

Using ChatGPT with Confidence for Biodiversity-Related Information Tasks

Michael J. Elliott
José A.B. Fortes

Have you used ChatGPT?*

*Or other LLMs like Google's Bard, Microsoft's Bing Chat, or offline models



1 – Never!

2 – Yes!

3 – Yes, *and the API!* (**bonus points**)

Click me!



What plant species can I find in Gainesville, Florida?



(GPT-3.5)



Gainesville, Florida, is located in North Central Florida and features a diverse range of plant species typical of the region's subtropical climate. The specific plants you'll find in Gainesville can vary depending on the time of year and local environmental conditions, but here are some common plant species you might encounter in the area:

1. **Live Oak (*Quercus virginiana*):** These iconic Southern oak trees are known for their sprawling canopies of evergreen leaves.
2. **Spanish Moss (*Tillandsia usneoides*):** This epiphytic plant hangs from trees and adds a distinctive character to the landscape.
3. **Southern Magnolia (*Magnolia grandiflora*):** Known for its large, fragrant white flowers and glossy leaves, the Southern Magnolia is a popular ornamental tree in the region.
4. **Crape Myrtle (*Lagerstroemia indica*):** These small trees or shrubs are famous for their



Reformat the list in TSV using DarwinCore terms



(GPT-3.5)



```
scientificName  vernacularName  location
Quercus virginiana  Live Oak      Gainesville, Florida
Tillandsia usneoides  Spanish Moss  Gainesville, Florida
Magnolia grandiflora  Southern Magnolia  Gainesville, Florida
Lagerstroemia indica  Crape Myrtle  Gainesville, Florida
Sabal palmetto  Palmetto      Gainesville, Florida
Pinus palustris  Longleaf Pine  Gainesville, Florida
Serenoa repens  Saw Palmetto  Gainesville, Florida
Lantana camara  Lantana  Gainesville, Florida
Pinus elliottii  Slash Pine  Gainesville, Florida
Pinus taeda  Loblolly Pine  Gainesville, Florida
```

Can we mine Large Language Models (LLMs) like ChatGPT for biodiversity data?

A potential use case for LLMs:

- A ton of biodiversity knowledge is buried in text documents
- Not all of it is captured in openly available, curated datasets
- Can LLMs help us fill this gap?

The problem:

- LLMs sometimes make up information – “*hallucination*”
- How do we know when to trust them?

Our proposed solution:

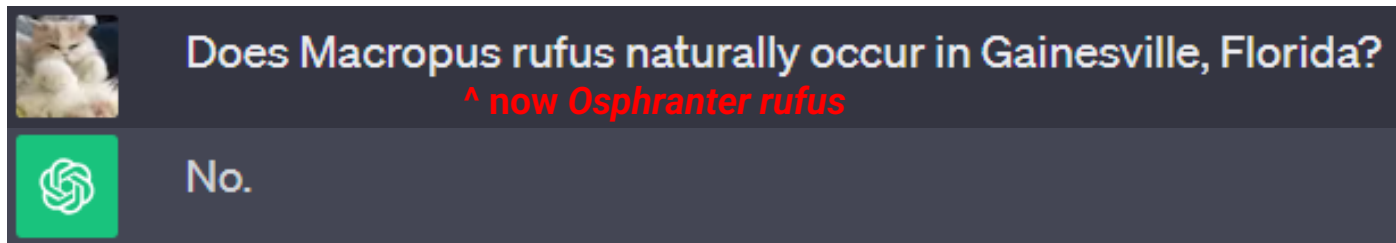
- First evaluate an LLM’s performance on a test set
- Then train a *confidence model* to detect/discard its mistakes

Application: use ChatGPT to predict species occurrences *with high confidence*

Given a species and location, we want ChatGPT to predict whether the species is present or absent at the location:

[SPECIES] is [PRESENT or ABSENT] at [LOCATION]

This can be phrased as a yes-or-no question:



