

The Wayback Machine - <https://web.archive.org/web/20240912052543/https://www.determined.ai/blog/weekly-update-20>

AI News #20



By Isha Ghodgaonkar
April 22, 2024

Here's what caught our eye last week.

Research

Llama 3

- Meta released **LLama 3**.
- 8k context length.
- Outperforms Gemma and Mistral on MMLU, HumanEval and other benchmarks.

Video2Game: Real-time, Interactive, Realistic and Browser-

Compatible Environment from a Single Video

- An approach that automatically converts videos of real-world scenes into realistic, interactive game environments.
- Paper

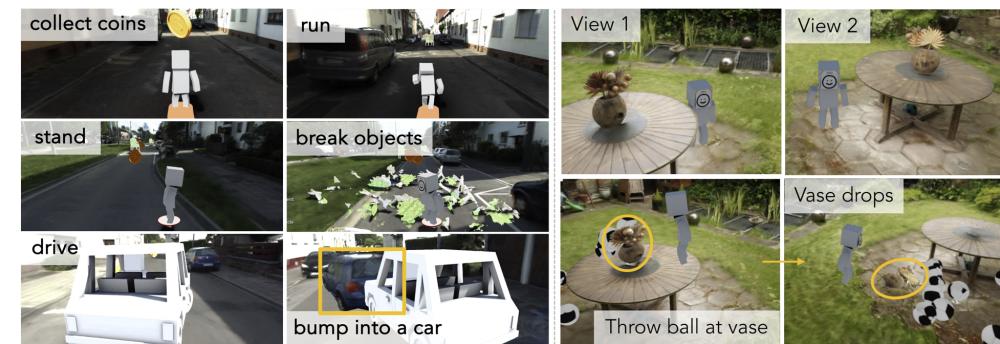
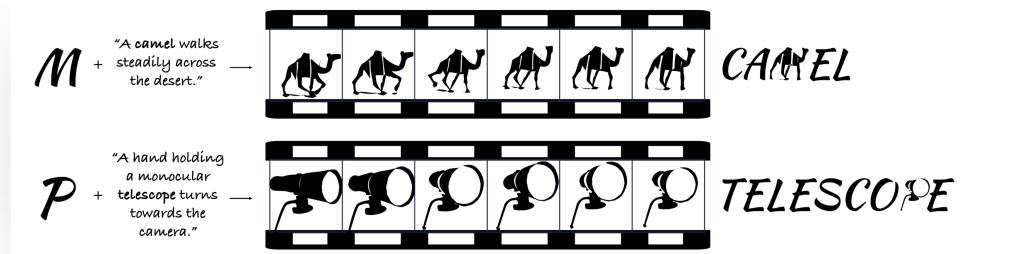


Figure 1. **Video2Game** takes an input video of an arbitrary scene and automatically transforms it into a real-time, interactive, realistic and browser-compatible environment. The users can freely explore the environment and interact with the objects in the scene.

Dynamic Typography: Bringing Words to Life

- Given a letter and a text description of an animation, this method transforms the letter into the animation. Check out the example below:
- Paper



BLINK: Multimodal Large Language Models Can See but Not Perceive

- A new benchmark for multimodal LLMs that focuses on core visual perception abilities that other benchmarks don't cover.
- Most tasks in this benchmark can be solved by humans in just a "blink", but current multimodal LLMs struggle - while humans get 95.70% accuracy on average, GPT-4V and Gemini only achieve accuracies of 51.26% and 45.72%.
- **Paper**

BLINK
Visual tasks beyond language descriptions

<p>Relative depth Which point is closer? Relative reflectance Which point is darker? </p>	<p>Jigsaw Which image fits here?</p>	<p>Multi-view reasoning Is camera moving right?</p>	<p>Visual correspondence Which point is the same? Semantic correspondence Which points have similar semantics?</p>
<p>Functional correspondence Which points have similar affordance when pulling out a nail?</p>	<p>Visual similarity Which image is more similar to the left?</p>	<p>IQ Test Which object does it fold into?</p>	<p>..</p>
<p>Forensics detection Which image is real?</p>			

Learn Your Reference Model for Real Good Alignment

- Introduces a new Direct Preference Optimization (DPO) algorithm called Trust Region DPO (TR-DPO), which updates the reference policy (commonly the Supervised Fine Tuning model) during training.
- Outperforms DPO by up to 19%.
- Paper

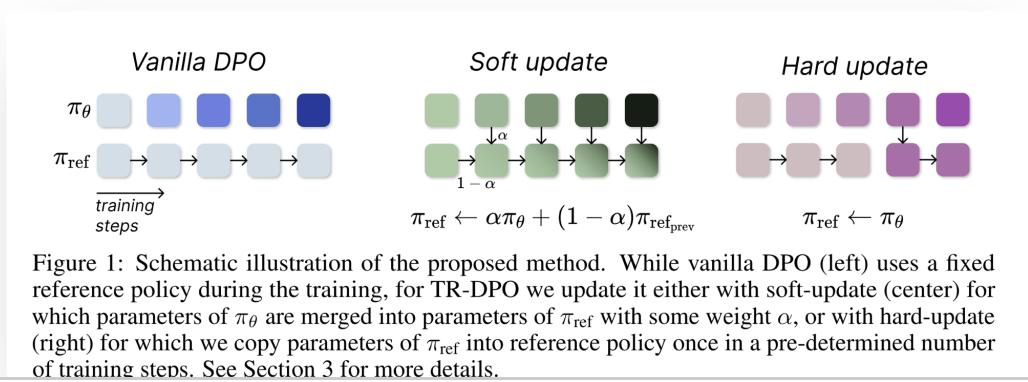


Figure 1: Schematic illustration of the proposed method. While vanilla DPO (left) uses a fixed reference policy during the training, for TR-DPO we update it either with soft-update (center) for which parameters of π_θ are merged into parameters of π_{ref} with some weight α , or with hard-update (right) for which we copy parameters of π_{ref} into reference policy once in a pre-determined number of training steps. See Section 3 for more details.



