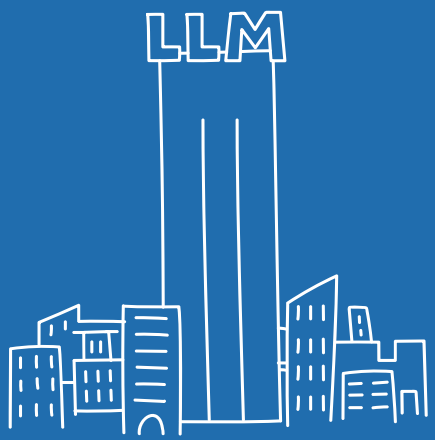
Teza

White paper – May 2023

The Landscape of Large Language Models

Part III

Misha Malyshev, Alexey Goldin, Qiang Zheng



This white paper represents the third installment in our transformer series. Our first installment delved into the history of machine learning and the emergence of transformers in finance. Our second installment, slightly more technical in nature, concentrated on the properties of GPT models while exploring the reach and limitations of ChatGPT. In this paper, we aim to provide a concise overview of the rapidly expanding realm of Large Language Models (LLM). This paper is designed for practitioners in the field and readers who seek to comprehend the landscape of this domain.

If you ask an LLM how to choose an LLM, it will give you an answer such as the following:

> Choosing the best large language model is a lot like speed dating: you ask it a few questions, hope it impresses you with its wit and intelligence, and then decide if you want to spend the rest of your project together!
>
> **GPT-4**

Part I: Large number of Large Language models

Given the recent surge in popularity of GPT models, new Large Language Models are emerging in the market nearly every week. As a class, they're distinguished from their predecessors by one core thing: size.

Classification dimensions:

1. **Training data size and source:**
   Distinct LLMs may utilize different training datasets, which can influence their performance and capabilities. For instance, some LLMs might rely on extensive, publicly available, open-source text datasets, whereas others, like those from OpenAI or Meta, might employ high-quality, proprietary datasets.

   Here are a few examples:
   a. Bloom: 1.6 terabytes (TB) of public datasets and 341 billion tokens
   b. LLaMA: roughly 5 TB of public datasets and 1.4 trillion tokens – including 82% from CommonCrawl and C4 (public data collected over years of web crawling), 18% from Github, Wikipedia, Arxiv, and books
   c. GPT-3: roughly 45 TB of public datasets and 500 billion tokens – including CommonCrawl, Webtext2, Wikipedia, Github, and books
   d. GPT-4: public datasets (combined from previous GPT models) + human annotated datasets – OpenAI hired 100+ employees to clean and annotate these datasets

   Why might some non-public datasets be better? First, improved data quality. A pre-trained model on a smaller yet higher-quality dataset can outperform a model trained with larger, mixed-quality data. Second, is the de-duplication of pre-training data – which prevents the pre-training model from memorizing or overfitting the same data multiple times – thereby enhancing the model's ability to generalize. Lastly, is dataset diversity. A diverse pre-training dataset may include domain diversity, format diversity (such as text, code, and tables), and language diversity. Diverse datasets further improve the performance and applicability of a model across many different scenarios.

2. **Model size and number of parameters:**
   Some LLMs may be larger than others, boasting more parameters and computational resources, which may enable them to perform better on certain tasks. However, this advantage may also entail longer training times and greater storage requirements. In 2020, OpenAI proposed a scaling law linking increased model size to improved model performance, suggesting that most of the budget should be allocated to scaling up the model. This paper directly sparked the trend of increasing model size.

   Nonetheless, indiscriminately increasing model size is not the optimal choice for enhancing model performance given limited budgets and memory constraints. In 2022, the DeepMind team published a paper comparing model size and training data, ultimately concluding that most language models are evidently **undertrained**. In other words, once the model size reaches a certain threshold, training models on larger datasets without increasing model size can yield significant benefits. This latest research indicates that, after extensive expansion, the newest LLMs tend to optimize existing large parameter sizes. Consequently, optimized "smaller models" (which can still possess tens of billions of parameters) can compete with models containing hundreds of billions of parameters across numerous tasks.