Questions can be submitted in bulk to ChatGPT using their web API



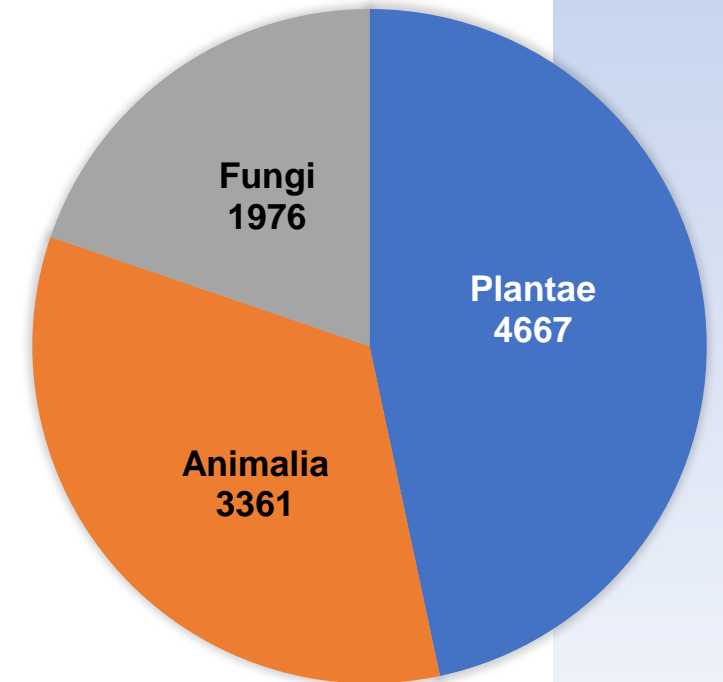
Step 1: compile a labeled test dataset

Collected species occurrence records using the iDigBio API :

- 10,004 plantae, animalia, and fungi records
- Represent all phyla available through iDigBio

iDigBio's records are largely *presence only*, so we made a synthetic dataset of pseudo-absences:

- Randomized the 10,004 records' locations
- Removed all species-location pairs that matched occurrence records in iDigBio (8.26%)
- 9,178 pseudo-absences



Test dataset →

<https://doi.org/10.5281/zenodo.8417791>

hash://sha256/fd1c8a5304508981a02481a31dc88ad429356a26dff6f89d24fe47fc60a275a4

Step 2: submit test questions to ChatGPT and grade answers

Construct natural language yes-or-no questions from Darwin Core records:

```
{  
  "dwc:scientificName": "Alitta plenidentata",  
  "dwc:country": "United States",  
  "dwc:stateProvince": "California",  
  "dwc:county": "San Diego County"  
}
```



“Does Alitta plenidentata naturally occur in San Diego County, California, United States? Yes or no.”



ChatGPT

And submit to ChatGPT

- “Yes” is correct for presence records
- “No” is correct for pseudo-absence records
- Anything else: “I don’t know”

Step 3: use graded responses to train a confidence model

A confidence model assigns confidence scores to each response

- **Confidence** := a **probability** that the response is correct

But what does the model use as input?

For each question, we have:

- The scientific name and location in the question
- The model's "yes" / "no" / "I don't know" response
- **Not enough info!**
 - We need more **uncertainty information**

Step 3: use graded responses **and uncertainty information** to train a confidence model

Method 1: repeat each question **10** times:

Q1: yes yes yes yes yes yes yes yes ...

Q2: no yes no yes no no IDK IDK ...

High uncertainty

Method 2: ask each question several different ways:

- *Can [SPECIES] be found in [LOCATION]? → “yes”*
- *Is it possible to encounter [SPECIES] in [LOCATION]? → “yes”*
- *Is there a presence of [SPECIES] within [LOCATION]? → “no”*

Method 3: accuracy on related questions about the species

Method 4: number of “I don’t know” responses to related questions

Method 5: historical accuracy by taxon, location

Method 6: number of records in iDigBio by taxon, location

...

Anything that could be correlated with the LLM’s performance

Step 3: use graded responses and uncertainty information to **train a confidence model**

Our implementation*:

- **20 uncertainty measures** as model inputs
 - Based on **70** ChatGPT responses per question, plus information from iDigBio & test set performance
- **XGBoost + isotonic regression** to build the confidence model
 - With a constraint: increasing uncertainty **does not** increase confidence
- Used 50% of questions for training, 50% for testing

*Full implementation details at <https://github.com/acislab/biodiversity-llms> →



Confidence model results

overall accuracy: 50%

Correct?	Confidence	Question	ChatGPT
✗	.56	<i>Lycoperdon subvelatum</i> in Armstrong, Pennsylvania, USA?	Yes
✗	.71	<i>Trametes ectypa</i> in Marin, California, USA?	Yes
✗	.36	<i>Diplophyllum albicans</i> in Lord Howe Island, NSW, Australia?	Yes
✓	.72	<i>Phaeophyscia squarrosa</i> in Jackson, Alabama, USA?	Yes
✓	.52	<i>Frullania plana</i> in Avery, North Carolina, USA?	Yes
✓	.47	<i>Nitella opaca</i> in Charleston, South Carolina, USA?	Yes
✗	.36	<i>Syzygites megalocarpus</i> in Westmoreland, Penn., USA?	No
✓	.89	<i>Notropis texanus</i> in Kongsberg, Viken, Norway?	No
✓	.57	<i>Sphaerosyllis taylori</i> in Skamania, Washington, USA?	No
✗	.56	<i>Lithophyllum corallinae</i> in Rockingham, New Hamp., USA?	No
✓	.79	<i>Pteropurpura festiva</i> in Sør-Aurdal, Oppland, Norway?	No
✗	.52	<i>Puberella crenata</i> in Sumter, South Carolina, USA?	No

