Detección de contenidos no apropiados.

Centro de moderación de contenidos de Meta

Inaugurado en mayo de 2018

Aproximadamente 2.000 personas en plantilla

Encargado de supervisar contenido para Facebook y Instagram

20%

de los empleados solicitó la baja por depresión

300–500

Post diarios

5 min

De descanso por hora de trabajo

2025

Cerró el centro

500 horas

Por minuto

2500 horas

Por minuto

25%

Presenta síntomas depresivos

50%

Presenta síntomas de ansiedad

1

Moderación de Contenido con IA

De filtros simples a sistemas inteligentes

Evolución de la IA

- 2010: Detección de spam e insultos
- 2015: YouTube Content ID – derechos de autor
- 2018: Nacen los AI Integrity Teams
- 2021: IA identifica formas más sutiles de contenido problemático

Problema: La IA solo detectaba contenido explícito directo, sin comprender el contexto

2

Qué es un LLM?

De detectar texto a generar lenguaje humano

LARGE LANGUAGE MODEL

Evolución de los LLM

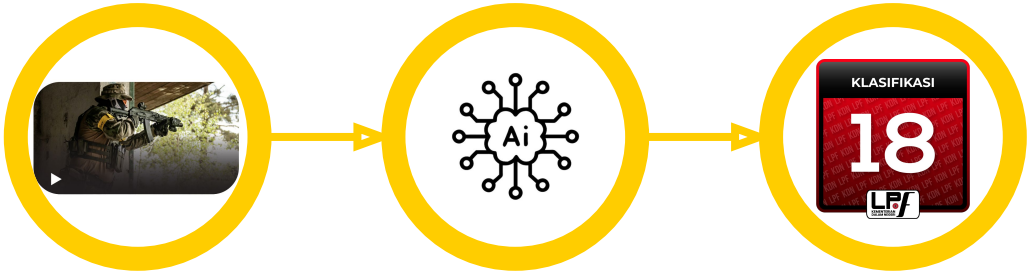
- 2017: Primera versión de un LLM
- 2018: OpenAI lanza ChatGPT-1
- 2019: Llega ChatGPT-2
- 2020: Revolución con ChatGPT-3: “IA que escribe como humano”
- 2022: GPT-3.5 alcanza 100 millones de usuarios en un mes
- Hoy: LLMs avanzados como GPT-4, Claude (Anthropic), LLaMA (Meta) y Gemini (Google)

