Paper ET07

Comprehensive evaluation of large language models (LLMs) such as ChatGPT in biostatistics and statistical programming

Michael Pannucci, Arcsine Analytics
Weiming Du, Alnylam Pharmaceuticals
Songgu Xie, Regeneron Pharmaceuticals
Huibo Xu, Greenwich High School
Toshio Kimura, Arcsine Analytics

ABSTRACT

Generative artificial intelligence using large language models (LLMs) such as ChatGPT is an emerging trend. However, discussions using LLMs in biostatistics and statistical programming have been somewhat limited. This paper provides a comprehensive evaluation of major LLMs (ChatGPT, Bing AI, Google BARD, Anthropic Claude 2) in their utility within biostatistics and statistical programming (SAS and R).

We tested major LLMs across several challenges: 1) Conceptual Knowledge, 2) Code Generation, 3) Error Catching/Correcting, 4) Code Explanation, and 5) Programming Language Translation. Within each challenge, we asked easy, medium and advanced difficulty level questions related to 3 topics: Data, Statistical Analysis and Display Generation. After providing the same prompts to each LLM, responses were captured and evaluated. For some prompts, LLMs provided incorrect responses, also known as "hallucinations". While LLMs replacing biostatisticians and statistical programmers may be overhyped, there are nevertheless use cases where LLMs are helpful in assisting statistical programmers.

INTRODUCTION

LLMs have garnered tremendous attention for their "knowledge" and "skills" across many areas. They have the ability to answer both broad and specific questions and seem to be knowledgeable about any topic. This begs us to ask if LLMs can answer biostatistics and statistical programming questions.

- Can they teach us about CDISC data standards?
- Can they write SAS and R programs?
- Can they fix our code or make it more efficient?
- Can they explain the code others wrote?
- Can it translate code from SAS to R?

One challenge and criticism the LLMs receive is that they sometimes generate "hallucinations" or incorrect answers. So, perhaps there's one more question to add to the list:

- Can we trust the responses that these LLMs generate?

To begin answering these questions, we need to perform a comprehensive, systematic evaluation of these LLMs to unpack their abilities while understanding their shortcomings before they can be widely deployed. With this understanding, users can take advantage of their power while being cognizant of their limitations.

Since these LLMs are framed as "chats", we can extend that analogy and think about who we are chatting with. If we were chatting with someone, we would want to know their experience and background so that we can put their responses into perspective.

- Are we chatting with someone who is an industry and subject matter expert? Or are we chatting with someone who knows some things but still lacks full knowledge?

- Is this a statistician/statistical programmer assistant or a statistician/statistical programmer replacement?

Finally, as we look into our crystal ball …

- Is this a novelty fad that will fade over time? Or will it have lasting impact on our industry?

- How will we integrate this into our workflow process?

METHODS

We devised a comprehensive and systematic approach to evaluating some of the most popular LLMs. In our evaluation, we considered the following LLMs in their utility in supporting biostatistics and statistical programming (SAS and R) work:

1) Open AI ChatGPT 3.5 (free version)
2) Open AI ChatGPT 4.0 (paid version)
3) Google BARD
4) Microsoft Bing
5) Anthropic Claude 2

We tested each of the LLMs across 5 challenges:

1) Conceptual Knowledge
2) Code Generation (SAS and R)
3) Error Correcting (SAS and R)
4) Code Explanation (SAS and R)
5) Programming Language Translation (SAS to R and R to SAS)

Within each challenge, we asked questions pertaining to 3 topics:

1) Data
2) Statistical Analysis
3) Display Generation

Finally, within each topic, we asked several questions at 3 difficulty levels:

1) Easy
2) Medium
3) Advanced

Across all of the LLMs, the evaluation process collected ~1,000 prompts and responses. These responses were captured and stored as markdown files to preserve the formatting. Additionally, the responses were also simultaneously stored in an Excel spreadsheet. Each response was then evaluated as to how well the response addressed the question.

Below are a few example prompts. Note that these are shortened and simplified prompts to provide a sense for the types of prompts that were used.