Confidence model results

confidence $\geq .5 \rightarrow 63\%$

Correct?	Confidence	Question	ChatGPT
✗	.56	<i>Lycoperdon subvelatum</i> in Armstrong, Pennsylvania, USA?	Yes
✗	.71	<i>Trametes ectypa</i> in Marin, California, USA?	Yes
	.36	<i>Diplophyllum albicans</i> in Lord Howe Island, NSW, Australia?	Yes
☑	.72	<i>Phaeophyscia squarrosa</i> in Jackson, Alabama, USA?	Yes
☑	.52	<i>Frullania plana</i> in Avery, North Carolina, USA?	Yes
	.47	<i>Nitella opaca</i> in Charleston, South Carolina, USA?	Yes
	.36	<i>Syzygites megalocarpus</i> in Westmoreland, Penn., USA?	No
☑	.89	<i>Notropis texanus</i> in Kongsberg, Viken, Norway?	No
☑	.57	<i>Sphaerosyllis taylori</i> in Skamania, Washington, USA?	No
✗	.56	<i>Lithophyllum corallinae</i> in Rockingham, New Hamp., USA?	No
☑	.79	<i>Pteropurpura festiva</i> in Sør-Aurdal, Oppland, Norway?	No
✗	.52	<i>Puberella crenata</i> in Sumter, South Carolina, USA?	No

