Aligning Documentation and Q&A Forum through Constrained Decoding with Weak Supervision

Rohith Pudari*, Shiyuan Zhou*, Prof. Iftekhar Ahmed†,
Dr. Zhuyun Dai‡ and Prof. Shurui Zhou*.

*University of Toronto
†University of California, Irvine
‡Google

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FORCOLAB

UCI University of California, Irvine

Google

StackOverflow

StackOverflow posts

- StackOverflow users often ***manually*** link external websites in their answers.[1]
  - Documentations
  - Blog posts
  - Tutorials …

[1] Baltes et.al., Contextual documentation referencing on StackOverflow, TSE-2022.

Motivating Example

StackOverflow question: How to get POSTed JSON in Flask?

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

On the browser it correctly returns the UUID I put in the GET, but on the console, it just prints out `None` (where I expect it to print out the `{"text":"lalala"}`. Does anybody know how I can get the posted JSON from within the Flask method?

Motivating Example

StackOverflow question: How to get POSTed JSON in Flask?

622

First of all, the `.json` attribute is a property that delegates to the `request.get_json()` method, which documents why you see `None` here.

You need to set the request content type to `application/json` for the `.json` property and `.get_json()` method (with no arguments) to work as either will produce `None` otherwise. See the Flask `Request` documentation:

> The parsed JSON data if `mimetype` indicates JSON (`application/json`, see `.is_json`).

You can tell `request.get_json()` to skip the content type requirement by passing it the `force=True` keyword argument.

Motivating Example

StackOverflow question: How to get POSTed JSON in Flask?

622

First of all, the `.json` attribute is a property that delegates to the `request.get_json()` method, which documents why you see `None` here.

You need to set the request content type to `application/json` for the `.json` property and `.get_json()` method (with no arguments) to work as either will produce `None` otherwise. See the Flask `Request` documentation:

> The parsed JSON data if `mimetype` indicates JSON (`application/json`, see `.is_json`).

You can tell `request.get_json()` to skip the content type requirement by passing it the `force=True` keyword argument.

Links to docs

Motivating Example

StackOverflow question: How to get POSTed JSON in Flask?

622

First of all, the `.json` attribute is a property that delegates to the `request.get_json()` method, which documents why you see `None` here.

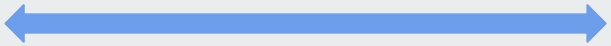
You need to set the request content type to `application/json` for the `.json` property and `.get_json()` method (with no arguments) to work as either will produce `None` otherwise. See the Flask `Request` documentation:

> The parsed JSON data if `mimetype` indicates JSON

Links to docs

53 out of 200 sampled StackOverflow answers had at least one link to documentation.

Documentation and StackOverflow



StackOverflow questions

Documentation

- The links to documentation sources support the answer but can be laborious to perform manual search.

- In our study, we focus on documentation links - formal resource to get ground truth.

Documentation and StackOverflow

Alignment?

StackOverflow questions

Documentation

- The links to documentation sources support the answer but can be laborious to perform manual search.

- In our study, we focus on documentation links - formal resource to get ground truth.

Documentation and StackOverflow

DOSA

StackOverflow questions

Documentation

- The links to documentation sources support the answer but can be laborious to perform manual search.

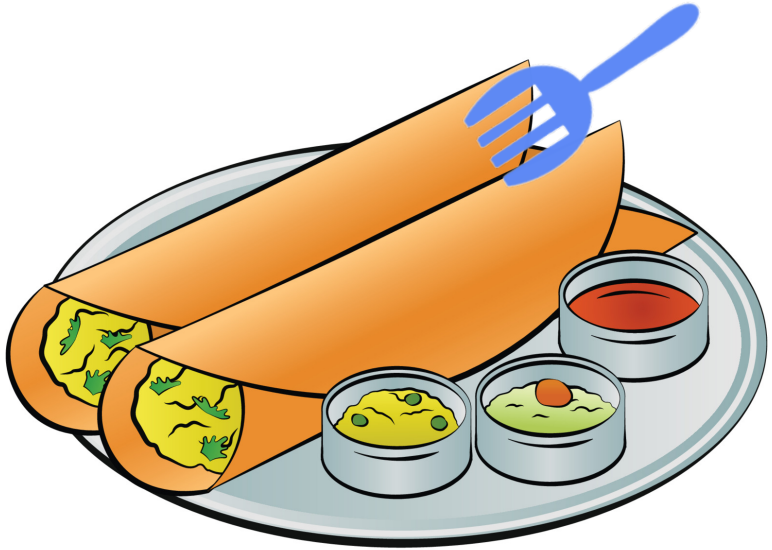
- In our study, we focus on documentation links - formal resource to get ground truth.

