LLMs on the Fly: Text-to-JSON for Custom API Calling

Miguel Escarda-Fernández[1], Iñigo López-Riobóo-Botana[1], Santiago Barro-Tojeiro[1], Lara Padrón-Cousillas[1], Sonia González-Vázquez[1], Antonio Carreiro-Alonso[1,2] and Pablo Gómez-Area[1,2]

[1] **Centro Tecnológico ITG**, Cantón Grande 9, Planta 3, A Coruña
[2] **Flythings®** Technologies – IoT for Solutions 4.0

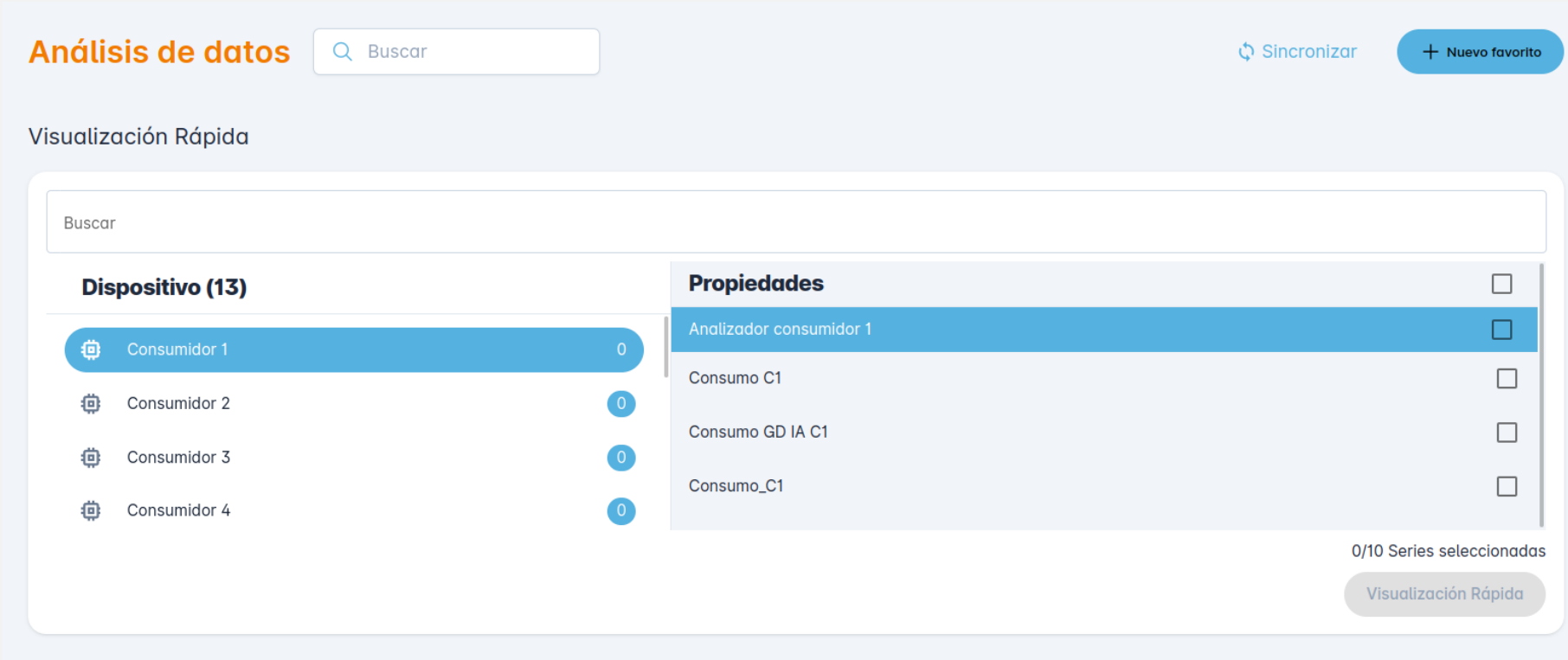
.grow
.transform
.imagine

Índice

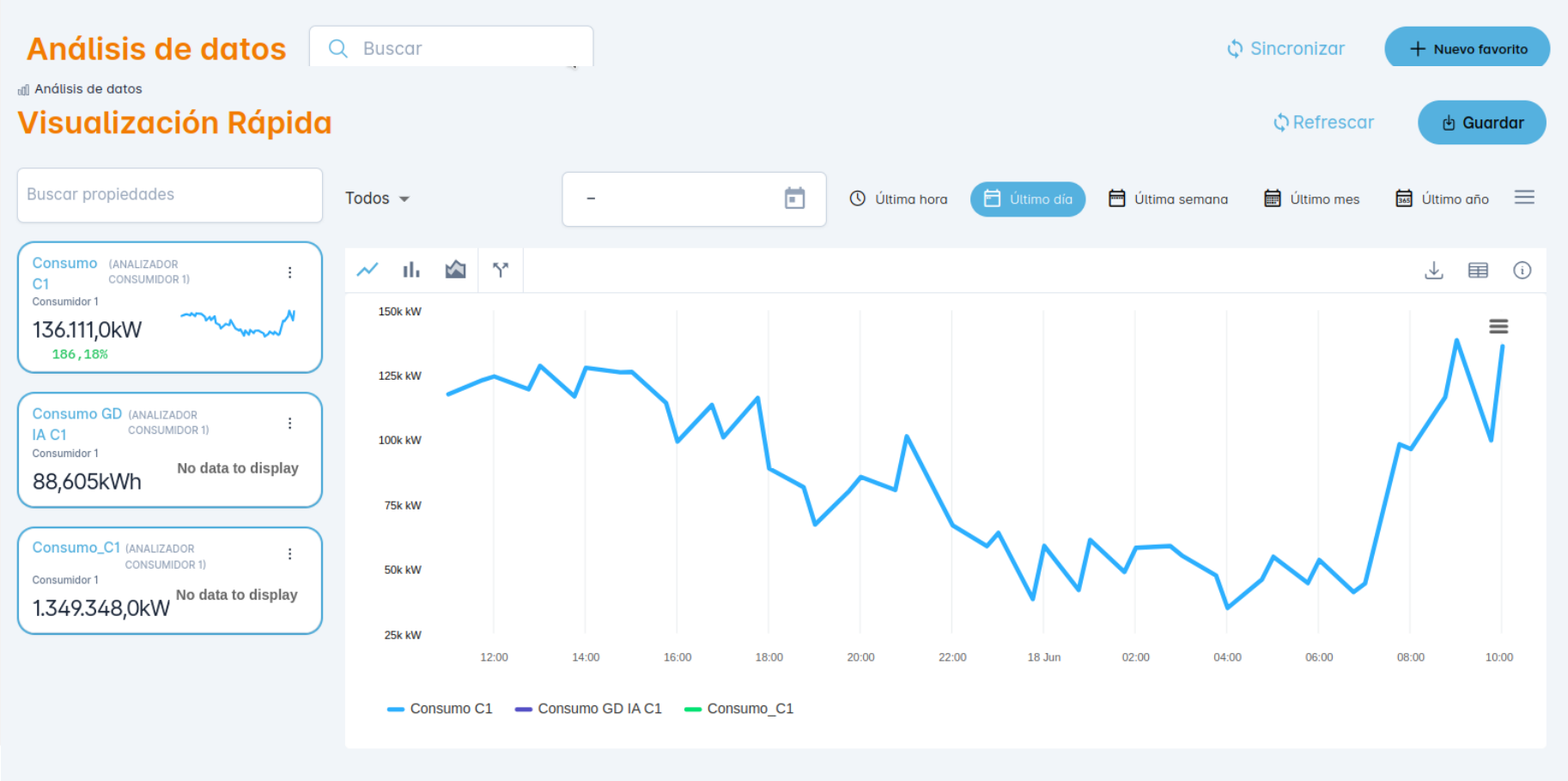
Introducción: Flythings + NLP

- **Flythings es una plataforma IoT** creada por ITG para **visualización, control, monitorización y análisis predictivo** sobre datos de **dispositivos IoT** en fábricas e instalaciones industriales.
- Hasta ahora, la web **Flythings IoT** integraba *dashboards* y visualizaciones mediante **paneles que únicamente podían ser filtrados/ajustados mediante widgets y elementos de UI**
- En este proyecto, **se busca simplificar el uso de la interfaz de la plataforma,** para añadir una **interfaz de lenguaje en formato chat que permita realizar las mismas consultas** de forma más ágil y natural