Confidence model results

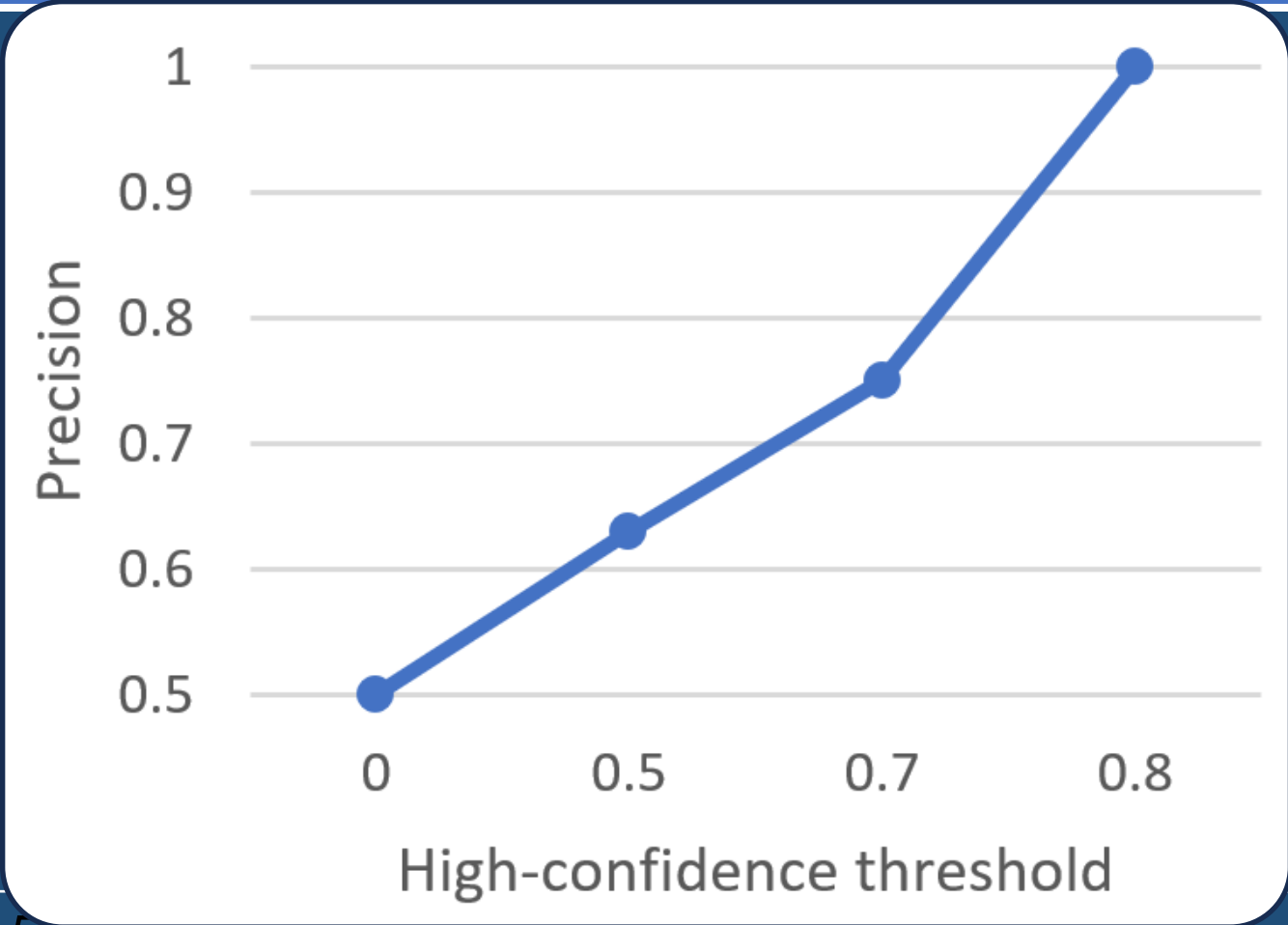
confidence $\geq .7 \rightarrow 75\%$

Correct?	Confidence	Question	ChatGPT
	.56	<i>Lycoperdon subvelatum</i> in Armstrong, Pennsylvania, USA?	Yes
✘	.71	<i>Trametes ectypa</i> in Marin, California, USA?	Yes
	.36	<i>Diplophyllum albicans</i> in Lord Howe Island, NSW, Australia?	Yes
☑	.72	<i>Phaeophyscia squarrosa</i> in Jackson, Alabama, USA?	Yes
	.52	<i>Frullania plana</i> in Avery, North Carolina, USA?	Yes
	.47	<i>Nitella opaca</i> in Charleston, South Carolina, USA?	Yes
	.36	<i>Syzygites megalocarpus</i> in Westmoreland, Penn., USA?	No
☑	.89	<i>Notropis texanus</i> in Kongsberg, Viken, Norway?	No
	.57	<i>Sphaerosyllis taylori</i> in Skamania, Washington, USA?	No
	.56	<i>Lithophyllum corallinae</i> in Rockingham, New Hamp., USA?	No
☑	.79	<i>Pteropurpura festiva</i> in Sør-Aurdal, Oppland, Norway?	No
	.52	<i>Puberella crenata</i> in Sumter, South Carolina, USA?	No

