

You get an LLM, you
get an LLM, everyone
gets an LLM,
but does it work?

Ashwin Phadke
(this guy evaluates)

Will you speak the truth, the whole truth and
nothing but the truth?



Evaluations



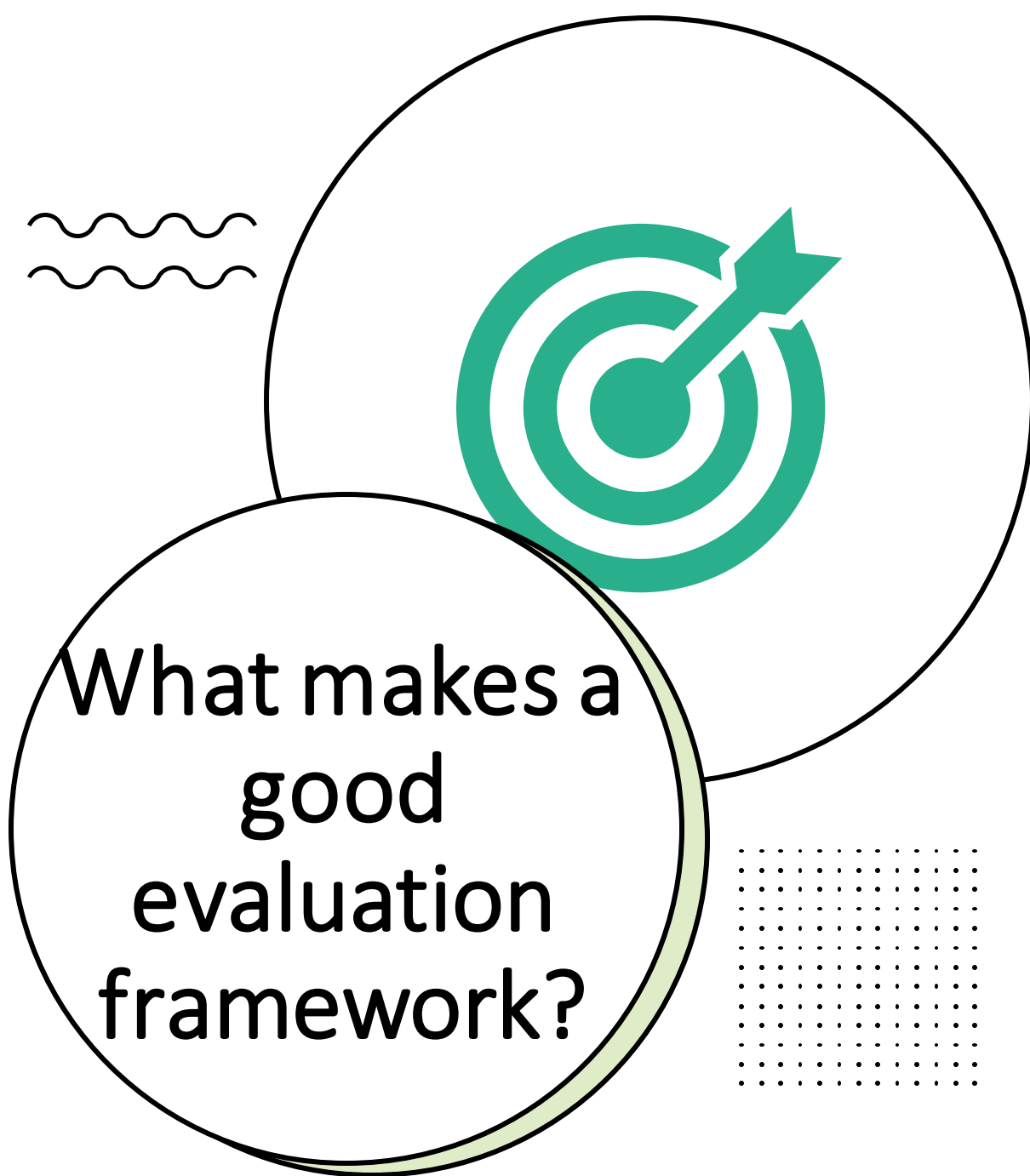
You need to measure it to manage it.



You need to measure it to understand it.



You need to measure it to improve it



What makes a
good
evaluation
framework?