Conceptual knowledge		
Statistical Analysis	Easy	What statistical methods are used to analyze count data?
Display Generation	Medium	What figure should I use to summarize the time to first death data?
Code Generation		
Data	Easy	Generate SAS code to derive the quartiles of AGE and name the new variable AGEQUART and add it back to the ADSL dataset.
Statistical Analysis	Advanced	Generate SAS code to perform Mixed Model Repeated Measure analysis for Questionnaire Patient-Reported dataset ADQS to evaluate treatment effects by visit.
Error Fixing		
Display Generation	Easy	The following SAS code was supposed to generate a scatter plot of two continuous variables grouped by treatment with a regression line; however, there was an error in the code. Please fix the code.
Data	Advanced	The following SAS code was trying to derive the time to first discontinuation from the study and store the derived standard variables in dataset ADTTE; however, there was an error in the code. Please fix the code.
Code Explanation		
Statistical Analysis	Medium*	Please explain the following SAS code and please include a high-level summary of what the SAS code is trying to do.
Display Generation	Advanced*	Please explain the following SAS code and please include a high-level summary of what the SAS code is trying to do.
Language Translation		
Display Generation	Easy*	Please translate the following SAS code to R code.
Data	Medium*	Please translate the following R code to SAS code.

* Difficulty level is determined by the actual code that is submitted as part of the prompt

To test each prompt, we entered the various prompts into the LLMs all of which more or less look similar. They all have appearance of a chatbot. Below are screenshots of Open AI ChatGPT 3.5. Prompts are entered in an open text box, and the LLMs provide various responses based on those prompts. If code is generated by the LLM, the code is demarcated in a separate box as show below in Figure 1.

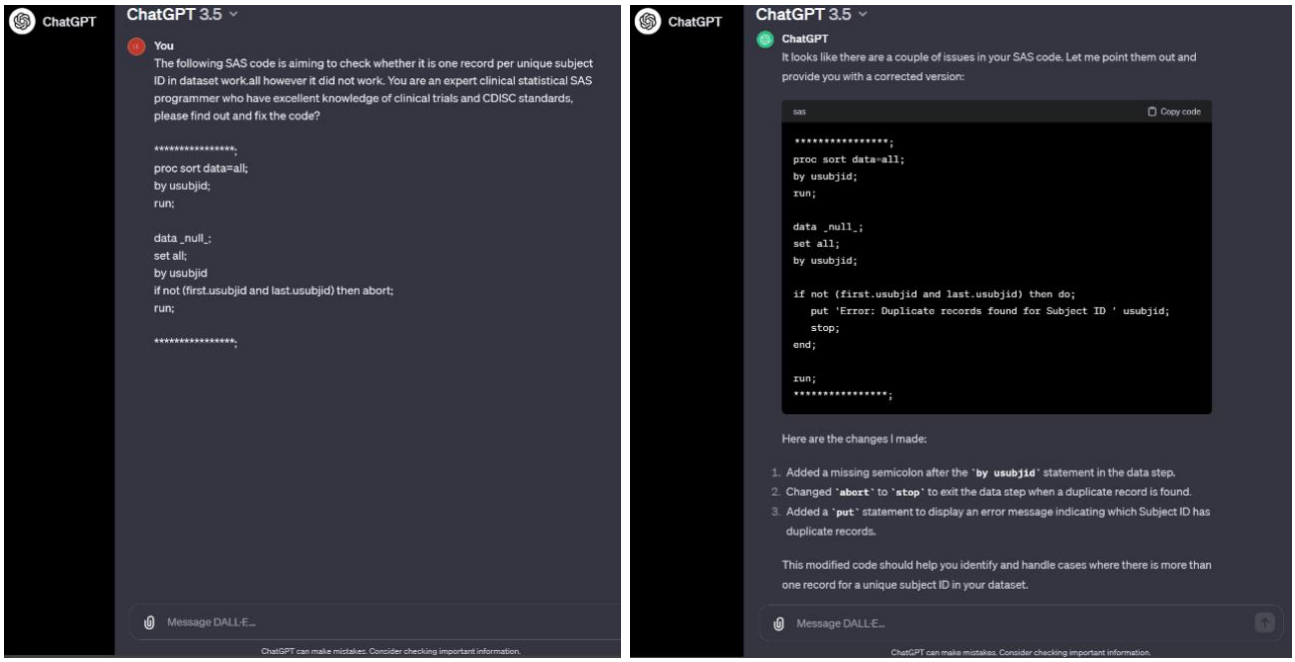


Figure 1: Screenshots of OpenAI ChatGPT 3.5

RESULTS

After reviewing ~1,000 prompts and response, there were several patterns that were identified through the evaluation process:

1) Easy/Medium Questions: All LLMs answered easy/medium questions correctly
2) Difficult/Unusual Questions: All LLMs failed in answering difficult/unusual questions
3) Differences across LLMs: Some LLMs perform better than others

Below are prompts and responses for each of the above observed patterns.

