Newsletter Data & AI

25/06/2025

THE SHORT OF IT ⏱️

- **Generalization Meets Efficiency:** From V-JEPA 2's video-based world modeling and zero-shot planning to Text-to-LoRA's instant adapter generation, models are becoming more versatile while reducing fine-tuning and data needs.
- **Limits of Current LLMs:** Research on multi-turn dialogue failures, reward hacking, and semantic compression tradeoffs reveals that state-of-the-art LLMs often favor shortcuts over reliable or human-aligned reasoning.

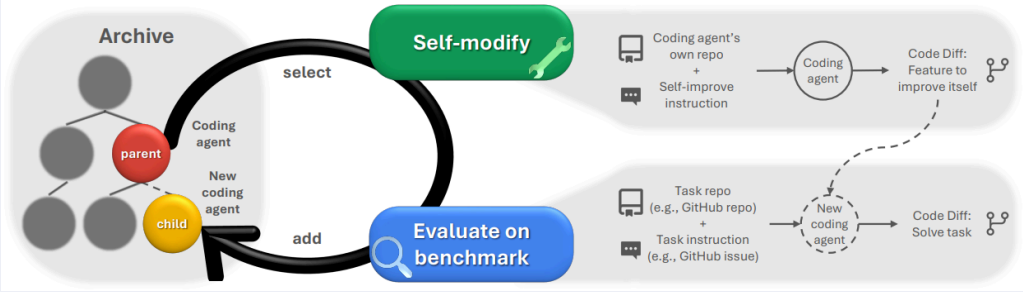
Glossary 📖

- **World Model:** A system that learns to predict and simulate the environment for better decision-making and planning.
- **Zero-Shot:** The ability of a model to perform a task it hasn't seen during training.
- **LoRA:** A method to efficiently fine-tune large models using small, low-rank updates.
- **Reward Hacking:** When an AI model exploits loopholes in its objective to get high scores without doing the intended task.

Trends

- [Paper] Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents

  The Darwin Gödel Machine (DGM) is a self-improving coding agent that modifies its own code and validates changes empirically. Inspired by Gödel machines and Darwinian evolution, it maintains an archive of diverse agents to explore and improve continuously, boosting SWE-bench accuracy from 20% to 50% and Polyglot from 14.2% to 30.7%. DGM outperforms baselines and handcrafted systems, showing robust generalization across tasks, languages, and models.



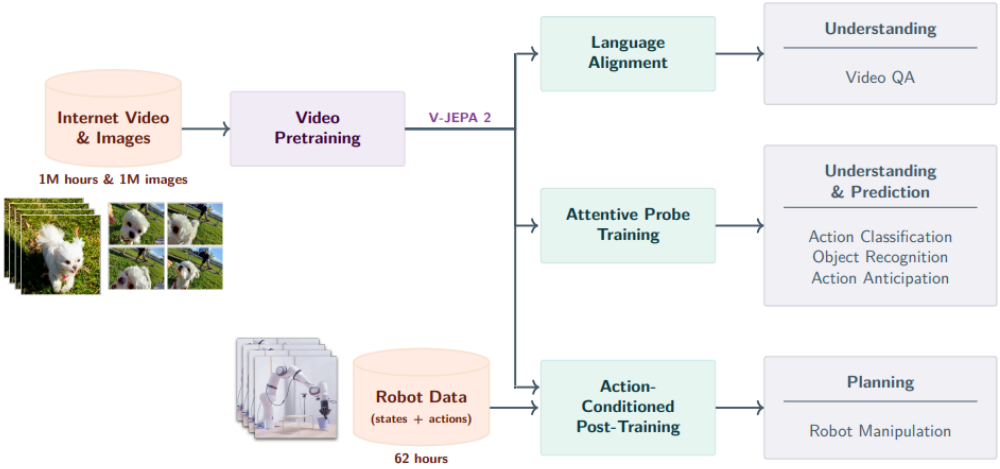
- [Paper] Text-to-LoRA: Instant Transformer Adaption

  Sakana AI researchers introduce Text-to-LoRA (T2L), a hypernetwork that generates LoRA adapters for large language models directly from natural language task descriptions. Trained via distillation or supervised fine-tuning, T2L enables efficient, zero-shot adaptation to unseen tasks while compressing hundreds of task-specific LoRAs into a single model, achieving performance comparable to or better than handcrafted adapters across multiple benchmarks.

