Gemini 2.5 è il modello di intelligenza artificiale Google più potente di sempre?

Maria Cattini | 01/04/2025 | Intelligenza Artificiale

□□ Più di un modello, un nuovo modo di ragionare

Google alza ancora una volta l'asticella dell'Intelligenza Artificiale con il lancio di **Gemini 2.5**, il modello più avanzato della sua famiglia AI. A guidare questa nuova generazione è Gemini 2.5 Pro Experimental, una versione che si distingue per prestazioni di alto livello e una caratteristica inedita: è un **modello che "pensa"**.

□□ Cosa significa "pensante"

Con Gemini 2.5 Google introduce un paradigma nuovo: l'IA non si limita a generare risposte, ma **ragiona**, analizza, deduce. Il modello è stato progettato per:

• Comprendere meglio il contesto e le sfumature
• Costruire catene logiche di pensiero prima di rispondere
• Affrontare problemi complessi con inferenze e analisi profonde

Questa capacità di "pensiero" si traduce in **risposte più accurate, rilevanti e umane**, superando i modelli basati solo su pattern recognition.

□□ Prestazioni da record

Gemini 2.5 Pro si posiziona **al primo posto su LMArena**, superando tutti gli altri modelli con un margine netto. Ma i traguardi non finiscono qui:

• Eccelle nei benchmark GPQA e AIME 2025 per matematica e scienze
• Ottiene punteggi elevati su Humanity's Last Exam, un dataset che misura la frontiera del pensiero umano

Questi risultati lo confermano come uno dei modelli più capaci mai sviluppati da Google.

□□□□ Codifica evoluta: oltre la generazione di codice

Gemini 2.5 fa anche un salto significativo in ambito programmazione:

• Eccelle nella creazione di app web interattive e accattivanti
• Dimostra abilità notevoli nel modificare e trasformare codice esistente
• Ottiene punteggi elevati nel benchmark SWE-Bench Verified, standard per la codifica agentica

Esempio concreto? Può **generare il codice di un videogioco giocabile da un prompt semplice**.

Le intelligenze artificiali di Google

Modello	Anno di rilascio	Caratteristiche principali	Contesto
Gemini 1	2023	Multimodalità base, competenze avanzate in testo e codice	Gemini Pro gratuito; Gemini Ultra riservato
Gemini 1.5	2024	Contesto esteso (1M token), capacità di ragionamento migliorate	Rilasciato per sviluppatori su Vertex AI
Gemini 2.0	inizio 2025	Ottimizzazione su codice, audio, video, testo, immagini	Incluso in Gemini Advanced
Gemini 2.5 Pro Experimental	marzo 2025	AI 'pensante', ragionamento logico, 1M token, codifica avanzata	Google AI Studio, app Gemini, in arrivo su Vertex AI

□□ Le due chiavi del successo: multimodalità e contesto

Gemini 2.5 si basa su due pilastri della nuova generazione di AI di Google:

□□ Multimodalità nativa

Il modello è in grado di comprendere e generare **testo, immagini, audio, video e codice** in modo fluido, integrando fonti diverse per risposte più ricche e intelligenti.

□□ Finestra di contesto estesa

Con una capacità attuale di **1 milione di token** (in crescita verso i 2 milioni), Gemini può elaborare e collegare **grandi volumi di dati complessi** in un’unica interazione.

□□ Un'evoluzione basata su tecniche avanzate

Gemini 2.5 nasce dall'integrazione di:

• Chain-of-thought prompting (ragionamento passo-passo)
• Apprendimento per rinforzo
• Post-training avanzato con focus su ragionamento e coerenza

Questa combinazione rende le sue risposte non solo intelligenti, ma **riflessive e consapevoli**.

□□ Disponibilità e accesso

Attualmente Gemini 2.5 Pro Experimental è disponibile per:

• Gemini Advanced tramite l’app ufficiale
• Google AI Studio per sperimentazioni dirette
• Vertex AI nelle prossime settimane, per aziende e sviluppatori

Google prevede di introdurre anche **piani tariffari per uso su larga scala** nei prossimi mesi.

□□ Obiettivo: rendere l’IA sempre più utile

Lo sviluppo di Gemini 2.5 è orientato a un fine chiaro: creare **IA capaci di comprendere, pensare e agire con consapevolezza contestuale**.

Migliorando a ritmo costante, Gemini rappresenta uno strumento chiave per affrontare sfide sempre più complesse.

Il feedback degli utenti sarà fondamentale per guidare la prossima evoluzione.

