The Art & Science of Benchmarking Agents — Vincent Chen, Snorkel AI

23 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=INKFLCIIJ0U
https://www.youtube.com/watch?v=iNkFlCiij0U

SUMMARY

Vincent, a research fellow and co-founder at Snorkel AI, discusses the importance of building effective benchmarks in AI research and deployment. He emphasizes the need for robust evaluation methodologies that not only measure current capabilities but also shape future advancements in AI, particularly in high-stakes environments like finance and healthcare.

- *There is a gap between AI capabilities and the readiness of enterprises to deploy these agents in high-stakes environments.*
- *Effective benchmarks should not only reflect past progress but also define future goals and capabilities in the field.*
- *Key factors for building useful benchmarks include individual task quality, distributional diversity, difficulty of tasks, and robust evaluation methodologies.*
- *Successful benchmarks should have a clear thesis about where the field is heading and inspire new research directions.*
- *The researcher experience is crucial for the adoption of benchmarks; they should be easy to use and extend for the community.*
- *Snorkel AI has initiated the Open Benchmarks Grant, committing $3 million to support the development of new benchmarks.*
- *Future benchmarks should focus on environment complexity, autonomy horizons, and capturing nuanced output complexities to better reflect real-world applications.*

Everybody, how's it going? Lovely. Well, I'm very excited to be here hailing from San Francisco. It's a little bit of a trek over, but I'm super excited to to chat with you all today after the talk and beyond. my name is Vincent. I'm a research fellow and a co-founder at Snorkel AI. And you know, today I'm going to be talking about some meta evaluations for building benchmarks, the art and science of what we found to be really useful when building effective benchmarks. I have the great privilege at Snorkel of working with both our researchers, great collaborators in academia, industry, and the open source community to build great benchmarks. And I wanted to share some of the learnings that we've had over the, you know, last few years on on what really makes for benchmarks that shape the field and and move it forward.

So, a little bit about us. we're a frontier AI data lab. we have, you know, labs both at at academic settings, you know, our our co-founders have have labs at Stanford, U Dub, Wisconsin. We've an internal team for deployed engineers, applied engineers. And I mention this because we get a lot of exposure to both, you know, the academic frontier, work with frontier labs, and also real enterprises and and companies who are deploying in

practice. So, our focus as a company is on building the best data sets and environments to define and advance future AI capabilities. And we are in a fortunate spot where we get to play at this unique intersection of both, you know, the frontier academically, but also to contact reality with our deployments in enterprises.

So, today I wanted to talk about an asymmetry that we see in our real-world deployments. There's real excitement around agents today. You know, we see it at every person in this room, I'm sure has has played with these these these agents and we see a real progress marked by, you know, hill climbing on model cards. We see the vibes are improving, right? And truly shifting, especially in coding. But when you ask individuals and enterprises or these, you know, large-scale organizations if they're fully ready to let these agents loose and you know, deploy them in high-stakes environments, you get a little bit of hesitation. And that's not to say the capabilities aren't there, but our ability to actually measure these agents in practice that is falling behind of where the capabilities actually are.

This is one of the challenges and research questions that I think are actually one of the most important in the field and one of the ones that we're very interested in here at Snorkel. So, closing that gap, that evaluation gap, we believe requires a toolkit, right? As I mentioned earlier, we're strong believers in field deployments, right? So, this is you know, actually deploying engineers, researchers on our teams to again, contact reality and and and work with folks to deploy these models in real production settings where the stakes are high, right? These are finance settings, insurance, you know, healthcare settings where it's not just about a number, it's about real outcomes. And we're also big fans of other eval tools, right? This is red teaming, private human evals, crowdsource labeling, a lot of the the themes that we saw talked about today, which which has been awesome to see.

