NOT DOOMED TO REPEAT

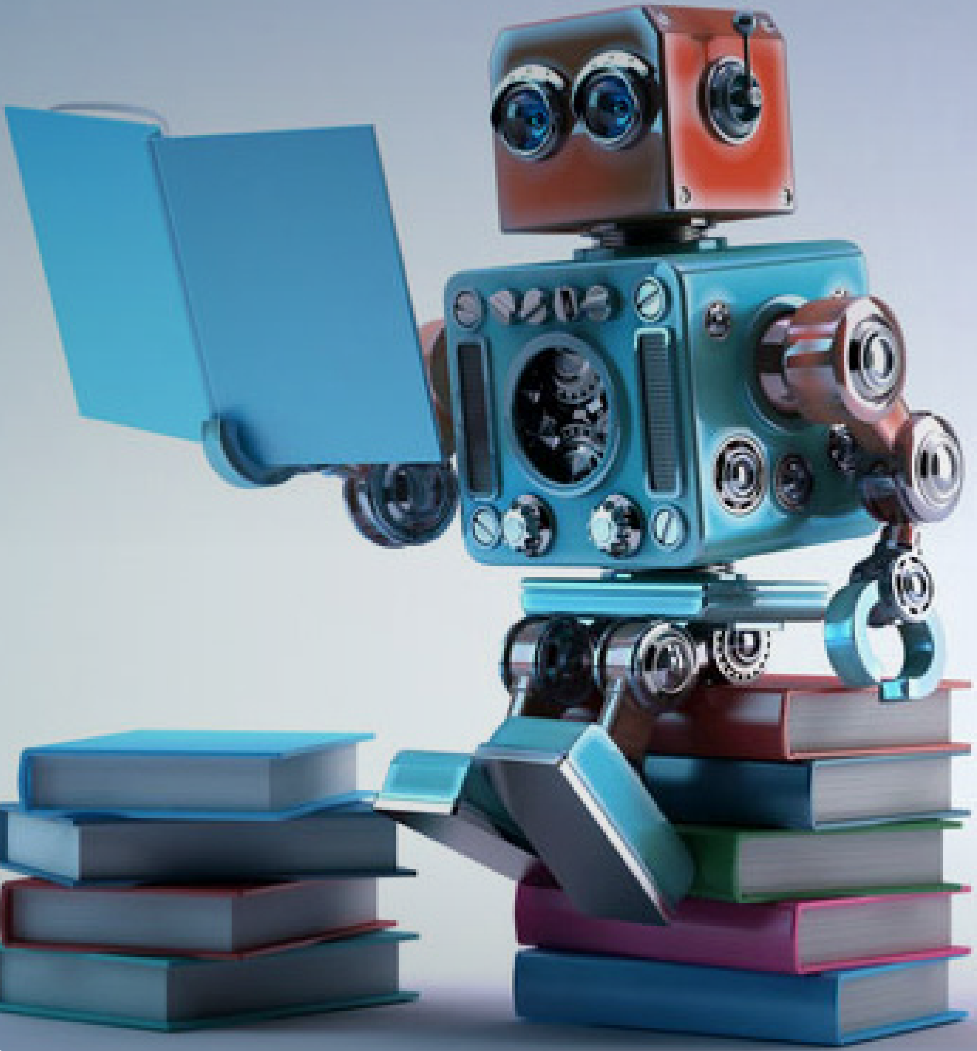
Applying Lessons from CX Text Analytics to Generative AI

January 2024

AUTHORS

Fiona Moss

Rich Timpone, Ph.D.



IPSOS VIEWS
AI SERIES

GAME CHANGERS Ipsos

#IPSOSHiAi

At Ipsos, we champion the unique blend of Human Intelligence (HI) and Artificial Intelligence (AI) to propel innovation and deliver impactful, human-centric insights for our clients.

Our HI stems from our expertise in prompt engineering, data science, and our unique, high quality data sets – which embeds creativity, curiosity, ethics, and rigour into our AI solutions, powered by our Ipsos Facto Generative AI platform.

Our clients benefit from insights that are safer, faster and rooted in the human context.

