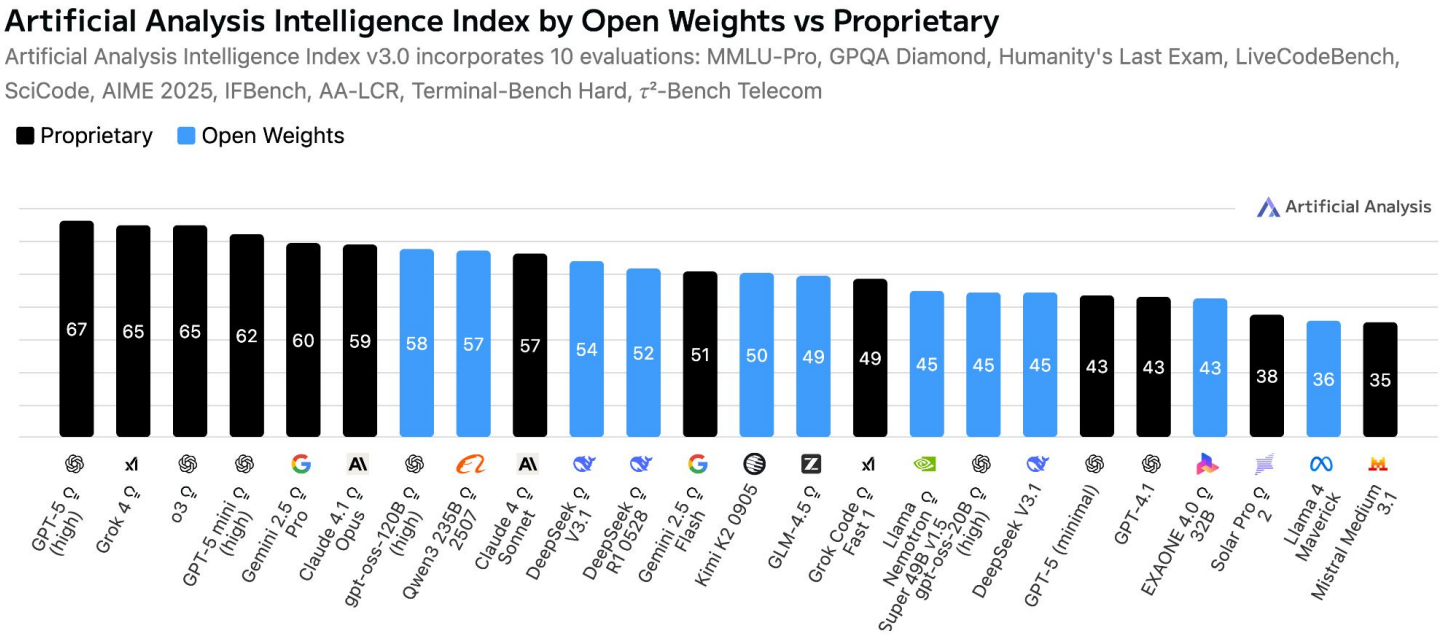
