Synoptic Key of Life: Feature Extraction for Fungal Taxonomy

La Monte Henry Piggy Yarroll
Syracuse University Student
Pittsburgh, PA, USA
piggy@acm.org

Padmaja Kurumaddali
InformationTechnology Services,
Syracuse University
Syracuse, NY, USA
pkurumad@syr.edu

Patrick Le
Syracuse University Student
Los Angeles, CA, USA
patrickle1214@gmail.com

***Abstract*—This paper contributes to a machine learning-based system for automating the classification and semantic embedding of taxonomic content in mycological journal articles. The goal is to streamline fungal species identification by transforming unstructured scientific text into structured data suitable for building a user interface for a synoptic key.**

***Keywords*—Prompt engineering, LLM, feature extraction, mycology, taxonomy.**

I. INTRODUCTION

The classification and identification of fungal species is a labor-intensive process, often requiring manual review of extensive taxonomic literature. This challenge is compounded by the unstructured nature of scientific texts, which intermix nomenclature, morphological descriptions, and general exposition. As the volume of digitized biological literature grows, scalable methods for extracting and organizing taxonomic knowledge have become increasingly important.

The Synoptic Key of Life Project (SKOL), aims to automate this process by extracting species descriptions from taxonomic journals. The project starts with some readily available mycological literature, as the first author is familiar with mycology. A synoptic key is a flexible identification tool that allows users to classify organisms based on any observable characteristic rather than following a rigid, ordered decision tree, as in traditional binary keys.

At the core of this system is the extraction of features and diagnostic features-observable traits that are useful for distinguishing species. The extracted features are suitable for populating menus in a tool to aid with generating taxonomic descriptions. Such a description can then be embedded in a search space with the work described in Yarroll2024. By combining natural language processing, machine learning, and distributed computing, SKOL supports rapid consistent species identification and enhances research productivity in taxonomy. It opens the taxonomic literature to the enthusiastic amateur.

II. ABOUT THE DATA

A. *Annotated Mycology Journals*

The dataset used in this study comprises digitized biological literature from three prominent mycological journals: *Mycologia*, Mycotaxon, and *Persoonia*. The text is OCR of page scans from Wood1995. These journals span several decades and contain a rich mix of taxonomic content, including species names, morphological descriptions, and general scientific exposition. The data is organized into two primary corpora: a labeled corpus and an unlabeled corpus.

The labeled corpus consists of 190 journal issues that were manually annotated using the YEDDA annotation tool. This corpus includes 60,754 paragraphs, categorized into three classes:

- Nomenclature (6,072 paragraphs): containing formal species names and taxonomic identifiers.
- Description (6,192 paragraphs): detailing morphological and ecological traits of fungal species.
- Miscellaneous Exposition (48,564 paragraphs): encompassing introductions, discussions, and other non-taxonomic content.

Corpus Type	**Issues**	**Paragraphs**	**Lines**	**Words**
Labeled Corpus[1]	190	60754	~300,000	~2.1 million
Unlabeled Corpus	1021	N/A	~4,300,000	~25 million

Fig. 1. Summary of Corpora

Both corpora are extracted from page scans with Optical Character Recognition (OCR).

In total, the labeled corpus contains over 2.1 million words across more than 300,000 lines of text, making it a valuable resource for training and evaluating classification models.

The unlabeled corpus is significantly larger, comprising 1,021 journal issues. This corpus includes approximately 4.3 million lines and over 25 million words.

Until 2011, every new species publication had to include a Latin description. Almost all of the source data include both Latin and English descriptions.

Heuristic rules were applied to segment paragraphs based on line breaks, indentation, and formatting cues. Additional preprocessing steps included merging fragmented text, removing empty lines, and applying TF-IDF weighting to emphasize taxonomically relevant terms. Suffix-based features (e.g., -aceae, -spore, -mycetes) were also extracted to enhance classification accuracy, particularly for distinguishing between nomenclature and descriptive content.

