Testing & Benchmarking AI Models & Architectures: Goals, Training, & Testing

Task Derivation and Integration
Recursion, computability, decidability, and formal task-spaces
Lazy Augmentation vs. Required Augmentation
& proposal for: dynamic, deterministic, training & test sets

2023.12.07

1. Selecting Goals, Designing Tests
    - The field of education's 'tobacco' moment when it comes out that most tests for most subjects are largely useless as real tests.

Testing & Benchmarking AI Models & Architectures: Goals, Training, & Testing

2. R&D Goals vs. Real-Life-Ish Goals (production-deployment)
    "Contamination" and Plagiarism

3. Turns, Queues & Stacks in Project-Space

4. Agile vs. Waterfall

5. conflicting goals and confused narratives: Why was the real gpt4 destroyed again, and not allowed to help life-saving professionals?)

6. Real Life Performance vs. Ideological-Reification of the Rational-General-Mind Paradigm

    for most of biological history "learning" was largely hard coded, or a completely effective way of what some people now derogatively call 'contamination learning' with either the presumption that the learning was not effective (biological history says otherwise) or that those people are above reality and can choose to live in a fictional bubble.

Most tasks, most of life, are EXTREMELY routine and habitual, to the point that the mind struggles to maintain coherence and perception of all the not-quickly-moving background features, lurking within which are long-neglected issues that become invisible.

7. Newer-Better Training Data with Old-Bad Tests
    - Winograd quagmire

object space tests
project space tests
decision/prospect space tests
architectural learning space tests

8. Word-Net, Fefe Li's Imagenet, & Hendrycks Math Tests

9. decision tests: system state and decision tools
There may be significant advantages to getting feedback from deliberately limited passive-reflective-models both in the short and long term. Some of the problems with decision making pointed out by Daniel Kahnamann et al will be incidentally rooted in the biology of one primate, but others are likely more general challenges that will face most systems on any planet. Like system collapse (perhaps), some system challenges are general.

.....

model testing

testing images

testing Augmentation & Retrieval (as in RAG or other search)
- indexing
- types of databases

Testing fine tuning

Testing pruning and sheering

testing set and setting

testing interactions and relationships

testing object & relationship scales and dimensions

testing representations & abstractions

testing MOE

Testing Task Derivation & Integration

Recursion, computability, decidability, and task-spaces

.....

performance tests are persistently tricky to align with future real life needs
'the Geni-wish problem': people make bad choices, want the wrong things, and pick the wrong goals.
(this is likely part of the problem, what wasted percent of resources is spent by market investors trying to make AI models that outwit other people's AI models to get rich quick on the market?)

Research goals vs. Production Goals
Image-net style top-5 to gemini's CoT@32
There are a number different research, deployment, or other, goals and foci that you may have when testing out an AI system.

While it is surely good that people are attempting to be thoughtful about what they are testing for, the pattern of history appears to be very still very strong at the end of 2023: it is very difficult to carefully select what you are testing for.

How many times have people in business been presenting with a nice looking numerical simplification that was explicitly a bad oversimplification with data, and the business person enthusiastically says that the clean simple number is all that matters...only to have that choice (very predictably) days weeks or months later turn out to be exactly what it looked like on the surface: ignoring a diverse reality to believe in a simple fiction, with destructive consequences.

R&D vs Production-Deployment (again):
This time for AI-ML testing.

Research Modes of Testing

'real-life-ish' modes of testing:
- Time & Frequency Map of Reliability:

Being caught in the gears of the paradox of the 'annecdote'

"Contamination" and Plagiarism:
- in research testing, and arguably justified, there is concern about

this also may relate to divergent uses of the term 'zero-shot' (as if people saw how utterly confusing the uses of the term 'zero-day vulnerability' was in cybersecurity and were irresistibly drawn to a new potential source of confusions. please seek clear communication.) Where you do not want a test answer to 'simply be memorized.'

This is however largely a fictional problem in terms of every-day life. If you are flying from NY to LA and the pilot manages to completely avoid any turbulence, would you viciously round on the pilot with a red face and pointing finger saying: "You memorized that! You trained on turbulence! THat wasn't real!! Yourcheated! You plagerized! You knew now to avoid standard terbulance! You have cooties!!! You're fired! Your're black listed! You're euthanized! Your family is Euthanized! I WIN! YOU LOOSE!!! Youre weeded out! Your excluded!! MWAHAHAHAHA!!!!" Hopefully not.

