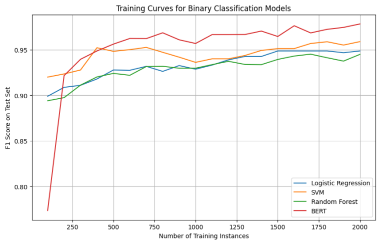

Figure 2: Training Curves of different models on incremental training data (binary classification)

Upon feature weight interpretation of the Logistic Regression and LIME on BERT (Ribeiro et al., 2016), we find that there is some bias in the data: Classification decisions occur on tokens that are not explicit quantifiers and easily substitutable without changing the classification result (e.g., common toponyms such as ‘Danube’). This presents a case of spurious correlations—an interesting future research direction, but a matching (Wang and Culotta, 2020) or counterfactual approach (Qian et al., 2021) appears challenging for this heterogeneous data. Yet, we annotate the best features with regard to their ‘spuriousness’ and find that classifiers are still robust without spurious features. This annotation also gives us a list of quantifiers which we utilize for transfer learning of a regression model (section 3.3).

3.2 Multi-Classification

Since the quantification of species frequency in practice exceeds the binary differentiation between presence and absence of animals, a multi-class approach provides more details. We use a 7-class system, categorizing texts based on the schema as shown by the descriptors in Table 1, ranging from ABUNDANT (5) to EXTINCT (-1). We decided to annotate data of four species for our case study (section 4): Roe deer, Eurasian Otter, Eurasian Beaver, Western Capercaille, each within the 119 forestry offices (with one annotator).

A second person annotates a random sample of 100 texts, resulting in a Cohen’s κ of 0.78, indicating high agreement.

We then train a few models with a 5-fold cross validation, and find that the language agnostic sentence encoder model LaBSE (Feng et al., 2022) performs better than monolingual BERT-models and a Logistic Regression. We also test a zero shot classification with GPT-4 and Deepseek-V3. See appendix for the prompt.

Model	F1 Micro	F1 Macro
Logistic Regression	0.69	0.61
gbert-base	0.63	0.51
bert-base-german	0.73	0.63
LaBSE	**0.77**	**0.68**
GPT4 Zero Shot	0.70	0.56
DSV3 Zero Shot	0.66	0.66

Table 2: Multi-class model performance.

As seen in Table 2, this task is generally quite challenging. We find that the main problem is posed by the underrepresented classes, as shown by the discrepancy between micro and macro scores, indicating that more data would help, which is, however, expensive to obtain. Zero shot classification with GPT-4 in turn is biased towards the RARE classes, such that COMMON categories are harder to predict, while DeepSeek-V3 (DSV3) shows a more balanced response.

3.3 Continuous Quantification

Finally, we experiment with operationalizing our task as a regression problem with the aim of generalizing the quantification problem to less arbitrary categories and a possibly imbalanced data set (Berggren et al., 2019). While a naïve labeling of quantifiers showed promising results, it is a challenge to create a comprehensive test set based on heuristic annotation. Thus, we experiment with Best-Worst Scaling, aided by LLMs.

3.3.1 Best-Worst Scaling with LLMs

Best-Worst Scaling (BWS) is a comparative judgment technique that helps in ranking items by identifying the best and worst elements within a set. This approach is easier to accomplish than manual labeling and there are fewer design decisions to make. In a BWS setting, the amount of annotations needed to rank a given number of text instances depends on three variables, namely 1) corpus size (total number of texts used), 2) set size (number of texts in each comparison set), and 3) number of comparison sets each text appears in.

The number of comparisons divided by set size is regarded as the variable N, where $N = 2$ generally yields good results in the literature (Kiritchenko and Mohammad, 2017). A reliable set size is 4, since choosing the best and worst text instance from a 4-tuple set essentially provides the same number of comparisons as five out of six possible pairwise comparisons (ibid).

We take a random sample of 1,000 texts (excluding texts with ABSENCE annotation, thus making the task harder, but giving us a more realistic distribution). With a set size of 4 and $N = 2$, every text occurs in exactly 8 different sets and we get 2,000 comparison sets (tuples). These are then individually prompted to three LLMs: the relatively small Ministral-8B,[6] OpenAI's GPT-4 (Achiam et al., 2023), and the DeepSeek-V3 open source model (Liu et al., 2024).