Benchmark		Gemini 2.5 Pro Experimental (03-25)	OpenAI o3-mini High	OpenAI GPT-4.5	Claude 3.7 Sonnet 64k Extended Thinking	Grok 3 Beta Extended Thinking	DeepSeek R1
Reasoning & knowledge Humanity's Last Exam (no tools)		**18.8%**	14.0%*	6.4%	8.9%	–	8.6%*
Science GPQA diamond	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	–	–	–	**84.8%**	84.6%	–
Mathematics AIME 2025	single attempt (pass@1)	**86.7%**	86.5%	–	49.5%	77.3%	70.0%
	multiple attempts	–	–	–	–	**93.3%**	–
Mathematics AIME 2024	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	–	–	–	80.0%	**93.3%**	–
Code generation LiveCodeBench v5	single attempt (pass@1)	70.4%	**74.1%**	–	–	70.6%	64.3%
	multiple attempts	–	–	–	–	**79.4%**	–
Code editing Aider Polyglot		**74.0% / 68.6%** whole / diff	60.4% diff	44.9% diff	64.9% diff	–	56.9% diff
Agentic coding SWE-bench verified		63.8%	49.3%	38.0%	**70.3%**	–	49.2%
Factuality SimpleQA		52.9%	13.8%	**62.5%**	–	43.6%	30.1%
Visual reasoning MMMU	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	–	no MM support	–	–	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		69.4%	no MM support	–	–	–	no MM support
Long context MRCR	128k	**91.5%**	36.3%	48.8%	–	–	–
	1M	83.1%	–	–			
Multilingual performance Global MMLU (Lite)		89.8%	–	–	–	–	–

Methodology

Gemini results: All Gemini 2.5 Pro scores are pass@1 (no majority voting or parallel test time compute unless indicated otherwise). They are all run with the AI Studio API for the model-id gemini-2.5-pro-exp-03-25 with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Vibe-Eval results are reported using Gemini as a judge.

Non-Gemini results: All the results for non-Gemini models are sourced from providers' self reported numbers. All SWE-bench Verified numbers follow official provider reports, using different scaffolding and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using model's own judgement.

Thinking vs non-thinking: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU come with 64k extended thinking. Aider with 32k, and HLE with 16k. Remaining results come from the non thinking model due to result availability. For Grok-3 all results come with extended reasoning except for SimpleQA (based on xAI reports).

Single attempt vs multiple attempts: When two numbers are reported for the same eval higher number uses majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks. Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam. AIME 2025 numbers are sourced from https://matharena.ai/. LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (10/1/2024 - 2/1/2025 in the UI). Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/

* Indicates evaluated on text problems only (without images)

□□ Più di un modello, un nuovo modo di ragionare

Google alza ancora una volta l'asticella dell'Intelligenza Artificiale con il lancio di **Gemini 2.5**, il modello più avanzato della sua famiglia AI. A guidare questa nuova generazione è Gemini 2.5 Pro Experimental, una versione che si distingue per prestazioni di alto livello e una caratteristica inedita: è un **modello che "pensa"**.

□□ Cosa significa "pensante"

Con Gemini 2.5 Google introduce un paradigma nuovo: l'IA non si limita a generare risposte, ma **ragiona**, analizza, deduce. Il modello è stato progettato per:

• Comprendere meglio il contesto e le sfumature
• Costruire catene logiche di pensiero prima di rispondere
• Affrontare problemi complessi con inferenze e analisi profonde

Questa capacità di "pensiero" si traduce in **risposte più accurate, rilevanti e umane**, superando i modelli basati solo su pattern recognition.

□□ Prestazioni da record

Gemini 2.5 Pro si posiziona **al primo posto su LMArena**, superando tutti gli altri modelli con un margine netto. Ma i traguardi non finiscono qui:

• Eccelle nei benchmark GPQA e AIME 2025 per matematica e scienze
• Ottiene punteggi elevati su Humanity's Last Exam, un dataset che misura la frontiera del pensiero umano

Questi risultati lo confermano come uno dei modelli più capaci mai sviluppati da Google.

□□□□ Codifica evoluta: oltre la generazione di codice

Gemini 2.5 fa anche un salto significativo in ambito programmazione:

• Eccelle nella creazione di app web interattive e accattivanti
• Dimostra abilità notevoli nel modificare e trasformare codice esistente
• Ottiene punteggi elevati nel benchmark SWE-Bench Verified, standard per la codifica agentica

Esempio concreto? Può **generare il codice di un videogioco giocabile da un prompt semplice**.

Le intelligenze artificiali di Google

Modello	Anno di rilascio	Caratteristiche principali	Contesto
Gemini 1	2023	Multimodalità base, competenze avanzate in testo e codice	Gemini Pro gratuito; Gemini Ultra riservato
Gemini 1.5	2024	Contesto esteso (1M token), capacità di ragionamento migliorate	Rilasciato per sviluppatori su Vertex AI
Gemini 2.0	inizio 2025	Ottimizzazione su codice, audio, video, testo, immagini	Incluso in Gemini Advanced
Gemini 2.5 Pro Experimental	marzo 2025	AI 'pensante', ragionamento logico, 1M token, codifica avanzata	Google AI Studio, app Gemini, in arrivo su Vertex AI

□□ Le due chiavi del successo: multimodalità e contesto

Gemini 2.5 si basa su due pilastri della nuova generazione di AI di Google:

□□ Multimodalità nativa

Il modello è in grado di comprendere e generare **testo, immagini, audio, video e codice** in modo fluido, integrando fonti diverse per risposte più ricche e intelligenti.