   Another important consideration pertains to the resources necessary for inference; the aforementioned papers do not account for this aspect. If the model is intended for deployment on a multitude of consumer devices with limited memory and computing resources, it may be advantageous to train smaller models on larger datasets. While these models may not perform as well as larger models with the same training compute budget, they could serve as better foundational models for offline use.

   Some examples for model sizing include:
   a. RoBERTa 2018: 300 million parameters (not really LLM, but a precursor)
   b. GPT-3 2020 and Bloom 2022:175 billion parameters (bigger is better!)
   c. LLaMA 2023: 65 billion parameters (smaller size but better performance!)

3. **Model structure and architecture:**
   Although all these models are based on the transformer architecture, they can employ different structures, training strategies, and hyper-parameter settings.

   Below are some examples of distinct architectures:
   a. GPT-style refers to a decoder-only autoregressive language model
   b. T5-style refers to an encoder-decoder language model
   c. GLM-style refers to the special model structure of GLM
   d. Multi-task refers to the model structure of ERNIE 3.0

4. **Supported tasks and applications:**
   Indeed, some LLMs are specifically designed to excel at particular tasks, while others are created for more general applications. Here are a few examples of specialized LLMs:
   a. Language translation: Models like Marian NMT, mBART, and T2T-ViT focus on translating text from one language to another while maintaining the original meaning and context
   b. Question-answering systems: BERT and its variants (e.g., RoBERTa, ALBERT) are often fine-tuned for question-answering tasks, enabling them to provide accurate and relevant responses to user queries
   c. Text generation: GPT and its successors (GPT-2, GPT-3) have been specifically designed for generating human-like text, making them ideal for tasks like content creation, summarization, and paraphrasing
   d. Text-to-image AI tools: Models like DALL-E and CLIP combine language understanding and image generation capabilities, enabling them to generate images from text descriptions or identify relevant images based on textual input

These specialized LLMs can outperform more general models in their respective domains, showcasing the importance of task-specific design and optimization. As research in the field progresses, it is likely that we will continue to see more LLMs tailored for specific applications, further enhancing their performance and utility.

Representative LLMs:

The competitive landscape of software-based technology companies vying to create the best LLM is rapidly evolving. The table below provides an illustration of this ongoing race, showcasing some of the key players and their respective contributions to the field:

Model Name	Publish time	Team	Language	Parameters	Tokens in training	Model style	Open-source
T5	2019-10	Google	English	13B	Unkown	T5-stvle	Yes
GPT-3	2020-05	OpenAI	Multiple languages	175B	300B	GPT-stvle	No
LaMDA	2021-05	Google	English	137B	2.8T	GPT-stvle	No
GPT-3.5	2021-06	OpenAI	Multiple languages	175B	Unkown	GPT-stvle	No
Jurassic	2021-08	AI21	English	178B	300B	GPT-stvle	No
MT-NIG	2021-10	Microsoft, NVIDIA	English	530B	270B	GPT-stvle	No
ERNIE 3.0 Titan	2021-12	Baidu	Chinese	260B	300B	Multi-task	No
Gopher	2021-12	DeepMind	English	280B	300B	GPT-style	No
Chinchilla	2022-04	DeepMind	English	70B	1.4T	GPT-style	No
PaLM	2022-04	Google	Multiple languages	540B	780B	GPT-style	No
OPT	2022-05	Meta	English	125M-175B	180B	GPT-style	Yes
BLOOM	2022-07	BigScience	Multiple languages	176B	366B	GPT-style	Yes
GLM-130B	2022-08	Tsinghua	English and Chinese	130B	400B	GPT-style	Yes
ChatGPT	2022-11	OpenAI	Multiple languages	173B	Unkown	GPT 3.5	No
LLaMA	2023-02	Meta	Multiple languages	7B-65B	1.4T	GPT-style	Yes
GPT-4	2023-03	OpenAI	Multiple languages	Unkown	Unkown	GPT-style	No
Alpaca	2023-03	Stanford	English	65B	Unkown	LLaMA	Yes
Bard	2023-03	Google	English	137B	Unkown	LLaMA	No

Picture 1: Comparison of recent LLMs

We could not mention all models here, as it seems that each day brings new ones. There are now models from Stability AI (those who gave us Stable Diffusion), Dolly from DataBricks, Cerebras-GPT from Cerebras.