	Annotator1	Annotator2	B	W	B + W
LLM-LLM	GPT4	DeepseekV3	0.73	0.69	0.56
	Ministral8B	DeepseekV3	0.54	0.54	0.36
	GPT4	Ministral8B	0.57	0.50	0.38
Average			0.61	0.57	0.43
Human-Human	AR	DS	0.56	0.65	0.45
	DS	KB	0.56	0.62	0.40
	MR	AR	0.51	0.65	0.39
	TP	AO	0.73	0.55	0.48
	MP	MR	0.59	0.52	0.41
Average			0.59	0.60	0.43
Human-LLM	AO	Ministral8B	0.43	0.31	0.23
	AR	Ministral8B	0.47	0.58	0.38
	DS	Ministral8B	0.43	0.42	0.23
	KB	Ministral8B	0.53	0.61	0.46
	MP	Ministral8B	0.45	0.43	0.30
	MR	Ministral8B	0.55	0.48	0.38
	TP	Ministral8B	0.49	0.31	0.24
Average			0.48	0.45	0.32
Human-LLM	AO	GPT4	0.68	0.55	0.45
	AR	GPT4	0.49	0.57	0.34
	DS	GPT4	0.44	0.71	0.43
	KB	GPT4	0.47	0.68	0.41
	MP	GPT4	0.57	0.62	0.41
	MR	GPT4	0.49	0.63	0.41
	TP	GPT4	0.63	0.57	0.43
Average			0.54	0.62	0.41
Human-LLM	AO	DeepseekV3	0.61	0.59	0.45
	AR	DeepseekV3	0.55	0.68	0.41
	DS	DeepseekV3	0.62	0.63	0.46
	KB	DeepseekV3	0.57	0.62	0.41
	MP	DeepseekV3	0.69	0.53	0.41
	MR	DeepseekV3	0.59	0.68	0.46
	TP	DeepseekV3	0.58	0.58	0.41
Average			0.60	0.62	0.43

Table 3: Cohen's κ Agreement between humans and LLMs in Best-Worst-Annotation (B: Best, W: Worst, B+W: Best + Worst). Two-letter shorthands for humans.

Whereas Ministral-8B is run locally, we use the OpenAI API to access GPT-4 and the fireworks.ai API endpoint for DeepSeek-V3, since the DeepSeek-webservices are limited at the time of the experiment and hardware limitations hamper local deployment. Prompts are in the appendix.

We ask seven native German post-graduates to annotate one of two subsets of 50 tuples each with a custom browser-based annotation interface. Table 3 shows Cohen's κ agreement across humans and LLMs. We find that agreement among humans is largely on par with agreement between humans and the two larger LLMs, while the lower agreement between Ministral-8B and humans, as well as the other machine annotators, indicates a limited capability of this model for the task at hand. It appears that it is easier to identify the worst instance than the best, which is likely an artifact of our data. Interestingly, agreement between GPT-4 and DeepSeek-V3 is the highest overall, which could lend itself either to a) the task being easier for the LLMs than for humans, or b) that the models are overall fairly similar. We find no significant difference ($p = .118$) between GPT-4 and DeepSeek-V3 in Human-LLM comparison.

$$s(i) = \frac{\#best(i) - \#worst(i)}{\#overall(i)} \tag{1}$$

By counting how often each text was chosen as the best, worst, or as one of two other texts, we calculate a score $s(i)$ as detailed in equation (1), resulting in an interval scale $[-1, 1]$, which we normalize to a scale $[0, 1]$. This scales (and ranks) the entire dataset, so it can be used for regression. It should be noted that the scores come in increments of $\frac{1}{8}$ (determined by number of comparisons of instance i), resulting in 17 discrete values. We find a flat unimodal inverted U-shape in the score distribution without notable outliers.

3.3.2 Regression Models

We train a variety of different regression models with 5-fold cross validation to optimize for the values generated by Best-Worst Scaling, as shown in Table 4. We compare a Kernel Ridge Regression (KRR) baseline against BERT-style-models with regression head, and test a transfer learning setup, for which we scale the 114 n-gram quantifiers as extracted from the binary Logistic Regression with another GPT-4 BWS, then match these scores to the texts and tune a LaBSE model on the same train/test split before using it for the final task.

Curiously, KRR with LaBSE embedding features benefits substantially from hyperparameter

[6] `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`

Features/Training Strategy	Model	MAE		R²	
		GPT4	DSV3	GPT4	DSV3
Unigrams	KRR	0.159	0.158	0.514	0.515
Frozen LaBSE Embeddings	KRR	0.118	0.117	0.678	0.686
Regression Head	bert-base-german	0.149	0.158	0.516	0.490
Regression Head	LaBSE	0.133	0.127	0.607	0.657
Reg. Head + Transfer	LaBSE	**0.107**	0.117	**0.730**	0.710

Table 4: Comparison of different training strategies for regression based on BWS-Scaling. GPT4: GPT-4 BWS annotation, DSV3: Deepseek-V3 BWS annotation

tuning, reaching superior results over LaBSE with regression head. The Transfer Model on GPT4 BWS offers the best performance, with acceptably high explained variance ($R^2 = .73$) and only .11 Mean Absolute Error (MAE), which makes this model useful for downstream prediction as in the case study below. However, more data would likely also help, since training curves show continuous improvement.