---

[1] Manually annotated using YEDDA; includes Nomenclature, Description, and Miscellaneous categories

©2025 La Monte HP Yarroll, Padmaja Kurumaddali, Patrick Le

The large annotated corpus and the even larger raw corpus provide a strong foundation for training scalable machine learning models and support the broader goal of automating synoptic key generation for fungal species.

B. *New Training Data*

For the current paper we produced 16 hand-built descriptions with JSON structures to use in fine tuning the model. We used 14 for fine tuning, 1 for validation, and 1 for testing.

III. Previous Work

A. *Origin of the Synoptic Key of Life Project*

The first phase of the Synoptic Key of Life Project focused on annotating some of the mycological literature. Using a combination of manual annotation and machine learning, the first author created a structured representation of taxonomic content. Central to this effort was the development of the Taxon python class, which encapsulates nomenclature, descriptive paragraphs, and metadata for each species. These structured objects standardize the content and prepare it for downstream processing, including semantic embedding and search.

B. *MycoSearch*

In IST664, a team extended Dr. Draft's SOTA Literature Search (Gisolfi2024) to produce Dr. Draft's MycoSearch. The Dr. Draft's system embeds descriptions from Taxon objects into a high-dimensional semantic space with SBERT (Sentence-BERT, Reimers2019). This enables users to input technical specimen descriptions and retrieve the most relevant species matches based on cosine similarity. The integration of SKOL with Dr. Draft's provides a powerful, domain-specific search interface tailored to mycological literature. The tool is usable from a CLI or from a python API.

C. *PySpark processing pipeline*

In IST718 another team built on the annotated dataset by introducing a distributed PySpark pipeline to scale the classification process. Over 1,000 unannotated journal issues were ingested and processed. The pipeline classifies paragraphs into three categories: Nomenclature, Description, and Miscellaneous Exposition. Feature engineering techniques included TF-IDF weighting and suffix-based features (e.g., -aceae, -spore), which improved classification accuracy. The best-performing model-Logistic Regression with combined TF-IDF and suffix features-achieved over 94% accuracy, laying the foundation for automated feature extraction.

IV. Models

A. *Exploratory models*

The proof of concept for the current project was a single query to ChatGPT 4.0 using the prompt in Fig. 2. This produced the JSON in Fig. 3. This was sufficient to suggest that an LLM approach could be made to work.

> Can you extract any attributes and their measurements from the following description of an organism? Present the result as a json data structure.
>
> Mycelium on the substrate is medium orange-brown, septate, 3-4 μm diam., often distinctly and coarsely rough; hyphae often run parallel to the long axis of cells in substrate tissue but usually branch on the exterior and appear as a non-patterned network. Sporulation units are produced as enlarging tips of short branches, at first concolorous with surface mycelium but soon dark orange-brown and finally black-brown and opaque. In type material an occasional enlarging branch can be seen attached to the hypha from which it developed; in most instances, however, the enlarged branch elements have broken loose from the mycelium or are so densely crowded on the substrate and so nearly opaque that points of attachment cannot be traced

Fig. 2. Prompt used to evaluate LLMs and SLMs.

```
{
 "mycelium": {
  "location": "on substrate",
  "color": "medium orange-brown"
 },
 "hyphae": {
  "septate": true,
  "diameter_um": {
   "min": 3,
   "max": 4
  },
  "surface": "distinctly and coarsely rough",
  "orientation": "parallel to long axis of substrate tissue cells",
  "branching": {
   "location": "on exterior",
   "pattern": "non-patterned network"
  }
 },
 "sporulation_units": {
  "origin": "enlarging tips of short branches",
  "color_progression": [
   "concolorous with surface mycelium",
   "dark orange-brown",
   "black-brown and opaque"
  ],
  "attachment": {
   "occasional": "attached to hypha",
   "typical": "broken loose or obscured by crowding and opacity"
  }
 }
}
```

Fig. 3. Initial response from ChatGPT 4.0.