While it's difficult to make a judgment on which is the best LLM yet, the GPT-4 model is widely considered to have better and more comprehensive performance than LLaMA and earlier fully open-sourced Bloom models. However, the GPT-4 model is completely closed and very large, so users can only access the product by paying to use the OpenAI servers. This makes open-source "small" models such as LLaMA and Alpaca, which can be deployed on small servers for enterprises and individuals, more popular. Moreover, users can also redevelop these open-source models based on their own needs, using their own data and better specializing in their downstream tasks.

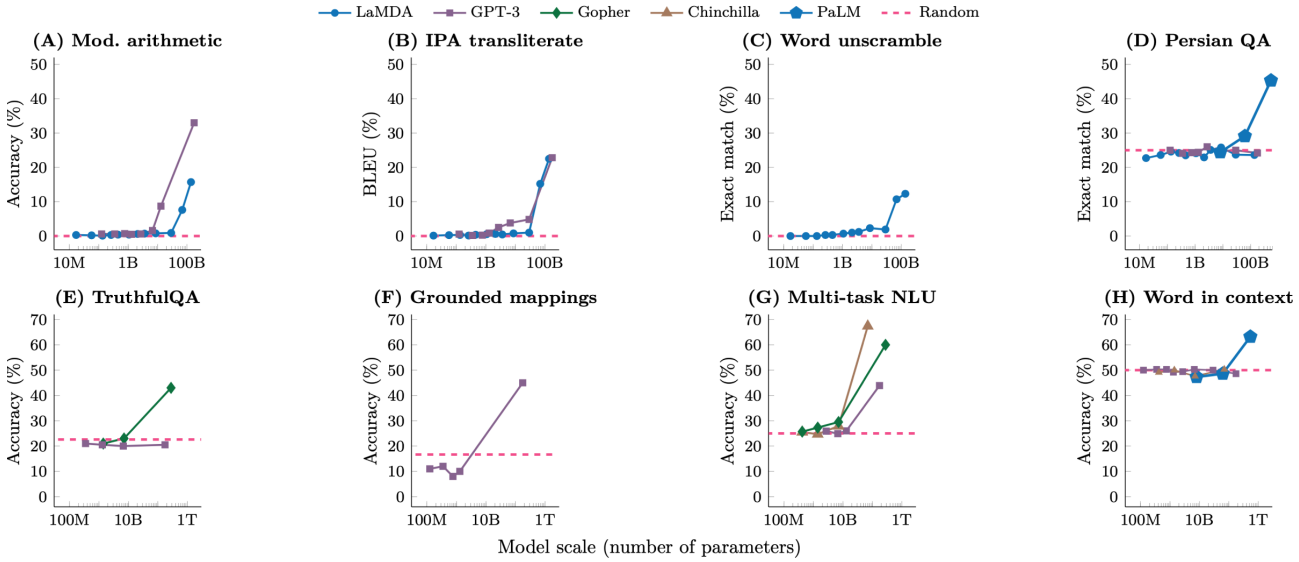
Part II: What makes these LLMs work so well?

As we have already discussed in our previous paper, there are certain characteristics of language transformer models that are available or even appeared with the increasing size of these models. Here we list the majority of them:

1. Emergent abilities

Perhaps the most surprising and least understood property of Large Language Models (LLMs) arises as the model size emerges. This often causes people to perceive these models as "thinking" or "creating" discussions about the emergence of Artificial General Intelligence (AGI).