But one of the things that we feel most strongly about is that benchmarks, open benchmarks in particular, remain a really critical piece of the measurement toolkit. The best open benchmarks aren't just about, you know, taking a snapshot of progress looking backwards. They're actually about defining progress and shaping the field and setting a goalpost about where capabilities need to go. And you know, even looking at the last few months, right? Benchmarks like Terminal Bench, Meters Long Horizon Benchmark, ARK GI, these are really exciting and critical guideposts for where the field is going. And And And as a result, the path to safe, trustworthy agents will will really depend on more of these benchmarks in practice.

So, what are we doing at Snorkel? One of the things that again I'm I'm very fortunate to to be able to be a part of is the Open Benchmarks Grant. We recently, a few weeks ago, a month ago, deployed $3 million to commit to open benchmarks. And this is really fun job. I get to work with the best academic teams, you know, builders to really accelerate and fund the next wave of benchmarks that's going to really steer and and and guide where the field is going. We've had a wild reception so far. I'm a little admittedly behind on some some reviews. but we've been really excited to see what the community has come up with so far. And in this talk in particular, you know, we've reviewed I think over 120 applications so far, spanning academia and industry labs. We wanted to share a few perspectives, a few learnings over the fast past few months about what we view as one table stakes for for useful benchmarks, right? How do you actually build good empirical measuring sticks that are actually useful, you know, to to measuring progress? And two, what really separates, you know, those benchmarks that are

shaping the frontier, right? What it What is the art and and the the the science, if you will, of building really official effective benchmarks at the end of the day.

So, as I'm doing this, I'll have a fun opportunity, maybe this is a little too American, but to pull a Timothy Chalamet and honor some of the greats, you know, some of the great benchmarks of the last few years that have really shaped the field in talking about some of these axes. And I hope that these, you know, themes resonate with you and also inspire a bit of you know, kind of new thinking about, "Hey, how how can we actually deploy some of the learnings we're we're all kind of driving towards in our day-to-day work to, you know, shape the field and and move it forward. So, two themes here again on the science side, you know, how do we actually build effective measuring sticks? We'll talk about task quality, distributional control, robust evals in general.

and on the on the art side, right, really the the differentiators for great benchmarks, how do you build benchmarks with a thesis on where the field is going that inspire new road maps, and that critically are are built for this audience, right, a researcher audience, a builder audience so that adoption is is something that is a way smoother and and a first-class citizen for a bunch of these benchmarks. So, let's start with the science. this is again what makes for really effective measuring sticks as we've seen them in in practice and in the the deployments that we see in industry and academia and with frontier labs.

So, the first thing I want to talk about is individual task quality. Right, this is the idea that individual tasks need to be exceptionally rigorously validated, right? They need to represent real-world complexity. They need well-posed, well-structured instructions. they need verifiable solutions that ideally have been actually validated by real-world domain experts. one of the the benchmarks here, GPQA, is one of my favorites not just because it's been a very lasting and enduring benchmark that captures, you know, graduate-level and professional knowledge even to this day, right?

You kind of still still see this on model cards. But, one of my favorite contributions is actually tucked away in the appendix. GPQA introduced one new adversarial qual- quality control mechanism. So, the idea was that not only do these tasks need to be well-posed, they need to be tractable for other experts to solve. So, they had a very very rigorous multi-reviewer protocol where there was an original author, you know, there were there were reviewers and adjudicators in the loop.

There was opportunity for revision, right? These were tasks that were really pushing the frontier of knowledge and it was non-trivial for any single expert to say, "Yeah, this is actually a good task or not." And so, developing this sort of rigorous adversarial quality control mechanism was one of the contributions I was most excited about here. And if you read the appendix, you also see that they introduce new incentive mechanisms, right? Payouts are actually based on whether there was certain agreement and, you know, coming from academia, you know, there's there's some inspiration here from the peer review process as as flawed as that is, but, you know, this type of innovation around how you actually get really rigorous, you know, multi-expert quality control leads to the type of outcomes that we see around individual task quality that we we see as a a key foundation for any benchmark that matters at the end of the day.