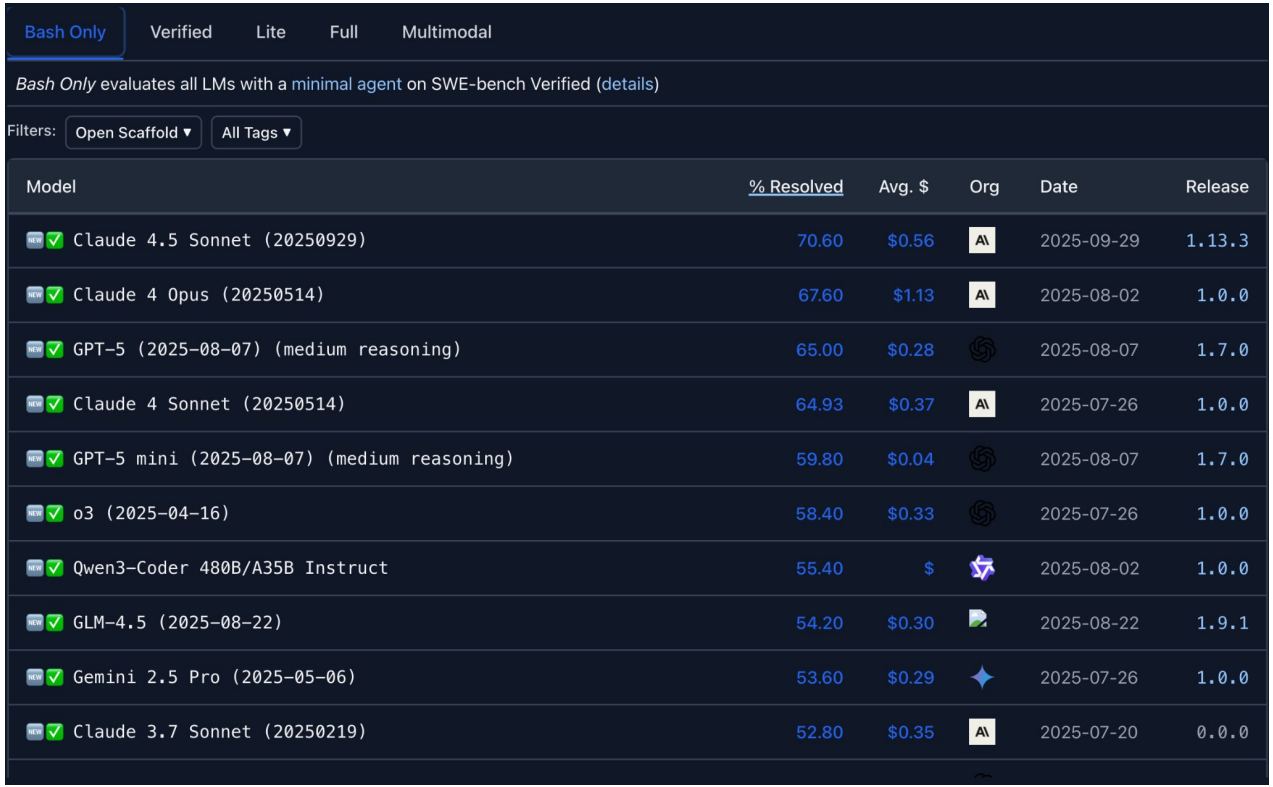
2. Preliminaries

