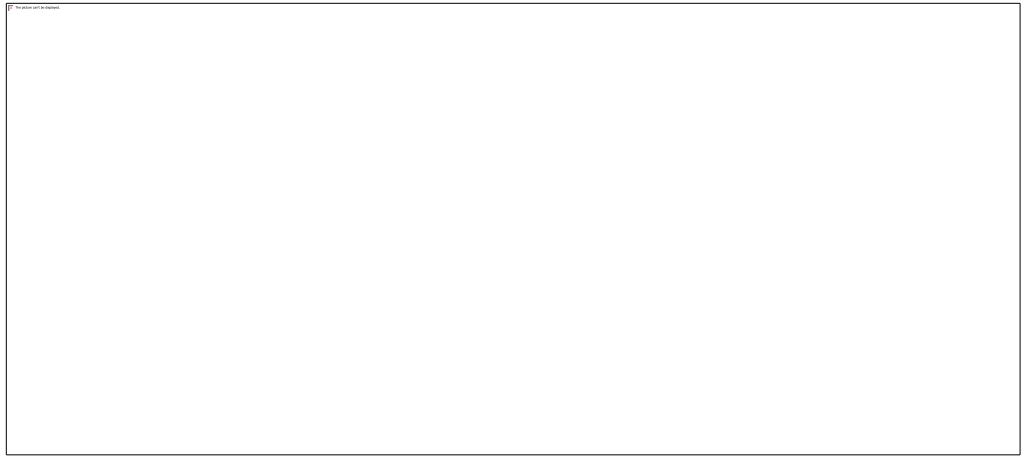
Use of LLMS for extracting statistical data from economic texts