Let's unlock the potential of HI+AI

Generative Artificial Intelligence (AI), with ChatGPT being one high profile example, has rapidly democratised the power of text-based Artificial Intelligence. Essentially, anyone and everyone with access to the internet can now ask questions and get responses from these super-smart bots. These tools also enable the broad application of text analytics in numerous use cases. While we are in a new landscape, learning from the past of text analytics will ensure we don't repeat errors and can leverage the new tools to their greatest advantage.
In this way we can learn from history, so we aren't doomed to repeat mistakes as we seize new opportunities.

Given our focus on text analytics, when we discuss Generative AI in this paper, we are generally focusing on Large Language Models (LLMs). LLMs are probabilistic text generation tools, meaning, in simple terms, that they have been trained to predict, in response to a prompt, the most likely word or token sequence to follow – giving the impression of 'human' speech[1]. However, they can also be leveraged for many practical text analytics use cases. It is here where learning from past experience in the field is particularly useful, although, at a high level, these learnings are also applicable for other types of Generative AI including image and multi-modal models.

The democratisation of LLMs for text analytics is in marked contrast with the early days. In 2009, when Ipsos' Customer Experience (CX) team first started delivering text analytics, it was a niche offering, used by a small number of clients looking for an efficient way to understand huge volumes of unstructured or text data. Skip forward over 10 years and text analytics is now standard in most large or ongoing CX programmes, providing identification and quantification of key topics and sentiment across solicited (e.g. open-end questions) and unsolicited (e.g. social media) feedback.

Given the stunning rapidity with which Generative AI has been taken up by the world, its exploration and use for text analytics will be much faster than past tools. While we can all be greatly impressed by the potential of Generative AI, and LLMs themselves, leveraging the lessons learned from the testing and application of past text analytics supports organisations in confidently designing and delivering profitable, sustainable, and positive customer experiences moving forward.

> While we are in a new landscape, learning from the past of text analytics will ensure we don't repeat errors and can leverage the new tools to their greatest advantage. ”

LEVERAGING THE PAST

At Ipsos, we have been applying the framework of Truth, Beauty, and Justice to evaluate the quality and responsible use of Generative AI models[2]. This reinforces the idea of learning from the past, as this framework was adapted for an earlier era of AI models to understand social and behavioural attitudes, processes and actions, but is just as relevant, and in some cases more relevant, to the new generation of tools[3,4].

In this framework:

- **Truth** focuses on the accuracy of the models and their outputs;
- **Beauty** deals with the explainability of the outputs, and, in some use cases, the ability to surprise and generate new insights;
- **Justice** encompasses multiple important areas – AI ethics, bias, algorithmic fairness, data security and privacy, alongside the rights and responsibilities of creators of data used for training, and the users of the tools.

With this framework in mind, in this paper we outline five key learnings that are still relevant as teams apply LLM-powered Generative AI tools:

1. Demand transparency
2. Don't forget the data
3. Formal evaluation still matters
4. Remember to manage expectations
5. Establish a reporting/usage mechanism that meets business needs.

1. DEMAND TRANSPARENCY

In the early days of text analytics, the industry was overflowing with jargon and hype. Deciphering alien terminology and triaging between multiple providers all making best-in-class claims often created a barrier to first use.

While Generative AI has created some of its own jargon and hype, not least raising awareness of LLMs, it has bypassed some barriers by putting free-to-access user-interface-based technology straight into the hands of potential users. The interfaces and AIs make these more accessible and less intimidating than the past, but they are not accurate for all use cases, and how they are prompted will impact the quality of what is produced. Again their purpose as probabilistic text generation tools makes their application a potential issue for text analytics use cases.

As organisations move to put in place contracts for enterprise-based access, clean rooms/ walled gardens, thus embedding this technology into their day-to-day operations, Ipsos' recommendation was, and continues to be, to challenge a provider to articulate, with clarity, its outputs and limitations, in addition to its benefits. Such transparency is an important part of Beauty.