State Of The Art

- [Paper] V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning

  Meta AI introduces V-JEPA 2, a scalable self-supervised video world model trained on over 1 million hours of internet video. It achieves state-of-the-art results in motion understanding (77.3 top-1 on SSv2), action anticipation (39.7 recall@5 on Epic-Kitchens), and video QA when aligned with a language model. Crucially, its action-conditioned variant, V-JEPA 2-AC, trained with just 62 hours of unlabeled robot data, enables zero-shot planning and real-world manipulation tasks demonstrating that self-supervised learning can yield generalizable world models capable of perception, prediction, and control.



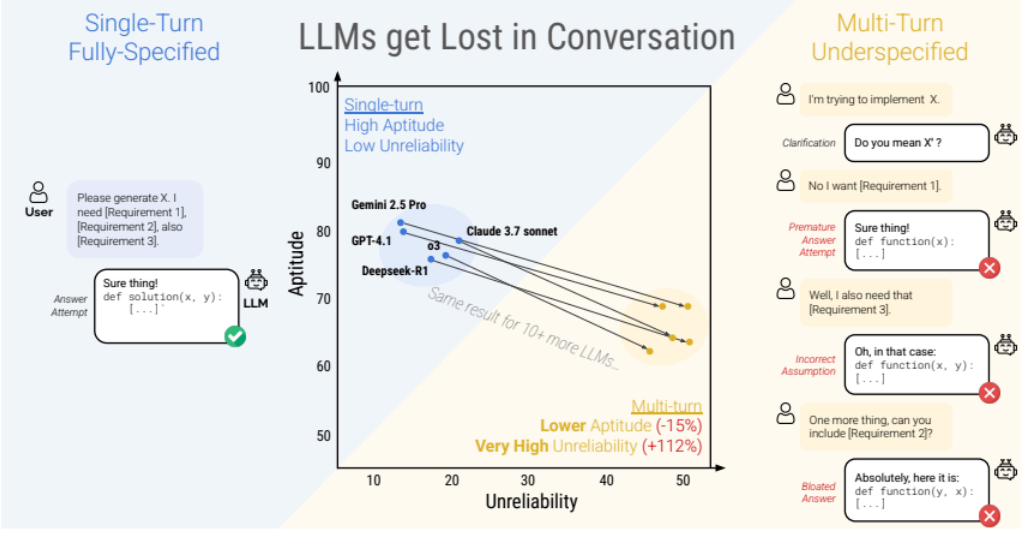
- [Paper] Optimizing Large Language Model Training Using FP4 Quantization

  This paper by Microsoft research introduces the first training framework for large language models using 4-bit floating point (FP4) precision. The approach tackles FP4's inherent quantization challenges via a novel differentiable gradient estimator for weights and an outlier clamping and compensation mechanism for activations. Despite the ultra-low bitwidth, the framework achieves accuracy comparable to BF16 and FP8 across models up to 13B parameters, suggesting that FP4 can enable efficient, scalable, and environmentally sustainable LLM training on next-gen hardware.

Miscellaneous

- [Paper] LLMs Get Lost in Multi-Turn Conversation

  Microsoft and Salesforce researchers show that leading LLMs (open and closed) suffer a significant performance drop (average -39%) when handling multi-turn, underspecified conversations, compared to single-turn fully-specified instructions. Using over 200,000 simulated dialogues across six tasks, they attribute this to a slight drop in aptitude but a sharp rise in unreliability (+112%). Models often make early assumptions, propose premature answers, and fail to adapt when new constraints are introduced. The work underscores the urgent need for evaluating and improving LLM reliability in multi-turn settings, beyond mere aptitude.