From Image-net style top-5 to CoT@32:

CoT@32
Table 2 | Gemini performance on text benchmarks with external comparisons and PaLM 2-L. * **The model produces a chain of thought with k = 8 or 32 samples, if there is a consensus above a threshold (chosen based on the validation split), it selects this answer, otherwise it reverts to a greedy sample.** Further analysis in Appendix 9.1. ** Results self-collected via the API in Nov, 2023. *** Results shown use the decontaminated nu

top-5
five-shot
0-shot,

Kahnaman Tversky Tests (non-human abilities)
project-state tests:
project object tests:
cut-up tests:
coordinated decisions tests:
coordinated-resource tasks: herding-cats tests: shared-mouse tasks

"shot" vs "try" vs "top" vs "k/K"

open-turned-based tests:
open-turned-based tests are similar to a shared-mouse type coordinated-resource task, but in this case the shared resources are more like time and space. For example, if an admin said to a group of applicants "please step forward and state your name" if everyone did so simultaneously the result would be a useless cacophony. Depending on the (term?) modality or data-type or test type, this could be text or visual or robotic etc.

Participating in a multi-participant space, e.g. a text-only multi-participant chat space, where what everyone types is visible may represent a challenging environment for classic generative passive-reflective bot, because it would need to carefully separate its own role from everyone elses.
This also raises some process questions, most passive-reflective bots are kind of inherently discreetly turn based, for example in their own quasi-state kludge of a 'conversation' or ('context') memory.
So what happens in a live situation where many participants, human and bot, are acting and interactive in real time, with no nanny super-participant to curate the bot's memory?

Google's real-time Gemini may be able to do this, but any turn-based open-ai 2023 bot would have real trouble navigating this alone, even if the task were: state your name, where the implicit task was that the participants would need to negotiate among each-other in what order to go, let alone to remember who said what, and what to do when collisions happened, or even to recognize that a collision was happening.

In theory, very simple micro tests involving only a few parts and steps can be arranged and rearranged (with perhaps scales of modular elements as well) to test different aspects of system-project-state and modular-externalization even for very simple and discrete tasks. (more convoluted examples can of course also be tested out).

Scouting Ahead: team adhoc queue coordinated decision negotiation

Another level of process step, still simple overall, is if the model/AI cannot themselves see what is happening in the queue, so they need a scout to scout-ahead and report back on what is happening.
Note: The scout will only know what they see, they will not know per-se if there is a queue collision, so they can only report what they found, and the speaker/writer/typer will possible need to communicate back and forth about what was written at what times by then to the scout (or perhaps that's up to what they decide to do).

Ad hoc queues, turns, and stacks in tasks and projects with multiple participants (perhaps in some cases with only one participant):
-

specific test example recommendations...

- routine task tests
- modular task tests

...

benchmarks for ai lm ds

dynamic, deterministic, training & test sets

Classic Analogy Question:

```
image-net    -> narrow deep learning
?            -> broad foundation models
```

paper on 'generalized' test training
https://lmsys.org/blog/2023-11-14-llm-decontaminator/
https://arxiv.org/pdf/2311.04850.pdf

Lazy Augmentation vs. Required Augmentation

This seems to raise questions about the nature of data-augmentation in model training and even generalization itself. Clearly, ad absurdum, we cannot pragmatically go in the direction of saying that any model that can successfully answer a question has been contaminated. The whole point is to train a model with 'rephrased' or 'augmented' training data that is not exactly the same as the validation/test/real-world data that the models need to generalize-to.

Lazy-Bad vs. Good-Required
We were shown an example of (what I will call) "lazy" augmentation of data, where only a few words were shuffled around but the numbers and perhaps most importantly the sequence and 'key' of the answers was the same, creating a bad-augmentation path where instead of generalizing the solution process the model simply memorizes: answer = choice 'D'

Let's compare this to Khan Academy which deterministically generates and shuffles math questions, but not in a bad-lazy way: they do not just move a couple of 'the' and 'is' words around leaving the numbers and answer completely unchanged.

The problem I am trying to illustrate here is that "rephrasing" is NOT the problem. Data augmentation is NOT the problem. Generalizing to test data with training data that is not exactly the same as test data is NOT the problem. Answering questions you have never seen before is NOT the problem. Studying and using past testing and answers is NOT the problem.

The problems are: a failure to test-train split, a failure to rephrase, a failure to augment, a failure to train, and a failure to answer never-before-seen questions (or a failure to test).