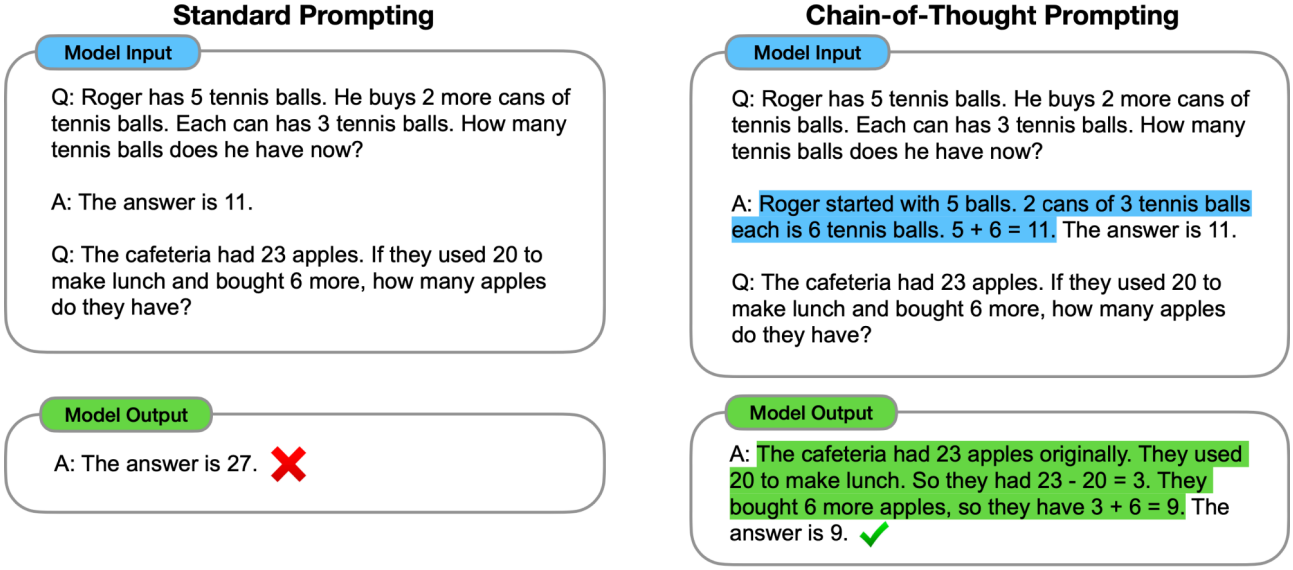
As the amount of training data expands and the number of model parameters surpasses a specific threshold, the model's performance experiences a sudden, significant improvement, ultimately exceeding the predicted scaling law. This phenomenon is described in the research paper, "Emergent Abilities of Large Language Models" (2022). The following illustration demonstrates this concept using various statistics and a range of LLMs.



Picture 2: Eight examples **of emergence in the few-shot prompting setting**

2. Prompt learning and chain of thought training

It is widely acknowledged that advanced LLMs, such as GPT-4, have developed the ability to employ a chain of thought process that deconstructs multi-step problems into separate, solvable intermediate steps. When tackling complex reasoning tasks, the generated thought chains mimic the human cognitive process. Although GPT-4 and similar models lack true consciousness or thinking capabilities, their use of thought chains resembling human reasoning significantly enhances their performance in reasoning tasks, overcoming the plateau effect of fine-tuning. GPT-4, with its capacity to create multimodal thinking chains, possesses a certain level of logical analysis ability, transcending the traditional vocabulary probability approximation model. Here is an example of a chain of thought:



Picture 3: Chain-of-thought **reasoning processes.**

Prompt learning refers to the design of a series of questions or tasks based on specific goals and contexts, in order to use large models to generate coherent and meaningful text related to a topic or subject area. The goal of Prompt Engineering is to improve the quality and relevance of generated text by carefully designing prompts to elicit the desired responses from the model. Prompt Engineering is closely related to the generation of thinking chains and is the theoretical basis for current natural language programming.

In a figurative sense, the training approach for models preceding GPT-3 involved large-scale text pre-training combined with local data fine-tuning. Prompt learning can be likened to a teacher guiding a student's response during a Q&A session, significantly reducing the reliance on data and manual labeling during the fine-tuning phase. Recent research, such as auto-prompt, has started to investigate enabling machines to automatically search for suitable prompt questions and answers using Masked Language Models (MLMs). This further minimizes human effort in crafting prompt questions and enhances unsupervised learning. Auto-prompting serves as a lightweight alternative to fine-tuning a model.

