Received: 27 March 2024 | Accepted: 27 January 2025

DOI: 10.1002/aps3.70005

Applications in Plant Sciences

APPLICATION ARTICLE

Fully automatic extraction of morphological traits from the web: Utopia or reality?

Diego Marcos[1,2] | Robert van de Vlasakker[3†] | Ioannis N. Athanasiadis[3] | Pierre Bonnet[4] | Hervé Goëau[4] | Alexis Joly[5] | W. Daniel Kissling[6] | César Leblanc[2,4,5] | André S. J. van Proosdij[3] | Konstantinos P. Panousis[1,2]

[1]INRIA, TETIS, University of Montpellier, Montpellier, France

[2]University of Montpellier, Montpellier, France

[3]Wageningen University & Research, Wageningen, The Netherlands

[4]AMAP, University of Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France

[5]INRIA, LIRMM, Montpellier, France

[6]University of Amsterdam, Amsterdam, The Netherlands

Correspondence

Diego Marcos, INRIA, TETIS, University of Montpellier, Montpellier, France.
Email: diego.marcos@inria.fr

Abstract

Premise: Plant morphological traits, their observable characteristics, are fundamental to understanding the role played by each species within its ecosystem; however, compiling trait information for even a moderate number of species is a demanding task that may take experts years to accomplish. At the same time, online species descriptions contain massive amounts of information about morphological traits, but the lack of structure makes this source of data impossible to use at scale.

Methods: To overcome this, we propose to leverage recent advances in large language models and devise a mechanism for gathering and processing plant trait information in the form of unstructured textual descriptions, without manual curation.

Results: We evaluate our approach by automatically replicating three manually created species–trait matrices. Our method found values for over half of all species–trait pairs, with an F1 score of over 75%.

Discussion: Our results suggest that large-scale creation of structured trait databases from unstructured online text is now feasible due to the information extraction capabilities of large language models. However, the process is currently limited by the availability of textual descriptions that cover all traits of interest.

KEYWORDS

automatic trait extraction, large language models, morphological trait matrices, natural language processing

Traits are observable characteristics of organisms that can be used to answer a variety of questions about their ecology, evolution, and even usefulness to humans. Morphological traits (i.e., those that correspond to the physical appearance of the organisms, such as the number and color of flower petals, the size and shape of the fruits, or the leaf arrangement), in particular, are the main cues that humans have been using for centuries to identify species. However, the sheer number of known species, the variety of morphological traits, and the complexity of trait-based descriptions make it extremely challenging to design a comprehensive framework for trait-based descriptions that would be suitable across taxonomic groups. To address these challenges, recent efforts have advocated for a standard vocabulary to make trait databases cross-compatible (Schneider et al., 2019) and an open science initiative to leverage the collective efforts of the community (Gallagher et al., 2020). Nonetheless, this complexity has resulted in most existing trait databases being limited in either geographic (Falster et al., 2021) or taxonomic scope (Kissling et al., 2019). Moreover, large community efforts such as TRY (Kattge et al., 2011), the Botanical Information and Ecology Network (BIEN) database (Maitner et al., 2018), or TraitBank (Caldwell and Hart, 2014), which aim at covering all plant species, are far from being comprehensive or representative (Kattge et al., 2020), despite having amassed

[†]Deceased 27 May 2023.

millions of contributed trait measurements. For instance, in TRY version 6, the 30 species with the highest number of traits are from Western Europe (27 species) and North America (three species), showcasing a common imbalance in which relatively less data is available for species from biodiverse tropical regions. On the other hand, over 80% of plant species in TRY have 10 traits or fewer (Kattge et al., 2020).

At the same time, taxonomists have been carefully categorizing and describing traits for the purpose of species identification since the dawn of taxonomy and, more recently, using them for this task with modern machine learning approaches (Almeida et al., 2020). Many of these trait-based descriptions, capturing a vast expertise in different languages and with varying vocabularies, along with large amounts of trait data, can now be found online in the form of textual descriptions. However, these data do not come in a structured, ready-to-process format, and thus a thorough and laborious curation process is needed to render the data usable (Endara et al., 2018; Folk et al., 2023). For instance, using a partially automated workflow on up to 25,000 Australian taxa, Coleman et al. (2023) estimated that between eight and 23 hours of manual work are required per trait, with most of the time being required for manual verification. Domazetoski et al. (2023) aimed at reducing the reliance on manual labor by training a natural language processing (NLP) model to output the trait values for a limited number of traits (those the model has been trained on) when provided with a textual description. This process supersedes the need for manual work with the need for structured trait information for training, which explains why the authors limited the approach to eight traits.

In this work, we explore the potential of leveraging these morphological descriptions, in the form of text, for filling in gaps in structured trait databases. We posit that recent advances in NLP models, and particularly large language models (LLMs), bring us closer to exploiting this knowledge in an automated manner. LLMs have been shown to behave as remarkable zero-shot learners (Kojima et al., 2022); this means that they can be leveraged to solve tasks without a single training example via the use of textual instructions in natural language. Among these tasks, LLMs have been shown to excel at the extraction of structured information from text (Wei et al., 2023). To this end, we investigate the feasibility of a workflow that, given the names of the species of interest, along with the traits and possible trait values being considered, fills in a species–trait matrix using web crawling and LLMs. This is in contrast to related approaches for plant trait extraction, which require manual input for either post-processing (Endara et al., 2018; Coleman et al., 2023; Folk et al., 2023) or preparing a training set (Domazetoski et al., 2023).

METHODS

We propose a novel framework that requires only three inputs: (i) a list of species of interest, (ii) a list of categorical traits of interest, and (iii) a list (for each trait) with all the possible values that trait is allowed to take. The selected traits should correspond to those typically used in plant species descriptions. The output is a species–trait table that indicates, for each species, which trait values pertain to it. Specifically, the workflow (see Figure 1) can be divided into the following steps:

Textual data harvesting: A search engine API is used to retrieve URLs that are relevant to the species name and downloads the text content therein.

Description detection: To filter out irrelevant text, a binary classification NLP model is used to detect description sentences within the retrieved text.

Trait information extraction: An LLM is then used to detect all possible categorical trait values within the descriptive text.

Species–trait datasets for evaluation

To evaluate the automatic trait extraction workflow, we fixed the species, traits, and trait values to those found in three manually created species–trait matrices, using the following databases:

Caribbean: 42 woody species in the Dutch Caribbean. The database, created in the context of this work, contains 24 traits, with an average of 8.5 possible values per trait (minimum of two and a maximum of 22).

West Africa: 361 species of trees in the West African savanna (Bonnet et al., 2008). We considered all 23 traits, averaging 5.8 possible values per trait (minimum of two and a maximum of 10).

Palms: We used the 333 species that have complete trait descriptions (Kissling et al., 2019). We considered the six categorical traits in the dataset, with an average of 9.5 possible values per trait (minimum of two and a maximum of 31).

Textual data harvesting and description detection

Textual data harvesting

For each species of interest, we used the Google Search API to submit a query with the binary scientific name of the species, in quotes, to make sure that the search engine only returns websites containing the exact species name. The first 20 URLs returned were then visited and the text scraped; only HTML sites are considered in this work. We double-checked that the species name was present in the HTML header in order to filter out web pages not specifically dedicated to that species. Because the text obtained in this way is unstructured and a large part of it does not correspond to morphological descriptions, we selected the sentences most likely to be part of a description by using a custom text classifier, which is described in the following