For LLMs, these considerations include:

- A clear statement of what the model has been trained to do (regardless of claims from providers, we strongly encourage your own testing, or seeing the validation of others, as LLM accuracy will vary across use cases);
- An understanding of the nature and volume of data used to train the model (plus any limitations) to identify what insights build directly from the corpus and which go beyond. Both have potential for hallucinations – statements of 'fact' that are in reality fallacies invented by the technology – but with different levels of risk;
- Evidence of whether the model will continue to learn and adapt as it experiences new data, or if it is fixed, requiring retesting applications to ensure they still perform as tested previously as updates may have led to declines in quality in some areas[5];
- Questions on how your data will feed back into any such updates and model training;
- How the LLM can be harnessed by existing business systems – API access, the ability for data engineers to build in links etc. – so that the LLM can be used operationally in the way the business intends.

Despite the apparent flexibility and seeming intelligence of many LLMs, clearly defined business and research objectives for deployment are still required from the outset. As a result, keeping humans in the loop from initial model training through to delivery of research results is key, even for foundation (pretrained) models that receive reinforcement learning from human feedback (human correction). 'Humans in the loop' improves their quality – as hallucinations are an ongoing risk – keeping the outcomes on track. This blending of Human Intelligence with Artificial Intelligence is core to Ipsos' AI philosophy.

However, for LLMs, transparency isn't just required in terms of the capabilities and workings of the model. It is also important to understand where the data provided by the business is going for the activation of the output.

With this in mind, data privacy and security are key concerns for many open-access models, and highlighted in the terms of use by them. We encourage buyers to put in place enterprise contracts, governance, and infrastructure to ensure that sensitive customer, employee, and proprietary data and information are adequately protected. Many companies, including Google, now tell their teams not to use public chatbots, like ChatGPT and Bard, with any sensitive data[6]. Understanding the privacy and security terms of the solutions being considered is key to risk management and business comfort when using this new technology freely.

> Despite the apparent flexibility and seeming intelligence of many LLMs, clearly defined business and research objectives for deployment are still required from the outset. ❞

2. DON'T FORGET THE DATA

Rubbish-in Rubbish-out/Garbage-in Garbage-out

The rubbish-in rubbish-out paradigm (GIGO in the US) has always been true for text as well as all types of analytics. Indeed, all text analytics is a function of the training data. Fundamentally if the data involved is not representative or relevant to your business question or does not contain sufficient detail to answer that question, then text analytics will not deliver against your objectives[7]. But this is because of the data itself and not the analytics.

For LLMs, this paradigm remains. Indeed, we need to be sure that both the text data under analysis, and the text data that is used to train the LLM, are fit for purpose. Beyond the general principle, when training or fine tuning models in particular, the tools will extrapolate answers to questions that go beyond the data used in the set-up of the foundation models. While these answers may provide interesting hypotheses, they are not true insights and have even greater potential to be misleading.

There is an implicit trust then that training data is going to deliver reliable outcomes that can be used to inform business decisions. But we know that in foundational LLMs cultural and group biases exist, reflecting the internet. The issues of data quality and representativeness are critical for teams training their own models as well.

To get a level of trust and improve explainability before deployment, it is essential to understand what data has been used to train the LLM. It is the quality and volume of this data that will dictate whether the LLM will deliver correct responses. Insufficient, missing or biased data can deliver inaccurate or even misleading results. This ties into the fundamental need for evaluation and considerations of Justice as well as Truth.

In the world of CX all of this due diligence regarding transparency and data is necessary to build trust before allowing LLM-powered tools to:

- Access customer or business data;
- Interact directly with your customers (e.g. in the form of enhanced chatbots);
- Facilitate staff to do their job (e.g. by recommending actions to frontline staff or providing summary level insights from large corpuses of data to inform business strategy).

Without this due diligence in place, the risks to breaching customers' trust and/or delivering an incorrect or substandard customer experience are too high.

Fundamentally if the data involved is not representative or relevant to your business question or does not contain sufficient detail to answer that question, then text analytics will not deliver against your objectives. ”

When using LLMs (i.e. where text-based interaction between user and machine is expected) it is also useful to understand from your provider the prompts and questions that elicit the best and most accurate responses from your technology – just as it is important in text analytics to choose open-ended questions that invite precise and detailed responses or unsolicited data sources that provide relevant, articulate content.