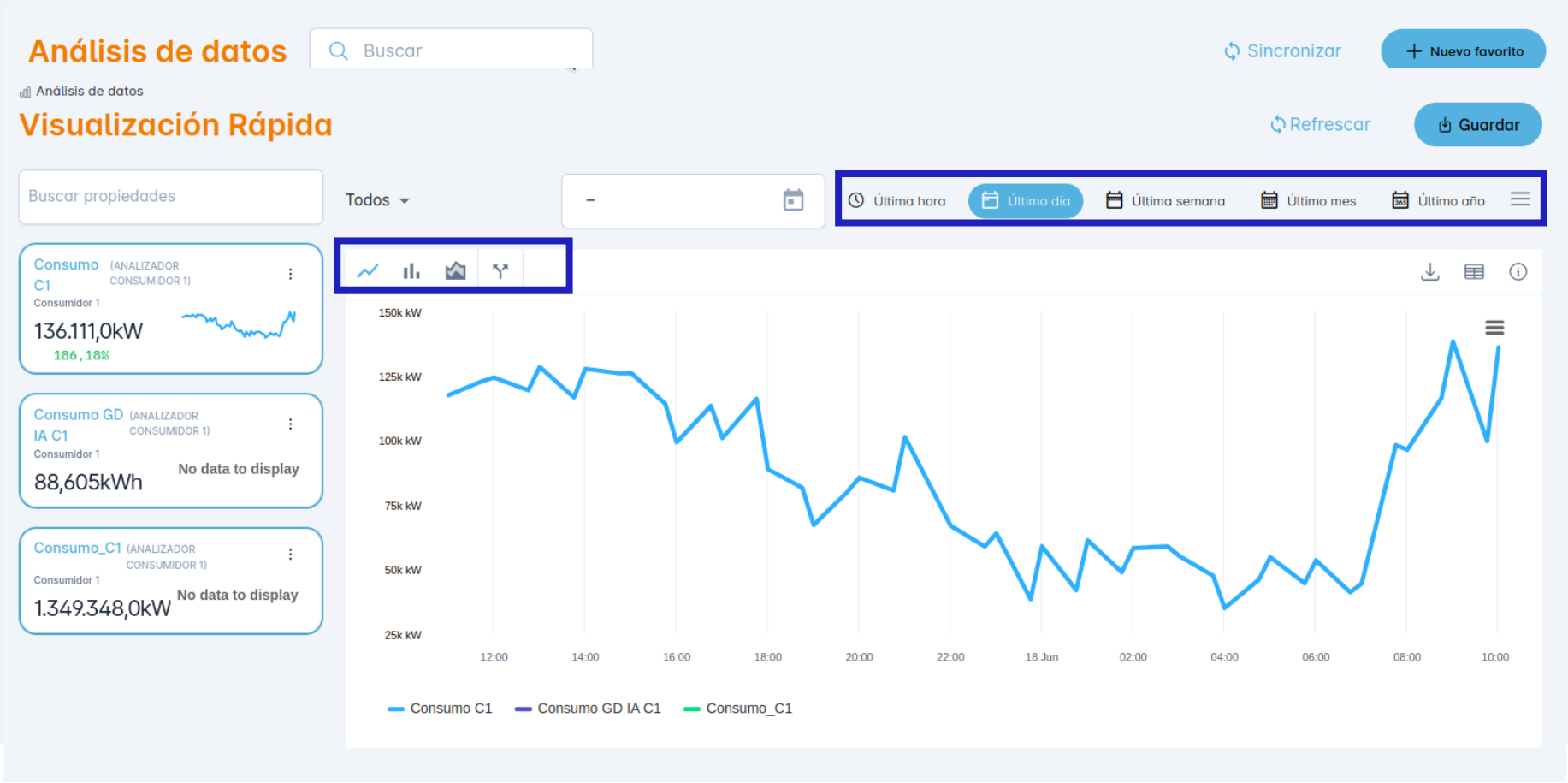
Introducción: Flythings + NLP



Introducción: Flythings + NLP



Introducción: Flythings + NLP



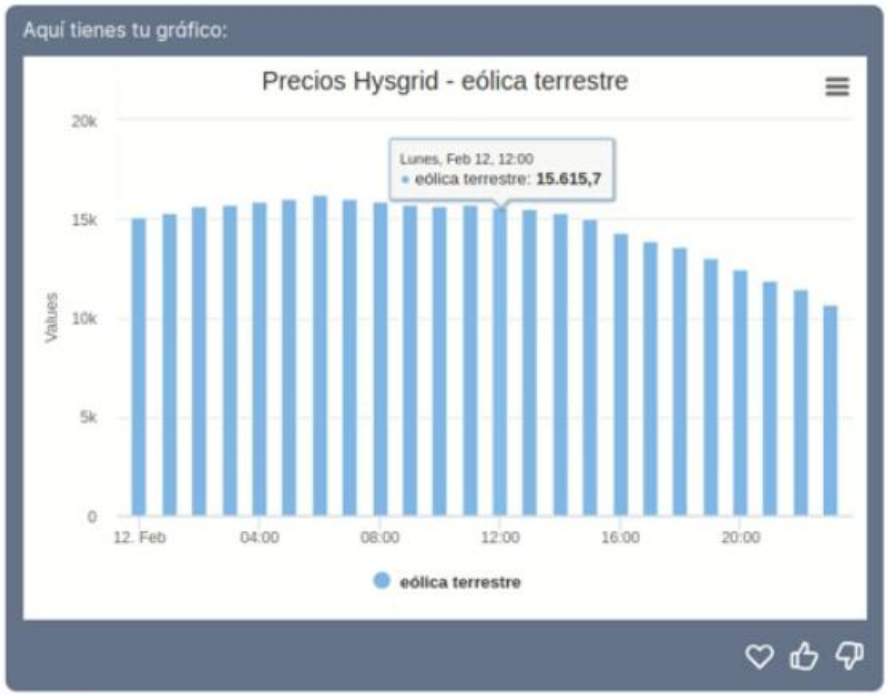
Introducción: Flythings + NLP

16:48

Dame una gráfica de barras de la propiedad eólica terrestre para el dispositivo 'precios Hysgrid'. Dame los resultados: máximos de la semana actual

16:49

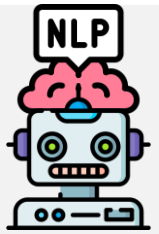
Aquí tienes tu gráfico:



16:49

Introducción: Flythings + NLP

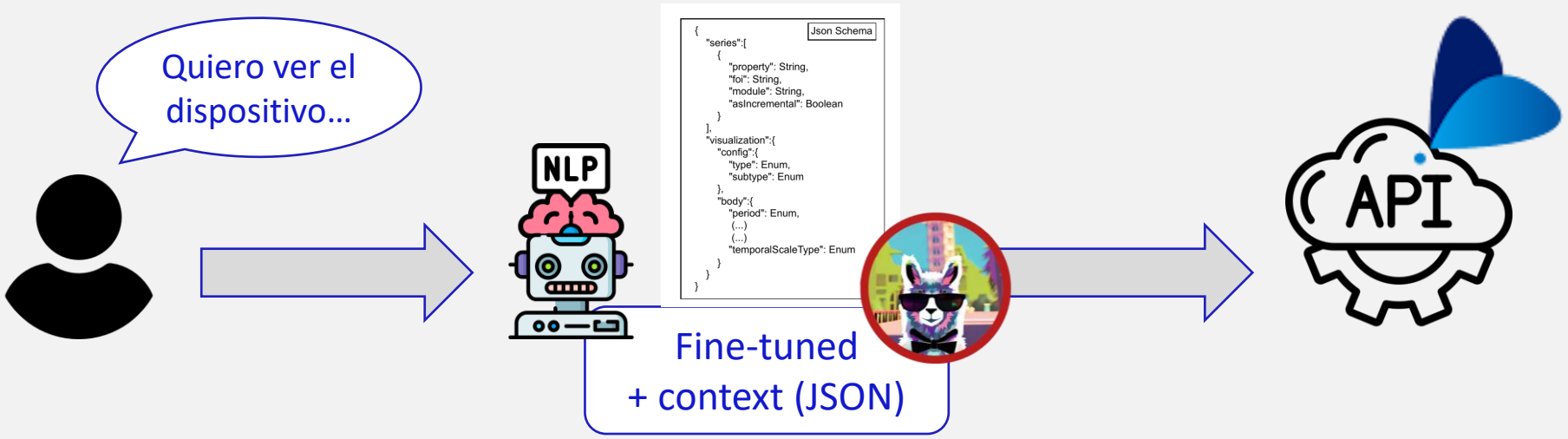
- ¿Qué **ideas** exploramos?
  - ***Instruction fine-tuned* LLMs** (*open source*)



  - Tarea de ***entity recognition*** + ***information extraction*** generalizada a un problema de *Natural Language Generation* (NLG) mediante uso de LLMs
    - **Ajuste de LLM** para generación de **salidas estructurada (JSON) identificando todos los campos** relevantes para **peticiones API REST específicas** de la plataforma Flythings (**ad hoc**)
  - Ideas de ***agentic workflows***
    - **LLM como "puente" entre peticiones de usuario y los objetos JSON adaptados a la plataforma Flythings**. Aquí el "agente" es una implementación de la plataforma que determina la visualización e información a mostrar sobre los dispositivos solicitados (**acción**).

