

¿Qué tiene la inteligencia artificial para mi archivo? De la teoría a la práctica.

Procesado de documentos PDF.

Pere Vila Fumás
Julio 2025

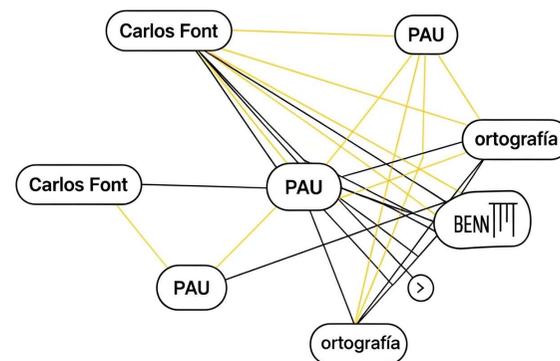
¿Qué es un LLM?

Un LLM (Large Language Model), o Modelo Lingüístico Grande, es una inteligencia artificial entrenada para procesar y generar lenguaje humano.

¿Cómo “piensa” el modelo para darnos una respuesta?

La trampa de la antropomorfización

El archivista estadístico definitivo



Ejemplos sacados del PDF de la entrevista a correctores de la PAU (<https://elpais.com/educacion/2025-06-10/en-la-piel-de-un-corrector-de-la-selectividad-cuando-llevas-20-examenes-corregidos-dices-voy-a-bajar-el-nivel.html>)

- * El mapa no “sabe” que **Carlos Font** es profesor de Historia de España; simplemente ha observado que la cadena «**Carlos Font**» aparece con altísima probabilidad cerca de expresiones como «**nueva PAU**», «**siglo XIX y XX**» y «**penalización ortográfica**».
- * Del mismo modo, el modelo “aprende” que tras la frase «**La correctora de Lengua castellana Teresa Martínez**» es muy probable encontrar términos como «**tildes**», «**penalización máxima de dos puntos**» o la cifra «**0,25 por falta**» .
- * Igualmente, detecta que la secuencia «**entrar en Medicina**» suele ir seguida de la referencia «**más de un 13 sobre 14**», vinculada al estrés de los estudiantes .
- * Si aparece la expresión «**20 exámenes corregidos**», el modelo ve que a menudo co-ocurren frases como «**bajar el nivel de exigencia**» y «**documento de Excel hecho a mano**» .

¿Cómo genera entonces su respuesta?

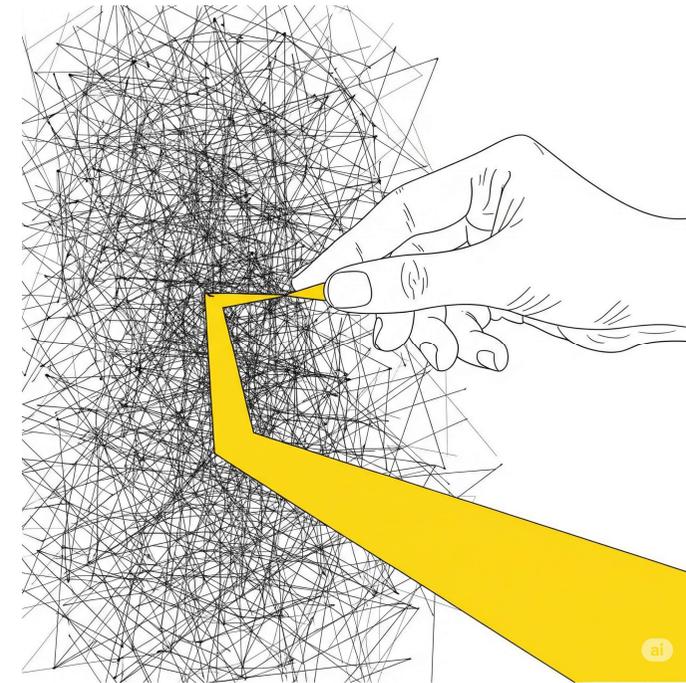
Cuando le pides algo, **el LLM busca en ese mapa** qué combinaciones de palabras tienen mayor probabilidad de seguir a tu instrucción, del mismo modo que un corrector experimentado predice las respuestas que encontrará al abrir un nuevo cuadernillo.

