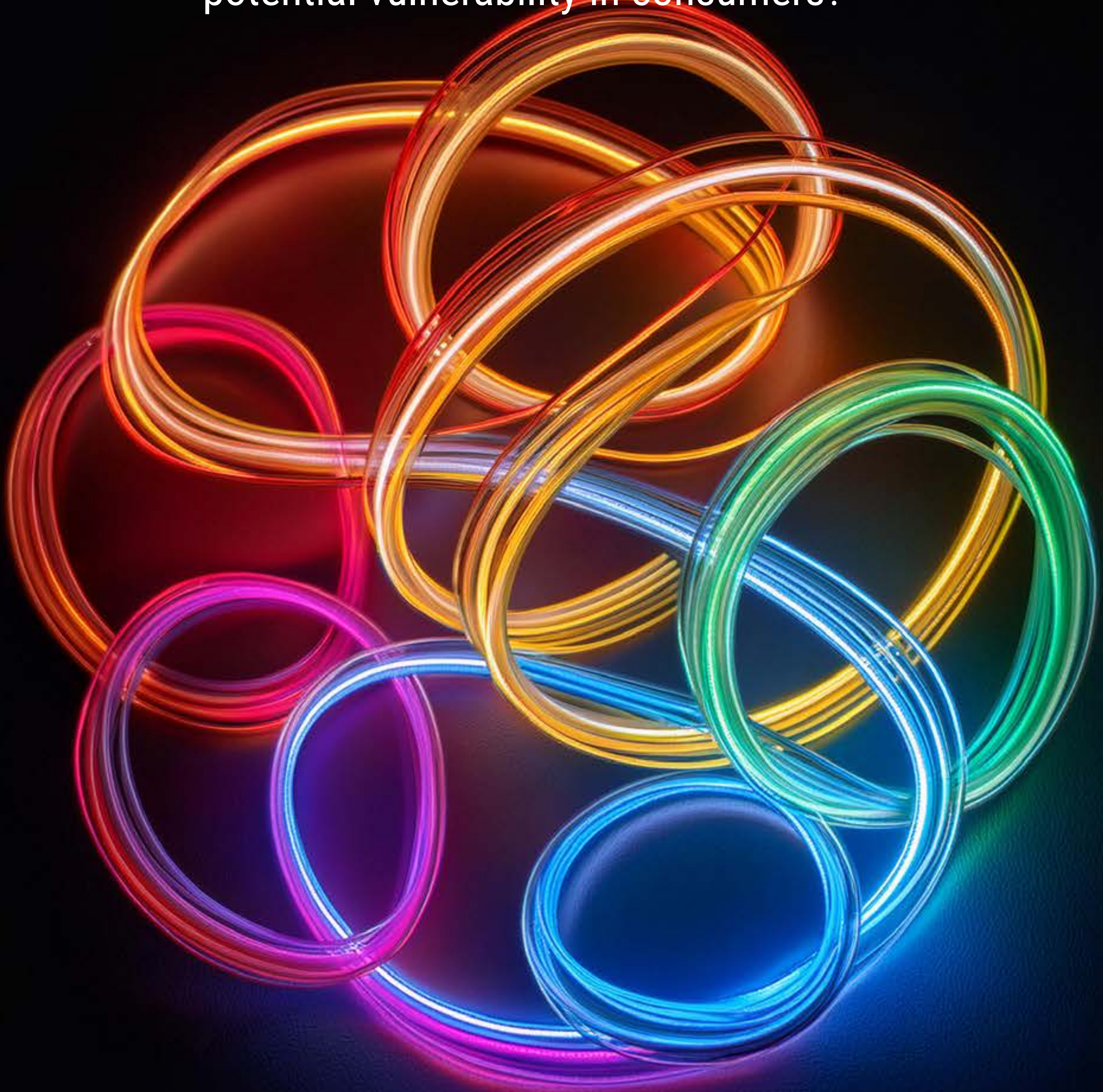
ContactEngine
A NICE company

The Fundamentals of AI for the Board series 2

GenAI and Vulnerability

How useful is GenAI in the identification of potential vulnerability in consumers?



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.

