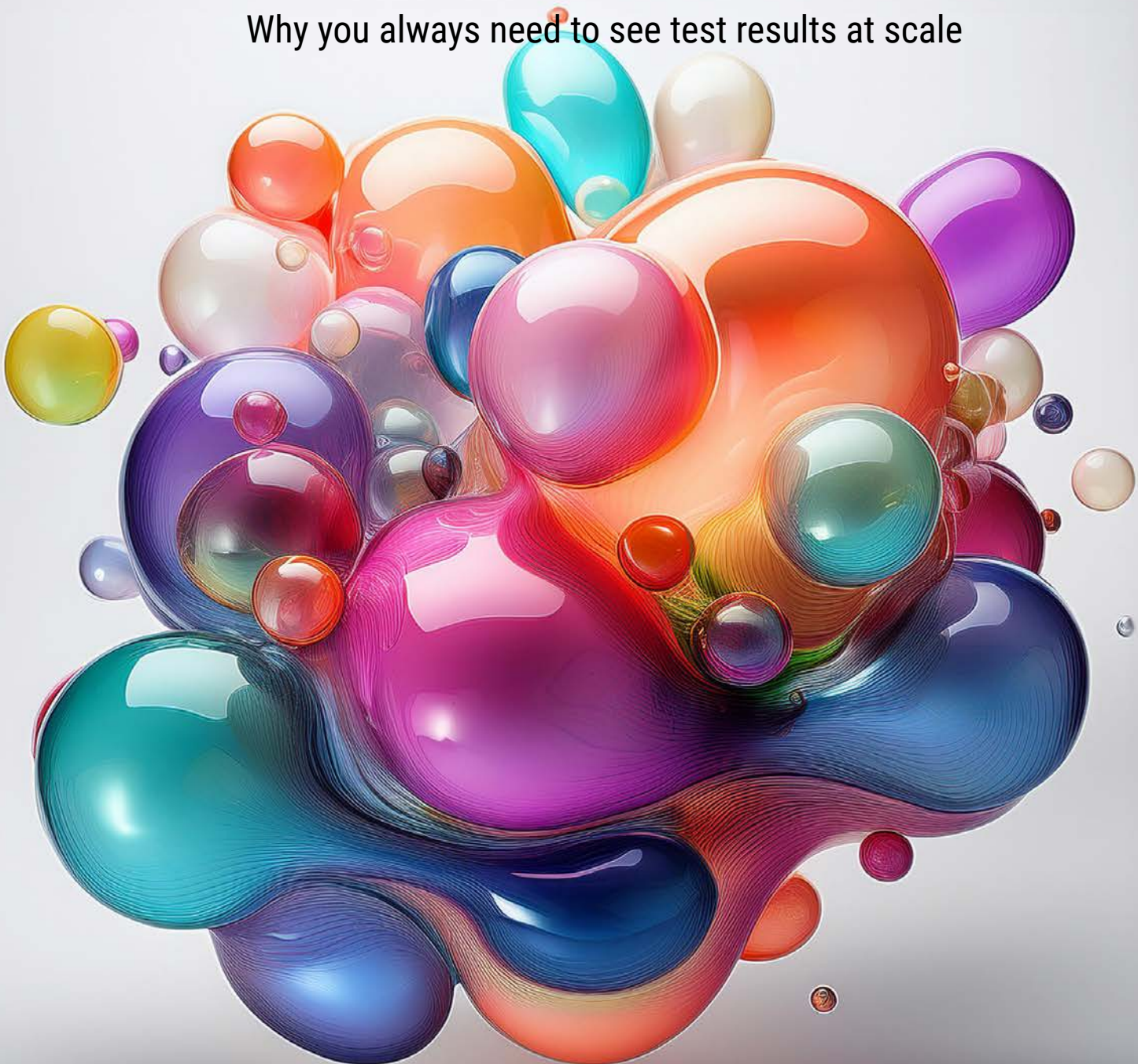


Getting the best from Generative AI

Why you always need to see test results at scale



Non-Executive Directors, C-suite, and other senior leaders are ultimately accountable for how **Artificial Intelligence ('AI')**, and now **Generative AI**, are implemented and procured in organisations. The ability to make informed judgements about these technologies, and what is right for the organisation, has never been more important.

The stakes are high, and whilst there's a lot of talk, there is not much consensus on the basics like sustainable use cases. For every headline about 'intelligent experience engines' and stratospheric ROI, there's another highlighting the difficulty that enterprises face in finding appropriate use cases, navigating legacy tech debt, and thwarting cyber threats – to name but a few! The complexity of the challenge you face is hard to overstate, perhaps only matched by the potential benefits to you, your colleagues and customers.



