

digitales\_

# Small Language Models (SLM)

Anexo al Libro Blanco de la IA Generativa

Grupo de trabajo IA Generativa, Transformación Digital.

DIGITALES



# Índice

<b>Prólogo</b>	<b>4</b>		
<b>1. Introducción</b>	<b>5</b>		
1.1. Definición	5		
1.2. Características principales	5		
1.3. Relevancia en el ámbito de la IA	6		
1.4. Cómo funcionan	7		
<b>2. Origen y evolución</b>	<b>9</b>		
2.1. Impacto actual de los SLM	9		
2.1.1. Adopción y accesibilidad en entornos diversos			
2.1.2. Contribución a la reducción del consumo energético			
2.1.3. Expansión de casos de uso en tiempo real			
2.2. Ejemplos de modelos publicados	11		
2.2.1. Lista de principales SLM creados			
2.2.2. Lista de principales LLM creados			
2.3. Comparativa práctica de respuestas SLM vs LLM para el consumidor	13		
2.3.1. Ámbito telecomunicaciones			
2.3.2. Ámbito banca			
<b>3. Ventajas y limitaciones de los SLM</b>	<b>18</b>		
3.1. Ventajas frente a los LLM	18		
3.2. Limitaciones	21		
<b>4. Casos de uso</b>	<b>24</b>		
4.1. Aplicaciones comunes de los SLM	24		
4.2. Aplicaciones específicas por industria	25		
4.2.1. Telecomunicaciones			
4.2.2. Aprendizaje y Desarrollo (L&D)			
4.2.3. Banca y servicios financieros			
4.2.4. Salud y bienestar (Healthcare)			
4.2.5. Retail y comercio electrónico			

<b>5. Retos y oportunidades tecnológicas</b>	<b>36</b>
5.2. Retos tecnológicos	36
5.3. Oportunidades de expansión y mejora	37
5.4. Futuro de los SLM en la IA generativa	37
<b>6. Conclusiones</b>	<b>39</b>
6.1. Definición	39
6.2. Origen y evolución	39
6.3. Características y relevancia	39
6.4. Ventajas y limitaciones	40
6.5. Aplicaciones y casos de uso	40
6.6. Retos y oportunidades	41
6.7. Democratización de la IA	41
<b>7. Bibliografía y referencias</b>	<b>42</b>
<b>8. Glosario</b>	<b>43</b>
<b>9. Créditos</b>	<b>44</b>

# Prólogo

La Inteligencia Artificial Generativa (IAG) ha revolucionado nuestra interacción con la tecnología, permitiendo la creación autónoma de contenido. Esta transformación se debe en gran medida a la diversidad de modelos de lenguaje disponibles, que varían en tamaño, capacidad y aplicación.

Este anexo al Libro Blanco de la IAG se centra en los Small Language Models (SLM), modelos que destacan por su capacidad para operar en entornos con recursos limitados, equilibrando el poder de la generación y comprensión del lenguaje con la eficiencia.

Estos modelos permiten una democratización de la inteligencia artificial, haciendo posible que organizaciones de todos los tamaños puedan beneficiarse de tecnologías avanzadas sin necesidad de grandes inversiones en infraestructura. Su capacidad para adaptarse a diversas aplicaciones tecnológicas los convierte en una opción ideal para enfrentar los desafíos actuales y futuros en industrias y contextos muy diversos.

La creación de este anexo ha sido posible gracias al esfuerzo y colaboración de un equipo de profesionales de varias compañías, que han aportado contenido y perspectivas diversas, esperando que contribuya como fuente de inspiración y colaboración para materializar las oportunidades que nos ofrece esta tecnología.

El lector de este documento encontrará a continuación una completa definición y revisión del estado del arte de esta tecnología, a través de un repaso de numerosos casos de uso de los SLM en industrias claves en nuestra economía.

**Borja Lizarraga**, Líder del Grupo de Trabajo IA Generativa, Línea de Trabajo Inteligencia Artificial, en DigitalES

# 1. INTRODUCCIÓN

Es necesario comenzar por abordar en primer lugar una definición clara y concisa de los SLM, y describir sus características principales, por cuanto es un elemento diferenciador, así como su funcionamiento. De este modo a continuación podemos analizar su relevancia en el ámbito de la Inteligencia Artificial, ya presente en nuestra actividad productiva y cotidiana durante la última década.

Debemos señalar que este documento parte como extensión del Libro Blanco de Inteligencia Artificial Generativa (incluido en el listado final de bibliografía y referencias), que fue publicado por DigitalES en 2024. Recomendamos su lectura para mayor comprensión del contexto global de la inteligencia artificial en el que se enmarcan los Small Language Models.

## 1.1 DEFINICIÓN

Los *Small Language Models* (SLM) son modelos de lenguaje entrenados para procesar y generar texto de forma similar a los *Large Language Models* (LLM), pero con un número menor de parámetros y, por lo general, un consumo de recursos computacionales más bajo.

Por tanto, los SLM (como Llama o Granite) son modelos de Inteligencia Artificial diseñados para procesar y generar texto de manera eficiente, con un menor número de parámetros en comparación con los LLM (como GPT-4 o PaLM).

El objetivo principal de los SLM es ofrecer capacidades de comprensión y generación de lenguaje natural que sean más ligeras, rápidas y fáciles de desplegar en entornos con recursos limitados (por ejemplo, dispositivos móviles, aplicaciones embebidas o sistemas IoT). Otros beneficios que nos ofrecen los SLM son mayor transparencia, menores alucinaciones y menor coste en su entrenamiento.

## 1.2. CARACTERÍSTICAS PRINCIPALES

Como se indica en la introducción, son precisamente las características particulares de los SLM las que les otorgan una naturaleza de aplicabilidad diferenciada. Antes de profundizar en escenarios de uso, es importante hacer un breve repaso de los elementos más técnicos. Así, los SLM están caracterizados por las siguientes propiedades principales: 1. Tamaño reducido de parámetros

- Mientras que los LLM pueden contar hasta con millones de millones de parámetros, los SLM se encuentran en rangos mucho más moderados (decenas o cientos de millones hasta unos pocos miles de millones, e incluso menos en algunos casos).
- Esta reducción se logra mediante técnicas de comprensión, knowledge distillation, cuantización o arquitecturas más eficientes que permiten que el modelo siga siendo útil en tareas de lenguaje sin requerir grandes recursos de hardware.

## 2. Velocidad y eficiencia de cómputo

- Al requerir menos memoria y potencia de procesamiento, los SLM pueden ejecutarse de forma más ágil, con menor latencia y en dispositivos con menos capacidad de cómputo.
- Esto los hace ideales para aplicaciones en tiempo real donde la inmediatez de la respuesta es esencial.

## 3. Menor coste de entrenamiento y despliegue

- Entrenar y ajustar modelos de gran tamaño supone elevados costes económicos, de infraestructura y energéticos.
- Un SLM puede entrenarse con menos recursos y desplegarse en mayor cantidad de escenarios, haciendo posible su adopción masiva, incluso en organizaciones o proyectos con menor presupuesto.

## 4. Aplicaciones focalizadas

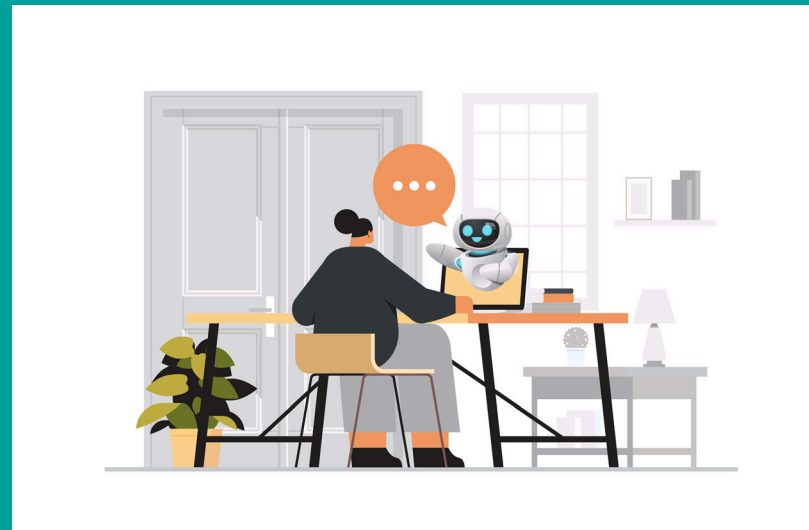
- Por lo general, los SLM están diseñados para tareas concretas o dominios específicos donde la amplitud de conocimiento de un gran modelo no es estrictamente necesaria.
- Ejemplos: asistentes virtuales en teléfonos móviles, aplicaciones de mensajería, correctores gramaticales ligeros, sistemas de respuesta en dispositivos de hogar inteligente, etc.

## 1.3. RELEVANCIA EN EL ÁMBITO DE LA IA

Los SLM representan una evolución clave en la inteligencia artificial porque no se limitan a producir respuestas extensas, sino que priorizan la precisión, la eficiencia y el ajuste a tareas específicas. Su irrupción es relevante porque permiten aplicaciones más accesibles, eficientes y seguras sin depender de infraestructuras de cómputo masivas.

### 1. Importancia frente a otras tecnologías

- A diferencia de los LLM, los SLM pueden ofrecer respuestas rápidas y a bajo coste energético, lo que facilita su implementación a gran escala.



- Comparados con enfoques tradicionales de IA basados en reglas o sistemas de pattern matching, los SLM siguen reteniendo la flexibilidad y capacidad de aprendizaje de las redes neuronales, pero de forma más eficiente.

## 2. Aceleración de la investigación en optimización de modelos

- Los SLM requieren menor cantidad de recursos computacionales para entrenar y ejecutar, lo que amplía su accesibilidad, y por tanto reducen los costes asociados al utilizar menos parámetros y técnicas de optimización, lo que permite su uso en hardware más accesible.
- El auge de los SLM ha impulsado la investigación en ámbitos como:
  - *Destilación (Distillation)*: Entrenamiento de un modelo pequeño (estudiante) utilizando la "guía" de un modelo grande (profesor).
  - *Poda (Pruning)*: Eliminación de conexiones o parámetros menos relevantes de un modelo grande sin degradar mucho el rendimiento.
  - *Cuantización (Quantization)*: Reducción del número de bits utilizados para representar cada parámetro.
  - Arquitecturas más ligeras, por ejemplo, MobileBERT, DistilBERT, TinyBERT y otras variantes.

### 3. Uso en computación perimetral (edge computing): privacidad y seguridad

- En muchos casos de uso industrial o de IoT, la latencia y la fiabilidad de la conexión a Internet son factores determinantes.
- Los SLM pueden procesar y generar texto localmente, sin depender de servidores remotos, garantizando la privacidad de los datos y la rapidez en la respuesta.

### 4. Personalización y adaptabilidad

- Los SLM pueden entrenarse o ajustarse para tareas específicas sin necesidad de grandes cantidades de datos o infraestructura avanzada.

### 5. Menor huella de carbono

- Los SLM reducen el consumo energético asociado al uso de Inteligencia Artificial, contribuyendo a un desarrollo más sostenible.

## 2.4. CÓMO FUNCIONAN

Al igual que los LLM, los SLM emplean una arquitectura basada en redes neuronales conocida como modelo transformador. Los transformadores son un tipo de arquitectura de red neuronal que transforma o cambia una secuencia de entrada en una secuencia de salida, y se han convertido

en fundamentales en el procesamiento del lenguaje natural (PLN).

He aquí una breve descripción de la arquitectura de los transformadores:

#### 1. Codificadores

- Los codificadores transforman las secuencias de entrada en representaciones numéricas llamadas incrustaciones, las cuales capturan la semántica y la posición de los tokens en la secuencia de entrada.
- Un token es una unidad fundamental de texto utilizada por los modelos de inteligencia artificial generativa para procesar y generar lenguaje.

#### 2. Mecanismo de autoatención

- Permite a los transformadores centrar su atención en los tokens más importantes de la secuencia de entrada, independientemente de su posición.

#### 3. Decodificadores

Los decodificadores utilizan este mecanismo de autoatención y las incrustaciones de los codificadores para generar la secuencia de salida estadísticamente más probable.

### Técnicas de compresión de modelos

Las técnicas de compresión de modelos se aplican para crear un modelo más ágil a partir de uno más grande. Comprimir un modelo implica reducir su tamaño y, al mismo

tiempo, conservar la mayor precisión posible. Estos son algunos métodos comunes de compresión de modelos:

- Poda
- Cuantificación o cuantización
- Factorización de bajo rango
- Destilación de conocimiento

#### Poda

La poda elimina parámetros menos cruciales, redundantes o innecesarios de una red neuronal.

Los modelos podados necesitarán a menudo ser afinados después de la poda para compensar cualquier pérdida de precisión. Es vital saber cuándo se han podado suficientes parámetros, ya que la poda excesiva puede degradar el rendimiento de un modelo.

#### Cuantificación o cuantización

La cuantificación convierte los datos de alta precisión en datos menos precisos. La cuantificación puede aligerar la carga computacional y acelerar la inferencia.

La cuantificación puede incorporarse al entrenamiento del modelo (conocido como entrenamiento consciente de la cuantificación o QAT) o realizarse después del entrenamiento (llamada cuantificación posterior al entrenamiento o PTQ). La PTQ no requiere tanta potencia computacional

y datos de entrenamiento como el QAT, pero el QAT puede producir un modelo más preciso.