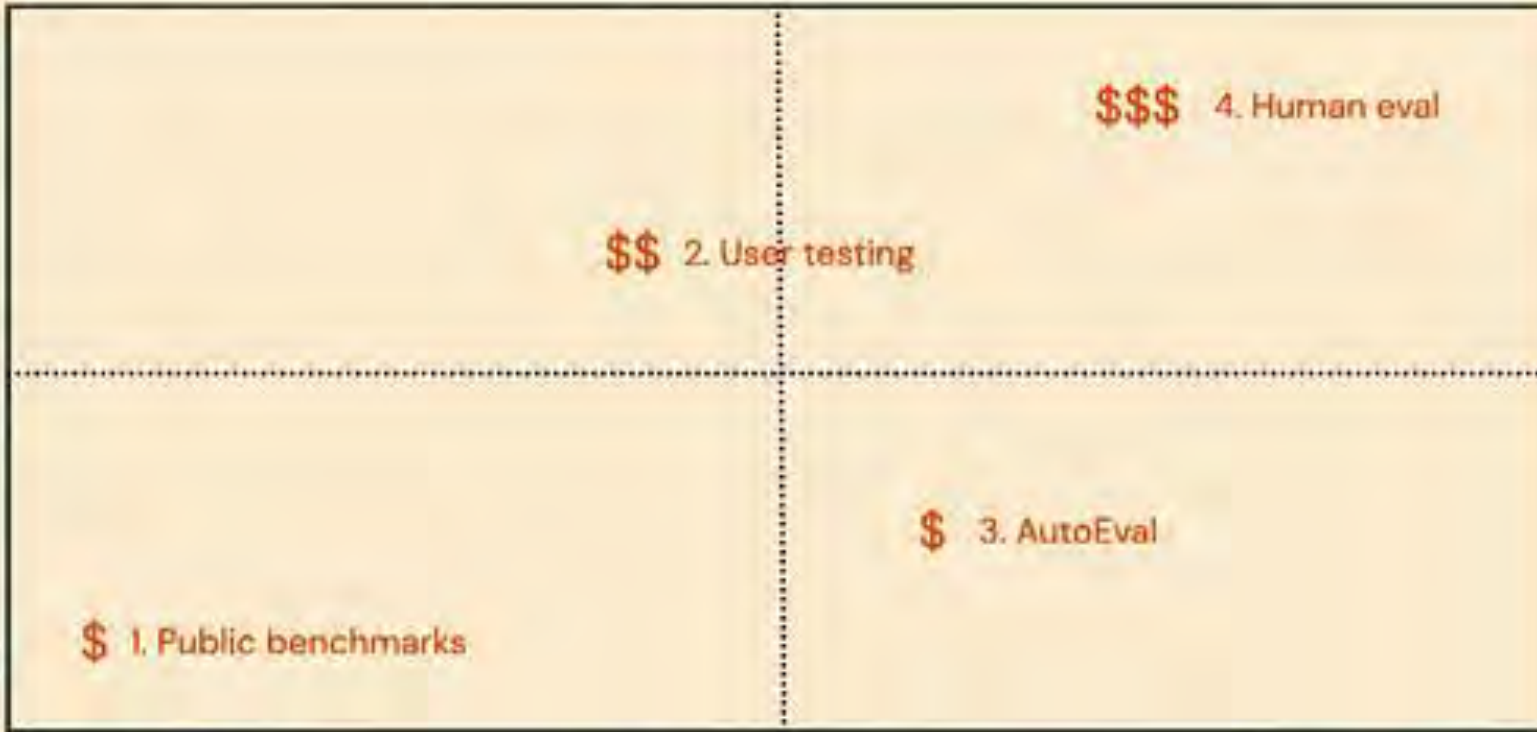
- **Task specific scores designed to measure the right outcome.**
- **A sample list of metrics to validate.**
- **Extensible, fast and easy to maintain**

More correlated
w/ outcomes

Evaluation metrics



Less correlated
w/ outcomes



Less like prod data



Evaluation data

More like prod data

- Human Evaluator

- Feedback on whether an outcome was right or wrong.
- Choices may be different based on geography and task.
- Defining success metrics becomes difficult if based on user choices.

- Auto Evaluator

- What part of the generated output matches the expectation.
- Choice solely depends on known outcomes.
- Well defined traditional and modern metrics work on a reference.

- Public Benchmark:

- Does its job fairly enough and will give you an idea about the generic direction of the model.
- Let's say it works 100 % of the time on a public benchmark, but why not for you?
- Generally task independent and mostly foundational models.
- HeLM, BigBench, SQuAd


- Golden Datasets:

- Your need and use case comes first.
- You'll know if a public model works for you as is or if it needs special care.
- Usually task specific and fine-tuned models. Also highly helpful in RAG flows.
- Semantic similarity, perplexity, rouge

Your use case is likely well defined.

- You know the answers for the following:
 - Questions
 - Summaries
 - Citations and References
 - Supporting document library/vector DB/just a vast collection.
- Scoping the solution to your use case.
 - Fine Tuning/Topic modeling/guardrails.