DOSA

Documentation and **S**tackOverflow **A**lignment

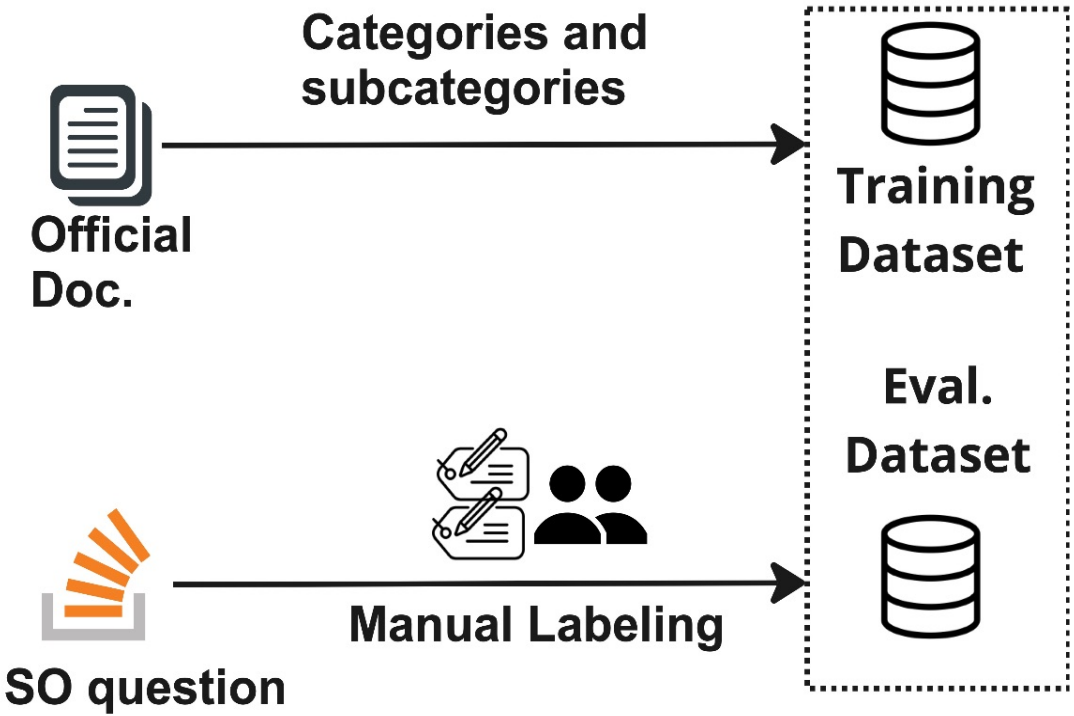
DOSA

Documentation and **S**tackOverflow **A**lignment

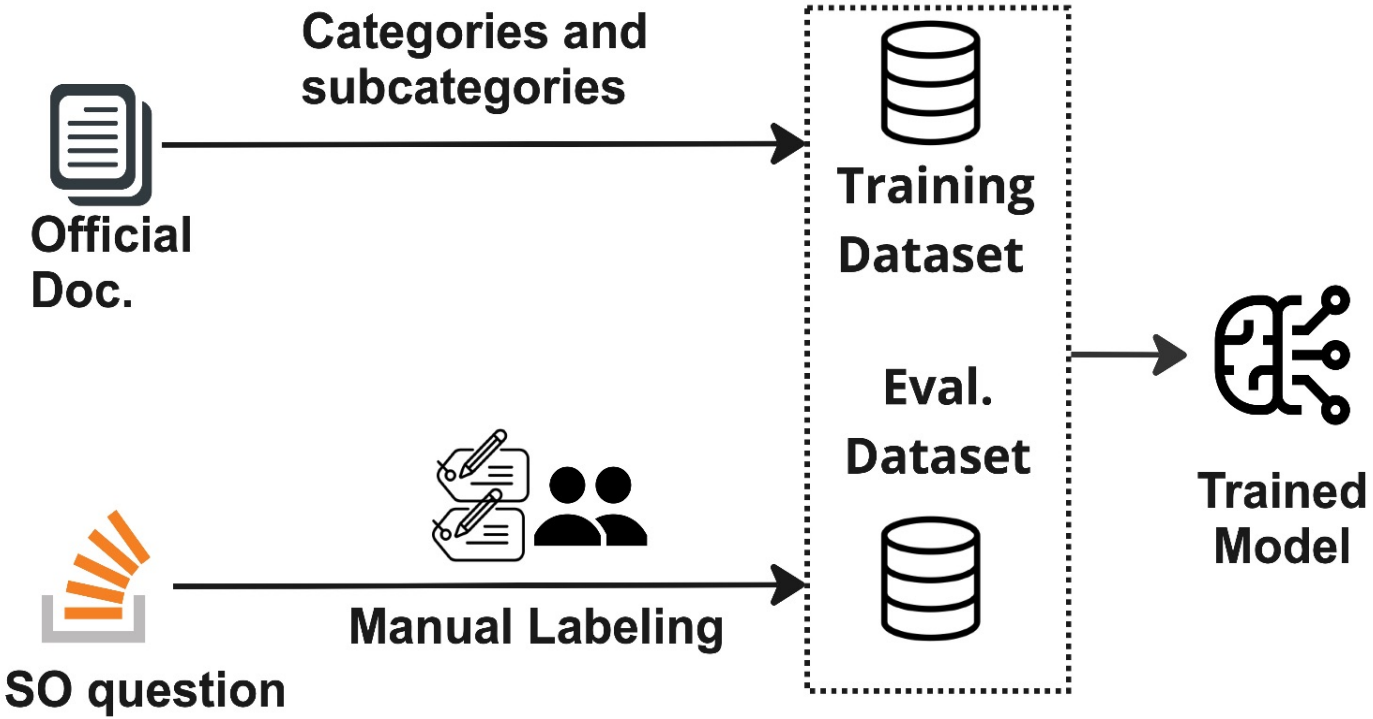


Dosa

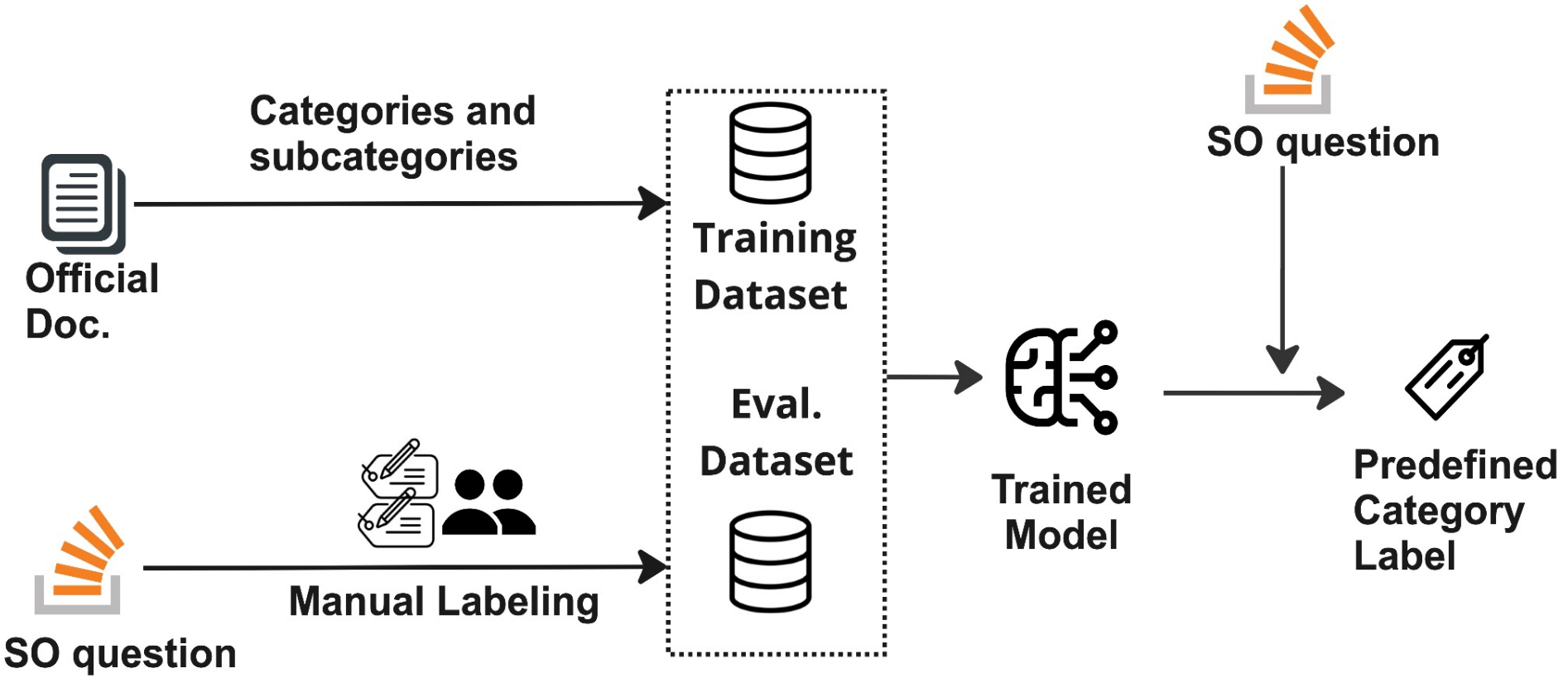
Our Approach - DOSA



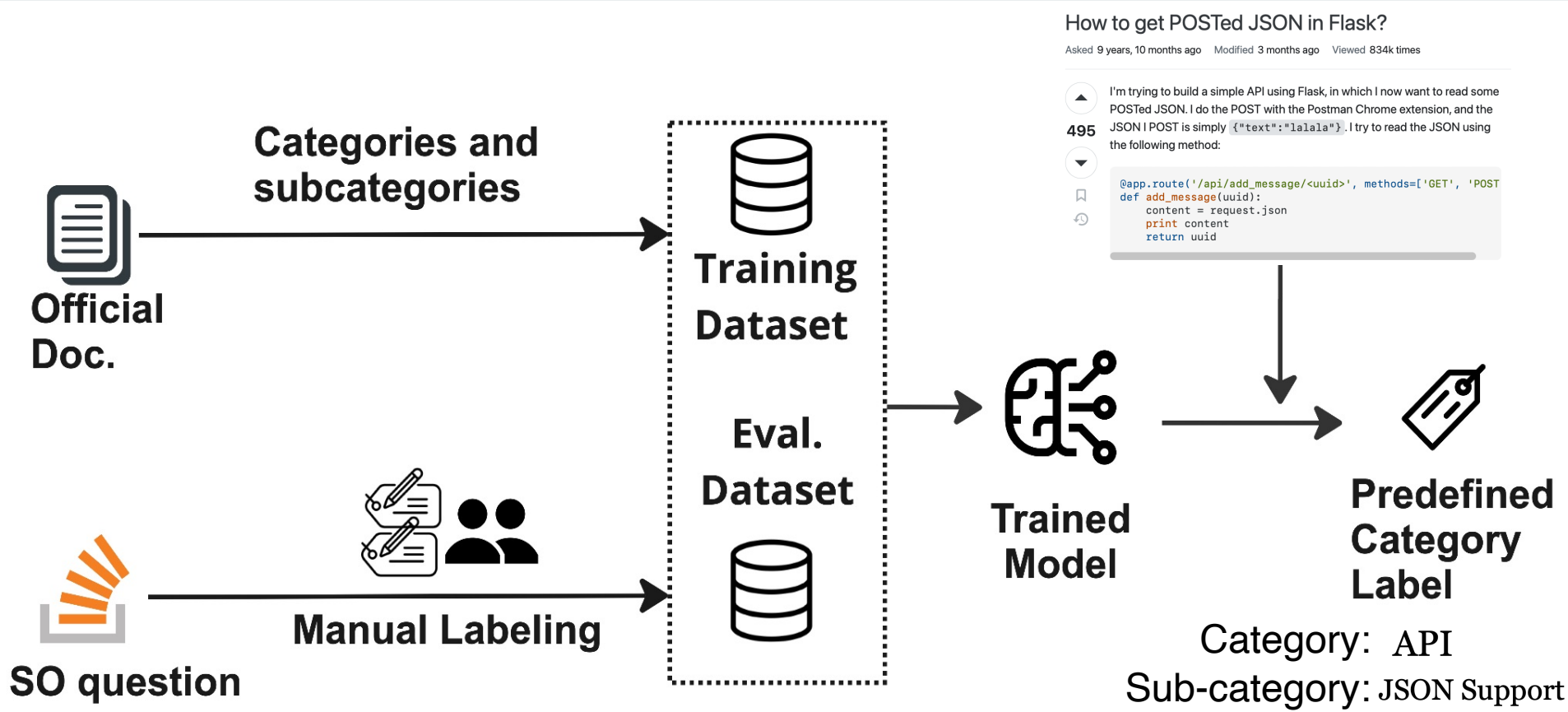
Our Approach - DOSA



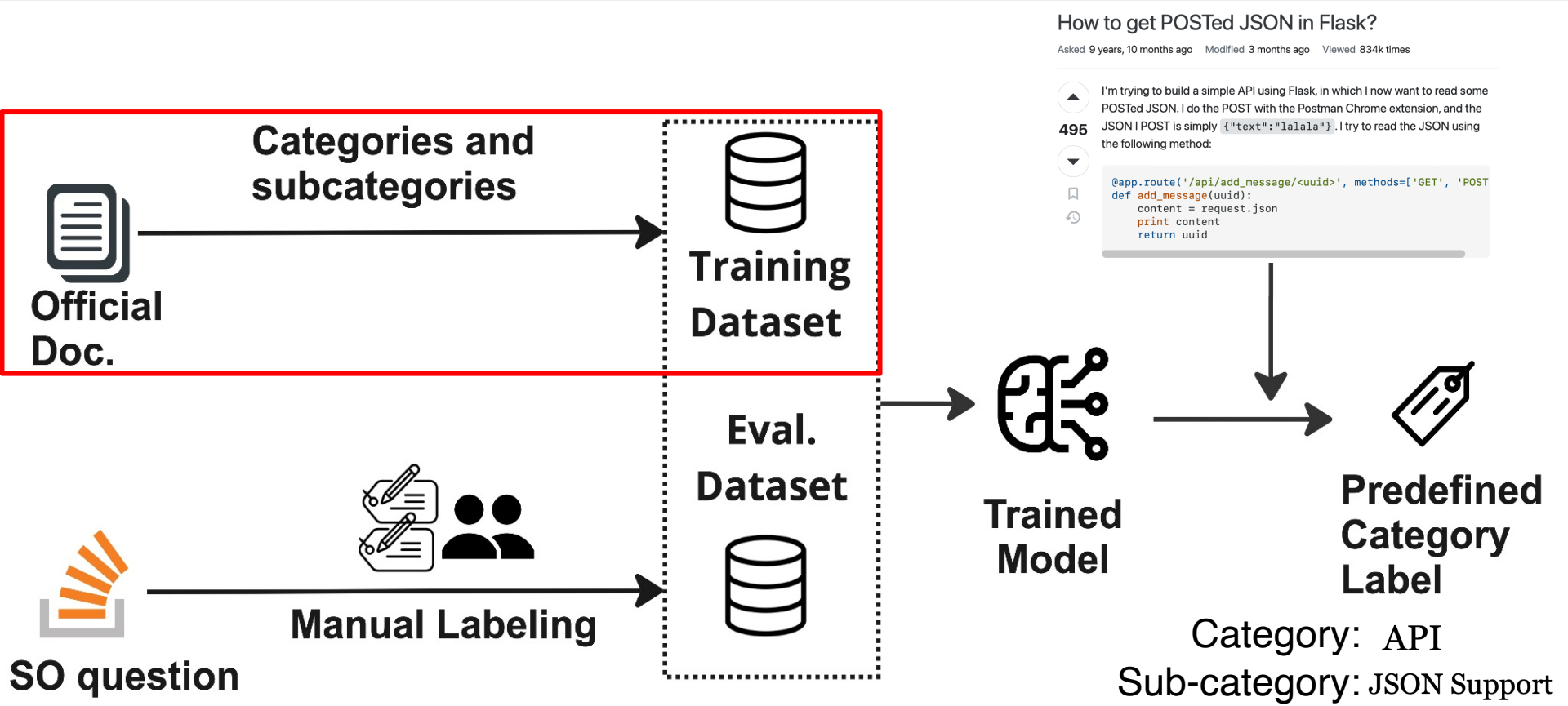
Our Approach - DOSA



Our Approach - DOSA



Our Approach - DOSA



Training Dataset

API

This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support

Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.

Training Dataset

API C1

This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support

Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.

Training Dataset

API C1

This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support SC1-1

Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.

Training Dataset

API C1

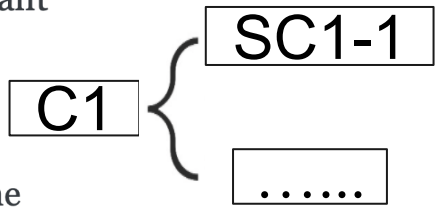
This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support SC1-1

Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.



C = Category,
SC = Sub-category,
s = Sentence.

Training Dataset

API C1

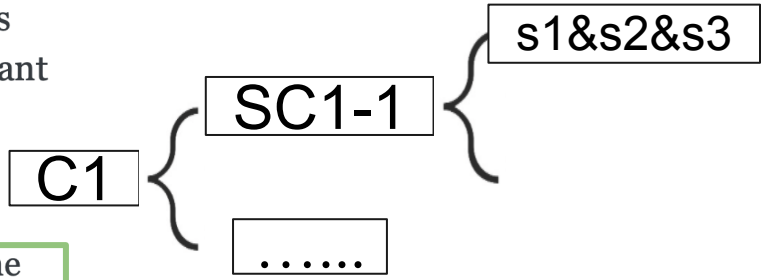
This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support SC1-1

Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.



C = Category,
SC = Sub-category,
s = Sentence.

Training Dataset

API C1

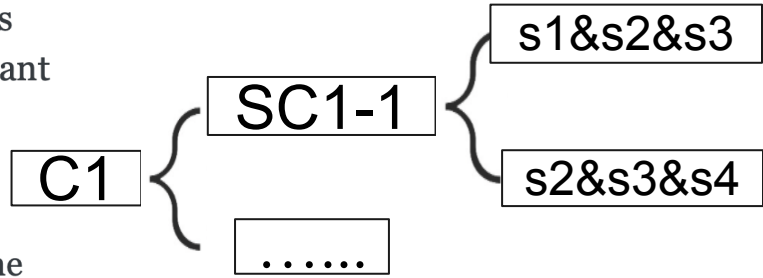
This part of the documentation covers all the interfaces of Flask. For parts where Flask depends on external libraries, we document the most important right here and provide links to the canonical documentation.

.....