Introducción: Flythings + NLP

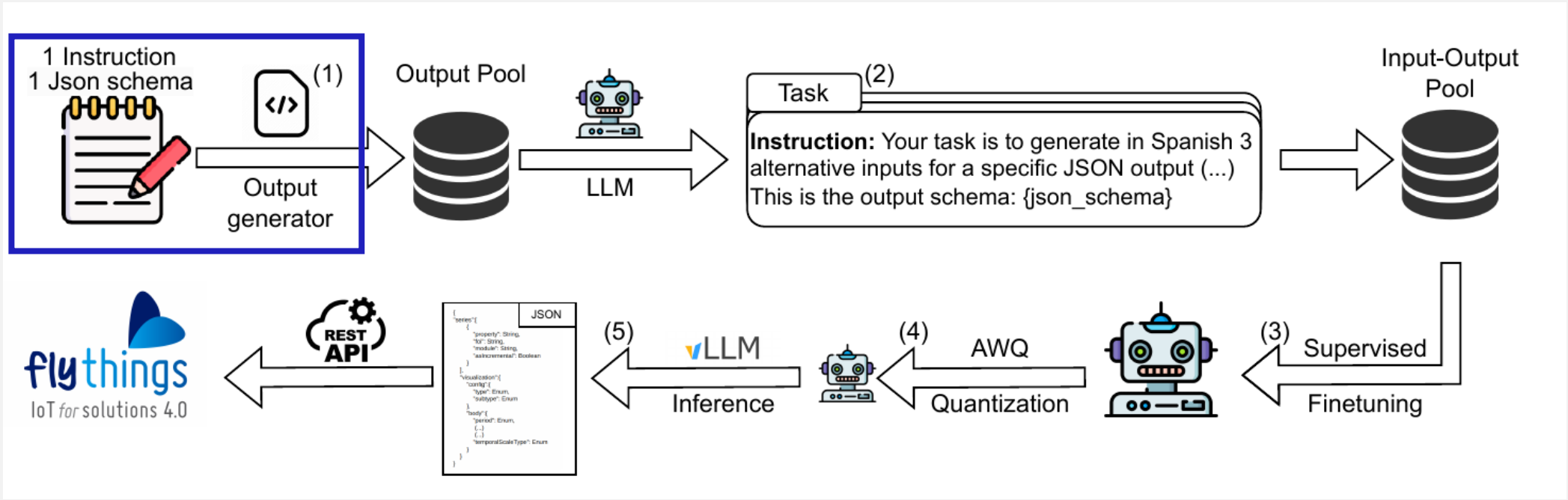
- **Interacción vía lenguaje natural con APIs de consulta de Flythings IoT**



Creación de conjuntos de datos

- La tarea es **muy específica de dominio y de plataforma** (solución **ad hoc** necesaria)
- **Exploración previa de *In-Context Learning*** (ICL) *few/many shot*
  - Problema de **alucinaciones** y **falta de control** de salida
    - Formato **específico de JSON** según schema
    - **Excepciones** y **casos especiales** de la plataforma Flythings
- Uso de ***Supervised Fine-tuning* (SFT)** sobre LLMs
  - Necesarias **secuencias de texto supervisadas** (pares input-ouput)
    - **Input** – fundamentalmente la **expresión del usuario (*queries*)** (hay más implicaciones realmente)
    - **Output** – **JSON** correspondiente que mapee cada entidad/elemento relevante para las peticiones

Creación de conjuntos de datos



Creación de conjuntos de datos

- ***Feedback* inicial** para la salida supervisada
  - El equipo de **Flythings proporciona primeros ejemplos**, con las **correspondencias input *query* – salida JSON**
  - Ejemplos de **cómo deben ser las salidas JSON y su estructura**
    - **Acordamos formato** mediante un **JSON schema** (como *golden rule*)
- **Conjunto semilla** inicial
  - Para **generar salidas JSON**: método basado en **plantillas según reglas del *schema***
    - **Generación de forma pseudoaleatoria**, atendiendo a los valores posibles de cada campo.
    - La mayoría de los campos son enumerados o contienen un grupo de categorías.
- **¿Cómo expresar en lenguaje natural las peticiones de usuario** (input queries) dado un JSON de salida?