Presenter: Issoufou Seidou Sanda

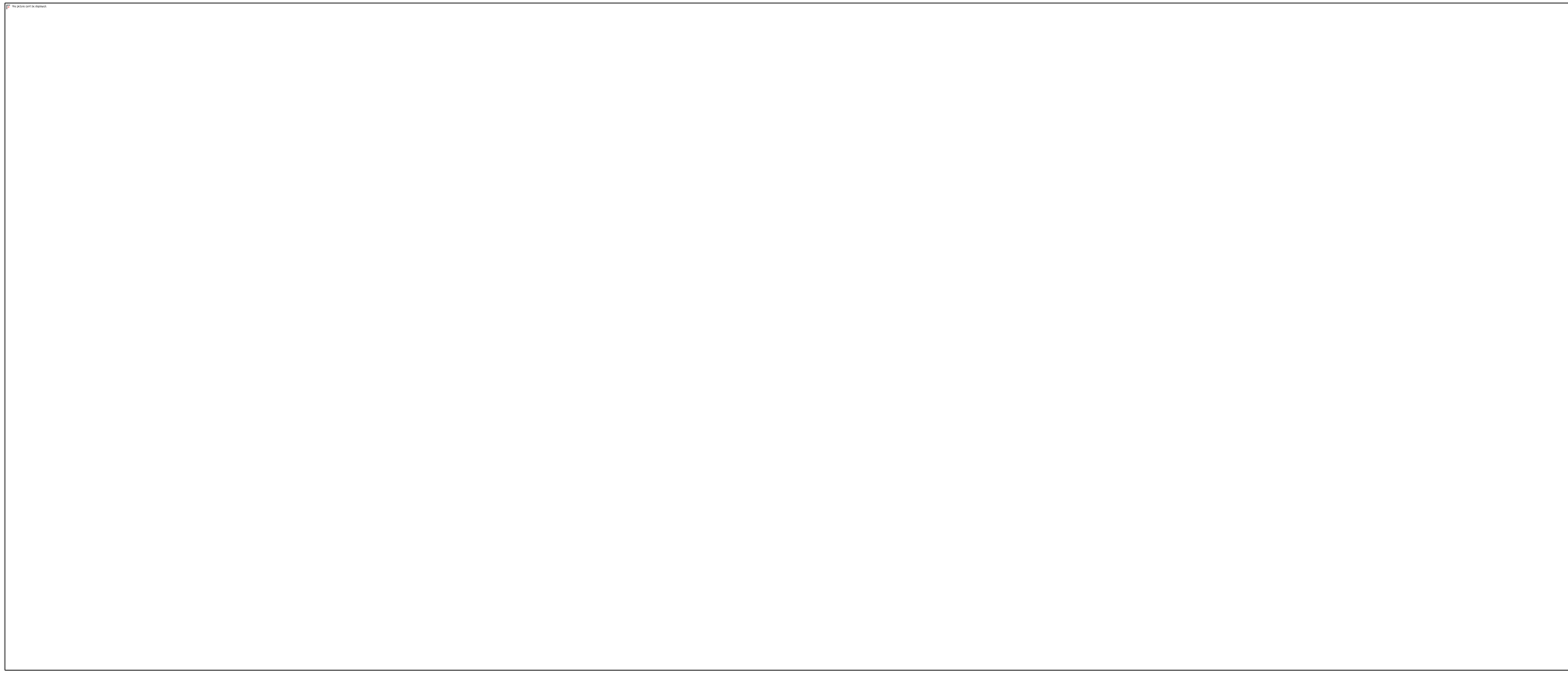
EIASS, ECA-ACS

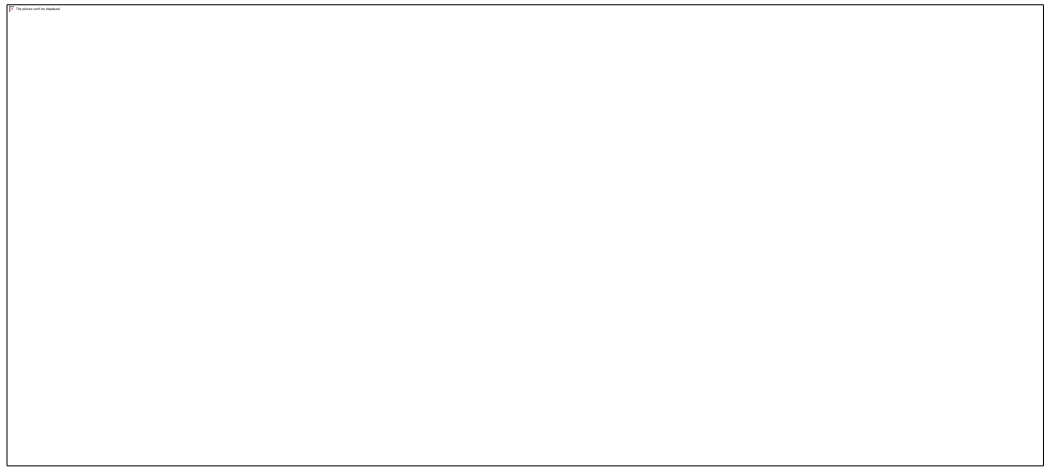
One of the most remarkable striking features of Large language models (LLMs) is their ability to understand the complexity and nuances of human language.

It makes it easy to extract useful information from unstructured format like text even when the information is hidden behind a very complex wording.



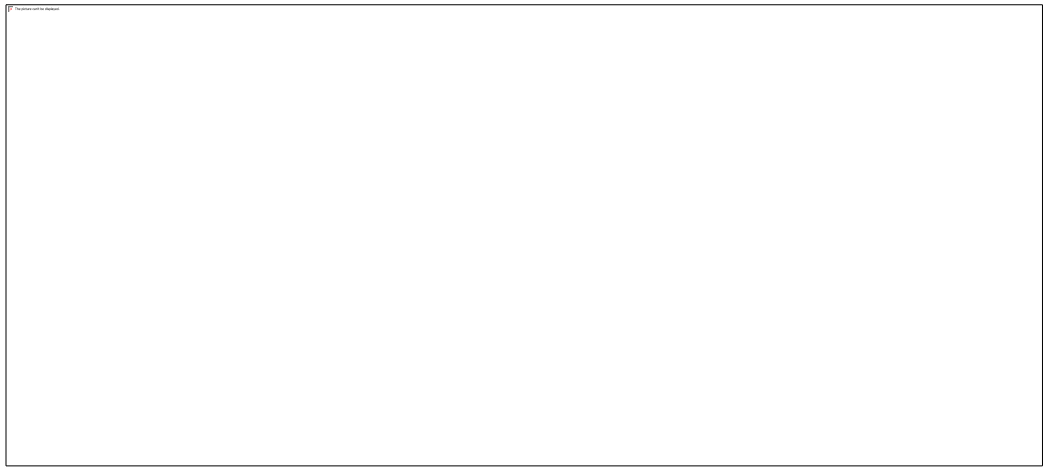
One of the most remarkable striking features of Large language models (LLMs) is their ability to understand the complexity and nuances of human language.



