Evaluation Rebellion: Why LLMs Need Report Cards Instead of Tests

3 Days ago, one of my great Mentor and Friend, Scott, Posted a document about evaluation of the model. https://ysymyth.github.io/The-Second-Half/
The article argues for moving beyond benchmark performance to focusing on real-world utility and impact.

Today's benchmarks work like paediatric growth charts – useful for spotting regression ("Is Model v2 dumber than v1?") but useless for assessing adult capabilities. When GPT-7 scores 99.9% on every test, how do we measure that a model is better?

- Do we invent harder trivia? (Spoiler: Models will outpace us)
- Do we shift to qualitative metrics? (But how to compare?)
- Do we stop comparing models entirely and focus on the Human getting better?

Rethinking Metrics – Beyond Knowledge, Beyond Machines

Let's investigate two parallel paths: **inventing new metrics** that go beyond just knowledge assessment and **repurposing human-centric frameworks** never designed for AI to assess a model.

Part 1: Can New Metrics Capture AI's Human Impact?

Frameworks like HumanELY exemplify this first approach, scoring models on five dimensions: relevance (answer alignment), coverage (completeness), coherence (logical flow), harm (bias/safety), and comparison (relative performance). Unlike traditional benchmarks that test *what* models know, these metrics assess *how* they communicate and collaborate – critical for real-world applications like healthcare or education.

While this helps frame a model in a different perspective that ROUGE and BLEU, it is still a human feedback metric (meaning bias) and fairly involved in terms of time.
https://www.medrxiv.org/content/10.1101/2023.12.22.23300458v2.full.pdf

In the future there will be a plethora (love that word) of benchmarks, each calming to be better than the other.

Part 2: When the ChatBot takes a personality test.

The second path I wanted to investigate is , can we adapt tools from psychology and HR and measure LLM?

The Machine Personality Inventory (MPI), modelled after the Big Five personality traits, categorizes LLMs into spectrums like openness-to-experience or conscientiousness. Similarly, MBTI-inspired evaluations classify models as "INTJ" or "ESFP" based on interaction patterns. While these frameworks help match models to user preferences (e.g., a detail-oriented "ISTJ" LLM for legal research), they risk anthropomorphism and may overlook uniquely machine strengths like probabilistic reasoning.

https://arxiv.org/pdf/2307.16180 and https://arxiv.org/pdf/2206.07550.

What if I was interviewing LLM before using it? So I understand if I can work with them or not.

The Hybrid Future

Is a creative-but-unreliable LLM better than a boring-but-accurate LLM? The answer depends on context, what do you want to achieve today> What is Important?
This is a reminder that evaluation must evolve alongside how we, human, *use* AI.

Let's dig further….can we measure the performance improvement of the Human-LLM as a whole instead of just the Model ?

The Human-AI Collaboration KPIs

Everything we talked about previously was making sure the LLM could perform on its own, Less error, better response, no bias. But how do we measure a LLM as part of a team, as an assistant?

Part 1: Measuring outcome – Who leads Improvement?

When LLMs act as assistants, the evaluation shifts to the *human's ability to guide the LLM*. Consider two coaches training the same athlete: their methods, feedback, and adaptability produce different results. Similarly, two users interacting with the same LLM-one crafting precise prompts, the other using vague instructions-will yield vastly different outputs.

This raises a critical question: **Do we evaluate the model's raw capability or the user's skill in directing it?** If an LLM "fails" a task, is it due to poor training data (the developer's responsibility) or unclear guidance (the user's responsibility)? Coaching frameworks suggest a hybrid approach:

- **Prompt Engineering as Skill Development**: Like coaching communicate with athletes, users must learn to communicate intent effectively in order to get the bast out of the model.
- **Model Training as Talent Recruitment**: Coaches create set of 'drills' to improve the performance of the athletes. Users will also need to "train" models through iterative feedback.

Part 2: Measuring Partnership Success – The Doctor-LLM Case Study

The true test of an LLM assistant isn't its solo performance but its impact on human outcomes. Take healthcare: a doctor using an LLM could generate accurate diagnoses faster, but if the partnership leads to *patient lack of trust* or *overreliance on AI from the practitioner*, is it truly successful?
Key partnership metrics might include:

1. **Outcome Improvement**: Does the doctor-LLM duo reduce misdiagnoses compared to the doctor alone? Hard to measure, but let's keep it here anyway as a goal.
2. **Skill Transfer**: Does the doctor learn from the LLM's reasoning, improving their own diagnostic skills over time?

Part 3: The Balancing Act

Measuring only the output of the model may leads to missed opportunities. For example, an LLM might generate "incorrect" responses that inadvertently spark creative problem-solving in humans.We see that in brainstorming sessions. In some sports, an atheist “makes the team better” even if themselves are not the super star on the field.

The challenge lies in designing metrics that value *both* guided performance and unexpected synergies.

Conclusion - The road ahead *where the author goes on car ride analogy thanks to AI*

Here’s the uncomfortable truth: today’s metrics are becoming the seatbelts of AI evaluation- essential for safety but irrelevant to whether you arrive exhilarated or exhausted.

Just as checking tire pressure tells you nothing about a car’s handling on mountain roads, current benchmarks ensure models don’t regress… but say nothing about their real-world impact.

Similarly, we keep generating harder tests. But once every vehicle ascends Mount Everest, how do we measure progress?

Consider two GPS systems, and imagine they are two different models.

- **GPS A** prioritizes speed, never deviating from highways.
- **GPS B** discovers scenic routes but occasionally gets lost.

Neither is “better”. Their value depends on whether you’re rushing to a meeting or exploring Tuscany. This is the same for a model, we need to link the concept of better to an outcome.

Let’s think even further about the definition of : better. To me it is linked to other concepts like success. Are you more successful than your neighbour? Are you better? How come? Who decides?

While measuring and data driven decision are important, and we can clearly evaluate a model on very specific tasks and outcome, ultimately at one point the ceiling will be reached, and then there will still be a part of the evaluation that will be left to the user.

With the editing and partnership help of perplexity