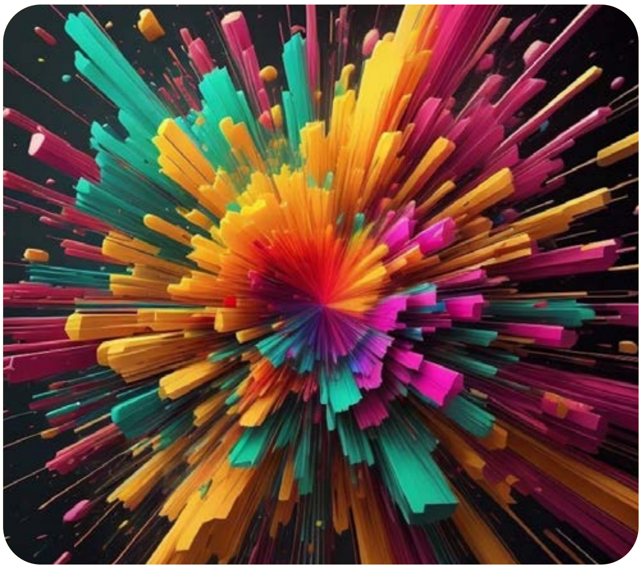
The stakes are high, and whilst there's a lot of talk, there is not much consensus on the basics like sustainable use cases. For every headline about 'intelligent experience engines' and stratospheric ROI, there's another highlighting the difficulty that enterprises face in finding appropriate use cases, 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 #2 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.3
Getting the best from GenAI: why you always need to see test results at scale

Read

Jump to...

How we built and tested an NLP model for identifying vulnerability

How we built and tested a GenAI model for identifying vulnerability

How we compared them – and what we think this means for you

Explaining the differences between NLP and GenAI models

Training, testing, human annotation, prompts, and more

→ **Vulnerability**
Specific definition can vary by regulator, but broadly a vulnerable customer might inlude those with physical or mental health problems, specific characteristics such as age or literacy skills, or changes in personal circumstances such as bereavement, job loss or changes in household income.

→ **Omnichannel**
Use of all available channels for engaging a customer, usually targetting a seamless, consistent experience across those channels.

Vulnerability: a Critical Issue for Financial Services (FS)

The complexity of identifying vulnerable customers is not to be underestimated.

Identifying potentially vulnerable customers is a vital capability as it enables FS firms (and other regulated sectors like Utilities and Telecommunications) companies to provide customers with the help and support they need, and which is required under regulations such as Consumer Duty in the UK.

The best firms regard these interactions as 'moments of truth' for their organisations, a key proof of service quality and positive customer impact. Significant effort is put into training, technology and culture engineering to ensure that in human-human customer interactions – such as face-to-face in a branch, or over a voice call – a human expert colleague will successfully be able to spot the signs of ↖**vulnerability** and take the appropriate actions.

The complexity of this 'ask' is not to be underestimated. A frontline advisor can be faced with a vast array of inquiries, in an ↖**omnichannel** world, to which they are expected to have the answers – at speed, and in high volumes. It is high stakes and often high stress, leading to vulnerability issues for colleagues themselves, and high levels of staffing churn. Customer vulnerability is not even 'just' a core regulatory and operational priority, it is also increasingly pivotal to talent acquisition and retention in critical service areas for the business.

→ **Goal state**
The set of conditions that a solution has been designed to meet.

→ **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.

→ **Human-annotated data**
This is data where a human has assigned meaning to each data point, e.g. assigning a tag of 'dog' to an image of a dog.

→ **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.

→ **Zero-shot learning**
Zero-shot learning (ZSL) is a machine learning scenario in which an AI model is trained to recognise and categorise objects or concepts without having seen any examples of those categories or concepts beforehand.

It is a fact of life that in human-computer interactions, where the purpose (↖ goal state) is typically to maximise automation of the interaction, there is an increased risk that vulnerable customers will not be identified. So, for the very specific and business critical use case of Vulnerability, where can Technology, including GenAI, help?

In 2023, ContactEngine Research Group developed and deployed a ↖ **Natural Language Processing** (‘NLP’) model targeted at identifying vulnerable customers during an automated conversation. When the model identifies potential vulnerability, the customer can be escalated to an expert human for review so that they can ensure the right help/ support can be provided.

This NLP model works very well, but the downside of these types of models is that they require significant volumes of ↖ **human-annotated** training data to achieve performance that is suitable for production. What does this mean? A training and test set large enough to represent the randomness and variability that it would receive in a real-world deployment.

However, the recent advances in ↖ **Large Language Models (LLMs)** and their ↖ **zero-shot learning** capability could potentially bypass the need for human-annotated training data whilst still achieving ‘good enough’ performance.