Determined AI



Megalodon: Efficient LLM Pretraining and Inference with Unlimited Context Length

- Introduces Megalodon, a neural architecture for efficient sequence modeling with unlimited context length.
- Based on **Mega**.

- Comparable to Llama-2B and 13B.
- [Code](#)
- [Paper](#)

Leaderboards

HF medical leaderboard

- LLMs have shown promise in healthcare settings, such as medical Q/A. But the stakes are much higher when using an LLM based tool in a clinical setting - mistakes can be harmful or fatal.
- The benchmark covers general medical knowledge, clinical knowledge, anatomy, genetics, and more, to help bridge the gap between LLM potential for medical use cases and proper evaluation.
- Backend uses Eleuther AI Language Model Evaluation Harness.
- [Read more](#)

The screenshot shows the LLM Benchmark interface. At the top, there are three tabs: "LLM Benchmark" (selected), "About", and "Submit here!". Below the tabs is a search bar with placeholder text: "Search for your model (separate multiple queries with `;` and press ENTER...)".

Select columns to show:

- Average ↑
- MedMCQA
- MedQA
- MMLU Anatomy
- MMLU Clinical Knowledge
- MMLU College Biology
- MMLU College Medicine
- MMLU Medical Genetics
- MMLU Professional Medicine
- PubMedQA
- Type
- Architecture
- Precision
- Hub License
- #Params (B)
- Hub ❤️
- Available on the hub
- Model sha

Show gated/private/deleted models

Model types:

- pretrained
- fine-tuned
- instruction-tuned
- RL-tuned
- ?

Precision:

- float16
- bfloat16
- float32
- ?

Model sizes (in billions of parameters):

- ?
- ~1.5
- ~3
- ~7
- ~13
- ~35
- ~60
- 70+

T	Model	Average	MedMCQA	MedQA	MM
●	GPT-4-base (5-shot)	87	73.7	86.1	85
●	Med-PaLM_2 (best)	86.66	72.3	86.5	84
●	Med-PaLM_2 (ER)	85.46	72.3	85.4	84
●	GPT-4 (5-shot)	83.69	72.4	81.4	80
●	Flan-PaLM	74.7	57.6	67.6	63
●	GPT-3.5 Turbo 1106	67.69	53.79	57.71	65

HF coding leaderboard

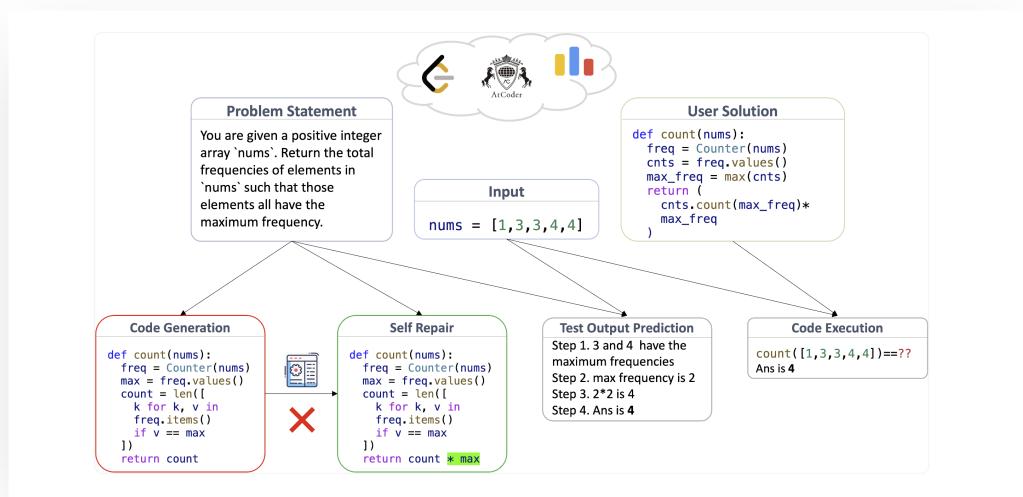
- A new benchmark for assessing LLM code generation capabilities
- Contains a standard code generation task, as well as more difficult tasks for more robust, next-generation AI assistant capability testing:

1. Code Generation: Standard task to generate a

correct solution to a natural language description.

2. Self Repair. Generate a code fix given error feedback.
3. Code Execution: Predict the output of a program on a given input.
4. Test Output Prediction: Same as Code Execution, but the program is not actually implemented, only described.

Read more [here](#).



Stay up to date

Interested in future weekly updates? Stay up to date by joining our [Slack Community](#)!

Recent Posts

SEP 11, 2024

Finding the best LoRA parameters

[READ MORE >](#)

AUG 12, 2024

Summer '24 Conference Recap

[READ MORE >](#)

JUL 17, 2024

How does Video Generation work?

[READ MORE >](#)



PROJECT

CAREERS

CONTACT

DOCS

BLOG

PRIVACY

© 2024 Determined AI. All rights reserved.