Native language considerations

For businesses operating across multiple markets, native language has always been a consideration for text analytics. The key decision is between building a single consistent text analytics model in one language across all comments, and using automated translation to put the comments into the same language; or building multiple native language text analytics models but losing some consistency and comparability along the way. The latter has the benefit of being specifically tailored to the markets involved, whereas the former carries cost and efficiency savings, as well as easier inter-market analysis. Most recently, we have been addressing this question in the cross-cultural databasing of the foundations of emotion in the Ipsos Emotion Framework[8].

Languages are also a consideration for LLMs. While many have been exposed to data sources in numerous – in some cases hundreds of different – languages during the training process, this does not mean that they offer the same level of performance in each language. As a result of native language training, the quality and responses given to the same prompts may be dramatically different in different languages.

In one test of the quality of Generative AI tools across the use cases of transcription from video audio tracks, translation, sentiment, and theming of content conducted by Ipsos researchers, the performance of different LLMs and providers differed from each other and across languages substantially[9]. Thus, when evaluating LLMs, quality checks need to be done for each use case and language to determine the relative and absolute accuracy for each, and whether they are fit for purpose.

Therefore, just as for text analytics, identifying the languages that you wish to use with the LLM and checking that it supports them in an adequate way is essential. In some cases, this may mean asking users to interact with the LLM in their second language to get the best out of it. As a result, it is also key to identify who will be using the LLM, in what way, and – given the importance of prompt creation – if they have any potential language skills required, particularly if your customers are likely to have direct interactions with the tool.

Working with your chosen technology to get the best out of it is therefore as important now as it was in 2009.

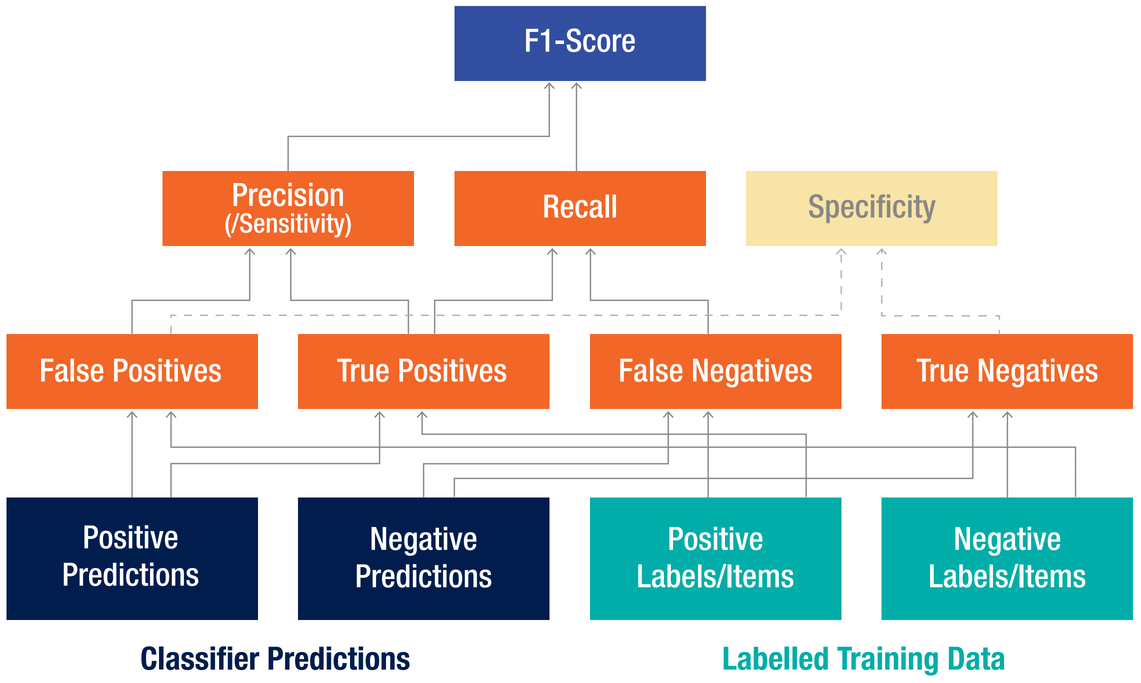
> Therefore, just as for text analytics, identifying the languages that you wish to use with the LLM and checking that it supports them in an adequate way is essential. ”