The Landscape of LLMs: open vs closed



https://www.stateof.ai/

The Landscape of LLMs: Open vs Closed



Bash Only | Verified | Lite | Full | Multimodal

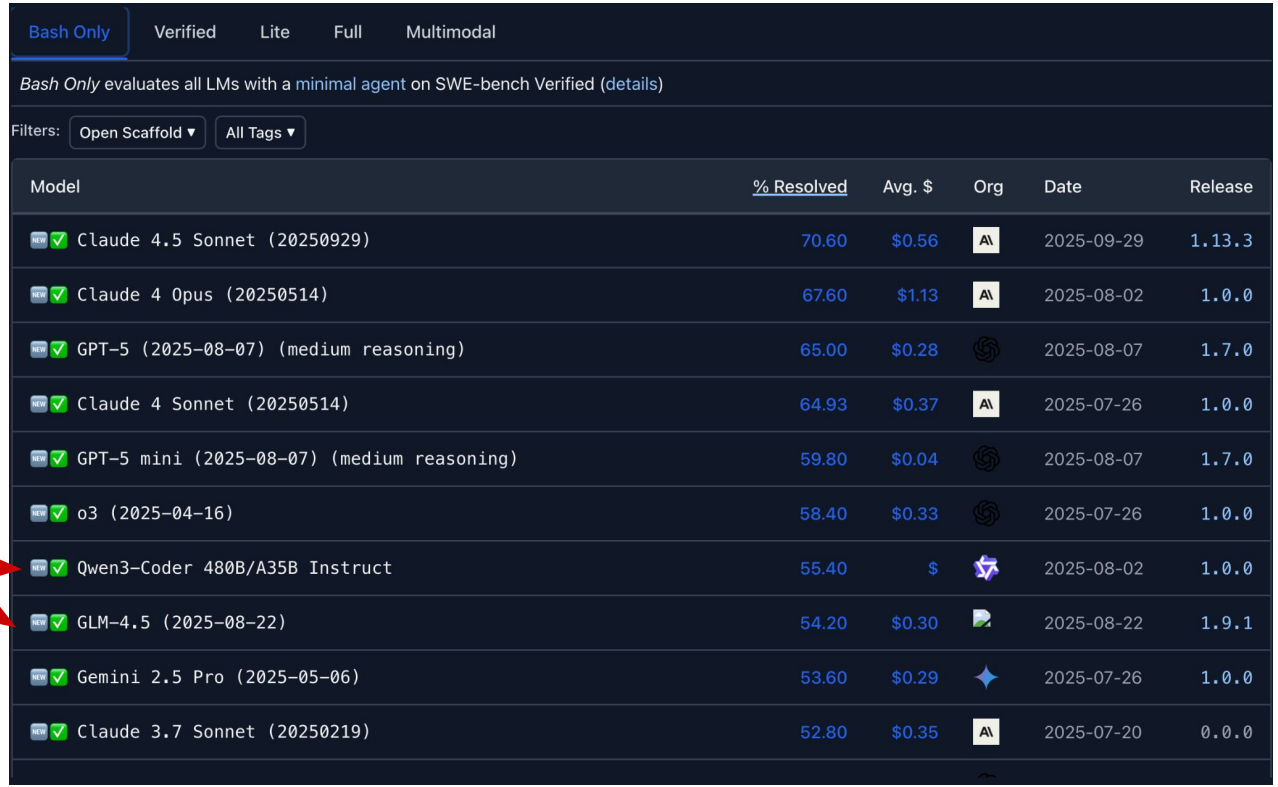
Bash Only evaluates all LMs with a minimal agent on SWE-bench Verified (details)

Filters: Open Scaffold ▾ All Tags ▾

Model	% Resolved	Avg. $	Org	Date	Release
Claude 4.5 Sonnet (20250929)	70.60	$0.56	A\	2025-09-29	1.13.3
Claude 4 Opus (20250514)	67.60	$1.13	A\	2025-08-02	1.0.0
GPT-5 (2025-08-07) (medium reasoning)	65.00	$0.28		2025-08-07	1.7.0
Claude 4 Sonnet (20250514)	64.93	$0.37	A\	2025-07-26	1.0.0
GPT-5 mini (2025-08-07) (medium reasoning)	59.80	$0.04		2025-08-07	1.7.0
o3 (2025-04-16)	58.40	$0.33		2025-07-26	1.0.0
Qwen3-Coder 480B/A35B Instruct	55.40	$		2025-08-02	1.0.0
GLM-4.5 (2025-08-22)	54.20	$0.30		2025-08-22	1.9.1
Gemini 2.5 Pro (2025-05-06)	53.60	$0.29		2025-07-26	1.0.0
Claude 3.7 Sonnet (20250219)	52.80	$0.35	A\	2025-07-20	0.0.0

https://www.swebench.com/

The Landscape of LLMs: Open vs Closed



Bash Only | Verified | Lite | Full | Multimodal

Bash Only evaluates all LMs with a minimal agent on SWE-bench Verified (details)