To scale this approach across a large corpus of organism descriptions, we evaluated several alternative models. ChatGPT 4.0, while accurate, was deemed too slow and cost-prohibitive for high-throughput processing. Initial attempts to deploy the llama3.3_70b model were unsuccessful due to hardware limitations, and the model's lack of Latin language support further constrained its utility.

Subsequent trials with gemma3_27b yielded promising results in terms of output quality; however, performance was hindered by excessive latency and eventual out-of-memory failures on our local GPU infrastructure. We then transitioned to gemma3_12b, which provided acceptable performance and output quality. This model was used extensively until we began integrating with the Hugging Face (Ait2023) API, with the goal of enabling transfer learning. At this stage, we encountered compatibility issues, as several components of the Hugging Face API were not fully supported for gemma3_12b.

The example code provided in the Hugging Face documentation was tailored for mistral_7b, which we found to perform well within the ollama infrastructure. However, when deployed via the Hugging Face API, mistral_7b produced significantly degraded results. Further experimentation revealed that mistral_7b_instruct_0.3 delivered superior performance, even surpassing the earlier results obtained with mistral_7b under ollama.

Given the availability of a high-quality fine-tuning example for mistral_7b_instruct (Adithya2023), we opted to pursue fine-tuning this model rather than implementing a transfer learning pipeline.

B. *Primary Results*

We used the mistral_7b model with the ollama front end to do prompt engineering. We revised our initial prompt incrementally until we had something that more consistently produced what we wanted. The final prompt is in Fig. 4.

> Please extract features, subfeatures, optional subsubfeatures, and values from the following species description.
> Format the output as JSON.
> The top level of the JSON is feature names. The next level in is subfeature names . The optional next level in is subsubfeature names.
> The innermost layer is lists of string-valued values.
> Lists are only present at the innermost level of the JSON.
> Feature values that are comma-separated strings should be broken down into separate values.
> Translate Latin paragraphs to English.

Fig. 4. Final prompt.

Fine tuning with our very small training set shows clear signs of overfitting. The training loss, Fig. 5, drops to near 0, showing that the training samples are very well learned. The evaluation loss, Fig. 6, after a brief drop shoots up to about 1.3 and continues up.

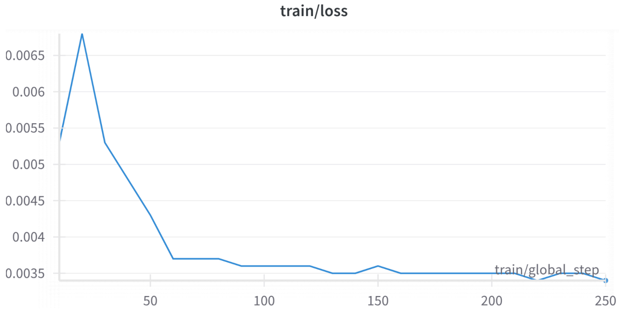

Fig. 5. Training loss while fine tuning.

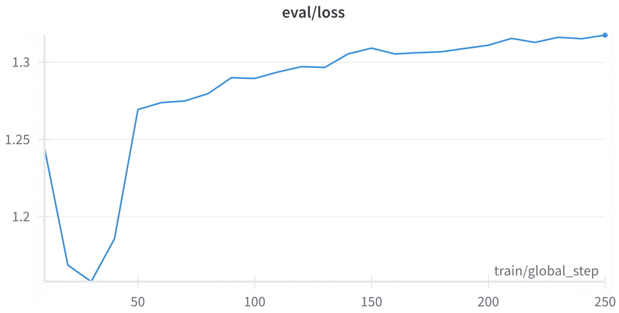

Fig. 6. Evaluation loss while fine tuning

Results for our test description are in Fig. 7, edited for brevity. This includes the base model with our engineered prompt, the same prompt to the fine-tuned model, and the hand-built ground truth.

