



Figure 1 | Cost-performance plot. Gemini 2.5 Pro is a marked improvement over Gemini 1.5 Pro, and has an LMArena score that is over 120 points higher than Gemini 1.5 Pro. Cost is a weighted average of input and output tokens pricing per million tokens. Source: [LMArena](#), imported on 2025-06-16.

Gemini 2.5 models build on the success of Gemini 1.5 in processing long-context queries, and incorporate new modeling advances allowing Gemini 2.5 Pro to surpass the performance of Gemini 1.5 Pro in processing long context input sequences of up to 1M tokens (see Table 3). Both Gemini 2.5 Pro and Gemini 2.5 Flash can process pieces of long-form text (such as the entirety of “Moby Dick” or “Don Quixote”), whole codebases, and long form audio and video data (see Appendix 8.5). Together with advancements in long-context abilities, architectural changes to Gemini 2.5 vision processing lead to a considerable improvement in image and video understanding capabilities, including being able to process 3-hour-long videos and the ability to convert demonstrative videos into interactive coding applications (see our recent blog post by [Baddepudi et al., 2025](#)).

The smaller models in the Gemini 2.5 series — Flash size and below — use distillation ([Anil et al., 2018](#); [Hinton et al., 2015](#)), as was done in the Gemini 1.5 series ([Gemini Team, 2024](#)). To reduce the cost associated with storing the teacher’s next token prediction distribution, we approximate it using a k-sparse distribution over the vocabulary. While this still increases training data throughput and storage demands by a factor of k, we find this to be a worthwhile trade-off given the significant quality improvement distillation has on our smaller models, leading to high-quality models with a reduced serving cost (see Figure 2).

2.2. Dataset

Our pre-training dataset is a large-scale, diverse collection of data encompassing a wide range of domains and modalities, which includes publicly available web documents, code (various programming languages), images, audio (including speech and other audio types) and video, with a cutoff date of June 2024 for 2.0 and January 2025 for 2.5. Compared to the Gemini 1.5 pre-training dataset