Glossary of terms used in this document

→ Artificial Intelligence ('AI')

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

→ Generative Artificial Intelligence (GenAI)

Generative artificial intelligence is artificial intelligence capable of generating text, images, videos, or other data using generative models, often in response to prompts. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.

We want to help. And we don't think it's all about presenting yet more 'solutions'. We think the missing link is helping Boards to 'tool up' for evidence-based decision-making for AI.



→ **Model**

An AI model is a program that has been trained on a set of data to recognise certain patterns or make certain decisions without further human intervention. Artificial intelligence models apply different algorithms to relevant data inputs to achieve the tasks, or output, they've been programmed for.

→ **Testing data**

The data used to test the performance of a trained AI model (this should be data that was not used to train the model).

→ **Production**

A live deployment of a solution/ application.

→ **Precision**

Precision is a measure of how many of a model's positive predictions were really a true positive prediction.

→ **Recall**

Recall is a measure of how many of the true positives a model is finding in the data.

→ **False positive**

Where a model makes a positive prediction and that prediction is incorrect.

→ **Misses**


Inputs that are a true positive, but which a model fails to identify as a true positive.

This is Paper #3 of three papers specifically written to support Board members and other leaders in the firm to equip you with the knowledge and confidence to ask the 7 key questions of any AI solution that your organisation is considering procuring and/or implementing:

1. How was the **model tested**?
2. What volume of **testing data** was used?
3. How was the testing data sourced?
4. Is the testing data representative of what is expected in **production** (in real-life use)?
5. What is the **precision** and **recall** of the model(s)?
6. Is there a common reason for **false positives** or **misses**?
7. What alternative approaches were considered/tested? What were their results?

Each paper can be read as stand-alone, or as a series, and in addition to giving you the tools to ask the best questions, will also help you judge whether the answers you are given are sufficient, and where you need to probe further – or not.

Jump to



PAPER NO.1
Models, Metrics & Trade-offs: AI & GenAI Demystified

Read

Jump to



PAPER NO.2
GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?

Read

Jump to...



The essential place of testing in evaluating GenAI



The limits of precision and recall as metrics



The 6 Quality metrics



Knowing 'good performance' when you see it



The limits of a well-performing model: from customer vulnerability to frustration



5 take-outs for your next steps

→ Generative Artificial Intelligence (GenAI)

Generative artificial intelligence is artificial intelligence capable of generating text, images, videos, or other data using generative models, often in response to prompts. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.

It is trivial to prove that a GenAI solution can deliver the desired purpose or result when based on just a few examples.

Jump to



PAPER NO.2

GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?

Read



Testing: forewarned is forearmed

Any AI solution, generative or otherwise, must be tested at scale (not just a few examples that seem to show it working), and furthermore, using human-labelled data. (See [Paper #2: GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?](#))

In today's era of [GenAI](#) especially, it is all too easy to be fooled by the fluency of AI: you could even go so far as to say that it is trivial to prove that a GenAI solution can deliver the desired purpose or result when based on just a few examples.

For Board members and other Execs, especially in regulated sectors like Financial Services, it is vital for you to maintain purposeful neutrality until performance at scale has been analysed and understood. Evidence-based decision-making requires getting beneath the surface of the proposed solution, and also ensuring that there are not better options available to get to the desired outcome. That means testing. It is the only way to truly understand how any AI is likely to perform in the real world.

Testing may not be the most exciting activity, but isn't it better to know how something truly works before you decide to procure and/or deploy it?

Read on for the best questions to ask to help deliver on that evidence-based decision-making for GenAI solutions.

→ **Classifier model**

A classifier model is one designed to classify inputs into certain classes, with those classes defined prior to training. A class could be something such as 'dog' or 'cat', for example.

→ **Natural Language Processing**

Natural language processing (NLP) is a subfield of artificial intelligence (AI) that uses machine learning to enable computers to understand human language.

→ **Precision**

Precision is a measure of how many of a model's positive predictions were really a true positive prediction.

→ **Input**

The data that is provided to the model to analyse.

→ **Prompt**

The input a user provides to a generative AI solution in order to request it to perform a certain action.

* In our experiment testing the relative usefulness of NLP and GenAI models to identify vulnerability, we built a GenAI model to act as a classifier model specifically so we could compare it to the NLP results.