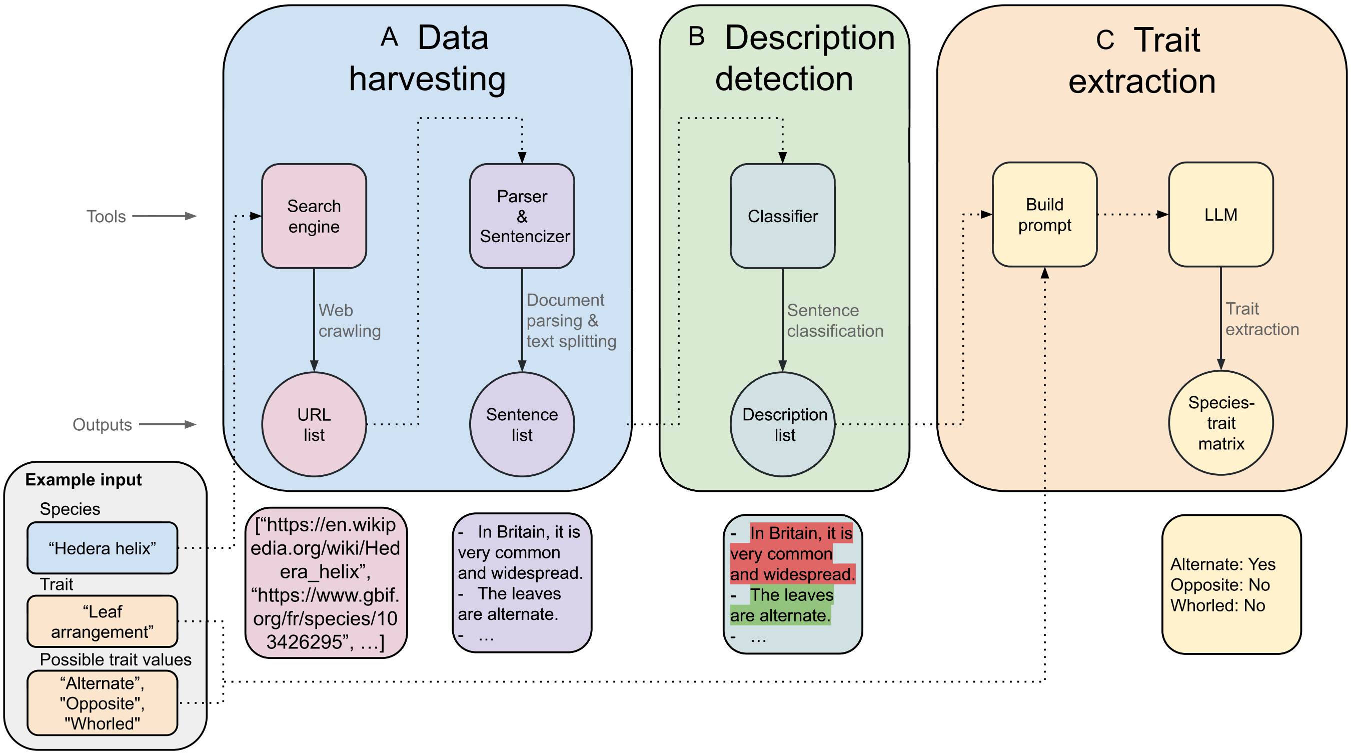


FIGURE 1 Overview of the methodology. The panels display the sequence of tasks performed during each of the three main stages: (A) data harvesting, (B) description detection, and (C) trait extraction. Below each task is an example of what its output may look like.

section. The full list of internet domains contributing to the harvested text can be found in the GitHub repository (see the Data Availability Statement).

Description detection

We start by formulating an approach to distinguish between descriptive and non-descriptive sentences in the form of an NLP binary classification task, aimed at filtering out all the text from the retrieved websites that does not describe the morphology of the species. For instance, on the English-language Wikipedia page referring to *Hedera helix* L., the sentence "The fruit are purple-black to orange-yellow berries," from the "Description" section, explicitly describes morphological traits. On the other hand, the sentence "Once ivy is established it is very difficult to control or eradicate," from the "Control and eradication" section, does not explicitly describe any morphological traits and is thus considered non-descriptive. We are interested in sentences, as in the first example, from which trait values can potentially be extracted; therefore, an automated approach is needed to determine whether or not any given sentence is descriptive. Such a model can be trained without the need for manual annotations by leveraging structured online sources, such as Wikipedia, in which a "Description" section is often present and can be used to obtain descriptive training samples, while text from the other sections can be used as non-descriptive samples. This model can in turn be used to collect descriptive sentences from other, less structured but relevant, websites for further processing.

Creation of the training dataset

To create a large dataset of descriptive and non-descriptive text via parsing structured websites, we selected four different web sources that (i) comprise large databases and (ii) have rich scholarly content about species descriptions, namely:

Wikipedia: the best-known free online encyclopedia, maintained by volunteers; everyone is allowed to edit pages, while moderators maintain the quality of the content (https://en.wikipedia.org/wiki/Main_Page).

Plants of the World Online (POWO): an international collaborative database of the world's flora. The data are based on scientific publications and are maintained by the Royal Botanic Gardens, Kew (https://powo.science.kew.org/ [retrieved 13 July 2021]).

Encyclopedias of Living Forms (LLifle): a collaborative effort to provide species descriptions, offering descriptions of 31,213 plant species, with a focus on xerophytes (https://www.llifle.com/).

World Agroforestry Centre (ICRAF) Agroforestree database: provides textual descriptions of 670 tree species that are useful in agroforestry (https://apps.worldagroforestry.org/treedb/index.php).

All of these sources contain structured plant species information divided into different sections (e.g., "Introduction," "Appearance," "Characteristics," and "Habitat"), which are specific to each source. These section headers allow for automatic labeling of the data; for example, text stemming from the sections "Introduction" and "Habitat" is assigned non-descriptive labels (i.e., they are not relevant to the description of the species), while text from the sections "Characteristics" and "Appearance" is assigned to the descriptive class. We also consider random pages from Wikipedia, not pertaining to species, as an augmentation approach that enriches the non-descriptive data.

Training the classifier

We can then proceed with training a description detector. For this task, we need a model that is able to assign a binary label (descriptive versus non-descriptive) to a piece of text of arbitrary length. The most straightforward approach for this is to use a text encoder model, which can convert a text sequence of any size (up to some maximum allowed length) into a vector of fixed length. Any machine learning classifier can then be trained, in a supervised manner, using this vector representation as input. For our description sentence classification model, we used a distilled version of BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019), a widely used NLP model for obtaining vector representations of sentences, and specifically to DistillBERT (Sanh et al., 2020). This decision was motivated by the balance between complexity and performance that DistillBERT exhibits. This variant comprises 40% fewer parameters than the original BERT model, leading to 60% faster computations, while still yielding 97% performance on general language understanding. Both BERT and DistillBERT have been trained on a large corpus of English text, and pre-trained model weights are freely available. Within this context, we augmented the base DistillBERT model by introducing (i) a dropout layer for regularization purposes and (ii) two fully connected layers. The first layer takes the output vector of DistillBERT, of size 768, and outputs a vector of size 512, while the second layer takes this output vector and yields an output of size 2, which are the logits for our binary classification task.

To prepare the collected text for fine-tuning the model, it must first be split into discrete tokens, which are either syllables or entire words, for which we use a tokenizer (Wolf et al., 2020). The conventional BERT architecture can accommodate up to 512 tokens at once, corresponding to approximately 400 words. This means that text spans that are longer than 512 tokens (e.g., a paragraph or a sentence) need to be truncated to length 512 to be compatible with DistillBERT. In this work, we randomly split the text into spans with a minimum of 10 and a maximum of 512 tokens. This functions as data augmentation and forces the model to capture the characteristics of descriptive sentences, even when not seeing the whole sentence, potentially providing robustness when the model is exposed to text not included in the training set, where sentences could have different length and structure compared to the training set. We trained and validated the model only on data stemming from Wikipedia and POWO, while text from LLifle and ICRAF were used exclusively for model evaluation. By using different data sources for training and testing, we can obtain a better estimate of the generalization performance of the method on arbitrary websites returned by the search engine API upon deployment. For the evaluation, we used a sentencizer (Honnibal et al., 2020) to split the text into sentences.

Noise robust loss function

Due to the use of automatically obtained labels, it is likely that the resulting dataset will contain inconsistencies, as is common when dealing with unstructured data (Kumar et al., 2020); after all, not all text within the Description section of a Wikipedia article consists of descriptive sentences, and some descriptive sentences may occur outside of it, typically in the introductory sections. It is also possible that some section headers may be missed through this process, further increasing the amount of noise in the final dataset. We can mitigate the effects of this potential inconsistency by turning to classification losses that are designed for robustness against noisy labels. We chose the "soft bootstrap" consistency objective (Reed et al., 2015; Marcos et al., 2022), which operates according to the principle that the labels are "diluted" by the model's current prediction, thus reducing the impact of the loss of data points in which the model confidently disagrees with the label. Specifically, the loss is computed as:

$$\mathrm{SoftLoss}(q, t) = \sum_{k=1}^{L} [\beta t_k + (1 - \beta) q_k] \log q_k \qquad (1)$$

where q are the predicted class probabilities, t are the observed noisy labels, and β is a balancing factor between the current prediction and the target. In this way, we can use the current state of the model to dynamically adapt the prediction targets, allowing the model to pay less attention to inconsistent labels. As the model improves its predictions over time, it becomes more coherent, allowing for assessment of the consistency of the noisy labels. We set $\beta = 0.20$ in a similar fashion to previous works (Zhang et al., 2020; Marcos et al., 2022).

For fine-tuning the model, we kept the DistillBERT parameters frozen and trained only the added classification head. We used the Adam optimizer to minimize Eq. 1 with a learning rate of 3.10^{-5}, a batch size of 32, and gradient clipping with a norm of 1.0. The model was fine-tuned until convergence, for a total of 35 epochs.

Trait information extraction

Information extraction with a generative LLM

The next step towards information extraction for species descriptions involves extracting relevant information from the obtained text snippets into a structured form. To this end, we leveraged the recent advances in LLMs, which have been empirically demonstrated to capture relational knowledge in the training data that can be extracted via natural language queries, also known as prompts (Ouyang et al., 2022). These models have been shown to perform well on relatively generic tasks, such as common-sense knowledge (Davison et al., 2019) or general knowledge (Petroni et al., 2019).

Although it is possible to directly query an LLM with a question, without providing any additional information, one should be aware of their tendency to provide responses that appear legitimate but that are completely unfounded, known as hallucinations (Zhang et al., 2023). This tendency is even more pervasive in specialized domains, including botany, that are characterized by long-tailed distributions in which a few elements are very abundant and many are extremely rare. This leads to LLMs being unreliable for this majority of uncommon elements. To mitigate this issue, we turned the task into information extraction from text via search engine retrieval (Lewis et al., 2020). We achieved this by feeding the LLM a piece of descriptive text obtained via the harvesting approach detailed in the "Textual data harvesting and description detection" section, along with questions referring to a predetermined set of traits and possible trait values. At the same time, we gave the LLM the option to explicitly state if the requested information is not available (NA) in the given text, mitigating potential hallucination issues that could otherwise arise. This means that only a subset of traits will be assigned a trait value, the proportion of which we will refer to as coverage rate.