Below are some examples of prompting:

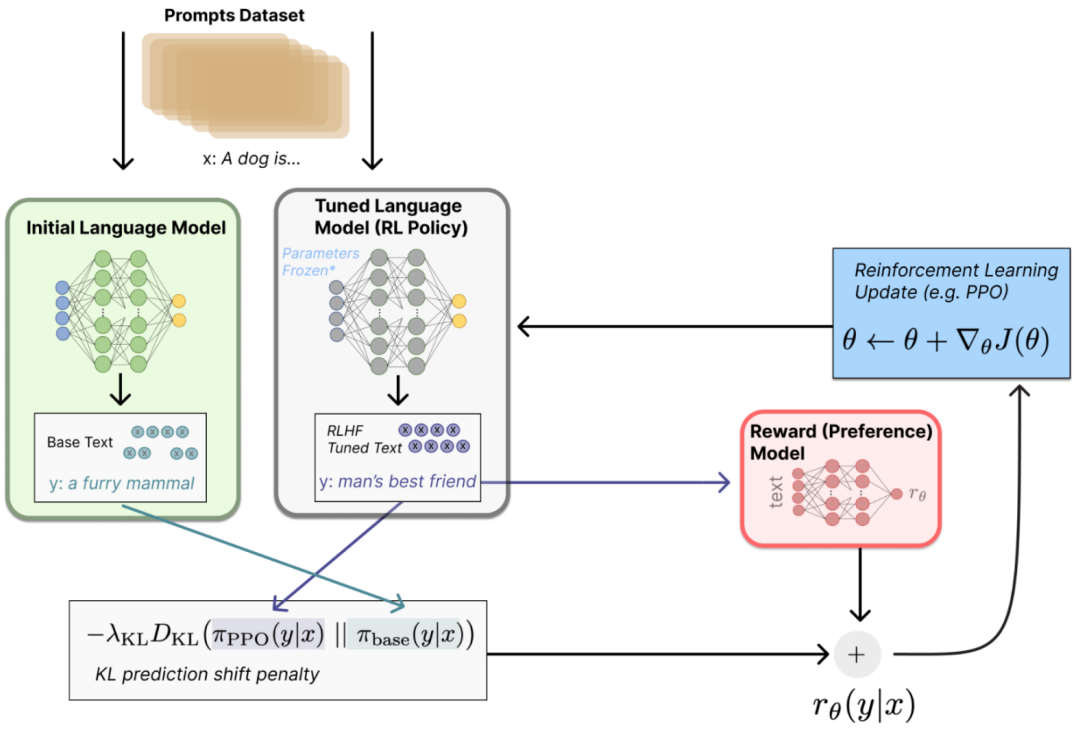
Name	Notation	Example	Description
Input	$\boldsymbol{x}$	I love this movie.	One or multiple texts
Output	$\boldsymbol{y}$	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(\boldsymbol{x})$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input $\boldsymbol{x}$ and adding a slot `[Z]` where answer $\boldsymbol{z}$ may be filled later.
Prompt	$\boldsymbol{x}'$	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input $\boldsymbol{x}$ but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z})$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	$\boldsymbol{z}$	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

Picture 4: Terminology and notation of prompting methods. z* represents answers that correspond to true output y*

3. Reinforcement Learning from Human Feedback (RLHF)

One other difference between GPT-4/3.5 and GPT-3 (in addition to size) is that a new technology called RLHF (Reinforcement Learning from Human Feedback) has been added. This training paradigm enhances human modulation of the model output intent and provides a more interpretable ranking of the results. Through RLHF technology, the model can prioritize high-quality answers, ensuring that its output is beneficial to humans and contributing to the model's safety. Furthermore, RLHF plays a crucial role in maintaining on-topic, multi-turn conversations. Ultimately, RLHF can help the model converge more rapidly, substantially reducing the time and resources required for each training session.