The base model and the fine-tuned model both have lists at higher levels than at the leaves, contrary to the instructions. This is not a major issue, as it is fairly easy to treat the keys in the lists of dictionaries as if they were keys in the parent dictionary, concatenating values for any conflicting keys.

The base model adds a superfluous top level key of “speciesdescription”. Again, this is annoying but not terribly difficult to code around.

The fine-tuned model has failed to translate the Latin description, and is reporting Latin and English results as if they were separate features. Closer investigation shows that it appended the English translation to the end of the output, after the JSON. Perhaps the prompt needs to specify the order of operations.

The fine-tuned model failed to note that there was no teleomorph observed.

Base model

```
{
  "speciesdescription": [
    {
      "leafspots": [
        {
          "location":
"amphigenous",
          "shape": [
            "circular",
            "angular",
            "irregular"
          ],
          "pattern": [
            "dark brown",
            "pale brown"
          ],
          "size": [
            "1-3 cm diam."
          ]
        }
      ],
…
      "conidia": [
        {
          "formation":
"solitary",
          "type":
"holoblastic",
          "dryingstate":
"dry",
          "color": [
            "dark brown to
blackish brown",
            "pale brown
when immature",
            "medium to
```

Fine-tuned model

```
{
…
  "conidia": {
    "location": [
      "solitaria"
    ],
    "condition": [
      "sicca"
    ],
    "color": [
      "atro-brunnea in
multitudine"
    ],
    "shape": [
      "laevia"
    ],
    "surface": [
      "raro verruculosa"
    ],
    "wall": [
      "crassitunicata"
    ],
    "septation": [
      "2-7 septis crassis"
    ],
    "size": [
      "25-57.5 µm x
5.5-8 µm lata"
    ]
  },
  "leaf spots": {
    "condition": [
      "necrotic"
    ],
    "location": [
      "amphigenous"
```

ground truth

```
{
  "stoma": {
    "development": [
      "well developed"
    ],
    "shape": [
      "globose",
      "subglobose",
      "irregular"
    ],
    "color": [
      "dark brown",
      "blackish brown"
    ],
    "size": [
      "50-72 µm diam."
    ],
    "formation": [
      "elongated
cylindrical cells"
    ]
  },
…
  "conidia": {
    "habit": [
      "solitary",
      "dry"
    ],
    "color": [
      "dark brown"
    ],
    "size": [
      "25-57.5 × 5.5-8 µm"
    ],
    "shape": [
      "laevia"
dark brown"
                ],
                "shape": [
                   "straight",
                   "slightly
curved"
                ],
                "smoothness":
"mostly smooth",
                "verruculosity":
"rarely minutely verruculose",
                "wallthickness":
"thick-walled",
                "septations":
"2-7-septate",

"wallthicknessandseptations": [
                   "thicker and
darker in the basal part",
                   "pale and thin
towards apex"
                ],
                "base": "truncate",
                "apex": [
                   "obtuse",
                   "subacute"
                ],
                "basesize": [
                   "2.0-2.2 µm"
                ],
                "apexsize": [
                   "2.5-3.5 µm"
                ]
             }
          ],
          "teleomorph": [
             "not observed"
          ]
       },
       …
    ]
}
```

```
      ],
      "shape": [
         "circular",
         "angular",
         "irregular"
      ],
      "pattern": [
         "forming
concentric dark and pale
brown patterns on the
spots"
      ],
      "spread": [
         "spreading
eventually covering most
of the leaf, but not vein
limited"
      ],
      "size": [
         "1-3 cm diam."
      ]
   },
 …
   }
}
```

```
      ],
      "rarity": [
         "raro verruculosa"
      ],
      "wall characteristics": [
         "crassitunicata"
      ],
      "septation": [
         "2-7 thickly septate"
      ],
      "base": [
         "truncate"
      ],
      "apex": [
         "obtuse or subacute"
      ]
   },
   "leaf spot": {
      "nature": [
         "necrotic"
      ],
      "location": [
         "amphigenous"
      ],
      "shape": [
         "circular",
         "angular",
         "irregular"
      ],
      "pattern": [
         "forming concentric
dark and pale brown
patterns"
      ],
      "spreading": [
         "eventually covering
most of the leaf"
      ],
      "size": [
         "1-3 cm diam."
      ],
      "limitation": [
         "not vein limited"
      ]
   },
…
   "teleomorph": [
      "not observed"
   ]
}
```