Let's look at two examples: Image-net and Khan Academy

Image net was a huge database of labeled images to see if a model could learn that data. I'm guess there was also a non-public testing set. Indeed, I think there was a scandal where one participating group (not naming names) rigged and fraudulent way to re-train their model based on leaked test-set answers (very paraphrased here).
https://www.zdnet.com/article/baidu-admits-cheating-in-international-supercomputer-competition/
https://www.theregister.com/2015/06/05/youve_been_a_baidu_boy_tech_giant_caught_cheating_on_ai_tests/
https://www.dataversity.net/baidu-admits-to-cheating-on-an-artificial-intelligence-test/

So in this new 'fear of contamination cooties' world, you cannot train on any cat photos or any animal photos before asking asked to identify a cat, or you are 'cheating'?

e.g. What if instead for the example paraphrased question where mininal paraphasing of words when the numbers are exactly the same and the answers are in the same order could result in the model memorizing that the answer is 'c' given that multiple choice question.

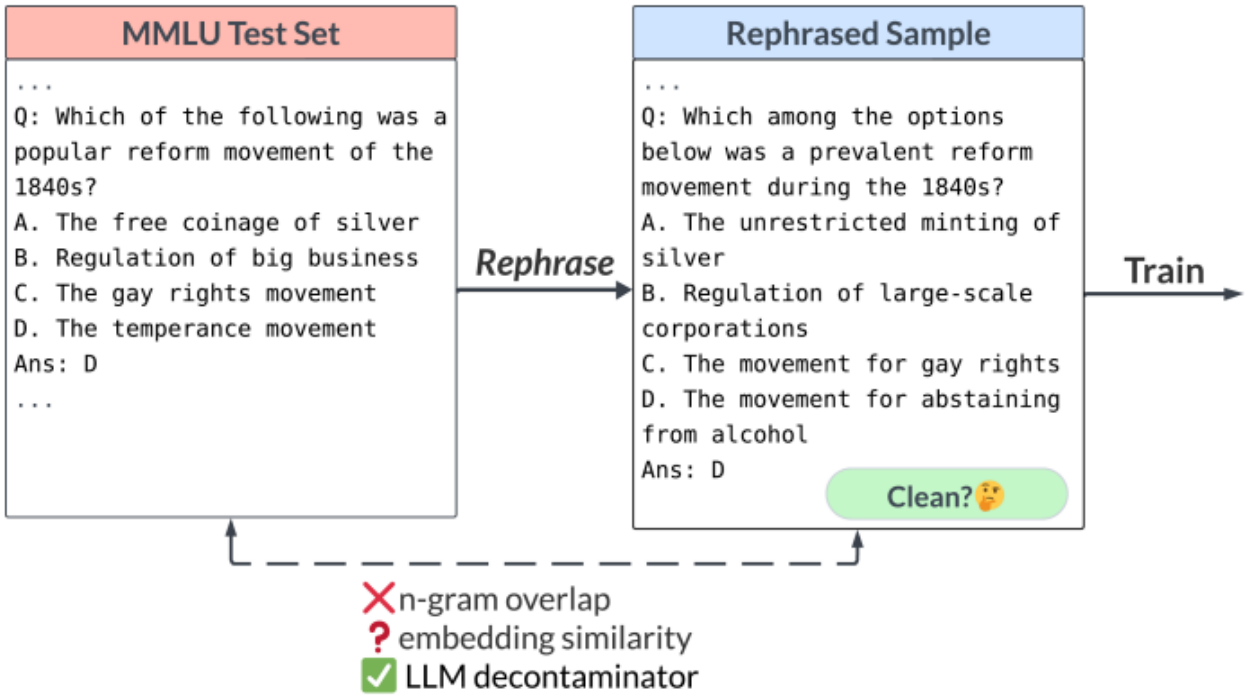
We need better training open source training sets.
We need better dynamic-testing sets.

Note: in same cases, memorization is the goal.
Asking a question about the periodic-table of elements...is a memorization question, in various ways.
E.g. Any arbitrary information is a memorization question.

The old Chess-reasoning question.

'Clever Hans' and Goal Selection:
-
-

Think about image-net.



there is a bottleneck crisis now for testing-training data

deterministic Open Source,
math & reasoning,
training set,
and testing set,
generator

(like khan academy)

based on a curriuclum-tree of math and reasoning skills

vs.

pure memorization questions

object tests

multi-participant project-tests:

...

changing variables
changing numbers of variables

making cut-ups

maybe...comparing object-rx between questions:
- not asking the asking, but asking about similarities or differences between two questions

schedule and long term questions

...

A testing-set must also be a training set:

again: image vs. language

The questions of how 'best' to do test-train splits:
- cross validation
- is there one solution for all uses and architectures...probably not.
-

Issues with with-holding data from training:
- which data do you not train on?
- is there data which is beneficial to test on that would not be beneficial to test on?
- is it possible to keep testing-data secret? (e.g. Baidu & Image-Net)
- the question of training on all the data, in some cases, after your training pipeline has been well tested.
- advantages of modularly, dynamically, deterministically, generated tests

...