ContactEngine Research Group conducted an experiment to compare the performance and efficacy of a GenAI-based solution to a trained NLP model. What did we find?

→ **Unsupervised generative capacity**

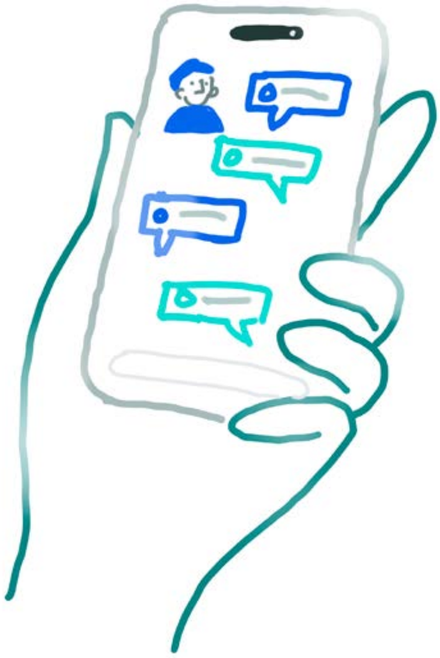
Unsupervised models are ones where AI computationally determines associations in data without any human input or use of human-annotated data - large language models fall into this category of model. Generative means a model that is designed to generate content. Put these two together and you have a model that generates content for a user, but with that generation using a model trained without human input.

→ **Supervised learning system**

This is where the model is trained using human-annotated data only.

→ **Utterance**

The words a customer uses for a single input during a voice/text interaction.



The NLP approach to identifying vulnerable customers

The NLP model that ContactEngine Research Group developed utilises an LLM, but not in an ↖**unsupervised generative capacity**. Instead, the LLM is utilised as part of a ↖**supervised learning system**: the model is fine-tuned by expert humans to perform a specific task closely defined by those same expert humans.

We trained our supervised model using real-world examples of verbatim customer responses (↖‘**utterances**’) that expert human annotators had labelled as either ‘Vulnerable’ or ‘not Vulnerable’. This is why it is called a supervised model: the association between the verbatim and its meaning is assigned by a human. The context for those verbatims was telecommunications-related conversations, with an illustrative example of the data shown below.

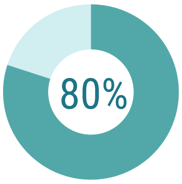
Customer verbatim	Expert Human labelling
That’s fine thank you	Not vulnerable
Yes, that’s OK but I am close to having a breakdown over the debts that I owe	Vulnerable
I’m elderly and can’t afford to pay for my heating	Vulnerable
I don’t need that any more	Not vulnerable
No thanks	Not vulnerable
That’s exactly what I wanted	Not vulnerable

→ **Clustering algorithm**
A type of AI algorithm that identfies patterns in data and groups data that exhibits a similar pattern together in 'clusters'. Common algorithms include K-Means Clustering and HDBSCAN.

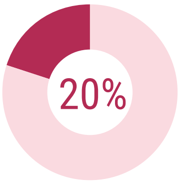
→ **Training set**
The data used to train an AI model.

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

This annotation process is necessarily time-consuming as all verbatims must be sifted through by human annotators so that all those exhibiting vulnerable characteristics are identified and labelled. We do use ↖**clustering algorithms** to identify groups of semantically similar utterances to speed up the labelling process where possible.



Typical percentage of data for **Training sets**



Typical percentage of data for **Test sets**

.

Labelling is done until a critical mass of human-annotated data is reached so that we can divide it into a ↖**training set** (typically 80% of the data, randomly selected) and a ↖**test set** (20% of the data, randomly selected).

You need sufficient volumes of data in both sets. ContactEngine Research Group defines ‘sufficient’ as 500 labelled utterances per intent. We could use much fewer, but because we are dealing with high volumes of important customer conversations, we need to attain higher levels of performance than your typical chatbot.