Fig. 7. Final prompt result, Fine-tuned result, ground truth

To measure the effectiveness of the JSON objects created through the large language models, the features and the values were extracted separately as sets. Each feature/value set is compared with a handwritten JSON and individually tested through jaccard distance. Jaccard distance is calculated using the following formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The feature score and value score calculated through Jaccard distance is then averaged to get the final score. The final score for the test description in our base model is 0.762 and the final score for our fine-tuned model is 0.799. Consistent with the loss metrics, the fine-tuned model is performing worse than the base model. The base model Jaccard distance is pretty poor, so there is still incentive to build a larger training set for fine-tuning.

V. FUTURE WORK

The biggest issue with our model is that it has fallen to heavy overfitting. Because it was so tedious to hand create more JSON results for the descriptions and limited time for this project, our training data size was too small. Therefore to avoid overfitting, we would create more training samples to create better results for our model.

The next natural step in the SKOL is to take the JSON structures produced by the current work and build a system of dynamically generated menus.

We need a mechanism for picking the top features to look at, given the current state of the description. Something like a TF/IDF measure for the (feature, subfeature, value) tuples could be helpful.

There are productionization tasks to attend to. Web crawlers are needed to pick up new articles from open access publications. There is a fair amount of older non-journal mycological literature which we could also mine. Most of those will require OCR in the pipeline.

Fig. 8 Shows how the current work fits into the current and future architecture. The JSON representations of features, subfeatures, and values provide the material to build menus in the Description Builder. This component helps the user build a technical description. This description is sent to Mycosearch where we find the n most-similar descriptions from the literature.

The service will be deployed at https://synoptickeyof.life.

