MCIT
Modeling, control &
information technologies

Generative AI text summarization performance analysis prospects

https://doi.org/10.31713/MCIT.2024.037

Oleksandr Tsypliak
Computer Science PhD student,
G.E. Pukhov Institute for Modelling in Energy Engineering
Kyiv, Ukraine
saifbant@gmail.com

Volodymyr Artemchuk
Doctor of Technical Sciences, Senior Researcher,
G.E. Pukhov Institute for Modelling in Energy Engineering
Kyiv, Ukraine
ak24avo@gmail.com

***Abstract* – AI investment world trends analysis in scope based on venture investments timeline by domain visualization and researches proves that not all AI-agent development projects are successful. The current paper describes an AI-agent performance analysis approach based on LLM-specific metrics and scenario-specific metrics. The AI-agent for highlights generation is in focus as an object of performance analysis.**

***Keywords* – large language model, performance analysis, metrics, perfroamnce optimization, generative artificial inteligence, LLAMA 2.**

I. Introduction

Generative artificial intelligence chatbots based on LLM (large language model) popularity increased awareness about its usage possible advances among stakeholders all over the world.

Academic researches [1] emphasize on prospects of its application for data processing related automations. When social media [2] call it “anything tool” and discuss its perspectives on different human work automation.

On the other hand, following Organization for Economic Cooperation and Development (OECD) AI statistics about worldwide venture investments in AI the peak was in 2021 and it covered all investigate domains including [3]:

- Education and training
- Digital security
- Logistics, wholesale and retail
- Robots, sensors, IT hardware
- Financial and insurance services
- Business processes and support services
- Media, social platforms, marketing
- IT infrastructure and hosting
- Healthcare, drugs and biotechnology
- Mobility and autonomous vehicles
- “Other” as separate category uncovered above

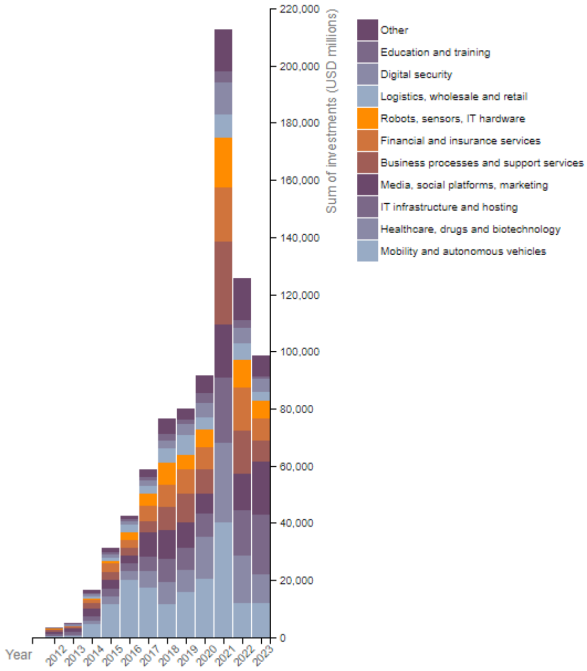

Figure 1. Venture capital investmens into AI by domains years trend in million dollars [3]

Since 2022 worldwide venture investments volume decreased, which can be partially explained by the ongoing Russo-Ukrainian War. But the trend of 2023 showed that all domains except “Media, social platforms, marketing” and “IT infrastructure and hosting” are getting less and less investments getting back to the level of 2020. Worth mentioning that yearly venture investments by years is a high-level metric, specifying only investments into new businesses and projects.

Considering the high complexity and the risks of each software development project involving LLM usage [4] possible key root cause of the investments volume decrease, a consequence of lower interest in starting new projects, may be back in 2021 initiated projects failures.

Among other reasons, there is a problem of the performance evaluation of the selected LLM used inside the designed AI agent application. There are four aspects of performance in this context:

- Ability to resolve specific task
- Ability to follow respond in required time boundaries
- Ability to be hosted in bounds of pre-calculated hardware resources
- Ability to serve the required number of users following functional and non-functional system requirements

The current paper is focused on the problem of an autonomous LLM as a component of AI agent performance metrical measurement in the context of a single specific prototype created for text summarization.

II. Detailed problem definition

As an example of an AI-agent for performance analysis let's consider REST API behind the application for highlights generation described in detail in the separate article [5].