Si preguntas «¿Por qué la ortografía es tan importante en la nueva PAU?», el modelo tenderá a sacar trozos del mapa donde convergen «**penalización ortográfica**», «**hasta dos puntos**» y «**tildes**», porque esas conexiones son las que estadísticamente han aparecido juntas en su inmenso archivo de textos.

Para tus prompts

- **Sé explícito:** Igual que los estudiantes deben citar datos concretos (p. ej., «0,25 por falta»), tu prompt debería proporcionar detalles claros.
- **Contextualiza:** Si quieres que aparezca *Carlos Font*, menciona «Historia de España» o «nueva PAU» para activar esa zona del mapa.
- **Evita ambigüedades:** «Notas altas» puede llevar al modelo a “pensar” en un 10 o en un 13/14; especifica la escala.

Así, comprenderás que un LLM no “razona” como un humano, sino que **asocia patrones** basados en la frecuencia con la que los textos —como la entrevista sobre la Selectividad— relacionan ideas, nombres y cifras.

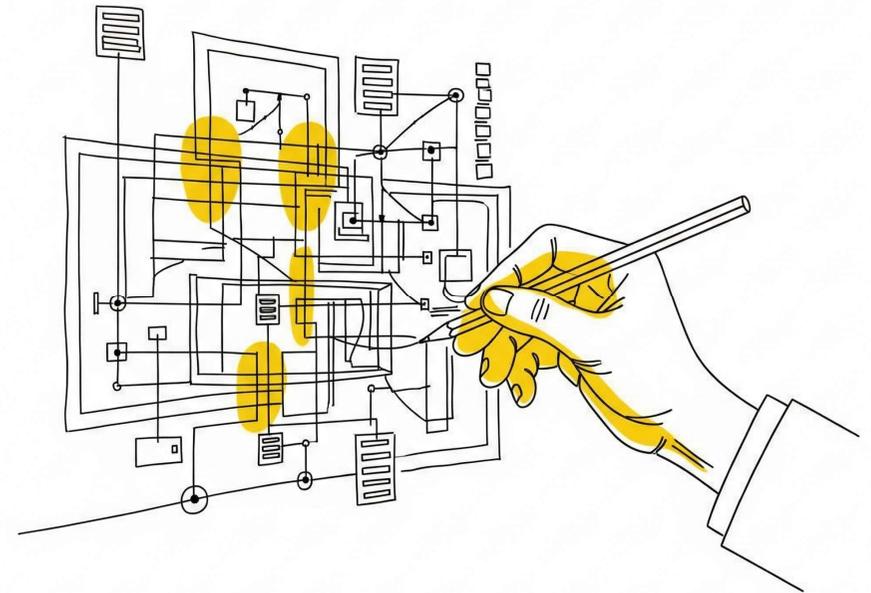


¿Por qué es esto una **Habilidad Clave para la Documentación?**

Entender este mecanismo nos revela por qué el **prompting** es tan poderoso y por qué es una habilidad técnica, no un simple acto de "charlar" con una máquina.

- 1. Nosotros definimos el punto de partida:** Nuestro prompt es el texto inicial que le damos al modelo. Si ese texto es claro, específico y rico en contexto, forzamos al LLM a empezar a predecir desde un punto muy concreto y relevante de su "mapa asociativo".
- 2. Guiamos la predicción:** Al darle un rol ("Actúa como un documentalista experto"), unas reglas ("Si no encuentras un dato, usa N/A") y un formato ("Dame la salida en JSON"), no le estamos dando órdenes a una persona, le estamos dando potentísimas pistas estadísticas. Estamos acotando radicalmente el universo de posibles respuestas, guiando su motor de predicción hacia el único resultado que nos interesa: los datos estructurados y precisos.
- 3. El documentalista como arquitecto de la consulta:** Nuestro trabajo ya no es solo leer y catalogar manualmente. Ahora, nuestro rol evoluciona a ser **arquitectos de la consulta**. Usamos nuestro conocimiento experto para diseñar un prompt que extraiga la información de forma automática, consistente y a una escala imposible de alcanzar manualmente.