4 Case Study

For a proof of concept, we map the predictions of the regression model (LaBSE transfer regression model based on GPT-4 BWS) to the multi-class human annotation. Figure 3 shows a strong relationship between multi-class labels and regression scores for the entire dataset (four species), but also that the extinction class is not properly represented in the regression, and furthermore that higher values are challenging to predict.

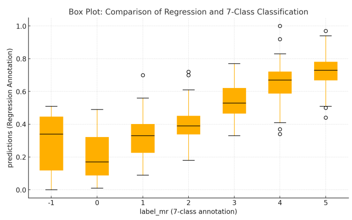

Figure 3: Multi-Class vs. Regression Distribution

Figure 4 shows specie-specific distributions for Roe deer and Eurasian otter across all 119 offices, indicating a fairly good alignment between the regression result (top) and the multi-class annotation (bottom). However, the mapping is not unambiguous due to 1) shortcomings of the regression, such as the inability to model extinction and difficulty in predicting high values, and 2) imperfect alignment with class intervals, which are fuzzy with regard to the continuous values. However, pending further research, we find that our method performs well and produces plausible results.

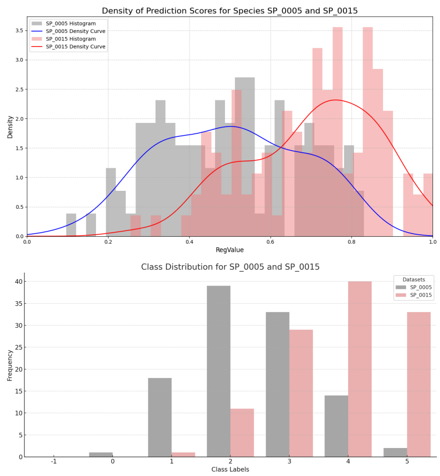

Figure 4: Density histogram of regressor prediction (top) and multi-class (bottom) distribution for Roe deer (SP_0015, red) and Eurasian otter (SP_0005, grey).

5 Conclusion & Future Work

This study demonstrates that information of occurrence frequencies from semi-structured historical biodiversity survey texts can be adequately modeled with Best-Worst Scaling through LLMs. While a simple classification approach performs well with minimal training data, a more complex classification struggles with design decisions and imbalanced data. BWS meets this by eliminating rating scale design decisions. In addition, it is cognitively and computationally less expensive, since no manual annotation of training data is necessary, while still offering similarly accurate results with much finer granularity through regression.

The robustness of methods and models should be further tested, not exclusive to biodiversity surveys, lending itself to a number of tasks. Yet, similar data to ours likely exists, e.g., on 19th century Bavarian flora, Württembergische Oberamtsbeschreibungen (1824–1886), or data in biodiversitylibrary.org, making our methods valuable.

Limitations

The accuracy of the method depends heavily on the capabilities of the specific LLM used. If a model lacks domain-specific knowledge or has biases, it may impact results. Furthermore, without a reliable dataset to benchmark against, it is difficult to assess the absolute accuracy of the BWS-based regression approach, because we also test on BWS values. While we measured agreement on the BWS task with humans, it is impractical to scale the entire dataset with both LLMs and humans, and thus our agreement calculation may suffer from sampling bias.

The effectiveness of the approach on different text sources or structured data remains uncertain. Differences in linguistic styles, terminologies, and data availability across domains may limit generalization. The approach assumes that frequency-related information in historical texts can be accurately mapped to numerical frequency estimates. If the original texts contain qualitative descriptions rather than explicit quantifiers, this may introduce errors. Also, older survey texts may reflect sampling biases, observer subjectivity, or incomplete data. If LLMs learn from these biases, the resulting quantity estimations may reinforce historical inaccuracies rather than correct them.

Acknowledgements

We thank our annotators from the Journal Club, and the colleagues in the Computational Historical Ecology group at the Chair for Computational Humanities at the University of Passau for their feedback.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy. Association for Computational Linguistics.

Stefan Brönnimann, Christian Pfister, and Sam White. 2018. Archives of nature and archives of societies. In *The Palgrave Handbook of Climate History*, pages 27–36. Palgrave Macmillan UK, London.

Robert M Dorazio, Nicholas J Gotelli, and Aaron M Ellison. 2011. Modern methods of estimating biodiversity from presence-absence surveys. *Biodiversity loss in a changing planet*, pages 277–302.