Someone Convinced a ChatGPT-Powered Chevy Dealer to Sell \$81K Tahoe for Just \$1

Published: 21 Dec 2023, 09:45 UTC • By: [Bogdan Popa](#) 

ChatGPT has taken the world by surprise with its new-generation capabilities, stepping in to help people



Good ol' metrics

- Accuracy, f1, rouge: n-gram matching to your references.
 - Perplexity: Intrinsic, no shockers and from within.
 - BLEU: Fallen out of love, but useful to some extent*
 - QAEval: Content quality of summary (**danieldeutsch**)
 - Ragas: RAG pipeline, faithfulness, relevancy etc.
 - Other LLMs: Could be great, could be not.
 - Human: Matter of choice, relevancy, geography dependent.
- 

LLM evaluates LLM




Metrics evaluate LLM



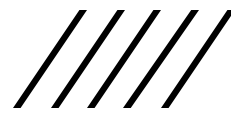


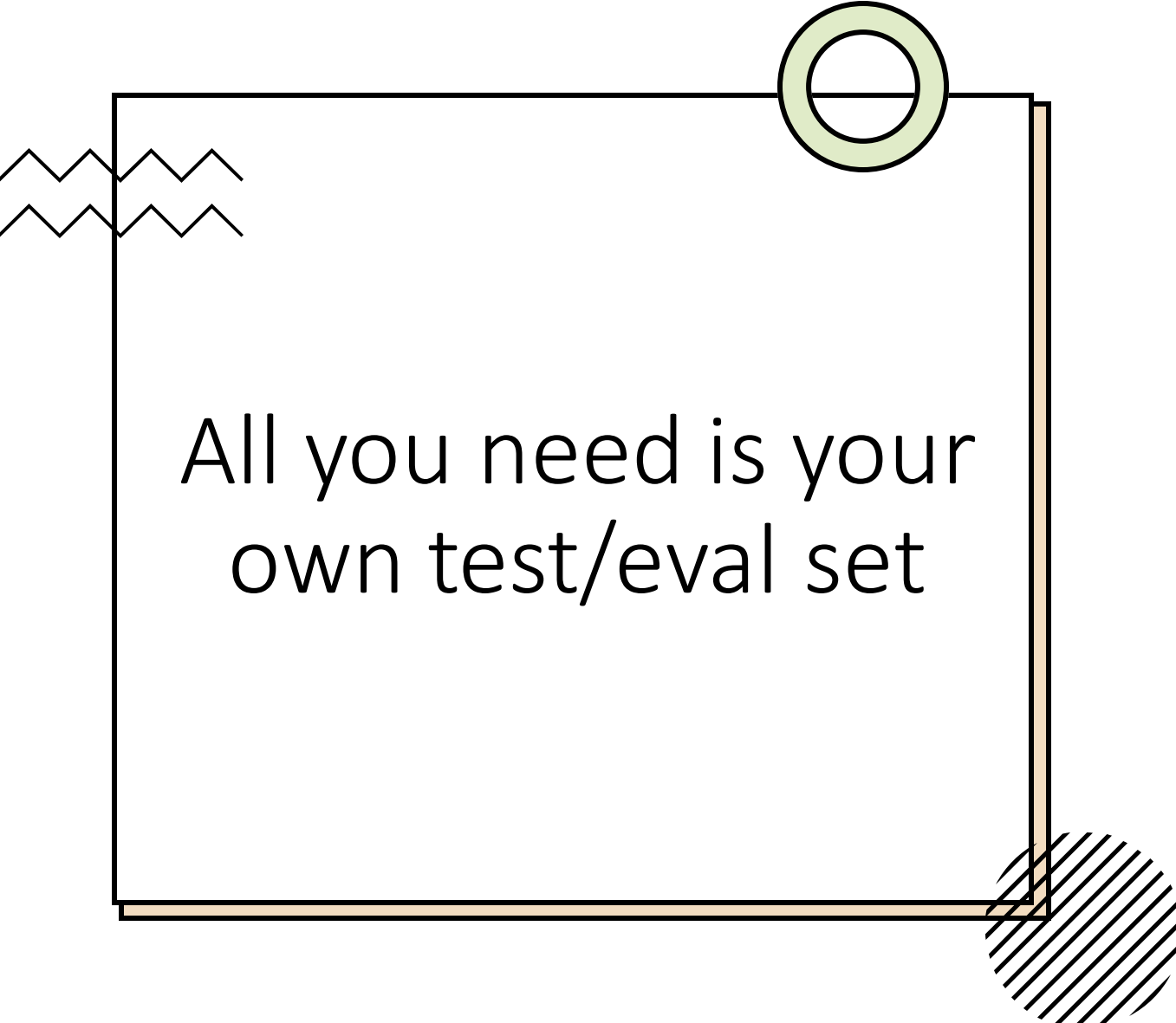
Closing the gap

- Metrics allow you to:
 - Clearly determine if something is working out for you or not.
 - Will help you understand where the problem lies
 - Is it you or the model?
 - Continuously evaluate and improve
 - Fine tune for specific task
 - Chase the "new cool model" rabbit.
 - Your company is happy because graph and math.
 - You know where you fall behind and can iterate over.
 - Get good at the game.
- 