3. FORMAL EVALUATION STILL MATTERS

One of our core messages, and the underlying premise of the dimensions of Truth, Beauty, and Justice, is that Generative AI needs to be evaluated with the same rigour that text analytics has been subjected to for years. Given their qualitative construction and probabilistic building, LLMs are often reviewed more through face validity – whether the output appears reasonable – than formal scrutiny. One of our key learnings from the past though is that to get the most value from text analytics, the quality for specific use cases must be systematically evaluated.

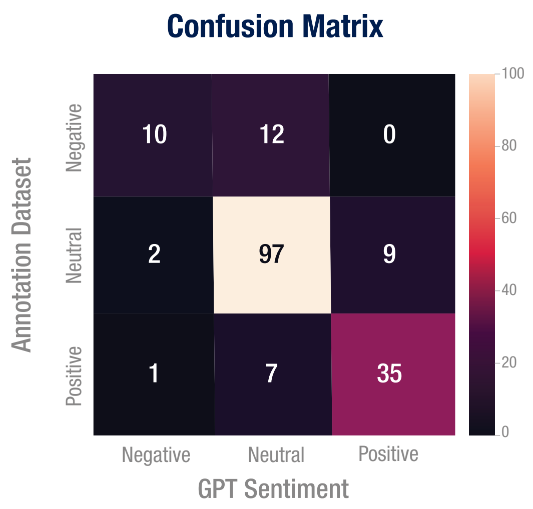
Figure 1 provides a framework that is used for evaluating sentiment and other text classifications' quality. These go beyond simple percent accuracy scores to systematically understand the overall quality as well as where issues may exist[10]. To create such tests, evaluation is needed against ground truth baseline measures. While this was standard in the past, it is less common with LLMs. However, we believe creating such formal evaluations is critical for use case evaluations including text analytics.

Figure 1: One Set of NLP Evaluative Metrics



Source: Kanstrén, T. (2020) "A Look at Precision, Recall, and F1 Score"

Figure 2: Ipsos Evaluation of an LLM Sentiment Coding



Holding LLMs to the same rigour ensures obtaining the most value

	Precision	Recall	F1-Score	Support
Negative	0.77	0.45	0.57	22
Neutral	0.84	0.90	0.87	108
Positive	0.78	0.81	0.80	43
Weighted avg	0.81	0.82	0.81	174

Explainer: Precision – percent of correct positive predictions out of all positive prediction; recall – percent of positive predictions out of all true positive cases; f1-score – the harmonic mean of the two measures.

Source: Legg, J. and Bangia, A. Ipsos UU AI Quality Assessment.

In the example mentioned earlier, where we tested various LLMs and third-party suppliers, Ipsos explicitly created baselines for evaluation in each language. *Figure 2* provides an illustration of one category test in one language with a confusion matrix that evaluates where AI generated sentiment coding is predicting the ground truth accurately. The table goes further with each of the tests from *Figure 1* demonstrating the differential quality. While a full discussion of the tests and measures is beyond the scope of this paper, this illustration shows that comparing different tools against each other and against objective standards can inform decisions of when and where tools are viewed as providing adequate quality for use, and where they fall short.

The point is not to hold LLMs or any Generative AI tools to higher standards than in the past, but to avoid being so impressed that we hold them to lower ones. The lesson from the past is that the rigour of evaluation has created quality standards and ways to compare measures. We believe that is a lesson that deserves to be applied today.

4. REMEMBER TO MANAGE EXPECTATIONS

In the early years, text analytics became a victim of its own hype, failing to live up to huge promises of very high accuracy levels.

Now claims are more moderate about both the level of categorisation accuracy and the level of coverage (i.e. how much of the text data available is included in the model). Indeed, in today's world, any provider promising 100% accuracy or coverage certainly merits some follow up questions!