From the architecture point of view, the application is a wrapper around the self-hosted LLAMA 2 model. Briefly about its functionality:

1. Agent cuts the full text of the application into paragraphs
2. Each paragraph text got sent to LLM with an inbuild prompt to LLAMA 2
3. The output of the prompt got analyzed using text algorism and sent back to LLAMA 2 with another prompt
4. Steps 2 – 3 repeated for the chain of prompts
5. Final output got recorded into the document with the original full text of the article as a highlight

LLM performance analysis in context of this prototype has 3 dimensions of metrics related to:

1. Model ability to serve specific generalized tasks
2. Scenario-focused functionality – evaluating the quality of the output
3. Hardware resource consumption by self-hosted LLM
4. Multithread usage of the agent by multiple users simultaneously

III. Research results

Let's review some of the most relevant approaches to each of these dimensions that are most relevant to the mentioned AI agent and design several more specific metrics.

A. Model ability to serve specific generalized tasks

Unlike early limited neural language models (NLMs) modern transformer-based neural language models contain many more parameters and are pretrained on the large amounts of data. These modern task-agnostic systems are referenced as LLMs (large language models). [6]

The problem with LLMs is that technically each of them can be used to resolve any text generation/analysis task. But a rich variety of available models has different architecture and different strong and weak sides. There are already academically described multiple examples of AI-agent development projects that failed because of wrong model selection. [7]

So, on the system design level the most suitable one should be selected based general tasks list inherited from the functional and non-functional requirements of the system.

In the context of highlights generating application, we can specify LLM capabilities used to resolve this task:

- Comprehension related
  - Text summarization
  - Text simplification
  - Reading comprehension
- Knowledge utilization
  - General world knowledge reference
  - Ability to use attached text resources
- Emergence related:
  - Common sense logical reasoning
  - In context learning

Understanding the scope of tasks to the cover optimal model can be selected by collecting metrics that will give each of the alternative LLMs a score. [8] To collect such metric steps below should be followed:

1. Select/design datasets relevant to the listed general tasks resolution
2. Apply data set to collect specific metrics in the context of each selected data set
3. Apply exactly same datasets and metrics collection techniques to alternative LLMs

There are publicly available "leaderboards" containing some most popular models in the context of metrics. For example, a board from "Artificial Analysis" platform [9]. However, using this data for research or commercial project is a high risk because the specification of process and raw data for calculating specific metric for specific model is not publicly available and consequently can't be fully trusted.

Moreover, each dataset unitization generates additional exclusive metrics that are not published at all for simplification.

Datasets most relevant to highlights generation AI-agent tasks mentioned above:

- MMLU [10] to assess general knowledge and problem-solving ability
- TriviaQA [11] to check Wikipedia knowledge high-quality access

- RACE [12] reading comprehension quality verification
- HellaSwag[13] accessing commonsense
- TruthfulQA [14] for common misunderstanding verification during text summarization and simplification

Most relevant metrics:

- GLUE score [15] for verification of logical reasoning using attached text and context learning
- ROUGE [16] summarization tasks evaluation
- Fluency – human-measured metric to measure clarity of the LLM answer

B. Scenario focused functionality

LLM utilization libraries allow to use alternative models once they are formatted properly without system redesign and even major configuration changes.

In the scope of scenario-based verification in the context of highlight generation AI-agent two aspects should be measured:

1. Answer generation time
2. Answer quality

For this purpose, should be designed a set of paragraphs with human written correct summaries. Initial should be created manually, possibly extended using potentially most suitable model and then verified by humans. Then use it for analysis:

1. Using an automated test each such paragraph should be summarized by the AI-agent using alternative LLMs.
2. Compare the quality of each answer using text algorithms compared to correct human-written one and measure processing time.
3. Use processing time in seconds of each answer in seconds by percentiles for aggregating statistics for all answers by specific LLM and the same for quality % percentile
4. Based on collected statistics select most suitable model.

C. Hardware resource consumption by self-hosted LLM

As mentioned in section B. run the AI-agent using an alternative model and using a hardware metrics monitoring agent like Telegraph to measure server-side performance metrics mentioned below. Store metrics in time series database like influx and the visualize the timeline using visualization tool like Grafana.

Metrics to measure that are mostly consumed by LLMs:

- Available CPU %
- Available RAM in megabytes
- Read/Write disc time % usage

Once monitoring is configured the test mentioned in section B can be executed and hardware metrics collected during periods of those tests can be compared as visualized timelines.