### Factorización de bajo rango

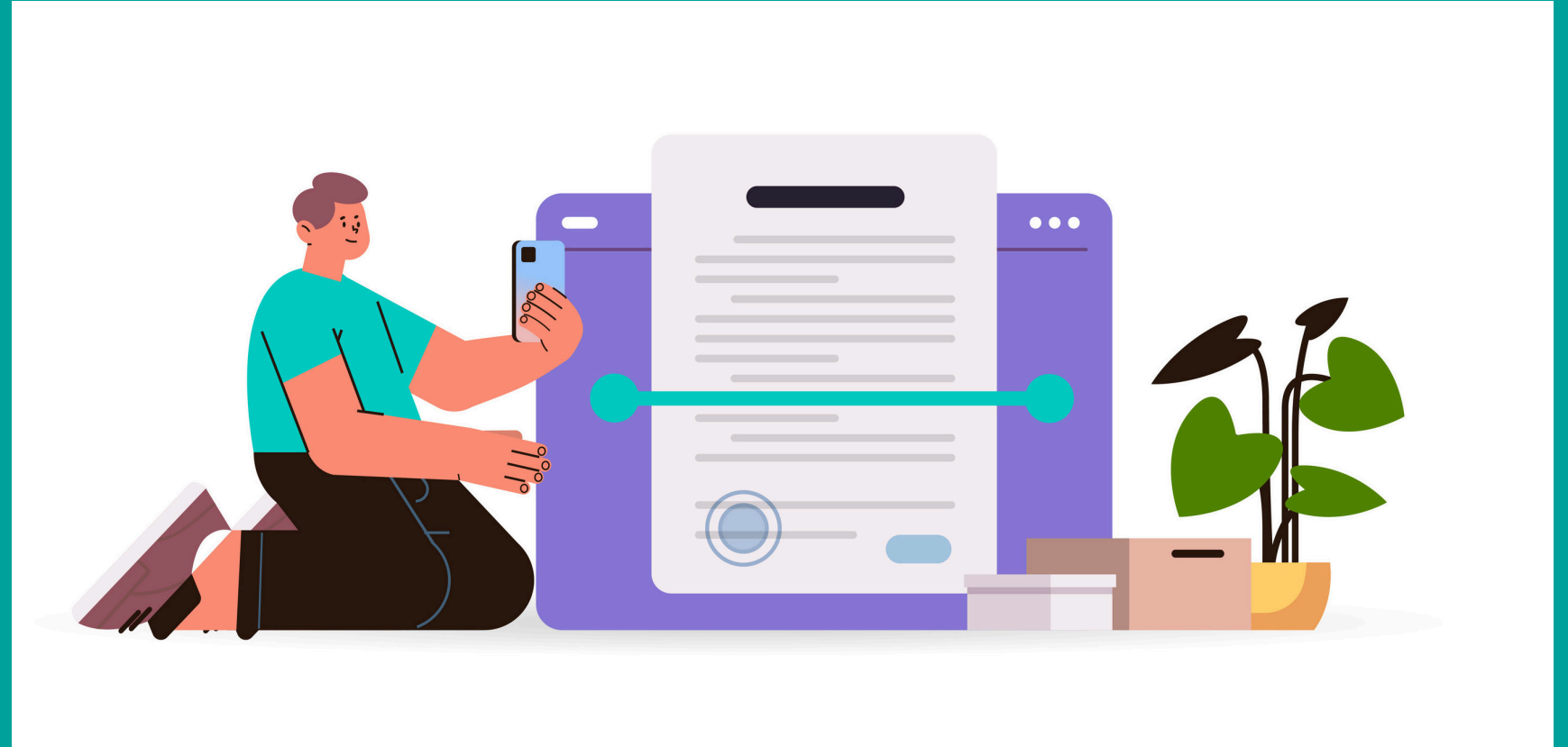
La factorización de bajo rango descompone una gran matriz de ponderaciones en una matriz más pequeña y de menor rango. Esta aproximación más compacta puede dar lugar a menos parámetros, disminuir el número de cálculos y simplificar las operaciones matriciales complejas.

Sin embargo, la factorización de bajo rango puede ser intensiva desde el punto de vista computacional y más difícil de implementar. Al igual que la poda, la red factorizada requerirá de afinación para recuperar cualquier pérdida de precisión.

### Destilación de conocimiento

La destilación de conocimiento implica normalmente transferir los aprendizajes de un "modelo maestro" pre-entrenado a un "modelo estudiante". El modelo estudiante está entrenado no solo para coincidir con las predicciones del modelo maestro, sino también para imitar su proceso subyacente de razonamiento. Como tal, el conocimiento de un modelo más grande se "destila" esencialmente en uno más pequeño.

La destilación de conocimiento es un enfoque popular para muchos SLM.



## 2. ORIGEN Y EVOLUCIÓN

Los SLM surgen como una respuesta a la necesidad de equilibrar el poder de los grandes modelos de lenguaje con la eficiencia, el coste y la escalabilidad en distintos entornos y dispositivos. Su relevancia radica en que permiten aprovechar las ventajas de la inteligencia artificial basada en redes neuronales, sin requerir la enorme infraestructura ni los gastos asociados a los modelos de gran tamaño.

En el futuro, conforme las técnicas de compresión, destilación y optimización sigan avanzando, es probable que veamos un crecimiento todavía mayor de estos modelos ligeros, capaces de abordar un amplio espectro de aplicaciones en el mundo real sin sacrificar en exceso la calidad de sus resultados.

### Necesidades y desafíos tecnológicos que han motivado la creación de SLM

#### 1. Recursos y coste

- Los LLM requieren infraestructura de alto nivel (GPUs de última generación, gran cantidad de memoria, etc.), mientras que los SLM están diseñados para funcionar con hardware limitado.

#### 2. Rendimiento en tareas generales

- Los LLM suelen destacarse en tareas abiertas y amplias, gracias a la gran variedad de datos con los que son entrenados.
- En cambio, los SLM pueden mostrar un rendimiento

cercano, o incluso igual, a modelos más grandes en dominios o tareas específicas, siempre y cuando estén bien optimizados o *fine-tuned* en esos contextos.

#### 3. Privacidad y despliegue local

- Un SLM puede ejecutarse en local, garantizando que los datos de los usuarios no salgan de su dispositivo, algo que puede ser muy relevante en entornos con restricciones de seguridad y privacidad.

#### 4. Tamaño de la comunidad y soporte

- Aunque los LLM han acaparado mayor atención mediática, existe una comunidad creciente que desarrolla y comparte modelos pequeños, ofreciendo bibliotecas, guías y herramientas de optimización.

### 2.1 IMPACTO ACTUAL DE LOS SLM

Los SLM han tenido un impacto significativo en la democratización de la inteligencia artificial y en la optimización de recursos tecnológicos. En concreto, los SLM han transformado el panorama de la inteligencia artificial al hacer que los modelos de lenguaje sean más accesibles, eficientes y versátiles.

Su capacidad para operar en entornos con recursos limitados y su impacto en la reducción del consumo energético los posiciona como una alternativa clave frente a los

modelos de lenguaje de gran escala, impulsando la innovación en diversos sectores.

### 2.1.1 Adopción y accesibilidad en entornos diversos

Los SLM han facilitado el acceso a modelos de lenguaje en escenarios donde los LLM tradicionales son difíciles de implementar debido a restricciones de recursos computacionales o costes operativos. Su impacto en la adopción y accesibilidad se puede ver en:

#### Dispositivos con capacidad limitada

- Los SLM están diseñados para ser más ligeros y eficientes en términos de recursos computacionales. Esto los hace ideales para su implementación en dispositivos con capacidades limitadas, como smartphones, tabletas y otros dispositivos IoT.
- Adicionalmente, los SLM pueden ejecutarse en dispositivos móviles, tabletas y computadoras personales sin necesidad de acceso constante a internet o a centros de datos.
- Esto ha impulsado aplicaciones como teclados inteligentes, asistentes de voz y herramientas de escritura con corrección gramatical que funcionan localmente.

#### Facilidad de implementación y mantenimiento

- Debido a su menor tamaño y complejidad, los SLM son más fáciles y rápidos de implementar y mantener. Esto

reduce significativamente el tiempo y el esfuerzo necesarios para su integración en diversas aplicaciones, permitiendo a las empresas pequeñas y startups adoptar tecnologías avanzadas de procesamiento de lenguaje natural (NLP) sin necesidad de grandes equipos de desarrollo.

#### Pequeñas y Medianas Empresas (PyMEs)

- Las empresas pequeñas, que a menudo no tienen acceso a infraestructuras de computación avanzadas, pueden beneficiarse de estos SLM sin incurrir en altos costos de hardware y energía.
- En particular, muchas organizaciones no pueden costear la infraestructura para ejecutar LLM, pero los SLM permiten incorporar IA en tareas como atención al cliente, análisis de datos y automatización sin grandes inversiones en servidores o servicios en la nube.
- Ejemplo: *chatbots* y asistentes virtuales ligeros para mejorar la interacción con clientes sin depender de procesamiento externo.

#### Accesibilidad en entornos educativos y comunitarios

- Los SLM también han encontrado un lugar en entornos educativos y comunitarios, donde los recursos pueden ser limitados. Por ejemplo, se utilizan en aplicaciones de aprendizaje a distancia y en herramientas educativas que requieren procesamiento de lenguaje natural,

proporcionando acceso a tecnologías avanzadas a estudiantes y educadores en áreas remotas o con menos recursos

#### Infraestructuras locales y entornos con conectividad limitada

- En regiones donde el acceso a internet es intermitente o costoso, los SLM permiten ejecutar modelos de IA sin depender de la nube, favoreciendo su uso en educación, salud y comunicación en áreas rurales o en países en desarrollo.

### 2.1.2 Contribución a la reducción del consumo energético

La implementación de SLM no solo optimiza el rendimiento y la eficiencia, sino que también juega un papel crucial en la reducción del consumo energético. Estos modelos, al requerir menos recursos computacionales en comparación con los LLM, permiten un desarrollo más sostenible y alineado con los objetivos de eficiencia energética.

#### Eficiencia energética y sostenibilidad

- Los LLM requieren centros de datos de alto consumo energético para su entrenamiento y ejecución. En contraste, los SLM reducen la huella de carbono al disminuir la necesidad de procesamiento masivo.
- Esto ha llevado a una mayor adopción de soluciones de IA en empresas comprometidas con la sostenibilidad,

ya que pueden implementar modelos de lenguaje sin generar altos costes ambientales. Esto es crucial en un contexto donde la sostenibilidad y la eficiencia energética son cada vez más importantes.

### 2.1.3 Expansión de casos de uso en tiempo real

La eficiencia y menor latencia de los SLM los convierten en una solución ideal para aplicaciones en tiempo real.

#### Casos de uso en tiempo real

- Debido a su eficiencia y menor latencia, los SLM son ideales para aplicaciones en tiempo real, permitiendo respuestas inmediatas en aplicaciones como traducción simultánea, reconocimiento de voz y asistentes de productividad sin depender de servidores externos.
- En sectores como la salud, han permitido diagnósticos más rápidos a través de herramientas de IA que procesan información localmente, sin riesgos de privacidad asociados a la transmisión de datos sensibles.
- En ciberseguridad, los SLM pueden ejecutarse en dispositivos de borde (*edge computing*) para detectar amenazas en tiempo real sin enviar información a la nube.

## 2.2 EJEMPLOS DE MODELOS PUBLICADOS

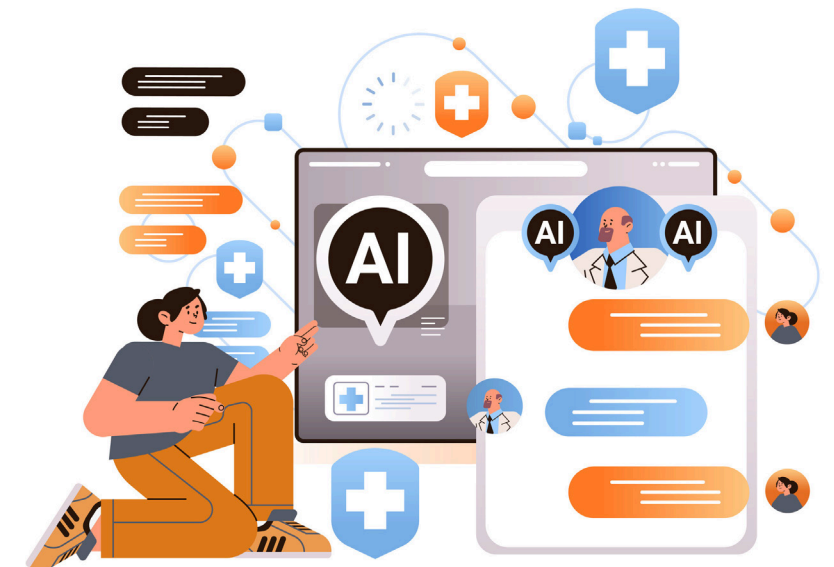
El año 2025 marca una etapa crucial en la inteligencia artificial, observándose una doble tendencia: el desarrollo de modelos fundacionales cada vez más grandes y potentes (LLM) y, simultáneamente, la proliferación de modelos más pequeños y eficientes (SLM). Los SLM ganan terreno por su rentabilidad y agilidad, mientras los LLM como GPT-4o y Gemini amplían fronteras en multimodalidad (texto, imagen, audio, vídeo), razonamiento complejo y ventanas de contexto extendidas, como se ve también en la serie IBM Granite y GPT-4.1. La creciente adopción de modelos de código abierto, como Mistral, Qwen, Deepseek, Llama y Granite, está democratizando el acceso a la IA y estimulando la innovación comunitaria.

Este panorama se caracteriza por escalar capacidades y, a la vez, optimizar la eficiencia y accesibilidad mediante SLM y código abierto. La multimodalidad y el razonamiento avanzado se consolidan como estándares en los LLM líderes. Mientras estos impulsan aplicaciones sofisticadas (*chatbots* avanzados, creación de contenido), los SLM son clave en tareas que demandan eficiencia y baja latencia, desplegándose en móviles, IoT y sistemas embebidos. El sector empresarial adopta ambos tipos para usos variados, desde atención al cliente hasta análisis especializados. Este informe ofrece una visión actualizada de SLM y LLM en 2025.

Presentamos algunos de los modelos de lenguaje pequeños (SLM) y grandes (LLM) más conocidos y recientes, indicando de forma aproximada su número de parámetros y la ventana de contexto (*tokens*).

### 2.2.1 Lista de principales SLM creados

En los últimos años, los SLM han evolucionado de curiosidades académicas a piezas clave en entornos con recursos limitados. Respondiendo a la necesidad de respuestas rápidas y bajo consumo energético, los SLM se han sofisticado para usos reales sin renunciar a la



calidad. Hoy en día, son herramientas estratégicas para acercar la inteligencia artificial al dispositivo (edge).

En 2025, la oferta de SLM presenta novedades significativas conocidas en los *leaderboards* de modelos, hasta la fecha de publicación de este informe. A saber:

- **Gemma 3:** Google lanzó Gemma 3 en marzo de 2025, con tamaños que van desde **1B** hasta **27B** parámetros (nota: B = *billion*, mil millones). Los modelos de **4B**, **12B** y **27B** son multimodales, aceptando texto e imagen, y soportan más de 140 idiomas. El modelo de **1B** se enfoca en texto en inglés. Todos ofrecen ventanas de contexto extendidas, hasta **128k tokens** en los modelos más grandes.
- **Granite 3:** IBM ha continuado su serie Granite con la versión 3.3 **8B**, anunciada en abril de 2025. Modelo instructivo con capacidades mejoradas de razonamiento, entrenado en datos de alta calidad optimizados para el lenguaje natural y los casos de uso empresarial con soporte de llenado en el medio. **128k tokens**. Adicionalmente la familia también cuenta con modelos de visión, de voz, seguridad incrustaciones de lenguaje.
- **Phi 4:** Microsoft presentó la familia Phi 4, incluyendo Phi-4 (**14B**), Phi-4-mini-instruct (**3.8B**) y Phi-4-multimodal-instruct (**5.6B**). Phi-4 se lanzó en diciembre de 2024, mientras que las versiones mini y multimodal se hicieron públicas en febrero de 2025. Destacan por sus

capacidades de razonamiento, con la versión multimodal integrando voz, visión y texto.