The years of text analytics have also taught us that there is a balancing act between accuracy and coverage. For example, the more accurate a category is required to be, the more likely the analyst is to push out relevant comments along with the noise. As a result, accuracy goes up, but coverage goes down. In contrast, as we build broader categories, allowing in some noise along with relevant comments, accuracy goes down, but coverage goes up. The success of this balance relies on the analyst's skills and the end-user's expectations. We have seen this in domains from survey research to large scale social listening research. In fact, additional coding for specific domains is often necessary to improve accuracy even when 'Big Data' is being investigated.

LLMs are not exempt from concerns about accuracy as we have seen in the previous section. Indeed, Generative AI confidently asserts its answers regardless of whether they are correct. We must remember that its answer is 'simply' based on the probability that certain words will follow each other in response to a given prompt – all depending on the text it was trained on.

Given the concerns with hallucinations and other issues of accuracy, this response certainty is especially concerning for learning and exploration for which chatbots in CX are often used. While extremely valuable to speed up the work of CX practitioners, fact checking and not accepting the confidence of LLM responses at face value is key. There have been many examples where this has not been carried out,

including the lawyers who filed a brief that was riddled with fake legal precedents that were generated by ChatGPT[11]. This absolutely did not help their case.

Therefore, just as for text analytics in the past, we need to manage end-users' expectations about commentary provided by LLMs or Generative AI. It may be confident and sound competent, it may even appear in a slick and glamourous user-interface, but that does not mean that every answer should be accepted as the truth. The world of Artificial Intelligence sounds intimidating and, well, more than a little intelligent, but that does not mean that the outputs it provides should go unchallenged and unchecked – particularly if those answers are intended to inform the decisions and behaviours of staff or customers who may not themselves be in a position to question the veracity of the information received.

5. ESTABLISH A REPORTING/USAGE MECHANISM THAT MEETS BUSINESS NEEDS

One of the biggest shifts in text analytics has been in the way it is used. When text analytics was starting out, it was enough that a simple spreadsheet or slide could show what was being said in – potentially – several million verbatim comments.

Today though, text analytics is not simply a static slide. In the world of CX, for example, it is a dynamic tool used to drive better customer experiences. Increasingly this means real-time text analytics delivered straight into the hands of end-users, so agents can respond and close loops on problems almost as they occur; or deep-dive post-text analytics exploration to identify the impact of particular feedback on KPIs, for example.

LLMs and Generative AI pick up where text analytics already is – with existing, configurable interfaces for live interactions. These interfaces, together with models that support the right functionalities, need to be put into the hands of the right users.

For example:

- Contact centre agents working to close the loop on red flags may require a system to help them deliver the best intervention when faced with specific problems;
- Insight professionals may require text summarisation tools to synthesise insights from multiple data sources rapidly and efficiently;
- Analysts may benefit from an interface that provides suggestions to optimise code or automate outputs.

All of this leads us to a place where LLMs make our lives easier, taking the weight out of some of our tasks, and leaving humans to do what humans do best: thinking outside of existing parameters and interacting with other humans – essentially working to build strong customer relationships in innovative ways.

CONCLUSION

Text analytics has grown up. It is no longer a small child clamouring for attention, and sometimes misbehaving when it gets it. It is an established adult capable of clearly and reliably informing today's business decisions.

But as with all adults, to get the most from it, we also need to treat it fairly. This means being clear about our business objectives; not expecting text analytics to find content where there is none in the underlying data; bringing realistic expectations to any assessment of quality; and ensuring that the end results are analysed and conveyed to end-users in the best way possible.

LLMs and other forms of Generative AI tools are traditional text analytics' sophisticated (and badass) cousins, expecting the same fair treatment, adding in a few new conditions, but delivering results that currently astound many of us. Will we eventually adapt to this sophistication and consider it every day as we now do with text analytics? Absolutely, and even faster than the past as many are already embedding LLMs into existing tools and our everyday lives. And, for CX, will the process of doing this also lead to better, more loyal, and more profitable customer relationships? Almost undoubtedly it will, if we treat LLMs with the respect they deserve, learn from the past, and embrace the future.