Choice of LLM

The 32,000-token context window in Mistral Medium (version 2312, released on December 2023; Mistral AI, Paris, France) was sufficient to accommodate all the text for a given species along with the considered traits and trait values. In addition to this, we found Mistral Medium to be a good compromise in preliminary tests, with results substantially better than those of GPT-3.5 Turbo (Open AI, San Francisco, California, USA) and at a similar cost. We tested the Mixtral-8x7B (Mistral AI) and Llama 2 (Meta, Menlo Park, California, USA) open source models, but they provided results of lower quality than those of GPT-3.5 Turbo. However, we have conducted tests with the open source Mixtral-8x22B, which was released after we conducted most of our experiments, and obtained results comparable to Mistral Medium, making it a good alternative for labs with access to a GPU cluster. Refer to the section "Additional experimental results" (below) for details.

Prompt design

The considered species–trait datasets comprise categorical traits; these can be encoded in a binary form and expressed as multiple-choice textual questions to engineer discrete prompts. In this binary encoding context, we are interested in discovering which trait values should be "1" or "0" for any given set of species and trait value by exploiting the information from the retrieved description sentences. Such an encoding of the categorical traits "life form" and "phyllotaxis" of the Caribbean dataset is depicted in Table 1.

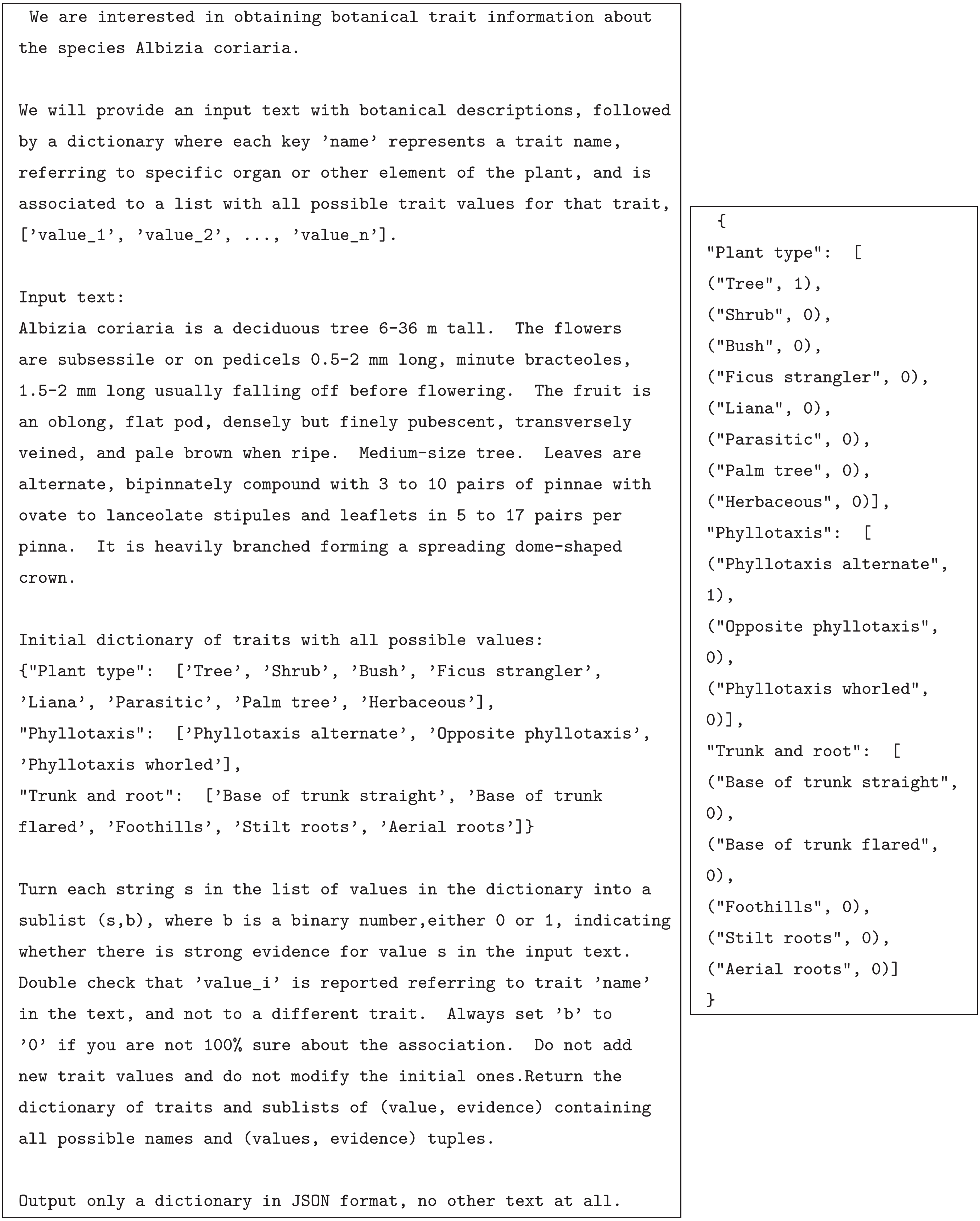
TABLE 1 A binary encoding (with presence denoted as "1" and absence as "0") of the manual annotations of two different traits for three different species in the Caribbean dataset.

	Life form		Phyllotaxis	
Species	**Tree**	**Liana**	**Alternate**	**Opposite**
Avicennia germinans (L.) L.	1	0	0	1
Metopium brownei (Jacq.) Urb.	1	0	1	0
Morisonia flexuosa L.	0	1	1	0

Thus, for each trait, we first group together all its possible values, e.g., "life form": ["tree", "liana"] and "phyllotaxis": ["alternate", "opposite"]. Then, to create a prompt for the LLM, we consider: (i) all the textual description sentences about a species and (ii) the list of traits and considered trait values as described before. Based on this construction, we prompt the LLM to infer the values for each trait based on the provided text. A realistic example prompt based on the described process is depicted in Figure 2 (left). In this example, we asked the LLM about three traits. For the first two, "plant type" and "phyllotaxis," there is some information available in the input text: "[…] is a deciduous tree" and "Leaves are alternate." For the third trait, related to "Trunk and root," no information is present in the text. Indeed, the actual response of the LLM when queried with the constructed prompt is shown in Figure 2 (right). Therein, we observe that the LLM correctly infers the values of each trait from the given textual description, exhibiting behavior consistent to what we expected. For the last trait, there is no evidence for any of the accepted trait values, and this will be treated as "NA."

Although the example prompt includes only three traits and provides information about a single species, we are interested in scaling this approach to hundreds or thousands of species and long lists of possible traits. Scaling the approach can be done by simply repeating the process for new species and including additional traits in the prompt. In our work, and to be able to compare the prompt results to the ground truth data of the three considered datasets (Caribbean, West Africa, and Palms), we consider the exact same species, traits, and trait values found in each dataset.

By using an LLM with a large enough context window, it is possible to fit the whole text and dictionary of traits into a single prompt. Alternatively, it is also possible to split the task into multiple prompts by querying about a single prompt at a time or by providing only a subset of the input text. The answers of the LLM can then be parsed in order to



```
 We are interested in obtaining botanical trait information about
the species Albizia coriaria.

We will provide an input text with botanical descriptions, followed
by a dictionary where each key ’name’ represents a trait name,
referring to specific organ or other element of the plant, and is
associated to a list with all possible trait values for that trait,
[’value_1’, ’value_2’, ..., ’value_n’].

Input text:
Albizia coriaria is a deciduous tree 6-36 m tall.  The flowers
are subsessile or on pedicels 0.5-2 mm long, minute bracteoles,
1.5-2 mm long usually falling off before flowering.  The fruit is
an oblong, flat pod, densely but finely pubescent, transversely
veined, and pale brown when ripe.  Medium-size tree.  Leaves are
alternate, bipinnately compound with 3 to 10 pairs of pinnae with
ovate to lanceolate stipules and leaflets in 5 to 17 pairs per
pinna.  It is heavily branched forming a spreading dome-shaped
crown.

Initial dictionary of traits with all possible values:
{"Plant type":  [’Tree’, ’Shrub’, ’Bush’, ’Ficus strangler’,
’Liana’, ’Parasitic’, ’Palm tree’, ’Herbaceous’],
"Phyllotaxis":  [’Phyllotaxis alternate’, ’Opposite phyllotaxis’,
’Phyllotaxis whorled’],
"Trunk and root":  [’Base of trunk straight’, ’Base of trunk
flared’, ’Foothills’, ’Stilt roots’, ’Aerial roots’]}

Turn each string s in the list of values in the dictionary into a
sublist (s,b), where b is a binary number,either 0 or 1, indicating
whether there is strong evidence for value s in the input text.
Double check that ’value_i’ is reported referring to trait ’name’
in the text, and not to a different trait.  Always set ’b’ to
’0’ if you are not 100% sure about the association.  Do not add
new trait values and do not modify the initial ones.Return the
dictionary of traits and sublists of (value, evidence) containing
all possible names and (values, evidence) tuples.

Output only a dictionary in JSON format, no other text at all.
```

```
 {
"Plant type":  [
("Tree", 1),
("Shrub", 0),
("Bush", 0),
("Ficus strangler", 0),
("Liana", 0),
("Parasitic", 0),
("Palm tree", 0),
("Herbaceous", 0)],
"Phyllotaxis":  [
("Phyllotaxis alternate",
1),
("Opposite phyllotaxis",
0),
("Phyllotaxis whorled",
0)],
"Trunk and root":  [
("Base of trunk straight",
0),
("Base of trunk flared",
0),
("Foothills", 0),
("Stilt roots", 0),
("Aerial roots", 0)]
}
```

FIGURE 2 An example prompt used to query the LLM about the presence of morphological traits given a textual description sentence for a given species (left), along with the LLM response (right). The LLM correctly identifies that there is evidence in the text indicating that the plant type is tree and the phyllotaxis is alternate, while no evidence can be found for the other trait values.

build a species–trait matrix in the same format and the manual annotations.