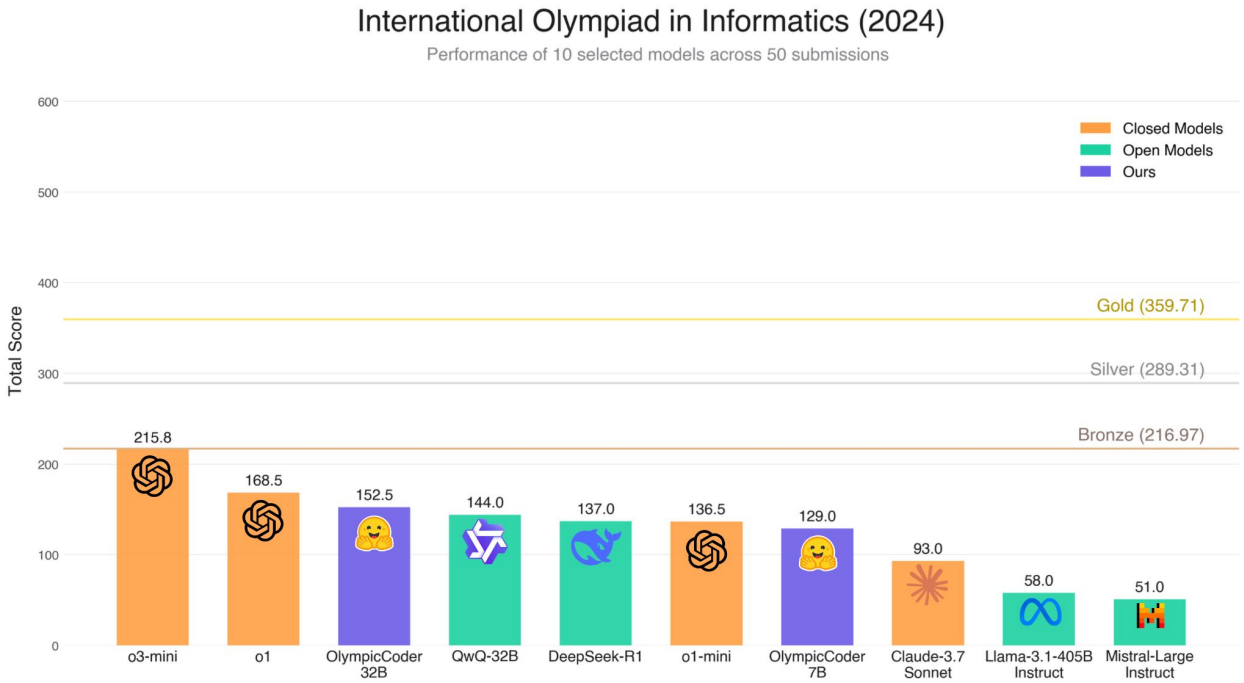
Filters: Open Scaffold ▾ All Tags ▾

Model	% Resolved	Avg. $	Org	Date	Release
Claude 4.5 Sonnet (20250929)	70.60	$0.56	A\	2025-09-29	1.13.3
Claude 4 Opus (20250514)	67.60	$1.13	A\	2025-08-02	1.0.0
GPT-5 (2025-08-07) (medium reasoning)	65.00	$0.28		2025-08-07	1.7.0
Claude 4 Sonnet (20250514)	64.93	$0.37	A\	2025-07-26	1.0.0
GPT-5 mini (2025-08-07) (medium reasoning)	59.80	$0.04		2025-08-07	1.7.0
o3 (2025-04-16)	58.40	$0.33		2025-07-26	1.0.0
Qwen3-Coder 480B/A35B Instruct	55.40	$		2025-08-02	1.0.0
GLM-4.5 (2025-08-22)	54.20	$0.30		2025-08-22	1.9.1
Gemini 2.5 Pro (2025-05-06)	53.60	$0.29		2025-07-26	1.0.0
Claude 3.7 Sonnet (20250219)	52.80	$0.35	A\	2025-07-20	0.0.0

Open models!

https://www.swebench.com/

The Landscape of LLMs: Open vs Closed



https://huggingface.co/blog/open-r1/update-3

What Are Code LMs?

Large Language Models **trained in part on source code;**
Or LLMs **specialized for coding.**

