Human Intelligence:
The Key to Harnessing Generative AI

»

opusresearch

July 2023

Dan Miller | Lead Analyst & Founder, Opus Research

Opus Research, Inc.
893 Hague Ave.
Saint Paul, MN 55104

www.opusresearch.net

How ChatGPT Changed Customer Expectations

Since November 2022, Generative AI has been experiencing its "TikTok Moment" because hundreds of end-users are taking turns using their own words to get results from ChatGPT, Bard, Claude, Firefly, or another low-cost front-end to the world of Conversational AI.

In very short order giants of enterprise IT, cloud-based infrastructure, digital marketing, and search have rushed to make these portals into omnipresent virtual assistants. These assistants are capable of taking notes during discussions, researching relevant topics, summarizing long discussions, or drafting follow-up emails.

ChatGPT, the high-profile offering from OpenAI, grew its user base to 100 million in record time. The user base continues to grow as giants of cloud-based IT, search, social networking, and commerce accelerated the introduction of new services that transform the way both customers and employees interact with the growing population of bots for customer care and employee productivity.

Defining the Conversational AI Lexicon

Here are often-used terms and concepts that are important to understand in order to overcome misinformation and disinformation surrounding Conversational AI while evaluating alternative approaches to incorporating it into Contact Center and IT infrastructure.

Deep Neural Networks (DNNs): Layers of computing units that simulate the human brains ability to recognize patterns and learn. They use sophisticated mathematical modeling to process data in complex and often non-explainable ways.

Large Language Models: Artificial intelligence models that result from ingesting both structured and unstructured digital information. GPT is one of the most conspicuous but is not alone with candidates from Google, Meta (LLaMA), Baidu (Ernie), and more to come.

Pretraining: Analytic and machine learning processes that detect patterns that, in effect, train an LLM to predict missing words in a sentence and generate the next word in a sequence. The objective of pretraining is to learn a representation for the LLM that can be fine-tuned for specific end-user tasks.

Generative AI: A class of AI that is able to generate new content that is similar to the data that it has been trained on.

We use the word "transform" on purpose. "Transformer" is the "T" word in "GPT". It is more accurate to call it "Transformational" because transformers are the foundational neural network architecture that make it possible to ingest massive amounts of information, aka "Large Language Models" (LLMs). These models detect patterns in ways that enable them to categorize, summarize, translate, predict, and ultimately generate rich content spanning text, voice, graphics, and video. With so many individuals putting generative resources to use for their own purposes, the amount of creative energy dedicated to improving results and outcomes is unprecedented.

AI CAN TURN CONTROL TO THE AGENT WHEN A PROBLEM BECOMES TOO COMPLEX FOR IT TO SOLVE. AN ADDED ADVANTAGE OF HAVING THE AGENT "IN-THE-LOOP" IS TO CHECK THE LLMs OUTPUTS FOR INACCURACIES.

Great Expectations: Use Cases for LLMs for Contact Centers, CX, and EX

The speed at which Generative AI and LLMs have been assimilated into everyday business workflows is stunning. Because resources are made available to the general public through portals that enable anyone to provide prompts "in their own words" to describe exactly what they want. Press coverage has unfortunately focused on the more sensational threats of Generative AI, citing job loss due to automation, manipulation of opinions expressed on social media. Fears also include a proliferation of deepfakes and computer-generated content overwhelming human-generated content by the sheer volume of pieces it is able to generate.

Thus far, rather than training their replacements, employees in almost every department in every vertical are training it to make them more efficient at their own jobs. Here are just a few examples:

➤ **Conversational Flow Design:** LLM-based resources have been used to ingest conversations between companies and their customers through both digital and voice channels in order to accelerate the development of conversational flow designs and customized language models that can be used to inform bots and live agents alike. The results are chatbots or voicebots that quickly recognize the intent of a customer's contact and speed them to task completion.

➤ **Call Summarization and Reporting:** One of the most "high-impact" early implementations that has proven positive impact on a company's bottom line because it provides accurate summaries of agent conversations, complete with notations regarding compliance with laws and internal guidelines, in a fraction of the time that an agent can perform such a task.