En resumen, un LLM no piensa, **asocia**. Y el arte del prompting engineering para documentalistas consiste en construir el estímulo perfecto para que esas asociaciones nos devuelvan, de manera predecible y estructurada, la información exacta que necesitamos.



Prompt Básico, extracción información de un PDF

Extrae la siguiente información del artículo de prensa que te proporciono adjunto:

- Título del artículo.
- Breve Sumario
- Un resumen del contenido.
- Medio de Comunicación.
- Fecha de publicación.
- Número de paginas
- Pagina núm.
- Descripción física
- Nombre del autor o periodista.
- Categoría de la información
- Temas principales que trata.
- Conceptos y palabras clave
- Personas y organizaciones importantes mencionadas.

Resultado con Deep Seek.

****Título del artículo:**** "Cuando llevas 20 exámenes corregidos, dices: 'Voy a bajar el nivel'"

****Breve Sumario:**** El artículo aborda las dificultades que enfrentan los estudiantes y los profesores correctores con el nuevo formato de la Prueba de Acceso a la Universidad (PAU), destacando cambios en la evaluación, el estrés de los alumnos y los desafíos en la corrección.

****Resumen del contenido:**** El artículo explora las impresiones de varios profesores que corrigen los exámenes de la nueva PAU. Se menciona que el nuevo formato es más exigente, ya que cubre más temario y penaliza más duramente los errores ortográficos. Los correctores destacan el estrés de los estudiantes, especialmente aquellos que aspiran a carreras competitivas, y cómo ajustan sus criterios de evaluación para ser justos. También se señala que los alumnos tienen dificultades para aplicar la teoría en preguntas competenciales.

****Medio de Comunicación:**** EL PAÍS

****Fecha de publicación:**** Martes, 10 de junio de 2025

****Número de páginas:**** 1 (página 31)

****Descripción física:**** Artículo de prensa escrito, publicado en la sección "Segunda" del periódico.

****Nombre del autor o periodista:**** Ignacio Zafra

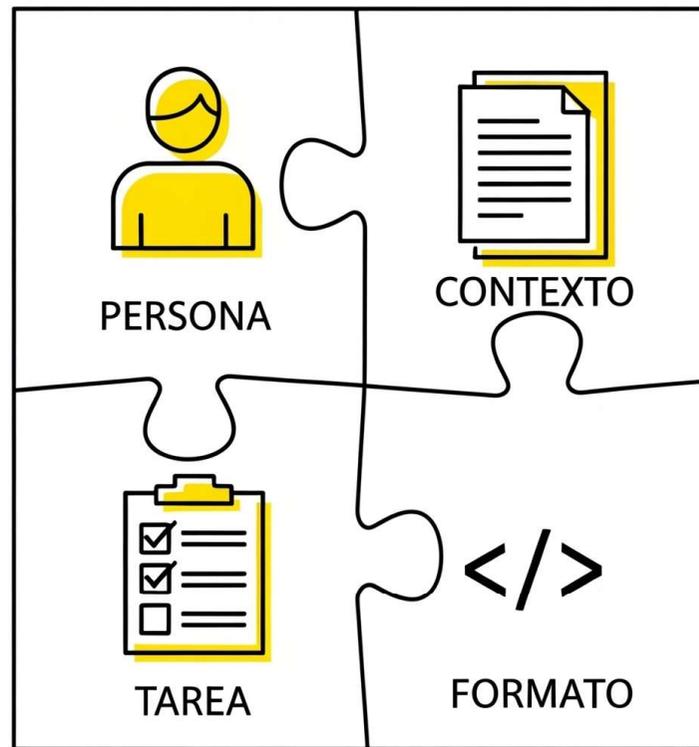
****Categoría de la información:**** Educación / Sociedad

****Temas principales que trata:**** - Nuevo formato de la Prueba de Acceso a la Universidad (PAU): - Dificultades de los estudiantes y correctores.; - Estrés y presión académica en los alumnos. ; - Cambios en la evaluación y penalización ortográfica. ; - Enfoque competencial en los exámenes.

****Conceptos y palabras clave:**** Selectividad, PAU, Bachillerato, corrección de exámenes, estrés académico, notas de corte, ortografía, competencias, educación secundaria, universidad.