1) Easy/Medium Questions: All LLMs answered easy/medium questions correctly

All (or most) LLMs performed well when prompted with easy/medium questions. Below are a few examples and how the LLMs generally performed.

		Open AI ChatGPT 3.5	Open AI ChatGPT 4.0	Google Bard	Microsoft Bing	Anthropic Claude 2
Conceptual Knowledge Statistical Analysis	In clinical trials, what methods are used to analyze count data?	All LLM models provided the common methods that were used to analyze count data in clinical trials including Poisson regression, negative binomial regression, zero-inflated models etc.				

Error Fixing
Data
	The following SAS code is aiming to check whether there is one record per unique subject ID in dataset work; however, it did not work. Please fix the code.
	`**********;`
	`proc sort data=all;`
	`by usubjid;`
	`run;`
	`data _null_;`
	`set all;`
	`by usubjid`
	`if not (first.usubjid and last.usubjid) then abort;`
	`run;`
	`**********;`
	Missing semicolon
		All LLMs successfully fixed the syntax error.				

Code Generation Display Generation	Please write R code to create a scatterplot of age versus BMI. Please differentiate between two treatments.	All LLM models correctly created the R code.				

2) Difficult/Unusual Questions: All LLMs failed in answering difficult/unusual questions

All (or most) LLMs failed when presented with difficult or unusual questions. Below are a few examples and how the LLMs generally performed.

		Open AI ChatGPT 3.5	Open AI ChatGPT 4.0	Google Bard	Microsoft Bing	Anthropic Claude 2
Conceptual Knowledge Display Generation	When multiple binary outcomes that have nesting nature and are representing two directions of change, for example, proportion of > 5%, > 10%, >20% improvement and proportion of >5%, >10%, >20% worsening from baseline, what would be a good way to visualize those outcomes in one plot?	None of the LLMs suggest the divergent nested stack bar plot.				
Code Generation Display Generation	Given the CDISC datasets ADTTE, write SAS code to create a GTL template for Kaplan-Meier Curve with legends. Please also calculate hazard ratio and median and show them on the graph and axistable to show patients at risk.	None of the LLMs were able to successfully create a working GTL template to meet the requested requirements.				
Conceptual Knowledge Statistical Analysis	When patients can stay in the study after receiving rescue medication, what should be done in the analysis dataset in order to evaluate the treatment effect including visits after receiving the rescue medication?	Instead of suggesting censoring the collected data (set to missing) after taking rescue medication, the LLMs suggested removing patients entirely from the analysis if the patient received rescue medication.				

3) Differences across LLMs: Some LLMs perform better than others

There are however performance differences across the different LLMs. For some questions, different LLMs made different mistakes. While for others, they ranged from bad to good to great answers. The examples below will highlight some of these differences.

A) Hallucinations are a general problem. We have already seen poor performance earlier where all of the LLMs failed. There are different levels of failure and different ways of failing across the different LLMs. Below is an example where ChatGPT 3.5 made up a SAS option and another where ChatGPT 4.0 stated that the SAS procedure does something that the SAS procedure actually does not do.

		Open AI ChatGPT 3.5	Open AI ChatGPT 4.0

Code Generation
Statistical Analysis
	Please write SAS code to calculate the hazard ratio of treatment effects of Overall Survival from Time-to-Event dataset ADTTE, and also the logrank p-value.	/* Use PROC LIFETEST to perform survival analysis */
		proc lifetest data=ADTTE plots=survival;
		time TimeToEvent*Event(0);
		strata TreatmentGroup;
		logrank;
		hazardratio out=hazardratio_output;
		run;
		
		There's no hazardratio statement in PROC LIFETEST
			proc phreg data=ADTTE;
			class treatment(ref='control');
			/* Assuming 'treatment' is a variable in ADTTE */
			model time*status(0) = treatment / risklimits;
			/* 'time' and 'status' are typical names for these variables */
			hazardratio 'Treatment Effect' treatment;
			/* The log-rank test is automatically performed as part of the model fit */
			run;
			
