Key Factors driving new LLM Models

Token Window Size

GPT 3.5 (4,096 tokens) —> GPT 40 (128,000 tokens)

Why is this important?

- Better grounding
- Less hallucinations

Training Parameters

8 billion parameters to 405 billion (Llama)

Why is this important?

- Better instruction following
- Wider vocabulary, hence quality

Model Size

GPT 4o vs 4o-mini | Gemini Ultra vs Nano

Why is this important?

- On-device embedding
- Better inference Speed

Multi-Modality

Chat completions to VQA

Why is this important?

- Towards Artificial General Intelligence
- Multimedia based interaction

Cost

50% to 78% reduction (consumption)

Why is this important?

- Adoption & Trust
- Competition

(general use)

GenAI Consumption

Prompt Engineering

... still plays a BIG role