The model consuming the least RAM, CPU and disk time will be the best alternative as most efficient and cheap for hosting.

D. Multithread usage of the agent by multiple users simultaneously

Using the scenario designed in the scope of section B run a multithread test and monitor hardware metrics as explained in section C.

But for the current highlight generation task on the current stage of development that was impossible to perform, because a single threaded run occupies 100% of CPU time. As is parallel process run leads to failure which shows prospects of optimization.

IV. Conclusion

Venture investment statistics clearly demonstrated, that not all domains that started to utilize LLM-based AI were able to keep their investors.

Multiple researches show that not all domains AI utilization kept the phase of growing after GPT 3 first release. This makes the problem of LLM-based AI-agents performance analysis worth attention.

The analysis should be conducted:

1. On the stage of design, to determine the best alternative LLM using general tasks specific metrics
2. On the stage of user acceptance testing, to collect metrics relevant to the specific scenario of AI agent usage.

The result of the work is determined process of performance analysis in the context of the specific AI-agent designed for highlights generation.

Implementation of such analysis with raw data collection and description is in the scope of further publications.

References

[1] Pallathadka, H., E.H. Ramirez-Asis, T.P. Loli-Poma, K. Kaliyaperumal, R.J.M. Ventayen, and M. Naved. “Applications of Artificial Intelligence in Business Management, e-Commerce and Finance.” Materials Today: Proceedings 80 (2023): 2610–13. https://doi.org/10.1016/j.matpr.2021.06.419.

[2] Huang, Haomiao. “How ChatGPT Turned Generative AI into an ‘Anything Tool.’” Ars Technica, August 2023. https://arstechnica.com/ai/2023/08/how-chatgpt-turned-generative-ai-into-an-anything-tool/.

[3] Co-operation, Organisation for Economic and Development. “Total VC Investments in AI by Country and Industry,” 2024. https://oecd.ai/en/data.

[4] Tsypliak O.O., Artemchuk V.O. Risks Using LLM-Based Platforms in Energy Digitalization Context. Energy security in the digital transformation era, V scientific-practical conference of the G.E. Pukhov Institute for Modeling in Energy Engineering National Academy of Sciences of Ukraine : materials (Kyiv, November 22, 2023). Kyiv: PIMEE NAS of Ukraine, 2023. p. 123-125.

[5] Tsypliak, Oleksandr, and Volodymyr Artemchuk. “Console Application Development for Articles` Highlights Generation

Based on Artificial Intelligence Designed Using Autonomous Large Language Model." In Information Technology for Education, Science, and Technics, 53–64. Springer Nature Switzerland, 2024. https://doi.org/10.1007/978-3-031-71801-4_5

[6] . Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep Learning Based Text Classification: A Comprehensive Review." arXiv, 2020. https://doi.org/10.48550/ARXIV.2004.03705.

[7] Fotheringham, D., and M.A. Wiles. "The Effect of Implementing Chatbot Customer Service on Stock Returns: An Event Study Analysis." Journal of the Academy of Marketing Science 51, no. 4 (2023): 802–22. https://doi.org/10.1007/s11747-022-00841-2.

[8] Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. "Large Language Models: A Survey." arXiv, 2024. https://doi.org/10.48550/ARXIV.2402.06196.

[9] Platform, Artificial Analysis. "LLM Leaderboard - Compare GPT-4o, Llama 3, Mistral, Gemini & Other Models \textbar Artificial Analysis," 2024. https://artificialanalysis.ai/leaderboards/models.

[10] Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. "Measuring Massive Multitask Language Understanding." arXiv, 2020. https://doi.org/10.48550/ARXIV.2009.03300.

[11] Joshi, Mandar, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension." arXiv, 2017. https://doi.org/10.48550/ARXIV.1705.03551.

[12] Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. "RACE: Large-Scale ReAding Comprehension Dataset From Examinations." arXiv, 2017. https://doi.org/10.48550/ARXIV.1704.04683.

[13] Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "HellaSwag: Can a Machine Really Finish Your Sentence?" arXiv, 2019. https://doi.org/10.48550/ARXIV.1905.07830.

[14] Lin, Stephanie, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." arXiv, 2021. https://doi.org/10.48550/ARXIV.2109.07958.

[15] Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." arXiv, 2018. https://doi.org/10.48550/ARXIV.1804.07461.