🗄 **Datasets:** ⬡ open-thoughts/**OpenThoughts-114k** ⧉　　♡ like　440

Follow　⬡ Open Thoughts　174

Modalities: 🔤 Text　Formats: 🗎 parquet　Size: 100K - 1M　Tags: curator　▨ Synthetic

Libraries: 🤗 Datasets　📕 Dask　🥐 Croissant　+ 1

📦 **Dataset card**　　⊞ Viewer　　▤ Files　　🤝 Community 11

⊞ Dataset Viewer

⟳ Auto-converted to Parquet　</> API　⊡ Embed

⊞ Full Screen Viewer

Subset (2)	Split (1)
default · 114k rows ⌄	train · 114k rows ⌄

🔍 Search this dataset　　SQL Console

system string · classes	conversations list · lengths
▬▬▬▬▬▬▬ 1 value	▬▬▬▬ 2 　　 2
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
Your role as an assistant involves…	[{ "from": "user", "value": "Generate an executable Python function generated from
"Your role as an assistant involves…	[{"from":"user","value":"Generate an execut Python function generated from the given p:
"Your role as an assistant involves…	[{"from":"user","value":"Generate an execut Python function generated from the given p:
"Your role as an assistant involves…	[{"from":"user","value":"Generate an execut Python function generated from the given p:
"Your role as an assistant involves…	[{"from":"user","value":"Generate an execut Python function generated from the given p:

Downloads last month　　　51,427

Use this dataset ⌄

✎ Edit dataset card

⋮

Homepage:
open-thoughts.ai

Repository:
github.com

Point of Contact:
Open Thoughts Team

Size of downloaded dataset files:
3.55 GB

Size of the auto-converted Parquet files:
3.55 GB

Number of rows:
227,914

📦 **Models trained or fine-tuned on**

⬡ open-thoughts/OpenThin…
✎ Text Generation · · ⭳ 3.35k · ♡

⚫ Quazim0t0/Phi4.Turn.R1…
Updated 1 day a… · ⭳ 2.46k · ♡ 5

☰ Open Thoughts

Open-Thoughts-114k

Open synthetic reasoning dataset with 114k high-quality examples covering math, science, code, and puzzles!

Available Subsets

default subset containing ready-to-train data used to finetune the OpenThinker-7B and OpenThinker-32B models:

```
ds = load_dataset("open-thoughts/OpenThoughts-114k",
```

metadata subset containing extra columns used in dataset construction:

- `problem`
- `ground_truth_solution`
- `deepseek_reasoning`
- `deepseek_solution`
- `domain`
- `source`
- `test_cases` (code only)
- `starter_code` (code only)

```
ds = load_dataset("open-thoughts/OpenThoughts-114k",
```

OpenThinker Models

The numbers reported in the tables below are evaluated with our open-source tool [Evalchemy](#).

	AIME24	MATH500	GPQA-Diamond	LCBv2 Easy	LCBv2 Medium	LCBv2 H...
OpenThinker-32B	66	90.6	61.6	95.1	70.9	26
OpenThinker-7B	31.3	83.0	42.4	75.3	28.6	6.
Bespoke-Stratos-7B	22.7	79.6	38.9	71.4	25.2	0.
DeepSeek-R1-Distill-Qwen-7B	60	88.2	46.9	79.7	45.1	14
gpt-4o-0513	8.7	75.8	46.5	87.4	42.7	8.
o1-mini	64	85.6	60	92.8	74.7	39

We are fully open-source. Our [model weights](#), [datasets](#), [data generation code](#), [evaluation code](#), and [training code](#) are all publicly available.

	Open Weights	Open Data	Open Code
OpenThinker-32B	✅	✅	✅
OpenThinker-7B	✅	✅	✅
Bespoke-Stratos-7B	✅	✅	✅
DeepSeek-R1-Distill models	✅	❌	❌

	Open Weights	Open Data	Open Code
OpenAI/Gemini	❌	❌	❌

We are actively working towards improving the dataset, so please stay tuned!

Data Curation Recipe

Code

- [BAAI/TACO](#)
- [codeparrot/apps](#)
- [deepmind/code_contests](#)
- [MatrixStudio/Codeforces-Python-Submissions](#)
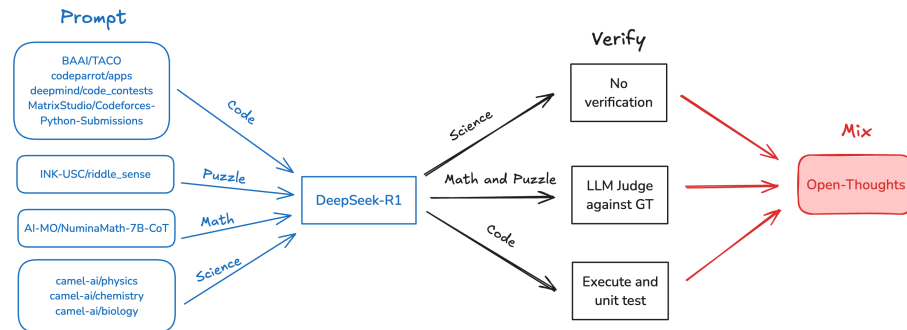
Math

- [AI-MO/NuminaMath-CoT](#)

Science

- [camel-ai/chemistry](#)
- [camel-ai/biology](#)
- [camel-ai/physics](#)

Puzzle

- [INK-USC/riddle_sense](#)

Using a curated mix of the datasets above, we generate reasoning traces from DeepSeek-R1 and verify correctness to construct the final dataset.

The full code for the data generation pipeline is publicly available
in our github repo.

Citation

```
@misc{openthoughts,
  author = {Team, OpenThoughts},
  month = jan,
  title = {{Open Thoughts}},
  howpublished = {https://open-thoughts.ai},
  year = {2025}
}
```

Links

- 📊 OpenThinker-32B Blog Post

- 📊 Measuing Reasoning with Evalchemy Blog Post

- 📊 Open Thoughts Launch Blog Post

- 💻 Open Thoughts GitHub Repository

- 🧠 OpenThoughts-114k dataset - this dataset.

- 🤖 OpenThinker-32B model

- 🤖 OpenThinker-7B model

- 📊 Bespoke-Stratos Blog Post

- 🧠 Bespoke-Stratos-17k dataset

- 🤖 Bespoke-Stratos-32B model

- Stratos-7B model

System theme   TOS   Privacy   About   Jobs   🤗   Models   Datasets   Spaces   Pricing   Docs

Visualization

All 114k examples, clustered by semantic similarity, can be explored in [Nomic Atlas](#).