Grandes modelos de lenguaje y context engineering

Próximo token, contexto y coste.

Conceptos clave

- Un LLM es un predictor de próximo token. La creatividad.[es] **sampling** sobre una distribución.
- Temperatura T: $T = 0$ greedy (extracción); $T \in [0{,}2,\ 0{,}5]$ asistente Q&A; $T \in [0{,}8,\ 1{,}1]$ creatividad.
- El contexto se divide en **cinco piezas**: sistema, memoria, RAG, histórico, pregunta.
- *Prompt caching* (Anthropic) cobra un 10 % del precio cuando el prefijo se reusa.

Ejemplo a mano

Coste de una sesión RAG empresarial con Claude Sonnet 4.6:

Componente	Tokens	$ /M	$ sesión
Instrucciones de sistema	2 000	3 in	0,006
RAG: 5 chunks × 1 000	5 000	3 in	0,015
Histórico (4 turnos)	3 000	3 in	0,009
Pregunta del usuario	200	3 in	0,001
Respuesta	800	15 out	0,012
Total por sesión	11 000	–	**0,043**

50 000 sesiones/mes $\Rightarrow \sim$ 2 150 $/mes. Con prompt caching la cifra baja a $\sim$ 800 $.

Concepto → aplicación: elección de LLM

Caso	Modelo	Latencia
Q&A para 50K empleados	Sonnet + caching	< 2 s
Borradores marketing	GPT-5 / Mistral Large	< 3 s
OCR documentos masivos	Pipeline VLM local	5–15 s
Asistente de código interno	Claude Code + Sonnet	< 5 s
Resumen masivo (millones docs)	DeepSeek-V3 (MoE)	10–30 s

Idea clave del Día 11. **No es *qué* prompt escribes, sino *qué* contexto orquestas.** Las cinco piezas (sistema, memoria, RAG, histórico, pregunta) son la habilidad técnica más valiosa en 2026. Toda respuesta crítica va con $T \leq 0{,}3$.

Para llevarte: lectura Anthropic, *Effective context engineering for AI agents.*