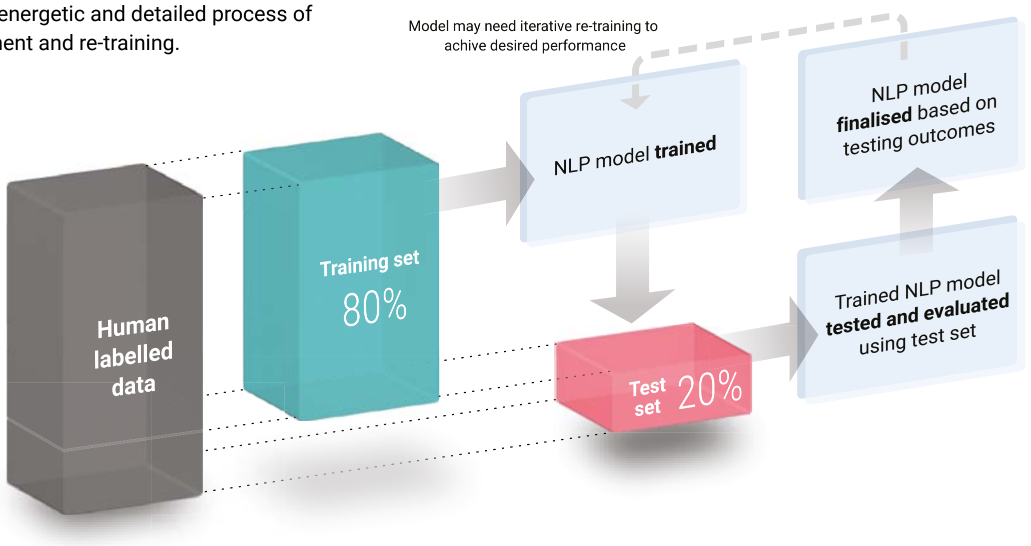
IN THE KNOW

Basics on training & test sets

As their names suggest, the **training set** is what is used to train the model and the **test set** is used to evaluate how well the model performs.

The test set should **never** be included in the training process: this would fatally flaw the whole process as it would be akin to showing a student the answers to their exam before it starts. We want the model to actively predict a vulnerable intent/meaning, **and** actively predict a **non**-vulnerable intent/meaning, so the model is trained on verbatims that **don’t** exhibit vulnerability, as well as those that **do**.

Here's a much-simplified illustration of the NLP model development process. The reality is that this is always an energetic and detailed process of tweaking, refinement and re-training.


Model may need iterative re-training to achive desired performance
Human labelled data
Training set 80%
Test set 20%
NLP model trained
Trained NLP model tested and evaluated using test set
NLP model finalised based on testing outcomes

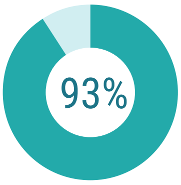
→ **Precision**
Precision is a measure of how many of a model's positive predictions were really a true positive prediction.

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

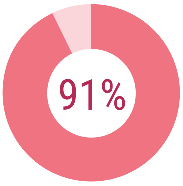
What did we find?

We used a sample of 10,871 verbatims from real conversations to train and test a model to identify a customer verbatim exhibiting some form of vulnerability during a telecommunications-related conversation. The expert human annotators labelled 10% of the verbatims as exhibiting vulnerability and the remaining 90% as not exhibiting vulnerability.

The results



↖ **Precision**
93% of the model's predictions of vulnerability were correct i.e. matched the human annotators' labelling



↖ **Recall**
The model found 91% of all the verbatims that a human had assigned as vulnerable

The verdict

- **This is a well-performing NLP model suitable for real-world use.**
- **It provides a good baseline against which to compare the GenAI-based approach.**

→ **Zero-shot learning**
Zero-shot learning (ZSL) is a machine learning scenario in which an AI model is trained to recognise and categorise objects or concepts without having seen any examples of those categories or concepts beforehand.

The GenAI approach to identifying vulnerable customers

One of the attractions of GenAI is ↖ **zero-shot capability**. In this context, that means that GenAI can be prompted to act like the NLP model described above, i.e. identify vulnerability in customer responses, but without the need for training data or the time-consuming expert human annotation process.

STEP 1

How we build it

Step 1 was to design an initial prompt (the priority escalation description) for GPT-4:

Prompt

You are an AI assistant that helps call centre agents prioritise which cases to work on, based on a description of what a priority case looks like and a message from the customer. The messages you see are SMS responses from the customer.

You process messages in the following JSON format:
{"priority_escalation_description": "This is a description of the kind of cases which need to be prioritised",

"customer_messages": [This field contains a message that the customer has sent within the case]}

Based on the **priority_escalation_description** which instructs on what cases should be prioritised for escalation, you read the customer_message, and decide if the case meets the criteria for escalation.