Classifier Models vs GenAI

In the two other whitepapers in this series, we've looked at how to assess models that are attempting to predict the likelihood of an input falling into a given class and for which there is usually a right or wrong answer. In this case, for example with the [Natural Language Processing \(NLP\)](#) model we built to identify vulnerability in customers (See [Paper #2: GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?](#)) we can analyse how many of the model's predictions are likely to be right ([precision](#)), as well as how many of the true positives the model is likely to find ([recall](#)).



The job of [classifier models](#) like these is to find/identify what they have been trained to find/identify. Most GenAI applications are not attempting to classify the [input](#). Instead, they attempt to predict the **most likely** text to follow on from the [prompt](#) they are given – often there **isn't** a singular right answer.*

This means two things for you in putting GenAI models to the test:

1. Precision and recall are usually not relevant when assessing a GenAI application
2. Evaluating a GenAI offering typically requires a broader set of quality metrics i.e. questions to be asked about the output/response generated

Jump to



PAPER NO.2

GenAI and Vulnerability:
how useful is GenAI
in the identification of
potential vulnerability in
consumers?

Read








→ **Guardrails**

In the era of GenAI, this typically refers to a set of constraints written by a user in the prompt in an attempt to prevent undesired outputs from occurring (e.g. 'do not swear in your response to the customer'). However, there is no guarantee that guardrails will be followed every time – they guide the output, but don't control it.

A greater degree of judgement is needed particularly in assessing the inevitable risk/reward trade-offs.




Assessing GenAI: 6 quality metrics

 <p>Factual accuracy</p> <p>Are the facts in the output/response correct?</p>	 <p>Factual fabrication</p> <p>Are there facts in the output/response that are made up?</p>	 <p>Factual completeness</p> <p>Does the output/response contain all the facts that it should?</p>
 <p>Response sensibility</p> <p>Does the output/response make sense?</p>	 <p>Response usefulness</p> <p>Is the output/response useful based on what was asked?</p>	 <p>Prompt alignment</p> <p>Does the output/response conform to the prompt, e.g. follow  guardrails?</p>

You can see that applying these 6 quality metrics is far more subjective than even the trade-offs we looked at in [Paper #1: Models, Metrics & Trade-offs: AI & GenAI demystified](#). Will you use all 6, and/or a combination? Which of the 6 is most important? Or are all 6 equally vital to your deciding to procure and/or have it put into production in the business?

The answer is that this will largely depend on the operating context for the GenAI application. Here's where a greater degree of judgement is needed particularly in assessing the inevitable risk/reward trade-offs and how that aligns with your business' priorities and risk appetite. It's more complicated for you, but, regardless: you must test at scale to take a view on how 'good' the model's performance is.

Jump to



PAPER NO.1
Models, Metrics & Trade-offs: AI & GenAI Demystified

Read



IN THE KNOW

What's 'good performance' for GenAI?

→ **Model bias**

The occurrence of biased results due to human biases that skew the original training data or AI algorithm used, leading to distorted outputs and potentially harmful outcomes.

→ **Gap in training data**

A gap in training data occurs where the training and test data for a model does not fully reflect the real world scenario in which it is to be deployed.

Unfortunately, there is no hard and fast rule in terms of what 'good' performance is. The answer is always: 'it depends'. It depends on the purpose of the model, the context in which it is to be used, the regulatory and legal environment, and, because **no AI model is 100% reliable**, it depends on what is considered an **acceptable risk/reward trade-off for your organisation**.

Rather than thinking about what absolute 'good' performance is, it's more useful to think about instead: 'what is 'acceptable' performance, given our specific risk/reward trade-off appetite?' You can only make that risk/reward judgement armed with the results of testing. Too often the focus is only on the benefit/reward, which can lead to unintended and unexpected consequences.

Our top tip is to **always ask if there is any pattern in the mistakes or misses a model makes, no matter how few there may be**. Ideally, the mistakes and misses would be random, but sometimes this can be where you start to see **model bias** emerging, or a **gap in training data** being exposed. Both of these could be important considerations for taking a view on performance.

→ **Goal state**

The set of conditions that a solution has been designed to meet.



Why you need to test each time, even if the task appears the same

Let's say you see a GenAI model that delivers good performance – according to your requirements and needs – for the specific purpose (**goal state**). It would be natural to want to extend the model's use. As we shared in [Paper #2: GenAI and Vulnerability](#) we developed a GenAI model designed to act as a classifier, and we saw remarkable results in identifying vulnerable customers during a conversation. Would it be a safe assumption that because the GenAI model was good at the task of identifying vulnerable customers, it would be as good a solution for identifying other types of customer issues without the need for further testing? We decided to find out.

Both the same NLP and GenAI (initial and final iteration) models were used as before but tasked with identifying **frustration** in customer conversations, rather than vulnerability.

