Published: March, 2026

Model Evaluation – Approach, Methodology & Results

Gemini 3.1 Flash-Lite

Approach: Gemini 3.1 Flash-Lite was evaluated across a range of benchmarks, including speed, reasoning, multimodal capabilities, factuality, agentic tool use, multi-lingual performance, coding, and long-context.

Methodology: All Gemini scores are single attempt (pass @1) except where otherwise noted. "Single attempt" settings allow no majority voting or parallel test-time compute. All of the results are all run with the Gemini API for the model-id gemini-3.1-flash-lite-preview with default sampling settings and high thinking unless indicated otherwise. To reduce variance, we average over multiple trials for smaller benchmarks.

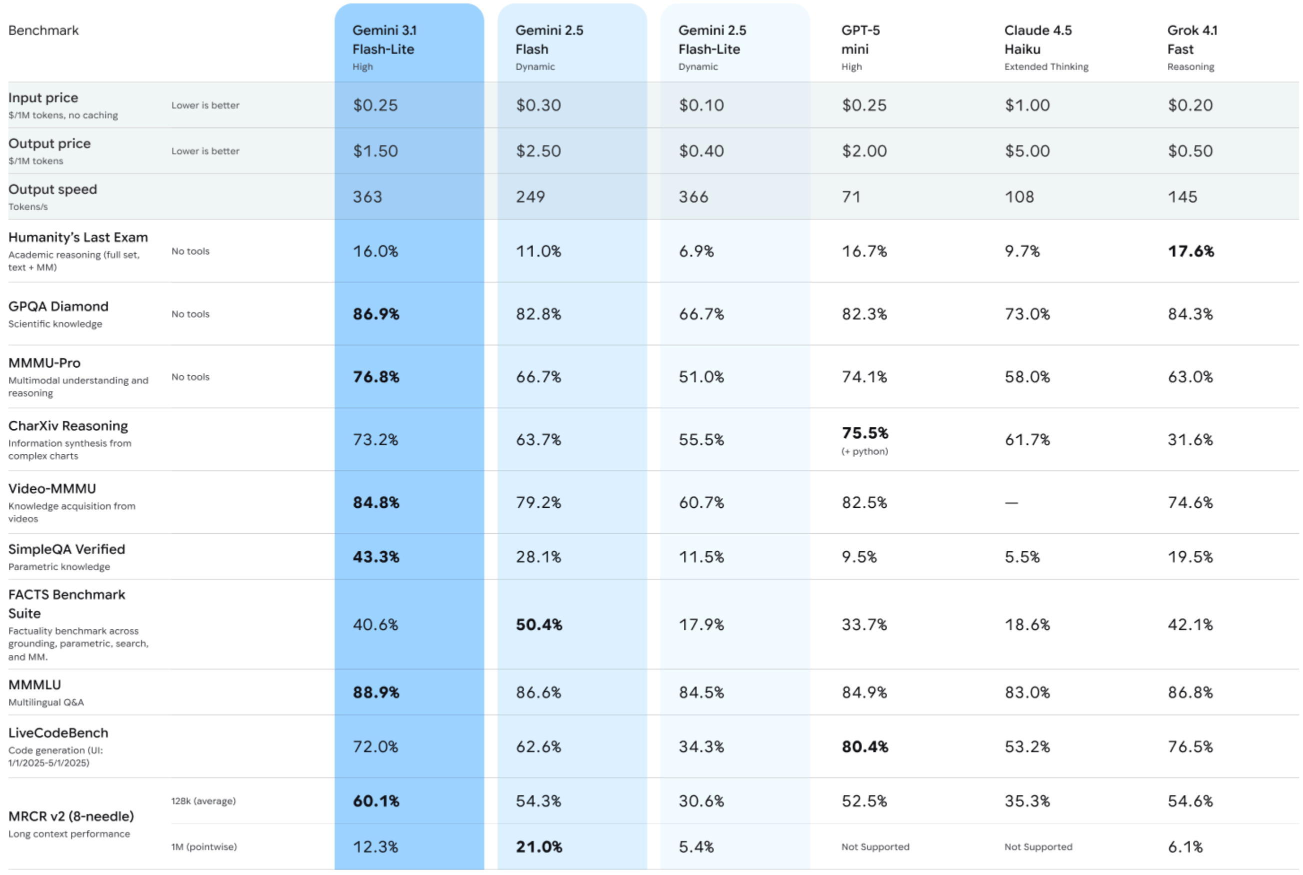
All the results for non-Gemini models are sourced from providers' self reported numbers unless mentioned otherwise below. For GPT-5 mini, Claude 4.5 Haiku, and Grok 4.1 Fast we default to reporting maximum thinking/reasoning settings available, but when reported results are not available we use best available reasoning results.