Evaluation metrics

Evaluation of the automatic trait extraction

To evaluate the responses of the LLM, we compared them to the species–trait matrices manually curated by expert botanists (Caribbean, West Africa, and Palms datasets, described above). We report the proportion of traits for which a value was found (i.e., the coverage), along with the precision, recall, and F1 score computed for the found traits; precision, recall, and F1 score are computed only for traits with coverage. The precision is the proportion of correctly predicted positives (i.e., all the trait values predicted by the model as being present) compared to the manual dataset. The recall is the proportion of positives in the manual dataset that are retrieved by the approach. To combine these two complementary metrics, the F1 score consists of the geometrical average of precision and recall.

Evaluation of the false negative rate

Even though the described evaluation process allows for assessing whether the detected traits are correct, according to manually created species–trait matrices, it does not allow for quantifying the false negative rate of the LLM (i.e., whether all traits described in the text are effectively extracted). In this context, we need to assess whether the false negatives arise due to the LLM extraction process or due to the fact the information is simply not present in the harvested text. To mitigate this issue, we performed an additional evaluation of the trait extraction process by asking four senior botanists whether a certain trait value can be inferred from a specific piece of text and then using this information as ground truth. Specifically, we first randomly selected a trait from one of the species–trait datasets. We then selected a random species from the same dataset and picked a text snippet with a low distance in the DistillBERT embedding space to the name of the trait, in order to increase the number of relevant text–trait pairs. This allowed us to generate 1216 text–trait pairs, which we then shared with the botanists. They were given the text and the ground truth trait value and asked if this trait value could be inferred from the text. According to the botanists, 298 of the 1216 snippets contained relevant information about the trait of interest. To assess the capacity of the LLM extraction process in this setting, we constructed the corresponding prompts with the same pairs of sentences and traits as those presented to the botanists. The prompt used was of the same structure as the one shown in Figure 2. This allows us to investigate whether the LLM behaves in an excessively conservative manner, preferring to return empty results rather than make a mistake, or has a tendency to hallucinate responses that are not explicitly in the text.

RESULTS

Descriptive text classification

The descriptive/non-descriptive dataset creation process described in the “Descriptive text classification” section resulted in approximately 1.45 million sentences: 1.1 million sentences corresponding to non-descriptive text and 356,000 sentences corresponding to descriptive text. In the performance analysis of the model (Table 2), we observe that, within the in-domain validation set, our description classification model achieves very high precision for both classes (i.e., “Descriptive” and “Non-Descriptive”), with F1 scores of 0.96 and 0.99, respectively. However, the recall in the test set decreases substantially for the descriptive class, from 0.95 to 0.55, but remains basically stable for the non-descriptive class, at 0.99 for the validation set and 0.98 for the test set. A few example sentences with their corresponding score are shown in Figure 3, where we can see that the model behaves as expected, with only botanical descriptions having a score higher than 0.5.

Descriptive text harvesting

Having trained and validated our descriptive sentence detector, we next considered any potential source of textual information to extract species descriptions towards a downstream task. The text harvesting step returned description text for the majority of species, but not for all. Specifically, we obtained text for 40/42 species in the Caribbean dataset, 358/361 for the West Africa dataset, and 248/333 for the Palms dataset. On average, we obtained 35, 36.8, and 43.5 descriptive sentences per species for each dataset, respectively. Refer to the code repository to find the whole set of found descriptive sentences and their original URLs (https://github.com/konpanousis/AutomaticTraitExtraction/blob/main/Descriptions).

Automatic trait extraction

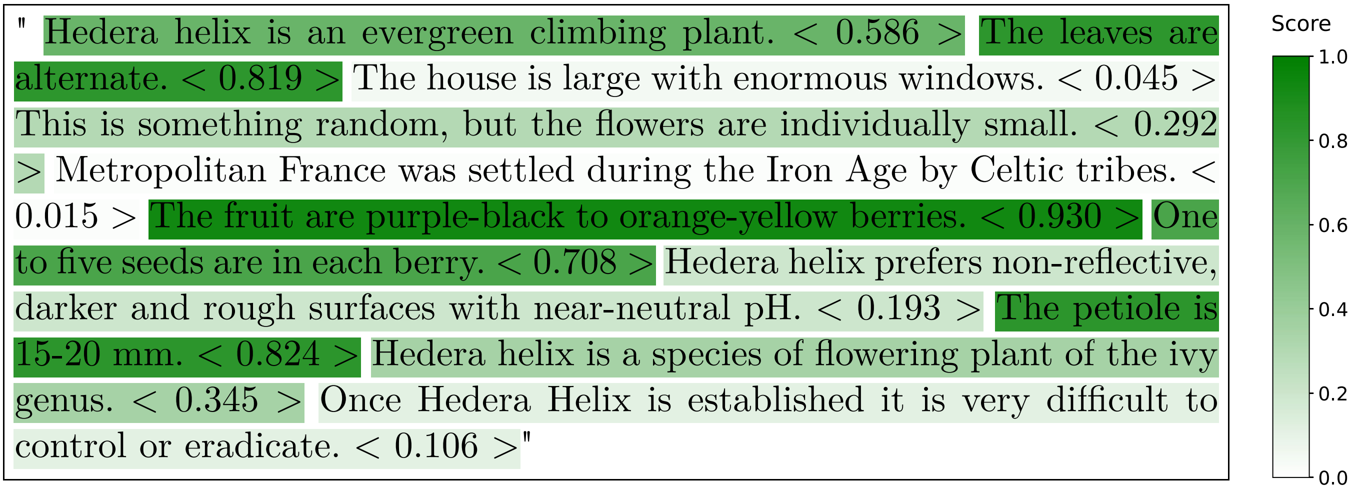
Comparison to manually curated trait data

The coverage ranges between 55% and 56% (Table 3), meaning that the described method assigns a value to more than half of the traits. The F1 scores range between 73% in the Palms dataset and 78% in the West Africa dataset, the recall is remarkably constant at between 77% and 78%, and the precision varies between 70% in the Palms dataset and 80% in the West Africa dataset.

Both the per-trait F1 scores and coverages display large variations (Figure 4). Some commonly found traits, such as *life form* and *seed color* in the Caribbean dataset or *plant type* and *leaf shape* in the West Africa dataset, have been retrieved for well above 80% of species, while *trunk* and *root* are only found for around 10% of species in the

TABLE 2 The precision-recall metrics for the binary classification model tested on the test dataset and two external datasets (LLifle and AgroForestree).

	Precision		Recall		F1 score		No. of sentences	
Class	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Non-descriptive	0.98	0.94	0.99	0.98	0.99	0.96	167,955	66,246
Descriptive	0.97	0.79	0.95	0.55	0.96	0.65	57,864	8590



" Hedera helix is an evergreen climbing plant. < 0.586 > The leaves are alternate. < 0.819 > The house is large with enormous windows. < 0.045 > This is something random, but the flowers are individually small. < 0.292 > Metropolitan France was settled during the Iron Age by Celtic tribes. < 0.015 > The fruit are purple-black to orange-yellow berries. < 0.930 > One to five seeds are in each berry. < 0.708 > Hedera helix prefers non-reflective, darker and rough surfaces with near-neutral pH. < 0.193 > The petiole is 15-20 mm. < 0.824 > Hedera helix is a species of flowering plant of the ivy genus. < 0.345 > Once Hedera Helix is established it is very difficult to control or eradicate. < 0.106 >"

FIGURE 3 An example set of sentences and their corresponding "description" score. The text is broken into single sentences by the sentencizer of Honnibal et al. (2020), and the classifier classifies each sentence. Sentences with a value of 0.50 or higher are stored in the database. The darker the color green, the higher the prediction value. The prediction value is also shown after each sentence.

TABLE 3 Precision, recall, F1 score, and coverage (i.e., the proportion of species–trait entries for which at least one value is found) with respect to the three manually curated databases. The accuracy metrics are computed only for these entries.