➤ **Agent Assistance and Bots:** Customer support agents can employ Conversational AI as personal assistants that can draft responses to customer queries that can be pasted into texts, emails, or spoken over a voice channel. Workflows are evolving to treat the agent as the subject matter expert who can review the virtual assistant's output and correct or personalize it before conveying it to the customer.

➤ **Content Generation:** In addition to emails or text messages between agents and customers, Generative AI resources are now called on to prepare drafts of business documents, such as financial summaries, interpretations of complex legal issues, promotions for products, services or events, and other marketing literature.

➤ **Research:** Employees in all departments can prompt an assistant to study various topics and find answers that are informed by the sum total of information in the hyperscalers' LLMs and augmented by Conversational Intelligence derived from the enterprise's own knowledge base as well as the content of conversations among employees and customers.

LLMs WILL NEVER BE TRANSPARENT. CORE DECISIONS ARE MADE IN A LAYER OF THE COMPUTE FABRIC THAT MIMICS THE HUMAN NERVOUS SYSTEM AND ARE BASICALLY "HIDDEN" FROM AUDITS.

➤ **Language Translation:** Multinational firms have found it very valuable to employ a virtual agent to sit in on virtual meetings and provide simultaneous translations of conversations. There are similar applications for customer care and eCommerce sales conversations involving multiple languages.

➤ **Anonymization and Redaction:** LLMs can be trained to recognize personal information or personally identifiable information (PII) to prevent unauthorized sharing or its use in training procedures.

In Case You Haven't Heard: Failure is More Than an Option

We are only months into the rapid adoption of Generative AI in the enterprise IT fabric. Yet developers have had several years to acquaint themselves with the underlying LLMs and their experiences inform a lot of objections that are shaping debate about whether end results are so flawed and, in some cases, dangerous that it's time to "push the pause button" on Conversational AI deployment.

➤ **Training Biases and Manipulation:** Dating back to Microsoft's introduction of the chatbot named Tay on Twitter in 2016, it has been clear that interactions with an organized group of end-users could exacerbate "training biases." In less than 24 hours, an anodyne conversational bot on Twitter was converted into a racist misogynist as it learned from the content of some 96,000 messages from followers. The tendency for a small group of bad actors to carry out efforts to bias the output of chatbots and virtual agents is not expected to go away. For this reason alone, there is a need for human intervention to monitor and amend the output of generative AI for the foreseeable future.

➤ **Incomplete Datasets:** Developers who cut their teeth on early renditions of LLMs from Deepmind, OpenAI, Baidu, and others recognized that the information contained was finite. There are beginning and end dates. Decisions were made on what should and should not be included. There was no effort to train on data in enterprise knowledge bases that might include product descriptions or inventory status.

➤ **Privacy Concerns:** To offer great service means to recognize, or sometimes to predict, a customer's intent and provide the response most closely matched to that prompt. This process relies on training the Generative AI on conversational data that includes personal data or personally identifiable information (PII). Companies must take extra care to comply with privacy laws when deploying Conversational AI. These elements should not be used to train the publicly available LLMs and, as noted above, a promising application for LLMs is to identify personal information and "anonymize" it if it is going to be used to train an LLM or to redact it if it should not be shared in any form.

Another overarching concern is "explainability" which is sometimes conflated with "transparency". Because they are built on computing architectures that employ "Deep Neural Networks" (DNNs), LLMs will never be transparent. Core decisions are made in a layer of the compute fabric and are basically "hidden" from audits. The ability to "explain" the AI process is attainable if humans are employed to evaluate multiple outcomes and describe how the LLM is likely to reply. In regulated industries, like financial services or insurance, this is a must.

The Case for Human Intelligence

Human intervention is required to compensate for the known shortcomings of Generative AI and LLMs. Individuals who have gained experience with one of the popular portals, like ChatGPT, Bard, Claude, and Dall-e, know from experience how to tailor prompts to generate the most useful response. In fact, educational institutions are rushing to develop the curricula for courses in Prompt Engineering.

➤ **Engineering Prompts:** As described above, tailoring prompts to get the best results from LLMs is emerging as a very important skill set for a new category of employee.