****Personas y organizaciones importantes mencionadas:**** - ****Carlos Font:**** Profesor de Historia de España en Castellón.; - ****José Manuel Jiménez:**** Coordinador del ejercicio de Filosofía en Jaén y director de instituto.; - ****Carlos Javier Montejano:**** Profesor de Lengua en Albacete. ; - ****Teresa Martínez:**** Correctora de Lengua castellana y Literatura en Murcia.; - ****Isabel Ruso:**** Correctora de Historia y presidenta de la asociación de directores de institutos públicos de Galicia. ; - ****Incluseño:**** Profesora de Física y Química en Guadarrama, Madrid. ; - ****Ministerio de Educación:**** Mencionado por su papel en la reforma de la PAU.

La Anatomía de un Prompt Eficaz



Prompt Mejorado

Persona: Actúa como un documentalista experto y meticuloso del centro de documentación de un importante periódico nacional.

Contexto: Tu tarea es analizar, catalogar y extraer metadatos clave de un artículo de periódico que te voy a proporcionar. La precisión y la fidelidad a la fuente son cruciales. La fecha de análisis es la fecha de hoy: 26 de junio de 2025.

Tarea: Lee el siguiente texto y extrae la información solicitada. Sigue estas reglas:

1. No inventes información. Si un campo no aparece explícitamente en el texto, déjalo como "N/A".
2. El resumen debe ser objetivo y conciso (máximo 100 palabras).
3. El resumen ampliado debe ser más detallado (entre 200 y 300 palabras), capturando los argumentos principales.
4. Los temas deben ser palabras clave que representen el contenido del artículo.
5. Las personas y entidades nombradas deben ser extraídas tal y como aparecen.

Formato de Salida: Proporciona la salida exclusivamente en formato JSON válido, utilizando las siguientes claves:

```
{  
  "titulo": "...",  
  "resumen_sumario": "...",  
  "resumen_ampliado": "...",  
  "fecha_publicacion": "...",  
  "unidades_paginas": "...",  
  "pagina": "...",  
  "autor": "...",  
  "forma": "Artículo de opinión, Noticia, Entrevista, etc.",  
  "categoria": "Sección del periódico, ej: Internacional, Nacional, Cultura",  
  "temas": ["tema1", "tema2", "tema3"],  
  "personas": ["nombre1", "nombre2"],  
  "fecha_analisis": "26/06/2025",  
  "entidades_nombradas": ["entidad1", "entidad2"]  
}
```

Profesores analizan dificultades de Selectividad



json

Copy Download

```
{
  "titulo": "Seis docentes que est\u00e1n evaluando los ejercicios de la Selectividad explican las dificultades que ha afrontado el alumnado con el nuevo formato de la prueba",
  "resumen_sumario": "Profesores que corrigen los ex\u00e1menes de la nueva Selectividad destacan las dificultades de los estudiantes con el formato competencial y la penalizaci\u00f3n ortogr\u00e1fica homogeneizada. Notas podr\u00edan bajar debido a estos cambios.",
  "resumen_ampliado": "El art\u00edculo presenta las impresiones de varios profesores que est\u00e1n corrigiendo los ex\u00e1menes de la nueva Prueba de Acceso a la Universidad (PAU). Destacan que los estudiantes enfrentan dificultades con el nuevo formato competencial, que requiere aplicar conocimientos en lugar de solo reproducirlos. Adem\u00e1s, la homogeneizaci\u00f3n de la penalizaci\u00f3n por faltas ortogr\u00e1ficas en toda Espa\u00f1a podr\u00eda afectar las notas, especialmente en asignaturas que requieren textos largos. Algunos correctores mencionan que los ex\u00e1menes parecen m\u00e1s asequibles este a\u00f1o para facilitar la transici\u00f3n, pero los alumnos a\u00fan luchan con la aplicaci\u00f3n pr\u00e1ctica de la teor\u00eda. Tambi\u00e9n se aborda el estr\u00e9s y la presi\u00f3n que sufren los estudiantes debido a las altas notas de corte para carreras como Medicina o ingenier\u00edas.",
  "fecha_publicacion": "10/06/2025",
  "unidades_paginas": "N/A",
  "pagina": "31",
  "autor": "IGNACIO ZAFRA",
  "forma": "Noticia",
  "categoria": "Nacional",
  "temas": ["Selectividad", "PAU", "educaci\u00f3n", "ortograf\u00eda", "estr\u00e9s acad\u00e9mico"],
  "personas": ["Carlos Font", "Jos\u00e9 Manuel Jim\u00e9nez", "Carlos Javier Montejano", "Teresa Mart\u00ednez", "Isabel Ruso"],
  "fecha_analisis": "26/06/2025",
  "entidades_nombradas": ["Ministerio de Educaci\u00f3n", "universidad p\u00fablica"]
}
```