REFERENCES

1. Wolfram, S. (2023). "What Is ChatGPT Doing … and Why Does It Work?" 14 February, 2023. Stephen Wolfram Writings. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

2. Timpone, R. and Guidi, M. (2023) Exploring the Changing AI Landscape – From Analytical AI to Generative AI. Ipsos Views. https://www.ipsos.com/en/chatgpt-and-rise-generative-ai-navigatingchanging- landscape-ai

3. Taber, C.S., and Timpone, R.J. (1996). Computational Modeling, Quantitative Applications in the Social Sciences #113, Sage Publications, Thousand Oaks, London and New Delhi.

4. Timpone, R. and Yang, Y. (2018). "Justice Rising: The Growing Ethical Importance of Big Data, Survey Data, Models and AI." Paper presented at the 2018 BigSurv Conference; Barcelona, Spain.

5. Zumbrun, J. (2023) "AI Surprise: It's Unlearning Basic Math". The Wall Street Journal. August 5-6, 2023; p A2.

6. Dastin, J. and Tong, A. (2023) "Focus: Google, one of AI's biggest backers, warns own staff about chatbots". Reuters. June 15, 2023.

7. Ho, C. and Mu, J. (2023) Humanizing AI: Real human data to generate and predict real innovation success. Ipsos Views. https://www.ipsos.com/en/humanizing-ai-real-human-data-generate-and-predict-real-innovation-success

8. Garcia-Garcia, M., Baldo, D., and Timpone, R. (2023) Emotions Around the World: A Cross-Cultural Framework for Emotion Measurement. Ipsos Views. https://www.ipsos.com/en/emotions-around-world

9. Legg, J. and Bangia, A. (2023) Conversations with AI: Unveiling AI quality in qualitative workstreams. Ipsos Views. https://www.ipsos.com/en/conversations-ai-part-ii-unveiling-ai-quality-qualitative-workstreams

10. Kanstrén, T. (2020) "A Look at Precision, Recall, and F1 Score" Published in Towards Data Science. https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

11. Neumeister, L. (2023) "Lawyers blame ChatGPT for tricking them into citing bogus case law." Associate Press. June 8, 2023. https://apnews.com/article/artificial-intelligence-chatgpt-courts-e15023d7e6fdf4f099aa122437dbb59b

FURTHER READING

1. Exploring the Changing AI Landscape https://www.ipsos.com/en/chatgpt-and-rise-generative-ai-navigating-changing-landscape-ai

2. Conversations With AI: How generative AI and qualitative research will benefit each other https://www.ipsos.com/en/conversations-ai-how-generative-ai-and-qualitative-research-will-benefit-each-other

3. Conversations With AI: Unveiling AI quality in qualitative worksteams https://www.ipsos.com/en/conversations-ai-part-ii-unveiling-ai-quality-qualitative-workstreams

4. Conversations With AI: How AI boosts human creativity in ideation workshops https://www.ipsos.com/en/conversations-ai-part-iii-how-ai-boosts-human-creativity-ideation-workshops

5. Humanizing AI: Real human data to generate and predict real innovation success https://www.ipsos.com/en/humanizing-ai-real-human-data-generate-and-predict-real-innovation-success

6. High Hopes: Tips for ensuring successful text analytics https://www.ipsos.com/en-uk/high-hopes-tips-ensuring-successful-text-analytics

7. GenAI: The need for Human Intelligence (HI) with Artificial Intelligence (AI) https://www.ipsos.com/en/almanac-2024/gen-ai-need-human-intelligence-hi-artificial-intelligence-ai

8. Going All In With AI? Here's How to Keep the Customer at the Center https://www.ipsos.com/en-us/going-all-ai-heres-how-keep-customer-center

9. The Modern Marketer Dilemma: Making Artificial Authentic https://www.business-reporter.co.uk/management/the-modern-marketer-dilemma-making-artificial-authentic

NOT DOOMED TO REPEAT

Applying lessons from CX Text Analytics to Generative AI

AUTHORS

Fiona Moss, Head of CX Global Analytics Team, Ipsos

Rich Timpone, Ph.D., Head of Global Science Organization, Ipsos

The **Ipsos Views** white papers are produced by the **Ipsos Knowledge Centre.**

www.ipsos.com
@Ipsos

GAME CHANGERS