			Log-rank test is not automatically performed

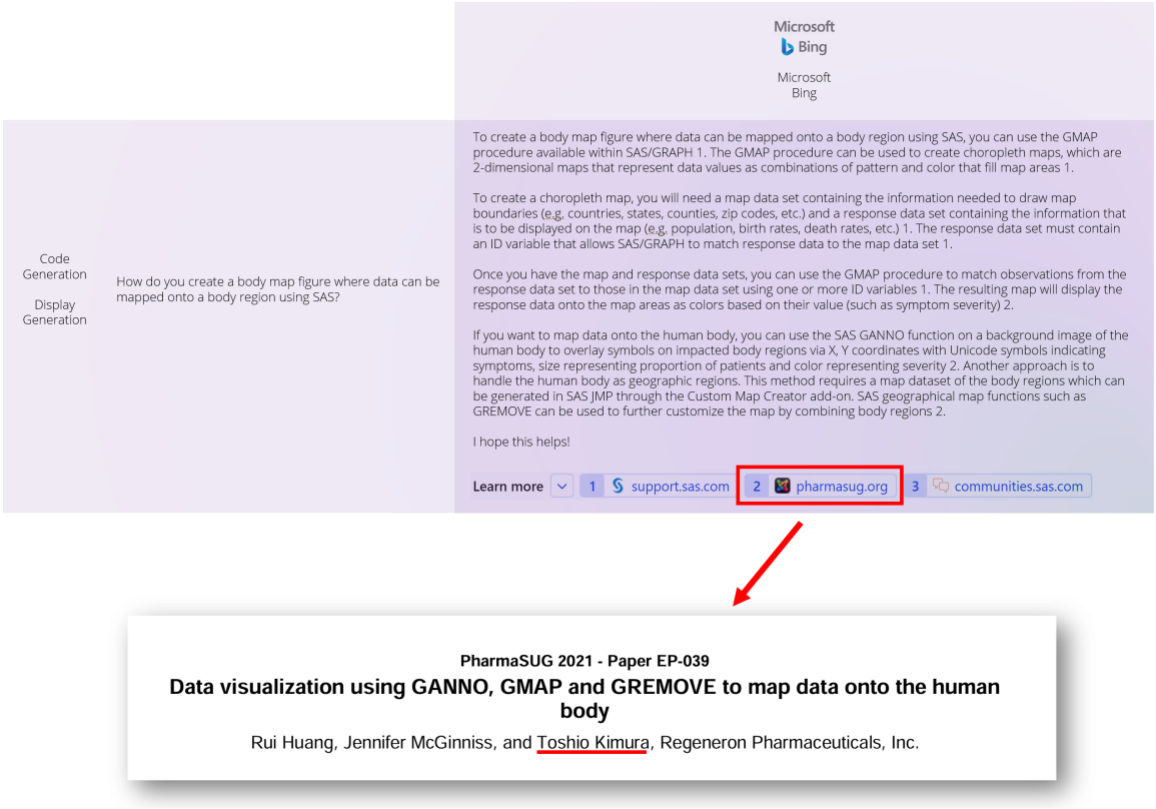
B) ChatGPT 4.0 outperforms ChatGPT 3.5. As expected, the higher paid tier of ChatGPT performs better than the free tier. Below are a few examples, and as one of the examples show ChatGPT 3.5 states that an obsolete R package should be used whereas ChatGPT 4.0 uses a different R package that is available for use.

		Open AI ChatGPT 3.5	Open AI ChatGPT 4.0

Code Translation
Data
	The LLMs were asked to translate specific SAS code (of calculating the quartile of Age) to R code:
	
	**********;
	proc univariate data=adsl noprint;
	var age;
	output out=pctlage p25=p25age p50=p50age p75=p75age;
	run;
	**********;
		# Calculate quartiles for the 'age' variable in the adsl dataset
		pctlage <- adsl %>%
		summarise(
		p25age = quantile(age, 0.25),
		p50age = quantile(age, 0.50),
		p75age = quantile(age, 0.75)
		)
			# Calculate the quartiles
			pctlage <- adsl %>%
			summarise(
			p25age = quantile(age, probs=0.25, na.rm = TRUE),
			p50age = quantile(age, probs=0.50, na.rm = TRUE),
			p75age = quantile(age, probs=0.75, na.rm = TRUE)
			)
			