- **Ministral:** Lanzado por Mistral AI en octubre de 2024, Ministral **8B** es un modelo de 8 mil millones de parámetros diseñado para eficiencia en el *edge* y en dispositivos. Cuenta con un patrón de atención de ventana deslizante entrelazada y soporta hasta **128k** de longitud de contexto, destacando en conocimiento y razonamiento. Mistral también dispone de más modalidades como Ministral **3B**, Pixtral.
- **QwQ-32B:** El equipo de Qwen de Alibaba Cloud lanzó QwQ-**32B** en marzo de 2025. Es un modelo de razonamiento experimental con **32.5 mil millones** de parámetros, basado en Qwen2.5-**32B**. Utiliza aprendizaje reforzado para mejorar sus habilidades de razonamiento y cuenta con una ventana de contexto de **130k tokens**. Es un modelo que por su tamaño estaría en un rango intermedio con los LLM de gran tamaño, pero que cuantizado sería otra opción a valorar. También dispone de versión **7B** Qwen-2.5-**7B**-Instruct.

La diversidad actual de SLM permite adaptar la elección del modelo a requisitos específicos de rendimiento, coste y caso de uso. Estos modelos se aplican ampliamente en sectores como salud (automatización administrativa, asistencia clínica), educación (*chatbots* educativos, personalización del aprendizaje), finanzas (detección de

fraude, análisis de cumplimiento) y tecnología móvil (asistentes offline, traducción en dispositivo), demostrando su versatilidad y valor práctico.

### 2.2.2 Lista de principales LLM creados

Si nos remontamos a los orígenes, los LLM continúan su escalada en capacidades. Modelos fundacionales como **GPT-3** (175B parámetros) y **BERT-large** (~340M parámetros) fueron hitos previos. **PaLM** de Google superó los 500B parámetros en sus versiones más grandes.

Las actualizaciones y lanzamientos clave en 2025 incluyen:

- **Familia GPT-4.1 (GPT-4.1, mini, nano):** Disponible vía API, supera a GPT-4o en codificación y seguimiento de instrucciones, soportando hasta **1 millón** de *tokens* de contexto.
- **Familia Gemini (Google DeepMind):** Con sus modelos **Gemini 2.5 Pro Experimental** lidera *benchmarks* de razonamiento complejas y con 2.5 Flash Preview. Soportando hasta **1 millón** de *tokens* de contexto.
- **Familia Claude (Anthropic):** Se enfoca en IA responsable. Los modelos actuales son **Claude 3.7 Sonnet** (híbrido, rápido y con buen razonamiento) y **Claude 3.5 Haiku** (el más rápido y compacto). Hasta **200k tokens** de contexto.
- **Llama 4 (Meta AI):** Lanzada en abril de 2025, con ar-

arquitectura MoE (Mezcla de Expertos), capacidades multimodales (texto/imagen -> texto) y multilingüismo. Incluye **Scout (109B** parámetros totales, 128 expertos, 17B activos), **Maverick** (400B parámetros totales, 128 expertos, 17B activos) y **Behemoth** (preview, 2T parámetros totales, 16 expertos con 288B activos). **Llama 3.1** (hasta 405B) se lanzó en julio de 2024. Hasta **10M tokens** de contexto.

- **DeepSeek R1:** Modelo de razonamiento de DeepSeek AI. Destacando por velocidad y precisión tras ajustes finos. Hasta **128k** tokens de contexto.
- **Qwen 2.5-Max:** Modelo MoE a gran escala de Alibaba, parte de la serie Qwen 2.5. Destaca en multimodalidad. Hasta **32k** tokens de contexto.
- **Grok-3:** Lanzado por xAI en febrero de 2025, con mejoras significativas en potencia computacional para razonamiento avanzado. Sigue a **Grok-2** (agosto 2024). Hasta **131k** tokens de contexto.

Nota: esta información se ha elaborado desde *Leaderboards* y evaluación del rendimiento en 2025 de *Hugging Face*.

Plataformas como **Hugging Face Open LLM Leaderboard**, **LM Arena**, **Artificial Analysis LLM Performance Leaderboard** y **LLMWorld** son referencias para evaluar LLM y SLM. Utilizan *benchmarks* diversos como MMLU

(conocimiento general), HumanEval (codificación), GPQA (conocimiento científico), SWE Bench (resolución de problemas reales), categoría de *opensource* o propietario, capacidades de idioma y métricas como velocidad, tamaño de contexto, latencia, coste, etc.

### 2.2.3 Perspectivas

El 2025 consolida un ecosistema de IA diverso. Los SLM ofrecen eficiencia y accesibilidad para tareas específicas y entornos restringidos, mientras los LLM siguen avanzando en capacidades cognitivas y multimodales.

Las tendencias futuras apuntan a una mayor eficiencia en todos los tamaños de modelo, expansión de la multimodalidad, ventanas de contexto aún más largas, consolidación del código abierto, un énfasis creciente en la seguridad y la ética, y una mayor especialización de modelos



para nichos industriales. La clave reside en comprender los requisitos específicos de cada caso (recursos, latencia, complejidad) para guiar la selección entre SLM y LLM, considerando tanto *benchmarks* como rendimiento en el mundo real, y aprovechando la flexibilidad y los desafíos del creciente ecosistema de código abierto y las soluciones empresariales enfocadas como Granite.

## 2.3 COMPARATIVA PRÁCTICA DE RESPUESTAS SLM VS LLM PARA EL CONSUMIDOR

A continuación, se incluye un par de ejemplos hipotéticos de cómo podrían responder un SLM y un LLM ante la misma consulta en dos industrias diferentes, elegidas por su amplia utilización actual en servicios de impacto directo al consumidor.

### 2.3.1 Ámbito telecomunicaciones

El escenario elegido es una consulta de un cliente que experimenta problemas con la velocidad de internet en casa y busca asistencia.

La respuesta del SLM estará caracterizada por:

- Mensaje breve y directo.
- Menos "razonamiento creativo" y menos detalles adicionales.

- Suele enfocarse únicamente en la información esencial y en el dominio específico (telecomunicaciones).

### Ejemplo de prompt:

El prompt o consulta del usuario:

*“Hola, tengo problemas con la velocidad de internet en casa, ¿pueden ayudarme a solucionarlo?”*

### Ejemplo de respuesta de un SLM:

*“Claro. Para mejorar la velocidad de tu internet, por favor revisa lo siguiente:*

- *Comprueba si tu router esté encendido y cerca de tus dispositivos.*
- *Reinicia el router apagándolo durante 10 segundos y volviendo a encenderlo.*
- *Asegúrate de que no haya muchos dispositivos conectados a la vez.*

*Si el problema continúa, escríbenos los indicadores de luces en tu router y tu número de cuenta para revisar tu línea.”*

En este ejemplo, el SLM:

- Es directo: ofrece **instrucciones** básicas para detectar y solucionar problemas comunes.
- Solicita **datos concretos** (indicadores de luces del rou-

ter, número de cuenta) que ayuden en diagnóstico, sin extenderse mucho.

- Muestra un **conocimiento específico** sobre un procedimiento típico de reinicio y chequeo de router en telecomunicaciones.

La respuesta del LLM estará caracterizada por:

- Mayor elaboración y contexto, incluso agregando consejos más detallados o explicaciones técnicas.
- Puede abarcar más opciones de solución (por ejemplo, configuración interna del router, ajustes de seguridad, etc.).
- Tono más conversacional y con referencias adicionales.

### Ejemplo de respuesta de un LLM:

*“Hola, con gusto te ayudo a resolver los problemas de velocidad de tu internet en casa. He aquí algunos pasos detallados que puedes seguir:*

*Verifica la ubicación del router: Asegúrate de que esté situado en un lugar abierto, lejos de paredes gruesas o electrodomésticos que puedan bloquear la señal.*

*Actualiza el firmware: Entra a la configuración de tu router mediante la dirección indicada (normalmente 192.168.1.1) y verifica si hay alguna actualización de fir-*

*mware pendiente. Esto puede mejorar notablemente la estabilidad y la velocidad de tu conexión.*

*Canales de frecuencia: Si usas Wi-Fi en la banda de 2,4 GHz, considera cambiar el canal a uno menos saturado (por ejemplo, canal 1, 6 u 11). Si tu router es de doble banda (2,4/5 GHz), prueba conectarte a la banda de 5 GHz para mayor velocidad.*

*Desconecta equipos innecesarios: Dispositivos en segundo plano consumen ancho de banda; prueba pausarlos o desconectarlos temporalmente.*

*Reinicia los equipos: Apaga y enciende tu router y módem, espera unos segundos antes de volver a encenderlos. Este procedimiento ayuda a recalibrar la conexión.*

*Mide tu velocidad: Tras estos pasos, usa una herramienta de medición de velocidad en línea para verificar mejoras (p. ej., Speedtest u otro servicio).*

*Si después de estos consejos sigues experimentando baja velocidad, te recomiendo escribirnos los resultados de tu última prueba de velocidad, la marca/modelo de tu router y cualquier código de error que aparezca en las luces de tu dispositivo. Con esa información, podemos verificar si hay problemas en la línea o necesidad de un reemplazo de hardware.*

*Espero que esta información te sea útil y puedas disfru-*

tar nuevamente de tu conexión sin interrupciones. ¡Estoy para ayudarte en lo que necesites!"

En este ejemplo, el LLM:

- Proporciona **explicaciones más extensas** y "motiva" por qué ciertos pasos funcionan (ej. interferencias, actualización de firmware).
- Menciona **herramientas de medición y técnicas más avanzadas** (cambio de canales, banda de 5 GHz).
- Usa un **tono más conversacional** y abarca un rango más amplio de posibles soluciones técnicas.

### Diferencias clave entre la respuesta del SLM y la del LLM:

#### 1. Nivel de detalle:

- El SLM se limita a pasos básicos y directos de solución de problemas.
- El LLM aporta contenido adicional (firmware, canales de frecuencia, banda 5 GHz).

#### 2. Contexto e información adicional:

- El SLM no profundiza mucho en la explicación de por qué ciertas acciones ayudan.
- El LLM añade explicaciones y recursos externos, asumiendo una mayor amplitud de conocimiento.

#### 3. Longitud y tono:

- La respuesta del SLM es más escueta y concisa.
- La del LLM adopta un tono más cercano y detallado, casi como si el usuario hablase con un experto técnico.

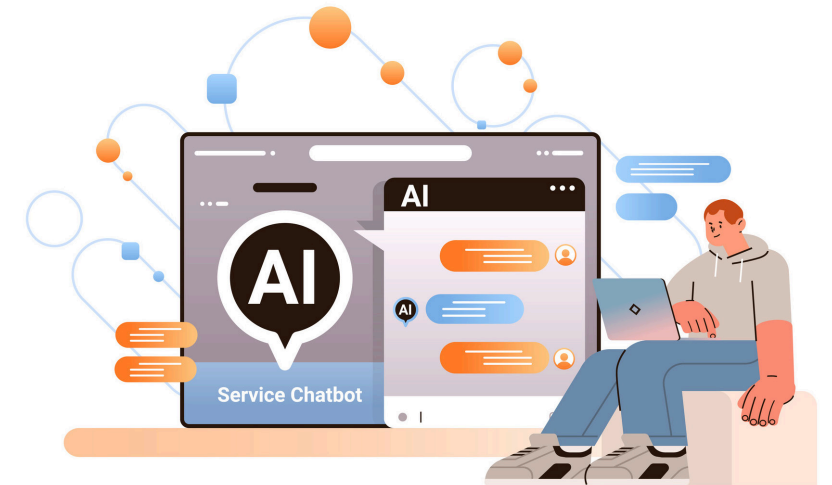
#### 4. Orientación a la acción vs. profundidad técnica:

- El SLM da instrucciones rápidas y sencillas.
- El LLM profundiza en causas potenciales e introduce pasos de diagnóstico más avanzados.

#### Conclusiones:

A modo de conclusión sobre el ejemplo presentado cabe resaltar:

- Este ejemplo ilustra cómo, ante la misma consulta en el **ámbito de las telecomunicaciones**, un **SLM** tiende a ofrecer **respuestas compactas** y concretas, mientras que un **LLM** amplía la respuesta con **más contexto y detalle**.
- La elección de uno u otro dependerá de factores como **recursos disponibles, finalidad de la asistencia** (rápida y específica vs. detallada y amplia) y **necesidades de los usuarios**.
- En todo caso, ambas soluciones pueden ser eficaces en un **centro de atención al cliente**, ya sea como **herramientas independientes o combinadas** (por ejemplo, entrenar un LLM y luego *destilar* sus capacidades en un SLM para un despliegue más ligero).



### 2.3.2 Ámbito banca

El escenario planteado es un cliente que desea revisar su estado de cuenta y entender qué comisiones se le han aplicado.

La respuesta del SLM estará caracterizada por:

- Enfoque **breve y directo** en la tarea.
- **Menos detalle** en términos de explicación o posibles variaciones
- Da **instrucciones concretas** para resolver la duda.

#### Ejemplo de *prompt*:

El *prompt* o consulta del usuario:

"Hola, necesito revisar mi estado de cuenta y saber qué comisiones me están cobrando. ¿Me pueden ayudar?"

#### Ejemplo de respuesta:

"Con gusto. Para revisar tu estado de cuenta y las comisiones:

- Ingresa a la banca en línea con tu usuario y contraseña.
- Selecciona la opción 'Estado de cuenta' o 'Movimientos'.
- Busca la sección 'Detalle de comisiones' para ver cada comisión aplicada.
- Si no la encuentras, comparte tu número de cuenta y la fecha aproximada del periodo que deseas consultar, y te enviaré un reporte detallado.

Avísame si tienes alguna duda adicional."

En este ejemplo, el SLM:

- Ofrece **pasos muy concretos** para ver el estado de cuenta.
- Menciona la sección de comisiones de manera **escueta**, sin profundizar en más explicaciones.
- **Sugiere un método adicional** para obtener el reporte (compartir datos de cuenta y periodo de consulta).

La respuesta del LLM estará caracterizada por:

- Mayor detalle y empatía en la comunicación.

- Expande posibles escenarios (ej. uso de la banca móvil, condiciones especiales).
- Agrega explicaciones sobre qué tipos de comisiones pueden existir y por qué se aplican.

#### Ejemplo de respuesta LLM:

"¡Claro que sí! Será un gusto ayudarte a revisar tu estado de cuenta y entender qué comisiones se te han aplicado. Aquí tienes algunos pasos e información que pueden serte útiles:

##### 1. Acceso a tu cuenta en línea o app móvil:

- Ingresa con tu usuario y contraseña en el portal web oficial o utiliza la app móvil.
- En caso de que no recuerdes tus credenciales, puedes restablecer tu contraseña haciendo clic en '¿Olvidaste tu contraseña?'.

##### 2. Ubicación del estado de cuenta:

- Una vez dentro, busca la sección 'Movimientos', 'Estados de cuenta' o 'Resumen de cuenta'. Ahí podrás ver tus transacciones recientes y pasadas.
- Normalmente, puedes seleccionar la fecha o rango de fechas para el periodo que te interese.