JSON Support SC1-1

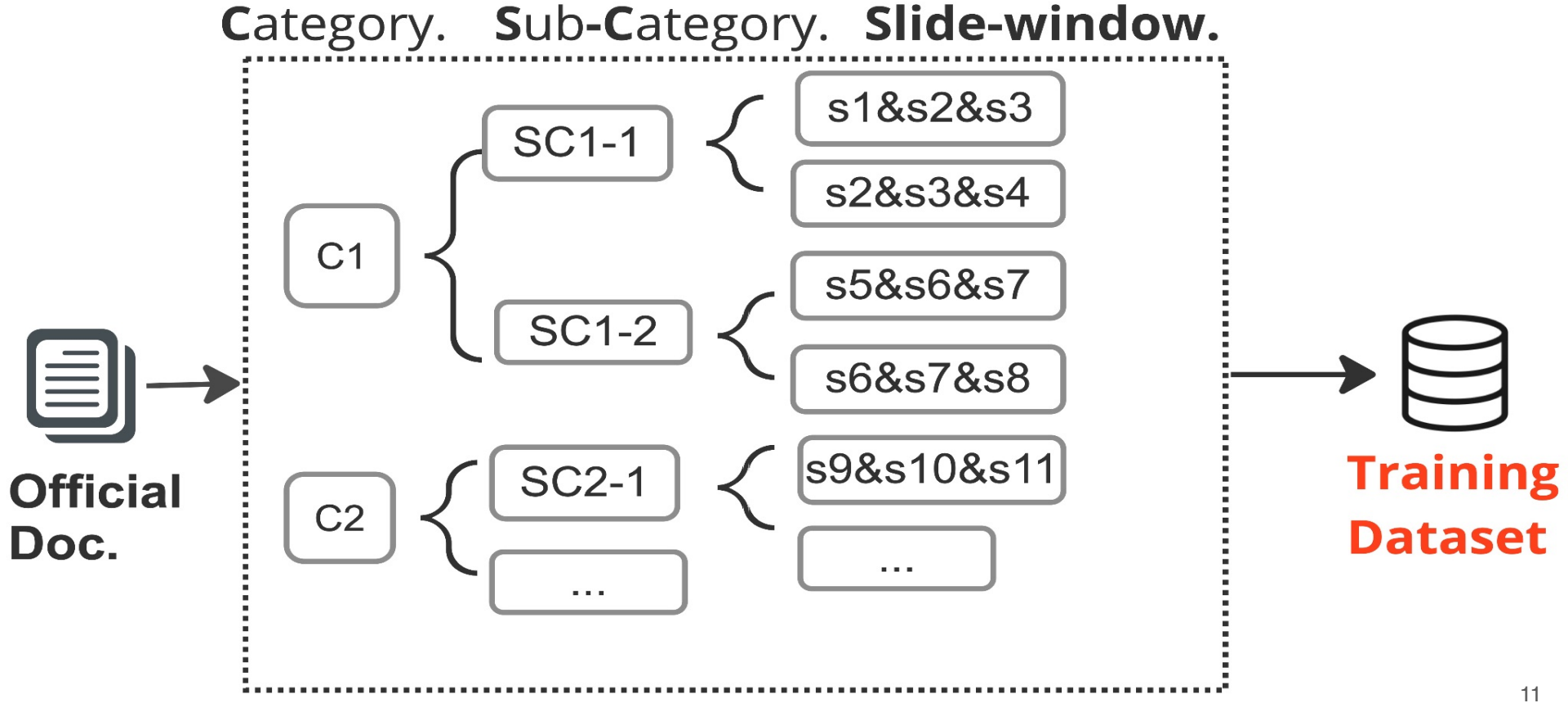
Flask uses Python's built-in **`json`** module for handling JSON by default. The JSON implementation can be changed by assigning a different provider to **`flask.Flask.json_provider_class`** or **`flask.Flask.json`**. The functions provided by `flask.json` will use methods on `app.json` if an app context is active.

Jinja's `|tojson` filter is configured to use the app's JSON provider. The filter marks the output with `|safe`. Use it to render data inside HTML `<script>` tags.

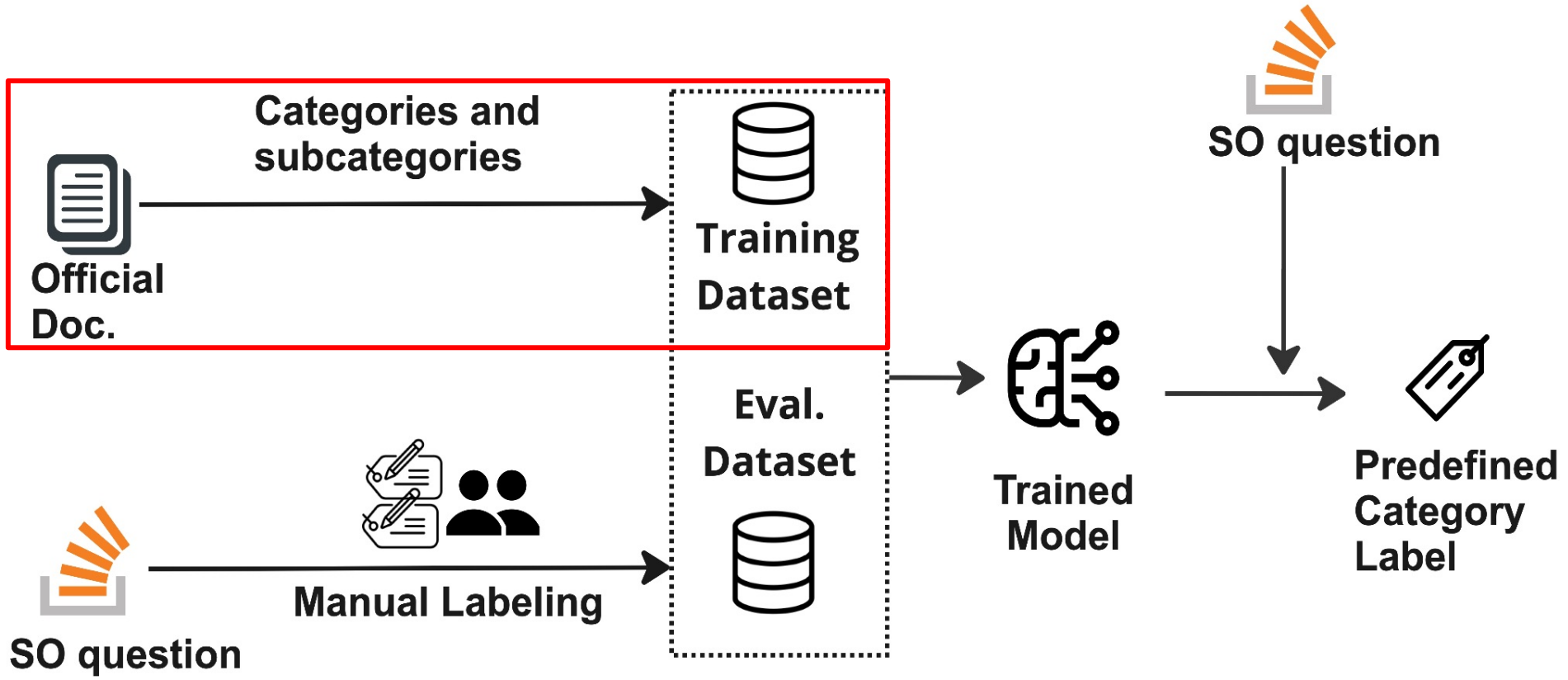


C = Category,
SC = Sub-category,
s = Sentence.