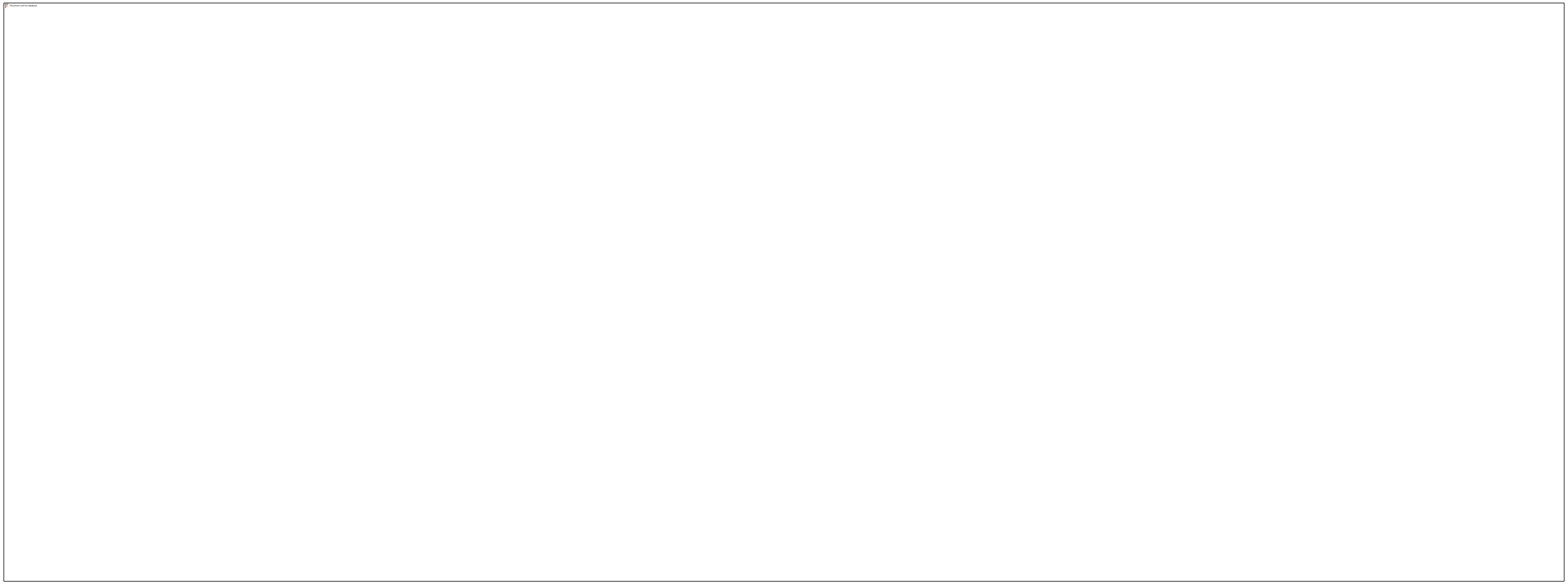


One of the most remarkable striking features of Large language models (LLMs) is their ability to understand the complexity and nuances of human language.

Based on the text above, give as a bulleted list what is threatening peace in Africa. Please limit yourself to the information given in the text above.



One of the most remarkable striking features of Large language models (LLMs) is their ability to understand the complexity and nuances of human language.



