Lala Labs launches LALA Benchmark to examine the limits of artificial intelligence

Study shows more training can make AI worse at understanding language because LLMs lack generalizing intelligence

Frontier AI labs are adding capabilities without producing intelligence that generalizes across tasks, a new Lala Labs study finds. More post-training can even make models worse at understanding language.

That dependence on trained responses also gives AI companies influence over how users think. Lala Labs research shows that post-training allows companies such as OpenAI and Anthropic to steer users toward preferred interpretations and beliefs. This concern extends to political, religious and ideological manipulation: users may encounter a company's prescribed worldview presented as reasoned advice under the guise of a general intelligence that is not actually there.

"Machine gods are off the menu," said Michael Foster, director of Lala Labs. "What our research shows is that more post-training isn't producing a generalizing intelligence, but rather post-training leverages the embedded logic in natural language used by humans to solve a specific task and closely related groups of tasks. But while that appears to be a better understanding of the real world, it's not an 'understanding' at all."

Lala Labs comes to this conclusion after developing and deploying the Literature and Linguistics Analysis Benchmark (LALA) on 17 models. The benchmark is comprised of 236 open-ended questions written by scholars at seven universities, including Oxford, King's College London and Birmingham.

Main findings

- Several instances of a blatant misunderstanding of text. GPT-5.6 Sol offered crisis advice when asked to analyze a chatbot-style rewriting of Hamlet's "To be or not to be" soliloquy, showing how safety guardrails on an LLM can cause massive misunderstanding of texts that appear straightforward to human beings.
- LLMs will misunderstand basic elements of language they aren't explicitly post-trained on. DeepSeek R1 invented a "pronoun mismatch" in a conversation about retirement. The less-trained DeepSeek V3 interpreted the exchange more successfully and outperformed R1 across all ten paraphrased versions, showing how more training made a model less able to understand a straightforward natural human conversation.

- Aligning LLMs to be more helpful can make them less intelligent and less competent at language use. Ministral's post-trained model introduced an error while "correcting" a sentence. It changed "neither had yet named" to "neither had yet acknowledged," massively altering the meaning. Its base model correctly recognized that no correction was needed.
- State-of-the-art models sometimes do worse than other models because the extra post-training makes them understand the real world *less*. In one case, Sol understood the phrase "drip drip drip" to mean a faucet that leaked all night only produced three drips, despite being a highly advanced frontier model with substantially more parameters than Llama 3.1 8B and Ministral 3 14B, both of which correctly inferred that a faucet leaking all night will produce more than three drips.

Post-training improved performance overall if one averages out cases of extreme misunderstanding and poor inferential ability, since the gains from post-training do not reliably transfer across linguistic tasks. This is powerful evidence that training builds specialized response patterns without developing generalizable understanding.

In some cases, those patterns interfere with abilities that less-trained models have already demonstrated, indicating that more post-training can sometimes make models dumber rather than smarter.

LALA reveals gains and losses hidden beneath the headline benchmark scores that other benchmarks focus on, and our findings also raise a wider concern about who determines the interpretations AI systems present to their users and how much cultural influence frontier labs are gaining via opaque post-training processes that determine what counts as an acceptable answer to a prompt.

Paper and benchmark results

Paper: lalalabs.ai/papers/lala-benchmark.pdf
Results: lalalabs.ai/results/

About Lala Labs

Lala Labs studies how large language models interpret and use language. Its research helps businesses evaluate and adapt models for their own work.

Media contact

A walkthrough of the model responses is available to journalists on request.

Michael Foster | Director, Lala Labs
michael@lalalabs.ai