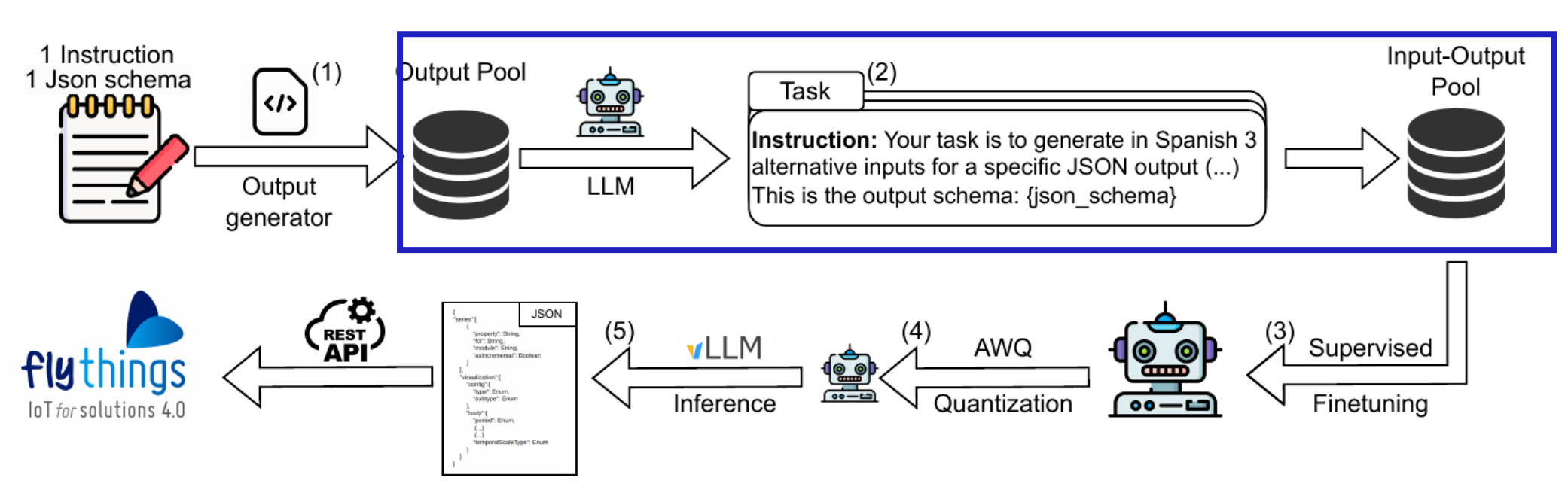
Creación de conjuntos de datos

Json Schema

```
{
   "series":[
      {
         "property": String,
         "foi": String,
         "module": String,
         "asIncremental": Boolean
      }
   ],
   "visualization":{
      "config":{
         "type": Enum,
         "subtype": Enum
      },
      "body":{
         "period": Enum,
          (...)
          (...)
         "temporalScaleType": Enum
      }
   }
}
```