PRESS RELEASE

Ghana, April 2023
CONSUMER PRICE INDEX AND INFLATION

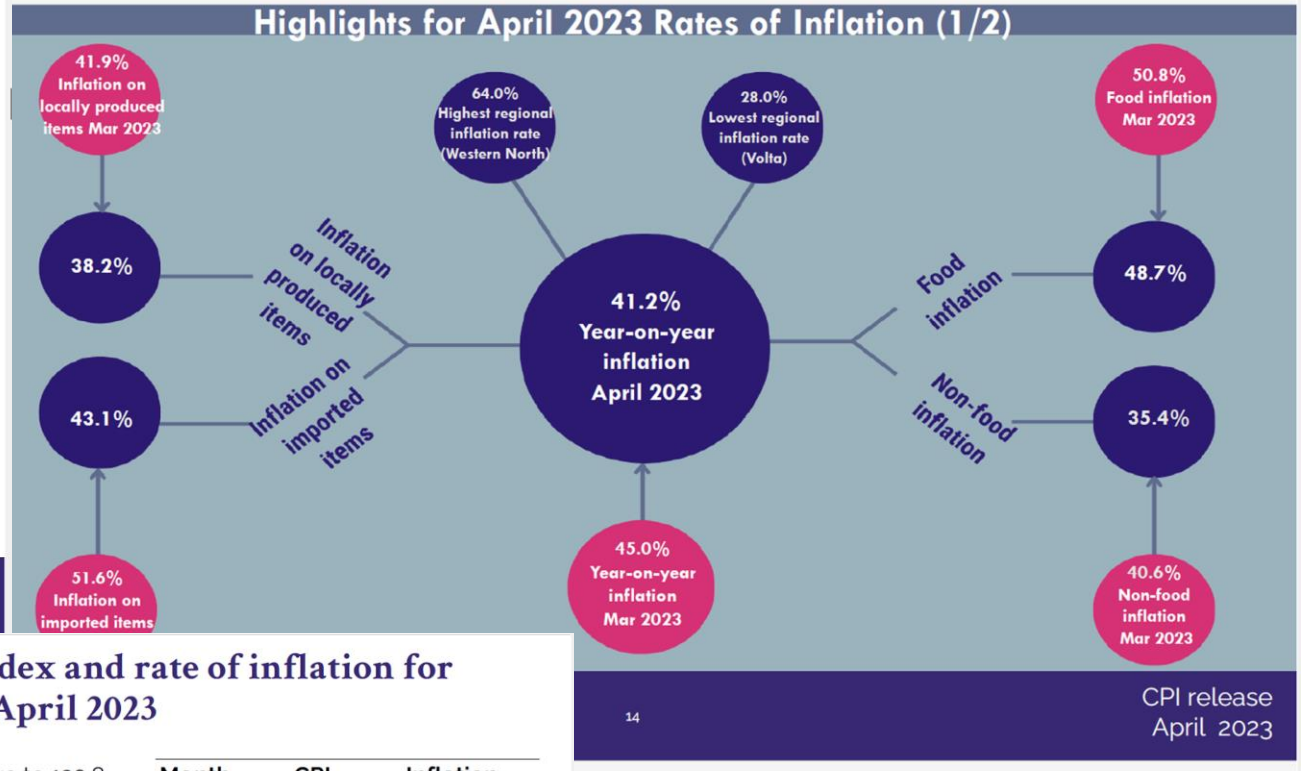
10th May 2023

GHANA
STATISTICAL SERVICE

Can LLMs understand these complex structures and automatically extract these indicators with their dimensions?

Definition and measurement of CPI and rate of inflation (3/3)

- Price collection is done in **57** markets
- Prices are collected from about **8,337** outlets.
- Prices are collected for **47,877** products every month from 16 regions.
- Products are ordered in a hierarchy of 13 Divisions, 44 Groups, 98 Classes, 156 Subclasses and 307 Items.
- Every Item can only be part of one Subclass, and every Subclass can only be part of one Class, etc.