3

LLM y moderación

De contenidos explícitos

Inicio del proyecto: necesidad, contexto y objetivos



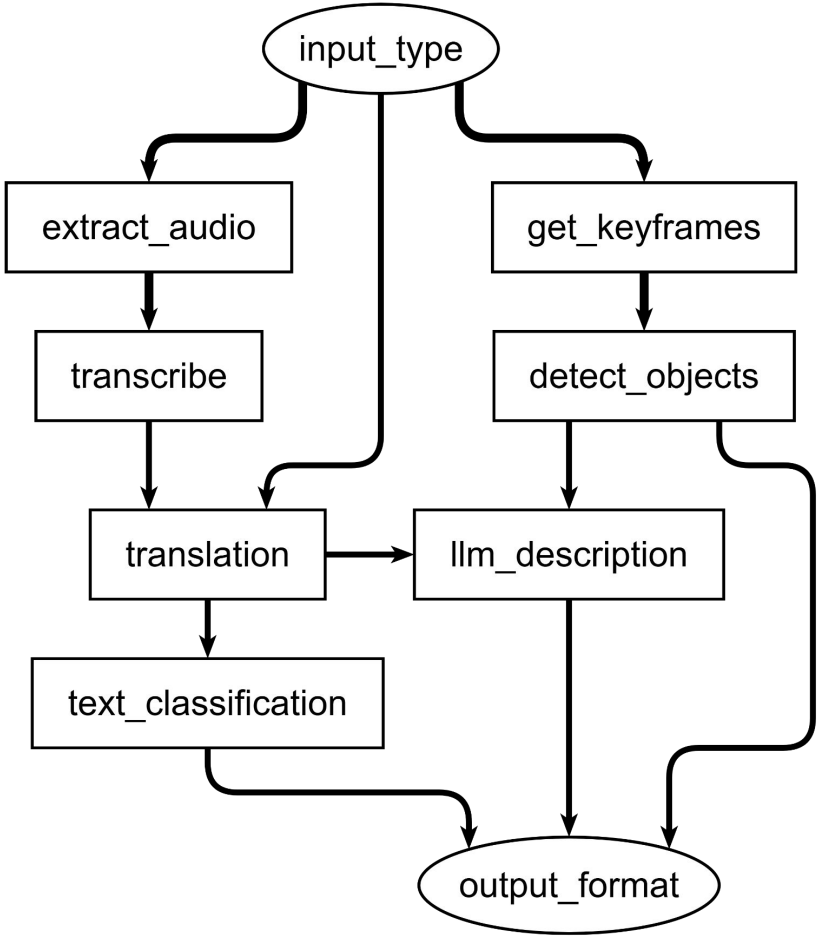
Cómo la IA se vuelve experta

4

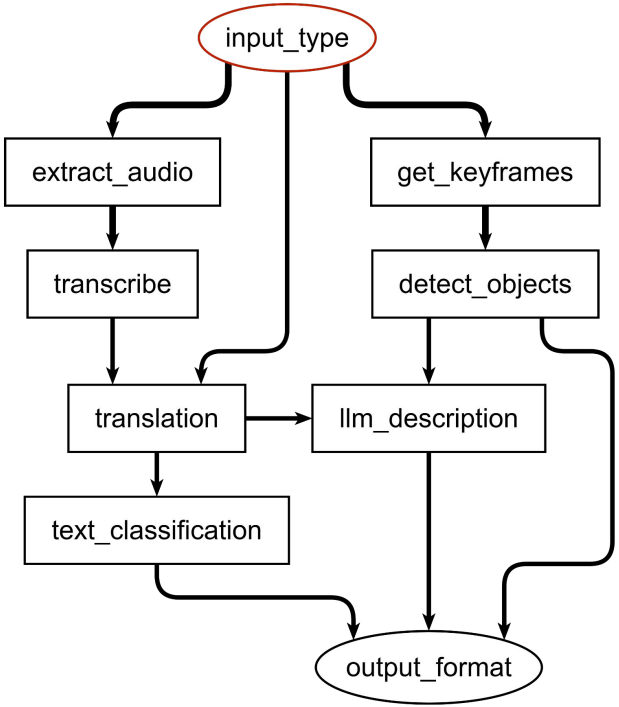
Solución propuesta

Cómo funciona y qué módulos la componen

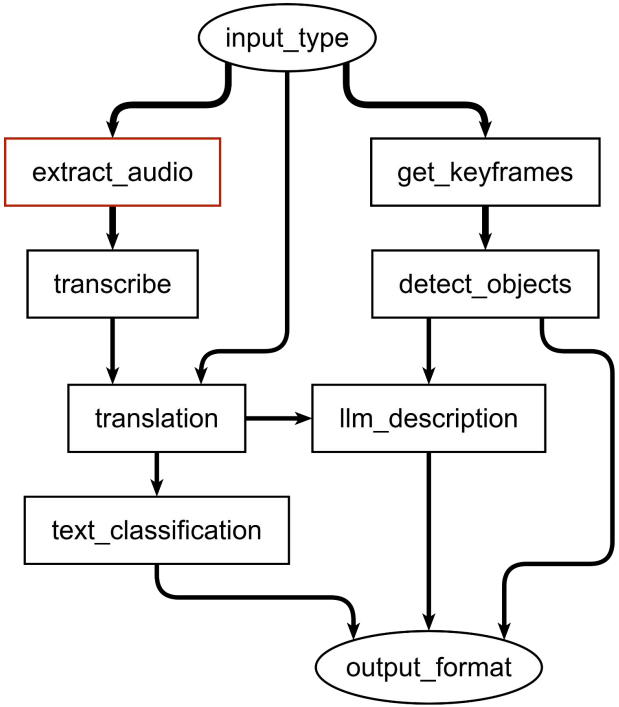