##### 3. Detalle de comisiones:

- En tu estado de cuenta, deberías encontrar una colum-

na o sección específica dedicada a comisiones o cargos por servicios.

- Entre las más comunes se encuentran comisiones por manejo de cuenta, retiros en cajeros de otros bancos o pagos tardíos de tarjeta de crédito.

- Si te parece que alguna comisión no corresponde, anota la fecha y el monto exacto para investigar su origen.

##### 4. Generar reporte y aclaraciones:

- Si necesitas un resumen más detallado (por ejemplo, de varios meses), haz clic en 'Descargar PDF' o 'Generar reporte'.
- Si encuentras alguna discrepancia, ponte en contacto con nuestro soporte especializado al teléfono que aparece en tu banca en línea o envíame más detalles aquí para que revise tu caso.

##### 5. Consulta de promociones o exenciones:

- A veces hay planes de exención de comisiones si cumples ciertos requisitos (saldo mínimo, número de movimientos, etc.). Puede ser útil revisar esas condiciones para que no se te apliquen comisiones innecesarias en el futuro.
- Te recomiendo revisar la información y anotar cualquier cargo que no reconozcas o que te resulte inesperado. Con gusto podemos investigar esos puntos

juntos. Si prefieres, déjame un mensaje con tu número de cuenta parcial (por seguridad, no compartas todos los dígitos), la fecha del cargo y la descripción de la comisión para que pueda ayudarte a revisarla a detalle. ¡Quedo atento a cualquier duda que tengas!"

En este ejemplo, el LLM:

- Proporciona **más contexto** acerca de los tipos de comisiones comunes.
- Ofrece **pasos más detallados**: menciona la posibilidad de descargar un reporte, investigar discrepancias y contactar al soporte especializado.
- Incluye consejos adicionales sobre **exenciones de comisiones** y sobre cómo comunicar los detalles de forma segura.

### Diferencias clave entre la respuesta del SLM y la del LLM:

Las diferencias principales entre las respuestas de los SLM y LLM son las siguientes.

#### 1. Nivel de detalle y explicación:

- El SLM da **instrucciones concretas** sin mucha profundidad.
- El LLM **amplía la información** con ejemplos de comisiones y recomendaciones adicionales.

#### 2. Longitud y tono:

- El SLM es más **conciso** y directo, se enfoca en resolver puntualmente.
- El LLM aporta un **tono más conversacional** y empático, abarcando varias situaciones que el cliente podría enfrentar.

#### 3. Conocimiento más amplio:

- El SLM se ciñe a lo esencial, **no profundiza** en las razones de las comisiones.
- El LLM **menciona** tipos de comisiones, posibilidad de exención y **hasta recomendaciones** de seguridad.

#### 4. Asistencia adicional:

- El SLM **invita a enviar información** si el cliente necesita más detalles.
- El LLM sugiere diferentes vías de soporte y **detalla cómo hacerlo**, incluso habla sobre contacto telefónico y descarga de reportes.

#### Conclusiones:

Este ejemplo ilustra cómo un **SLM** respondería de forma **rápida y concreta**, mientras un **LLM** podría ofrecer un **abanico más amplio de explicaciones y soluciones**.

- En banca, un SLM cubriría la necesidad básica de guiar al cliente en la consulta de su estado de cuenta y comisiones.

Un LLM, en cambio, enriquecería la conversación con detalles adicionales, posibles escenarios y *tips* de seguridad o ahorro de costos. La elección entre SLM y LLM dependerá de factores como la infraestructura disponible, la complejidad del soporte requerido y la experiencia que se quiera ofrecer a los clientes.



# 3. VENTAJAS Y LIMITACIONES DE LOS SLM

Los SLM surgen como una solución ante las dificultades que presentan los LLM, tales como GPT y Llama. Los SLM establecen un equilibrio entre desempeño, eficiencia, coste y privacidad, llegando incluso a igualar o superar a los LLM en determinadas situaciones. Este capítulo tiene como objetivo analizar las ventajas, las limitaciones y la función estratégica de los SLM en la aplicación práctica de la Inteligencia Artificial.

## 3.1 VENTAJAS FRENTE A LOS LLM

Aunque los LLM han demostrado excelente capacidad en una amplia gama de tareas, su gran tamaño supone barreras significativas. Los SLM, en cambio, presentan un conjunto de ventajas que los hacen más atractivos para contextos específicos ampliando el alcance de la IA generativa.

### 3.1.1 Eficiencia y rendimiento optimizado

La ventaja fundamental de los SLM radica en su reducido tamaño en comparación con los LLM. Mientras que los LLM a menudo superan los 100 mil millones o incluso alcanzan billones de parámetros, los SLM suelen operar en rangos muy inferiores como hemos visto en capítulos anteriores. Esta diferencia de escala se traduce en beneficios tangibles:

- **Menor consumo de recursos:** La reducción en el nú-

mero de parámetros disminuye drásticamente la necesidad de potencia computacional (CPU/GPU), memoria (RAM) y capacidad de almacenamiento. Esta eficiencia no solo reduce los costes operativos directos, sino que también disminuye significativamente el consumo energético, alineándose con los crecientes objetivos de sostenibilidad.

- **Menor latencia:** La arquitectura más ligera de los SLM permite tiempos de inferencia (el tiempo que tarda el modelo en generar una respuesta) significativamente más rápidos. Esta baja latencia es un factor crítico para aplicaciones que requieren respuestas en tiempo real, como asistentes de voz interactivos o sistemas de respuesta automática en chatbots donde cualquier retraso empeoraría la experiencia del usuario o haría inviable la aplicación.
- **Velocidad de entrenamiento y ajuste (*fine-tuning*):** Entrenar un SLM desde cero o, lo que es más habitual, ajustarlo para una tarea específica (*fine-tuning*) requiere mucho menos tiempo, datos y recursos computacionales que realizar el mismo proceso con un LLM. Esta agilidad permite a las organizaciones iterar más rápidamente sobre sus modelos, adaptarlos a necesidades cambiantes con mayor flexibilidad y experimentar con diferentes enfoques sin incurrir en costes excesivamente altos.

- **Mejor adaptabilidad a contextos locales o dominios específicos:** Gracias a su menor tamaño y a la reducción de requerimientos técnicos, los SLM pueden desplegarse con mayor facilidad en entornos locales o en dispositivos periféricos (*edge devices*), lo que los hace especialmente adecuados para escenarios donde la privacidad de los datos es crítica o donde no se dispone de conexión constante a servidores remotos. Asimismo, su menor complejidad facilita una personalización más precisa a dominios concretos (legal, médico, industrial, etc.), mejorando la relevancia y precisión de sus respuestas en contextos altamente especializados.

Las ventajas de los SLM, como el bajo consumo, la velocidad y la rápida adaptación, impulsan la innovación al democratizar el acceso a la IA generativa. Las PYMES, las *startups* y los investigadores ahora pueden utilizar hardware de menor coste para desarrollar e implementar soluciones de IA, lo que está acelerando la innovación en variedad de sectores.

### 3.1.2 Accesibilidad y reducción de costes

La eficiencia inherente de los SLM se traduce directamente en una mayor accesibilidad económica y técnica:

- **Menores costes de infraestructura y despliegue:** La reducida demanda de hardware y recursos de cómputo, vistas en el punto anterior, hace que el despliegue



y mantenimiento de SLM sea significativamente más económico que el de LLM.

- **Democratización de la IA:** Al reducir las barreras económicas y técnicas, los SLM facilitan que un abanico más amplio de desarrolladores, organizaciones e individuos pueda participar activamente en el campo de la IA generativa.
- **Privacidad y cumplimiento normativo:** El despliegue local de SLM también permite un mayor control sobre los datos tratados, algo fundamental en sectores regulados o que manejan información sensible. Al evitar el envío de datos a servicios externos, las organizaciones pueden mejorar su cumplimiento con normativas como el RGPD o la LOPDGDD, mitigando riesgos lega-

les y reputacionales. Además del ahorro, desplegar SLM localmente da independencia a las organizaciones. Al no depender de APIs (Application Programming Interface) externas, evitan la volatilidad de precios, cambios en políticas de consumo y posibles interrupciones. Operar SLM internamente da mayor control y previsibilidad de costes.

### 3.1.3 Viabilidad de la IA en Edge Computing

Una de las ventajas transformadoras de los SLM es su capacidad para ejecutarse directamente en dispositivos con recursos computacionales limitados, lo que sucede cuando hablamos de *Edge Computing*.

- **Habilitación de la IA en el edge:** Los SLM, gracias a su eficiencia, pueden operar en una amplia gama de dis-

positivos fuera de los centros de datos tradicionales: *smartphones*, tabletas, sensores IoT, sistemas embebidos en vehículos, equipos en plantas de fabricación o servidores locales.

- **Casos de uso:** Esta capacidad permite innovaciones como diagnósticos médicos portátiles en zonas remotas, inspección de calidad automatizada en producción, o navegación autónoma.
- **Beneficios del procesamiento local:** Ejecutar la IA en dispositivos *edge* ofrece múltiples ventajas sobre el procesamiento en la nube: latencia drásticamente reducida (esencial para interacciones instantáneas), menor consumo de ancho de banda (los datos se procesan donde se generan), mayor fiabilidad (la aplicación puede funcionar incluso sin conexión a internet o con conectividad intermitente) y, crucialmente, mejoras significativas en privacidad y seguridad (los datos sensibles no necesitan abandonar el dispositivo).

### 3.1.4 Especialización y adaptabilidad

Mientras que los LLM destacan por la amplitud de su conocimiento, los SLM ofrecen la posibilidad de una mayor especialización:

- **Alto rendimiento en tareas específicas:** Un SLM adecuadamente entrenado o ajustado (*fine-tuned*) para un dominio o tarea concreta puede alcanzar un ren-

dimiento comparable al de un LLM de mayor tamaño dentro de ese nicho específico.

- **Ejemplos de especialización:** Los SLM especializados pueden ser adaptados a tareas específicas, como análisis de sentimiento, traducción automática, análisis de datos médicos, detección de fraude o atención al cliente.
- **Facilidad de personalización:** Ajustar un SLM para un área de conocimiento o tarea específica es más rápido, económico y requiere menos datos que adaptar un LLM, lo que hace que los modelos personalizados sean viables para implementar en un mayor número de empresas y organizaciones.

El interés en los SLM está creciendo a medida que el mercado de IA madura. Las empresas están pasando de los LLM generalistas a los SLM especializados, que son más fácilmente personalizables, eficientes y rentables para tareas específicas.

### 3.1.5 Privacidad y seguridad mejoradas

La capacidad de los SLM para operar localmente ofrece ventajas sustanciales en términos de privacidad y seguridad de los datos:

- **Procesamiento local (*On-Premise* / *On-Device*):** La eficiencia de los SLM, ya comentada en puntos anteriores, permite desplegarlos directamente en los dis-

positivos de los usuarios (*on-device*) o dentro de la infraestructura propia de una organización (*on-premise*). Esto es fundamental cuando se manejan datos sensibles, como información médica, financiera, datos personales de clientes o secretos comerciales, ya que evita la necesidad de enviar dicha información a servidores externos o plataformas en la nube gestionadas por terceros.

- **Minimización de riesgos:** Al mantener los datos dentro de un perímetro controlado (el dispositivo del usuario o la red de la empresa), el procesamiento local reduce significativamente las opciones de un potencial ataque y los riesgos asociados con la transmisión de datos a través de redes públicas y el almacenamiento en sistemas externos. Disminuyen las posibilidades de exposición accidental, filtraciones de datos, accesos no autorizados o violaciones de la privacidad.
- **Cumplimiento normativo:** Para organizaciones que operan en sectores altamente regulados, el procesamiento local facilita enormemente el cumplimiento de normativas estrictas de protección de datos como el Reglamento General de Protección de Datos (GDPR) en Europa u otras regulaciones sectoriales. La organización mantiene el control total sobre el flujo y almacenamiento de los datos, simplificando las auditorías y la demostración del cumplimiento.

- **Menor dependencia de APIs externas:** Como se mencionó previamente, el despliegue interno reduce la dependencia de servicios de terceros, lo cual no solo tiene implicaciones estratégicas y de coste, sino que también mejora la postura de seguridad al limitar la exposición a posibles vulnerabilidades en las plataformas externas.

### 3.2 Limitaciones

A pesar de sus numerosas ventajas, los SLM también presentan limitaciones inherentes que deben ser consideradas al evaluar su idoneidad para una aplicación determinada. Estas restricciones derivan principalmente de su tamaño reducido y su enfoque especializado.

#### 3.2.1 Menor capacidad generalista y de transferencia de conocimiento

La especialización que confiere a los SLM su potencia en tareas específicas también limita su amplitud:

- **Alcance de conocimiento limitado:** A diferencia de los LLM, que a menudo se entrenan con cantidades masivas de datos extraídos de internet, los SLM poseen una base de conocimiento del mundo más restringida y menos detallada. Pueden tener dificultades significativas para responder preguntas sobre temas generales diversos, comprender referencias culturales poco

comunes o realizar tareas que requieran una base de conocimiento fuera de su dominio de entrenamiento o ajuste.

- **Dificultad en tareas “Zero-Shot” o “Few-Shot”:** La capacidad de los LLM para generalizar y realizar tareas completamente nuevas para las que no han sido entrenados explícitamente (aprendizaje “zero-shot”, o con muy pocos ejemplos, “few-shot”) es una de sus características más notables. Los SLM, por lo general, muestran una capacidad inferior en este tipo de aprendizaje. Requieren, en la mayoría de los casos, un proceso de ajuste (*fine-tuning*) más concreto y con datos específicos para adaptarse con éxito a nuevos problemas o tareas.
- **Menor transferencia entre áreas de conocimiento:** Aplicar un SLM que ha sido altamente optimizado para un área de conocimiento específica (por ejemplo, análisis de contratos legales) a un área completamente diferente (como la interpretación de datos clínicos) probablemente resultará en una caída drástica del rendimiento. Su conocimiento está, por diseño, altamente especializado, y la transferencia de ese conocimiento a contextos no relacionados es limitada. Los LLM, con un entrenamiento más amplio, suelen ser más robustos en la transferencia de aprendizajes entre áreas distintas.



- **Dependencia del diseño y del caso de uso:** Es importante subrayar que estas limitaciones no implican una deficiencia per se, sino una consecuencia directa del enfoque de diseño de los SLM. Están concebidos para maximizar el rendimiento en contextos acotados y bien definidos, donde los requisitos de generalización son secundarios frente a la eficiencia, la privacidad o el control. En estos entornos, su especialización no solo es suficiente, sino deseable. Por tanto, el éxito de un SLM depende en gran medida de su alineación con el caso de uso concreto, más que de su capacidad de generalización universal.

Esta limitación representa la otra cara de la moneda de la especialización. El mismo enfoque que permite a los SLM alcanzar una alta eficiencia y precisión en un área concre-

ta restringe su capacidad para operar eficazmente fuera de ella. Cuando una aplicación requiere una comprensión amplia del mundo, la capacidad de responder a preguntas imprevistas sobre cualquier tema, o la flexibilidad para adaptarse sobre la marcha a tareas radicalmente nuevas sin entrenamiento adicional, las limitaciones generalistas de los SLM se hacen patentes. Es el clásico compromiso entre profundidad y amplitud.