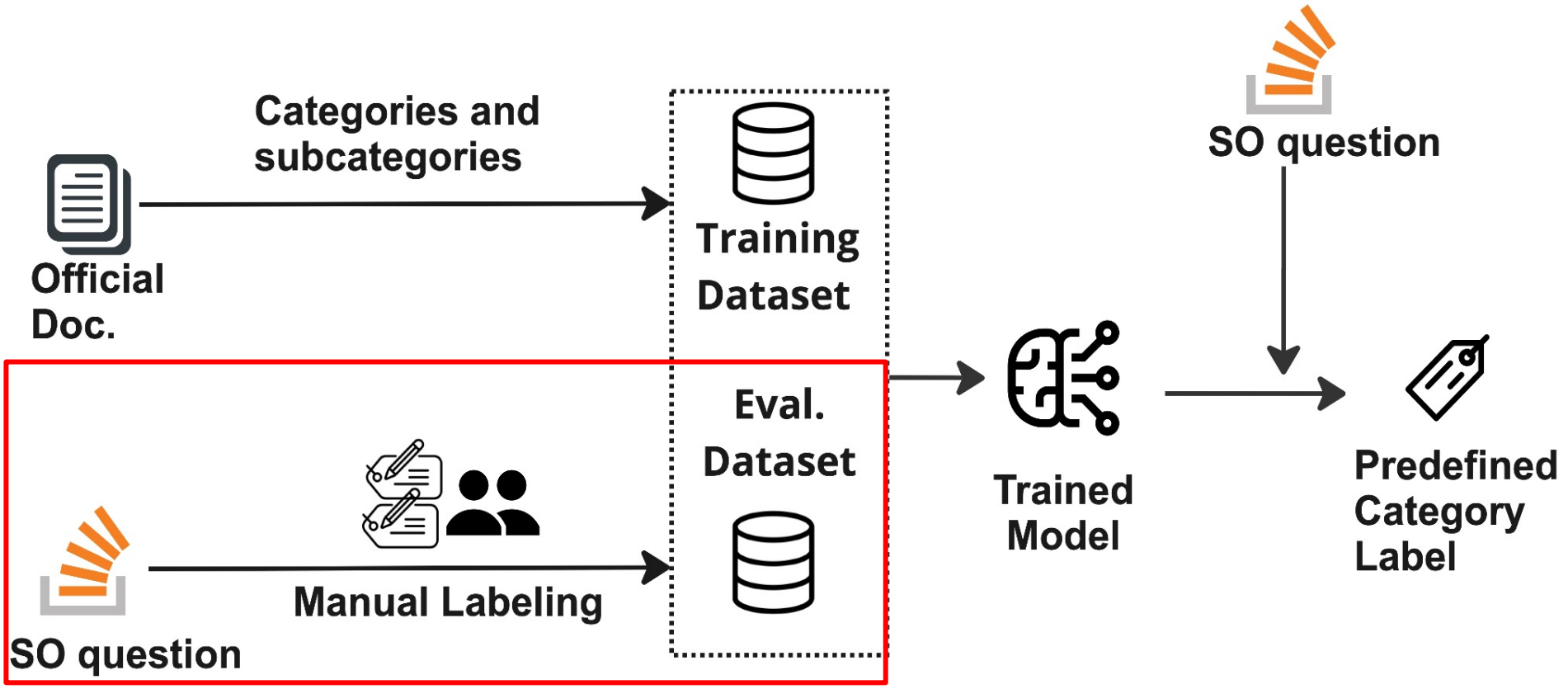
Training Dataset



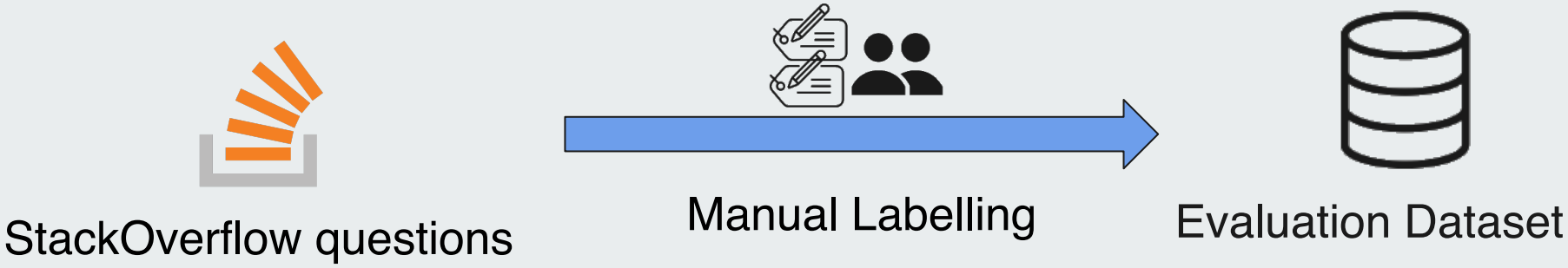
Our Approach - DOSA



Our Approach - DOSA

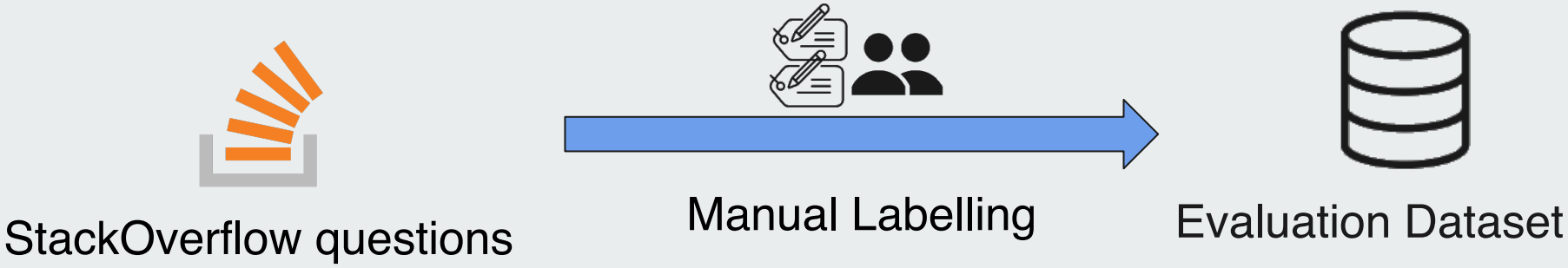


Evaluation Dataset



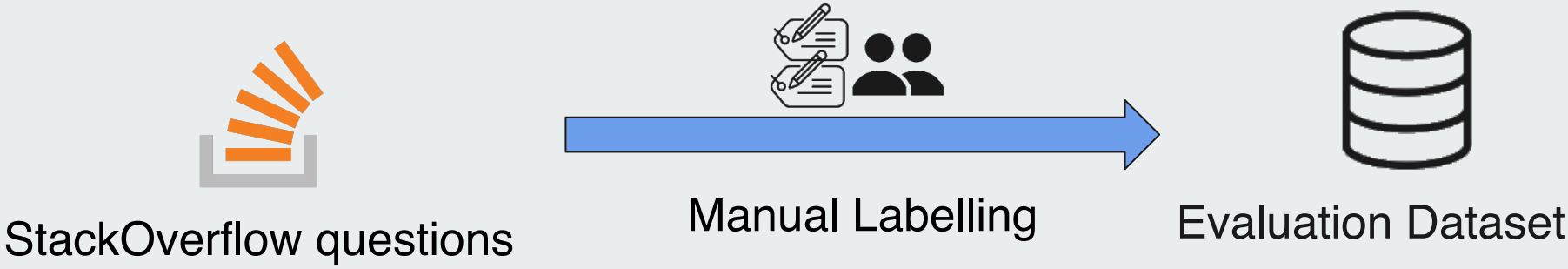
- Random sampled 200 StackOverflow questions for each of the ‘Python’ and ‘Flask’ tags.

Evaluation Dataset



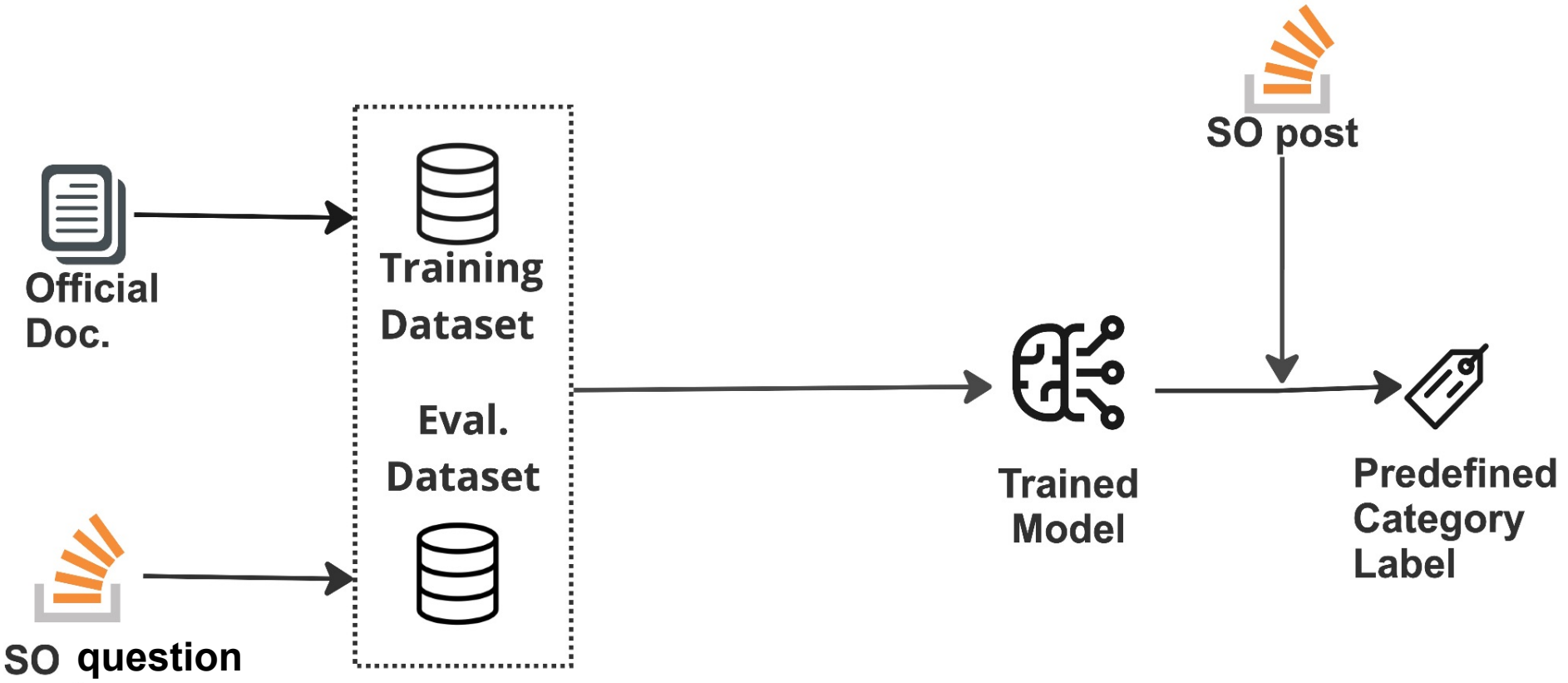
- Random sampled 200 StackOverflow questions for each of the ‘Python’ and ‘Flask’ tags.
- Two of the authors independently labelled sampled questions.