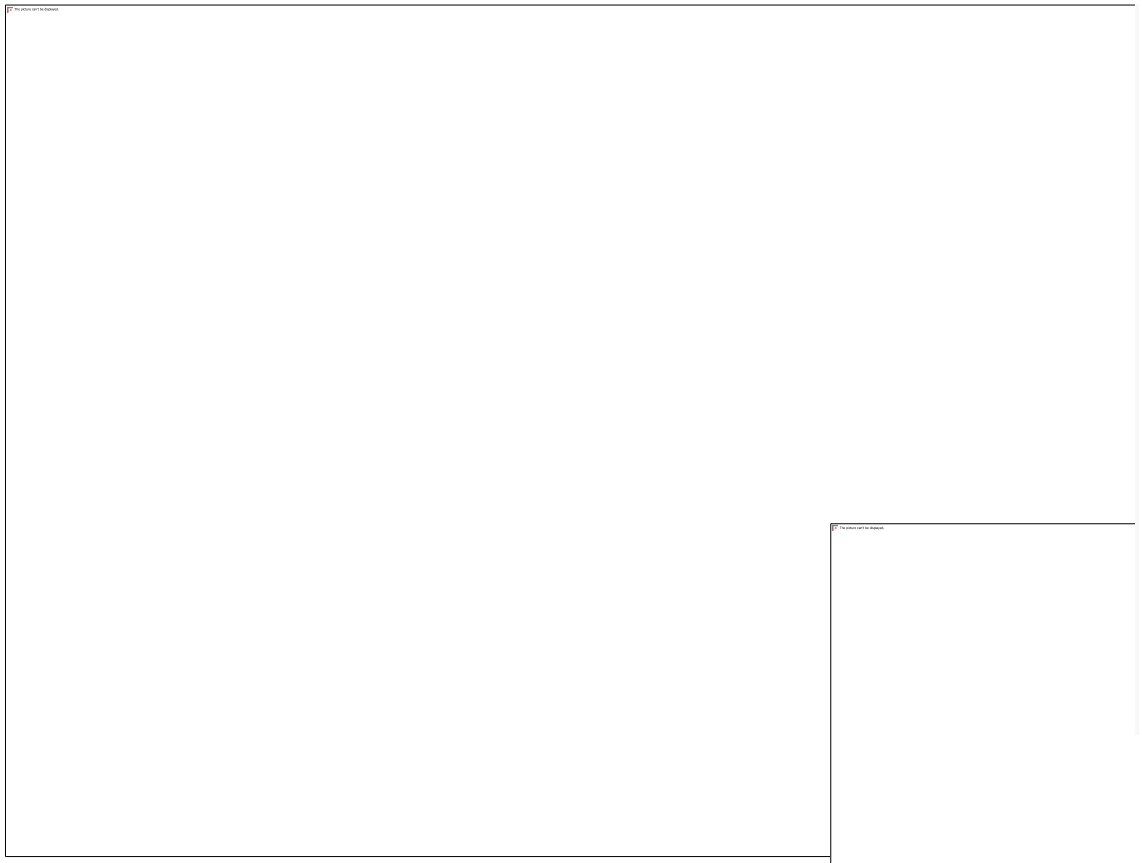


Consumer Price Index and rate of inflation for April 2023

- CPI for April 2023 was 170.5 relative to 120.8 for April 2022 using the linked series
- Year-on-year inflation rate for April 2023 was 41.2%
- This means that in the month of April 2023 the general price level was 41.2% higher than April 2022
- Month-on-month inflation between March 2023 and April 2023 was 2.4%

Month	CPI	Inflation	
		Monthly	Yearly
Apr- 2022	120.8	5.4%	23.6%
Oct-2022	144.4	2.7%	40.4%
Nov-2022	156.8	8.6%	50.3%
Dec-2022	162.8	3.8%	54.1%
Jan-2023	165.6	1.7%	53.6%
Feb-2023	168.7	1.9%	52.8%
Mar-2023	166.6	-1.2%	45.0%
Apr-2023	**170.5**	**2.4%**	**41.2%**

The data extraction program



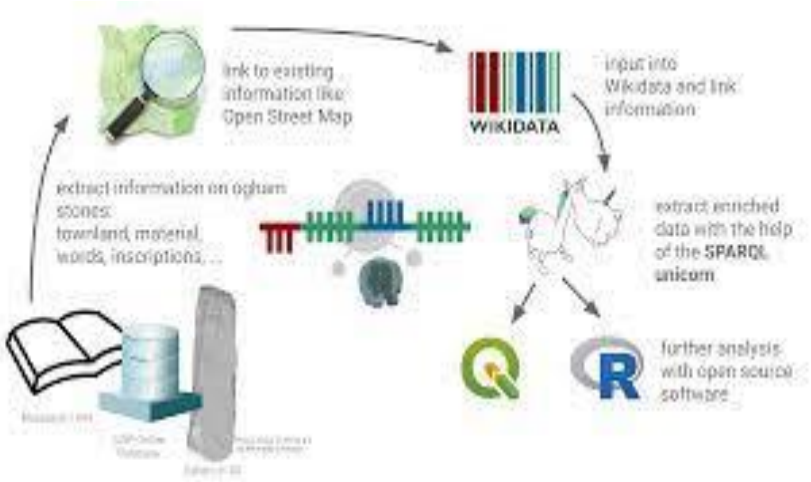
Disaggregation of May 2023 Rate of Inflation

- Food inflation (0.437) was 51.8%
  - Last month's Food inflation was 48.7%
  - Month-on-month Food inflation was 6.2%.