Confidence model results

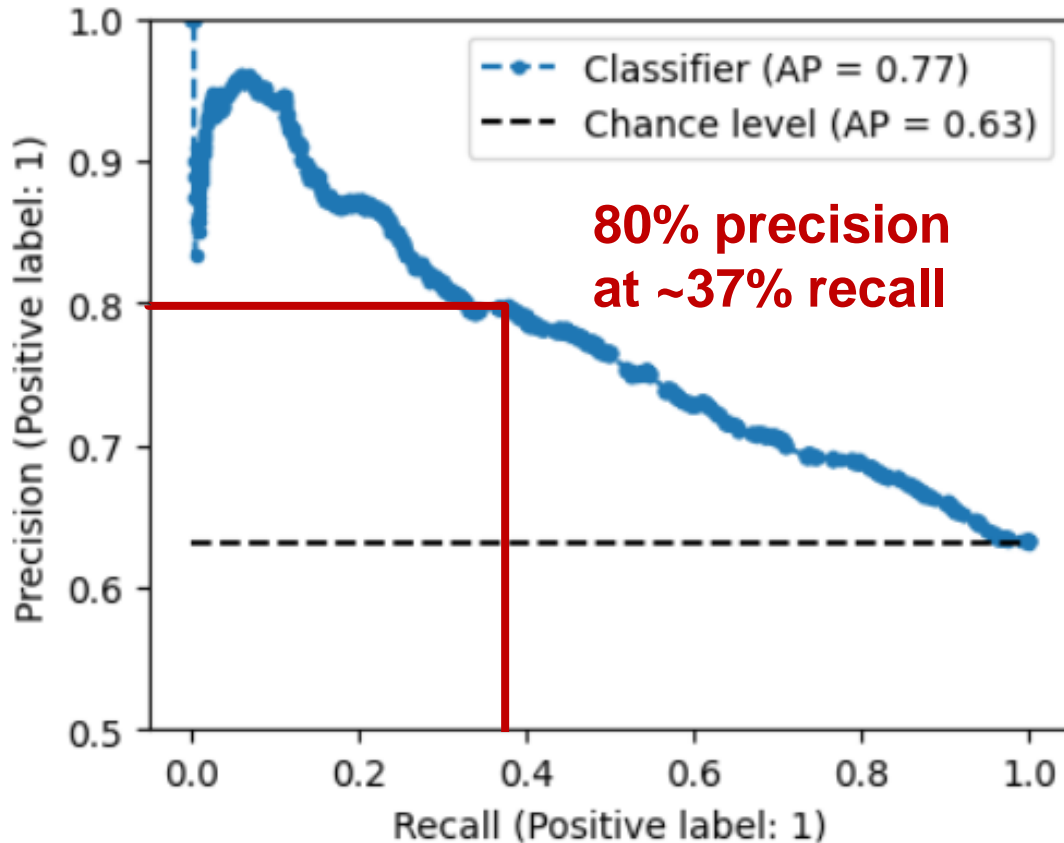
confidence $\geq .8 \rightarrow 100\%$

Correct?	Confidence	Question	ChatGPT
		...A?	Yes
			Yes
		...ia?	Yes
			Yes
			Yes
			Yes
			No
<input checked="" type="checkbox"/>			No
			No
		...?	No
			No
			No
.52		<i>Trubereia crenata</i> in Sumter, South Carolina, USA?	No