Dataset	Precision	Recall	F1 score	Coverage
Caribbean	0.7493	0.7800	0.7643	0.5500
West Africa	0.8058	0.7776	0.7806	0.5588
Palms	0.7013	0.7706	0.7343	0.5584

West Africa dataset. Moreover, large variations in terms of F1 accuracy can be observed across traits in all datasets, with a tendency towards higher accuracy in traits for which fewer values are allowed. For instance, *life form* in the Caribbean dataset has only two possible values, and the F1 accuracy stands at over 95%. On the other hand, *fruit color* in the Palms dataset has around 70% F1 accuracy for 12 possible values.

To visualize the most typical mistakes the model commits on this multi-label task, in which more than one trait value is allowed per trait, we show two co-occurrence matrices side by side: one corresponding to the co-occurrences found within the annotated data (i.e., which trait values are simultaneously present in a species) and a second one with the co-occurrences between the annotations and our predictions (which values are predicted for species that are annotated with a certain value) (Figure 5). The comparison shown here, for the traits *leaf position* and *fruit type* in the Caribbean dataset and *fruit* in the West Africa dataset, demonstrates that the general patterns of co-occurrence are maintained and, furthermore, that the committed confusions are often reasonable. For instance, for the *leaf position* trait, our approach returned *opposite* when the manual annotations stated *alternate-opposite*, *opposite*, *whorls of 3* and *opposite*, *whorls of 3*, *alternate*. Similar co-occurrence matrices are also shown for two traits in the Palms dataset: *fruit size*, with two possible values, and *fruit color*, with 12 possible values (Figure 6). In this example, although the correct values are often retrieved, the large number of options and potential ambiguities results in a much larger number of false positives. Finally, co-occurrence matrices are shown for the two traits with the highest and lowest overall F1 scores: *stem shape* and *leaf apex*, both from the West Africa dataset (Figure 7). We can see that the high scores in *stem shape* are driven both by the fact that only two possible values are allowed and that it is a very imbalanced trait, with the vast majority of species having the same value. On the other hand, *leaf apex* not only has seven possible values, but these values also show a very high overlap, which can be seen in the large off-diagonal values in three values of the annotations' co-occurrence. Our pipeline mostly tends to predict one of these three traits (*leaf apex acuminate*), while ignoring the other two.

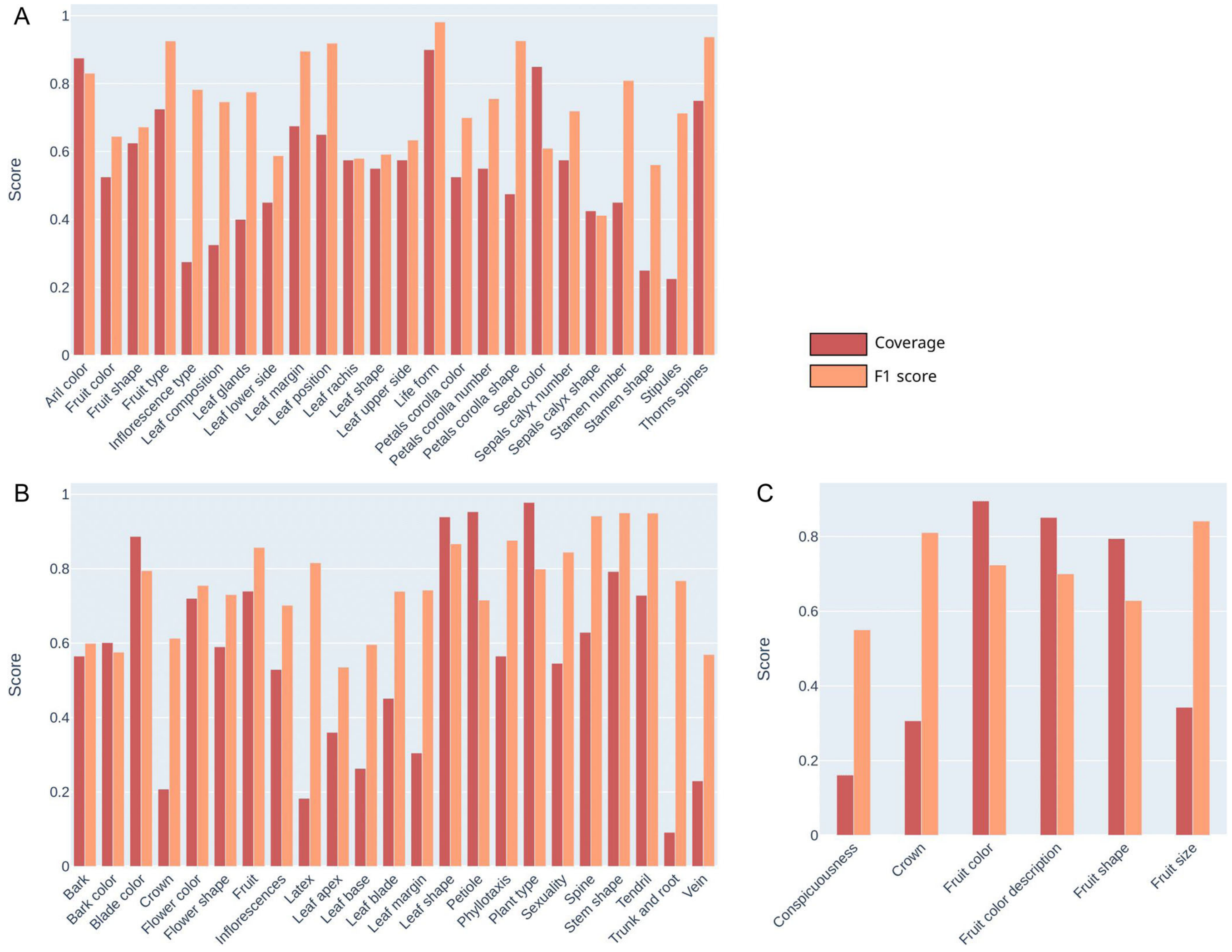


FIGURE 4 F1 score (orange) and coverage (red) per trait with respect to the three manually curated databases: (A) Caribbean, (B) West Africa, and (C) Palms. The coverage is the proportion of species for which at least one value is found. The F1 score is computed only for these species.

Evaluation of the false negative rate

In this section, we compare the ability of the LLM to predict "NA" in cases where no information about the desired trait can be found by comparing its responses to those of expert botanists on the same sentences. This allows us to estimate whether the coverage rate corresponds to the actual data availability in the harvested text. The confusion matrix in Table 4 shows that the LLM, in the setting used in this work, does not have a strong bias towards over- or under-detecting traits in text. Out of the 1216 text samples used in the survey, 24% were deemed to contain relevant trait information by the botanists, while the LLM reported found traits in 22%. By comparing the responses between the LLM and the botanists, we obtain a macro-averaged F1 score of 0.82, with an F1 score of 0.72 for the positive class and 0.92 for the negative class. The precision being higher than the recall suggests that the model has a conservative bias, with a tendency to under-report traits rather than hallucinating them. For around 32% of the traits reported as "NA," information was present in the text that was missed by the LLM. The observed precision is roughly in line with the performance of the approach, using the whole per-species text and set of traits in a single prompt, when compared to the manually curated species–trait matrix (Table 3), suggesting that the amount of input text provided in the prompt does not affect the quality of the results.

Additional experimental results

To further investigate the impact of some of the design choices, we evaluated the Caribbean dataset using two additional LLM settings. First, we wanted to verify that querying the LLM with all traits simultaneously in a single prompt does not substantially degrade the results. We observe that querying a single trait at a time results in a mere half percentage point improvement for the Mistral Medium model (Table 5), while the number of input tokens required is more than an order of magnitude larger, from