Creación de conjuntos de datos

- **Nos faltan los input (*queries*) de los usuarios**



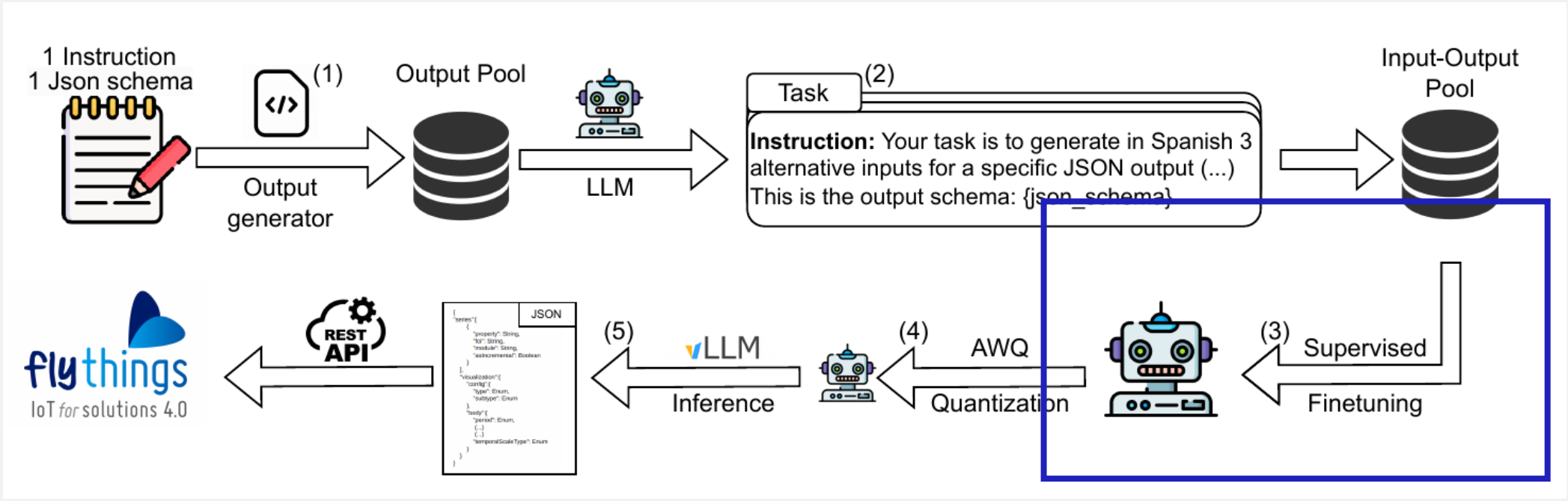
Creación de conjuntos de datos

- ***Data augmentation*** para variabilidad de **input de usuario**
  - Usando LLM vía ***few-shot ICL***
  - **MoE Mixtral-8x7B-Instruct-v0.1**
  - **3 inputs por cada salida** JSON
  - **350-400** muestras

Instruction: Your task is to generate 3 alternative inputs for a specific JSON output. ***{rules_to_follow}***
This is the output schema:
{"series": [{ "property": "tap 2", "foi": "greehouse water", "asIncremental": True }], "visualization": {"config": {"type": "chart", "subtype": "line"}, "body":{ "temporalScale": "DAILY", "temporalScaleType": "CHANGES" }}}

Input1: View the accumulated status changes for tap 2 of the greenhouse water device on a daily graph.
Input2: Observe the daily graph that displays the collective status alterations of tap 2 in the greenhouse watering device.
Input3: Examine the daily chart showing the aggregate changes in the status of greenhouse water device's tap 2.