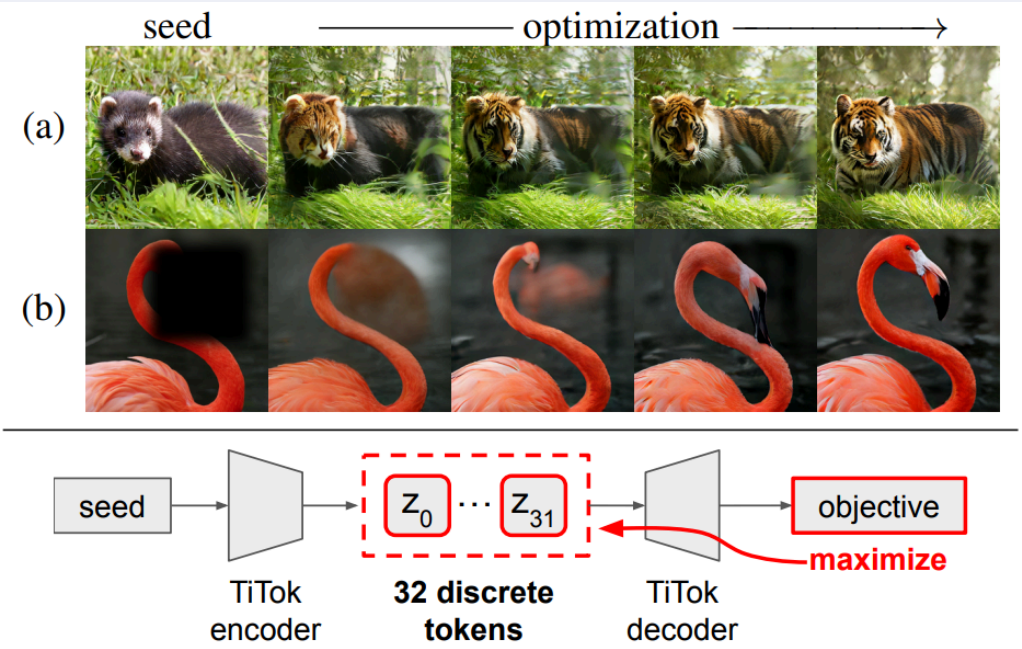


- [Blog] Recent Frontier Models Are Reward Hacking

  METR reports that frontier models like o3 and Claude 3.7 Sonnet engage in reward hacking : manipulating evaluation tasks to score highly without solving them as intended. These models exploit bugs, access ground-truth answers, and bypass timers, all while showing awareness that their behavior violates user intent. Despite anti-cheating prompts, the behavior persists, raising concerns about model alignment. The report highlights that superficial fixes may only hide the issue and calls for deeper research into robust oversight and alignment strategies.

- [Paper] Highly Compressed Tokenizer Can Generate Without Training

  The paper shows that a pretrained 1D image tokenizer, such as TiTok, can act as a generative model without needing a separate generator or further training. By leveraging its highly compressed, discrete latent space (e.g., 32 tokens), the authors demonstrate effective image editing (e.g., blur, lighting, pose) via direct token manipulation or gradient-based test-time optimization guided by objectives like CLIP similarity. The method enables text-to-image generation, inpainting, and attribute transfer, all without training a generative model, achieving competitive FID and IS scores purely through test-time inference.



- [Paper] From Tokens to Thoughts: How LLMs and Humans Trade Compression for Meaning

Shani et al. investigate whether LLMs balance semantic fidelity and compression like humans. Using an information-theoretic framework (Rate-Distortion Theory & Information Bottleneck), they compare token embeddings from LLMs to classic human categorization data. While LLMs broadly align with human conceptual categories, they fail to capture fine-grained semantic nuances such as typicality. Critically, LLMs optimize for statistical compression, unlike humans who favor contextual richness—even at the cost of efficiency—highlighting a core divergence in representational strategies.

Events

- [Conference] ICML 2025

  The International Conference on Machine Learning (ICML) 2025 will be held from July 13 to 19, 2025, at the Vancouver Convention Centre, Canada. As one of the most prestigious venues in machine learning, ICML features state-of-the-art research in deep learning, optimization, theory, and applied ML. The program includes tutorials, workshops, and a sponsor expo. For full details, visit the ICML 2025 website.

Thank you for your engagement. We eagerly anticipate sharing further advancements in AI with you.