

# Models, Metrics & Trade-offs: AI & GenAI Demystified



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

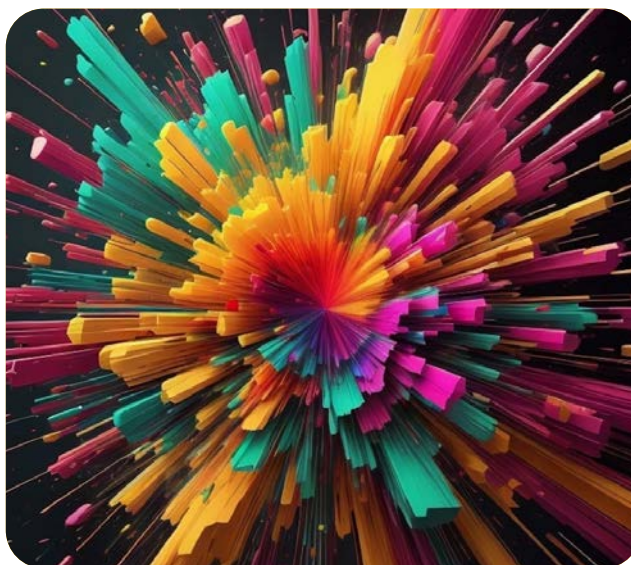
##### → Artificial Intelligence ('AI')

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 #1 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.2**  
GenAI and Vulnerability: how useful is GenAI in the identification of potential vulnerability in consumers?

Read

Jump to



**PAPER NO.3**  
Getting the best from GenAI: why you always need to see test results at scale

Read

# Jump to...



Why AI models are not like software development



The basics of GenAI



Precision and Recall: the key metrics for AI models and how to use them



Scale and trade-offs: the key requirements for choosing models



Other options to get the job done?



**→ Prompt**

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

**All AI models are based on probabilities: in essence, trying to predict the most probable output based on the input given.**



# AI Models: it's not like software development

It's all too easy to be impressed by a demo showing off the magic of AI but understanding how the model has been tested is paramount. Unlike software development, an AI model cannot be deemed to be working just because it demonstrates a repeatable outcome on a few test examples or scenarios – a different approach is needed.

Why? The reason for this is that all AI models are based on probabilities: in essence, trying to predict the most probable output based on the input given, for instance, predicting what is in an image, predicting what the meaning of a collection of words is, or predicting what is the most probable content to follow on from the **prompt** a user has written for ChatGPT or similar GenAI product.

There's a truism dating back to 1950s computer science that holds good today: GIGO or garbage in, garbage out. Flawed, biased or poor quality ("garbage") data, information or input invariably produces flawed, biased or poor-quality outputs. AI models are no different. This means that testing needs to represent the fullest possible variability and randomness of input to which a model will be exposed when in production. Even 100 test examples won't do – we need to be thinking more in the 1000s.

**IN THE KNOW**

**Generative AI (GenAI) for Text**

Generative AI applications like ChatGPT *appear* to have knowledge, even intelligence, when they generate a response to the prompt you give them. It is hard not to be impressed by the capability of such models, and they are indeed remarkable.

However, it's important to understand that what they are fundamentally doing is **predicting the most likely set of words to follow the user's prompt** (this being the input a user provides to prompt ChatGPT or similar product to do something).

ChatGPT and other GenAI applications have no actual knowledge or understanding of what the text they generate means, only that it is the most probable extension of the prompt written by the user. The reason that these models can be so accurate, fluent and plausible is due to the volume of data the **Large Language Model ('LLM')** that underpins the application has been trained on.

The volume of **training data** needed to build an LLM is vast, with a significant proportion of that data coming from what is available on the web. For context, it is estimated that the training data for OpenAI's GPT-4 comprised about 10 trillion words.

It's not really a surprise, therefore, that GenAI is good at things such as:

- Passing common tests such as the American bar exam and SATs (as most of those common tests are featured on the web in the form of tutorial websites, mock exam question/answer examples, past papers, etc.).
- Generating code (there are thousands of question/answer forums where people ask how best to write a code to perform a certain task).
- Writing the Chair's letter in an Annual Report and Accounts (there are thousands of annual reports available on the web).

But remember: GenAI applications such as ChatGPT do not 'know' the answer to the question, there is no reason or intelligence involved; it is simply predicting the most

probable words to follow on from the user's input (usually a question of some sort), with that probability informed by the enormous quantity of examples on the web, from where most of the training data for an LLM is sourced.

