# **Informe Técnico: Variaciones en la Entrada y Salida de Modelos de IA**

## **1. Campo output\_format: Propósito y Funcionamiento**

El campo output\_format en la configuración JSON de un modelo de IA define la forma estructural en la que el modelo devuelve sus salidas (raw\_output) luego de la inferencia. Su objetivo es permitir que el sistema sepa de antemano qué tipo de postprocesamiento o desempaquetado debe aplicar antes de ejecutar el adaptador de salida (output\_adapter).

### **Posibles Valores:**

| **Valor** | **Descripción** |
| --- | --- |
| raw | El modelo devuelve un solo tensor plano con todos los datos ya organizados por detección. Cada fila representa una detección. |
| multihead | El modelo devuelve una lista de múltiples tensores. Cada tensor representa una "cabeza" de salida (por ejemplo: una para cajas, otra para clases, otra para máscaras). Es necesario combinarlos. |
| softmax | El modelo devuelve probabilidades (usualmente clasificación) que deben ser procesadas para obtener la clase más probable. |

Este campo permite que el sistema automatice el preprocesamiento de raw\_output sin depender del framework (ONNX, TFLite, etc.), desempaquetando las salidas en estructuras estandarizadas.

## **2. Informe General de Variaciones desde Inputs hasta Outputs**

### **Sección input**

* width: Ancho en píxeles requerido por el modelo.
* height: Alto en píxeles.
* channels: Cantidad de canales (1 para escala de grises, 3 para color RGB/BGR).
* color\_order: Define el orden de los canales: "RGB" o "BGR".
* normalize: Aplica normalización por canal.
* mean y std: Valores por canal para normalizar (ej. Imagenet).
* scale: Divide los valores por 255 si es true.
* letterbox: Si se aplica resize + padding para conservar aspecto.
* auto\_pad\_color: Color del relleno.
* preserve\_aspect\_ratio: Mantiene proporciones originales.

#### **Subestructura input\_tensor**

* layout: Representación del tensor. Valores:
  + HWC, CHW, NHWC, NCHW
* dtype: Tipo de dato esperado: float32, uint8, int8.
* quantized: Si el modelo está cuantizado.

### **Sección output**

* confidence\_threshold: Umbral para descartar predicciones con baja confianza.
* apply\_nms: Si se aplica supresión de no-máximos (NMS).
* nms\_threshold: Umbral de IoU para el NMS.

#### **Subestructura tensor\_structure**

* box\_format: Formato de las coordenadas de salida. Valores:
  + xyxy, cxcywh, yxyx
* coordinates: Diccionario con los índices de las coordenadas.
* confidence\_index: Índice donde se encuentra la confianza.
* class\_index: Índice donde se encuentra el ID de clase.
* num\_classes: Total de clases detectables.

#### **Subestructura output\_tensor**

* layout: Disposición del tensor (igual que input\_tensor).
* dtype: Tipo de dato de salida: float32, int8, etc.
* output\_format: Estructura de salida: raw, multihead, softmax.

Este documento estandariza todas las formas de entrada y salida que puede tener una IA, y permite adaptarse con precisión a cada una. Cualquier motor de inferencia que se construya sobre este esquema tendrá soporte universal para modelos modernos en diversos frameworks (en teoría).