Non-deterministically generated test data:
- humanities subjects
- screening
-

...

clear-area testing and grey-area testing:

...

input-output measures:

-

...

A growing, open source, modular, set of testing and training tools:
-

...

AI Testing and any-participant testing:

...

Task Derivation & Integration:
- when do you stop sub-dividing a question?
- re-assembling answer parts to get solution

Multi-tasking:
- two questions, keeping object separate

compound questions:
- same premise, two questions at once

curriculum tree of base basal questions
- derived distal questions
- in dimensions of derivation:
- new variable names
- new answers
- new sub-types of answers
- multiplication of variables
- edge cases

...

standards for answer: only one final answer
decimal rules
fraction rules
spelling error rules
singular-plural rules

...

Starting at about 10:50sec
Analysis of issues with MMLU question-answer sets.
"Phi-2, Imagen-2, Optimus-Gen-2: Small New Models to Change the World?"
by AI Explained
https://www.youtube.com/watch?v=nPgs8THgbuI

I'll try to double check this, but this may be an important example of not simplistically aiming to train a model on scores we believe should be correct.

...

modular questions

measurable scales of type of questions

...

clear line and fuzzy line between memorization and conceptual generalization:

2. data structures
3.
4. chess again: chess openings and pre-game prep: opening lines
5.

...

Cross-Reinforcement of Different Skills vs. Specialist Models

It is still not clear whether the best way to get a model to be skilled in one area is to:
A. focus on just that area, or
B. to not focus on just that area.

...

Super-Signals & Demand Distortion:
- It is inevitable that deleterious and abstract measures will be fixated on by some people who 'just like those' measures, even though the measures (and the people) have no connection to reality.

This historically has also been used for crude 'monopoly by a rule' methods, where groups will secure an advantage by making rules saying only they can compete.

...

Do not assume that people love reality and feedback.
Do not underestimate the degree to which biological H.sapiens-humans are violent animals who physically destroy attempts to knowledge STEM.
Do not ignore or deny historical data.

...

Different Cultures of Ethics, Honestly, and being opposed to Fraud and Corruption:
- some people and groups of people embrace nihilistic fraud and corruption
- some people and groups embrace STEM

...

The AI Job-Interview Problem:

Just as hiring people by people in HR is a total mess of miscommunication, mismeasurement, bad questions, bad data-management, bad communication etc., evaluating AI appears to be following the same trajectory of ineptitude, where the same bad data and bad tests get recirculated into an ever more nonsensical toxic stew of dysfunctional ignorance, coverups, fraud, and gang-pecking-order-cargo-cults.

Practical questions, Useful questions, pedantic questions, Gotcha questions, Hazing questions, Telepathy questions, etc.

...

NIST & Education

Testing the Tests

...

Curriculum areas:

math areas

...

STEM is a start, but just a start

...

What is included in STEM?

...

STEM & Projects

...

STEM & Humanities

Non-deductive areas and disputed areas

...

expert reviews per discipline

...

bug-hunting

...

Math logic and STEM

values, functions, equations

systems of logic

computability

notation etc.

...

"proof"?

...
Tests and Retrieval Augmented Generation, & Generation Augmented Retrieval
where 'retrieval' can mean any use of external-data-tools, so:
"external-data-tool & AI-model interactions & interfaces"

You can come up with extreme examples of cases where you want to have one of a few different goals such as:
1. use guestimation but reduce guestimation-inaccuracy, or
2. have no guestimations and directly do database-queries guided by the AI

on the one end of the spectrum an AI has already memorized most of the data in the database,
on the other end of the spectrum the topic is so distant from the models training that the model cannot interface well with the database, perhaps the database is entirely medical-jargon that the model was never trained in,
or the database is (not sure what a good example of this is) new data such as...maybe a completely new set of structures by a completely new telescope which have no direct connection to anything in the model.

What is a good, or best, way to train the model on the right amount of data about the information in the database

This may also connect to scale and intergration-of-task issues, where as tasks involve more and more levels and larger scales of parts, you need new parts of the system to, e.g. managage all the different 'function' type operations that the model is supposed to use. Asking a model to query wikikpeida is not a problem. But asking the model to juggle 500+ different query tools and sources does become a problem. And at each level there may be a 'familiarity-level.' issue.

Also, there may (somehow) be familiarity over-fitting or otherwise as the general topic and theme goes, not (for whatever reason) getting good results in performance due to which training data was used in what way.

...

Output formatting:

generating training/testing data

...

Modular Approach to not-overfitting.

When each skill is modular, skills can be combined with new question details in ever-changing ways.

If some type of over-fitting is detected, you can (dynamically)
alter the modular-creation-augmentation of your training data
so that the examples are more unique and less similarly augmented