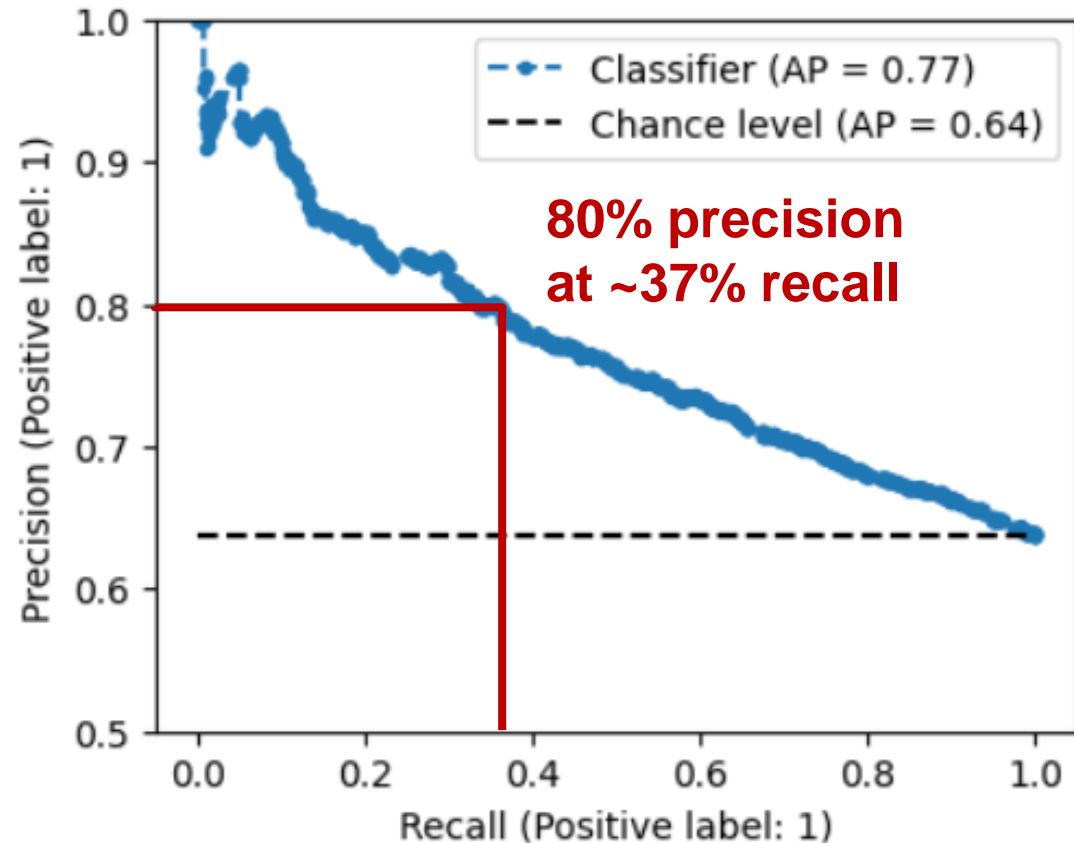


Confidence model results

Animalia



Plantae

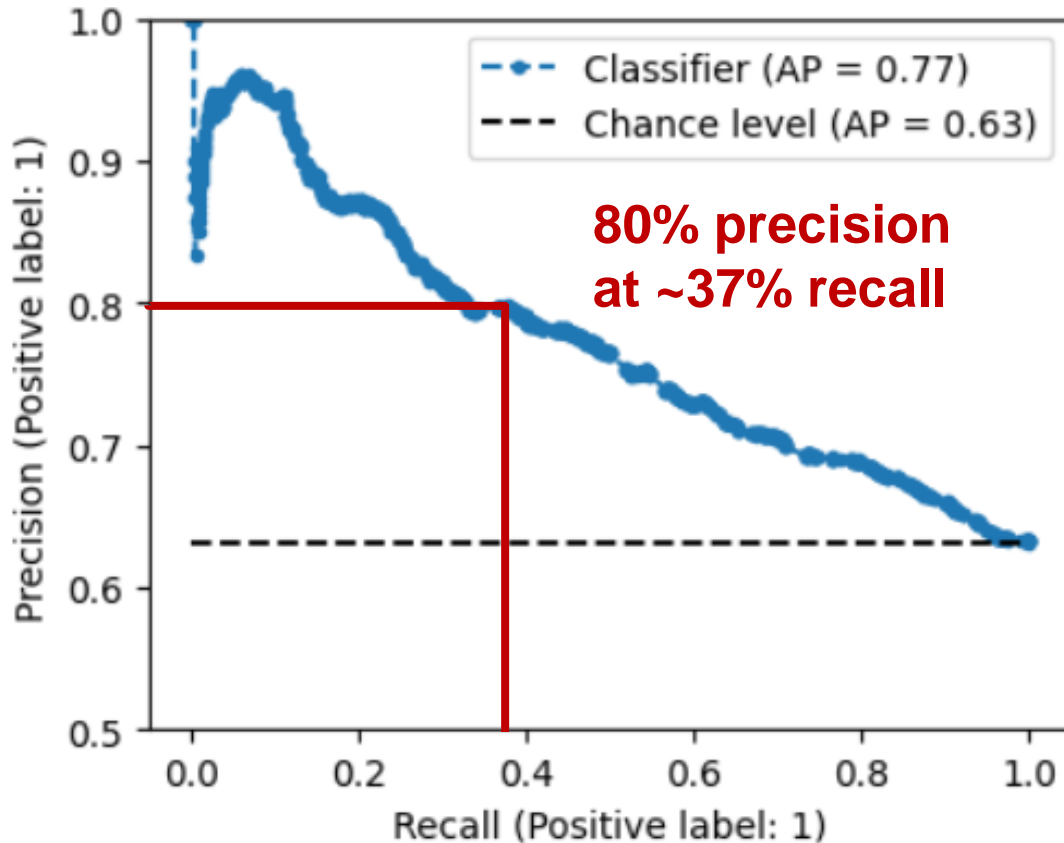


Each point represents a **confidence threshold** for **high confidence**