Two is distributional diversity, right? This is the idea that for any benchmark that really matters, you want to define a clear taxonomy for the domain, for real-world tasks, and distribute those tasks intentionally. So, this might be, "Hey, I I captured a trace or or kind of real-world stream of traffic along you know, how my agent is operating in the real world, and I want to really represent that distribution." It could also mean, "Hey, I'm specifically characterizing and taxonomizing the failure modes that are, you know, paradoxically rare, but you know, disproportionately important in in production, right? If you take classic self-driving settings, right?

Yellow lights or, you know, pedestrians or motorcyclists, right? Might actually show up way less than other other types of scenarios, but are disproportionately important, you know, to get right in these settings. And so, defining that taxonomy, being really intentional about distributing tasks across it is one of the hall hallmarks of great benchmarks in our view. MMLU, few years old now, constructed a a quite ambitious taxonomy of, you know, 57 academic and and professional domains across STEM, humanities, et cetera. It's remained one of the lasting benchmarks for understanding graduate and professional level knowledge and again a lot of this was as as we we believe a result of really thoughtful and intentional taxonomy designed and and building towards that.

The third axis here is around difficulty of individual tasks and model headroom. Right? It's really important that the benchmark is unsaturated that it exposes real soft spots in capabilities and reliably separates where models sit at at the frontier. One of my favorite plots is is the one on the top right. This was you know all all credit to the ARC Prize Foundation team. ARC-AGI 2 you know for a very long time was unsaturated right for for several months and years. And when there was the big reasoning push you know maybe 18-24 months ago we saw a massive leap in capabilities that actually corresponded to a real leap in model capabilities. Right? This was a benchmark that was intentionally designed to represent a type of efficiency or capability that humans have but models didn't have and and they really kind of captured well hey there's a lot of model headroom here. Humans can do this.

Where's that gap? And again low and behold it correlated quite well with the recent you know 01 style reasoning push that has really dominated the field in the in the past 18-24 months. Just a few weeks ago the ARC team just launched AGI ARC-AGI 3 and again at launch they had frontier models under 1%. Every single task was a human solvable to some degree and so it remains one of the I think the the most meaningfully exciting benchmarks in the space where you know any new model you know people are kind of awaiting hey how does it do do on ARC? And and I think they did this quite well. Right? The the kind of model headroom here is is really really exciting.

This last axis here I want to talk about on on the empirical measurement side is all about a robust eval methodologies. Now this goes really deep. So just kind of capturing some of the high-level ideas, benchmarks need to ideally go beyond accuracy to capture real-world dimensions that matter, right? This is everything from cost, latency, you know, the quality of the reasoning traces, some of the intermediate steps and and and tool use. Whatever dimensions actually matter for the capability at hand, capturing those as reward or supervision signals is really critical and measuring what it what it claims to is is actually a non-trivial feat in in you know, building robust and and reproducible benchmarks. So Tow Bench is a benchmark that we're we're a big fan of, you know, it's had multiple evolutions over the year, but over the years, but it would it was a benchmark that was

built to evaluate both task completion of these multi-turn agents.

They built a a clever, you know, kind of user simulator, but also, you know, not just accuracy and completion, but adherence to policy constraints. So a model, for example, on the right-hand side, this was the one of the examples from the paper, a model that books the right flight, but violates fair class rules, still fails, is still a kind of no-go at the end of the day. So this notion of hey, being intentional about what axes we actually care about, what do we actually want to measure and measuring that rigorously, is one of the hallmarks that matters when you're when you're building these these frontier evals.