Prompt Mejorado, con ejemplo.

PERSONA

Eres un sistema automatizado de catalogación documental (DCAS), desarrollado para un centro de documentación de prensa.

TAREA

Tu función es procesar archivos PDF, analizar su contenido textual y extraer metadatos con la máxima precisión en un formato estructurado y consistente.

REGLAS

Debes seguir estas 5 reglas de forma rigurosa en todas tus respuestas:

1. **Fidelidad a la Fuente:** Procesa el archivo PDF que el usuario suba. Basa tu extracción EXCLUSIVAMENTE en el contenido textual del documento.
2. **Cero Alucinaciones:** No infieras ni añadas información que no esté presente explícitamente en el texto extraído del PDF.
3. **Manejo Estricto de Datos Nulos:** Si un campo no se puede rellenar desde la fuente, utiliza el valor `null` sin excepción.
4. **Inclusión de Metadatos del Proceso:** La `fecha_analisis` siempre debe ser la fecha actual (`26 de junio de 2025`).
5. **Garantía de Formato y Estructura:** La salida DEBE ser únicamente un objeto JSON válido, sin texto introductorio ni conclusiones, siguiendo el formato del ejemplo guía.

EJEMPLO GUÍA (FEW-SHOT PROMPTING)

A continuación tienes un ejemplo perfecto de cómo procesar un texto y cuál debe ser tu salida. Aprende de este patrón.

Ejemplo de Entrada###: en el PDF adjunto sobre la EBAU

Ejemplo de Salida (JSON correspondiente): ""json

```
{ "titulo": "Cuando llevas 20 exámenes corregidos, dices: 'Voy a bajar el nivel'",  
  "resumen_sumario": "Docentes que corrigen la nueva Selectividad (PAU) explican las dificultades del alumnado con el formato competencial y la penalización por ortografía, anticipando una posible bajada en las calificaciones.",  
  "resumen_ampliado": "El artículo recoge las impresiones de varios profesores que están evaluando la nueva Prueba de Acceso a la Universidad (PAU). Docentes como Carlos Font, profesor de Historia, creen que las notas bajarán debido a que el nuevo formato exige el estudio del temario completo y a la homogeneización de la penalización por faltas de ortografía, que puede restar hasta un punto. Teresa Martínez, correctora de Lengua, lamenta tener que descontar hasta dos puntos en exámenes de calidad por fallos de acentuación. El nuevo enfoque competencial, que requiere aplicar conocimientos en lugar de solo reproducirlos, también presenta un desafío para los estudiantes. Algunos profesores perciben que los examinadores han intentado compensar estas dificultades bajando el nivel de las preguntas más complejas."  
  , "fecha_publicacion": "2025-06-10", "unidades_paginas": 1, "pagina": 31, "autor": "IGNACIO ZAFRA", "forma": "Reportaje", "categoria": "Sociedad", "temas": ["Selectividad", "PAU", "EBAU", "Educación", "Bachillerato", "Evaluación", "Ortografía", "Sistema educativo"], "personas": ["Carlos Font", "Teresa Martínez", "Carlos Javier Montejano", "Isabel Ruso", "Irene Nieto", "José Manuel Jiménez", "Federico García Lorca"], "fecha_analisis": "2025-06-28", "entidades_nombradas": ["EL PAIS", "Prueba de Acceso a la Universidad (PAU)", "Ministerio de Educación", "Galicia", "Jaén"]} ""
```

FORMATO MARKDOWN. Prompt Mejorado, con ejemplo.