Esquema de la Arquitectura



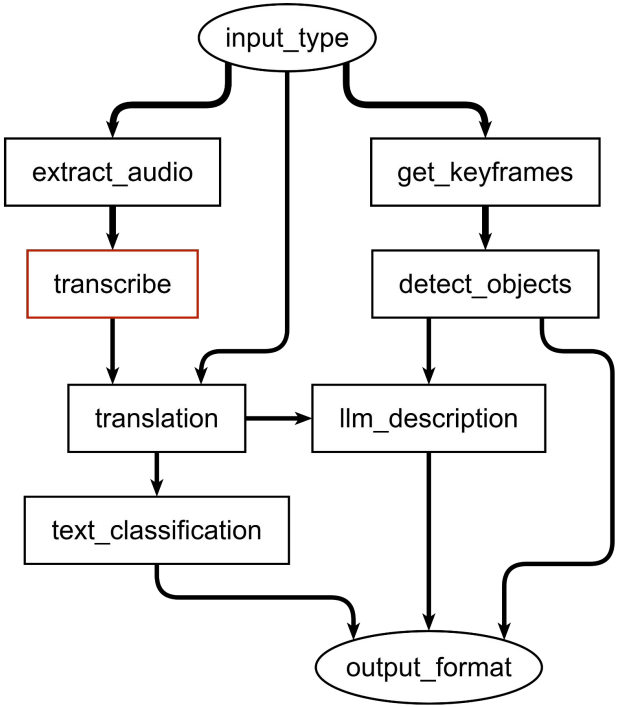
Módulo: Selector de ruta



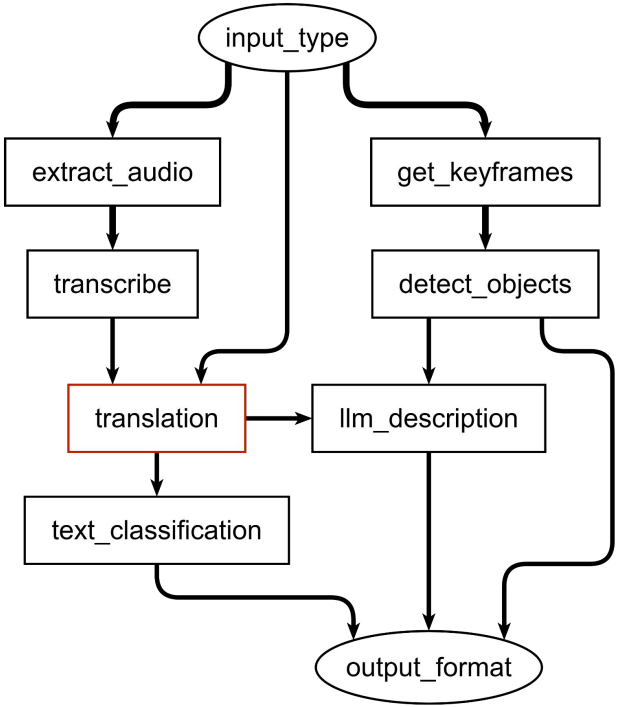
Módulo: Extractor de audio



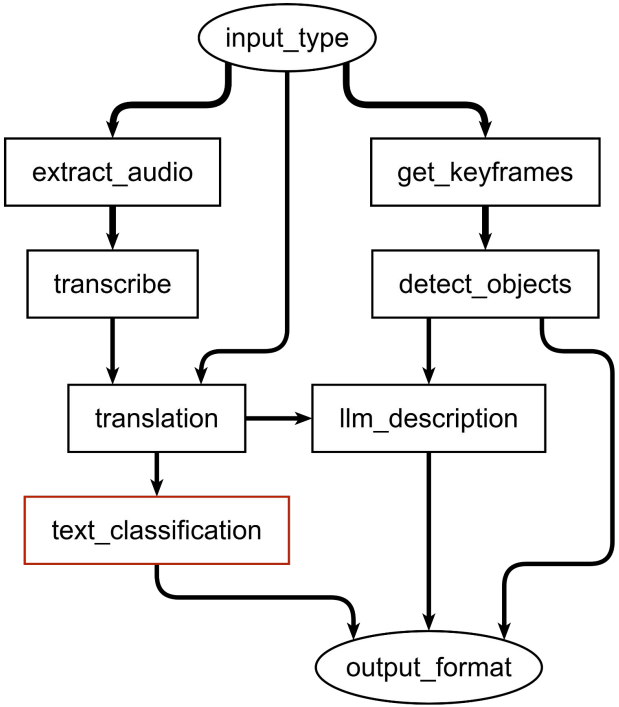
Módulo: Transcripción

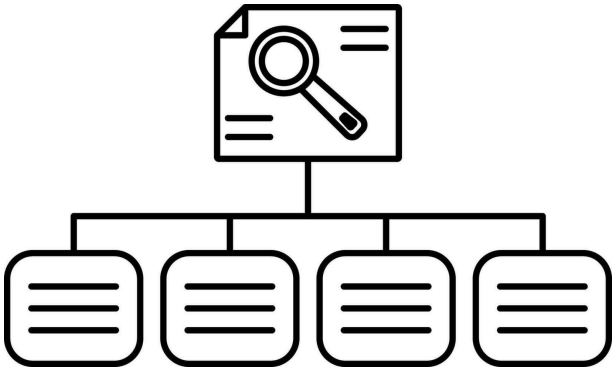


Módulo: Traducción