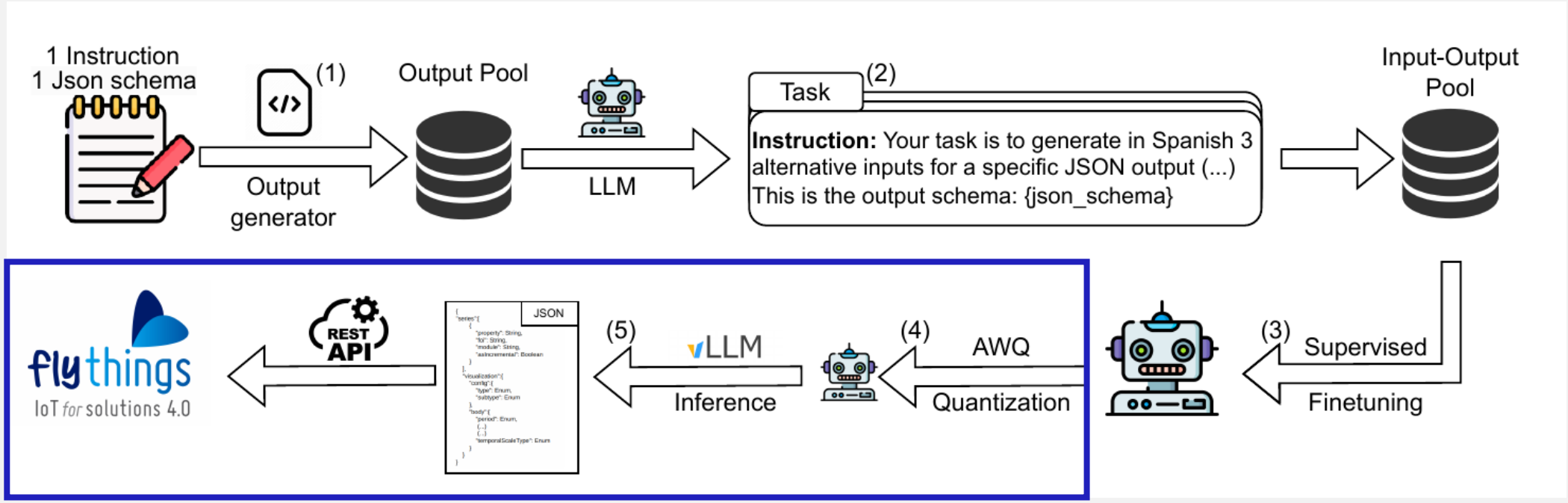
Aprendizaje supervisado en LLMs



Aprendizaje supervisado en LLMs

- SFT optimizado mediante QLoRA
- Partimos de *instruction fine-tuned* *teknium/OpenHermes-2.5-Mistral-7B*
- Datos de la fase previa de *augmentation*

Optimizaciones de inferencia



Optimizaciones de inferencia

- **Reducir uso computacional de recursos *on-premise*** de nuestro *cluster* GPU
  - Técnicas de (post-)**cuantización (AWQ)**
- Necesario **reducir latencias durante inferencia** del modelo
  - **vLLM como motor de inferencia** de LLMs
- Despliegue en servidores propietarios en una RTX A6000 48 GiB GDDR6

Demo: entorno Flythings

https://youtu.be/qHs47rcmpHU

¿En qué puedo ayudarte?

Quiero un gráfico anual de barras para la propiedad Consumo C1 del dispositivo Consumidor 1 del módulo Analizador consumidor 1.

Ver respuesta

Muestra una tabla, registrando el último valor para Consumo C1 del dispositivo Consumidor 1 en el módulo Analizador consumidor 1.

Ver respuesta

Quiero un gráfico de linea para la propiedad Consumo C1 del dispositivo Consumidor 1 del módulo Analizador consumidor 1 del último mes.

Ver respuesta

¿Cuál fue el último valor de Consumo C1 del dispositivo Consumidor 1 con módulo Analizador consumidor 1 del último mes?

Ver respuesta

Escribe tu pregunta aquí...

Limitaciones

- **Primera versión** para la demo (fase temprana de desarrollo)
- **Centramos el *pipeline* exclusivamente** en *data augmentation*, *fine-tuning* y *deployment*
- **Pendiente evaluación exhaustiva** de las salidas
  - ¿*Benchmarking* interno? ¿*Evals* propios? ¿*LLM as a judge o* manual?
- **Robustez a variabilidad de input**
  - **Alineamos distribución de lenguaje a un formato de tarea específico mediante SFT**, **pero no abarcamos la infinidad de formas de expresar queries por parte de usuarios**, lo que puede originar outputs incongruentes o JSON mal formados
  - **Medida de contingencia**: validación de output y, de ser el caso, comunicación de error al usuario

Conclusiones

- **Nueva interfaz basada exclusivamente en lenguaje natural** para consultar los servicios Flythings **sin necesidad de manipular widgets de UI**
- **Pipeline completo** para generación y preparación de **datos, fine-tuning, optimizaciones** de inferencia y **despliegue** de modelo LLM
- Pipeline **generalizable a otras tareas** *text-to-JSON o text-to-API*

Trabajo futuro

- Investigación en **técnicas de *implicit reward modelling*** (DPO, KTO)
  - Añadir **función de *reward* basada en preferencias humanas** (RLHF)
  - Tomar **datos reales mediante *feedback* binario** (*like*/*dislike*) **sobre las respuestas que obtienen los usuarios**
- ***Benchmarking* y *evals* internos** para nuestra tarea *text-to-JSON*
- **Integraciones** futuras en el contexto de ***Virtual Reality* (VR)**
- **Añadir más acciones** además de visualización (gestión dispositivos, alertas…)

Contacto

Iñigo López-Riobóo Botana

@Inigo_LRB ibotana@itg.es

https://flythings.io

info@flythings.es

Cantón Grande 9, Planta 3. 15003. A Coruña
T. +34 981 173 206 · itg@itg.es

¡Gracias!
¿Preguntas?

Centro Tecnológico Nacional

Cantón Grande 9. Planta 3.
15003. A Coruña, España.
T. +34 981 173 206
itg@itg.es - itg.es

.grow
.transform
.imagine