- Non-food Inflation (0.563) was 34.6%
  - Last month's Non-Food inflation was 35.4 %
  - Month-on-month Non-Food inflation was 3.5%

- Inflation for locally produced items was 36.2%
- Inflation for imported items was 43.8%

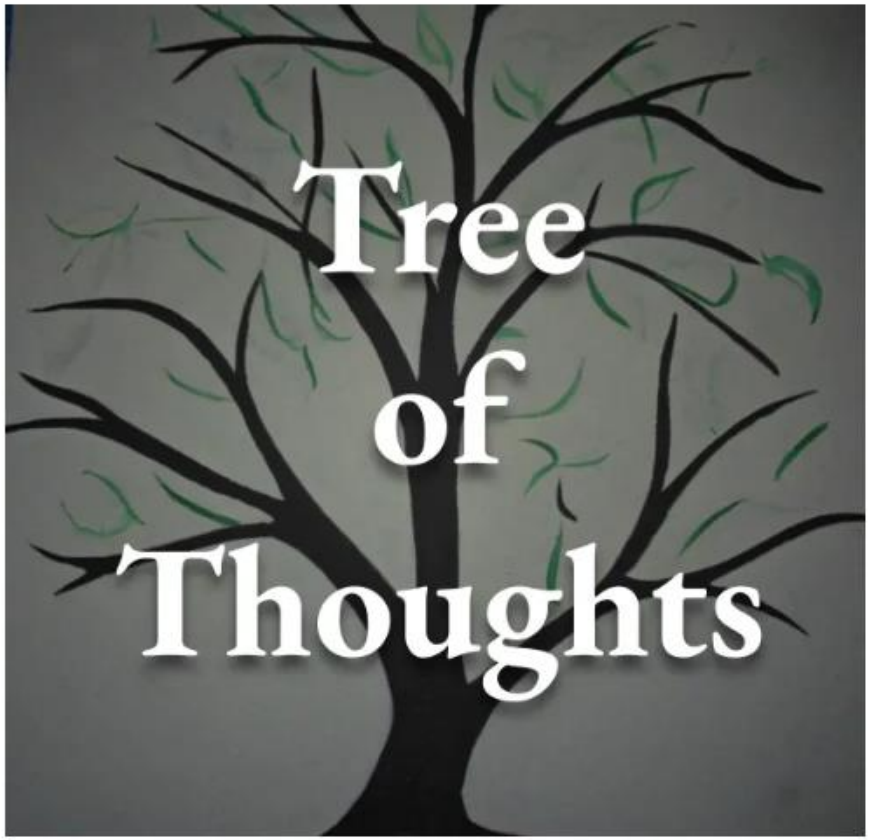
Even in a complex representation like this one,
The LLM was often able to get the numbers and their meanings (indicators dimension)
More powerful LLM = Higher accuracy (maybe a LLM fine-tuned for the task is the best solution)



Automated information is therefore much easier nowadays with Large Language Models. For example retrieving statistical information from free text (optimized for human reading rather than computing) used to be possible only through slow and tedious human processing. One of the major costs of data production. Now it can be done with Large Language Models.

Tree of Thoughts: An Improvement of Chain of Thoughts (Paper Review)

gArtist · Follow
Published in Better Programming · 7 min read · May 26



Background: Generated by Stable Diffusion 2.1.