□□ Finestra di contesto estesa

Con una capacità attuale di **1 milione di token** (in crescita verso i 2 milioni), Gemini può elaborare e collegare **grandi volumi di dati complessi** in un'unica interazione.

□□ Un'evoluzione basata su tecniche avanzate

Gemini 2.5 nasce dall'integrazione di:

• Chain-of-thought prompting (ragionamento passo-passo)
• Apprendimento per rinforzo
• Post-training avanzato con focus su ragionamento e coerenza

Questa combinazione rende le sue risposte non solo intelligenti, ma **riflessive e consapevoli**.

□□ Disponibilità e accesso

Attualmente Gemini 2.5 Pro Experimental è disponibile per:

• Gemini Advanced tramite l'app ufficiale
• Google AI Studio per sperimentazioni dirette
• Vertex AI nelle prossime settimane, per aziende e sviluppatori

Google prevede di introdurre anche **piani tariffari per uso su larga scala** nei prossimi mesi.

□□ Obiettivo: rendere l'IA sempre più utile

Lo sviluppo di Gemini 2.5 è orientato a un fine chiaro: creare **IA capaci di comprendere, pensare e agire con consapevolezza contestuale**.

Migliorando a ritmo costante, Gemini rappresenta uno strumento chiave per affrontare sfide sempre più complesse.

Il feedback degli utenti sarà fondamentale per guidare la prossima evoluzione.

Benchmark		Gemini 2.5 Pro Experimental (03-25)	OpenAI o3-mini High	OpenAI GPT-4.5	Claude 3.7 Sonnet 64k Extended Thinking	Grok 3 Beta Extended Thinking	DeepSeek R1
Reasoning & knowledge Humanity's Last Exam (no tools)		**18.8%**	14.0%*	6.4%	8.9%	–	8.6%*
Science GPQA diamond	single attempt (pass@1)	**84.0%**	79.7%	71.4%	78.2%	80.2%	71.5%
	multiple attempts	–	–	–	**84.8%**	84.6%	–
Mathematics AIME 2025	single attempt (pass@1)	**86.7%**	86.5%	–	49.5%	77.3%	70.0%
	multiple attempts	–	–	–	–	**93.3%**	–
Mathematics AIME 2024	single attempt (pass@1)	**92.0%**	87.3%	36.7%	61.3%	83.9%	79.8%
	multiple attempts	–	–	–	80.0%	**93.3%**	–
Code generation LiveCodeBench v5	single attempt (pass@1)	70.4%	**74.1%**	–	–	70.6%	64.3%
	multiple attempts	–	–	–	–	**79.4%**	–
Code editing Aider Polyglot		**74.0% / 68.6%** whole / diff	60.4% diff	44.9% diff	64.9% diff	–	56.9% diff
Agentic coding SWE-bench verified		63.8%	49.3%	38.0%	**70.3%**	–	49.2%
Factuality SimpleQA		52.9%	13.8%	**62.5%**	–	43.6%	30.1%
Visual reasoning MMMU	single attempt (pass@1)	**81.7%**	no MM support	74.4%	75.0%	76.0%	no MM support
	multiple attempts	–	no MM support	–	–	78.0%	no MM support
Image understanding Vibe-Eval (Reka)		69.4%	no MM support	–	–	–	no MM support
Long context MRCR	128k	**91.5%**	36.3%	48.8%	–	–	–
	1M	83.1%	–	–			
Multilingual performance Global MMLU (Lite)		89.8%	–	–	–	–	–

Methodology

Gemini results: All Gemini 2.5 Pro scores are pass@1 (no majority voting or parallel test time compute unless indicated otherwise). They are all run with the AI Studio API for the model-id gemini-2.5-pro-exp-03-25 with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Vibe-Eval results are reported using Gemini as a judge.

Non-Gemini results: All the results for non-Gemini models are sourced from providers' self reported numbers. All SWE-bench Verified numbers follow official provider reports, using different scaffolding and infrastructure. Google's scaffolding includes drawing multiple trajectories and re-scoring them using model's own judgement.

Thinking vs non-thinking: For Claude 3.7 Sonnet: GPQA, AIME 2024, MMMU come with 64k extended thinking. Aider with 32k, and HLE with 16k. Remaining results come from the non thinking model due to result availability. For Grok-3 all results come with extended reasoning except for SimpleQA (based on xAI report).

Single attempt vs multiple attempts: When two numbers are reported for the same eval higher number uses majority voting with n=64 for Grok models and internal scoring with parallel test time compute for Anthropic models.

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks. Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam. AIME 2025 numbers are sourced from https://matharena.ai/. LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (10/1/2024 - 2/1/2025 in the UI). Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/

* Indicates evaluated on text problems only (without images)