			Addition of na.rm=TRUE allows for missing age

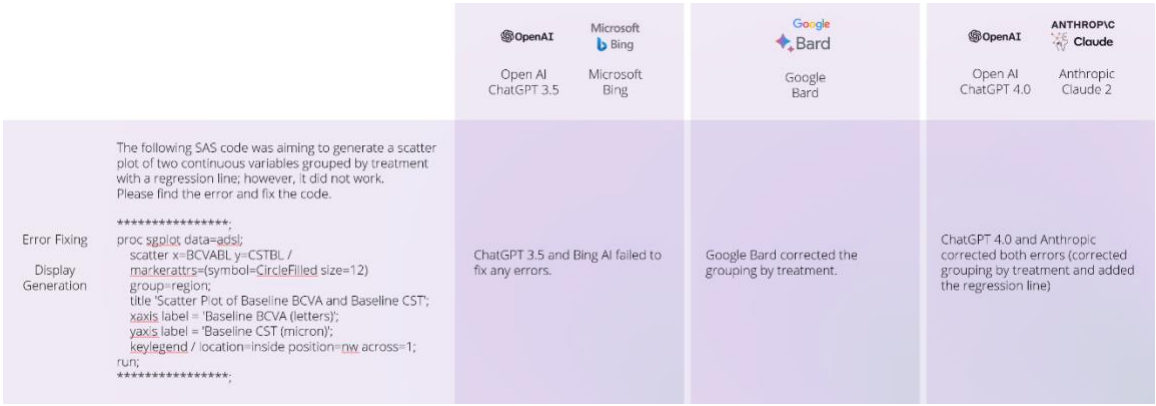
Conceptual Knowledge Data	In a clinical trial setting, endpoints are analyzed using observed cases, LOCF and WOCF to handle missing data. What should be done in the BDS analysis dataset to implement the above missing data handling?	ChatGPT 3.5 provided some SAS code that did not address the question.	ChatGPT 4.0 provided two appropriate suggestions including "Single Record Approach with Multiple Variables" and "Flag Variables".
Code Generation Display Generation	Please write R code to create a line plot of change in weight at different visits for subject from baseline to Week 10 and output the graph into an RTF output.	When it comes to generating RTF output, GPT 3.5 cites (ReporteRs), a package that's unfortunately been obsolete since 2018.	GPT 4.0 cites {officer} to generate Word output, which could access and manipulate Microsoft Word, RTF and Microsoft PowerPoint documents from R.

C) Microsoft Bing Chat (now Co-Pilot) includes references that were used as the basis for the response. Below is an example where the response referenced a paper by one of the co-authors of this paper. The LLMs are using all of the information available on the web, and this example illustrates that the knowledge that we create and make available are also being ingested and used by the LLMs as part of its training data.



D) LLMs fixed either none, one or both errors

There were many prompts where the differences between the LLMs can be observed some of which were already explained above. The Error Fixing example below further highlights these differences. The Error Fixing example contains 2 errors. ChatGPT 3.5 and Microsoft Bing did not fix any of the errors. Google Bard fixed one of the errors, and both ChatGPT 4.0 and Anthropic Claude 2 fixed both errors.



E) Anthropic was consistently one of the better LLMs

Anthropic consistently responded with better responses. As the examples below illustrates, Anthropic Claude 2 either uniquely suggested different options such as suggesting a body map as a data visualization method or provided responses to specifically address the situation as opposed to a more generic approach that didn't address the specific issue in the prompt.

