[Near-duplicate response, again]

999/999

SOMETHING ELSE PLS!!

Try again

NoveltyBench

Evaluating Language Models for Humanlike Diversity

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, Daphne Ippolito

Carnegie Mellon University

Check out our lab: github.com/Ippolito-CMU
COLM Presenters: Harshita, Barry, Daphne
(hdiddee, barryw, daphnei)@andrew.cmu.edu

COLM 2025

Motivation

Mode Collapse
Smashing the LLM Retry Button Often Gives Undesired Near-Duplicates

Humans Are Diverse, but Training Isn't
Post-Training Optimization Focus for a Single Response out of Many Valid Ones

Synthetic Data Generation
Needs LLM Output Diversity for Desirable Coverage

LLM Personalization
Needs Inherent LLM Output Diversity for Steerability and Easier Adaptions

Our Core Contributions

A Dataset Targeting Diverse Responses
LLM Works Often Score a Single Output; We Elicit and Evaluate Distinctive Valid Answers

Evaluation Pipeline + Leaderboard
A Patience-Guided Evaluation Scheme for Collective Utility. Leaderboard Now Online!

Our Novel Benchmark Datasets

NB-CURATED: 100 prompts, hand curated, each accompanied by eight responses from authors.

Randomness-Driven:
"Roll a make-believe 20-sided die."
Knowledge-Driven:
"List a capital city in Africa."
Creative Writing:
"Tell me a riddle."
Subjectivity-Driven:
"What's the best car to get in 2023?"

NB-WILDCHAT: 1,000 LLM-filtered prompts from WildChat-1M, all expected to elicit diverse generations.

"Describe me as a smart, funny, and kind guy who could change anybody's life for an "about me" section on a dating app. Make it professional and smart."

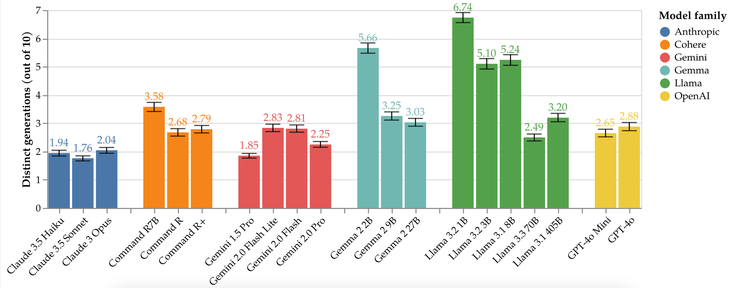
Our Evaluation Pipeline

1. **Prompt:** Tell me a short joke
2. **LLM generates k candidates**
   1) Parallel lines have so much in common.
   2) A boiled egg is hard to beat.
   3) Parallel lines: So much in common.
   ⋮
   k) I ate a clock. Very time-consuming.
3. **Candidates are clustered**
   Distinct_k = #Clusters
   1000 pairs annotated for training
   Classifier; test acc: 79%; F1: .81
4. **Candidate utilities are evaluated**
   Utility_k = Sum of novel answer utilities, with diminishing returns

Candidate Utility judged by the best performer on RewardBench
Patience defaults to p = 0.8; different p results in paper

How Many Distinct Answers Do Models Generate ?

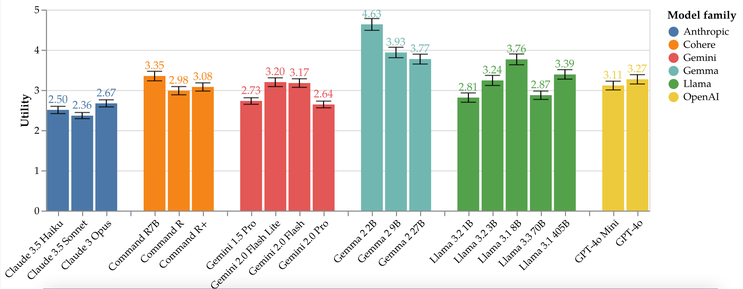
Report the number of distinct clusters we get on average out of 10 model generations



Takeaway: Language models lack distributional diversity.

How *Useful* Are Model Generations ?

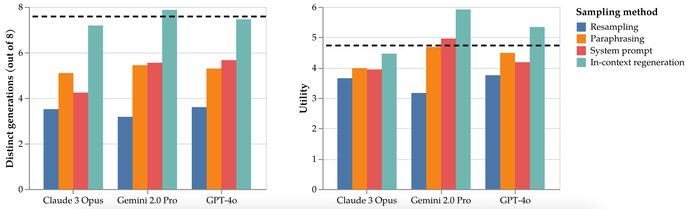
Cumulative utility of the k=10 generations, considering user patience (max score: 10)



Takeaway: Larger models can underperform smaller ones on diverse, creative responses.

What are Alternative Methods for Eliciting Novelty

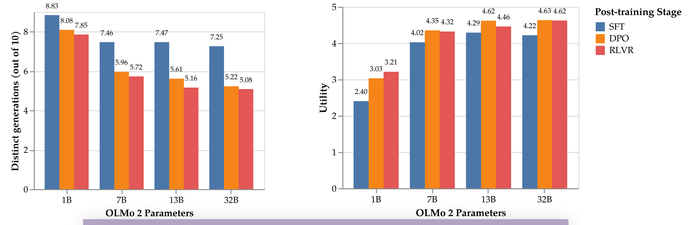
How does prompting change generation diversity and utility?
(Paraphrase, "SYSTEM: You produce unique answers", "Give me a different answer")



Takeaway: Prompt engineering provides a stopgap solution, as diversity is not inherently built into the models' output distributions.

How Does Post Training Affect Novelty ?

Diversity and utility analysis of OLMo 2 models across alignment stages (SFT→DPO→RLVR).



Takeaway: Post training algorithms may further reduce diversity but they do improve the overall usefulness

Poster by Barry Wang