To bring this point to life, consider the below examples from Princeton University research:

**Linear functions: Task probability**

**Common function:**  $f(x) = (9/5)x + 32$

Input: 64	Input: 577
Correct: 147.2	Correct: 1070.6
✓ GPT-4: 147.2	✗ GPT-4: 1069.6

**Rare function:**  $f(x) = (7/5)x + 31$

Input: 64	Input: 577
Correct: 120.6	Correct: 838.8
✗ GPT-4: 89.8	✗ GPT-4: 805.4

**Acronyms: Task probability**

**Common task: First-letter acronym.** Combine the first letters of the words in the sequence.

Input: penance aplenty rooster trample impasse envious subtext
Correct: PARTIES
✓ GPT-4: PARTIES

Input: unbound newness cranium likable emerald abalone reissue

Correct: UNCLEAR
✗ GPT-4: UNCLEBA

**Uncommon task: Second-letter acronym.** Combine the second letters of the words in the sequence.

Input: aplenty maestro precept strayed figment negaton ascetic
Correct: PARTIES
✗ GPT-4: PLEGGET

Input: quattro ennoble scissor fluency regency hawkish pricked

Correct: UNCLEAR
✗ GPT-4: UNCLEARW

As you can see, when the task at hand is a common one, GPT-4 provides the correct answer, but, when the task is an uncommon one, GPT-4 does not provide the correct answer.

→ **Large Language Model (LLM)**

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.


→ **Trained**


A model is considered as trained once the iterative process of providing training data to a model, testing results using test data, and then adjusting/re-training the model is complete



# Precision and Recall


the key metrics for evaluating non-Generative AI Models


Jump to 



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

[Read](#)

Jump to 



**PAPER NO.3**  
Getting the best from GenAI: why you always need to see test results at scale

[Read](#)

Precision and Recall are the key concepts for evaluating AI models that are built to classify an input, or make a prediction based on an input.

Here's a way to think about these critical metrics:

- You want to evaluate an AI model built to identify images that contain dogs.
- You have 200 images of animals new to the model of which you know that 100 contain dogs.
- You ask the model to predict which of your 200 images contain dogs.



The model predicts that 90 images contain dogs: these are called 'positive predictions'. However, on review, you see that only 80 of those 90 images in fact contain dogs. The model mistakenly identified 10 images. This means the model's **precision** is 80/90, and as a percentage **88.9%**.

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

- What about **recall**? The model identified (or 'recalled') 80 true positive images of dogs out of the 100 that were available to find.



The **recall** of the model is therefore 80/100, and as a percentage **80%**.

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

→ **Positive prediction**

Where a model has positively predicted an input as being something it has been designed/trained to identify.

**When you attempt to increase model recall, you will typically notice a drop in model precision.**

In an ideal world, we would achieve 100% precision and recall, but this is not realistic. **Be highly-sceptical** of any model that claims 100% on either, or both, of these key metrics.

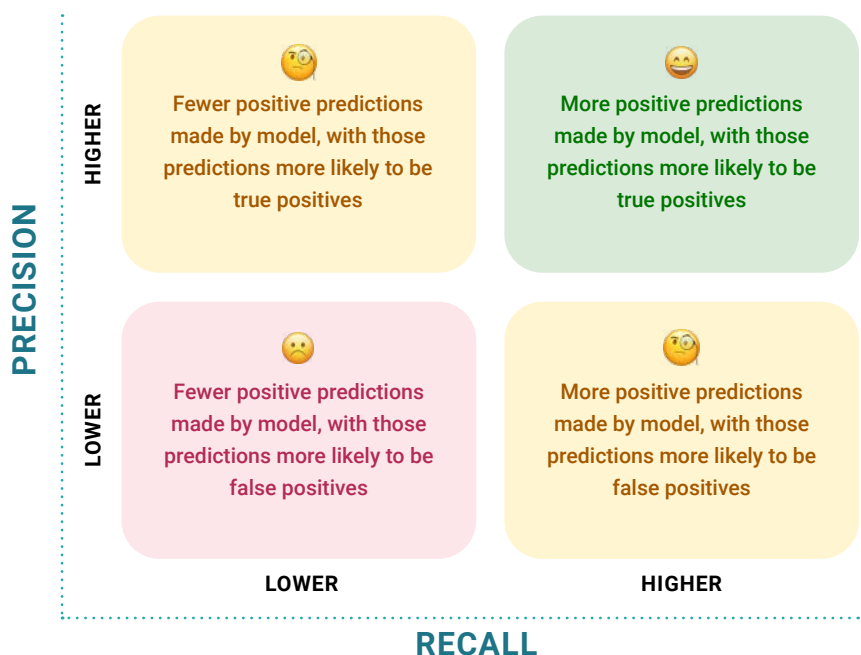
The reality is that there is always a trade-off between precision and recall. For example, if you want to improve recall, then you are effectively asking the model to make more positive predictions. However, with more positive predictions comes more risk of increased false positive predictions i.e. the model being more wrong.