### 3.2.2 Desempeño en tareas de alta complejidad y creatividad

Las tareas que requieren un alto grado de razonamiento abstracto o creatividad pueden exceder las capacidades actuales de muchos SLM:

- **Generación de texto menos refinada:** Si bien los SLM pueden generar texto coherente y útil para tareas específicas (como resúmenes, respuestas a preguntas frecuentes o descripciones de productos), generalmente no alcanzan el nivel de fluidez, riqueza, complejidad estructural y coherencia a largo plazo que pueden exhibir los LLM más avanzados en tareas de generación de texto abierto. Escribir artículos extensos y matizados, componer historias complejas con tramas o generar código sofisticado para sistemas complejos suelen ser tareas donde los LLM mantienen una ventaja significativa.
- **Razonamiento complejo limitado:** Las tareas que

requieren razonamiento complejo, comprensión profunda o planificación estratégica son difíciles para los SLM, los LLM suelen ser mejores en estas tareas. Aunque los SLM están mejorando en razonamiento, su solidez frente a problemas nuevos y complejos aún podría ser una ventaja de los LLM.

### 3.2.3 Dependencia crítica de datos de calidad para la especialización

El alto rendimiento de los SLM en tareas especializadas no surge espontáneamente, depende de manera crucial de los datos utilizados en su ajuste como la **calidad, precisión y representatividad del conjunto de datos con el que se entrenan o ajustan define los límites de su comprensión y capacidad operativa**. Al tratarse de modelos de menor escala, los SLM no disponen del respaldo de grandes volúmenes de información generalista que puedan compensar lagunas o errores en el entrenamiento. Por el contrario, su eficacia reside en el conocimiento profundo y específico que extraen de los datos concretos que se les proporcionan.

- **Necesidad de *Fine-Tuning* específico:** Para que un SLM alcance su máximo potencial en un área particular, requiere de un proceso de *fine-tuning* con datos relevantes y de alta calidad específicos para esa tarea o área. Si no se dispone de un conjunto de datos adecuado y correctamente validado, el modelo resultante

puede no alcanzar el nivel de desempeño deseado, limitando su utilidad práctica.

- **Sensibilidad al sesgo en los datos:** Al igual que cualquier modelo de IA, los SLM son susceptibles de heredar y potencialmente amplificar los sesgos presentes en los datos con los que se entrenan y ajustan. Dado que los conjuntos de datos utilizados para el *fine-tuning* suelen ser considerablemente más pequeños que los de preentrenamiento de los LLM, la representatividad y la equidad de estos datos adquieren una importancia aún mayor. Es fundamental implementar estrategias activas de mitigación de sesgos durante la preparación de los datos de *fine-tuning* para evitar que el SLM especializado genere resultados injustos, discriminatorios o perpetúe estereotipos.
- **Importancia crítica de la calidad de los datos:** La calidad general de los datos de *fine-tuning* es primordial. Datos incompletos, inconsistentes o irrelevantes derivarán a un modelo con bajo rendimiento, propenso a errores y alucinaciones. Por ello, las prácticas revisión y validación de datos y el uso de datos sintéticos de alta calidad generados cuidadosamente son esenciales.

### 3.2.4 Madurez del ecosistema y estandarización

El entorno de herramientas, modelos pre-entrenados y generación de comunidad que rodea a los SLM está evolu-

cionando rápidamente, pero puede presentar diferencias respecto al ecosistema más establecido de los LLM más populares.

- **Estado actual del ecosistema:** Se observa un crecimiento exponencial en la disponibilidad de SLM de código abierto de alta calidad. Plataformas como *Hugging Face* se han convertido en referencias para descubrir, compartir y colaborar en el desarrollo de estos modelos, ofreciendo también herramientas para su despliegue y evaluación. La comunidad está en crecimiento y participa activamente.
- **Comparación con LLM:** A pesar de este rápido crecimiento, en algunos nichos específicos o para ciertas herramientas de desarrollo, MLOps o de evaluación muy especializadas, el ecosistema de SLM podría percibirse como menos maduro o estandarizado en comparación con la infraestructura robusta y el soporte comercial que rodea a los LLM ofrecidos como servicios API por grandes proveedores.
- **Implicaciones para la adopción:** Esta posible falta de madurez o estandarización en ciertas áreas puede traducirse en una curva de aprendizaje más pronunciada o en la necesidad de un mayor esfuerzo de desarrollo, integración y mantenimiento por parte de las organizaciones que optan por construir soluciones basadas

en SLM, en comparación con la relativa facilidad de uso de algunas plataformas LLM comerciales.



# 4. CASOS DE USO

Los Small Language Models (SLM) están empezando a cambiar la manera en que diferentes industrias usan la inteligencia artificial. A diferencia de los modelos gigantes (LLM), estos sistemas más pequeños son ágiles, fáciles de adaptar y no requieren tantos recursos para funcionar. Eso los hace ideales para resolver problemas concretos en la vida real. En este capítulo veremos cómo distintas áreas están encontrando en los SLM una herramienta práctica y cercana para innovar.

## 4.1 APLICACIONES COMUNES DE LOS SLM

Los SLM han demostrado ser versátiles y eficientes para una gran variedad de aplicaciones. A continuación, se incluye un listado no exhaustivo de aplicaciones más comunes de los SLM.

### 1. Chatbots y asistentes virtuales

- **Resolución de preguntas frecuentes (FAQ)**
- **Atención al cliente 24/7** con respuestas rápidas y contextualmente adecuadas.
- **Gestión de solicitudes rutinarias** para liberar carga del personal humano.

### 2. Análisis de sentimientos (*Sentiment Analysis*)

- **Monitoreo de reputación** en redes sociales y foros.

- **Mejora de la experiencia de cliente** al detectar rápidamente percepciones negativas o positivas sobre productos o servicios.
  - **Detección de estados emocionales** (positivos, neutros, negativos) para planes de marketing o campañas específicas.
- ### 3. Clasificación de textos
- **Filtrado de spam o contenido no deseado.**
  - **Categorización de documentos** de acuerdo con temas, dominios o niveles de confidencialidad.
  - **Automatización de procesos administrativos**, por ejemplo, el enrutamiento de correos electrónicos o solicitudes a departamentos correspondientes.

### 4. Otras tareas específicas

- **Reconocimiento de entidades nombradas (NER)**: extracción de nombres propios, organizaciones o lugares de documentos.
- **Resumen de textos (*text summarization*)**: obtener versiones reducidas de informes, artículos o publicaciones extensas.
- **Traducción automática** (limitada a dominios específicos, con vocabulario focalizado).

## 4.2 APLICACIONES ESPECÍFICAS POR INDUSTRIA

En esta sección se desarrollan potenciales casos de uso de los SLM en diferentes sectores o industrias, describiendo su aplicación concreta.

### 4.2.1 Telecomunicaciones

Los SLM pueden optimizar diversos procesos específicos en el ámbito de telecomunicaciones, como son el caso del análisis de llamadas, la personalización de servicios y la gestión de consultas y soporte técnico.

#### CASOS DE USO

Descripción de potenciales casos de uso en el ámbito de Telecomunicaciones.

#### 1. ANÁLISIS DE LLAMADAS

##### Transcripción y clasificación de llamadas

- **Transcripción automática local:** Un SLM de reconocimiento de voz y procesamiento de texto, entrenado o ajustado para vocabulario propio de la empresa de telecomunicaciones, puede transcribir las llamadas en tiempo real sin requerir gran potencia de cómputo.
- **Clasificación y etiquetado:** Mediante técnicas de *clasificación de textos*, se puede etiquetar la llamada según su contenido (quejas, solicitudes de soporte, cambios de plan, etc.) de forma automática.
- **Ejemplo práctico:** Una operadora recibe miles de llama-

das diarias. Cada llamada se transcribe en tiempo real gracias al modelo local (sin mandar datos a la nube) y luego se etiqueta. Así, se identifican patrones frecuentes (por ejemplo, zonas con problemas de señal, quejas sobre facturación) y se genera un reporte inmediato para el equipo técnico, mejorando la capacidad de respuesta.

##### Análisis de sentimientos y satisfacción

- **Detección de emociones:** Un SLM especializado en análisis de sentimiento puede detectar si el cliente está irritado, confuso o satisfecho, lo cual desencadena protocolos de atención específicos.
- **Seguimiento de calidad:** Los supervisores pueden obtener métricas de satisfacción de forma rápida y en tiempo real, sin necesidad de revisar manualmente la conversación completa.
- **Ejemplo práctico:** Una vez transcrita la llamada, el SLM identifica palabras clave que indiquen frustración (repetitivas menciones de "no funciona", "error", "muy lento"), activando una alerta en el panel de un supervisor para intervenir o realizar un seguimiento más personalizado.

#### 2. PERSONALIZACIÓN DE SERVICIOS

##### Recomendaciones de planes y paquetes

- **Perfilado de uso:** Un SLM puede procesar texto rela-

cionado con el historial del cliente, sus interacciones previas y preguntas frecuentes, para generar un perfil de uso (tiempo de llamadas, datos consumidos, uso de *roaming*, etc.).

- **Ofertas personalizadas:** Basado en dicho perfil, se sugieren planes optimizados (por ejemplo, agregar más datos móviles a quien muestra alto consumo en redes sociales o videollamadas).
- **Ejemplo práctico:** Un *chatbot*, integrado en la aplicación de la compañía telefónica, detecta que un cliente pregunta con frecuencia sobre "internet adicional". El SLM, conociendo el historial de consumo y el contexto de las consultas, ofrece directamente un plan con más gigas o paquetes temporales para cubrir la necesidad puntual sin saturar la factura mensual.

##### Chatbots contextuales y asistencia en tiempo real

- **Asistentes virtuales ligeros:** Un SLM puede integrarse en la web o app de la operadora para resolver consultas de forma rápida (horarios, procedimientos para reportar robo de dispositivo, etc.).
- **Contexto individualizado:** El asistente "recuerda" interacciones anteriores y propone soluciones acordes al historial del usuario, mejorando la experiencia global.
- **Ejemplo práctico:** Un cliente inicia una sesión de chat solicitando ayuda sobre la configuración del router.

El *chatbot*, gracias al contexto del historial del cliente (router adquirido, tipo de conexión, incidencias previas), provee instrucciones detalladas y simples, reduciendo la necesidad de contactar a un agente humano.

### 3. GESTIÓN DE CONSULTAS Y SOPORTE TÉCNICO

#### Enrutamiento inteligente de consultas

- **Clasificación automática de tickets:** Cuando un cliente envía un correo o rellena un formulario de asistencia, un SLM asigna la consulta al departamento correcto (facturación, soporte técnico, alta de servicios, etc.)
- **Automatización de respuestas básicas:** Muchas dudas recurrentes (cómo recuperar el PIN, desbloquear el teléfono, revisar facturas) pueden resolverse con respuestas preconfiguradas, reduciendo la carga en el *call center*.
- **Ejemplo práctico:** En lugar de distribuir manualmente miles de correos de usuarios, un SLM clasifica cada uno en función del texto y deriva el caso directamente al área apropiada. Por ejemplo, si detecta menciones sobre "errores en la factura" o "cargos duplicados", envía la consulta a la sección de facturación. Esto disminuye el tiempo de resolución y la saturación de equipos.

#### Soporte a agentes de atención al cliente

- **Sugerencias de diagnóstico:** Durante la llamada, el SLM puede "escuchar" la conversación (transcrita) y sugerir al agente soluciones basadas en la base de conocimiento interna.
- **Resumen de la incidencia:** Al finalizar la llamada, el SLM genera automáticamente un breve resumen de lo ocurrido (incidencia, pasos de solución, nivel de satisfacción del cliente), agilizando la documentación.
- **Ejemplo práctico:** Un agente atiende una queja de baja velocidad de internet. El SLM analiza la transcripción y, al reconocer expresiones como "mi internet está lento desde ayer", "descargas muy lentas", sugiere al agente revisar la configuración del módem, posible saturación de la red o problemas geográficos en la zona. El agente ve las sugerencias en pantalla y avanza rápidamente en la solución.

### 4. ANÁLISIS DE DATOS DE REDES DE VOZ

En esta sección se describe en mayor profundidad y detalle un caso de uso particular dentro del ámbito de Telecomunicaciones, como es la monitorización y optimización de redes SIP.

#### Las comunicaciones de voz hablan SIP

Seguramente el término VoIP resulte familiar (*Voice over Internet Protocol*), al menos mucho más familiar que el



protocolo SIP (*Session Initiation Protocol*). VoIP se refiere a la tecnología que permite hacer llamadas telefónicas a través de una conexión de banda ancha, usando internet, en lugar de una línea telefónica analógica. La mayor parte de los operadores de telefonía móvil utilizan hoy en día VoIP, también las llamadas de voz desde WhatsApp usan esta tecnología.

Pues bien, el protocolo SIP es el *standard* (definido por IETF hace más de 25 años) que permite establecer una llamada VoIP entre dos o más usuarios, estableciendo y terminando sesiones entre dos o más dispositivos de comunicación (que ya no tienen por qué ser exclusivamente teléfonos, si no que pueden ser también *tablets*, ordenadores, dispositivos IOT, ... en definitiva, potencialmente, cualquier dispositivo conectado a Internet). Estas sesiones SIP pueden incluir llamadas de voz, videoconferencias u otras sesiones de comunicación multimedia. Este proceso, implica a su vez diversos protocolos SIP como SDP (*Session Description Protocol*); UDP (*User Datagram Protocol*), RTP (*Real-time Transport Protocol*).

### Protocolo SIP: diseñado para la comunicación eficiente entre máquinas

SIP es un protocolo basado en texto, que define una serie de mensajes que se utilizan para establecer una sesión multimedia: como INVITE (para invitar a otro dispositivo a una sesión), ACK (para confirmar que se ha recibido un

mensaje), BYE (para finalizar una sesión), y otros mensajes, incluso numéricos, para mostrar el estado de la llamada (p.e. mensajes de error en el caso de ausencia de respuesta por la otra parte ante un intento de establecimiento de llamada).

SIP también se encarga de negociar los parámetros de la sesión, tales como el tipo de códec a utilizar para la transmisión de audio o video; y los detalles de la sesión, como el número de teléfono del destinatario (o del originador) de la llamada, que puede ser alterado por los diferentes elementos intermedios que procesan este tipo de tráfico.