Additional Details: Our benchmarks span several capabilities as of Mar, 2026:

- **Speed**
  - *Output speed (tokens / s)*: Results for Gemini 3.1 Flash-Lite are sourced from Artificial Analysis. Results for all other models are sourced from the Artificial Analysis Output Speed Variance leaderboard, accurate as at March 3, 2026.
- **Reasoning and Academic Knowledge:**
  - *Humanity's Last Exam* results for Gemini models are self-computed.
  - *GPQA Diamond* results for Gemini models are self-computed. Grok 4.1 Fast and Claude 4.5 Haiku results are taken from the Artificial Analysis leaderboard.
- **Image**
  - *MMMU-Pro* scores are averaged across the Standard (10 options) and Vision settings. Claude 4.5 Haiku results are taken from the Artificial Analysis leaderboard.
  - *CharXiv Reasoning* results for Gemini models as well as Claude 4.5 Haiku and Grok 4.1 Fast are self-computed.
- **Video**
  - *Video-MMMU* results for Gemini models as well as Claude 4.5 Haiku and Grok 4.1 Fast are self-computed.
- **Factuality**
  - *SimpleQA Verified* results for all models are self-computed. When evaluating Claude 4.5 Haiku, we noticed that the model consistently refused to answer, responding “I don't have reliable information about [topic]”.
  - *FACTS Benchmark Suite* results for all models are self-computed.
- **Multilinguality**
  - *Multilingual MMLU* results for Gemini models, GPT-5 mini, and Grok 4.1 Fast are self-computed.

- **Coding**
  - *LiveCodeBench Code generation* results use the 175 UI problems from the date range: 1/1/2025-5/1/2025. Results for all models are self-computed.
- **Long Context**
  - *MRCR v2* results include 128k results as a cumulative score to ensure they can be comparable with other models and a pointwise value for 1M context window to show the capability of the models at full length. The full dataset is available for reproducibility in our repository: https://github.com/google-deepmind/eval_hub/tree/master/eval_hub/mrcr_v2

Results: Gemini 3.1 Flash-Lite results as of March, 2026 are below:



Benchmark		Gemini 3.1 Flash-Lite High	Gemini 2.5 Flash Dynamic	Gemini 2.5 Flash-Lite Dynamic	GPT-5 mini High	Claude 4.5 Haiku Extended Thinking	Grok 4.1 Fast Reasoning
Input price $/1M tokens, no caching	Lower is better	$0.25	$0.30	$0.10	$0.25	$1.00	$0.20
Output price $/1M tokens	Lower is better	$1.50	$2.50	$0.40	$2.00	$5.00	$0.50
Output speed Tokens/s		363	249	366	71	108	145
Humanity's Last Exam Academic reasoning (full set, text + MM)	No tools	16.0%	11.0%	6.9%	16.7%	9.7%	**17.6%**
GPQA Diamond Scientific knowledge	No tools	**86.9%**	82.8%	66.7%	82.3%	73.0%	84.3%
MMMU-Pro Multimodal understanding and reasoning	No tools	**76.8%**	66.7%	51.0%	74.1%	58.0%	63.0%
CharXiv Reasoning Information synthesis from complex charts		73.2%	63.7%	55.5%	**75.5%** (+ python)	61.7%	31.6%
Video-MMMU Knowledge acquisition from videos		**84.8%**	79.2%	60.7%	82.5%	—	74.6%
SimpleQA Verified Parametric knowledge		**43.3%**	28.1%	11.5%	9.5%	5.5%	19.5%
FACTS Benchmark Suite Factuality benchmark across grounding, parametric, search, and MM.		40.6%	**50.4%**	17.9%	33.7%	18.6%	42.1%
MMMLU Multilingual Q&A		**88.9%**	86.6%	84.5%	84.9%	83.0%	86.8%
LiveCodeBench Code generation (UI: 1/1/2025-5/1/2025)		72.0%	62.6%	34.3%	**80.4%**	53.2%	76.5%
MRCR v2 (8-needle) Long context performance	128k (average)	**60.1%**	54.3%	30.6%	52.5%	35.3%	54.6%
	1M (pointwise)	12.3%	**21.0%**	5.4%	Not Supported	Not Supported	6.1%

Methodology: deepmind.google/models/evals-methodology/gemini-3-1-flash-lite