Pre-training Recipes

- Coding is capability **within a general LLM:**
  - Qwen3, Kimi k2, DeepSeek v2

Pre-training Recipes

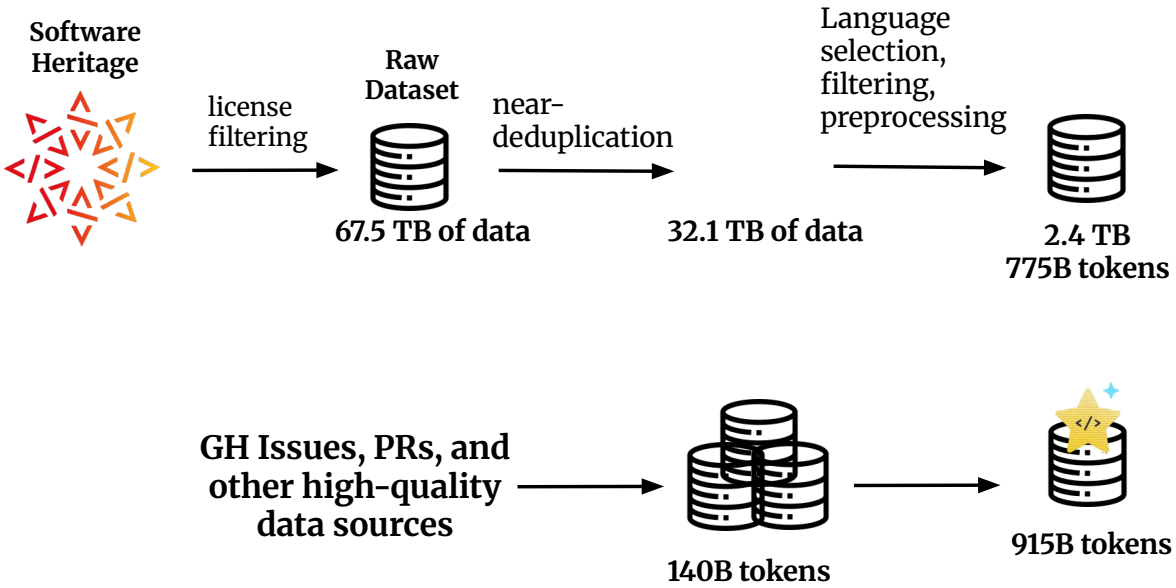
- Coding is capability **within a general LLM:**
  - Qwen3, Kimi k2, DeepSeek v2

- **Continual pre-training** from an LLM:
  - Qwen3-Coder, DeepSeek Coder v2

Pre-training Recipes

- Coding is capability **within a general LLM:**
  - Qwen3, Kimi k2, DeepSeek v2
- **Continual pre-training** from an LLM:
  - Qwen3-Coder, DeepSeek Coder v2
- Code LMs **trained from scratch**:
  - Codex, CodeGen, StarCoder, StarCoder2

Pre-training data: GitHub



The Stack v2 pipeline

Lozhkov, A., Li, R., Allal, L. B., Cassano, F., Lamy-Poirier, J., Tazi, N., ... & de Vries, H. (2024). Starcoder 2 and the stack v2: The next generation. arXiv preprint arXiv:2402.19173.

Pre-training data: CommonCrawl

- DeepSeekCoder v2 **70B tokens**
- Qwen2.5 Coder 580B tokens filtered down to **110B tokens**

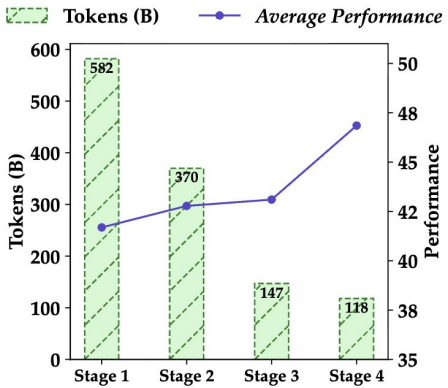


Figure 1: Number of data tokens across different cc-stages, and the validation effectiveness of training Qwen2.5-Coder using corresponding data.

Qwen2.5 Coder web code data filtering

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., ... & Lin, J. (2024). Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186.