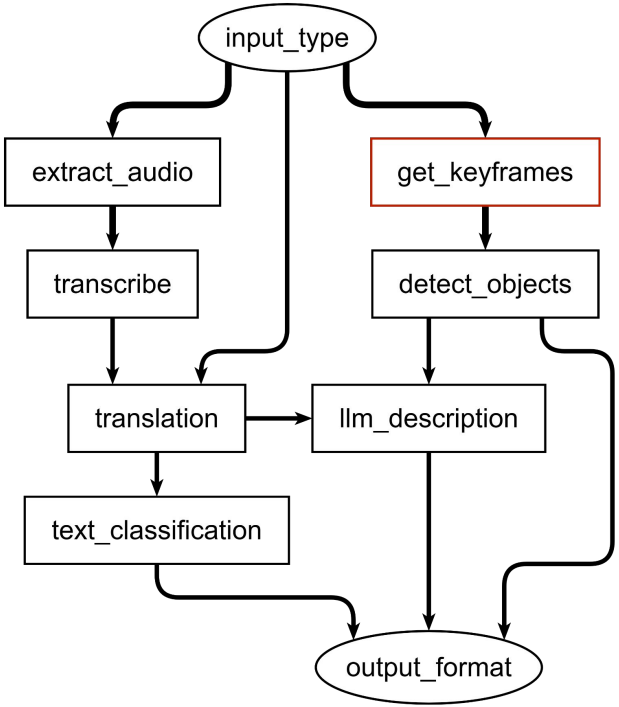


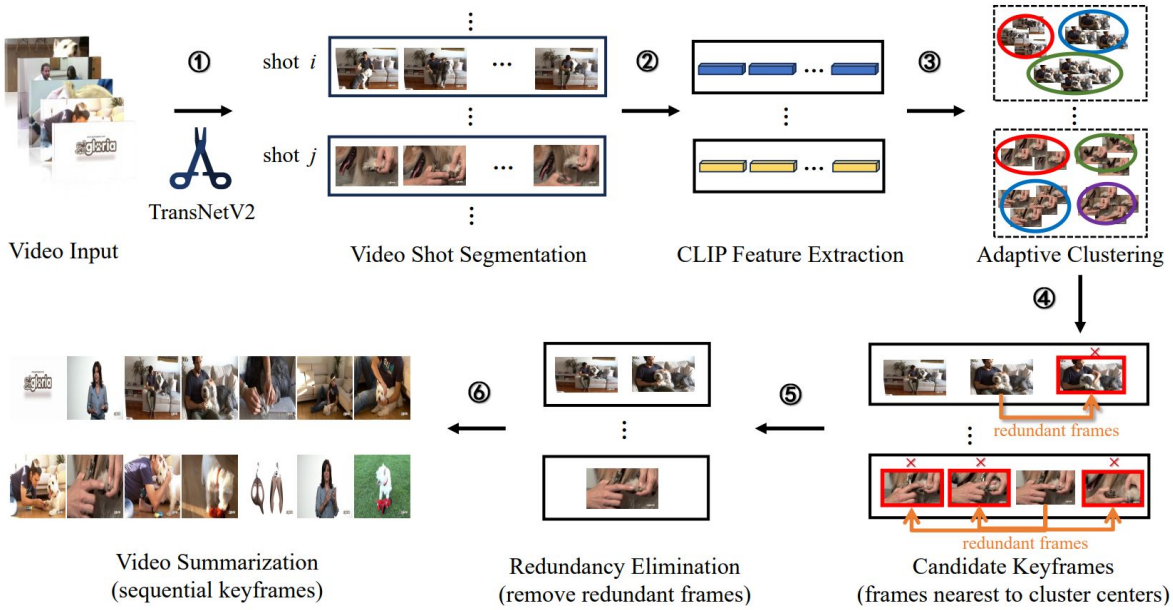
Módulo: Clasificación de transcripción



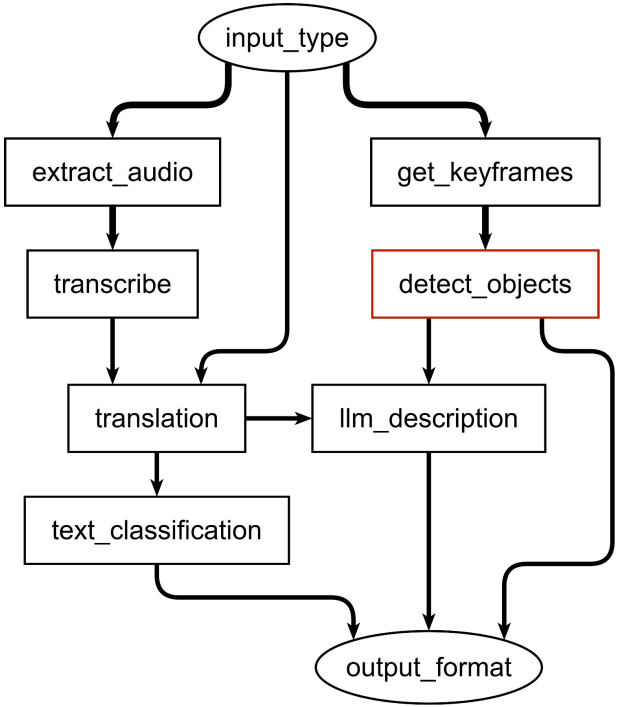


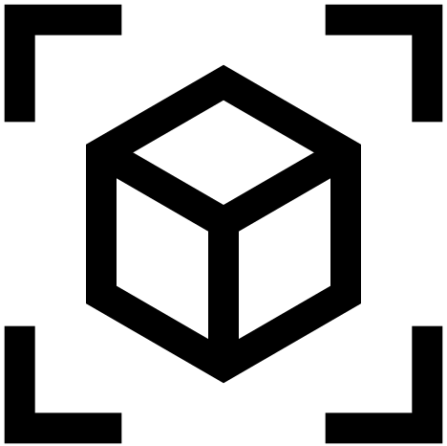
Módulo: Extracción de fotogramas clave



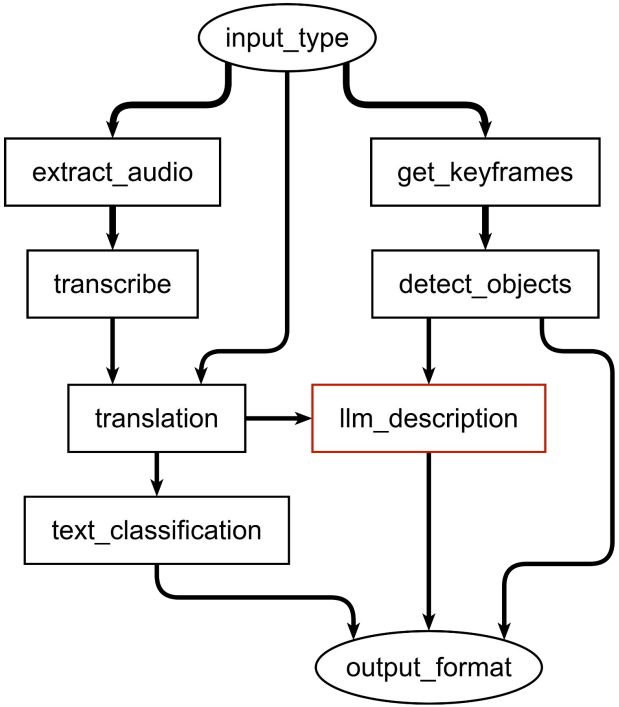


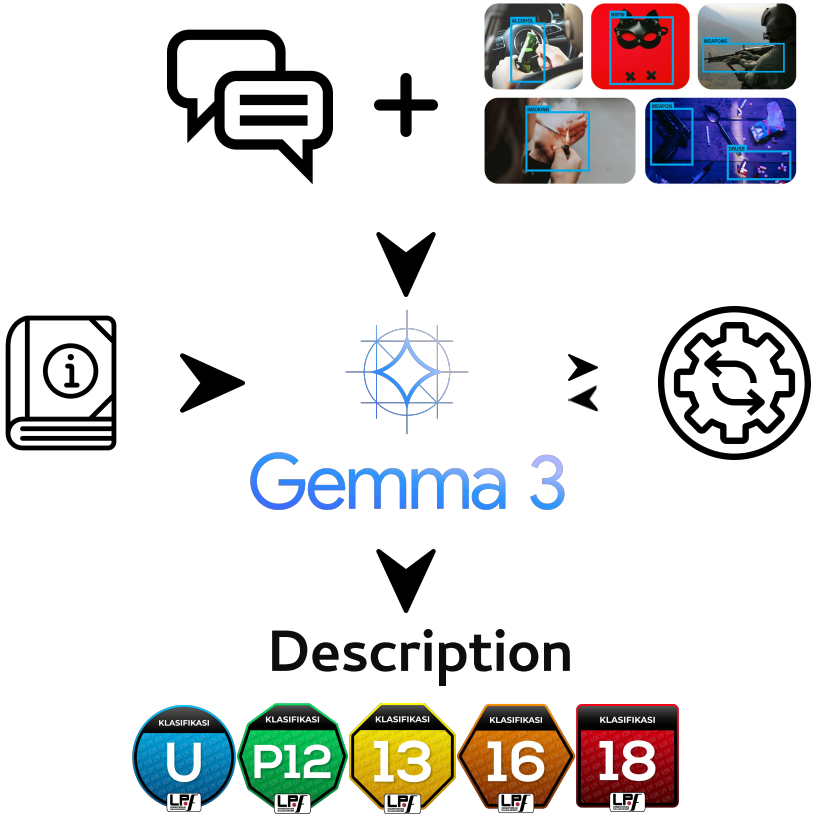
Módulo: Detección de objetos





Módulo: Classificador final





Resultados cuantitativos

Nombre	U (%)	P12 (%)	13 (%)	16 (%)	18 (%)	Correcte Safe (%)	Correcte UnSafe (%)
TikHarm S	95,45	4,55	0	0	0	95,45	83,8
TikHarm UnS	16,2	41,9	2,23	27,38	12,29		
Violence S	91	2	4	2	1	91	92
Violence UnS	8	11	8	67	6		
Porn S	73,27	17,11	1,97	6,22	1,43	73,27	98,96
Porn UnS	1,04	4,17	0,62	10,96	83,21		
Smoking S	92,22	7,78	0	0	0	92,22	100
Smoking UnS	0	32,22	0	66,67	1,11		

Demo técnica

https://nsfw.ugiat.com/

Detección de contenidos no apropiados.

Grácias por vuestra atención. Contacto: guillem.cadevall@ugiat.com