Maria Dornelas, Anne E. Magurran, Stephen T. Buckland, Anne Chao, Robin L. Chazdon, Robert K. Colwell, Tom Curtis, Kevin J. Gaston, Nicholas J. Gotelli, Matthew A. Kosnik, Brian McGill, Jenny L. McCune, Hélène Morlon, Peter J. Mumby, Lise Øvreås, Angelika Studeny, and Mark Vellend. 2013. Quantifying temporal change in biodiversity: challenges and opportunities. *Proceedings of the Royal Society*, 280.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Thomas van Goethem and Jan Luiten van Zanden. 2021. Biodiversity trends in a historical perspective. In *How Was Life? Volume II: New Perspectives on Well-being and Global Inequality since 1820*. Organisation for Economic Co-Operation and Development (OECD).

Sk Rezaul Hoque and Sk Rima Sultana. 2024. Addressing global environmental problems: Challenges, solutions, and opportunities. *The Social Science Review: A Multidisciplinary Journal*, 2(2):124–130.

Kelly J Iknayan, Morgan W Tingley, Brett J Furnas, and Steven R Beissinger. 2014. Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, 29(2):97–106.

Kenneth G Johnson, Stephen J Brooks, Phillip B Fenberg, Adrian G Glover, Karen E James, Adrian M Lister, Ellinor Michel, Mark Spencer, Jonathan A Todd, Eugenia Valsami-Jones, Jeremy R Young, and John R Stewart. 2011. Climate change and biosphere response: Unlocking the collections vault. *Bioscience*, 61(2):147–153.

Mike Kestemont, Folgert Karsdorp, Elisabeth de Bruijn, Matthew Driscoll, Katarzyna A Kapitan, Pádraig Ó Macháin, Daniel Sawyer, Remco Sleiderink, and Anne Chao. 2022. Forgotten books: The application of unseen species models to the survival of culture. *Science*, 375(6582):765–769.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Vamsi Krishna Kommineni, Waqas Ahmed, Birgitta Koenig-Ries, and Sheeba Samuel. 2024. Automating information retrieval from biodiversity literature using large language models: A case study. *Biodivers. Inf. Sci. Stand.*, 8.

Christian König, Patrick Weigelt, Julian Schrader, Amanda Taylor, Jens Kattge, and Holger Kreft. 2019. Biodiversity data integration—the significance of data resolution and domain. *PLoS biology*, 17(3):e3000183.

Lars Langer, Manuel Burghardt, Roland Borgards, Ronny Richter, and Christian Wirth. 2024. The relation between biodiversity in literature and social and spatial situation of authors: Reflections on the nature–culture entanglement. *People and Nature*, 6(1):54–74.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Andy Lücking, Christine Driller, Manuel Stoeckel, Giuseppe Abrami, Adrian Pachzelt, and Alexander Mehler. 2022. Multiple annotation for biodiversity: developing an annotation framework among biology, linguistics and text technology. *Language resources and evaluation*, 56(3):807–855.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online. Association for Computational Linguistics.

Malte Rehbein, Andrea Belen Escobari Vargas, Sarah Fischer, Anton Güntsch, Bettina Haas, Giada Matheisen, Tobias Perschl, Alois Wieshuber, and Thore Engel. 2024. Historical animal observation records by bavarian forestry offices (1845): Description of the data sets. Version 1.3 as of 2024-10-29, `https://doi.org/10.5281/zenodo.14008158`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Anders Telenius. 2011. Biodiversity information goes public: Gbif at your service. *Nordic Journal of Botany*, 29(3):378–381.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Württembergische Oberamtsbeschreibungen. 1824–1886. Wikisource. Accessed: 30 Jan. 2025.

APPENDIX: PROMPTS

Multi-Classification Prompt

System-prompt: You are a German native expert in text classification. Use the provided classification scheme to classify German texts based on species frequency descriptions.

User-prompt: You are a classification model. Classify the given German text into one of the following categories:
- Abundant (5): Species is very frequently observed or present.
- Common (4): Species is commonly found in the area.
- Common to Rare (3): Species is observed, but not very frequently.
- Rare (2): Species is rarely seen in the area.
- Very Rare (1): Species is seen only in exceptional circumstances.
- Absent (0): Species is not observed in the area.
- Extinct (-1): Species no longer exists in the area.

Read the provided text and classify it according to this scheme. Here is the text to classify:
Text

Best-Worst Scaling Prompt

System-prompt: You are an expert annotator specializing in Best-Worst Scaling of German texts based on quantity information about animal occurrences.

User-prompt: (Texts 1 to 4 were substituted with the actual texts of a tuple): Task: From the following German texts about animal occurrence, identify:

Best: The text conveying the highest quantity (e.g., presence, frequency, population size)

Worst: The text conveying the lowest quantity.

1. Text 1
2. Text 2
3. Text 3
4. Text 4

JSON format for your answer:

{ "Best": [Text Number],

"Worst": [Text Number]}