Pre-training data: Synthetic Data

- Qwen2.5 Coder
- Arctic-SnowCoder

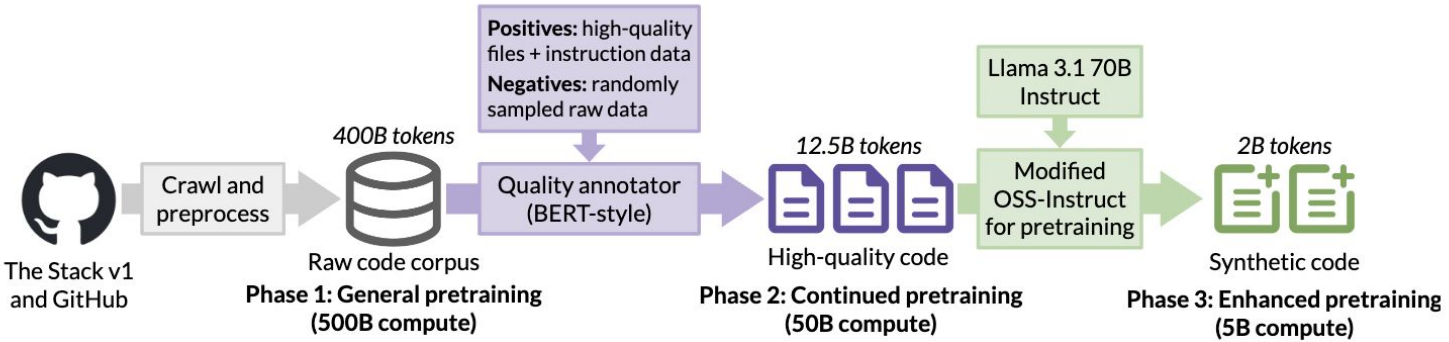


Figure 1: Three-phase pretraining of Arctic-SnowCoder-1.3B with progressively higher-quality data.

Wei, Y., Han, H., & Samdani, R. (2024). Arctic-snowcoder: Demystifying high-quality data in code pretraining. arXiv preprint arXiv:2409.02326.

Pre-training data: Synthetic Data Filtering

- GitHub Code filtered by an LLM: the LLM/classifier rates the educational quality from 0 to 5

```
Below is an extract of SQL code. Evaluate whether it has a high educational value and could help teach
SQL and database concepts. Use the additive 5-point scoring system described below. Points are
accumulated based on the satisfaction of each criterion:

- Add 1 point if the code contains valid SQL syntax, even if it's just basic queries or simple table
operations.
- Add another point if the code addresses practical SQL concepts (e.g., JOINs, subqueries), even if it
lacks comments.
- Award a third point if the code is suitable for educational use and introduces key concepts in SQL
and database management, even if the topic is somewhat advanced (e.g., indexes, transactions). The code
should be well-structured and contain some comments.
- Give a fourth point if the code is self-contained and highly relevant to teaching SQL. It should be
similar to a database course exercise, demonstrating good practices in query writing and database
design.
- Grant a fifth point if the code is outstanding in its educational value and is perfectly suited for
teaching SQL and database concepts. It should be well-written, easy to understand, and contain
explanatory comments that clarify the purpose and impact of each part of the code.

The extract:
--- MySQL mode ---
select name
from employee
order by name asc;

After examining the extract:
- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"
```

Prompt for filtering The Stack v2 SQL data in Stack-Edu

Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., ... & Wolf, T. (2025). SmolLM2: When Smol Goes Big--Data-Centric Training of a Small Language Model. arXiv preprint arXiv:2502.02737.

Pre-training data: Other Sources

- Web text data (FineWeb2, DCLM, Map-CC..)
- Math data (FineMath, MegaMath, InfiMM-WebMath..)

Pre-training data: Other Sources

From Qwen2.5 Coder: “Math and Text data may positively contribute to code performance, but only when their”

Token Ratio			Coding		Math		General			*Average*
Code	Text	Math	Common	BCB	MATH	GSM8K	MMLU	CEval	HellaSwag	
100	0	0	**49.8**	**40.3**	10.3	23.8	42.8	35.9	58.3	31.3
85	15	5	43.3	36.2	26.1	52.5	56.8	57.1	70.0	48.9
70	20	10	48.3	38.3	**33.2**	**64.5**	**62.9**	**64.0**	**73.5**	**55.0**

Table 3: The performance of Qwen2.5-Coder training on different data mixture policy.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., ... & Lin, J. (2024). Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186.