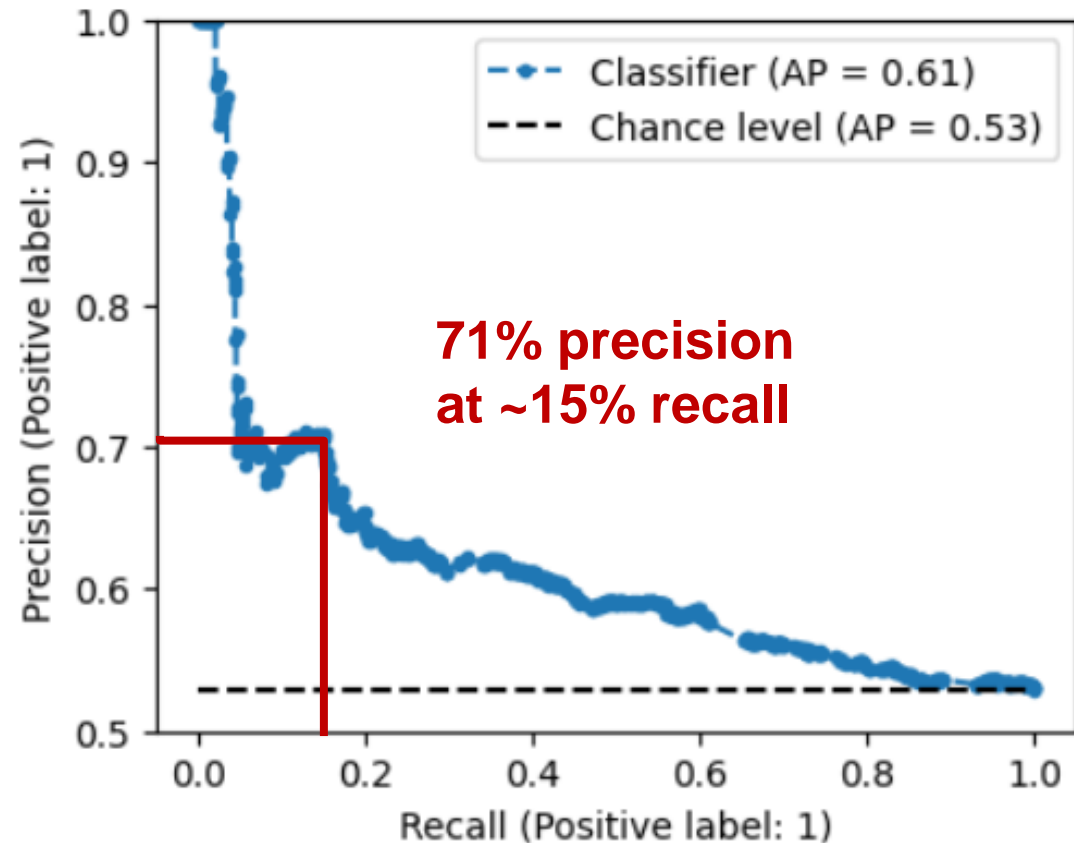
- **Precision:** ChatGPT's accuracy for responses with high confidence
- **Recall:** Percentage of correct responses that are assigned high confidence