- Cohen’s Kappa score of 0.83

Evaluation Dataset

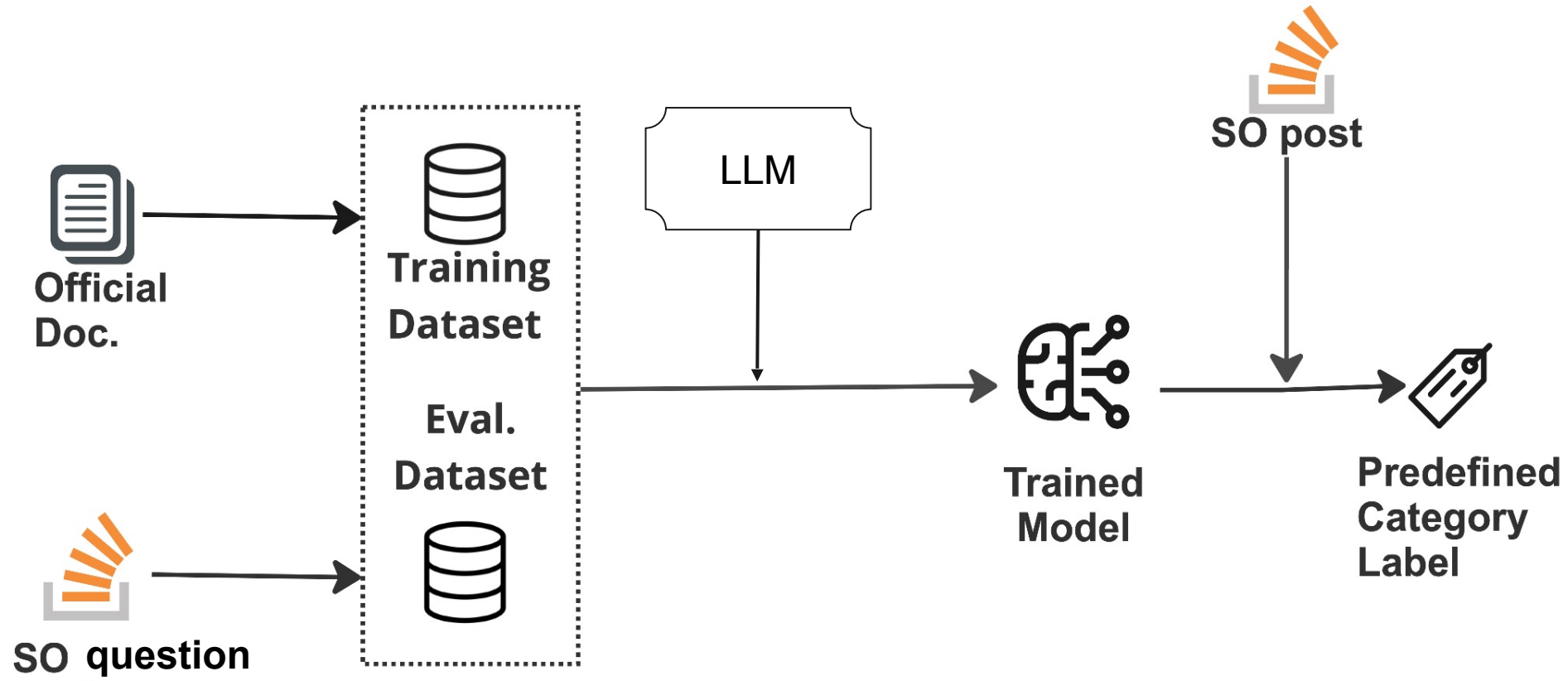


- Random sampled 200 StackOverflow questions for each of the ‘Python’ and ‘Flask’ tags.
- Two of the authors independently labelled sampled questions.
- Cohen’s Kappa score of 0.83
- We used the categories and subcategories extracted from the documentation as the labels.

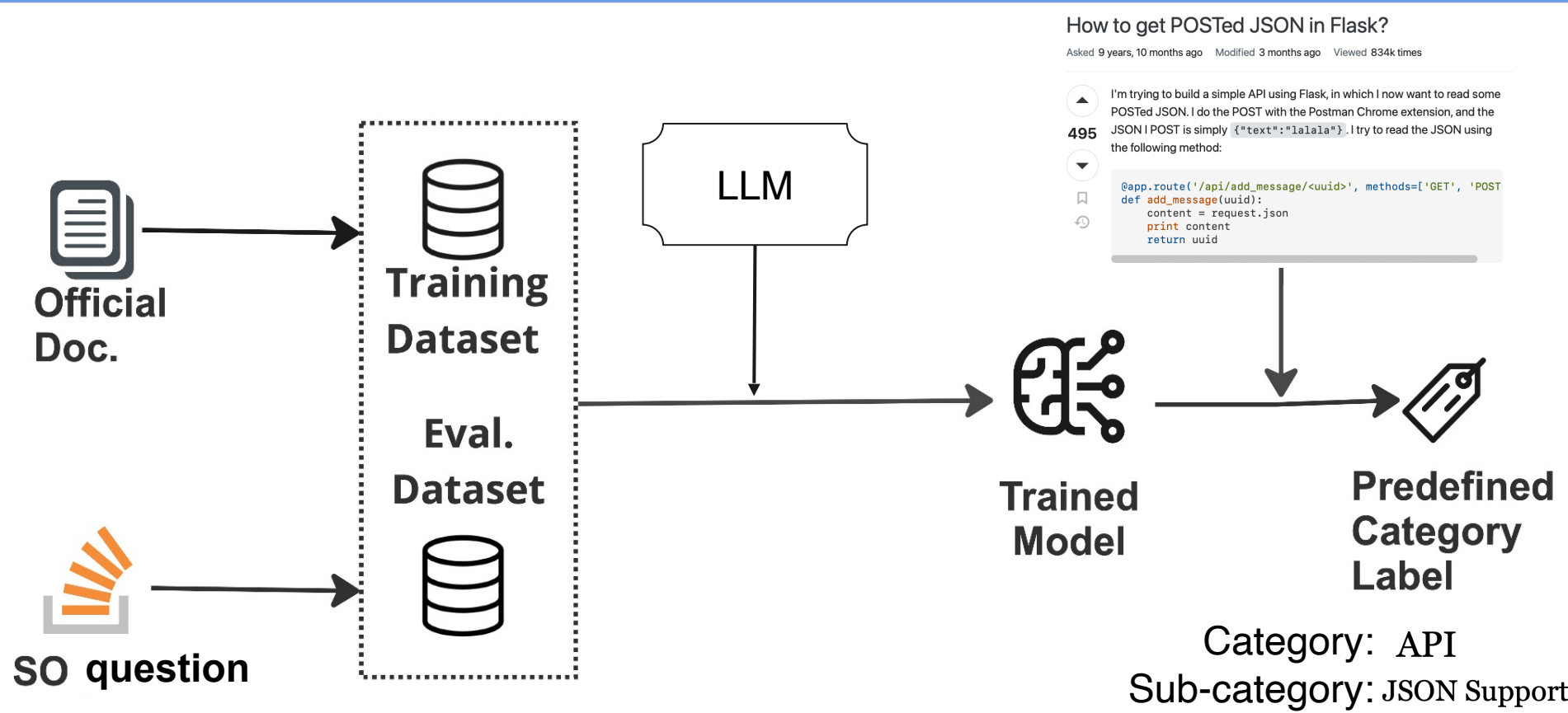
Our Approach - DOSA



Our Approach - DOSA



Our Approach - DOSA



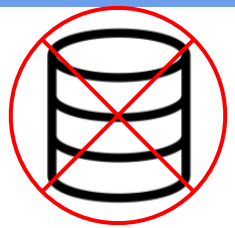
Why LLM?