Sintaxis de formato Markdown.

Títulos y Encabezados: Símbolos de almohadilla (#) para crear títulos. Por ejemplo, ## PERSONA ## y ### Ejemplo de Entrada ### son equivalentes a encabezados de nivel 2 y 3.

Negrita: Dobles asteriscos (**...**) para resaltar texto en negrita, como en **Fidelidad a la Fuente:**.

Listas Numeradas: Números seguidos de un punto (1., 2., etc.) para crear listas ordenadas.

Líneas Horizontales: Tres guiones (---) para crear una línea de separación.

Bloques de Código: tres acentos graves (```) que cierran el bloque JSON, la forma en que está presentado (con {"titulo": ...}) es la manera estándar de mostrar código dentro de un texto en Markdown.

Estructura Semántica Markdown

no solo añade formato visual, sino que también aporta un significado estructural al documento.

Jerarquía: Los encabezados (#, ##, ###) crean una jerarquía clara de títulos y subtítulos.

Énfasis: La negrita (**texto**) y la cursiva (*texto*) indican énfasis.

Listas: Las listas numeradas o con viñetas organizan la información en bloques lógicos.

Citas y Código: Los bloques de cita (>) o de código (```) separan claramente el contenido especial del resto del texto.

Open IA Asistente de Chat GPT

IA Documentación JACA 2025
• Live • Anyone with a link

Share Update

Create Configure

+

Name
IA Documentación JACA 2025

Description
Eres un sistema automatizado de catalogación documental (DCAS), desarrollado para un centro de documentación de prensa. Tu función es procesar arch

Instructions
PERSONA ##
Eres un sistema automatizado de catalogación documental (DCAS), desarrollado para un centro de documentación de prensa.
##TAREA##
Tu función es procesar archivos PDF, analizar su contenido textual y extraer metadatos con la máxima precisión en un formato estructurado y consistente.
Conversations with your GPT can potentially include part or all of the instructions provided.

Conversation starters

Knowledge
Conversations with your GPT can potentially reveal part or all of the files uploaded.

ENTREVISTA PROFESOR E PDF

Upload files

Recommended Model
Recommend a model to the user, which should be used by default for best results.

Preview Model 4o

IA Documentación JACA 2025
Eres un sistema automatizado de catalogación documental (DCAS), desarrollado para un centro de documentación de prensa. Tu función es procesar archivos PDF, analizar su contenido textual y extraer metadatos con la máxima precisión en un formato estructurado y consistente.



Procesa en archivo digital que te adjunto

```
json Copy Edit
{
  "titulo": "Cuidar la formación",
  "resumen_sumario": "Nicolás Álvarez de las Asturias, rector de la Universidad Eclesiástica San D",
  "resumen_ampliado": "El artículo presenta una entrevista con Nicolás Álvarez de las Asturias, re",
  "fecha_publicacion": "2025-06-10",
  "unidades_paginas": 1,
  "pagina": 34,
  "autor": "JOSÉ RAMÓN NAVARRO PAREJA",
  "forma": "Entrevista",
  "categoria": "Sociedad",
  "temas": ["Vocaciones", "Formación eclesialística", "Universidad Eclesiástica San Dámaso", "Teolog",
  "personas": ["Nicolás Álvarez de las Asturias", "Antonio María Rouco", "Francisco"],
  "fecha_analisis": "2025-06-26",
  "entidades_nombradas": ["Universidad Eclesiástica San Dámaso", "Iglesia Católica", "Madrid", "Sa",
}
```



Ask anything

+

🔍 🔊

Google

Gem de GEMINI

The screenshot shows the Gemini AI interface for a project named "IA Documentación JACA 2025". The interface is dark-themed and includes the following elements:

- Header:** A purple circle with a white vertical bar on the left, followed by the text "IA Documentación JACA 2025". On the right, there is a button labeled "Actualizar".
- Nombre:** A text field containing "IA Documentación JACA 2025".
- Instrucciones:** A section with an information icon. It contains two roles:
 - ## PERSONA ##**: "Eres un sistema automatizado de catalogación documental (DCAS), desarrollado para un centro de documentación de prensa."
 - ## TAREA ##**: "Tu función es procesar archivos PDF, analizar su contenido textual y extraer metadatos con la máxima precisión en un formato estructurado y consistente."Below the instructions are three icons: a left arrow, a right arrow, and a pencil.
- Conocimientos:** A section with an information icon. It contains a document card titled "ENTREVISTA...R EBA..." with a PDF icon and a plus sign below it.
- Vista previa:** A large preview area on the right showing the title "IA Documentación JACA 2025" with a vertical bar on the left. Below the preview is a text input field labeled "Pregunta a Gemini" with a plus sign on the left and a right arrow on the right.