A pesar de tratarse de texto, es un desafío para un humano (incluso para uno entrenado), poder interpretar esa información con la agilidad y escala que una infraestructura de telecomunicaciones requiere. Para un humano puede ser arduo interpretar grandes volúmenes de mensajes con un formato de este estilo:

```
INVITE sip:14107154522807@10.105.80.174:5060 SIP/2.0
Via: SIP/2.0/UDP 10.105.80.114:5060;branch=z9hG-4bK98e4117d52a6
From: "Solihull" <sip:01214248526@10.105.80.114>;-tag=25526~ffa80926-5fac-4dd6-b405-2dbb-c56ae9a2-551664735
To: <sip:14107584528207@10.105.80.174>
Date: Mon, 02 Apr 2012 18:12:31 GMT
```

```
Call-ID: 68781700-f791ec0f-2d26-e28690a@10.105.80.114
Supported: timer,resource-priority,replaces
Min-SE: 1800
User-Agent: Cisco-CUCM8.6
Allow: INVITE, OPTIONS, INFO, BYE, CANCEL, ACK, PRACK, UPDATE, REFER, SUBSCRIBE, NOTIFY
CSeq: 101 INVITE
Expires: 180
Allow-Events: presence, kpml
Supported: X-cisco-srtp-fallback
Supported: Geolocation
Call-Info: <sip:10.105.80.114:5060>;method="NOTIFY;Event=telephone-event;Duration=500"
Cisco-Guid: 1752700672-0000065536-0000007823-0237529354
Session-Expires: 84600
Contact: <sip:01214248526@10.105.80.114:5060>
Max-Forwards: 70
Content-Length: 0
Content-Type: application/sdp
Content-Length: 238
```

Como se puede apreciar en el ejemplo, se trata de un texto con bastante densidad de información, bien estructura-

da y estandarizada. Una información muy valiosa que se almacena (SIP logs o CDRs) para identificar eventos relevantes como intentos de acceso no permitidos; fraude telefónico; intentos de establecimiento de llamada frustrados; detección de *robocalls* o simplemente como referencia para la facturación de servicios de voz.

Además, cada sesión (una llamada, por ejemplo) da lugar al intercambio de multitud de mensajes SIP. A escala de un operador esto pueden dar lugar a Tb de información en muy poco tiempo:

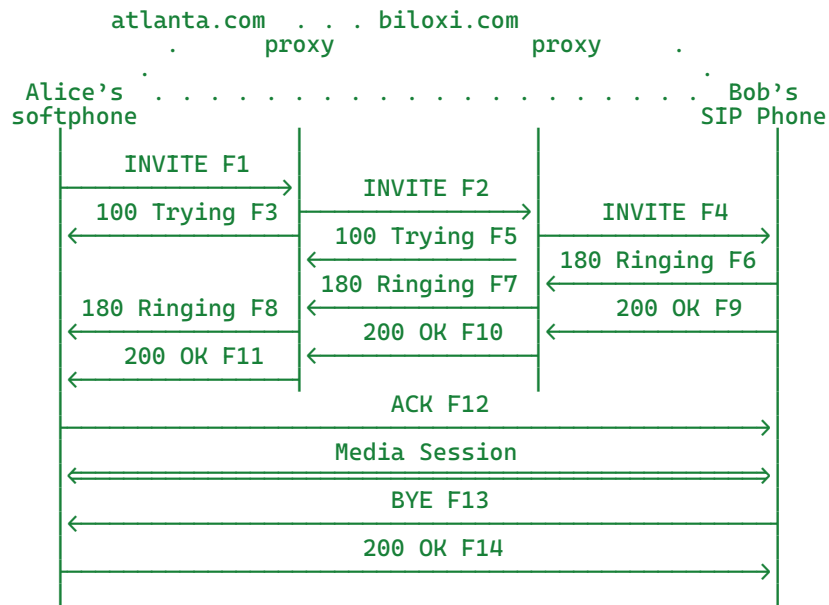


Figure 1: SIP session setup example with SIP trapezoid

En este ejemplo, de una simple llamada, se puede apreciar cómo cada línea da lugar a un mensaje SIP.

Es fácil imaginarse, el enorme volumen de mensajes que se pueden llegar a generar en una organización con varios miles de trabajadores, con llamadas ocurriendo en paralelo, y con la intervención de elementos de red como PBX (centralitas) o SBCs, que a su vez añaden más información a estos mensajes.

Si esta complejidad no resultase suficiente, hay que entender que, aunque SIP es un protocolo estándar y abierto, lo que significa que puede ser implementado en una amplia variedad de plataformas y dispositivos, las implementaciones son dispares (SIP over UDP, SIP over TCP, SIP over TLS) por lo que es preciso elementos intermedios que permitan interoperar entre dispositivos.

### SLM para la comunicación entre la red de telecomunicaciones y su administrador

Los SLM están adquiriendo tracción en el ámbito de la traducción de idiomas, ya que hacen posible la traducción instantánea, ya sea de conversaciones o imágenes (p.e. carteles). Al manejar menos parámetros que un LLM, sus tiempos de respuesta son mucho más rápidos (algo crítico en contextos que requieren baja latencia como el de las comunicaciones en tiempo real), sus costes más bajos y su adaptación a dominios especializados ofrecen un mejor rendimiento.



Precisamente, cuando hablamos de comunicaciones hombre-máquina, el desarrollo de un SLM especialmente orientado a la interpretación del protocolo SIP, es un caso de uso que se adapta perfectamente a las características de un SML.

Entendiendo las características del protocolo SIP, podemos diseñar una serie de *prompts*, que pueden estar predefinidos en el sistema, agilizando al máximo las respuestas a consultas recurrentes (tabla 1).

Es en este punto en el que el desarrollo de una interfaz con la información en lenguaje natural puede resultar crítica para el administrador de la red, reduciendo la exigencia de experiencia previa en SIP.

### Optimización de redes SIP y monitorización

El fin último de esta interfaz basada en un SLM es la optimización de la red. En este ámbito ya existen aplicaciones de tecnologías habilitadoras como el *big data*, la integración de otras fuentes de información mediante APIs, indicadores SNMP (*Simple Network Monitoring Protocol*), ... que expanden aún más el potencial del uso de IA, tanto en la identificación de eventos relevantes como en la optimización de la red a partir de dicha información.

El principal beneficio de la adopción de un SML es que muchas de estas operaciones caían del lado de sistemas automatizados que no son 100% transparentes para el ad-

Tabla 1

CARACTERÍSTICAS DEL PROTOCOLO SIP	PROMPTS DEFINIDOS PARA EL SML
<b>Media Negotiation - SDP (Session Description Protocol)</b>	<ul style="list-style-type: none"> <li>• ¿No hay dispositivos compatibles en la red?</li> <li>• ¿Cuál es el códec más común en la red?</li> <li>• ¿Solo hay usuarios de audio?</li> </ul>
<b>Security - SRTP (Secure Real-time Transport Protocol) and TLS (Transport Layer Security)</b>	<ul style="list-style-type: none"> <li>• ¿Cuántos intentos no autorizados has detectado?</li> <li>• ¿Hay medios sin cifrar?</li> <li>• ¿No hay mensajes de señalización cifrados?</li> </ul>
<b>Presence</b>	<ul style="list-style-type: none"> <li>• ¿Hay usuarios disponibles en la red?</li> <li>• ¿Cuánto tiempo ha estado un usuario disponible en la red?</li> <li>• ¿Cuál es el rango de horas disponible más frecuente para los usuarios?</li> </ul>
<b>Conferencing - multi-party audio, video, and screen sharing sessions</b>	<ul style="list-style-type: none"> <li>• ¿Cuántas veces se comparte la pantalla?</li> <li>• ¿Cuántos usuarios hay en la conferencia?</li> </ul>

ministrador, quien se ve obligado a recurrir al análisis de datos históricos, o al empleo de KPIs poco intuitivos como los definidos por el IETF (p.e. latencias en las respuestas de mensajes ISP), para obtener una imagen más precisa de lo sucedido, siendo esta actividad exigente en términos de tiempo y recursos de análisis.

Un SLM dedicado al protocolo SIP democratizaría el uso de los datos generados por la red, permitiendo un buen aprovechamiento de esta incluso por parte de perfiles

técnicos más inexpertos. Además, permite relacionarse con la información en vivo, facilitando la intervención sobre sesiones en curso (por ejemplo, expulsar de la red a un *"user agent"* sospechoso de estar realizando llamadas fraudulentas).

### SLM y privacidad

La posibilidad de desplegar estos modelos en infraestructuras tecnológicamente más accesibles y económicas, permiten mantener los datos a salvo en los servidores

del proveedor de servicio, evitando la circulación de información sensible a través de diferentes infraestructuras (como centros de datos especializados en el procesado de información).

## **BENEFICIOS**

Los principales beneficios de los SLM en Telecomunicaciones son los siguientes.

### **Eficiencia operativa**

- Menor tiempo de espera para el cliente.
- Mayor agilidad en la clasificación de incidencias y derivación al departamento adecuado.
- Ahorro en recursos humanos y optimización del *call center*.

### **Mejora de la experiencia de cliente**

- Respuestas más rápidas y personalizadas, aumentando la satisfacción global.
- Reducción de la frustración al disminuir repeticiones de pasos o cuestionarios innecesarios.
- Posibilidad de atender consultas 24/7 sin un gran coste adicional.

### **Toma de decisiones basada en datos**

- Análisis de grandes volúmenes de información textual (transcripciones, chats, correos) para detectar patro-

nes de incidencias frecuentes o necesidades emergentes.

- Identificación proactiva de áreas de mejora en productos y servicios, a partir de indicadores de sentimiento y palabras clave en llamadas.

### **Coste reducido y escalabilidad**

- SLM requieren menos recursos que los modelos gigantes, por lo que pueden ejecutarse localmente o en infraestructura moderada, facilitando la adopción en compañías de distintos tamaños.
- Su menor complejidad permite un mantenimiento más sencillo, una actualización rápida de vocabulario y la integración con otros sistemas corporativos sin grandes desarrollos.

### **Privacidad y cumplimiento**

- Al ejecutar o *fine-tunear* modelos de menor tamaño dentro de la propia infraestructura, se limita el envío de datos sensibles a servidores externos, lo que ayuda a cumplir normas y regulaciones de protección de datos.

## **CONCLUSIONES**

La adopción de *Small Language Models* en el sector de las telecomunicaciones brinda soluciones ligeras y eficientes para optimizar procesos tan diversos como la transcrip-

ción y el análisis de llamadas, la personalización de servicios o la gestión de consultas de soporte. Estas capacidades permiten:

- **Mejorar la experiencia del cliente**, al ofrecer respuestas rápidas y personalizadas.
- **Elevar la eficiencia operativa**, al reducir costes y acelerar la gestión de incidencias.
- **Fomentar la toma de decisiones basada en datos**, mediante el análisis automatizado de grandes volúmenes de información textual.

Gracias a su bajo consumo de recursos y facilidad de despliegue, los SLM resultan especialmente adecuados para entornos de telecomunicaciones que buscan soluciones escalables, rentables y alineadas con requisitos de privacidad y rendimiento en tiempo real.

### **4.2.2 Aprendizaje y Desarrollo (L&D)**

Los SLM optimizan varios procesos en el ámbito del Aprendizaje y Desarrollo como la automatización de tutorías; la personalización de contenidos y la gestión de consultas.

## **CASOS DE USO**

Descripción de potenciales casos de uso en el ámbito de Aprendizaje y Desarrollo.

### **1. Chatbots y asistentes virtuales de aprendizaje**

### Tutorías automatizadas y resolución de dudas

- **Asistentes virtuales ligeros:** Un SLM puede integrarse en plataformas LMS (*Learning Management Systems*) para ofrecer asistencia instantánea a empleados/participantes durante cursos de formación, respondiendo preguntas frecuentes y guiando en procedimientos habituales.
- **Guía interactiva y personalizada:** Los SLM son capaces de personalizar las respuestas en función del progreso y perfil del empleado, mejorando la experiencia de aprendizaje.
- **Ejemplo práctico:** Un empleado inicia un curso de ciberseguridad y está teniendo problemas accediendo a la evaluación final. El SLM, integrado en la plataforma, ofrece instrucciones claras y personalizadas de cómo realizar dicho trámite, reduciendo la necesidad de soporte humano.

### Soporte en tiempo real y contextualizado

- **Chatbots** capaces de recordar interacciones anteriores y ofrecer respuestas adaptadas al contexto del usuario.
- **Ejemplo práctico:** Un empleado retoma un curso de formación online tras varias semanas sin actividad. El *chatbot*, recordando el punto donde quedó, sugiere re-

tomar desde ese módulo y ofrece materiales complementarios o actualizados.

### 2. Análisis de desempeño y personalización del aprendizaje

#### Seguimiento automatizado del progreso

- **Monitorización de avance:** Un SLM analiza las interacciones y progreso del empleado, generando reportes personalizados.
- **Adaptación de itinerarios:** En función del desempeño de cada persona, sugiere nuevos módulos, refuerzos o contenidos complementarios.
- **Ejemplo práctico:** Un empleado que falla repetidamente en un módulo recibe recomendaciones personalizadas y acceso a materiales adicionales para reforzar ese contenido.
- Evaluación de conocimientos con retroalimentación instantánea
- **Personalización de las interacciones:** Los SLM ayudan a generar preguntas adaptadas al nivel del usuario y ofrecen retroalimentación inmediata.
- **Ejemplo práctico:** Tras completar una lección, el SLM presenta una prueba de nivel personalizada. En función de las respuestas y del resultado obtenido, sugiere volver a revisar algunos puntos.



### 3. Generación y gestión de contenido educativo

#### Creación automatizada de materiales

- **Generación** de resúmenes, guías y cuestionarios adaptados al perfil del usuario.
- **Ejemplo práctico:** Al actualizar una política interna, el SLM genera automáticamente un resumen formativo y preguntas para evaluar su comprensión.
- Traducción y adaptación a distintos perfiles
- **Segmentación:** SLM adaptan contenidos para distintos niveles de conocimiento o roles dentro de la empresa.

- **Ejemplo práctico:** Un mismo curso sobre ciberseguridad se adapta automáticamente para perfiles técnicos y no técnicos, simplificando o profundizando según el perfil del destinatario.

#### 4. Evaluación y retroalimentación automatizada

##### Análisis de encuestas y *feedback*

- **Procesamiento de respuestas** a encuestas para detectar puntos débiles y áreas de mejora en programas formativos.
  - **Ejemplo práctico:** Tras la realización de un curso por un determinado colectivo, el SLM analiza comentarios abiertos de los empleados, detectando menciones frecuentes a dificultades técnicas o falta de ejemplos prácticos.
- #### 5. Recomendaciones personalizadas
- **Generación automática** de sugerencias para próximos cursos o módulos, basadas en el historial y desempeño del usuario.
  - **Ejemplo práctico:** Un empleado que completó exitosamente un curso sobre protección de datos recibe la recomendación de un módulo avanzado relacionado con esta materia.

#### **BENEFICIOS**

Los principales beneficios de los SLM en aprendizaje y desarrollo son los siguientes.

##### 1. Eficiencia operativa

- Reducción del tiempo de respuesta ante consultas frecuentes.
- Automatización de tutorías y evaluaciones personalizadas.
- Ahorro en recursos humanos al disminuir la carga sobre los equipos de formación.