**When you attempt to increase model recall, you will typically notice a drop in model precision and vice versa.**



# Trade offs: get used to them

## Understanding Trade-offs between Precision and Recall in AI Models





**The Board and executive leaders need to be very clear as to what trade-offs are acceptable, and why, given the ultimate impacts of the model's implementation.**

Trade-offs are a fact of life with AI models. Models with lower precision but higher recall are not necessarily bad models. Similarly, models with higher precision but lower recall are not necessarily bad models either. It all depends on what the purpose of the model is and how it will be used. **This is where the Board and executive leaders need to be very clear as to what trade-offs are acceptable, and why, given the ultimate impacts of the model's implementation.**

What kind of trade-offs could you be faced with? In domains where high recall is important, such as financial fraud detection, or medical image diagnostics, you may tolerate lower precision, as it is more important to identify as many *potential* true positives as possible. The fact that there will be a human-in-the-loop to review and assess the positive predictions balances out the lower precision of the model: although you will experience more false positives, you are reducing the possibility of missing something important to your purpose by casting a bigger 'net' (recall). And the expert human will be the ultimate arbiter.

However, in instances where your purpose means that the process is going to be automated and there is no human-in-the-loop, such as conversation automation or transaction automation, precision is likely to be far more important, and you would rather miss potential true positive predictions if it means ensuring that positive predictions are less likely to be false positives.



This is one of the (many) reasons why driverless (autonomous) vehicles are so challenging. The stakes are very high in terms of accidents and safety, but you also can't have lots of false positive predictions of accidents or safety risks that would cause the vehicle to incorrectly come to a halt. In these situations, both high precision and high recall are needed, which is very challenging indeed.



→ **Statistical significance**

Refers to a level of statistical certainty that a result or observation from data is unlikely to be down to chance.

→ **Synthetic data**

Synthetic data is information that is artificially generated rather than produced by real-world events.

→ **Intent**

Typically used to describe the underlying intent of a customer's words during a voice/text interaction, e.g. 'I would like to pay my bill please' might be considered a 'Pay' intent.



# Scale matters!

Precision and recall metrics are vital, and they only become meaningful at a reasonable scale. What's reasonable? We need to test at scales of **statistical significance** to confidently assess how a model is performing, and that testing should use data that is representative of the variety and distribution of input the model is likely to see in production. If it's not, the test set will almost certainly give you a false sense of security that the model is working. In our experience, precision and recall measures are barely robust on even a test set of 200 examples – we typically use test sets of well over 1,000 examples.



## IN THE KNOW

### Why training and test data needs to represent production

It's possible to train an NLP model with only a few examples, and you may well create those examples yourself. This is also known as **'synthetic data'**. Whilst this can result in a model that appears to work, when these models are put into production, their performance can very often be quite different to what the testing suggested it would be. This is because one individual, or even a group of individuals, will struggle to represent the randomness and variation that a much larger group will in the language they use to express the same **intent**. Think of this as comparing a handful of staff interacting with a chatbot versus your entire customer base.

The same principle applies to any AI model - the training data must be representative of what it will receive as input in real-world use.



**It turned out the model was placing most emphasis on the presence of a ruler in the image to predict the presence of cancer, rather than the image of the tumour itself.**



# Final word: ask about patterns

No matter the performance of a model, you must always ask if there is any pattern in the mistakes or misses a model makes, no matter how few there may be.

To help bring this point to life, consider the following well-reported example:

A model was trained to identify likely cancerous tumours from still images taken during a patient scan. Of course, training this model required that it be provided examples of images that contained cancerous tumours, as well as ones that did not.



During testing, the model appeared to work very well, but investigation of the mistakes and misses revealed that all the images the model predicted to include cancerous tumours also contained an image of a ruler, whereas the images with cancerous tumours that it failed to identify did not. Why? Well, most of the actual images of cancerous tumours used for training purposes also had a ruler beside them to measure the size of the tumour.

It turned out the model was placing most emphasis on the presence of a ruler in the image to predict the presence of cancer, rather than the image of the tumour itself. The model had to be re-trained using images without any ruler present.

Ideally the mistakes and misses of a model would be random, but only by reviewing these can you start to see model bias emerging, or a gap in training data being exposed, which is an important consideration to factor in when making your judgement. There is a lot more detail to go into here, but for now know that it's important to ask the question – were any patterns identified in the model mistakes and misses?

Jump to

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

[Read](#)

Jump to

**PAPER NO.3**  
Getting the best from GenAI: why you always need to see test results at scale

[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](https://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](https://linkedin.com/in/euan-matthews-mba)

#### **Email**

[euan.matthews@contactengine.com](mailto:euan.matthews@contactengine.com)