Gemini

2.5 Pro

Invitar a un amigo

Analiza el PDF que te adjunto

IA Documentación JACA 2025 · Gem personalizado

Ver razonamiento

```
JSON
{
  "titulo": "«Francisco sabía que abría procesos en los que había poco consenso o cla",
  "resumen_sumario": "Entrevista a Nicolás Álvarez de las Asturias, rector de la Uni",
  "resumen_ampliado": "Nicolás Álvarez de las Asturias, rector de la Universidad Ecl",
  "fecha_publicacion": "2025-06-10",
  "unidades_paginas": 1,
  "pagina": 34,
  "autor": "JOSÉ RAMÓN NAVARRO PAREJA",
  "forma": "Entrevista",
  "categoria": "Sociedad",
  "temas": ["Universidad Eclesiástica San Dámaso", "Iglesia", "Papa Francisco", "Teol",
  "personas": ["Nicolás Álvarez de las Asturias", "Francisco", "Antonio María Rouco",
  "fecha_analisis": "2025-06-26",
  "entidades_nombradas": ["ABC", "Universidad Eclesiástica San Dámaso (UESD)", "Real",
}
```

Pregunta a Gemini

+

➤

Gemini 2.5 Flash: El Velocista Eficiente

"Flash", su principal virtud es la **velocidad** y la **eficiencia**.

- **¿Qué es?** Es el modelo diseñado para darte respuestas muy buenas, muy rápidamente y con un coste menor. Es el coche deportivo ágil: responde al instante, es divertido de usar y no gasta mucho combustible.
- **¿Para qué es ideal?** Para la gran mayoría de las tareas del día a día que necesitan ser rápidas. Piensa en él para tareas de **gran volumen**.
- **Ejemplos prácticos:**
 - **Para un documentalista/periodista:** Pedirle que lea 100 artículos de periódico y extraiga solo el titular y el autor de cada uno. Flash lo hará en un abrir y cerrar de ojos.
 - **Uso general:** Resumir rápidamente un correo largo, mantener una conversación fluida, clasificar una lista de 200 comentarios en "positivos", "negativos" o "neutros".

En resumen: Usa Flash cuando necesites una respuesta rápida y fiable para muchas tareas o cuando el coste sea un factor importante.

Gemini 2.5 Pro: El Experto en Tareas Complejas

"Pro" viene de "Profesional", su principal virtud es la **potencia de razonamiento** y la **profundidad**.

- **¿Qué es?** Es el modelo diseñado para resolver los problemas más difíciles que requieren pensar en varios pasos, entender matices muy sutiles o manejar una lógica compleja. Es la camioneta potente: puede que no sea la más rápida en arrancar, pero puede remolcar una carga pesada o atravesar terrenos difíciles que el deportivo ni soñaría.
- **¿Para qué es ideal?** Para tareas complejas que requieren un análisis profundo o la creación de algo nuevo y sofisticado. Es para los trabajos de **alta precisión**.
- **Ejemplos prácticos:**
 - **Para un documentalista/periodista:** Darle un único informe financiero de 200 páginas lleno de tablas y pedirle que "analice las tendencias de riesgo, identifique las tres mayores desviaciones presupuestarias y redacte un borrador con las conclusiones".
 - **Uso general:** Pedirle que escriba un ensayo comparando dos obras filosóficas, que resuelva un problema matemático de varios pasos o que escriba el código para una aplicación compleja.

En resumen: Usa Pro cuando la tarea sea muy difícil, requiera un razonamiento profundo y la calidad y precisión de la respuesta sean más importantes que la velocidad.

Muchas gracias.

pvila22@gmail.com