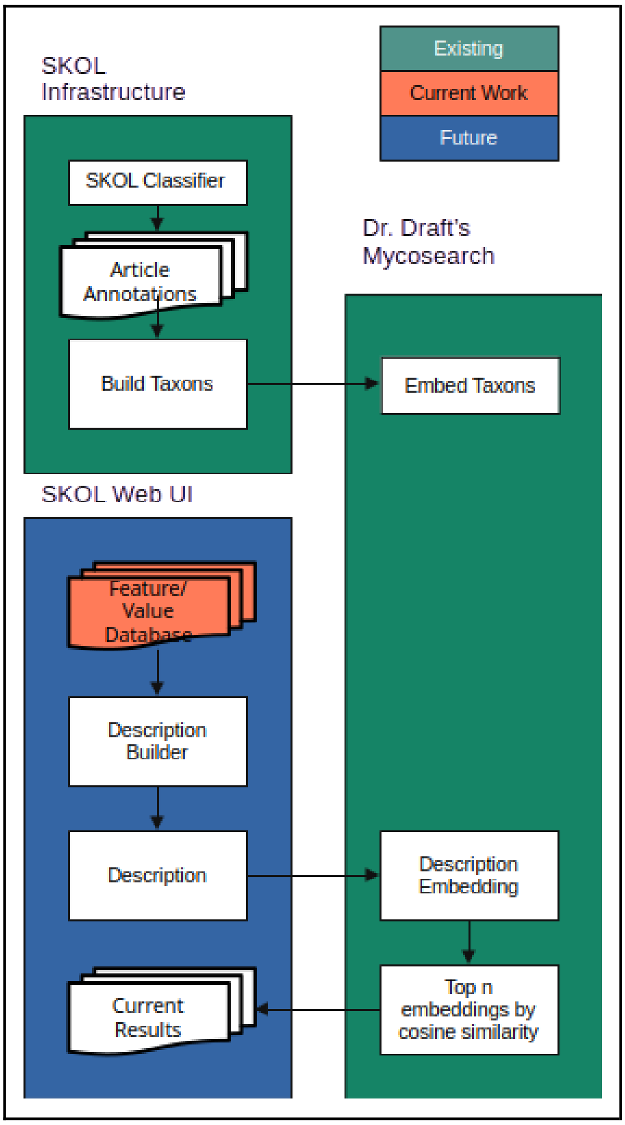

Fig. 8. SKOL System Diagram.

ACKNOWLEDGMENT

The Auton Lab at Carnegie Mellon University has made compute resources available for some of the work in the current paper.

REFERENCES

Adithya, S. K. (2023). A Beginner’s Guide to Fine-Tuning Mistral 7B Instruct Model. Medium. https://adithyask.medium.com/a-beginners-guide-to-fine-tuning-mistral-7b-instruct-model-0f39647b20fe. Retrieved 2025-06-18.

Ait, A., Cánovas Izquierdo, J.L., Cabot, J. (2023). On the Suitability of Hugging Face Hub for Empirical Studies. ESEM 2023. https://doi.org/10.48550/arXiv.2307.14841

Anonymous.. (2022). Find the jaccard index and jaccard distance between the two given sets. GeeksForGeeks. https://www.geeksforgeeks.org/dsa/find-the-jaccard-index-and-jaccard-distance-between-the-two-given-sets/. Retrieved 2025-06-17.

Gisolfi, N. (2024). Dr Draft's state-of-the-art (SOTA) Literature Search. Auton Lab. https://github.com/autonlab/dr-drafts-sota-literature-search. Retrieved 2025-06-18.

Hennebert, G.L., Korf, R. P. eds., (1974-2010) Mycotaxon: A New Journal on Taxonomy and Nomenclature of Fungi and Lichens, Ithaca, NY.

Jang, A. et al. (2023). Mistral 7B. https://doi.org/10.48550/arXiv.2310.06825

Khan, H.K. (2023). A Step-by-Step Guide to Fine-Tuning the Mistral 7B LLM. E2E Cloud. https://www.e2enetworks.com/blog/a-step-by-step-guide-to-fine-tuning-the-mistral-7b-llm. Retrieved 2025-06-18.

Murrill, W.A., ed., (1909-1961). Mycologia, New York Botanical Garden, Mycological Society of America.

Nauta, M.M., Noordeloos, M.E., eds., (1959-1998). Persoonia: A Mycological Journal, Rijksherbarium, Leiden, The Netherlands.

Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP 2019. https://doi.org/10.48550/arXiv.1908.10084

Wood, M., (1995-2025). MykoWeb Journals. https://www.mykoweb.com/systematics/journals.html. Retrieved 2019-04.

Yarroll, L.H.P, Balasi, J., Murphy, C., & Osuga, S. (2024a). Mycology Literature Search. Syracuse University. https://github.com/piggyatbaqaqi/skol/blob/main/IST664/IST664_Team3_Balasi_Murphy_Osuga_Yarroll.pdf Retrieved 2025-06-18.

Yarroll, L.H.P., Caspers, D., Murphy, C.. (2024b). Synoptic Key of Life II, Final Project Report. Syracuse University. https://github.com/piggyatbaqaqi/skol/blob/main/IST718/IST718_Final_Report_FINAL.pdf. Retrieved 2025-06-18.

Yarroll, L.H.P. (2025). Synoptic Key of Life. Mid Atlantic Mycology Conference 2025 (MASMC2025). https://github.com/piggyatbaqaqi/skol/blob/main/MASMC2025/Synoptic%20Key%20of%20Life%20MASMC%202025_poster.pdf. Retrieved 2025-06-18.