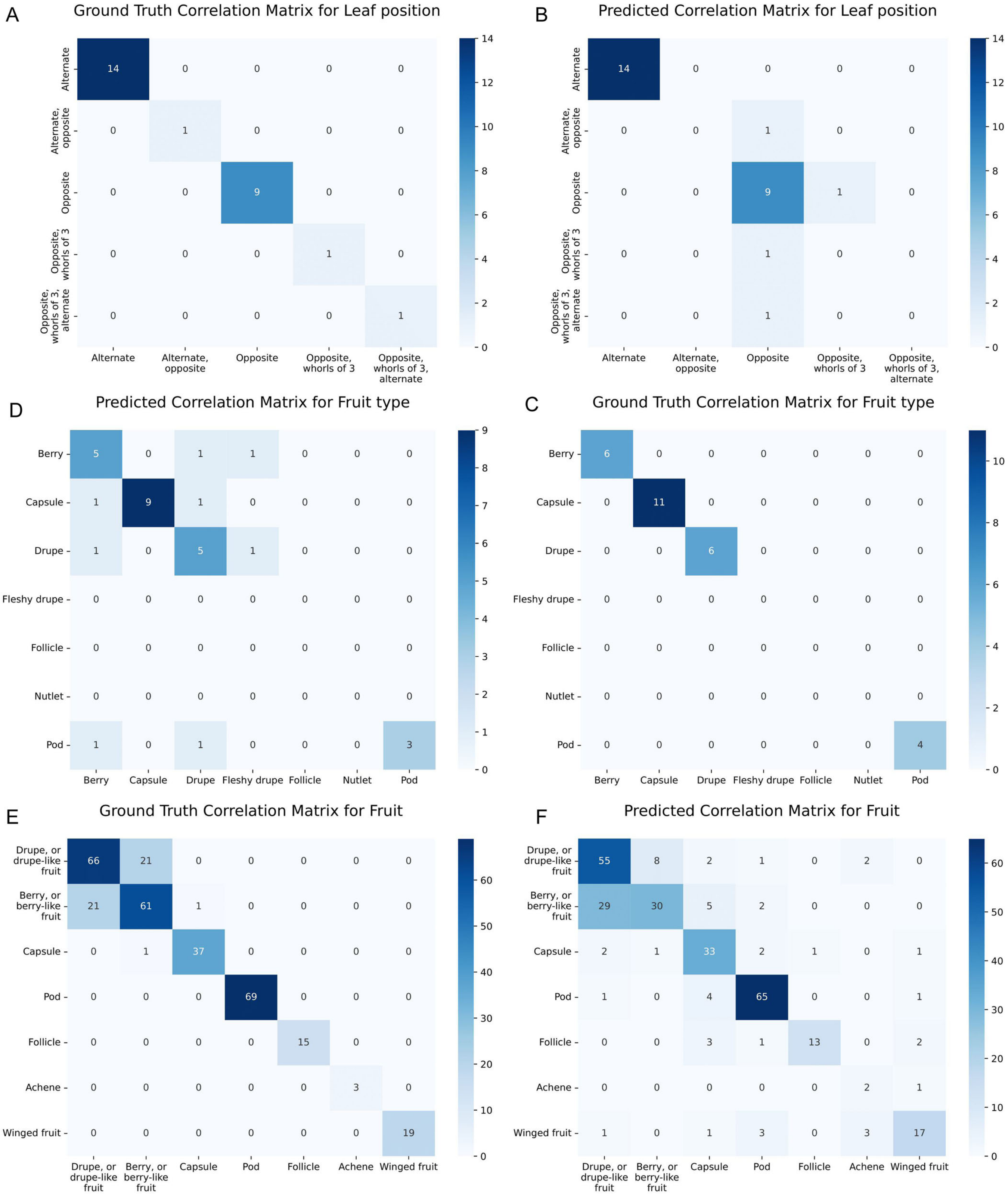


FIGURE 5 Co-occurrence matrices for every pair of trait values in *leaf position* (A, B) and *fruit type* (C, D) in the Caribbean dataset, and *fruit* in the West Africa dataset (E, F). For each trait, we compare the co-occurrences within the annotations (left) and the co-occurrences between the predictions (columns) and the annotated values (rows) (right). We can see that the patterns of co-occurrence are maintained.

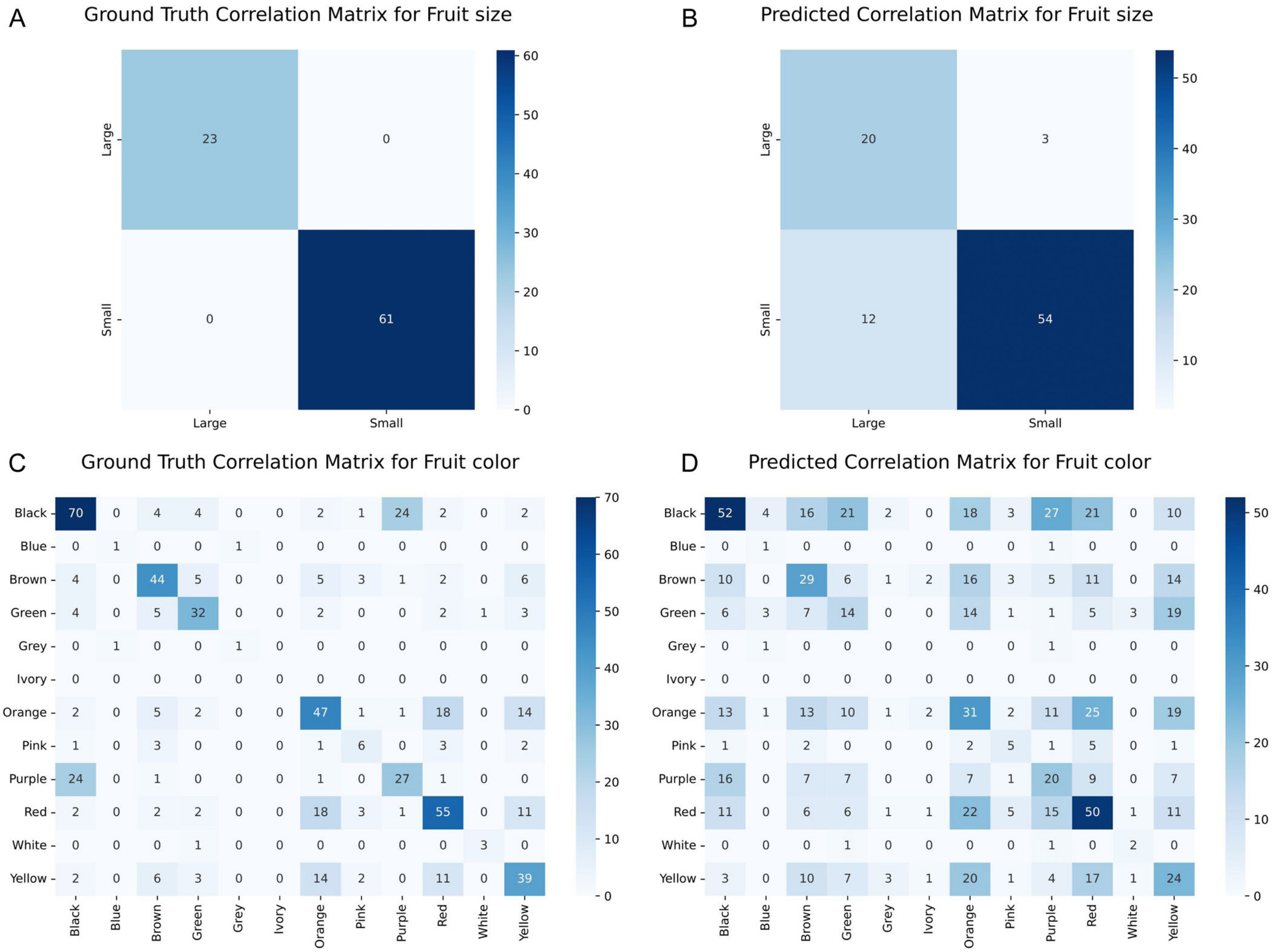


FIGURE 6 Co-occurrence matrices for *fruit size* (A, B) and *fruit color* (C, D) in the Palms dataset. For each trait, we compare the co-occurrences within the annotations (left) and the co-occurrences between the predictions (columns) and the annotated values (rows) (right). We can see that the patterns of co-occurrence are maintained.

150,000 input tokens to 1.67 million, due to having to provide the input text and instructions as many times per species as there are traits. The main improvement is in coverage, which increases from 55% to almost 58%. It should be noted that running our experiments with all traits at once on the over 700 plant species that comprise the three datasets only required around $30 USD in Mistral AI credits. Second, we evaluated the applicability of the open source model Mixtral-8x22B, which was released after we had run the main set of experiments, and found that this model is able to provide comparable results to Mistral Medium, with only a 0.4% lower F1 score and an improved coverage of over 60% (Table 5).

DISCUSSION

Our study reveals that while most species yielded useful descriptive text, a significant portion of species in the Palms dataset provided no sentences at all. This issue is partly attributable to the study's focus on English-language HTML websites and could be mitigated with the inclusion of non-English text, as many botanical descriptions are available in local languages. Expanding the language scope would be particularly feasible for languages with a significant online presence and a history of botanical use, such as French, Spanish, and Portuguese. Broadening the range of structured online resources used to train the description detector is one alternative for this multilingual expansion. Although modern LLMs are multilingual, the most significant challenge lies in extending our descriptive text detection approach to multiple languages. This could be addressed either by training on structured websites in various languages or by employing a higher-capacity LLM in this step via in-context learning (Brown et al., 2020), which would leverage the multilingual nature of most LLMs. With this last approach, a few examples of descriptions or a definition of the characteristics of descriptions could suffice, removing the need to train on a large set of multilingual description examples. The markedly lower recall compared to precision