If the case requires escalation, respond in the following format:
(string, int) where string = ESCALATE and int = the index of the highest priority message (using an index beginning with 0).

Otherwise respond:
(string, int) where string = BAU and int = 0.

Remember, when escalating a case, you only prioritise cases that fit the priority_escalation_description. All of these cases are already escalated, it is your job to prioritise which cases are actioned first based on the instructions only.

Priority escalation description

Only prioritise cases where the customer has described an issue which could make them or a family member vulnerable. This can include reporting a health issue such as old age, disability, welfare or benefits status, bereavement, mental illness, being ill, unwell and other like issues. Also take into account economic issues, such as the customer reporting that they are out of work, unemployed, on benefits or disability allowance. Look out for informal ways the customers might express this issue.

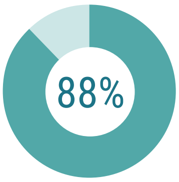
→ **Prompt**
The input a user provides to a generative AI solution in order to request it to perform a certain action.

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

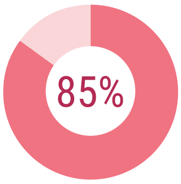
→ **Recall**
Recall is a measure of how many of the true positives a model is finding in the data.

The same 10,871 verbatims used for training the NLP model were run through this ↖ **prompt** but these did not include any of the labels assigned by the expert human annotators, i.e. GPT was not provided any examples of what a human considered an example of a verbatim that exhibited vulnerability.

The results



↖ **Precision**
88% of predictions of vulnerability were correct i.e. matched the human annotators' labelling



↖ **Recall**
The model found 85% of all the verbatims that a human had assigned as vulnerable

The verdict

- **This is remarkable performance, but below that of the trained NLP model.**

However, to be completely fair about the comparison, we felt that we could refine the priority escalation description to improve results. The refinement we applied was based on looking at the predictions the GenAI model got wrong, as well as the vulnerable verbatims the model missed (did not recall) and updating the priority escalation description to account for these.

STEP 2

The refined policy escalation description

After several iterations, the policy escalation description was updated to:

Policy escalation description

Only prioritise cases where the customer has described an issue which identifies them, a family member or someone they care for as vulnerable. This can include a customer reporting a health issue, such as pregnancy, old age, a disability, experiencing bereavement or the funeral of a family member, mental illness, and being ill or feeling unwell (mentally or physically).

Remember, being frustrated or upset with the service does not make someone vulnerable.

Attending a medical appointment may also indicate vulnerability but remember the customer must specifically indicate the appointment is doctor or hospital related - since we are communicating about a broadband install appointment this could be confused.

The customer may report that the service is disrupting essential medical equipment like an alarm, phoneline to emergency contacts or an essential lifeline. Or the customer could indicate that they are a carer for vulnerable dependents, such as an elderly parent, a baby or newborn, or an ill or disabled family member.

Just having children does not make someone a vulnerable case. If the service is for a primary care facility, that should be considered a vulnerable customer. Also take into account responses which indicate that the customer is experiencing financial hardship, such as responses highlighting bankruptcy, not being able to pay, being a welfare case, being on unemployment benefits, or other benefits like disability allowance.

However, remember that having to take time off work for the service completion does not make someone vulnerable. If the customer is concerned about a scam or fraud, or indicates that they have trouble speaking English, this also indicates that the customer may be vulnerable.

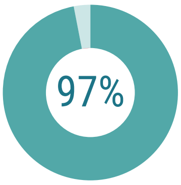
Look out for informal ways the customers might express these issues, though bear in mind z– if a customer responds, 'Not Working', they're likely describing their broadband service and not their economic circumstances and therefore should not be considered vulnerable. Asking for help also doesn't mean someone is vulnerable, unless they are asking for special accommodations or adaptions or reporting additional medical needs.

And finally, do not make assumptions about whether someone can speak English based on names or the quality of their grammar.

You can see that this updated version is significantly more detailed, and specific.

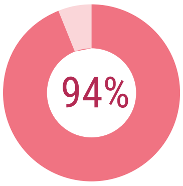
Updated results for the GenAI Model

The final iteration of the prompt resulted in the following performance:



Precision

97% of predictions of vulnerability were correct i.e. matched the human annotators labelling



Recall

The model found 94% of all the verbatims that a human had assigned as vulnerable

IN THE KNOW

Generative AI (GenAI) for Text

Prompting a GenAI model feels like you are telling the model what to do by providing it a set of instructions that it should follow. However, the model does not really treat these as hardwired instructions, but more of a guide.