➤ **Curating the Knowledge Base:** As Conversational AI evolves to respond to prompts addressing multiple domains and subject matter areas. LLMs from hyperscalers (e.g. Google, Amazon, Microsoft) will be invoked as needed for tasks that do not require local knowledge. Employees with subject matter expertise can be involved to define and refine the information contained in specialized knowledge bases that include corporate data, customer histories, and the like.

➤ **Refining Responses:** In self-service situations, individual customers craft their own prompts. In these cases, there is significant concern that, if unfiltered, a Generative AI is likely to provide an inaccurate or inappropriate response. That's why another major role of human intelligence is in monitoring and refining the responses from an automated assistant. They are the first drafts of responses that can be refined to assure accuracy and compliance.

➤ **Adding Real Empathy:** LLMs are best thought of as mimics. Even though end-users assign human-like qualities to them, they do not have feelings. Therefore it is important to design interactions and conversation flows so that a live agent is involved when appropriate to show true concern about a customer's well-being or sentiment.

Use Cases That Strike a Balance Between Human and Artificial Intelligence

At [24]7.ai, LLMs are being evaluated and put to use for the following use cases. Each promises to have a positive impact on key performance measures and, ultimately, the bottom line:

➤ **Agent Co Pilot:** [24]7 Copilot allows contact center agents to accelerate each visitor's problem resolution or task completion by monitoring ongoing chats and fetching relevant information from a database of past, successful conversations. An LLM can converse on behalf of the agent, based on context from the current chat and guidelines inferred from past chats. It can turn control to the agent when a problem becomes too complex for it to solve. An added advantage of having the agent "in-the-loop" to check the LLM's outputs for inaccuracies. The key challenge here is goal-orientation. Visitors often seek precise solutions to their problems, and crisp conversations are preferred.

THE NEW GENERATION OF GENERATIVE AI IS PROOF POSITIVE OF [24]7.AI'S "HI+AI" APPROACH. HUMAN INTELLIGENCE IS VITAL FOR IMPROVING ARTIFICIAL INTELLIGENCE, AND VICE VERSA.

> **Real-Time Conversational FAQ:** This is [24]7.ai's solution providing a dynamic version of a conversational FAQ. It enables a visitor to ask a question and present an answer from a preconfigured knowledge base of question-answer pairs. While this use case may seem to be a subset of the previous one, it poses a key challenge because there is no human-in-the-loop. Additional guardrails need to be present to prevent the LLM from hallucinating, or in general, producing incorrect information.

> **Intent Classification:** Many LLMs are proving to be excellent "few-shot" learners. They can produce accurate predictions with just a handful of examples and thus enable [24]7.ai to onboard clients quickly onto their current conversation solution AIVA. Solutions that combine fast training with an LLM's conversational abilities can engineer an automated backtracking system. For instance, an LLM might realize that there are multiple potential intents applicable at a point, and to pick one. Or it may automatically frame a disambiguation question for the visitor: "Are you looking for X, or Y?".

> **Simulation-Based Agent Training:** Agent turnover is a long-standing and expensive problem among contact center operators. When new agents are hired to support a client, a common high-effort activity is to train them to handle client-specific requirements. [24]7.ai has begun to look at the possibility of simulating visitors for specific intents and use cases. This way an agent can chat with such a simulator and upskill themselves quickly. This would reduce, or maybe even bypass, the need to conduct time-consuming training sessions.

Applications That Assist, Rather Than Replace, Humans

These represent a small sample of the many ways in which LLMs improve upon traditional implementations of solutions within [24]7.ai. In all cases, Conversational AI is put to use to enhance human performance rather than to replace humans. Yet each has a built-in business benefit that supports a business plan and brings strong potential to improve bottom-line results. Collectively, they speed up the time it takes to train both live and automated assistants, fine-tune results to assure relevance, and accelerate the workflows of customers, agents, and other employees.

About Opus Research

Opus Research is a diversified advisory and analysis firm providing critical insight on software and services that support digital transformation. Opus Research is focused on the merging of natural language understanding, machine learning, conversational AI, LLMs, conversational intelligence, intelligent authentication, and digital commerce. **www.opusresearch.net**