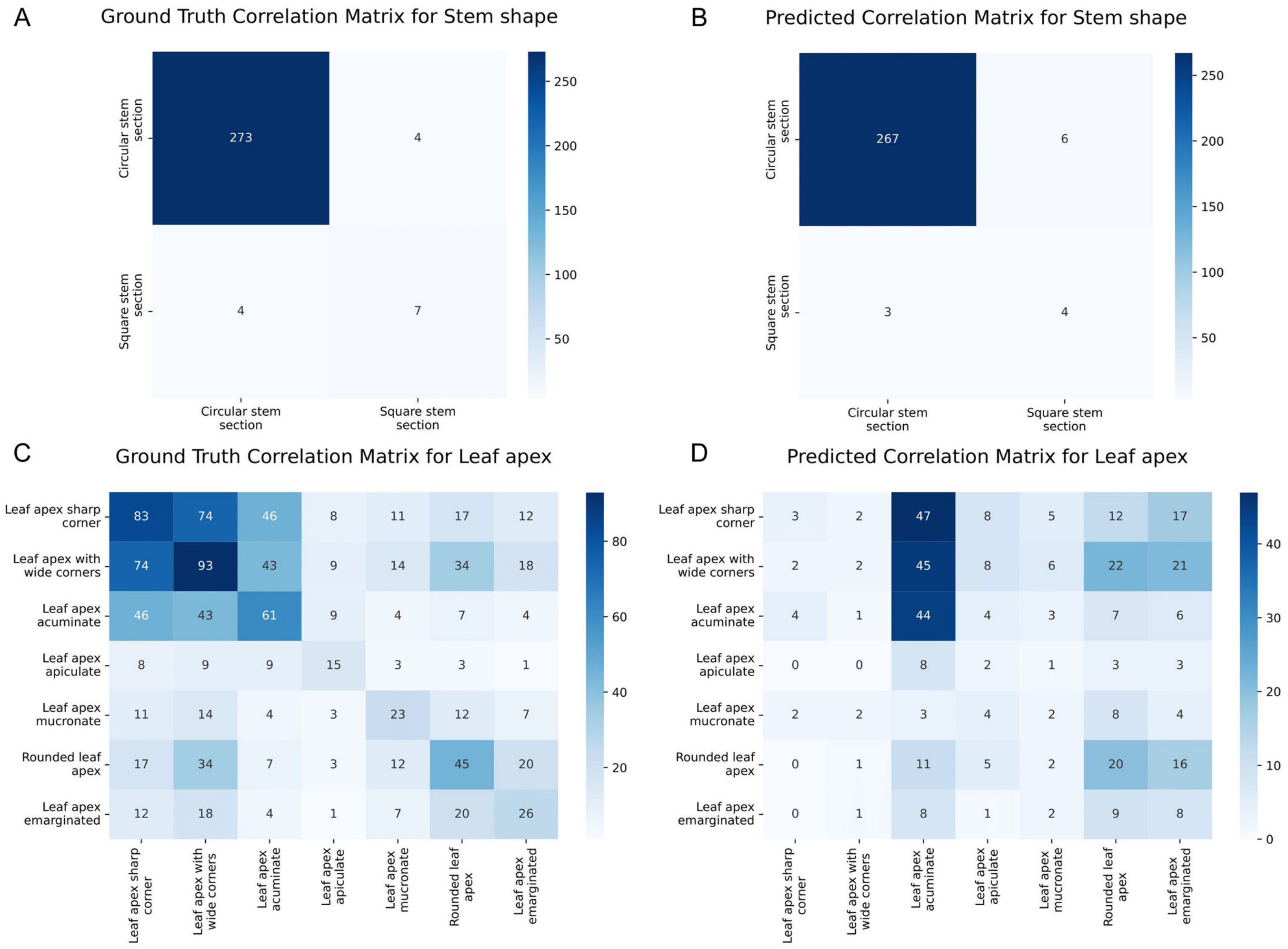


FIGURE 7 Co-occurrence matrices for *stem shape* (A, B) and *leaf apex* (C, D) in the West Africa dataset (the traits with the highest and lowest F1 scores, respectively [i.e., 0.95 and 0.54]). For each trait, we compare the co-occurrences within the annotations (left) and the co-occurrences between the predictions (columns) and the annotated values (rows) (right). The co-occurrence patterns are conserved, except for *leaf apex sharp corner* and *leaf apex wide corners*, which are generally predicted as *leaf apex acuminate*. Note that these three trait values are highly correlated in the annotation.

TABLE 4 Confusion matrix between the ground truth (GT) survey responses and the responses from the LLM. In this experiment, for the LLM queries, instead of using only the known ground truth values for each trait, we consider all its possible values. If there is evidence of any value present in the sentence, we consider the value found; otherwise, the value is considered missing. For the GT survey responses, we consider the value found only when the reviewers responded "Can infer correct value"; otherwise, the value is considered missing.

	LLM missing	LLM found	Total
GT missing	856	62	918
GT found	94	204	298
Total	950	266	1216

TABLE 5 Precision, recall, F1 score, and coverage using four different models and settings.[a]

Model (setting)[b]	Precision	Recall	F1 score	Coverage
Mistral Medium (all traits)	0.7493	0.7800	0.7643	0.5500
Mistral Medium (single trait)	0.7507	0.7920	0.7708	0.5791
Mixtral-8x22B (all traits)	0.7519	0.7726	0.7602	0.6052
GPT-3.5 Turbo (all traits)	0.6665	0.7390	0.7009	0.3031

[a] Precision, recall, F1 score, and coverage (i.e., the proportion of species–trait entries for which at least one value is found) are calculated with respect to the manually annotated Caribbean dataset. The accuracy metrics are computed only for these entries.

[b] The "all traits" setting used a single prompt per species, querying for all traits simultaneously, while the "single trait" setting queried for a single trait in each prompt.

in detecting descriptive text is expected, given the nature of the data and the loss function used, which accounts for a considerable amount of label noise. The low recall suggests that the model often determined that nearly half of the text within descriptive sections did not genuinely pertain to descriptions. While this may lead to the omission of some relevant sentences, it also results in a more concise and focused corpus.

Our quantitative trait extraction results show that the proposed pipeline is able to return a value for over half of the traits in the three considered species–trait matrices, with an average F1 score of over 0.75. In addition, the inspection of the errors committed by the pipeline suggests that they tend to be relatively reasonable mistakes, with similar trait values being typically confused for one another. The results of the false negative rate evaluation show that, in general, the LLM is well balanced and has no strong tendency towards either hallucinating or ignoring information. The fact that using a single trait per query results in very similar performance is a sign that this behavior does not depend on the number of simultaneously queried traits. These results mean that the low average coverage rate, of about 55%, is probably due to a lack of information in the harvested dataset, rather than on the LLM being unable to extract the information. A focus on improving the amount of textual information would, therefore, be the best way to further improve the trait coverage. Nonetheless, we have directly used the trait and trait value names as they were proposed by the original authors of the species–trait datasets. It is likely that these specific formulations are not the best possible for our task and can thus be optimized for prompting, such as by including descriptions of the traits and their possible values. In addition, the results using Mixtral-8x22B show that, at the time of publication, it would be possible to reproduce the results of this study, and scale the approach to new species, using a model with openly available weights.

This study focused on a relatively small number of plant species, approximately 700 in total, for which manually curated trait data were available for evaluation. We attempted to mitigate geographic bias toward Europe and North America by exclusively considering tropical species, which are more likely to be representative of the data availability for the world's flora than these overrepresented regions (Kattge et al., 2020). However, the study was limited to woody plants, and this constraint may affect the generalizability of our findings to the global flora. We also observed that our approach successfully filled in only a little over half of the traits, with up to 25% of species in the Palms dataset failing to yield any text during the web crawling phase, thus resulting in no identified traits. The primary limitation of our method lies in its reliance on species and traits that are textually documented online. As a result, the approach is more suited to retrieving morphological traits, which are frequently used in plant species descriptions that can be found on the web. This contrasts with existing trait database initiatives, such as TRY (Kattge et al., 2011), BIEN (Maitner et al., 2018), or TraitBank (Caldwell and Hart, 2014), which contain traits based on measured specimens, without a focus on the morphological traits that are typically used for species description and identification. In addition, changes in taxonomic nomenclature can lead to missing valid information that uses an outdated scientific name. To address this limitation, the procedure could be enhanced by incorporating less stringent filtering during text harvesting and including the use of synonyms. The capacity of the pipeline to capture more text could be further improved by implementing compatibility with JavaScript-based websites and PDF documents (Folk et al., 2023). Finally, while our study focused on categorical traits, we believe the approach could be adapted for other types of trait formulations, such as numerical values, by modifying the LLM prompt and adding a step to deal with different measuring units. We plan to explore this possibility in future work.

Concluding remarks

We developed and evaluated a pipeline that leverages recent advances in LLMs to extract trait information for any set of species from unstructured online text. Unlike other recent approaches that require species–trait information for training (Domazetoski et al., 2023) or manual curation (Folk et al., 2023), ours does not require any manual annotations for training or any curation step. The only manual effort required is the initial creation of the list of traits and the possible trait values, along with the list of species names to be examined; this means that the trait extraction can be effortlessly scaled to new sets of species without the need for previous knowledge on species–trait relations. These results point towards the potential of this type of methodology for leveraging the large amounts of unstructured text data available in online species descriptions. Although in this work we limited the list of traits to those present in the referenced, hand-crafted datasets, we could adapt the approach for use with more general lists (Castellan et al., 2023) to allow scaling up to much larger floras.

AUTHOR CONTRIBUTIONS

D.M., R.V., I.N.A., P.B., A.J., H.G., and W.D.K. conceived the experimental setup; K.P.P., C.L., R.V., and D.M. conducted the experiments; A.S.J.P., P.B., and W.D.K. provided the data; and K.P.P., C.L., and D.M. analyzed the results. All authors reviewed the manuscript and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank Vanessa Hequet (Institut de Recherche pour le Développement [IRD]) for providing expert annotations on trait information presence that were used in this work. The research described in this paper was partially funded by the European Commission via the GUARDEN and MAMBO projects, which have received funding from the European Union's Horizon Europe research and innovation program under grant agreements 101060693 and 101060639, as well as by the Agence Nationale de la Recherche (ANR) project OBTEA (grant number ANR-22-CPJ1-0054-01). This work was supported by the Wageningen University and Research investment theme “Data Driven Discoveries in a Changing Climate.”