##### 2. Mejora de la experiencia de aprendizaje

- Respuestas personalizadas y en tiempo real.
- Itinerarios adaptativos según desempeño y necesidades del empleado.
- Acceso a tutorías automatizadas 24/7.

##### 3. Toma de decisiones basada en datos

- Análisis de interacciones y progresos para identificar patrones de aprendizaje.
- Detección temprana de dificultades formativas y necesidades emergentes.

##### 4. Coste reducido y escalabilidad

- Los SLM funcionan con menos recursos que modelos grandes, facilitando su integración en plataformas LMS.
- Fácil mantenimiento y actualización de contenidos.
- Adaptación a distintos tamaños de empresa.

##### 5. Privacidad y cumplimiento

- Ejecución local dentro de las plataformas corporativas, evitando la exposición de datos sensibles.
- Cumplimiento con normativas de protección de datos al evitar enviar información a servidores externos.

#### **CONCLUSIONES**

La adopción de SLM en el sector de Aprendizaje ofrece soluciones ligeras y eficientes que permiten a las empresas mejorar la experiencia formativa, haciendo que los empleados adquieran las capacidades, más adecuadas a sus necesidades formativas, y mejoren sus resultados.

Entre las ventajas que ofrecen los SLM encontramos:

- Facilitar el acceso a la formación de forma individualizada
- Resolver dudas en tiempo real.
- Optimizar procesos formativos, reduciendo costes y plazos.

- Mejorar la toma de decisiones y la calidad de la formación mediante el análisis de datos.
- Gracias a su bajo consumo de recursos y facilidad de despliegue e integración, los SLM son ideales para organizaciones que buscan escalar la formación de manera rentable, eficiente y alineada con las necesidades del negocio.

#### 4.2.3 Banca y servicios financieros

Los SLM pueden tener aplicación en procesos concretos del sector de banca, como es el caso de la atención al cliente, la detección de fraude o el análisis de sentimiento.

#### CASOS DE USO

Descripción de potenciales casos de uso en el ámbito de Banca y servicios financieros.

##### 1. Chatbots para atención al cliente

- **Asistentes de cuenta:** Permiten a los usuarios consultar sus saldos, últimos movimientos o información básica de productos financieros sin necesidad de ingresar al portal web.
- **Soporte a solicitudes simples:** Por ejemplo, bloquear una tarjeta o solicitar un nuevo PIN a través del *chatbot*.

##### 2. Detección de fraude y clasificación de transacciones



- **Filtrado de transacciones sospechosas:** Clasifican movimientos bancarios según patrones de riesgo o comparan descripciones de transacciones para identificar anomalías.
  - **Alertas en tiempo real:** Un SLM entrenado en datos financieros específicos puede procesar grandes volúmenes de textos (descripciones, reportes) y emitir alertas cuando detecta indicios de actividades fraudulentas.
- ##### 3. Análisis de sentimiento en redes y foros especializados
- **Monitoreo de la reputación de marca:** Identificar qué opinan los clientes sobre un banco o compañía de seguros en redes sociales o portales de quejas.

- **Evaluación de riesgo crediticio:** Aunque es un uso más avanzado, algunos bancos pueden analizar lenguaje de solicitudes y correos para detectar señales tempranas de riesgos.

#### BENEFICIOS

Los SLM son útiles en Banca y servicios financieros pues permiten:

- **Eficiencia en entornos con alta seguridad:** Un modelo pequeño puede ejecutarse de manera local dentro de la infraestructura bancaria, cumpliendo requisitos de privacidad y confidencialidad.
- **Costes controlados:** Implementar soluciones basadas en SLM reduce la necesidad de grandes servidores o costosas APIs externas.
- **Cumplimiento normativo:** El procesamiento local minimiza la exposición de datos sensibles a entornos externos, factor clave en el cumplimiento de normativas como GDPR.

#### CONCLUSIONES

Los Small Language Models se han convertido en una herramienta clave para múltiples industrias debido a su eficiencia, menor demanda de recursos y facilidad de despliegue. En el caso particular de Banca, ayudan a automatizar la atención al cliente, clasificar transacciones y

reforzar la detección de fraude sin exponer información sensible.

#### 4.2.4 Salud y bienestar (Healthcare)

Los SLM pueden contribuir a facilitar procesos específicos en el ámbito de Salud, como es el caso de *chatbot* y triaje virtual, la gestión de documentación clínica o el análisis de sentimiento.

##### CASOS DE USO

Descripción de potenciales casos de uso en el ámbito de Salud y bienestar.

##### 1. Chatbots médicos básicos y triaje virtual

- **Recolección de síntomas primarios:** Un asistente virtual puede guiar al paciente con preguntas sobre su estado, clasificar la urgencia y derivarlo al especialista adecuado.
- **Recordatorios y seguimiento:** Notificaciones de citas médicas, recordatorios de toma de medicamentos o seguimiento postoperatorio.

##### 2. Clasificación y análisis de documentación clínica

- **Gestión de historias clínicas:** Los SLM pueden etiquetar o clasificar documentos médicos según su contenido (resultados de laboratorio, informes radiológicos, notas de evolución, etc.).

- **Filtrado de contenido relevante:** Permite a los profesionales encontrar rápidamente la información importante en grandes repositorios de documentos.

##### 3. Análisis de sentimiento en encuestas de satisfacción del paciente

- **Monitoreo de la experiencia de pacientes** para mejorar la calidad del servicio.
- **Detección temprana de descontento o problemas** en la atención, facilitando la intervención proactiva.

##### BENEFICIOS

Los SLM son útiles en Salud pues facilitan:

- **Privacidad y cumplimiento legal:** La posibilidad de ejecutar modelos localmente (en un servidor *on-premise* de la institución médica) reduce la necesidad de enviar datos sensibles a la nube.
- **Bajo consumo de recursos:** Clínicas y hospitales pueden desplegar estos sistemas sin requerir grandes inversiones en supercomputadoras.
- **Respuesta en tiempo real:** Ideal para triaje virtual y asistencia rápida a pacientes.

##### CONCLUSIONES

En el dominio de Salud y bienestar los SLM permiten de-

sarrollar sistemas de *triaje* virtual, análisis de documentos clínicos y seguimiento de pacientes sin comprometer datos privados.

#### 4.2.5 Retail y comercio electrónico

Los SLM pueden incrementar la eficiencia de diferentes procesos concretos en el dominio de *retail* y comercio electrónico, como la atención postventa, el análisis de opiniones o la categorización de los productos.

##### CASOS DE USO

Descripción de potenciales casos de uso en el ámbito de *retail* y comercio electrónico.

##### 1. Chatbots de atención postventa y soporte

- **Asistentes de compra:** Responden preguntas frecuentes sobre productos, estado de órdenes, devoluciones y reembolsos.
- **Recomendaciones personalizadas:** Basadas en un conjunto acotado de productos y preferencias del cliente.
- Análisis de opiniones de productos (reviews)
- **Detección de sentimiento en reseñas:** Distingue entre críticas negativas y positivas de manera rápida y ayuda a categorizar las opiniones para generar informes de satisfacción.

- **Identificación de tendencias o problemas comunes:** Permite detectar fallos en la cadena de suministro o en la calidad de productos.
- ## 2. Categorización de productos y descripciones
- **Clasificación automática en categorías y subcategorías:** Apoyo en la gestión de inventario y la mejora de la experiencia de búsqueda para el cliente.
  - **Generación de descripciones simplificadas:** Para *marketplaces* con gran variedad de artículos, un SLM puede ayudar a crear descripciones cortas y claras.

### **BENEFICIOS**

Los SLM son útiles en *Retail* y comercio electrónico pues habilitan:

- **Escalabilidad:** Muchas empresas de *retail* poseen catálogos dinámicos y requieren procesar grandes cantidades de datos de forma rápida. Los SLM permiten hacerlo sin necesitar infraestructura tan potente como la que exigen los LLM.
- **Menor latencia para atención al cliente:** El despliegue local o con pocos recursos otorga respuestas prácticamente instantáneas, mejorando la experiencia del usuario.
- **Reducción de costos operativos:** Un modelo más pe-

queño y optimizado evita invertir en servidores costosos y reduce costos de mantenimiento.

### **CONCLUSIONES**

En el ámbito de *Retail* y comercio electrónico los SLM permiten dar soporte a *chatbots* de atención al cliente, análisis de reseñas y categorización de productos, contribuyendo a la automatización y mejora de la experiencia de compra.



# 5. RETOS Y OPORTUNIDADES TECNOLÓGICAS

## 5.1 RETOS TECNOLÓGICOS

Los SLM representan un gran paso adelante en la evolución de la inteligencia artificial, ya que ofrecen soluciones más ligeras, rápidas y accesibles, al tiempo que mantienen un rendimiento competitivo. Gracias a su flexibilidad y a sus reducidos requisitos de recursos, los SLM abren nuevas perspectivas para una gran variedad de aplicaciones, desde entornos con recursos limitados hasta industrias preocupadas por la sostenibilidad. A medida que evolucionen las tecnologías, estos modelos prometen desempeñar un papel central en el futuro de la IA.

Podemos decir que los principales retos tecnológicos y desafíos técnicos a los que se enfrentan los SLM son los siguientes:

- **Sesgo y equidad:** Los SLM pueden tener los mismos o parecidos sesgos que tienen los LLM, al entrenarse con la misma naturaleza de datos existentes. Por lo tanto, es muy importante desarrollar métodos para minimizar los posibles sesgos tanto en los modelos de entrenamiento como en los propios modelos.
- **Calidad y relevancia de los datos:** Los SLM, a diferencia de los LLM, dependen más de la calidad y relevancia de los conjuntos de datos que de su volumen. Esto significa que la limpieza, la especialización y la selección de datos son cruciales para el rendimiento del SLM.
- **Gestión eficiente de datos:** Esta gestión es esencial (incluyendo la selección, almacenamiento y procesamiento) para que los SLM puedan operar de manera óptima.
- **Personalización y adaptación:** La capacidad de personalizar los SLM y adaptarlos a tareas y dominios específicos es un desafío importante.
- **Accesibilidad:** Aunque los SLM son más accesibles que los LLM en términos de recursos computacionales, la necesidad de acceso a Internet de alta velocidad y dispositivos adecuados sigue siendo un desafío, especialmente en áreas rurales y comunidades con menos recursos.
- **Ciberseguridad y privacidad:** La protección de la información y la privacidad de los datos son fundamentales para el desarrollo y la implementación de cualquier tecnología, incluyendo los SLM.
- **Infraestructura:** La necesidad de una infraestructura adecuada para el procesamiento y almacenamiento de datos, así como para la comunicación entre dispositivos, es un reto constante en el desarrollo de la tecnología.
- **Alucinaciones:** Al igual que sucede con los LLM, es vital validar los resultados de los SLM para asegurarse de que lo que producen es objetivamente correcto.



## 5.2 OPORTUNIDADES DE EXPANSIÓN Y MEJORA

La evolución de los SLM en combinación con LLM representa una de las principales áreas y oportunidades de expansión de los SLM.

Todos los desarrollos de IA realizados hasta ahora podemos decir que han dado lugar a enfoques de optimización que maximizan el poder conjunto de los LLM y los SLM, por ejemplo:

- **Patrón de IA híbrida:** Un modelo de IA híbrida puede hacer que modelos más pequeños se ejecuten en el

entorno local y accedan a los LLM de la nube pública cuando se necesita un corpus de datos más grande para responder a una instrucción.

- **Enrutamiento inteligente:** El enrutamiento inteligente puede aplicarse para distribuir de forma más eficiente las cargas de trabajo de la IA. Se puede crear un módulo de enrutamiento que acepte consultas, las evalúe y elija el modelo más adecuado al que dirigir las consultas. Los modelos de lenguaje pequeños pueden gestionar las consultas básicas, mientras que los mo-

delos de lenguaje de gran tamaño pueden abordar las más complicadas.

## 5.3 FUTURO DE LOS SLM EN LA IA GENERATIVA

El potencial de los SLM debe centrarse en aplicaciones para tareas específicas que, aunque sean más limitadas, se vuelven más ágiles, al igual que algunos casos de uso sectorizados se verían beneficiados.

Más allá de un sector concreto, el hecho de elegir entre un LLM y un SLM depende de las necesidades específicas de aplicación de cada empresa o industria. Y la decisión sobre qué modelo seguir debe tener en cuenta factores como el tamaño del modelo, la capacidad deseada, el costo, la interpretabilidad y la generalización.

Las empresas pueden afinar los SLM mediante conjuntos de datos específicos de un dominio para personalizarlos según sus necesidades concretas. Esta adaptabilidad implica que se pueden emplear modelos de lenguaje pequeños para una **gran variedad de aplicaciones del mundo real:**

- **Chatbots:** Gracias a su baja latencia y a sus capacidades de IA conversacional, los SLM pueden impulsar los *chatbots* de atención al cliente y responder con rapidez a las consultas en tiempo real.
- **Resumen de contenidos:** Los SLM pueden utilizarse

para resumir conversaciones en un smartphone y crear elementos de acción como eventos del calendario.

- **IA generativa:** Se pueden implementar modelos compactos para completar y generar texto y código de software, por ejemplo, pueden utilizarse para generar, explicar y traducir código a partir de una instrucción en lenguaje natural.
- **Traducción de idiomas:** Muchos modelos de lenguaje SLM son multilingües y se han entrenado en idiomas distintos del inglés, por lo que pueden traducir entre idiomas de manera rápida. Debido a su capacidad para comprender el contexto, pueden producir traducciones muy precisas que conservan los matices y el significado del texto original.
- **Mantenimiento predictivo:** Los modelos ligeros son lo suficientemente pequeños como para implementarse directamente en dispositivos *edge* locales como sensores o dispositivos de Internet de las cosas (IoT). Esto significa que los fabricantes pueden tratar los SLM como herramientas que recopilan datos de sensores instalados en maquinaria y equipos, y analizan esos datos en tiempo real para predecir las necesidades de mantenimiento.
- **Análisis de sentimiento:** Además de procesar y comprender el lenguaje, los SLM también son hábiles para ordenar y clasificar grandes volúmenes de texto de

manera objetiva. Esto los hace adecuados para analizar textos y calibrar el sentimiento que hay detrás de ellos, lo que ayuda a comprender el *feedback* de los clientes.

- **Asistencia a la navegación de vehículos:** Un modelo tan rápido y compacto como un SLM puede funcionar en los ordenadores de a bordo de un vehículo. Debido a sus capacidades multimodales, los modelos de lenguaje pequeños pueden combinar comandos de voz con clasificación de imágenes, por ejemplo, para identificar obstáculos alrededor de un vehículo. Incluso pueden aprovechar sus capacidades de RAG (Retrieval-Augmented Generation), recuperando detalles de los códigos o las normas de circulación para ayudar a los conductores a tomar decisiones de conducción más seguras e informadas.