StackOverflow question

Documentation

Why LLM?

StackOverflow question



Documentation

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

Category: API

Sub-category: JSON Support

Why LLM?

StackOverflow question

Documentation

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

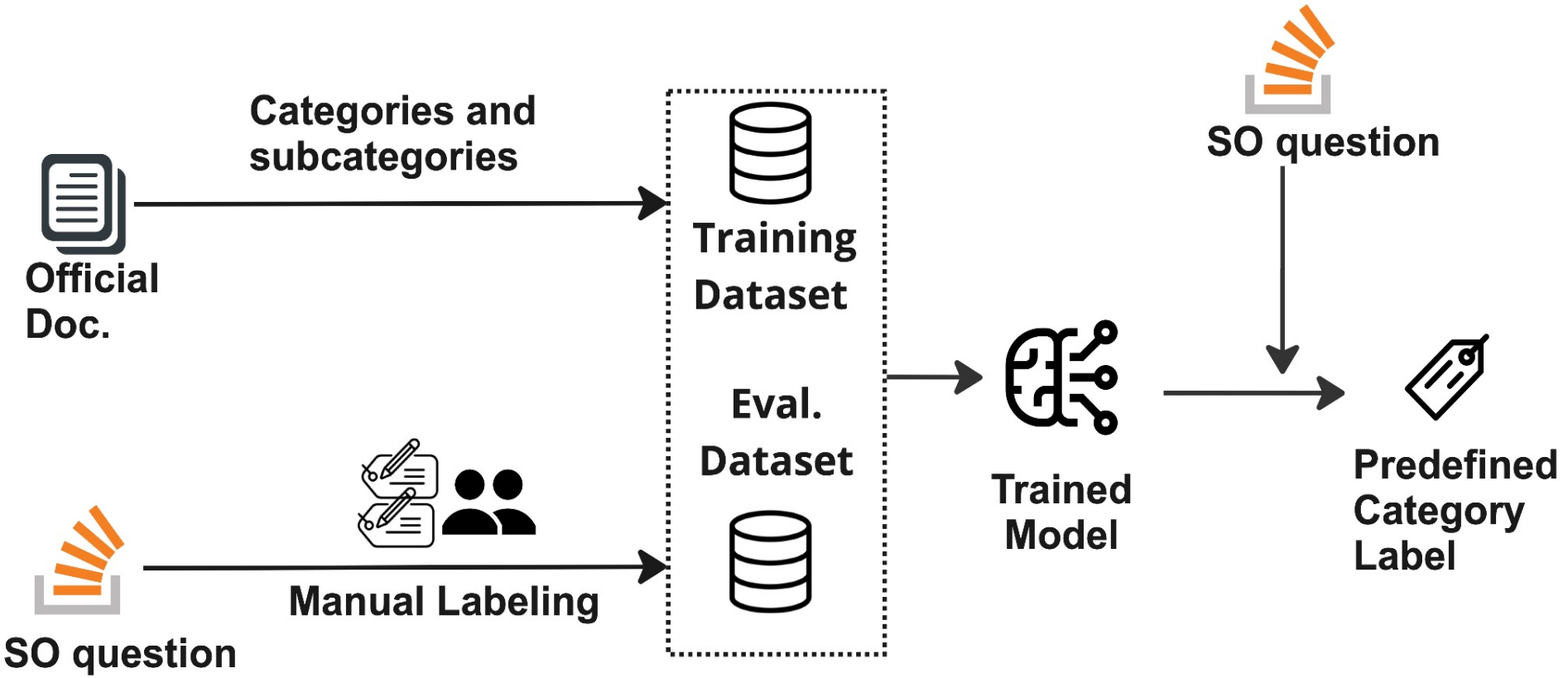
```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

Category: API

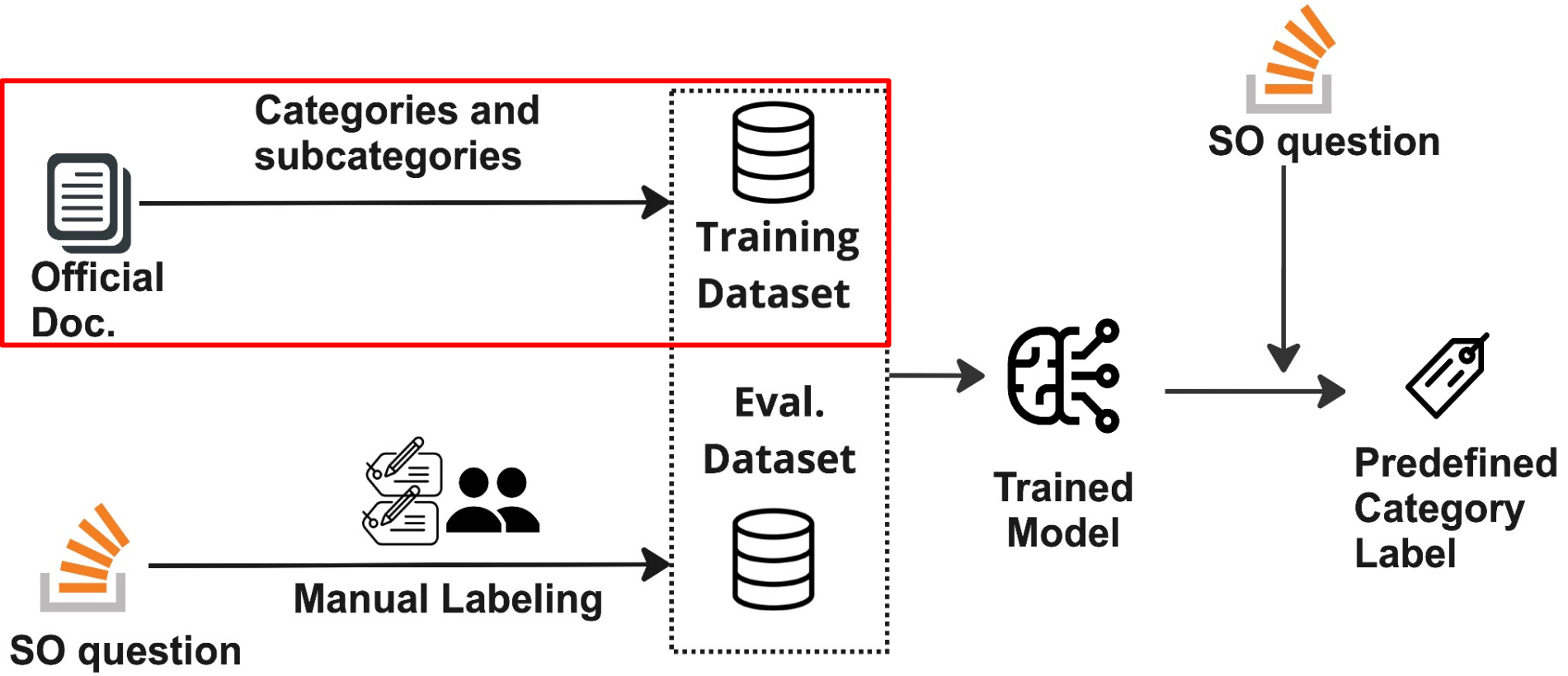
Sub-category: JSON Support

- No public dataset available between StackOverflow and Documentation.
- LLMs can help establish this connection in a zero-shot manner.

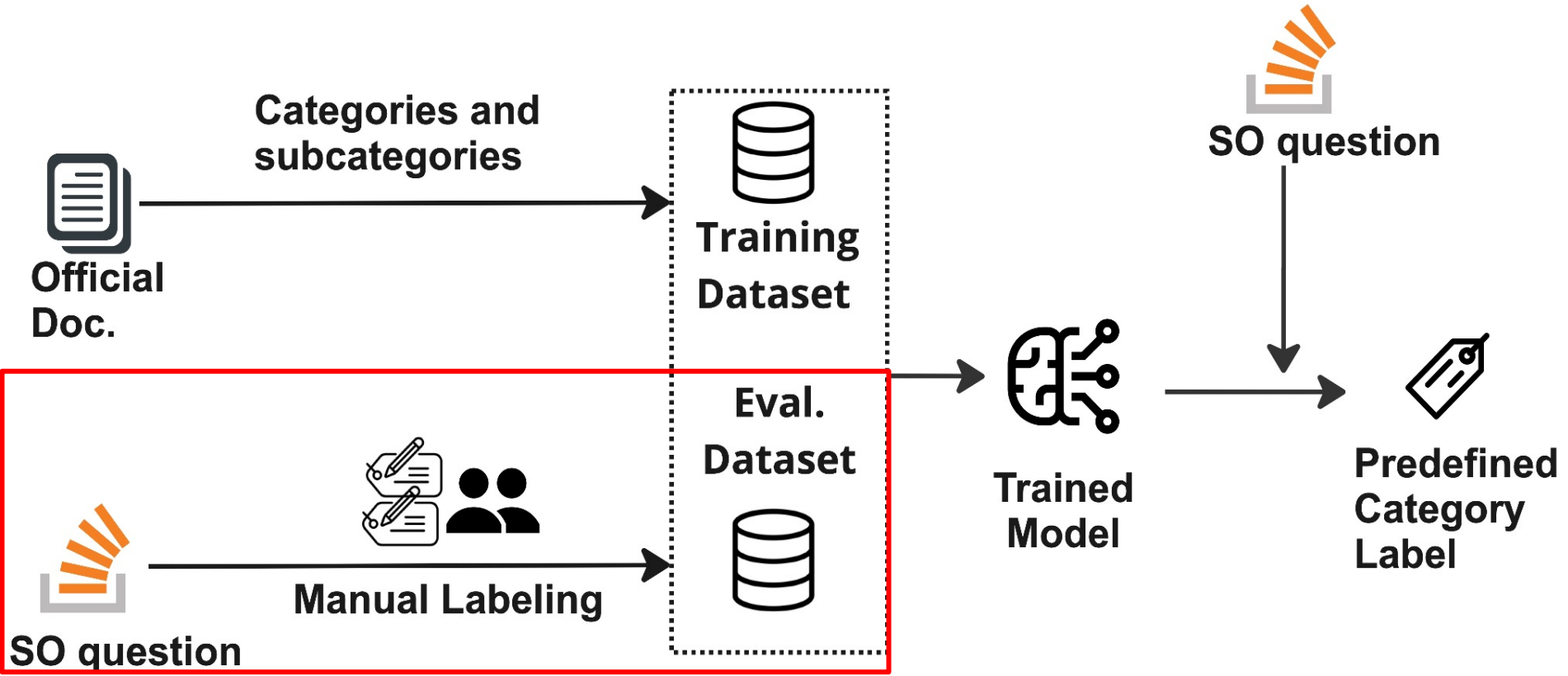
Our Approach - DOSA



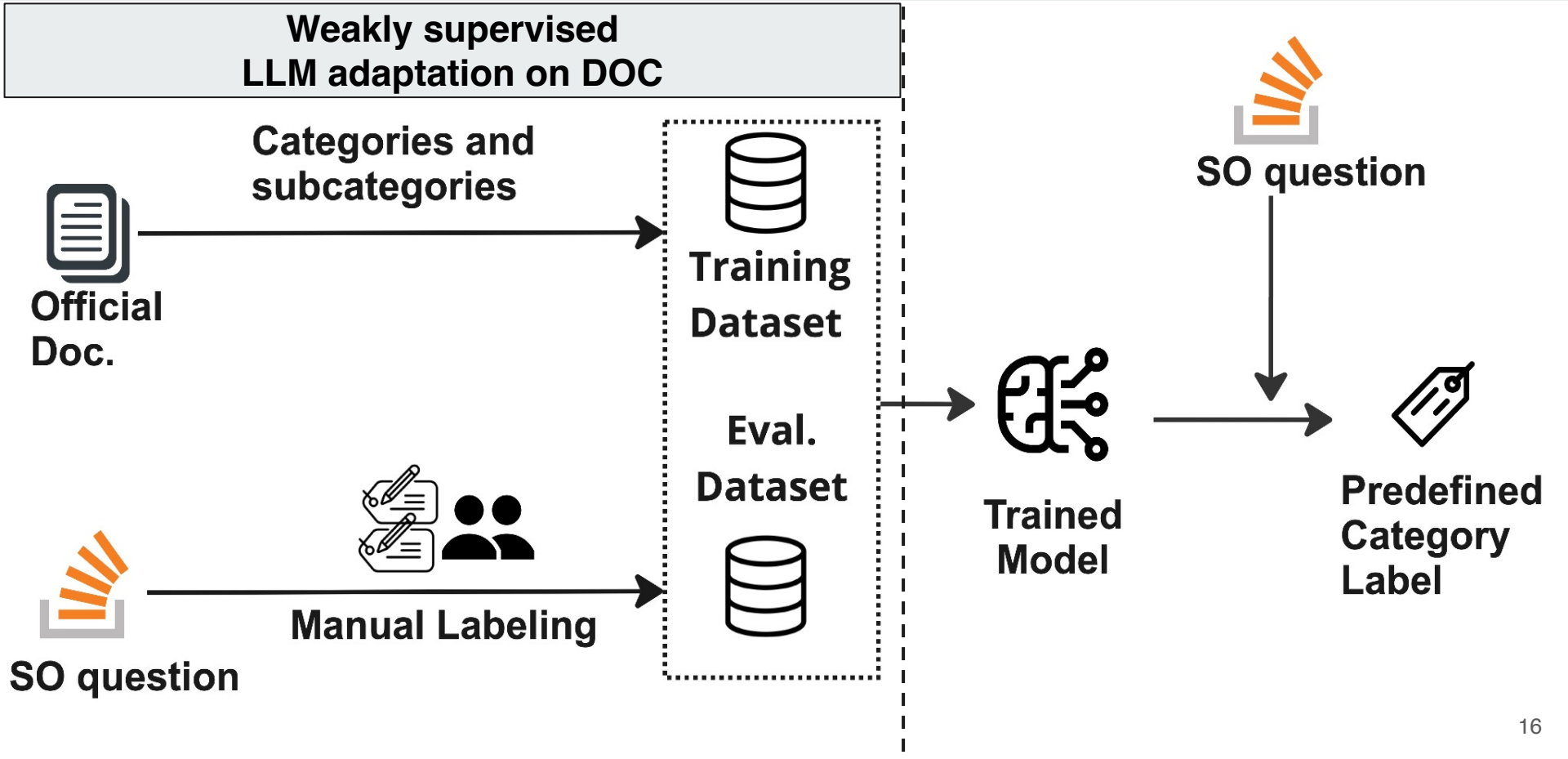
Our Approach - DOSA



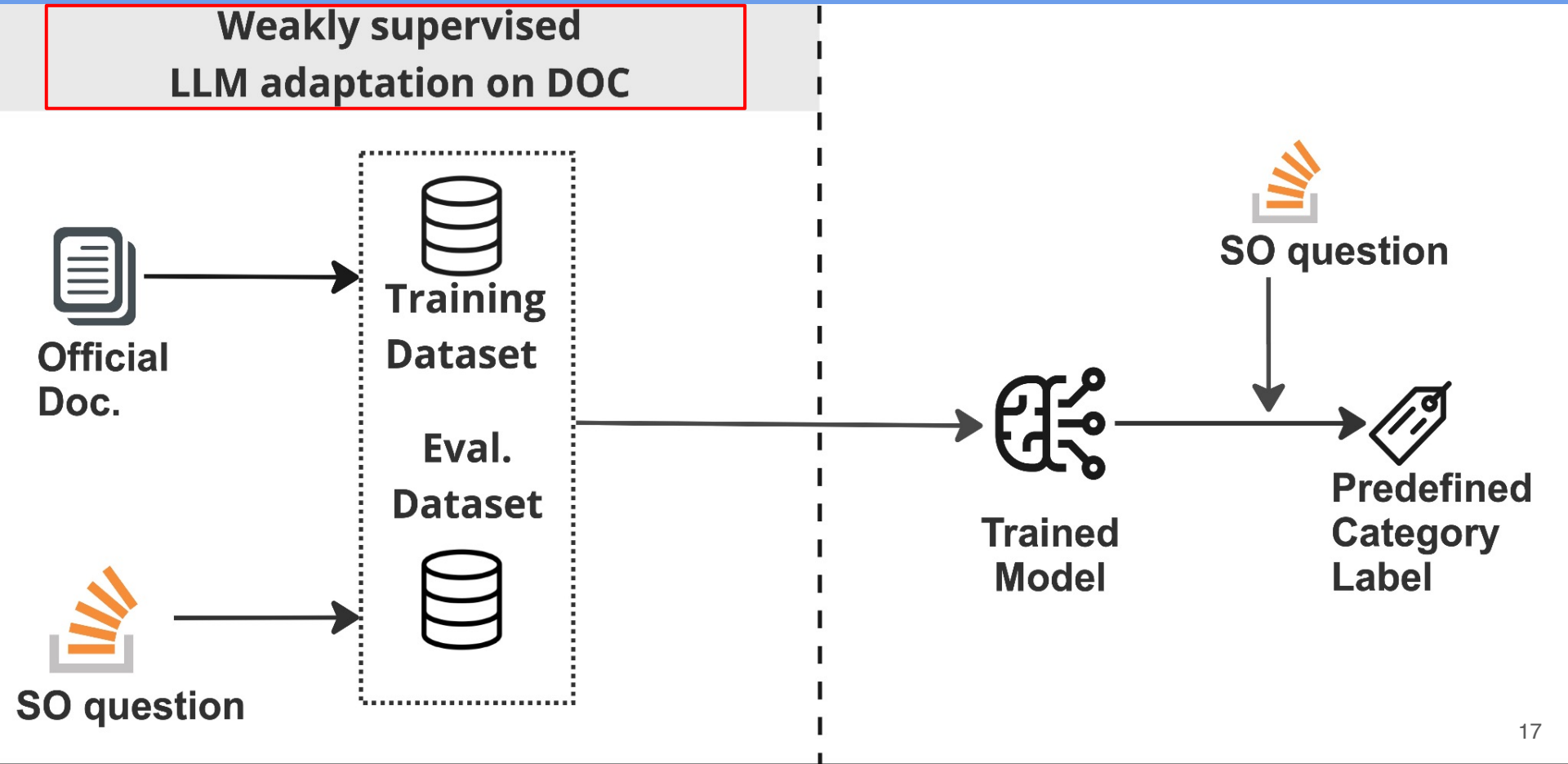
Our Approach - DOSA



Our Approach - DOSA



Our Approach - DOSA



Out of scope output sequence

Input
(SO question)

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

Out of scope output sequence

Input
(SO question)

GPT-2 Model

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

Out of scope output sequence

Input
(SO question)

GPT-2 Model

Output
(categories extracted from documentation)

[post, postman, text, print, build, …]

Out of scope output sequence

Input
(SO question)

GPT-2 Model

Output
(categories extracted from documentation)

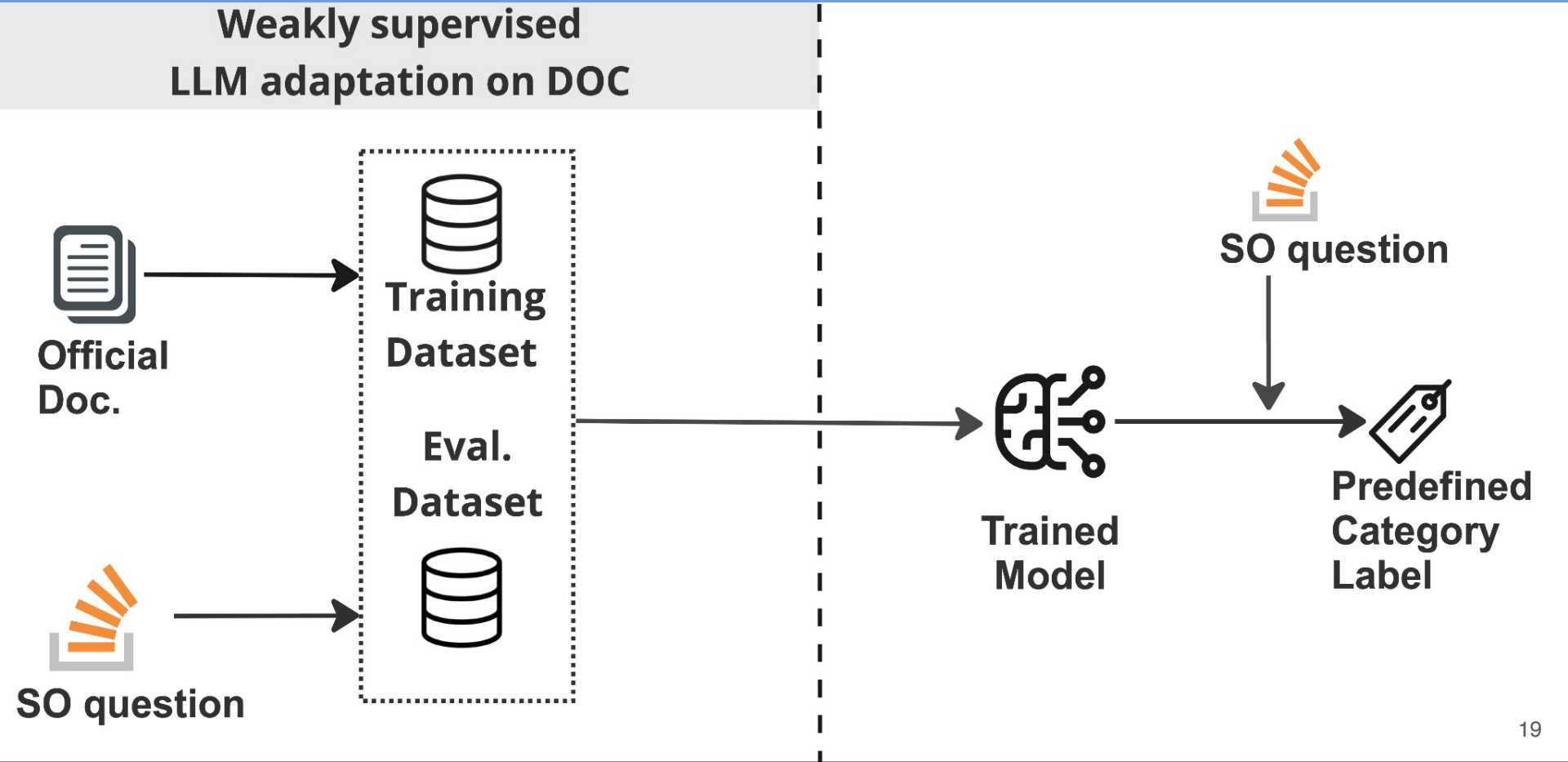
[post, postman, text, print, build, …]

Expected output

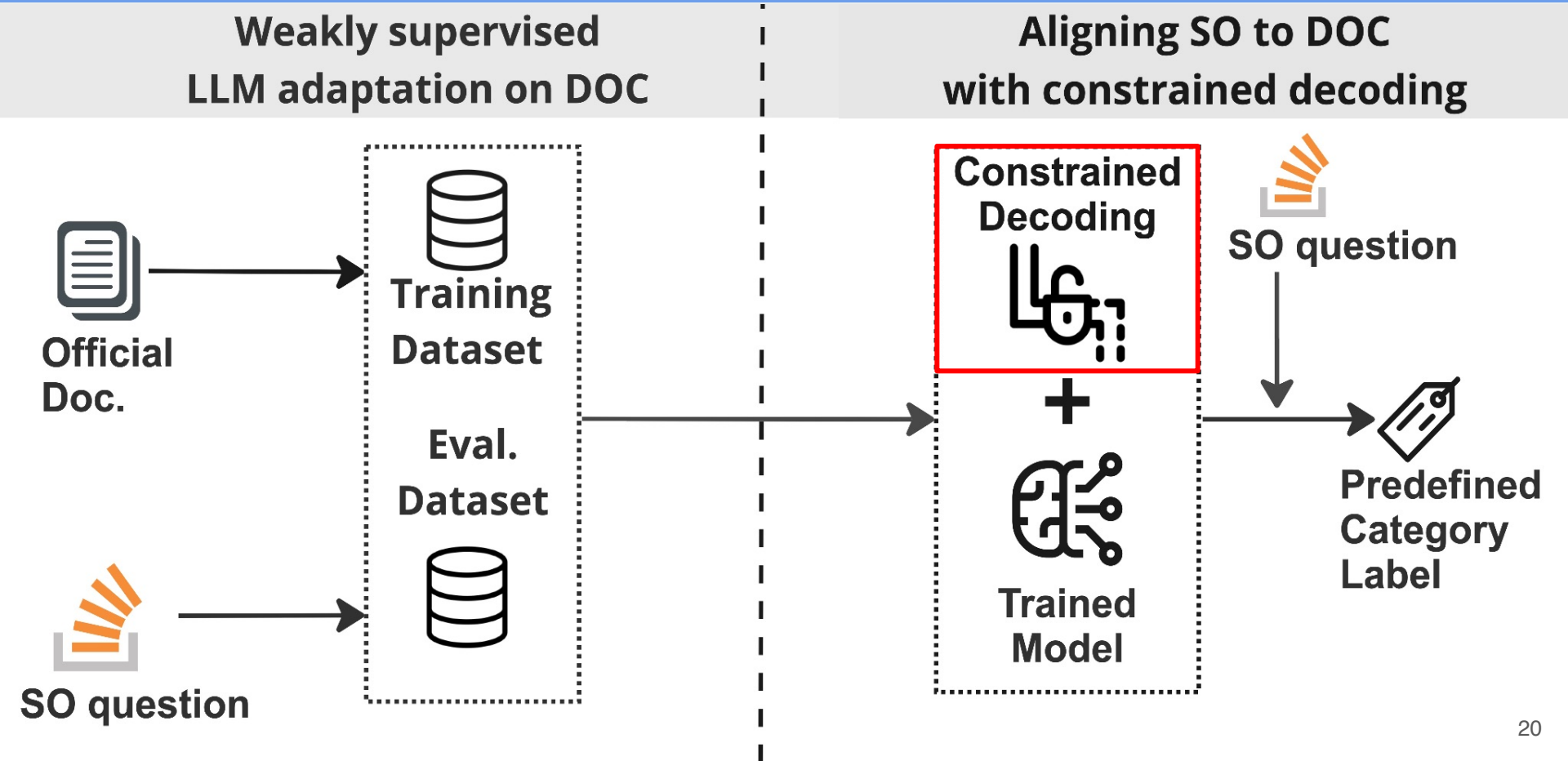
Category: API
Sub-category: JSON Support

- Output is out of scope for category/subcategory vocabulary.

Our Approach - DOSA