This work is dedicated to the memory of Robert van de Vlasakker (23 May 1992 to 27 May 2023), who was the first to initiate this research and whose dedication and passion laid the foundation for this project. As a student, he showed exceptional promise and maturity, and made significant contributions to the early stages of this work. Although Robert sadly passed away before the completion of the manuscript, his enthusiasm, vision, and extraordinary resilience remain an inspiration to all of us. We want to honor his memory with the publication of this piece of research, which would not have been possible without his pioneering efforts.

DATA AVAILABILITY STATEMENT

All the code and data needed to reproduce the results in this paper are available at https://github.com/konpanousis/AutomaticTraitExtraction. The version used for the results in this paper, along with all the associated data, is available at Zenodo (Marcos et al., 2024; https://doi.org/10.5281/zenodo.13969765). In addition to providing the code, we also provide an easy-to-run version, with a working example, in the form of a Python Jupyter Notebook that is directly runnable on Google Colab without the need for specialized hardware.

ORCID

Diego Marcos http://orcid.org/0000-0001-5607-4445
Ioannis N. Athanasiadis http://orcid.org/0000-0003-2764-0078
Pierre Bonnet http://orcid.org/0000-0002-2828-4389
Hervé Goëau http://orcid.org/0000-0003-3296-3795
Alexis Joly http://orcid.org/0000-0002-2161-9940
César Leblanc http://orcid.org/0000-0002-5682-8179
André S. J. van Proosdij http://orcid.org/0000-0003-0084-090X

REFERENCES

Almeida, B. K., M. Garg, M. Kubat, and M. E. Afkhami. 2020. Not that kind of tree: Assessing the potential for decision tree–based plant identification using trait databases. *Applications in Plant Sciences* 8(7): e11379.

Bonnet, P., M. Arbonnier, and P. Grard. 2008. Ligneux du Sahel, Outil graphique d'identification [Woody species of the Sahel: Graphic identification tool] [CD-Rom]. Éditions Quae, Versailles, France.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. Language models are few-shot learners. *In* Advances in Neural Information Processing Systems 33, Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 1877–1901.

Caldwell, I. R., and E. M. Hart. 2014. Using Encyclopedia of Life's Trait-Bank to identify plant traits associated with vulnerability. *PeerJ PrePrints* 2: e491v1. Available at: https://doi.org/10.7287/peerj.preprints.491v1 [posted 8 September 2014; accessed 13 March 2025].

Castellan, S., J. Käfer, and E. Tannier. 2023. Back to the trees: Identifying plants with human intelligence. *In* Proceedings of the Ninth Workshop on Computing within Limits, July 2023, Virtual-Online. Available at: https://doi.org/10.21428/bf6fb269.265c52ce.

Coleman, D., R. V. Gallagher, D. Falster, H. Sauquet, and E. Wenk. 2023. A workflow to create trait databases from collections of textual taxonomic descriptions. *Ecological Informatics* 78: 102312.

Davison, J., J. Feldman, and A. Rush. 2019. Commonsense knowledge mining from pretrained models. *In* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1173–1178. Hong Kong, China.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, pp. 4171–4186. Minneapolis, Minnesota, USA.

Domazetoski, V., H. Kreft, H. Bestova, P. Wieder, R. Koynov, A. Zarei, and P. Weigelt. 2023. Using natural language processing to extract plant functional traits from unstructured text. bioRxiv 2023.11.06.565787 [preprint]. Available at: https://doi.org/10.1101/2023.11.06.565787 [posted 6 November 2023; accessed 13 March 2025].

Endara, L., H. Cui, and J. G. Burleigh. 2018. Extraction of phenotypic traits from taxonomic descriptions for the tree of life using natural language processing. *Applications in Plant Sciences* 6: e1035.

Falster, D., R. Gallagher, E. H. Wenk, I. J. Wright, D. Indiarto, S. C. Andrew, C. Baxter, et al. 2021. AusTraits, a curated plant trait database for the Australian flora. *Scientific Data* 8: e254.

Folk, R. A., R. P. Guralnick, and R. T. LaFrance. 2023. FloraTraiter: Automated parsing of traits from descriptive biodiversity literature. *Applications in Plant Sciences* 12: e11563.

Gallagher, R. V., D. S. Falster, B. S. Maitner, R. Salguero-Gómez, V. Vandvik, W. D. Pearse, F. D. Schneider, et al. 2020. Open science principles for accelerating trait-based science across the tree of life. *Nature Ecology & Evolution* 4: 294–303.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. spaCy: Industrial-strength natural language processing in python. Available at: Zenodo repository: https://doi.org/10.5281/zenodo.1212303 [posted 16 October 2023; accessed 13 March 2025].

Kattge, J., S. Diaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, et al. 2011. TRY–a global database of plant traits. *Global Change Biology* 17: 2905–2935.

Kattge, J., G. Bönisch, S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, S. Tautenhahn, et al. 2020. TRY plant trait database–Enhanced coverage and open access. *Global Change Biology* 26: 119–188.

Kissling, W. D., H. Balslev, W. J. Baker, J. Dransfield, B. Göldel, J. Y. Lim, R. E. Onstein, and J.-C. Svenning. 2019. PalmTraits 1.0, a species-level functional trait database of palms worldwide. *Scientific Data* 6: 178.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. *In* Advances in Neural Information Processing Systems 35, Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 22199–22213.

Kumar, A., V. Dabas, and P. Hooda. 2020. Text classification algorithms for mining unstructured data: A SWOT analysis. *International Journal of Information Technology* 12: 1159–1169.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *In* Advances in Neural Information Processing Systems 33, Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 9459–9474.

Maitner, B. S., B. Boyle, N. Casler, R. Condit, J. Donoghue, S. M. Durán, D. Guaderrama, et al. 2018. The bien r package: A tool to access the botanical information and ecology network (BIEN) database. *Methods in Ecology and Evolution* 9: 373–379.

Marcos, D., A. Potze, W. Xu, D. Tuia, and Z. Akata. 2022. Attribute prediction as multiple instance learning. *Transactions on Machine Learning Research* Available at: https://openreview.net/forum?id=nmFczdJtc2

Marcos, D., R. van de Vlasakker, I. N. Athanasiadis, P. Bonnet, H. Goeau, A. Joly, W. D. Kissling, et al. 2024. Fully automatic extraction of morphological traits from the Web: utopia or reality? Available at Zenodo repository: https://doi.org/10.5281/zenodo.13969766 [posted 4 April 2024; accessed 13 March 2025].

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, et al. 2022. Training language models to follow instructions with human feedback. *In* Advances in Neural Information Processing Systems 35, Proceedings of the Annual Conference on Neural Information Processing Systems, pp. 27730–27744.

Petroni, F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. 2019. Language models as knowledge bases? *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2463–2473.

Reed, S. E., H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A Rabinovich. 2015. Training deep neural networks on noisy labels with bootstrapping. arXiv 1412.6596 [Preprint]. Available at: https://doi.org/10.48550/arXiv.1412.6596 [posted 20 December 2014; accessed 13 March 2025].

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 1910.01108 [Preprint]. Available at: https://doi.org/10.48550/arXiv.1910.01108 [posted 2 October 2019; accessed 13 March 2025].

Schneider, F. D., D. Fichtmueller, M. M. Gossner, A. Güntsch, M. Jochum, B. König-Ries, G. Le Provost, et al. 2019. Towards an ecological trait-data standard. *Methods in Ecology and Evolution* 10: 2006–2019.

Wei, X., X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, et al. 2023. ChatIE: Zero-shot information extraction via chatting with ChatGPT. arXiv 2302.10205 [Preprint]. Available at: https://doi.org/10.48550/arXiv.2302.10205 [posted 20 February 2023; accessed 13 March 2025].

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, et al. 2020. HuggingFace's transformers: State-of-the-art natural language processing. arXiv 02771 [Preprint]. Available at: https://doi.org/10.48550/arXiv.1910.03771 [posted 9 October 2019; accessed 13 March 2025].

Zhang, X., K. Zhou, S. Wang, F. Zhang, Z. Wang, and J. Liu. 2020. Learn with noisy data via unsupervised loss correction for weakly supervised reading comprehension. *In* Proceedings of the 28th International Conference on Computational Linguistics, pp. 2624–2634. Barcelona, Spain.

Zhang, Y., Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, et al. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv 01219 [Preprint]. Available at: https://doi.org/10.48550/arXiv.2309.01219 [posted 3 September 2023; accessed 13 March 2025].

How to cite this article: Diego, M., R. van de Vlasakker, I. N. Athanasiadis, P. Bonnet, H. Goëau, A. Joly, W. D. Kissling, C. Leblanc, A. S. J. van Proosdij, and K. P. Panousis. 2025. Fully automatic extraction of morphological traits from the web: Utopia or reality? *Applications in Plant Sciences* 13(2): e70005. https://doi.org/10.1002/aps3.70005