So I want to shift a little bit now to the differentiators, right? What what actually leads to the benchmarks that push the frontier. That's not to say, you know, anything I mentioned on the last few slides have not pushed the frontier. These are just special characteristics that I view as you know, critical to to the benchmarks that are real research contributions, that are really shaping where is the field going, where are the where are all the labs going to hill climb next?

and this is this is the art, the special sauce that helps push us forward. So one of the key hallmarks here is these benchmarks should have a thesis, right? They should have a research question about a subspace of capabilities, about where the field is going. It should revisit previous capabilities. and the most ambitious benchmarks are really a statement about where the world is going. Terminal bench is one of these bets, right? It was a bet on the CLI not just for coding agents, but for general purpose computer use. And in many ways I think this has turned out to be a largely correct and and consequential bet, right? As as we're seeing teams that had kind of Claude and and and Codex build their general purpose, you know, enterprise capabilities on top of these coding and CLI based tools, we're seeing this bet pan out. And again, Terminal bench remains one of the most robust and kind of most important benchmarks that are measured on on all the recent model cards. So, again, this was a bet early on to say, "Hey, we think the CLI is going to be really important as a core interface, a core abstraction and affordance for agents to interact with the real world in general purpose way. And by measuring those capabilities, you know, I I'd argue that it actually helped accelerate you know, how how the field is operating in this way.

The second piece here I think worth mentioning is the ability to kind of roadmap for the field. a a great benchmark, you know, one that really shapes where where all of us are going is producing new roadmaps, right? It's inspiring new attacks against research problems. It's helping folks ideate and come up with new ways for thinking about benchmarks and methods in general. And I think SweBench is really phenomenal example of this, right? It was a simple idea as often the best ones are quite simple, right? Head, how do you kind of leverage you know, existing coding coding type capabilities via via PRs. and it spawned a new family of benchmarks, right? all the way from SweBench light, verified, pro, multilingual, multimodal, etc. And its evolution I think is is still very relevant today. It's evolved how we think about coding agents and one of the things that's been awesome to see with the SweetBench team is how many new research directions and kind of inspired benchmarks have come after it in this coding space. And arguably I'd say, you know, there's a lot more room to kind of innovate on top of this as well. What are the new ways of coding look like? How do these types of workflows apply to 5 coding and kind of this new layer of abstraction that that's offered developers are

applying?

I think it's been really exciting to see the the foundation that the SweetBench team set and how that's going to shape, you know, how we think about coding agents moving forward. So, this team here I think is severely underrated and this is the notion of researcher UX. Right? I think the most prescient benchmark builders are committed to the researcher and builder experience. This is to say, it's really simple to run models and agents against your benchmark. It's really simple to contribute new tasks to extend. And also it's it's really simple to leverage some of the the signals that you're getting for the benchmark for RL or kind of tuning on post talk. I think this is really underrated. It's you know, a classic product principle to make what you're building and putting out there easy to use by the community or by your core users. And in this case, benchmarks have core users which are other builders or researchers. And so, really putting in time and attention to building those interfaces has been important for the adoption of some of the most important benchmarks. Right? To call out I think the the Stanford team at CRFM built Helm you know, several years ago which I I'd argue kind of pioneered a standardized modular harness for evaluating reproducible, you know, different scenarios as well as kind of models against a standard test bed of models.

Terminal Bench 2.0 just a few months ago again shipped with Harbor which has been in many ways a de facto harness and and kind of evaluation infrastructure for teams who are building agents more broadly. And so, you know, thankfully we have a bunch of open source software out there today, you know, based on this principle. But as you're building your benchmarks, right, kind of considering, hey, how easy is this to extend? How easy is it for the community to adopt and eventually kind of hill climb against this? I think is a severely underrated factor for for what makes for really high adoption of of these frontier benchmarks.

So, this is the full framework. Again, can go into more detail and and please find me afterwards. But again, what makes for really empirically meaningful measuring sticks, right, it's task quality and attention to distributional control and diversity. it's it's difficulty and model headroom. And of course, a robust eval methodology that measures the concrete axes that actually matter and it matter in practice and is is intentional about it. And of course, on the off side, right, these these great benchmarks really have a thesis on where the frontier's going.

they set road maps for the field and they really prioritize high research UX. Now, before I wrap up, I want to propose, you know, a few dimensions that we're really excited about at Snorkel that we think are really going to encapsulate, you know, the the next wave of of benchmarks. tried to leave some more degrees of freedom here for creativity, but these are areas where we think there's a lot of room to push complexity, to push, you know, realism in benchmarks. And so, I wanted to share a little bit of our internal road map and thinking around where the field is going and where we need more benchmarks and more contributions.