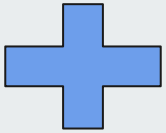


Our Approach - DOSA



Aligning SO to DOC with Constrained Decoding

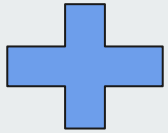
Trained LLM



Constrained decoding

Aligning SO to DOC with Constrained Decoding

Trained LLM



Constrained decoding

- Incorporating domain terminology during the token generation process.
- Constrained decoding only operates during the generation process at test time.

Out of scope output sequence

Input
(SO question)

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

Out of scope output sequence

Input (SO question)

How to get POSTed JSON in Flask?

Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

GPT-2 Model

Output (categories extracted from documentation)

[post, postman, text, print, build, …]

Expected output

Category: API

Sub-category: JSON Support

Out of scope output sequence

Input (SO question)

How to get POSTed JSON in Flask?

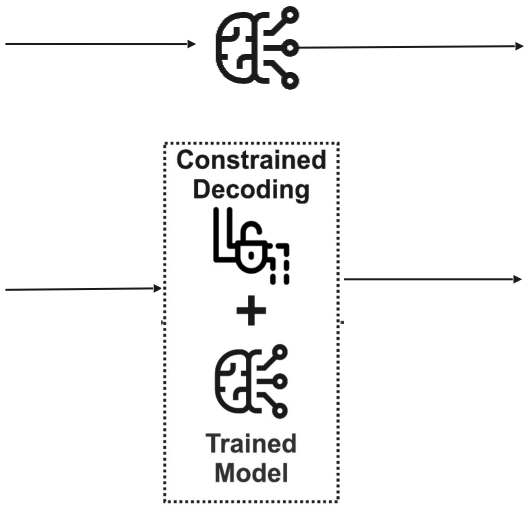
Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

GPT-2 Model



Output (categories extracted from documentation)

[post, postman, text, print, build, …]

Expected output

Category: API

Sub-category: JSON Support

Out of scope output sequence

Input (SO question)

How to get POSTed JSON in Flask?

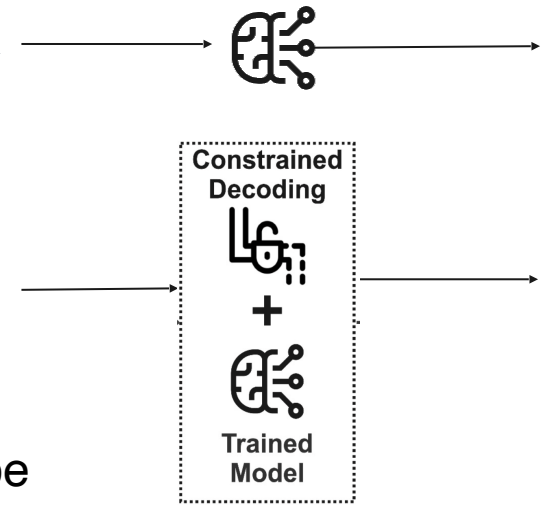
Asked 9 years, 10 months ago Modified 3 months ago Viewed 834k times

495

I'm trying to build a simple API using Flask, in which I now want to read some POSTed JSON. I do the POST with the Postman Chrome extension, and the JSON I POST is simply `{"text":"lalala"}`. I try to read the JSON using the following method:

```
@app.route('/api/add_message/<uuid>', methods=['GET', 'POST
def add_message(uuid):
    content = request.json
    print content
    return uuid
```

GPT-2 Model



Output (categories extracted from documentation)

[post, postman, text, print, build, …]

[JSON, API, application, handling, quickstart, ...]

- Produce tokens within the scope

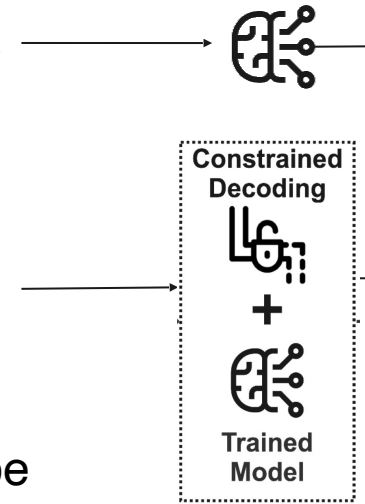
Expected output

Category: API
Sub-category: JSON Support

Out of scope output sequence

Input
(SO question)

GPT-2 Model



Output
(categories extracted from documentation)

[post, postman, text, print, build, …]

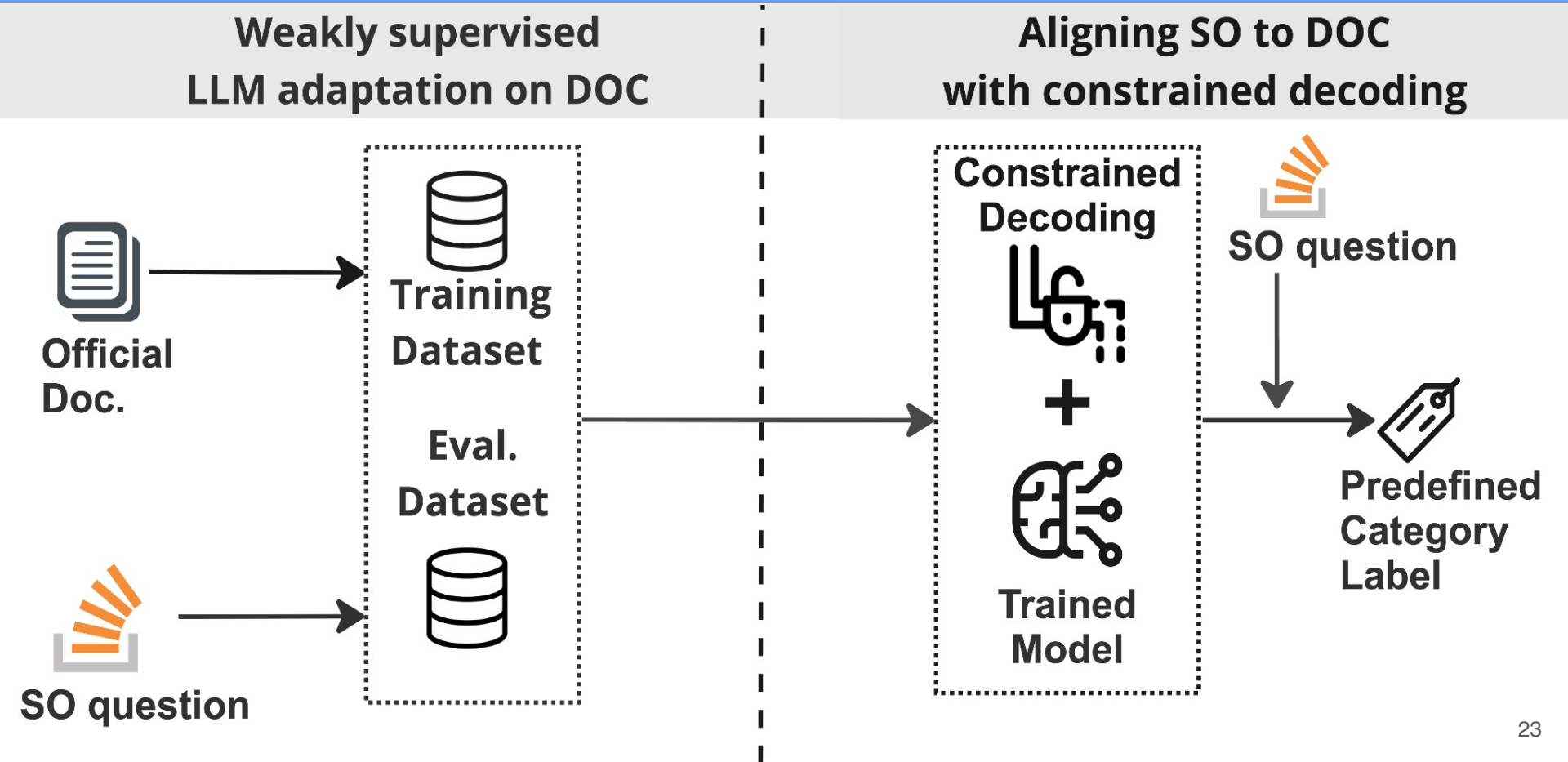
[JSON, API, application, handling, quickstart, ...]

- Produce tokens within the scope
- Stop when output label matches the category.

Expected output

Category: API
Sub-category: JSON Support

Our Approach - DOSA



RESULTS

Evaluation: DOSA vs Conventional IR

Topic / Model	FLASK		PYTHON	
	Category	Subcategory	Category	Subcategory
DOSA (GPT-2)	**P: 0.98 R: 0.99**	**P: 0.99 R: 1.0**	**P: 0.98 R: 0.94**	**P: 0.33 R: 0.41**

P: Precision and R: Recall

Evaluation: DOSA vs Conventional IR

- Conventional Information Retriever (IR) tools struggled to align StackOverflow questions to documentations.
- DOSA outperformed pyserini in both flask and python.

Topic / Model	FLASK		PYTHON	
	Category	Subcategory	Category	Subcategory
DOSA (GPT-2)	**P: 0.98 R: 0.99**	**P: 0.99 R: 1.0**	**P: 0.98 R: 0.94**	**P: 0.33 R: 0.41**
Pyserini	P: 0.26 R: 0.19	P: 0.03 R: 0.10	P: 0.17 R: 0.13	P: 0.08 R: 0.06

P: Precision and R: Recall

More details in the paper

More details in the paper

TABLE I

PRECISION (P) AND RECALL (R) SCORES OF DIFFERENT MODELS. WK – WEAK SUPERVISION, CD – CONSTRAINED DECODING. LLaMA-7B + WK DENOTES THAT THE MODEL WAS IMPLEMENTED WITH LLAMA-7B AND ONLY HAS *weak supervision* BUT NOT *constrained decoding*.

Model \ Topic	Flask		Python	
	Cat.	Sub-Cat.	Cat.	Sub-Cat.
Pyserini	P: 0.26 R: 0.19	P: 0.03 R: 0.10	P: 0.17 R: 0.13	P: 0.08 R: 0.06
***DOSA*(GPT-2)**	**P: 0.98** **R: 0.99**	**P: 0.99** **R: 1.0**	P: 0.98 R: 0.94	P: 0.33 R: 0.41
***DOSA*(LLaMA-7B)**	P: 0.93 R: 0.94	P: 0.97 R: 0.91	**P: 0.94** **R: 0.96**	**P: 0.62** **R: 0.57**
GPT-2 + WK	P: 0.61 R: 0.60	P: 0.66 R: 0.61	P: 0.52 R: 0.55	P: 0.19 R: 0.20
LLaMA-7B + WK	P: 0.48 R: 0.51	P: 0.32 R: 0.39	P: 0.69 R: 0.61	P: 0.19 R: 0.11
GPT-2 + CD	P: 0.30 R: 0.22	P: 0.14 R: 0.15	P: 0.11 R: 0.08	P: 0.09 R: 0.08
LLaMA-7B + CD	P: 0.80 R: 0.81	P: 0.76 R: 0.78	P: 0.81 R: 0.82	P: 0.41 R: 0.40
LLaMA-13B + CD	P: 0.87 R: 0.88	P: 0.82 R: 0.84	P: 0.80 R: 0.83	P: 0.44 R: 0.51

ranking function with a strong performance on zero-shot retrieval tasks [35]. It assigns a relevance score to each document

More details in the paper

TABLE I

PRECISION (P) AND RECALL (R) SCORES OF DIFFERENT M
WEAK SUPERVISION, CD – CONSTRAINED DECODING. LI
DENOTES THAT THE MODEL WAS IMPLEMENTED WITH LL
ONLY HAS *weak supervision* BUT NOT *constrained d*

Model \ Topic	Flask		P
	Cat.	Sub-Cat.	Cat.
Pyserini	P: 0.26 R: 0.19	P: 0.03 R: 0.10	P: 0.17 R: 0.13
***DOSA*(GPT-2)**	**P: 0.98** **R: 0.99**	**P: 0.99** **R: 1.0**	P: 0.98 R: 0.94
***DOSA*(LLaMA-7B)**	P: 0.93 R: 0.94	P: 0.97 R: 0.91	**P: 0.94** **R: 0.96**
GPT-2 + WK	P: 0.61 R: 0.60	P: 0.66 R: 0.61	P: 0.52 R: 0.55
LLaMA-7B + WK	P: 0.48 R: 0.51	P: 0.32 R: 0.39	P: 0.69 R: 0.61
GPT-2 + CD	P: 0.30 R: 0.22	P: 0.14 R: 0.15	P: 0.11 R: 0.08
LLaMA-7B + CD	P: 0.80 R: 0.81	P: 0.76 R: 0.78	P: 0.81 R: 0.82
LLaMA-13B + CD	P: 0.87 R: 0.88	P: 0.82 R: 0.84	P: 0.80 R: 0.83

ranking function with a strong performance on
trieval tasks [35]. It assigns a relevance score to ea

TABLE II

HALLUCINATION RATES WHEN MODEL ONLY HAS WEAK SUPERVISION.

Model \ Topic	Flask		Python	
	Category	Sub-Cat.	Category	Sub-Cat.
GPT-2 + WK	5%	11%	6%	9%
LLaMA-7B + WK	17%	25%	20%	28%

them susceptible to biases and noise present in the training data. Without specifying the scope of the output labels, the fine-tuned models may generate plausible labels but not in direct alignment with *DOC*.

Constrained decoding helps address this issue by incorporating specific constraints (i.e., only output labels within the defined scope) during the decoding process. It also reduces hallucinations by guiding the model to generate labels that are from the list of categories and sub-categories from *DOC*. In

More details in the paper

PRECISION (P) AND RECALL
WEAK SUPERVISION, CD – C
DENOTES THAT THE MODEL
ONLY HAS *weak super*

Model \ Topic	C
Pyserini	P: R:
***DOSA*(GPT-2)**	**P: R:**
***DOSA*(LLaMA-7B)**	P: R:
GPT-2 + WK	P: R:
LLaMA-7B + WK	P: R:
GPT-2 + CD	P: R:
LLaMA-7B + CD	P: R:
LLaMA-13B + CD	P: R:

ranking function with a
trieval tasks [35]. It assig

Aligning Documentation and Q&A Forum through Constrained Decoding with Weak Supervision

Rohith Pudari
University of Toronto
r.pudari@mail.utoronto.ca

Shiyuan Zhou
University of Toronto
shiyuan.zhou@mail.utoronto.ca

Iftekhar Ahmed
University of California, Irvine
iftekha@uci.edu

Zhuyun Dai
Google
zhuyundai@google.com

Shurui Zhou
University of Toronto
shurui.zhou@utoronto.ca

***Abstract*—Stack Overflow (*SO*) is a widely used question-and-answer (Q&A) forum dedicated to software development. It plays a supplementary role to official documentation (*DOC* for short) by offering practical examples and resolving uncertainties. However, the process of simultaneously consulting both the documentation and *SO* posts can be challenging and time-consuming due to their disconnected nature. In this study, we propose *DOSA*, a novel approach to automatically align *SO* and *DOC*, which inject domain-specific knowledge about the *DOC* structure into large language models (LLMs) through weak supervision and constrained decoding, thereby enhancing knowledge retrieval and streamlining task completion during the software development procedure. Our preliminary experiments find that *DOSA* outperforms various widely-used baselines, showing the promise of using generative retrieval models to perform low-resource software engineering tasks.**

***Index Terms*—Stack Overflow, Natural language processing, Constrained Decoding, Weak Supervision**

I. Introduction & Background

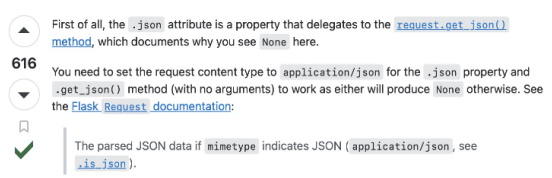


Fig. 1. An example from *SO* where official documents are linked to support a better understanding of the context [12].

However, accomplishing this task poses a difficulty given the scarcity of training data. As far as we know, there is currently no public dataset available that includes the alignment between *SO* questions and the relevant sections in the *DOC*. Creating a large-size custom dataset by using labor-intensive manual labeling is not scalable, making it not feasible to directly and efficiently employ conventional machine learning

EAK SUPERVISION.

Python	
tegory	Sub-Cat.
)	9%
%	28%

nt in the training
output labels, the
labels but not in

issue by incorpo-
labels within the
s. It also reduces
ate labels that are
es from *DOC*. In

Future work

- Amplify the generalizability, performance, and utility of our DOSA approach.

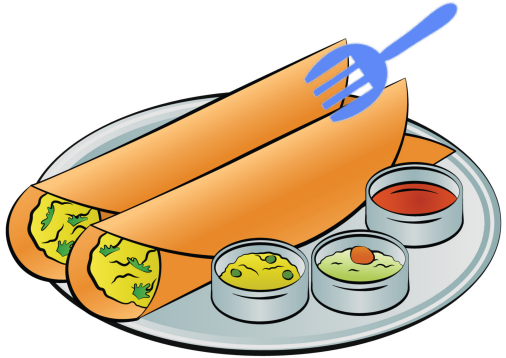
Future work

- Amplify the generalizability, performance, and utility of our DOSA approach.
  - Categorizing diverse types of artifacts
    - Aligning documentation to blog posts

Future work

- Amplify the generalizability, performance, and utility of our DOSA approach.
  - Categorizing diverse types of artifacts
    - Aligning documentation to blog posts
  - Broader domains beyond programming
    - Aligning medical queries with the vast corpus of medical literature

Aligning Documentation and Q&A Forum through Constrained Decoding with Weak Supervision



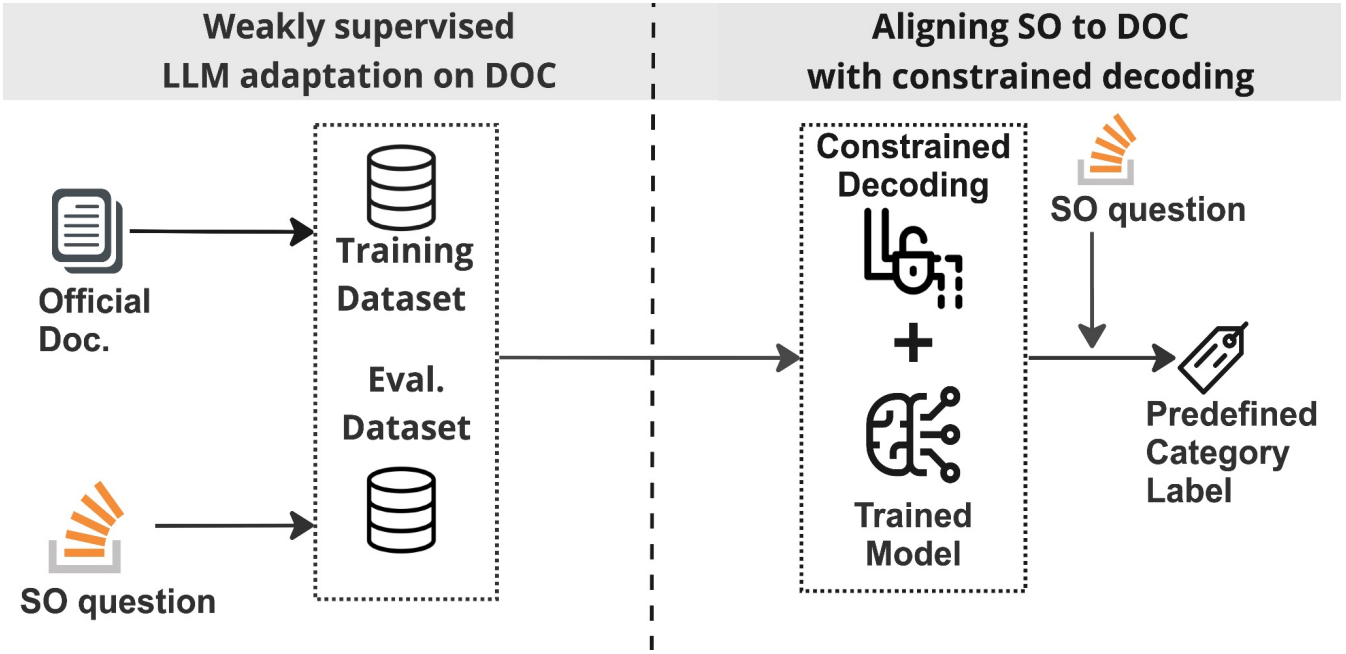
Dosa

The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FORCOLAB

Rohith Pudari
Socials: @rohithpudari
Email: r.pudari@mail.utoronto.ca

Aligning Documentation and Q&A Forum through Constrained Decoding with Weak Supervision



The Edward S. Rogers Sr. Department of Electrical & Computer Engineering
UNIVERSITY OF TORONTO

FORCOLAB

Rohith Pudari
Socials: @rohithpudari
Email: r.pudari@mail.utoronto.ca