# 6. CONCLUSIONES

Los SLM se han consolidado como una alternativa esencial en la **inteligencia artificial actual**, ya que equilibran el poder de la generación y comprensión del lenguaje con la eficiencia y la posibilidad de ejecutarse en entornos con recursos limitados.

El año 2025 consolida un ecosistema de IA diverso. Los SLM ofrecen eficiencia y accesibilidad para tareas específicas y entornos restringidos, mientras los LLM siguen avanzando en capacidades cognitivas y multimodales.

Las tendencias futuras apuntan a una mayor eficiencia en todos los tamaños de modelo, expansión de la multimodalidad, ventanas de contexto aún más largas, consolidación del código abierto, un énfasis creciente en la seguridad y la ética, y una mayor especialización de modelos para nichos industriales. La clave reside en comprender los requisitos específicos de cada caso (recursos, latencia, complejidad) para guiar la selección entre SLM y LLM, considerando tanto benchmarks como rendimiento en el mundo real, y aprovechando la flexibilidad y los desafíos del creciente ecosistema de código abierto y las soluciones empresariales enfocadas como Granite.

## 6.1 DEFINICIÓN

Los Small Language Models (SLM) son modelos de lenguaje entrenados para procesar y generar texto de forma similar a los Large Language Models (LLM), pero con un

número menor de parámetros y, por lo general, un consumo de recursos computacionales más bajo.

El objetivo principal de los SLM es ofrecer capacidades de comprensión y generación de lenguaje natural que sean más ligeras, rápidas y fáciles de desplegar en entornos con recursos limitados (por ejemplo, dispositivos móviles, aplicaciones embebidas o sistemas IoT).

## 6.2 ORIGEN Y EVOLUCIÓN

Los SLM **surgieron como respuesta a la necesidad** de equilibrar el poder de los grandes modelos de lenguaje con la eficiencia, el coste y la escalabilidad en distintos entornos y dispositivos. Permiten aprovechar las ventajas de la inteligencia artificial basada en redes neuronales, sin requerir la enorme infraestructura ni los gastos asociados a los modelos de gran tamaño.

**En el futuro**, conforme las técnicas de compresión y optimización sigan avanzando, es probable que veamos un crecimiento todavía mayor de estos modelos ligeros, capaces de abordar un amplio espectro de aplicaciones en el mundo real sin sacrificar en exceso la calidad de sus resultados.

## 6.3 CARACTERÍSTICAS Y RELEVANCIA

La Inteligencia Artificial Generativa (IAG) ha revolucionado la interacción con la tecnología, permitiendo la creación autónoma y eficiente de contenido.

En este documento, se han explorado los SLM, que **destacan por su relevancia en entornos con recursos limitados y su capacidad para ofrecer soluciones eficientes** y escalables sin sacrificar la calidad. Estos modelos están diseñados para procesar y generar texto de manera eficiente, con un menor número de parámetros en comparación con los LLM.

#### 6.4 VENTAJAS Y LIMITACIONES

Los SLM presentan una serie de **ventajas frente a los LLM**, destacando por su eficiencia y rendimiento optimizado. Al tener un tamaño reducido, los SLM requieren menos recursos computacionales, lo que se traduce en menor consumo de memoria y potencia de procesamiento. Esto permite tiempos de inferencia más rápidos, ideales para aplicaciones en tiempo real como asistentes de voz inte-



ractivos o sistemas de respuesta automática en *chatbots*. Además, los SLM son más accesibles económicamente, facilitando su adopción por parte de PyMEs y startups.

Otra ventaja significativa es la viabilidad de la IA en *Edge Computing*, ya que los SLM pueden operar en dispositivos con recursos limitados, como *smartphones*, sensores IoT y sistemas embebidos. Esto ofrece beneficios como menor latencia, mayor privacidad y seguridad, y reducción del consumo de ancho de banda. Los SLM también destacan por su especialización y adaptabilidad, permitiendo un alto rendimiento en tareas específicas como análisis de sentimiento, traducción automática y atención al cliente.

Sin embargo, los SLM **también presentan limitaciones**. Su capacidad de conocimiento general es menor en comparación con los LLM, lo que puede dificultar la respuesta a preguntas sobre temas diversos o la realización de tareas que requieran una base de conocimiento amplia. Además, los SLM tienen un rendimiento limitado en tareas de alta complejidad y creatividad, como la generación de texto extenso y matizado o el razonamiento complejo. La calidad de los datos utilizados para el ajuste (*fine-tuning*) es crucial para el rendimiento de los SLM, y cualquier sesgo presente en estos datos puede afectar negativamente los resultados.

En resumen, los SLM ofrecen ventajas significativas en términos de eficiencia, accesibilidad y especialización,

pero también enfrentan desafíos relacionados con su capacidad de conocimiento general y rendimiento en tareas complejas.

#### 6.5 APLICACIONES Y CASOS DE USO

Los SLM han demostrado ser versátiles y eficientes para una **gran variedad de aplicaciones en diferentes industrias y sectores**.

- En el ámbito de **telecomunicaciones**, los SLM pueden optimizar procesos como la transcripción y clasificación de llamadas, el análisis de sentimientos y satisfacción, la personalización de servicios y la gestión de consultas y soporte técnico.
- En el sector de **aprendizaje y desarrollo**, los SLM pueden facilitar la automatización de tutorías, la personalización de contenidos y la gestión de consultas.
- En banca y **servicios financieros**, los SLM pueden ayudar en la atención al cliente, la detección de fraude y el análisis de sentimiento.
- En **salud y bienestar**, los SLM pueden contribuir a la recolección de síntomas primarios, la gestión de documentación clínica y el análisis de sentimiento en encuestas de satisfacción del paciente.
- En el caso de **retail y comercio electrónico**, los SLM pueden incrementar la eficiencia en la atención pos-

tventa, el análisis de opiniones y la categorización de productos.

## 6.6 RETOS Y OPORTUNIDADES

Los SLM representan un gran avance en la inteligencia artificial, ofreciendo soluciones más ligeras, rápidas y accesibles, manteniendo un rendimiento competitivo. Su flexibilidad y bajos requisitos de recursos abren nuevas perspectivas para diversas aplicaciones, desde entornos con recursos limitados hasta industrias preocupadas por la sostenibilidad. **A medida que las tecnologías evolucionan**, estos modelos prometen desempeñar un papel central en el futuro de la IA.

En cuanto a oportunidades de expansión y mejora, la **evolución de los SLM en combinación con los LLM** representa una de las principales áreas de expansión. Los desarrollos de IA han dado lugar a enfoques de optimización que maximizan el poder conjunto de los LLM y los SLM.

## 6.7. DEMOCRATIZACIÓN DE LA IA

La capacidad de los SLM para comprender y generar lenguaje natural en entornos de recursos limitados está propiciando la **transformación de diferentes industrias**, suponiendo un **impacto de presente y futuro** en sectores que antes no podían acceder a estas tecnologías.

En concreto, al disminuir las barreras de entrada y los costes asociados, **los SLM democratizan la inteligencia**

**artificial**, abriendo oportunidades a proyectos y organizaciones que antes no podían costear soluciones basadas en LLM. La **combinación de eficiencia y eficacia** hace que los SLM resulten particularmente atractivos en la actualidad, donde muchas organizaciones necesitan soluciones potentes pero asequibles y escalables.

**En el futuro**, se prevé que la investigación en optimización de modelos y las aplicaciones prácticas de los SLM sigan

expandiéndose, **impulsando la innovación tecnológica** y mejorando la relación potencia y tamaño de estos modelos.

En definitiva, los SLM tienen el potencial de **transformar numerosos sectores** gracias a su **eficiencia, accesibilidad y alto retorno de inversión**.



# Bibliografía y referencias

En esta sección se presentan las fuentes y referencias más relevantes utilizadas para la elaboración del presente análisis.

CAPÍTULO	BIBLIOGRAFÍA
<b>1. Introducción</b>	<ul style="list-style-type: none"> <li>DigitalES, Libro Blanco de Inteligencia Artificial Generativa, 2024. <a href="https://www.digitales.es/publicacion/libro-blanco-de-la-ia-generativa/">https://www.digitales.es/publicacion/libro-blanco-de-la-ia-generativa/</a></li> <li>Accenture, Technology Vision 2025, AI: A Declaration of Autonomy <a href="https://www.accenture.com/content/dam/accenture/final/accenture-com/document-3/Accenture-Tech-Vision-2025.pdf#zoom=40">https://www.accenture.com/content/dam/accenture/final/accenture-com/document-3/Accenture-Tech-Vision-2025.pdf#zoom=40</a></li> <li>Accenture, Understanding use of gen AI in an AI system <a href="https://www.accenture.com/us-en/blogs/data-ai/understanding-use-gen-ai">https://www.accenture.com/us-en/blogs/data-ai/understanding-use-gen-ai</a></li> </ul>
<b>2. Origen y evolución</b>	<ul style="list-style-type: none"> <li>IBM, What are small language models? <a href="https://www.ibm.com/think/topics/small-language-models?mhsrc=ibmsearch_a&amp;mhq=small%20language%20models">https://www.ibm.com/think/topics/small-language-models?mhsrc=ibmsearch_a&amp;mhq=small%20language%20models</a></li> <li>Chatbot Arena, LLM Leaderboard <a href="https://lmarena.ai/leaderboard">https://lmarena.ai/leaderboard</a></li> </ul>
<b>3. Ventajas y limitaciones de los SLM</b>	<ul style="list-style-type: none"> <li>Deloitte, Tech Trends 2025 <a href="https://www2.deloitte.com/content/dam/insights/articles/us187540_tech-trends-2025/DI_Tech-trends-2025.pdf">https://www2.deloitte.com/content/dam/insights/articles/us187540_tech-trends-2025/DI_Tech-trends-2025.pdf</a></li> <li>NTT DATA, AI Governance <a href="https://es.nttdata.com/documents/ntt-data-ai-governance-v05.pdf">https://es.nttdata.com/documents/ntt-data-ai-governance-v05.pdf</a></li> <li>NTT DATA, Informe Global GenAI <a href="https://es.nttdata.com/insights/reports/global-genai-report#form">https://es.nttdata.com/insights/reports/global-genai-report#form</a></li> </ul>
<b>4. Casos de uso</b>	<ul style="list-style-type: none"> <li>Accenture, Overlap, Quobis Research, 2025 IEFT, SIP: Session Initiation Protocol (2002) <a href="https://www.ietf.org/rfc/rfc3261.html">https://www.ietf.org/rfc/rfc3261.html</a></li> <li>Elastic, Telecommunications observability with the Elastic Stack: Monitoring voice traffic data <a href="https://www.elastic.co/blog/telecommunications-observability-elastic-stack-monitoring-voice-traffic-data">https://www.elastic.co/blog/telecommunications-observability-elastic-stack-monitoring-voice-traffic-data</a></li> </ul>
<b>5. Retos y oportunidades tecnológicas</b>	<ul style="list-style-type: none"> <li>Nokia Research, 2025</li> </ul>

## Glosario

A continuación, se incluye un listado de los términos y acrónimos utilizados a lo largo del informe, proporcionando claridad y facilitando la comprensión de los términos técnicos y específicos empleados.

- **Número de parámetros:** Cuantos más parámetros, por lo general, mayor capacidad de modelado del lenguaje y de razonamiento en contextos amplios, pero también implica mayor uso de memoria y costes computacionales.
- **Ventana de contexto (tokens):** Indica la longitud máxima de secuencia que el modelo puede procesar de una sola vez. Suele ser menor en los modelos más antiguos o ligeros (512, 1024, 2048 tokens) y mayor en los modelos más recientes y potentes (4k, 8k, 32k, ...).

**CDR:** *Call Detail Records*. Registros detallados de llamadas que contienen información sobre cada llamada realizada, como duración, origen y destino.

**IAG:** *Inteligencia Artificial Generativa*. En el contexto del presente análisis, tecnología que permite la creación autónoma de contenido mediante modelos de lenguaje.

**IoT:** *Internet of Things*. Red de dispositivos interconectados que pueden comunicarse y compartir datos entre sí.

**LLM:** *Large Language Models*. Modelos de lenguaje de gran tamaño entrenados con grandes cantidades de datos para generar y comprender texto.

**MoE:** *Mixture of Experts*. Arquitectura de modelos de IA que utiliza múltiples expertos especializados para mejorar el rendimiento en tareas específicas.

**PLN:** *Procesamiento del Lenguaje Natural*. Rama de la inteligencia artificial que se enfoca en la interacción entre computadoras y el lenguaje humano.

**PTQ:** *Post-Training Quantization*. Técnica de cuantización aplicada después del entrenamiento de un modelo para reducir su tamaño y mejorar su eficiencia.

**QAT:** *Quantization Aware Training*. Técnica de entrenamiento que incorpora la cuantización para mejorar la eficiencia de un modelo sin sacrificar su precisión.

**RTP:** *Real-time Transport Protocol*. Protocolo utilizado para la transmisión de datos en tiempo real, como audio y video.

**SIP:** *Session Initiation Protocol*. Protocolo utilizado para iniciar,

mantener y finalizar sesiones de comunicación multimedia, como llamadas de voz y videoconferencias.

**SLM:** *Small Language Models*. Modelos de lenguaje de menor tamaño diseñados para ser eficientes y escalables en entornos con recursos limitados.

**SNMP:** *Simple Network Monitoring Protocol*. Protocolo utilizado para la gestión y monitorización de dispositivos en una red.

**TLS:** *Transport Layer Security*. Protocolo de seguridad que proporciona comunicaciones seguras a través de una red informática.

**UDP:** *User Datagram Protocol*. Protocolo de comunicación que permite el envío de datagramas sin necesidad de establecer una conexión previa.

**VoIP:** *Voice over Internet Protocol*. Tecnología que permite realizar llamadas de voz a través de una conexión a Internet en lugar de una línea telefónica analógica.

## CRÉDITOS

Este informe ha sido elaborado por el Grupo de Trabajo IA Generativa, liderado por Accenture. Grupo que pertenece a la Línea de Trabajo Inteligencia Artificial de DigitalES.

### Accenture

- Borja Lizarraga, Communications, Media and Technology (CMT) Managing Director
- José Luis Peris, AI Strategy Senior Manager
- Celia de la Fuente, CMT Client Account Lead
- Miguel Ángel Caballo, CMT Senior Manager
- Elena Sánchez, Consulting Analyst

### Deloitte

- Mafalda Vizcaino, Asociada Digital Law
- Rodrigo González, Socio Digital Law

### IBM

- Victoria Gómez, Líder IA, IBM España, Portugal, Grecia e Israel
- David Pérez, Experto en IA, IBM

### Nokia

- Juan Pérez, Nokia

### NTT Data

- Álvaro Pulido, Director NTT Data
- Ángel Fraga, Project Manager NTT Data

### Overlap

- Javier Hernández, Consultor
- Maximiliano Scapachini, Consultor I+D+i

### Quobis

- Iago Fernández, Marketing Manager en Quobis
- Yudani Riobó, Head of Innovation en Quobis

### DigitalES

- Beatriz Arias, Directora de Transformación Digital

digitales\_