In simple terms, RLHF incorporates human input into the reward function and fine-tunes the language model using reinforcement learning. In practical implementation, human annotators assume the roles of users and AI assistants in a dialogue. They provide dialogue samples for the model to generate responses, and then the annotators rank the response options by scoring them, offering feedback to the model. The model learns from both types of feedback—human reinforcement and model prediction—as a unified system. It fine-tunes the model through reward policies and continues to iterate through the process.



Picture 5: Illustrating Reinforcement Learning from Human Feedback (RLHF)

Part III: Specific industry (Finance) LLM applications

An obvious question that every industry practitioner might ask is: "Okay, these LLMs are impressive conversationalists, but how can they help me do my job? In our case, can they help a trader make better trades?" As discussed in our second paper, LLMs don't truly possess a thinking brain and merely (though extremely impressively) predict the next word in a sentence. If they were trained on sentences specifically about finance, would their performance on the topic improve?

On March 30, 2023, Bloomberg introduced a large language model named BloombergGPT, specifically designed for the financial industry, demonstrating the application of LLMs in the financial vertical field. Based on the results, BloombergGPT has outperformed GPT-3 level LLM models in benchmark financial tasks while using only 1/3 of the parameters, achieved by retraining with a substantial amount of financial training datasets (refer to the table from Bloomberg's paper below). This evidence supports the notion that enhanced downstream performance can be obtained through high-quality, specialized data.

	BLOOMBERGGPT	GPT-NeoX	$\mathrm{OPT}_{66\mathrm{B}}$	$\mathrm{BLOOM}_{176\mathrm{B}}$
ConvFinQA	**43.41**	30.06	27.88	36.31
FiQA SA	**75.07**	50.59	51.60	53.12
FPB	**51.07**	44.64	48.67	50.25
Headline	**82.20**	73.22	79.41	76.51
NER	60.82	**60.98**	57.49	55.56
All Tasks *(avg)*	**62.51**	51.90	53.01	54.35
All Tasks *(WR)*	**0.93**	0.27	0.33	0.47

Picture 6: Results **on financial domain tasks.**

Indeed, as observed in Part II, an LLM's performance improves with more parameters and more data. However, the results from Part III also suggest that specialization contributes positively. A logical conclusion to draw from this is that a combination of both—increasing parameters and data while focusing on specialization—would likely yield the best results for an LLM's performance.

Perhaps we can learn a lesson from another area – image generating neural networks, such as DALL-E (OpenAI), Imagen (Google), MidJourney, Stable Diffusion (Stability AI). Of these models Stable Diffusion is the smallest and, out of the box, has the lowest quality of generated images. However, lack of restrictions and ability to run on consumer hardware made thousands of fine tuned versions possible. In specific domains, such as portraits, photorealistic images, anime, or styles inspired by particular artists, hand-tuned Stable Diffusion models sometimes rival the quality of larger, more general models. Impressively, these accomplishments are driven by enthusiastic fans working without institutional support.

It is indeed possible that we will see similar dynamics with language models. Larger cloud-based models like GPT-4 will consistently be more intelligent. However, smaller models with more permissive licenses can be fine-tuned, for instance, on a company's internal documentation that cannot be shared with OpenAI or tailored to a specific domain. With relatively modest effort, the LLaMa derivative model, Vicuna, was taught to reason about images (a capability GPT-4 possesses, but is not yet publicly available). Thus, in the future, users may have the option to choose more intelligent, highly capable, and expensive-to-run large models or turn to cheaper-to-run, more flexible derivatives of smaller, less restrictive models.

Conclusion

A revolution is happening in the world of language-based transformer models. Large language models continue to increase in size and improve performance, surprising even their creators. There is no doubt that the proliferation of these models and their disruption to multiple industries (including finance) will only accelerate in the near future. Different models, based on their large "parents" will be further fine-tuned for specific tasks, performing even better than the generic ones that exist today.