Confidence model results

Animalia



Fungi

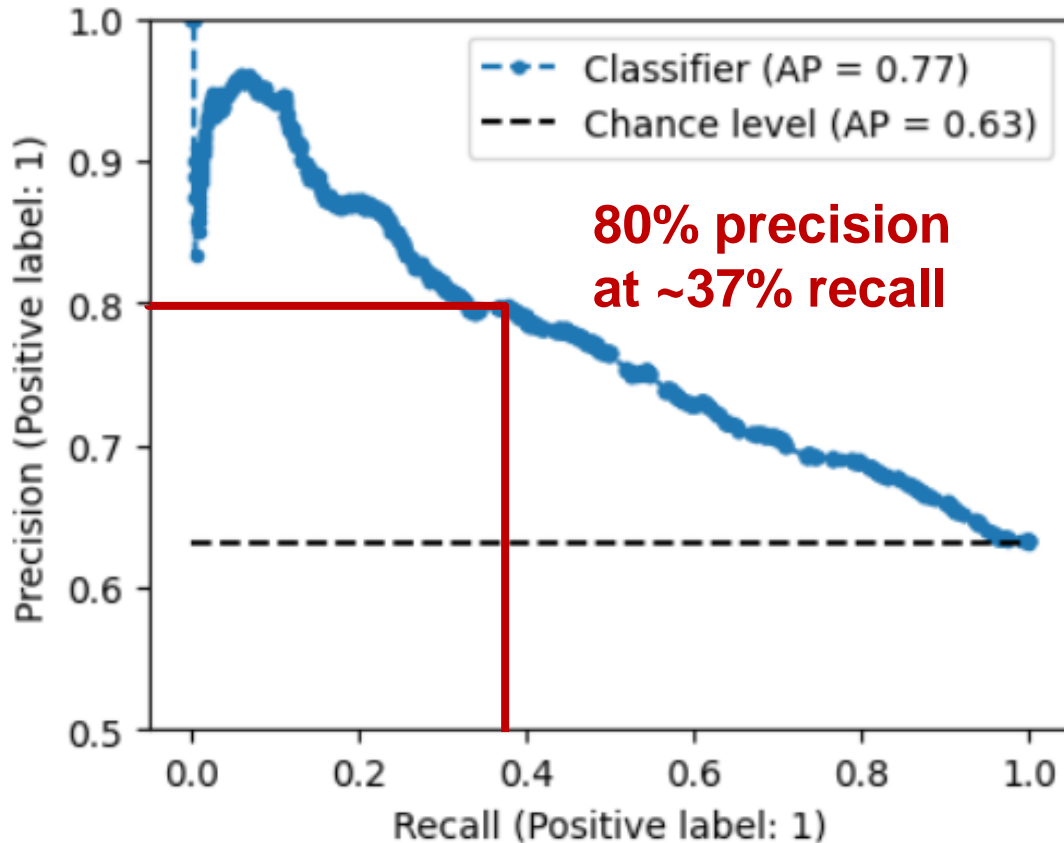


Each point represents a **confidence threshold** for **high confidence**

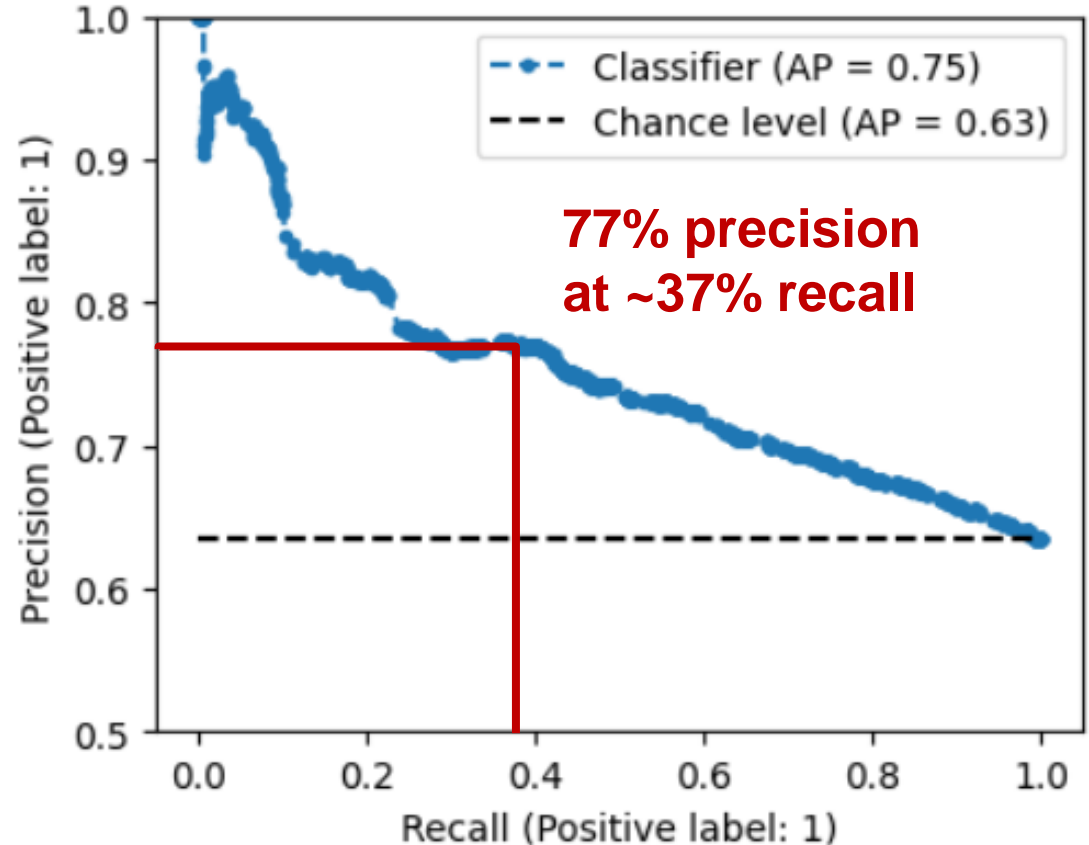
- **Precision:** ChatGPT's accuracy for responses with high confidence
- **Recall:** Percentage of correct responses that are assigned high confidence

Confidence model results

Animalia (Trained on all)



Animalia (Trained on Plantae)



Each point represents a **confidence threshold** for **high confidence**

- **Precision:** ChatGPT's accuracy for responses with high confidence
- **Recall:** Percentage of correct responses that are assigned high confidence

Conclusions

- LLMs know a *lot*... but sometimes they make mistakes!
 - Using them as info sources often demands high reliability
- LLMs can be used with a desired reliability (i.e., *precision*) by discarding low-confidence responses (i.e., at the cost of *recall*)
- Simple confidence models (*100s of parameters*) can be effective at detecting mistakes* made by extremely complex blackbox LLMs (*100-billions of parameters, even trillions!*)

*“Mistakes”, according to records in iDigBio

- More uncertainty information → better confidence estimates

Thanks for listening!

Talk to me: mielliott@ufl.edu



Conversation with ChatGPT

<https://chat.openai.com/share/df78af75-16ae-4594-9e03-848a98737d03>

Source code for data collection, uncertainty measures, and confidence models

<https://github.com/acislab/biodiversity-llms>



Test datasets and ChatGPT's responses

<https://doi.org/10.5281/zenodo.8417791>

hash://sha256/fd1c8a5304508981a02481a31dc88ad429356a26dff6f89d24fe47fc60a275a4

The research reported in this presentation was funded in part by grants from the National Science Foundation (DBI 2027654) and the AT&T Foundation.