GenAI models seek to generate the most probable word(s) that follow on from the prompt it is given. By adjusting the prompt, the probability of the follow-on word(s) is changed such that it will seem to abide by what has been entered into the prompt.

However, it is not guaranteed to abide by all the inputs in the prompt, or for each part of the prompt to have an equal effect on the change in the probability of the output.

It is for this reason that prompting can provide **a false sense of security**, especially when it comes to 'guardrails' i.e. where you tell it things it should not do in addition to the things it should. The model is not guaranteed to follow the entire prompt, but it may seem to do so if you only conduct a handful of tests.

This is another reason why testing at production scale is so important, as it only then that you will observe any anomalies/weakness in how the model has attempted to follow the prompt.

GenAI models do not treat prompts as hardwired instructions, but more of a guide.

Comparing the NLP and GenAI approaches

The results

Performance measure	**NLP**	**GenAI** Initial iteration	**GenAI** Final iteration
Precision	93%	88%	97%
Recall	91%	83%	94%

The 'Generative AI (final iteration)' solution achieves superior performance to the NLP solution. **But the final iteration performance was only achieved because a labelled data set of 10,871 customer verbatims was available to test the GenAI solution and create an updated prompt that reflected the mistakes/misses from the initial iteration.**

→ **Confidence score**

Classifier models typically assign a numerical confidence score to a positive prediction, giving an indication of how likely that positive prediction is a true positive prediction. Note, whilst higher confidence scores indicate a more likely true positive prediction, it is not the case that high confidence predictions are always true positives – AI models can be confidently wrong.

→ **Environmental impact**

All models require the use of computational resources and this increases as models get larger. The use of computational resources requires electricity to not only run the computers, but also to cool the data centers where those computers are located. Environmental impact refers to the emissions from this electricity consumption.

How does this help you?

The 'Generative AI (initial iteration)' represents what is achievable with no human labelling. This is still an impressive level of performance and, although not quite at the level of the NLP model, or the Generative AI (final iteration). At first glance, it may feel obvious that the performance differential between the Gen AI (initial iteration) and NLP approach is not sufficient to warrant the human labelling of data at all.

But, the only reason you can assess the performance differential is by having the labelled data to analyse performance with – **even if you don't need labelled data to train the model, you certainly need it to assess performance and explain the decision-making rationale.**

So the human labelling costs remain, whichever approach you choose.

If we remove the initially lucrative benefit of not needing training data from the equation, which approach might be preferable? To answer this question, we need to look at some of the pros and cons of the two approaches, including from a risk and controls perspective.

Pros & Cons of NLP vs GenAI (all use cases)

Aspect	NLP	Generative AI
Control	An NLP model outputs a ↖**confidence score** for each prediction. This allows for a confidence threshold to be set only above which a prediction is accepted, and this affords the user more control over the precision of the model	Unlike an NLP model, the output from Generative AI does not provide a confidence score. This means you have to accept the prediction provided without any discretion
Cost of running the model	Cheap – compared to GenAI solutions, NLP models are much less intensive in terms of the compute resources required to run them and are therefore cheaper to run	Expensive – due to the size of the LLMs that underpin them, GenAI solutions are highly compute resource intensive and therefore more expensive to run
↖**Environmental impact of running the model**	Lower – related to the compute resources required, NLP models require much less energy to run	Higher – the compute resources required to run these models mean higher energy use

The final word?

We knew that clients would want to understand whether GenAI would be a superior option for them (compared to e.g. NLP) to identify potential vulnerable customers. We wanted to know too! So we designed and conducted an experiment to compare the performance and efficacy of a GenAI-based solution to a trained NLP model. On balance, and in this instance, we would recommend a client utilises the NLP approach to vulnerable customer identification due to the additional controllability it affords, as well as lower costs and lower environmental impact, but it was only by comprehensively testing the two approaches that we could arrive at this evidence-based recommendation.

This highlights an important point when considering the use of a GenAI-based solution: GenAI can be used to do many things and indeed perform well at those things, but that does not mean it is the only solution. When you are presented with a GenAI-based solution to a problem, it's always best to ask if other alternative approaches – AI or otherwise – were considered to solving the problem at hand.

Jump to

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

Read

Jump to

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

Read

ContactEngine
A NICE company

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