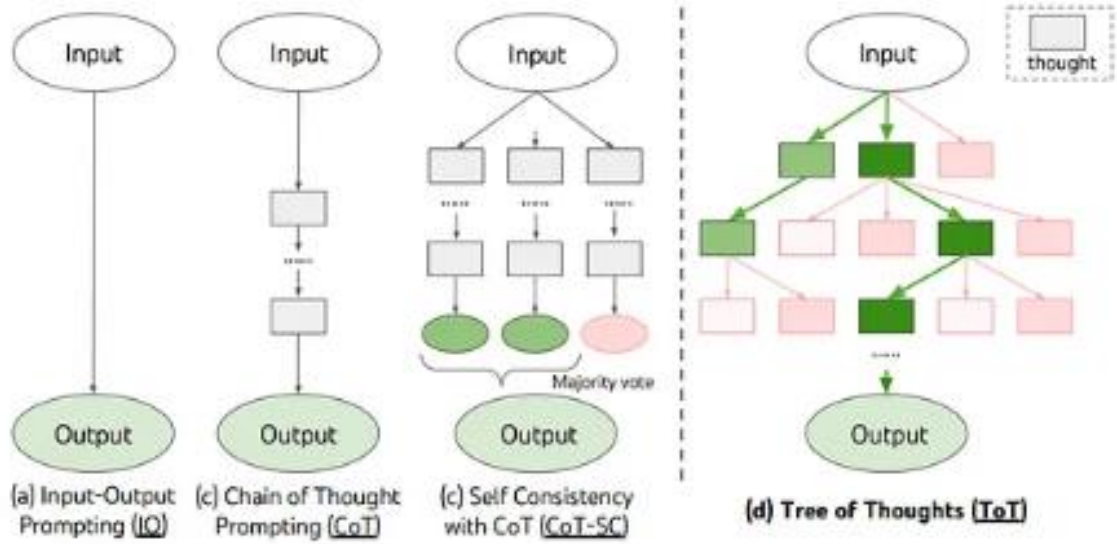


Figure 1: Schematic illustrating various approaches to problem solving with LLMs. Each rectangle box represents a *thought*, which is a coherent language sequence that serves as an intermediate step toward problem solving. See concrete examples of how thoughts are generated, evaluated, and searched in Figures 2,4,6.

Source: https://betterprogramming.pub/tree-of-thoughts-an-improvement-o

The data confidentiality problem with LLMS: what are the possible solutions?

Most official statistics and most economic analysis and policy documents are already free and public = = > No real data confidentiality issues

- ✓ However some documents contain information that we don't want to send to third parties.

Solution 1: Running 'distilled' large language models on PC or laptop.

-> Several techniques exist to reduce the memory requirement of LLMs to run on a single PC but they come at the cost of loss of precision or specialisation in a single task.

Solution 2: Anonymisation:

-> Make the data go through a module that hide all private and identifiable information in a reversible way, then send the anonymized text to LLM.

Solution 3: is to run a LLM on your own servers.

-> Some large organisations can afford that.

The data confidentiality problem with LLMS: what are the possible solutions?

Solution 4: Ensemble methods (1)

In machine learning, ensemble methods refer to techniques that combine multiple models to obtain better predictive performance (for a given cost/ risk level) than could be obtained from any of the constituent models alone. The primary motivation behind ensemble methods is that, by combining multiple models, one can leverage the strengths and compensate for the weaknesses of individual models. The idea is that while each model might make different errors or have various biases, by appropriately combining their predictions, these errors and biases can be averaged out or reduced.

Bagging (Bootstrap Aggregating):

Involves training multiple instances of the same classifier on different subsets of the training data. The subsets are obtained by sampling the training data with replacement.
Predictions from all the models are averaged (for regression) or taken by majority vote (for classification).
The most famous instance of bagging is the Random Forest algorithm, where decision trees are used as the base classifiers.

The data confidentiality problem with LLMS: what are the possible solutions?

Solution 4: Ensemble methods (2)

Boosting:
The idea is to correct the errors of previous models iteratively. Each new model tries to correct the errors made by its predecessor. Models are given weights based on their accuracy and the final prediction is a weighted sum (or vote) of all the models.
Popular algorithms include AdaBoost, Gradient Boosting Machine (GBM), XGBoost, and LightGBM.

Stacking (Stacked Generalization):
Multiple different models (could be of different types) are trained on the training set. A meta-model (or meta-learner) is then trained on the predictions (and possibly the original features) of these models to make the final prediction. For instance, you might have predictions from a linear regression, decision tree, and SVM. A logistic regression (the meta-model) then learns the best way to combine these predictions.