Disclosures

This document is provided solely for informational and educational purposes, and there is no consideration given to the specific investment needs, objectives, or tolerances of any recipient. This document is not investment research and should not be treated as such, nor does it represent a formal or official view of Teza. Additionally, Teza's investment positions may, and often will, vary from its conclusions discussed herein based on any number of factors, including client investment guidelines and restrictions. No representation is given that any statements made in this document are accurate or that Teza's objectives will be achieved. This document contains Teza's opinions, and such opinions are subject to change without notice.

This document does not constitute an offer to sell or the solicitation of an offer to purchase any security or investment product (each, a "Product") and should not be relied on in making any investment decision. Any such solicitation or offering may only be made by means of delivery of an approved offering document and relevant subscription documents, all of which must be read in their entirety. No offer to purchase shares in a Product will be made or accepted prior to receipt by the offeree of such documents and the completion of all appropriate documentation. No offer to sell (or solicitation of an offer to buy) will be made in any jurisdiction in which such offer or solicitation would be unlawful.

It should not be assumed that investments described herein will be profitable. Nothing described herein is intended to imply that an investment with Teza is safe, conservative, risk free or risk averse. An investment with Teza entails substantial risks, and a prospective investor should carefully consider the summary of risk factors included in Teza's Form ADV Brochure (and the relevant offering document) in determining whether an investment with Teza is suitable. The risk of loss in trading futures is substantial. This document does not consider the specific investment objective, financial situation or particular needs of any investor and an investment with Teza is not suitable for all investors. Prospective investors should not rely upon this document for tax, accounting or legal advice. Prospective investors should consult their own tax, legal, accounting or other advisors about the issues discussed herein. Historic market trends are not reliable indicators of actual future market behavior or future performance of any particular investment which may differ materially, and should not be relied upon as such. Investors are also reminded that past performance should not be seen as indication of future performance and that they may lose the entirety of their investment. No recommendation is made positive or otherwise regarding individual securities, futures, strategies or other investment products mentioned herein. Information provided about positions, if any, and attributable performance is intended to provide a balanced commentary, with examples of both profitable and loss-making positions; however, this cannot be guaranteed. Certain data and analyses contained herein are based on theoretical and/or backtested model portfolios and are not representative of the performance of accounts that Teza currently manages. The information provided herein is not intended to provide a sufficient basis on which to make an investment decision, and investment decisions should not be based on simulated, hypothetical or illustrative information that have inherent limitations. Unlike an actual performance record, simulated or hypothetical results do not represent actual trading or the actual costs of management and may have under or over compensated for the impact of certain market risk factors. Teza makes no representation that any account will or is likely to achieve returns similar to those shown. Gross performance results do not reflect the deduction of investment advisory fees, which would reduce an investor's actual return. There can be no assurance that any Product advised by Teza will implement the strategies or trading signals referred to herein, or that if implemented any such strategies or signals achieve their investment objectives.

Certain information contained in this document constitutes "forward-looking statements," which can be identified by use of forward-looking terminology such as "may," "will," "should," "expect," "anticipate," "project," "target," "estimate," "intend," "continue," or "believe" or the negatives thereof or other variations thereon or other comparable terminology. Such statements are based on the current expectations and certain assumptions of Teza, and are, therefore, subject to certain risks and uncertainties. A variety of factors, many of which are beyond Teza's control, affect the operations, performance, business strategy and results of the accounts that Teza manages and could cause the actual results, performance or achievements of such accounts to be materially different from any future results, performance or achievements that may be expressed or implied by such forward-looking statements or anticipated on the basis of historical trends.

See disclosures on pp. 9-10

Tables, charts and commentary contained in this document have been prepared on a best efforts basis by Teza using sources it believes to be reliable, although it does not guarantee the accuracy of the information on account of possible errors or omissions in the constituent data or calculations. Further, the information herein may be superseded by subsequent market events or for other reasons. Teza does not assume any duty to, nor does it undertake to, update the information herein. Charts and graphs provided herein are for illustrative purposes only. No part of this document may be divulged to any other person, distributed, resold and/or reproduced without the prior written permission of Teza.

* * *