# REST API Endpoints Requirements

Instructor: Dr. Charbel Azzi
SYDE 770 - Deployment of Deep Learning Models
University of Waterloo
Due: Friday April 4, 2025

## REST API Endpoints

The endpoints are essential for the basic functionality of deploying and using a YOLO model with RESTful API using a framework of your choice.

- **Inference Endpoint**
  - **Endpoint:** `/predict`
  - **Method:** `POST`
  - **Request Body:** `'image' : file, 'model' : text`
  - **Description:** Triggers the YOLO model to recognize objects in the uploaded image. Includes a body parameter to specify which specific YOLO model to use for making its inference as an optional body parameter. Default model should be used if body parameter is not included.
  - **Response:** JSON object containing the exact keys (you must not modify the names at all in your API) for: 1) predictions key "predictions": detected objects key "label", confidence levels key "confidence",bounding boxes key "bbox", and 2) model used key "model_used". Example JSON with the required key names:

```
{
  "predictions": [
    {
      "label": "timmies",
      "confidence": 0.91,
      "bbox": [42, 58, 172, 310]
    },
    {
      "label": "paper_cup",
      "confidence": 0.88,
      "bbox": [200, 120, 320, 250]
    }
  ],
  "model_used": "model_0"
}
```

- **Health Check Endpoint**
  - **Endpoint:** `/health-status`
  - **Method:** `GET`
  - **Description:** Checks the status of your server.
  - **Response:** JSON object containing the exact keys (you must not modify the names at all in your API) for: 1) status key "Status" with output "OK" or "Healthy" message if the server is functioning correctly, 2) server key "server", and 3) uptime key "uptime". Example JSON with the required key names:

```
{
  "status": "Healthy",
  "server": "FastAPI",
  "uptime": "3 days, 2 hours",
}
```

- **Model Management Endpoint (List)**
  - **Endpoint:** `/management/models`
  - **Method:** `GET`
  - **Description:** Lists all available models
  - **Response:** JSON object listing the models identification key 'available_models'. Example JSON with the required key names:

```
{
  "available_models": ["model_0", "model_1", "model_2"]
}
```

- **Group Info Endpoint**
  - **Endpoint:** `/group-info`
  - **Method:** `GET`
  - **Description:** Lists the group info.
  - **Response:** JSON object listing group info with the following key (you must not modify the names at all in your API): 1) group key "group", and 2) members key "members". Example JSON with the required key names:

```
{
"group": "group3",
"members": ["Alice", "Bob", "Charlie"],
}
```

- **Metrics Endpoint**
  - **Endpoint:** `/metrics`
  - **Method:** `GET`
  - **Description:** Retrieves high-level performance metrics for the deployed models.
  - **Response:** JSON object containing metrics keys (you must not modify the names at all in your API): 1) the request rate per minute key "request_rate_per_minute", 2) the average latency key "avg_latency", 3) the maximum latency "max_latency", and 4) total requests made key "total_requests. Example JSON with the required key names:

```
{
  "request_rate_per_minute": 42,
  "avg_latency_ms": 153.4,
  "max_latency_ms": 312.7,
  "total_requests": 1200
}
```

- **Model Info Endpoint**
  - **Endpoint:** `/management/models/{model}/describe`
  - **Method:** `GET`
  - **Description:** Provides detailed information about a model.
  - **Response:** JSON object with model details info with the following key (you must not modify the names at all in your API): 1) model key "model", 2) training configurations key "config", and 3) date of registration key "date_registered". Example JSON with the required key names:

```
{
  "model": "model_0",
  "config": {
    "input_size": [640, 640],
    "batch_size": 16,
    "confidence_threshold": 0.25
  },
  "date_registered": "2025-03-15"
}
```

- **Change Default Model Endpoint**
  - **Endpoint:** `/management/models/{model}/set-default`

- **Method:** `GET`
- **Description:** Sets the specified model as the default model for future inferences.
- **Response:** Success message indicating that the default model has been updated and the model that was updated too with the following key (you must not modify the names at all in your API): 1) success message key "success", and 2) a message indicating what model was selected to become the default model "default_model". Example JSON with the required key names:

```
{
  "success": true,
  "default\_model": "model_1"
}
```