The data confidentiality problem with LLMS: what are the possible solutions?

Solution 4: Ensemble methods (3)

Voting:
Several models are trained, and their predictions are combined by majority voting (for classification) or averaging (for regression).
Hard Voting: Predictions are taken by majority vote.
Soft Voting: Probabilities are averaged, and the class with the highest probability is chosen.

Bayesian Model Averaging (BMA) and Bayesian Model Combination (BMC):
In BMA, the predictions from different models are combined based on their posterior probabilities. In BMC, rather than averaging over models, new models are created by sampling from the space of possible model combinations.

Ongoing data science work at ACS and at the Africa Regional Hub on Big Data and Data Science

Launch of the website of the United Nations Regional Hub for Big Data and Data Science: In collaboration with the National Statistics Institute of Rwanda and ONS UK, ACS has launched the website of the United Nations Regional Hub for Big Data and Data Science on 24 August 2023. This platform facilitates cross-border collaboration on projects that apply big data and data science to complement official statistics, provide knowledge on newly developed methods, algorithms, and tools, and offer training in the use of big data and data science for the community of official statisticians in Africa.

Publication of the first report of the survey on assessing the readiness of African National Statistical Offices (NSOs) in using big data and data science in the production of official statistics: In collaboration with the national Statistics Institute of Rwanda and ONS UK, ACS has published the report of a survey on assessing the readiness of African National Statistical Offices (NSOs) in using big data and data science in the production of official statistics. The questionnaire was designed to update the United Nations Big Data Regional Hub for Africa on the readiness of National Statistical Offices (NSOs) to use Big Data for improving Official Statistics.

Ongoing data science work at ACS and at the Africa Regional Hub on Big Data and Data Science (2)

Webinar on Web scrapping for price data: On 13 July 2023, ACS, in collaboration with the World Bank, the National Statistics Institute of Rwanda and ONS UK, and through the UN Regional Hub for Africa for Big data and Data science organized a webinar to train experts from African countries on the use of web scraping to collect data for the production of the Consumer Price Index.

Launch of a project on "Using big data to estimate the spatial distribution of poverty, well-being, and other socio-economic and environmental indicators to support SDGs monitoring, climate action and the beyond GDP agenda": ACS has launched a project called "Using big data to estimate the spatial distribution of poverty, well-being, and other socio-economic and environmental indicators to support SDGs monitoring, climate action and the beyond GDP agenda". The objectives of the project are to increase the capacity of African national statistical systems to produce spatial distribution of poverty, well-being and other socio-economic and environmental indicators from satellite imagery and other alternative data sources.

Ongoing data science work at ACS and at the Africa Regional Hub on Big Data and Data Science (3)

Webinar to present the first results of the use of satellite imagery to produce spatial estimations of poverty - October 2023: The ACS data science team has been working on refining the methodology, aligning it with the satellite maps that are accessible for various African nations. Within this context, Namibia has been selected as the exemplary pilot country to empirically test the feasibility and efficacy of this approach. Scholars and stakeholders will have the opportunity to gain insights into the preliminary outcomes of this endeavor during an upcoming academic webinar.

Webinar on Assessing NSO Capacity for Using Big Data and Administrative Records for Official Statistics: November 2023: In collaboration with the US Census Bureau, the National Statistics Institute of Rwanda and ONS UK, and through the UN Regional Hub for Africa for Big data and Data science, ACS will organize a webinar Assessing NSO Capacity for Using Big Data and Administrative Records for Official Statistics in November 2023. The main goal of this workshop will be to share country experiences, assess the needs and capacities of NSOs in using big data and administrative records, and develop an action plan towards incorporating big data and administrative data in official statistics.

Leverage technology and your expertise to develop innovative solutions for a more sustainable and equitable future

Join us on this journey of innovation, exploration, and problem-solving as we harness the power of data to create a more sustainable and resilient world.

Datathon participants will develop innovative data-driven applications, tools or statistical models combining geospatial data with other data sources to help advance the implementation of the Sustainable Development Goals.

REGISTER TODAY!
DEADLINE 30 SEP 2023

Thank you for your attention!