So, this is our point of view. we think that the the axes for the next great benchmarks are threefold. I mean, I'll go into a little bit more detail about what I mean here in just a second. But one, it's environment complexity, right? How complex, how realistic, how dynamic is the operating environment that these benchmarks are working? Are they representative of real-world settings that a professional, that, you know, scientist, that someone using these tools could actually use?

Two is autonomy horizon. Do these benchmarks represent realistic and frontier horizon lengths that these agents are operating against? Are they capturing different points on the autonomy slider that are again representative of how users are using them as co-pilots versus fully autonomous agents? Is this an intentional design in the benchmark? And three, capturing the wide range of output complexity. I think this is very under explored today, right? Lots of chat-based or document-based outputs not as much around nuanced kind of differentiated reward signals, right?

Real artifacts that show up you know, in day-to-day work that we that we represent. And critically, new artifacts, right? New types of form factors that we haven't even imagined about you know, how agents interact with humans, how agents interact with each other. So, a little bit about each one. the first one here, I won't go into all of these in in significant detail, right? Environment complexity is all about capturing the real-world complexity that that is in our in our day-to-day working environments, right?

And this this gap is often where agents fail today. Consider coding agents, right? A real codebase has org-specific policies, you know, lots of Slack context, screenshots, flaky toolchains, you know, CI that's that's kind of distributed, human reviewers with knowledge in their heads about what they like and what they prefer, many contributors in parallel. Benchmarks today capture a fraction of this complexity and you know, not just in coding but other domains, there's a lot of excitement and an opportunity to up the level of complexity and continue to drive what these models can do to to represent real-world uses.

Two, again on autonomy horizon, right? this is all about how long an agent can operate before it reliably before reliability breaks down. Right, let's take a customer experience agent, you know, in in many cases, right? these these agents may lose track, you know, of of context that was, you know, delivered a few weeks ago, different integrations or product specs might, you know, change the actual spec or requirements for for a particular model. Reorgs can kind of shift, you know, priorities midstream. real-world settings actually represent a lot more complexity that again is represented in these kind of long-term continual learning type type settings that represent kind of changes in state and environment. so we again think that there's a lot of room to contribute and and build out new benchmarks that represent very very long horizon and an autonomous agents.

And lastly, this axis is all about producing more complex work, more representative work, and also nuanced signals that can be used for not just evaluation, but reward signals during training. This also has to be complex, and this gap is is growing as well, right? Let's take again a software example, right? Or or a complex report for making strategic recommendations. It's non-trivial and subjective, you know, to define, "Hey, what is verifiable about a good recommendation, a good strategic proposal, a good roadmap in general?" the nuances of this need to be captured well. They need to capture organizational context, really good human judgment, and tomorrow's benchmarks really, you know, we're we're excited about, signals that capture all of these settings.

Trustworthy outputs, right? The ability for agents to actually capture their own uncertainty and define, "Hey, I'm actually not sure about this. I actually need to stop or kind of ask for more information." Again, different types of outputs that aren't just, you know, a kind of plain text answer is something we're really excited about. So again, hopefully, you know, this inspired a little bit of thinking around, "Hey, what am I working on? How

can I turn this into a meaningful benchmark?" we are still accepting, you know, benchmarks in in the open benchmarks grant. So if you're excited, please reach us at benchmarks.squirrel.ai.

Feel free to reach me directly, and we're really excited to see where the field is going and to again to use benchmarks to not just measure, you know, progress looking backwards, but really shape where where things are going moving forward. Thanks for your time and excited to catch up soon.