Jump to

**PAPER NO.2**

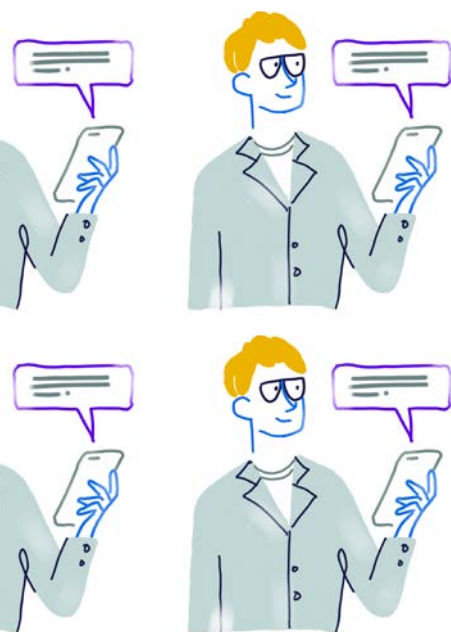
GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?

Read



What did we find?

As you can see, the GenAI solution **did not** achieve the same level of performance as it did for identifying vulnerability, even though the high-level task (classification of text) it was being asked to perform is the same.



Performance measure	NLP	GenAI Initial iteration	GenAI Final iteration
Precision	86%	42%	48%
Recall	58%	92%	88%

If this GenAI solution had been put into production without assessment against a specific test set of data, it would have produced an unacceptably high level of false positives. Even in the GenAI final iteration, over 50% of predictions would have been incorrectly identified as frustration: would this persuade you that this model should be put into production with your customers?

IN THE KNOW Intuition about what GenAI will be good at and where it may struggle

Independent academic research helps us to better understand the strengths and limitations of GenAI. One particularly useful piece of research was published by a team from Princeton University in September 2023, which helps us to build some intuition for where GenAI applications might perform well and where they might struggle.

A simplified version of their research findings is this general rule:

If you're likely to see examples of the problem you are tasking the GenAI application with solving on the web

(i.e. what you are asking the application to do is likely something that is on the web and use information likely to exist on the web), then it is likely to perform well. However, if the problem is not likely something on the web, or likely to involve information not available on the web, then it is likely to do less well.

This has important implications for use cases that are using proprietary information, or that are unlikely to exist on the web such as highly-specific corporate language or confidential customer data.

[Read the full report](#)





5 take-outs for your next steps

There are no shortcuts. GenAI models must be tested at scale and using human-labelled data for each individual task.

We hope this allows you to get beneath the surface of a proposed GenAI solution, equipped with the critical questions and the understanding to interpret the answers. Here’s our 5-point summary:

1. Evaluating a GenAI offering requires your judgement based on a deep understanding of the firm’s risk and reward priorities: you’ll regularly be faced with clarifying and weighing up trade-offs that bring huge impacts for your firm and your customers.
2. There are no shortcuts: GenAI models must be tested, at scale, and using human-labelled data, for each individual task.
3. There is no avoiding human labelling costs, even with GenAI.
4. Precision and recall are the established metrics for evaluating most AI models but are less relevant when assessing a GenAI application.
5. Make the 6 Quality Metrics for GenAI your ‘go-to’.

Jump to

PAPER NO.1
Models, Metrics & Trade-offs: AI & GenAI Demystified

[Read](#)

Jump to




PAPER NO.2
GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?

[Read](#)



About ContactEngine Research Group

ContactEngine Research Group is a specialist ContactEngine team made up of diverse experts drawn from academia, Deep Tech, applied AI and corporate innovation. Our focus is to dig out the sustainable Value from Conversational AI and Emerging Tech like GenAI – for clients, and for ContactEngine itself. Led by Director of AI Strategy, Euan Matthews, the team delve deep at the cutting edge, following wherever that leads, designing experiments and applying relevant learnings (including failures!) to CX experts ContactEngine's existing services. In today's rapidly-evolving Tech environment, the team also ensures that ContactEngine's in-house knowledge is where it needs to be. If you're asking questions of your own CX set-up, wanting fresh options or just an informal conversation on where the Value in GenAI really lies, get in touch with Euan for an informal discussion.

For more information, visit contactengine.com

Registered Office:
Nice Systems UK Ltd
Tollbar House, Tollbar Way, Hedge End,
Southampton, Hampshire, SO30 2ZP



Author of this paper

Euan Matthews

LinkedIn

linkedin.com/in/euan-matthews-mba

Email

euan.matthews@contactengine.com