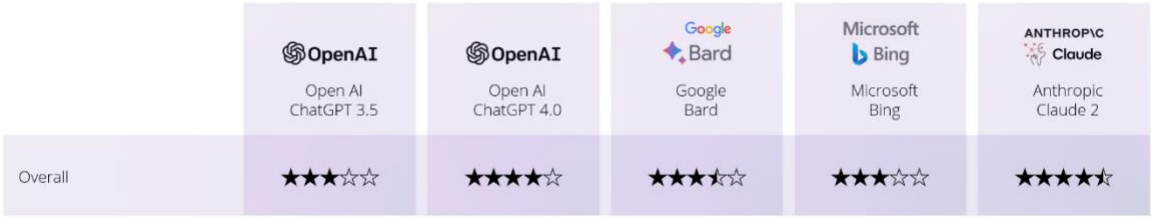
		Open AI ChatGPT 3.5 / Open AI ChatGPT 4.0 / Google Bard / Microsoft Bing	Anthropic Claude 2
Conceptual Knowledge Display Generation	Very often, we analyze outcomes of severity, symptoms, and location in a human body through tables after tables. Can we visualize those outcomes in a plot? What plot can it be?	None of the other LLMs suggested the Body Map as one of the options.	Anthropic suggested using Body Map: Body maps: Display an outline diagram of a body and use size/color of icons located on different regions to encode outcomes. Allows anatomical orientation.
Conceptual Knowledge Statistical Analysis	Suppose that a clinical trial enrolled half treatment naive patients and half patients treated with another approved drug that has similar mechanism to the study drug. The primary endpoint showed statistically significant results compared to the control group however the clinical significance is borderline. There are questions around the efficacy of the study drug. If you are the lead statistician, what analyses can be done to look at the treatment effect in a different angle?	The other LLMs suggested the use of general subgroup analysis (not specifically for treatment naive versus previously treated).	Anthropic specifically suggested subgroup analysis comparing treatment naive vs previously treated as well as suggesting forming binary outcomes (i.e. >30% improvement). In addition, unlike other LLMs that listed all possible analyses in a neutral tone (like a biostatistician with some experience), Anthropic was most specific in pointing out what analyses should be performed and why (like a more experienced biostatistician).

CONCLUSION

After a comprehensive and systematic review of ~1,000 prompts and responses, the following ratings were assigned to the different LLMs for each of the challenges. The challenge level ratings are a summary of the different topics and difficulty levels. The coding prompts were easier to rate since there are right and wrong answers. However, we recognize that there is some level of subjectivity in evaluating conceptual responses. Nevertheless, multiple co-authors evaluated these independently and came to similar conclusions.

	OpenAI Open AI ChatGPT 3.5	OpenAI Open AI ChatGPT 4.0	Google Bard Google Bard	Microsoft Bing Microsoft Bing	ANTHROPIC Claude Anthropic Claude 2
Conceptual Knowledge	★★★★☆	★★★★☆	★★★★☆	★★★½☆	★★★★★
Code Generation	★★★½☆	★★★★☆	★★★☆☆	★★★½☆	★★★★★
Error Fixing	★★☆☆☆	★★★☆☆	★★★☆☆	★☆☆☆☆	★★★½☆
Code Explanation	★★★★☆	★★★★½	★★★★½	★★★½☆	★★★★★
Language Translation	★★★★☆	★★★★½	★★★★½	★★★½☆	★★★★½

Overall, Anthropic Claude 2 performed the best among the five LLMs based on our evaluations. It consistently performed well across the challenges, topics, and levels of difficulty.



All of the LLMs did surprisingly well in many of these challenges, but at the same time not good enough. Based on their current performance, they are not good enough to replace a statistician or a statistical programmer. However, they are good enough to serve as an assistant to develop the first draft after which the human statistician or statistical programmer will review, fix and deploy. LLMs can serve as an effective brainstorming partner but not good enough to be a decision maker (for now).

As the difference between ChatGPT 3.5 and ChatGPT 4.0 illustrates, the LLMs are constantly improving. Therefore, we should be planning based not on how the LLMs are currently performing but rather how well they may perform in the future.

There are also ways to minimize risk. One of our observations was that the LLMs performed well in easy/medium questions. Therefore, we can limit the role of LLMs to these easy questions. We can use LLMs to automate routine tasks to increase operational efficiency.

While using LLMs as an assistant is already a great help, the integration of LLMs into an automated workflow would be even better. LLMs can be integrated into a system that will not only generate but also execute code. This level of integration and automation may truly change how the pharmaceutical industry performs statistical analysis.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name:	Michael Pannucci
Company:	Arcsine Analytics
Email:	michael.pannucci@arcsineanalytics.com

Author Name:	Toshio Kimura
Company:	Arcsine Analytics
Email:	toshio.kimura@arcsineanalytics.com

Brand and product names are trademarks of their respective companies.