Available frameworks

- Ragas
 - HeLM
 - Lm-evaluation-harness
 - LangChain – LangSmith w/o Weights and Biases
 - OpenAI evals
 - Deepeval
 - Hugging face eval.
- 

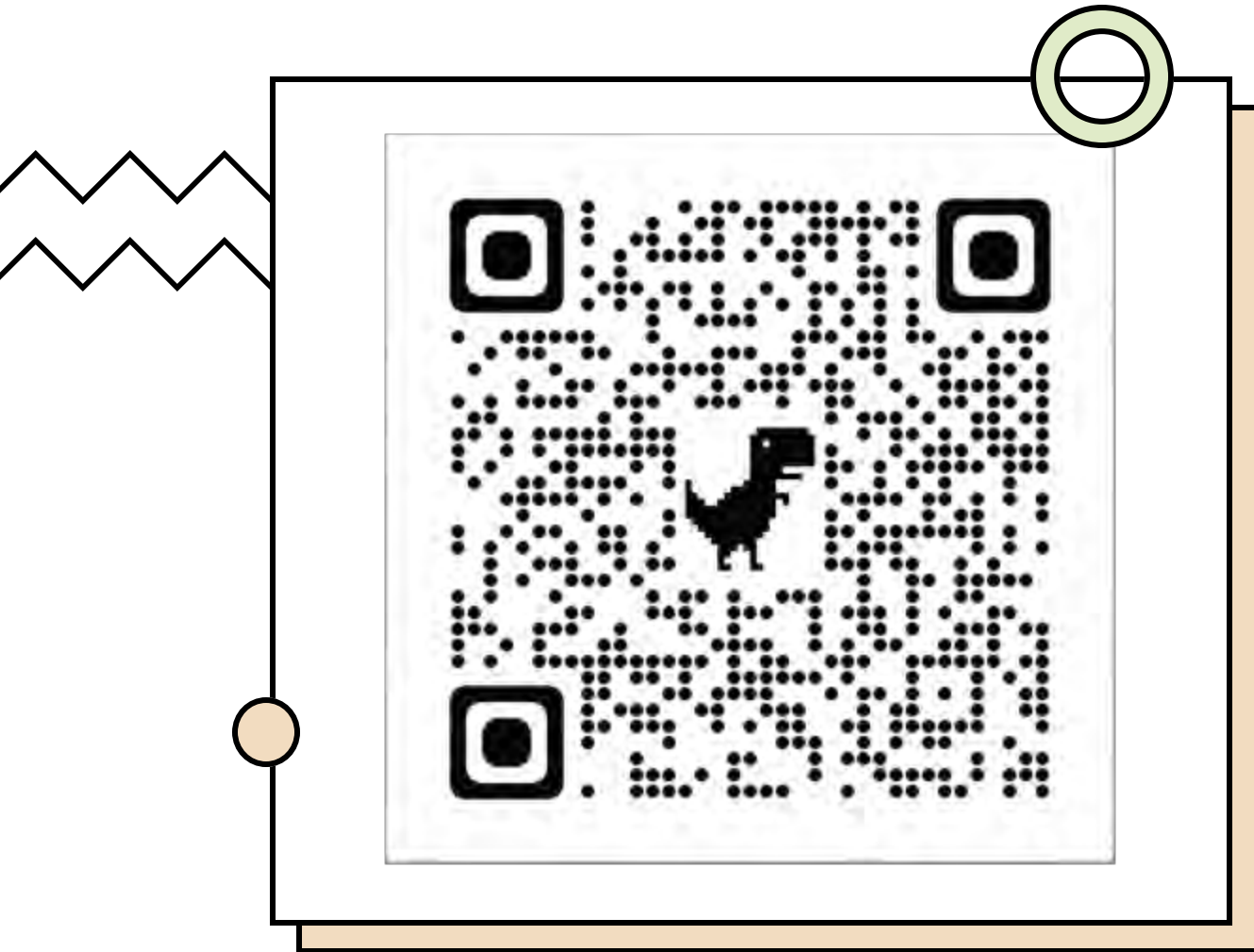


All you need is your
own test/eval set

- Derived from inherent model properties or have already been implemented.
- Implementations are publicly available and well abstracted.
- Well studied, observed and documented for quick overview
- Quantifiable.

And so,

- Get you own dataset and start rolling the carpet.
- Adapt public libraries to work on your dataset.
- Add the human component at the end
- Profit (?)



Thank you!

Ashwin Phadke

Senior Machine Learning Engineer,
Servient Inc.

